

Linköping Studies in Science and Technology. Dissertations
No. 1102

The Origin–Destination Matrix Estimation Problem — Analysis and Computations

Anders Peterson



Linköping University
INSTITUTE OF TECHNOLOGY

Department of Science and Technology
Linköpings universitet, SE-601 74 Norrköping, Sweden

Norrköping 2007

Linköping Studies in Science and Technology. Dissertations, No. 1102

The Origin–Destination Matrix Estimation Problem

— Analysis and Computations

Anders Peterson andpe@itn.liu.se <http://www.itn.liu.se>

The cover illustrates the equipment for collecting link flow observations. Here, the traffic entering and leaving the car park at the commercial Ekholmen centrum in Linköping, Sweden, is being counted. Photo: April 8, 2007, by Inger Munkhammar.

ISBN 978-91-85831-95-1 ISSN 0345-7524

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-8859>

Copyright © 2007, Anders Peterson, unless otherwise noted.

Printed by LiU-Tryck, Linköping, Sweden, 2007.

Acknowledgements

Most important for the performance of a post-graduate study are the supervisors, and I sincerely acknowledge Jan Lundgren and Torbjörn Larsson for advice and support over the years. Jan has introduced me to traffic modeling and especially the Origin–Destination matrix estimation problem. He has been a good advisor by all occasions and I am very grateful for his help in defining and structuring the research work.

Torbjörn has guided me through the process of scientific writing and re-writing. He has patiently offered a lot of time for the details and the nuances in the thesis.

I thank all colleagues at the division of Communications and Transport systems, for discussions, inspiration and companionship. Henrik Andersson kindly read and commented my work. Clas Rydberg has been a good fellow in research, teaching — and train commuting. Carl Henrik Häll, with whom I used to share my office room, has always had time for a word or two whenever the research process has troubled me.

The benevolent people at the division of Optimization in Linköping, where my post-graduate studies once started, are acknowledged for hosting many research meetings.

Finally, I would like to express my greatest gratitude to my family and my friends. I have appreciated their encouragement and understanding, especially in situations where I did not manage to fully explain the problems troubling me.

Norrköping, May 2007
Anders Peterson

Abstract

For most kind of analyses in the field of traffic planning, there is a need for origin-destination (OD) matrices, which specify the travel demands between the origin and destination nodes in the network. This thesis concerns the *OD-matrix estimation problem*, that is, the calculation of OD-matrices using observed link flows. Both time-independent and time-dependent models are considered, and we also study the placement of link flow detectors.

Many methods have been suggested for OD-matrix estimation in time-independent models, which describe an average traffic situation. We assume a user equilibrium to hold for the link flows in the network and recognize a bilevel structure of the estimation problem. A descent heuristic is proposed, in which special attention is given to the issue of calculating the change of a link flow with respect to a change of the travel demand in a certain pair of origin and destination nodes.

When a time-dimension is considered, the estimation problem becomes more complex. Besides the problem of distributing the travel demand onto routes, the flow propagation in time and space must also be handled. The time-dependent OD-matrix estimation problem is the subject for two studies. The first is a case study, where the conventional estimation technique is improved through introducing pre-adjustment schemes, which exploit the structure of the information contained in the OD-matrix and the link flow observations. In the second study, an algorithm for time-independent estimation is extended to the time-dependent case and tested for a network from Stockholm, Sweden.

Finally, we study the underlying problem of finding those links where traffic flow observations are to be performed, in order to ensure the best possible quality of the estimated OD-matrix. There are different ways of quantifying a common goal to cover as much traffic as possible, and we create an experimental framework in which they can be evaluated. Presupposing that consistent flow observations from all the links in the network yields the best estimate of the OD-matrix, the lack of observations from some links results in a relaxation of the estimation problem, and a poorer estimate. We formulate the problem to place link flow detectors as to achieve the least relaxation with a limited number of detectors.

Populärvetenskaplig sammanfattning

Matematiska modeller används av väghållare runt om i världen för att beskriva, planera och styra trafiken. Det kan handla om att beräkna restider, miljöpåverkan och tillgänglighet, eller att prognostisera effekterna av nybyggnationer, ändrad framkomlighet och vägavgifter. Ofta räcker det med att beskriva trafiken med genomsnittsförhållanden, men för att exempelvis producera realtidsinformation, kontrollera kövarningssystem och skapa handlingsplaner för olika olycksscenarier behövs tidsberoende modeller.

Noggranna uppgifter om antal resande från varje startpunkt (eng. origin) till varje målpunkt (eng. destination) i ett trafiknät är viktiga indata till de flesta av dessa modeller, och de brukar organiseras i s.k. OD-matrizer (eng. Origin–Destination matrices). Att utarbeta OD-matrizer utifrån resvaneundersökningar och olika statistiska sammanställningar är dyrbart och görs sällan. Observerade trafikflöden, som samlas in genom slangmätningar, vägkameror, tullportaler, etc., utgör däremot billig och lättillgänglig information om den rådande trafiksituationen. Denna avhandling behandlar problemet att skatta OD-matrizer utifrån trafikflödesobservationer.

Den skattade OD-matrisen ska tillsammans med den ruttvalsprincip som antagits för trafiken återskapa de observerade trafikflödena så bra som möjligt. Det finns olika principer för att beskriva trafikanternas ruttväl. Vanligtvis antar man att varje resenär väljer den väg som tar kortast tid utifrån den trängsel och de körtider som råder i den givna trafiksituationen. När OD-matrisen ändras kommer trafiksituationen också att ändras, vilket påverkar körtider och ruttväl. Ju hårdare trafikerat gatunätet är, desto större blir förändringarna. I en tidsberoende modell måste dessutom flödet av trafik återges korrekt i både tid och rum.

Att förutse hur ruttväl och flödesfortplantning beror av OD-matrisen är svårt och kräver omfattande analyser av hur känslig en given trafiksituation är för förändringar i resandet. I den här avhandlingen utvecklas och utvärderas olika sätt att modellera detta beroende.

En intressant fråga är var i gatunätet flödesmätningar ska utföras för att ge så hel täckande information som möjligt av trafiken. I avhandlingen jämförs olika placeringar med avseende på hur väl man lyckas skatta OD-matrisen och vi kan dra generella slutsatser om vilka placeringsstrategier som bör användas.

Avhandlingen innehåller både generella beräkningsrutiner, som i framtiden kan integreras i programvara, och mer specifika metoder, som redan kommit till nytta, bl.a. för att skatta tidsuppdelade OD-matrizer för Göteborgs-området. De metoder som utvecklats är intressanta både för dem som tar fram den programvara som används inom branschen och för specifika frågeställningar hos enskilda väghållare, dvs. i Sverige främst Vägverket och de större kommunerna.

Introduction and Overview

1 Background

The number of cars increased heavily in the decades following World War II. During the 1950s the number of registered passenger cars in Sweden was more than quadrupled and ended with 1.1 Million (Statistics Sweden, 2007). The situation was similar in all western countries and opened for the use of mathematical models in traffic planning: How can the available infrastructure be used in the most efficient way and what investments would give the best effects? These questions are still fundamental in traffic planning, though the valuation of accessibility, congestion, emission, travel time, etc. have changed.

To be able to evaluate different engineering alternatives, there is a need for mathematical models. The outcome of such models is important support for making decisions on how links in the traffic network should be built or rebuilt, and, in a broader sense, how the city should be arranged with new commercial, industrial and living areas. Except for estimating the utility of building new roads, traffic models can be useful for evaluating the effects of for example changing speed limit and capacity (number of lanes), introducing road tolls and re-designing the intersections (turning lanes, signals, roundabouts, etc.). Another application is the development of plans of actions for taking care of traffic interruptions that are caused by incidents.

A fundamental issue in this thesis is the mathematical modeling of route choices for vehicles in a congested road network, which we recognize as *the traffic assignment problem*. Models for uncogested road networks, as well as models including alternative travel modes, are not covered in this thesis.

Mathematical modeling of traffic requires a lot of data and other information about the road network and the travel demand. The intended user of the models and methods, that are discussed in the thesis, is the road administrator for a medium or large sized city. For Swedish conditions we identify the Swedish National Road Administration (Vägverket), which is responsible for the major national road network, and the municipal traffic administrators in the larger citites as important users.

The accuracy of the modeled traffic situation depends on the quality of the available information, and how this data is combined and weighted from different sources. The travel demand is a key component and nearly every traffic model requires a tableau

specifying the travel demand between different places in the network. Such a tableau is called an *Origin–Destination matrix*, or *OD-matrix* for short; synonymously used terms are *trip table* or *(origin–destination) trip matrix*. This thesis is devoted to the problem of estimating reliable OD-matrices, in a reliable way.

A major distinction between the different types of traffic models is drawn with respect to their level of detail. A *macroscopic model* uses fluid variables, such as flow and density, and does not model individual vehicles — these are aggregated into continuous variables. A *microscopic model* describes vehicles (and often even drivers) individually. A *mesoscopic model* is in between and combines the ideas from macroscopic and microscopic models. Typically it uses macroscopic speed–flow relations to depict the motions of individual vehicles.

Macroscopic models are traditionally used for planning in larger networks over a longer time period: How should the city network be developed in the coming 5–10 years, with respect to some assumption of population growth in different sub-areas of the city? A microscopic model on the other hand is used for a smaller network and the result is used for more specific measures: How can the available space be allocated to lanes in the best way, to handle a troublesome intersection or sequence of intersections? A microscopic model often uses data from a macroscopic model as input, for example by stating an average situation. A mesoscopic model can be used to catch the overall changes in the traffic induced by some detailed changes of the infrastructure.

Another distinction is made between *time-dependent* (dynamic) and *time-independent* (static) models. For recovering how a traffic scenario is developed over time, we need a dynamic model, which can reproduce the reaction of the traffic to a current situation. A time-independent model can be described as a steady state in a dynamic model, i.e. a situation where reactions and contra-reactions are balanced. Time-independent models give average descriptions of the traffic situations. They require less input data, and well-established analytical models exist. For time-dependent models, the input data must provide more details on the traffic situation and we must make assumptions on how the traffic flow propagates in both time and space. Until today most time-dependent models are based on simulation. Traditionally, macroscopic models are time-independent, whereas we need a smaller and more detailed network of microscopic type to analyze time-dependent effects.

In this thesis both time-dependent and time-independent traffic models are discussed. We start in the next Section with an overview of macroscopic models for the time-independent case, by introducing the traditional *four-stage model*. We also briefly describe the problems which arise when a time-dimension is introduced. Thereafter, we introduce the OD-matrix estimation problem, which is the subject of this thesis. This problem is considered for the time-independent case and the time-dependent case in Section 3 and Section 4, respectively. Finally, in Section 5, the contribution of this thesis is presented. We discuss the purpose and motivation for the work presented and give a summary of the five annotated papers.

2 The four-stage model

The traffic planning process traditionally follows four sequential stages: trip generation, trip distribution, modal split and traffic assignment. *The four-stage model* was originally developed during the 1950s and 1960s for the planning of major highway facilities (Papacostas and Prevedouros, 2001). Soon, however, the model was applied also in other traffic planning situations and recognized as a standard for macroscopic modeling. Many software tools for traffic planning are still based on the ideas from the four-stage model.

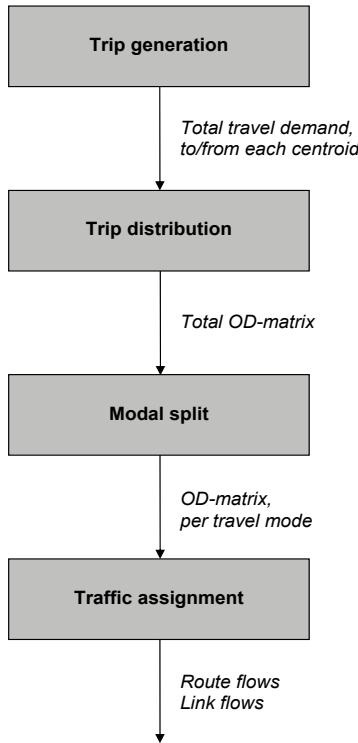


Figure 1: The four-stage model in its basic form.

Depending on the situation, some stages might not be applicable; for example, if no alternative travel modes are available. Over the years several alternative and/or extended planning schemes have been presented and some of them are implemented as options in the available software tools. Typical alternatives are to combine two or more stages, solve stages in reversed order, or to compute them iteratively for higher accuracy. Especially the second stage, deciding the distribution of travel

demand between origins and destinations, and the third stage, where the split onto different travel modes are computed, are often performed together as one stage. In the following the basic concept of the four-stage model, which is illustrated in Figure 1, will be presented to give a background to the OD-matrix estimation problem and its model setting. At the end of the section, we will briefly time-dependence as an extension to the model concept.

Many authors have already described the four-stage model. The following presentation has been inspired by the books of Wilson (1974), Patriksson (1994), Khisty and Lall (1998), Wright and Ashford (1998), Garber and Hoel (1999), Ortúzar and Willumsen (2001), and Papacostas and Prevedouros (2001), as well as the history by Bates (2000) and the thesis by Lundgren (1989).

2.1 Trip generation

The first of the four stages is the *trip generation*. The aim of this procedure is to determine how many trips there will be originating (trip production) or terminating (trip attraction) at each zone in the network. The size of these zones must be defined with an appropriate accuracy with respect to the purpose of the traffic model, and could range from some blocks in a city center up to a complete conurbation in a model for intercity travel demand forecasts. Each zone is represented by a single node in the model, which we will refer to as a *centroid*. For performing this stage a broad variety of *survey data* is collected, concerning characteristics of the trip makers for each centroid, such as age, sex, income, auto ownership, trip-rate, land-use, and travel mode. In general it is a major project to collect the required data and it is therefore advantageous to coordinate the investigations to a specific base year, or, synonymously, a *target year*. In Sweden the database “Folk- och Bostadsräkningen”, carried out by the Swedish authority for statistics, has been a valuable source. The latest “Folk- och Bostadsräkningen” was performed in 1990 (Statistics Sweden, 2007).

Depending on the purpose, it is possible to use different descriptions of the travel demand for different categories, trip purposes, travel modes, and time periods (time of day, day in week, week in season, etc.). Methodologically the two most frequently used techniques for trip generation analysis are cross-classification analysis (or category analysis) and multiple regression analysis. In the latter method values are aggregated for each centroid, whereas the central assumption in the cross-classification analysis is to disaggregate the demand to each household and calculate its generated trips by some pre-defined quota concerning the above given parameter types. It seems that most researchers today have agreed on the cross-classification analysis, essentially because it is based on census data (a priori given quota parameters). As a consequence, it is easy to compare and transfer models between cities, and the target year matrix can be used for validation, which is not the case in a regression analysis based model where the quota parameters depend on the data.

To conclude the description of the trip generation stage: Let P and Q denote the sets

of origin and destination centroids respectively. The output of the trip generation procedure is the sum of trips starting at all origins, o_p , $p \in P$, and the sum of trips ending at all destinations, d_q , $q \in Q$. (In most cases all centroids both produce and attract trips, which means that $P = Q$.)

2.2 Trip distribution

The second stage is the *trip distribution*. In this stage the generated sums of trips starting and ending at the centroids are connected to each other to form *travel demands* (OD-demands, OD-flows, or trip interchanges), for the OD-pairs. The aim of this stage is to find a trip distribution $g = \{g_{pq}\}$, $(p, q) \in P \times Q$ such that the aggregated information from the trip generation stage holds, i.e. such that $o_p = \sum_{q \in Q} g_{pq}$, $p \in P$ and $d_q = \sum_{p \in P} g_{pq}$, $q \in Q$. The trip distribution procedure is traditionally performed with some sort of *gravity model*. Alternatively, growth factor models and logit models can be used.

The name gravity model comes from Newton's law of gravitation, which states that the force of attraction between two bodies is directly proportional to the product of the masses of the two bodies and inversely proportional to the square of the distance between them. In the traffic distribution case it is assumed that the travel demand in OD-pair (p, q) is directly proportional to the trip ends production at the origin node, o_p , times the trip ends attraction at the destination node, d_q , weighted with a *deterrence function* (friction function), $f(\pi_{pq})$, of the travel impedance (time, distance, cost, etc.) between the two centroids. The travel demand from p to q can thus be expressed as

$$g_{pq} = k o_p d_q f(\pi_{pq}), \quad (1)$$

where k is some suitable weighting parameter. The deterrence $f(\pi_{pq})$ should be monotonically decreasing function of the travel impedance, π_{pq} , between p and q . The set $\Pi = \{\pi_{pq}\}$, $(p, q) \in P \times Q$ is known as the *skim table* (Papacostas and Prevedouros, 2001). Commonly a polynomial expression of the form $f(\pi_{pq}) = \pi_{pq}^{-\beta}$ is used. Here $\beta \geq 0$ is a parameter to be calibrated. (By setting $\beta = 2$, equation (1) states Newton's law of gravitation.) Alternatively, an exponential expression of the form $f(\pi_{pq}) = e^{-\beta\pi_{pq}}$ can be used.

In some models, it is also accounted for the relative attractiveness (accessibility) for each destination and the individual socioeconomic deterrence, k_{pq} , for each OD-pair. This leads to the following reformulation of equation (1):

$$g_{pq} = o_p \left(\frac{d_q f(\pi_{pq}) k_{pq}}{\sum_{q \in Q} d_q f(\pi_{pq}) k_{pq}} \right).$$

The bracketed expression can be interpreted as the probability that a trip originating at p will terminate at destination q .

There are many ways to calibrate the parameters required in the trip distribution procedure. Typically this is performed in an iterative process until a matrix close enough to the target year matrix is reproduced from the survey data. Most of these methods are of a heuristical nature; a common simplification is to estimate the value of each $f_{pq} = f(\pi_{pq})$ directly, instead of defining and calibrating the deterrence function.

For further reading on the use of gravity models in connection with transportation analysis, the reader is referred to the survey by Erlander and Stewart (1990).

The growth factor models are rather rough and cannot capture time-of-the-day variations. Instead, it is typically assumed that the travel demand is equal in both direction for all OD-pairs, i.e. that $g_{pq} = g_{qp}$, $p \in P$, $q \in Q$, and $P = Q$ (each centroid both produces and attracts trips).

To conclude the description of the trip distribution stage: Given the generated and attracted number of trips at each centroid, o_p and d_q , respectively, from the trip generation stage, the trip distribution procedure distributes the generated and attracted sums of trips onto travel demand g_{pq} , $(p, q) \in P \times Q$, such that $o_p = \sum_{q \in Q} g_{pq}$, $p \in P$ and $d_q = \sum_{p \in P} g_{pq}$, $q \in Q$. This distribution is typically performed with a gravity or a growth factor model.

2.3 Modal split

The third stage is the *modal split*. In the mode split (or mode choice) procedure the travel demand for each OD-pair is partitioned into different travel modes. In the simplest case there are only two travel modes available: private car and transit, but the travel modes can also be specified in more detail, for example “sharing car with another person” might be a separate mode. The transit travelers can be classified into different sub-modes according to how they get to the bus stop or railway station (walk, bicycle, car, etc.). In some situations also the purpose of the trip (“home-to-work”, “work-to-kindergarten”, etc.) is considered, since, for example, it seems more likely that a person would choose to travel to his work with the transit system, but prefer a private car, if available, for social trips. Further, the purpose of the trip can be important since it might affect the acceptance for a delay or route guidance information. Besides the consideration of available modes and trip purpose, some models also consider the socioeconomic status of the trip-maker. For some persons, the travel time is more important than the travel cost, whereas the situation is the opposite for some other.

To determine how to disaggregate the OD-matrix into different travel modes, the *utility* of each mode must be calculated. The utility is a weighted sum of different attributes, like for example travel time, cost and comfort. When analyzing public

transport systems, parameters as walking time to transit line stop, waiting time, in-vehicle time, transit line frequency, transfer, and/or transfer waiting time can be included (Sjöstrand, 2001).

Let x_{hk} be the measured value of attribute $h \in H$ for travel mode $k \in K$ and let α_h , $h \in H$ be the corresponding weighting parameter. The weighted sum of all measured values, denoted by v_k , together with a random error term, ε_k , states the utility for travel mode $k \in K$:

$$u_k = v_k + \varepsilon_k = \sum_{h \in H} \alpha_h x_{hk} + \varepsilon_k. \quad (2)$$

The random error term expresses other attributes of travel mode k than those captured in H , the overall uncertainty of the measured values and also the variability in preferences among the individuals. Equation (2) is sometimes referred to as the *utility function*, which states the utility of travel mode k . We should mention that it is possible to specify a utility function separately for each centroid or OD-pair. Such a disaggregation can be used to capture effects that a trip maker might be more likely to use the transit system when traveling into the city center, with a shortage of parking places, than when traveling to another destination.

The utility for a travel mode is used to calculate the probability, p_k , that a certain traveler chooses travel mode $k \in K$, or, more precisely, the probability that a certain traveler perceives the highest utility by choosing that travel mode. In most applications of utility functions, the error term in equation (2) is assumed to be Gumbel (or Weibull) distributed, whereby the probability can be calculated through a *logit model*. The simplest case, where the error terms are independent and have equal variance, is the *multinomial logit model*, in which

$$p_k = \frac{e^{\mu v_k}}{\sum_{k \in K} e^{\mu v_k}}, \quad (3)$$

where v_k denotes the weighted sum measured values for travel mode $k \in K$, as defined in equation (2), and $\mu > 0$ is a scale parameter. A derivation of (3) was first proposed by McFadden (1973) and a complete derivation is presented by Domencich and McFadden (1975).

In the simple form of the logit model, i.e. in equation (3), the similarity between the travel modes, and thus the definition of the travel mode classes, is very important. Since the utility will be used in relation to the alternatives, the alternatives must be significantly different. Suppose, for example, that car, bus and tram are the three available travel modes in a system and suppose all of them have the same utility. By letting K consist of these three modes, the probability that a traveler chooses car will be 1/3. On the other hand, if bus and tram together are considered as “transit”, this probability increases to 1/2.

To overcome this unwanted property the travel modes are often ordered in a hier-

archy, where the split of the travel demand on each level, has its own probability distribution. The probability for “tram” in the example above would be calculated as the product of the probability for “transit” and the probability for “tram, given transit”. This, so-called *nested logit model* can of course have more than two levels, for example when, in addition to the type of “transit”, also the way to access the transit network separates the travel modes. For further reading on nested logit models, see for example Ben-Akiva and Lerman (1985) and Oppenheim (1995).

Given the probability for a certain mode, it is easy to split the OD-matrix. The probability that a certain traveler chooses a certain travel mode can be interpreted as the *proportion* of all travelers choosing that mode. Thus the OD-demand for travel mode $k \in K$ in OD-pair $(p, q) \in P \times Q$ can simply be evaluated as $g_{pq}^k = p_k g_{pq}$.

To conclude the description of the modal split stage: Given the travel demand for each OD-pair g_{pq} , the modal split procedure determines how this is disaggregated into different travel modes, $k \in K$, such that $g_{pq} = \sum_{k \in K} g_{pq}^k$, $(p, q) \in P \times Q$. Normally this is done by finding a split proportion p_k for each mode $k \in K$.

2.4 Traffic assignment

The fourth and final stage is the *traffic assignment*. In the traffic (or trip) assignment the OD-matrix for each mode is assigned onto the traffic network, according to some principle. The aim of this procedure is to calculate the link flow volumes.

A traffic network can be represented by (N, A) , which are the sets of its *nodes* and *links* respectively. Each node $n \in N$ is either a centroid or an intermediate node, modeling a network intersection. Thus the relation $P, Q \subseteq N$ holds. Each link $a \in A$ is either an actual street section, a transit connection or a generalized relation, for example an artificial “street” connecting a living area to the main network or symbolizing the walking path to the bus stop. All links are *directed*, and two-way streets are simply modeled as two separate links.

As indicated, each link must be a bearer of one or more travel modes and in the assignment procedure, this must be taken into account. A bus line, for example, is a link sequence itself, but each vehicle must also be assigned onto the street network, together with the private cars. In the rest of this thesis, we will only consider one travel mode, private cars. For notational convenience, the mode subscript k is therefore left out of the following presentation.

There are in general many possible *routes* (or synonymously, paths) from one node to another. Let R_{pq} denote the set of acyclic routes from $p \in P$ to $q \in Q$. The set R_{pq} is finite, but typically very large. Therefore the assignment procedure must follow some assumption on how the routes are chosen.

The most common assumption is that each traveler will choose a route with the least instantaneous travel impedance. The travel impedance, π_{pq} , was introduced in

the trip distribution stage as a generalized measurement of time, distance, cost, etc., between the two centroids $p \in P$ and $q \in Q$. The term *generalized cost* is a more accurate description of the travel impedance, which instantaneously is experienced on a certain link, and is denoted by $c_a, a \in A$.

The generalized cost for a route is simply assumed to be the sum of the travel times on the included links, that is, travel impedance for passing an intersection must be expressed by the generalized cost on the links belonging to this intersection. The generalized link cost is a function of the *free flow travel time*, c_a^0 , representing the constant link characteristics and, for a road link, the congestion level in the network, which is expressed by the link flow volumes, $v = \{v_a\}, a \in A$.

If each *link cost function* (link performance function) c_a is assumed to be independent of the flows on all other links, i.e. if $c_a(v) = c_a(v_a), a \in A$, the link cost functions are said to be *separable*. In most models, the link performance function $c_a(v_a)$ is an exponential or higher order polynomial function, which heavily grows rapidly as the link flow approaches the maximum capacity (typically around 1,800 vehicles per lane and hour). In many models, it is explicitly required that $c_a(v_a)$ is a monotonically increasing function.

The most widely spread method to distribute travel demand over alternative routes is the criterion “Equal Times”, stated by Wardrop (1952):

“The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.”

An assignment fulfilling this criterion, will be referred to as a *user equilibrium*. Introducing the non-negative route flow variables $h = \{h_r\}, r \in R_{pq}$, $(p, q) \in P \times Q$ and the link-route incidence variables $\delta_{ar}, a \in A, r \in R_{pq}$, $(p, q) \in P \times Q$, being 1 if link a is included in route r and 0 otherwise, the user equilibrium assignment is equivalent to the optimal solution to the following mathematical program:

$$\begin{aligned} \min_v f(v) &= \sum_{a \in A} \int_0^{v_a} c_a(s) ds, \\ \text{s.t. } \sum_{r \in R_{pq}} h_r &= g_{pq}, \quad \forall (p, q) \in P \times Q, \\ \sum_{(p, q) \in P \times Q} \sum_{r \in R_{pq}} \delta_{ar} h_r &= v_a, \quad \forall a \in A, \\ h_r &\geq 0, \quad \forall r \in R_{pq}, \forall (p, q) \in P \times Q. \end{aligned} \tag{4}$$

This program was first stated by Beckmann et al. (1956) and is referred to as the *traffic assignment problem*. A complete derivation of the Wardrop’s equilibrium principle from this program is given in, for example, Patriksson (1994).

If the link cost is a monotonically increasing function of the link flow, the optimum solution to program (4) is uniquely determined in terms of link flows. The Frank–Wolfe method (Frank and Wolfe, 1956) is a well-known technique for linearly constrained convex problems and it has been successfully implemented for the assignment problem (4); for example the software tool Emme/2 (1999) is based on this method. For an overview of other techniques, proposed for the basic problem and its extensions, see Patriksson (1994).

Though the assignment problem has a unique link flow solution (as long as the link cost is a monotonically increasing function of the link flow) there are in general many corresponding route flows. This is one of the major obstacles when adjusting OD-matrices from observed link flows, assuming a user equilibrium assignment.

The assignment model discussed so far is *deterministic* in the sense that all travelers are assumed to drive the shortest path, i.e. one of the paths having the least generalized cost. This assumption takes no account for the variation in the travelers' different perception of travel cost. Daganzo and Sheffi (1977) extended the Wardropian user equilibrium condition to the principle of *stochastic user equilibrium* by stating that:

"In a stochastic user equilibrium network no user believes he can improve his travel time by unilaterally changing routes."

Mathematically this is modelled by adding a error term to the generalized cost for each route. Depending on which distributions and which mutual dependencies that are assumed for these terms, different mathematical programs can be formulated. If the random terms are assumed to be independent and identically distributed Gumbel variates, a multinomial logit model can be used. An advantage of the stochastic assignment model compared to the deterministic model is that the route flows of an assignment are uniquely determined.

A consequence of the stochasticits assumption is that each route has a strictly positive flow. For a network including a cycle there is no upper limit for the number of routes, and thus an infinite number of route flows has to be considered, at least theoretically. In practice of course, the number of routes has to be restricted somehow. A remaining problem therefore is how to generate a restricted set of routes, which is representative for all possible routes. When just a few routes are generated for an OD-pair, these tend to be rather similar, i.e. to have a number of links in common. The overlap between the routes then gives an incorrect value for the proportion of a certain OD-demand passing each link. For further details on stochastic assignment models, see Sheffi (1985).

To conclude the description of the traffic assignment stage: Given the travel demand for each mode in each OD-pair, g_{pq}^k , $k \in K$, $(p, q) \in P \times Q$, the traffic assignment procedure assigns the travelers to routes in the network and predicts the traffic situation in terms of the link flows v_a for all links in the network, $a \in A$.

2.5 Time-dependent traffic models

Although the four-stage model is almost fifty years old, it is still fundamental for strategic traffic planning, such as, for example, identifying bottlenecks in the network, or dimensioning the infrastructure to a new living area. The outcome is, however, always a static description of the situation, which does not provide any information on how the traffic fluctuates over time. A time-dependent (dynamic) model must consider the influence from traffic conditions in a certain time period to any succeeding time period.

Developing plans of actions for different emergency situations, and studying the effect of time-varying road tolls, are examples where time-dependent models are applied for strategic planning. In the operational planning, time-dependent models running in real-time can be used for providing information to variable message signs, and controlling traffic-lights and other traffic facilities.

In a time-dependent traffic assignment model not only the route choice and the resulting route flows must be described, but also the interaction in time between vehicle streams. Beside the equilibrium assignment rules, which essentially are time-dependent extensions of the rules for a time-independent case, there must be a model for the flow propagation in the network. This model is often controlled by a special subroutine, called a dynamic network loading procedure.

Several attempts to extend Wardrop's "Equal time" criterion for a time-dependent model have been proposed. A major distinction is drawn between models relying on the instantaneous and actual travel time, respectively. The instantaneous travel time for a route is computed as the sum of the link travel times on the links included in the route at that time, when the journey is started, whereas the actual travel time relies on a forecast on what the travel time on a link will be at that time, when the traveler reaches the link. The instantaneous travel time might be combined with a model where re-routing during the trip is allowed.

In comparison to the static assignment models, where most of the researchers have agreed on the four-stage model and the basic ideas in each of its stages, the area of dynamic models is more unexplored. Though the first models for dynamic traffic assignment were proposed some thirty years ago (Merchant and Nemhauser, 1978a, 1978b), there is still no standard model framework. The time-dependent models are also surprisingly poor described in books and survey articles. Two exceptions are the literature overview by Peeta and Ziliaskopoulos (2001) and the foundations on dynamic modeling given by Han (2003).

3 Estimation of time-independent OD-matrices

The OD-matrix estimation problem amounts to finding an OD-matrix which, when it is assigned onto the network, induces link flows close to those which have been

observed in traffic counts. Most models also require some kind of information about the magnitude of the prior travel demand for each OD-pair, i.e. a *target* OD-matrix. The target OD-matrix is typically an old (out-dated) OD-matrix, possibly updated by some growth factor, or a matrix generated through a trip distribution procedure. When a target OD-matrix is used as input, the OD-matrix estimation problem is often referred to as a *calibration* or *adjustment* problem.

The problem of finding an OD-matrix which corresponds to some given link flow observations, can be seen as an inverse of the traffic assignment problem. However, there are some factors, which make the estimation problem much more complicated to handle. First of all, there is in general not possible to observe the flows on all links in the network. And even if all link flows could indeed be measured, the data would most likely be neither error-free nor consistent. Secondly, even though correct and error-free link flow data would be available for all links in the network, there are still many OD-matrices which assigned onto the network would induce the observed link flows. This motivates the use of a target matrix.

The OD-matrix estimation problem for the time-independent case has been well studied in the past decades. The following overview is inspired by the literature surveys by Willumsen (1981), Barceló (1997), Bell and Iida (1997), Abrahamsson (1998), and Ortúzar and Willumsen (2001).

3.1 Problem description for the time-independent case

In the OD-matrix estimation problem we are interested in finding a feasible OD-matrix $g \in \Omega$, where $g = \{g_i\}$, $i \in I$, consists of the demands for all OD pairs. (We use the well-established term OD-matrix for g , although, for a convenient description, the elements are organized in a vector.) The assignment of the OD-matrix onto the links in the network follows an assignment proportion matrix $P = \{p_{ai}\}$, $i \in I$, $a \in A$, with p_{ai} being the proportion of the OD demand g_i that uses link a . We will use the notation $P = P(g)$ to emphasize that, in general, because of congestion, these proportions depend on the traffic volumes, i.e. on the OD-matrix.

When assigned to the network, the OD-matrix induces a flow $v = \{v_a\}$, $a \in A$, on the links in the network. We assume that observed flows, $\tilde{v} = \{\tilde{v}_a\}$, are available for a subset of the links, $a \in \tilde{A} \subseteq A$, and that a target matrix $\hat{g} \in \Omega$ also is available.

The OD-matrix estimation problem can now be formulated as

$$\begin{aligned} \min_{g,v} F(g,v) &= \gamma_1 F_1(g, \hat{g}) + \gamma_2 F_2(v, \tilde{v}), \\ \text{s.t. } \sum_{i \in I} p_{ai}(g) g_i &= v_a, \quad \forall a \in \tilde{A}, \\ g &\in \Omega. \end{aligned} \tag{5}$$

The functions $F_1(g, \hat{g})$ and $F_2(v, \tilde{v})$ are generalized distance measures between the estimated OD-matrix g and the given target matrix \hat{g} , and between the estimated link flows v and the observed link flows \tilde{v} , respectively. These functions are assumed to be convex and they can be designed to account for variations in the quality of the given data.

The parameters $\gamma_1 > 0$ and $\gamma_2 > 0$ reflect the relative belief (or uncertainty) in the information provided by \hat{g} and \tilde{v} , and the problem thus can be interpreted as a two-objective problem, where the two objectives are expressed in F_1 and F_2 , and γ_1 and γ_2 are the corresponding weighting factors. In one extreme case, using $\gamma_1 = 0$, the target matrix will have no influence, and in the other, using $\gamma_2 = 0$, the target matrix will be reproduced and the observed link flows will have no influence.

The set of feasible OD-matrices, Ω , usually consists of all non-negative OD-matrices, but it can also be restricted to the matrices that are within a certain deviation from the target matrix, i.e. $\Omega = \{g \geq 0 | (1 - \alpha)\hat{g} \leq g \leq (1 + \alpha)\hat{g}\}$, for some parameter $\alpha > 0$ stating the tolerance level. An analogous restriction could be used to instead state a maximum allowed deviation from the link flow observations $\Omega = \{g \geq 0 | (1 - \beta)\tilde{v}_a \leq v_a \leq (1 + \beta)\tilde{v}_a, a \in \tilde{A}\}$, where $\beta > 0$.

The set Ω can also be constrained by relations between flows on different links, for example turn proportions at some intersections. Often intersections are designed with different lanes for different turning movements, which gives an opportunity to measure the turn proportions directly. This information can be added to strengthen the estimation problem. However, these proportions must be used carefully to avoid inconsistency with the observed link flows.

Another possibility is to restrict the total travel demand in all OD-pairs originating or terminating at a certain centroid, which in the four-stage model would represent an adjustment of the trip distribution with respect to the trip generation. A weaker requirement would be a total number of trips for the entire OD-matrix. In any case, the set of feasible OD-matrices Ω will remain convex, and be easily handed.

3.2 Different choices of objective

In the objective of the estimation problem (5) the deviations from the target matrix and the link observations are minimized. Obviously the resulting OD-matrix depends on the choice of distance measure, and we will therefore discuss this choice in more detail.

One type of distance measure is the *maximum entropy* function, which can be formulated as

$$F_1(g, \hat{g}) = \sum_{i \in I} g_i (\log g_i - 1). \quad (6)$$

With this choice, the target matrix is of no importance A special entropy formulation is derived from the principle of *minimum information* and was originally proposed by van Zuylen (1978) as

$$F_1(g, \hat{g}) = \sum_{i \in I} g_i (\log \frac{g_i}{\hat{g}_i} - 1). \quad (7)$$

If no target OD-matrix is available, it is sensible to assume that all values are equally likely, and by replacing \hat{g}_i by unit weights for all $i \in I$, the expression in (7) reduces to that in (6). In this case, the OD-matrix estimation, however, turns into a trip distribution procedure.

Maximum entropy functions are used in the estimation models presented by Jörnsten and Nguyen (1979), van Zuylen and Willumsen (1980), Bell (1983), Bell (1984), Fisk (1988), Brenninger-Göthe et al. (1989), and Tamin and Willumsen (1989).

A second type of distance measure is the *maximum likelihood*, which states the likelihood of observing the target OD-matrix by the estimate. It is here assumed that the elements of the target OD-matrix are obtained as observations of a set of random variables. For a Poisson distributed system with a sampling factor ρ the distance measure can be formulated as

$$F_1(g, \hat{g}) = \sum_{i \in I} (\rho_i g_i - \hat{g}_i \log g_i). \quad (8)$$

Various types of maximum likelihood measures are used in the models proposed by Ben-Akiva (1987), Spiess (1987), and Tamin and Willumsen (1989). The maximum likelihood formulation proposed by van Zuylen and Branston (1982) is based on an assumption of normal distributed deviations.

Though the presentation here has focused on the deviation from the target OD-matrix only, both the maximum likelihood and the maximum entropy measures, of course, can be used for the deviation from the link flow observations as well, by formulating F_2 analogously to (8) and (6), respectively. Some of the refereed models do this, but there are also other possibilities.

The type of objective which is most common in the models proposed in the last decade is the *least-square* formulation. The least-square is a well-known deviation measure used in many types of estimation problems and is given by

$$F_2(v, \tilde{v}) = \frac{1}{2} \sum_{a \in \bar{A}} (v_a - \tilde{v}_a)^2. \quad (9)$$

It is of course possible to give individual weights to the single deviations. One way of choosing the weights is to utilize information on the reliability of each observation. The measurements contained in \tilde{v} are normally computed as means from a set of

observations for each link. In this case we can use the variance σ_a^2 among the measurements to account for how important each link observation is, and reformulate (9) as

$$F_2(v, \tilde{v}) = \frac{1}{2} \sum_{a \in \tilde{A}} \frac{1}{\sigma_a^2} (v_a - \tilde{v}_a)^2. \quad (10)$$

It is of course also possible to take the covariance between the observations on different links into account. Such measures are referred to as *generalized least-squares*. An analogous formulation can capture the deviation from the target OD-matrix, i.e. $F_1(g, \hat{g})$.

Various types of generalized least-square formulations have been used in the models proposed by Carey et al. (1981), Cascetta (1984), McNeil and Hendrickson (1985), Spiess (1990), Bierlaire and Toint (1995), Florian and Chen (1995), Maher and Zhang (1999), Bianco et al. (2001), Cascetta and Postorino (2001), Maher et al. (2001), Codina and Barceló (2004), Doblas and Benitez (2005), and Nie et al. (2005) among others.

To summarize this discussion, the functions $F_1(g, \hat{g})$ and $F_2(v, \tilde{v})$ represent distance measures, between the estimated OD-matrix g and the given target matrix \hat{g} , and between the assigned link flow v and the observed link flow \tilde{v} , respectively. Typically some combination of maximum entropy, maximum likelihood and least-square expressions are used, and they can be designed to take account varying data quality. We conclude that, in any case, the functions to be minimized are continuous, convex and at least two times differentiable with respect to their respective arguments.

3.3 Characteristics of the constraints

We have already mentioned that there are in general many OD-matrices which, when they are assigned to the network, induce the same link flows. In this section we will further explore the set of constraints defining the relationship between the travel demand (the OD-matrix) and the link flows.

Consider again the equations from the problem description (5), which connect the OD-matrix to the link flows:

$$\sum_{i \in I} p_{ai}(g) g_i = v_a, \forall a \in \tilde{A},$$

There is one equation for every link flow observation. This equation system is underdetermined as long as the number of OD-pairs, $|I|$, is greater than the number of link flow observations, $|\tilde{A}|$.

In general the number of OD-pairs is much greater than the number of link flow observations, especially for real-world networks. The number of OD-pairs is a subset of all possible node pairs, i.e. $I \subseteq N \times N$. Typically, the number of centroids are proportional to the number of nodes, and each pair of centroids defines an OD-pair. Therefore $|I| \propto |N|^2$ holds, which means that the number of unknown travel demands grows quadratically with the network size.

However, the mean number of links connected at each intersection, i.e. node, is independent of the network size. If we assume that some portion of the link flows are observed, the relation $|\tilde{A}| \propto |A| \propto |N|$ holds. Thus, the number of equations is proportional to the network size. We therefore conclude that the OD-matrix estimation problem has a greater freedom of choice, the greater the network is.

Topological dependencies in the network further delimits how well the OD-matrix can be determined by the equation system. Kirchoff's law, well-known from physics, states that the sum of incoming and outgoing flows at any intermediate node must coincide. This means that, for each intersection, at least one link flow is directly given from the others, which results in a row-wise dependency for the equation system.

The non-zero elements $p_{ai}(g)$ arise from one or more paths generated for OD-pair $i \in I$. However, since every subpath of a path is also a path, every pair of nodes along a certain path is also connected through parts of this path. This results in a column-wise dependency for the equation system.

We can conclude that the equation system most likely is not of full rank, which further increases the freedom of choice in the OD-matrix estimation problem.

3.4 Detector allocation

When OD-matrices are to be estimated from the information contained in link flow observations, the choice of links where detectors are placed is of course very crucial for the result. Therefore, some attention should be given to the problem of how to choose the set of detector links, aiming for a good and reliable estimate of the OD-matrix. The main reference for the overview in this section is the survey part in Paper IV of this thesis.

When link flow detectors are allocated to the traffic network with the aim to estimate a reliable OD-matrix, we want to maximize the coverage of the traffic (link flows, route flows, OD-pairs, OD-demands, etc.) in one way or the other. Clearly, there are different ways to specify how this coverage should be accomplished.

First of all, we must define under which conditions we consider a certain link to cover the travel demand in an OD-pair. There are basically two different definitions in the literature. In the first, and most commonly used approach, we consider an OD-pair to be covered, as soon as a certain portion of the travel demand passes at least one

link with a detector. Clearly, this approach requires some assumption about how the travel demand is distributed onto routes through the network. This route choice information is typically supplied from an assignment of the target OD-matrix.

In the other definition of coverage, we ignore the route choice and consider an OD-pair to be covered if and only if every possible route from the origin node to the destination node passes at least one detector. In practice this approach tends to allocate detectors to bridges and tunnels along natural boarders, such as rivers or railways.

Another aspect to take into account is whether we want to cover many OD-pairs, or many travelers. In the first case, all OD-pairs are equally important, whereas the second case counts every traveler, and thus, give more importance to the OD-pairs with a greater travel demand. Alternatively, we can formulate the detector allocation problem as to maximize the coverage of routes or route flows. The most commonly used strategy seems to be OD-pair coverage; see the survey part in Paper IV of this thesis.

The simplest approach for detector allocation is to choose those links where we expect the maximum flow to be observed, without taking account for any travelers being counted twice. Detectors are then typically allocated along one major road and most of the routes passing one detector also pass some other.

In practice, of course, link flow observations are not being performed only for providing information to the OD-matrix estimation problem. Detectors are also being placed for other purposes, such as the management of traffic signals, different real-time information systems, or road tolls. Therefore, when modeling the detector allocation problem, it is natural to consider the case where the placement of some link detectors are already at hand.

3.5 Constant assignment

From a modelling point of view, the most distinguishing difference between the approaches for the OD-matrix estimation problem, is how the assignment proportions in P are calculated and re-calculated throughout the estimation procedure. Especially it is crucial if the assignment matrix $P(g)$ is assumed to depend on g , i.e. if route choices are made with respect to congestion, or not. In the latter case, $P(g) = P$ is a constant assignment matrix, and the first set of constraints in the generic description (5), can be formulated as

$$\sum_{i \in I} p_{ai} g_i = v_a, \quad \forall a \in \tilde{A}. \quad (11)$$

The assumption that the assignment, i.e. the route choice, is independent of the load on the links, is realistic in a network with very low congestion rate, or in networks where in practice only one route can be used in each OD-pair. An example of this

is a corridor network, modeling a motorway through a city and its entrances and exits. In such a case the shortest path between origin and destination is uniquely determined, independent of the travel times.

A bit more sophisticated are the OD-matrix estimation methods where the used assignment matrix is taken from a carefully computed assignment of the target matrix, $P(\hat{g})$. If the OD-matrix to be estimated is close enough to the target matrix, this is a good approximation even for congested networks.

The models proposed by van Zuylen (1978), van Zuylen and Willumsen (1980), Carey et al. (1981), Willumsen (1981), Bell (1983), Cascetta (1984), McNeil and Hendrickson (1985), Spiess (1987), Brenninger-Göthe et al. (1989), Bierlaire and Toint (1995), Bianco et al. (2001) and Bierlaire (2002) all assume that the route proportions P are kept constant in the estimation problem. In the model by van Zuylen and Branston (1982) the route proportions P are replaced with flow conservation constraints for each node.

3.6 Equilibrium assignment

In case the network is congested, and the routes are chosen with respect to the current travel times, the OD-matrix estimation problem is more complicated. The route proportions depend of the current traffic situation (travel times/link flows), which in turn depends of the OD-matrix. Thus, the relationship between the route proportions P and the OD-matrix g can only be implicitly defined. The set of feasible solutions to the estimation problem (5), is then defined as all the points (g, v) where v is the link flow solution satisfying an assignment of the corresponding OD-matrix $g \in \Omega$.

Nguyen (1977) presented the first model of this type and an extended version was proposed by Jörnsten and Nguyen (1979). Gur et al. (1980) suggested a way to obtain unique OD-matrices. Erlander et al. (1979) and Fisk and Boyce (1983) have proposed OD-matrix estimation methods based on a combined distribution and assignment model. In all these estimation procedures it is assumed that the assignment is made according to the deterministic user equilibrium assumption. This assumption will hold also in the following presentation.

The deterministic equilibrium assignment is an *inferior* problem to the *superior* problem of estimating the OD-matrix. A problem, which can be separated into one superior (or synonymously *outer* or *upper*) part, and one inferior (or *inner* or *lower*) part, is called a *bilevel problem*. The generic OD-matrix estimation problem given in (5), can be reformulated as a bilevel programming problem in the following way.

In the superior problem, the OD-matrix g defines the decision variables and we want to minimize $F(g, v)$ subject to $g \in \Omega$, that is

$$\begin{aligned} \min_g F(g, v) &= \gamma_1 F_1(g, \hat{g}) + \gamma_2 F_2(v, \tilde{v}), \\ \text{s.t.} \quad g &\in \Omega. \end{aligned} \tag{12}$$

The link flow v must satisfy the equilibrium assignment conditions, given the OD-matrix g . These conditions are fulfilled by solving the nonlinear inferior problem in which the link (and route) flows are decision variables:

$$\min_v f(v) = \sum_{a \in A} \int_0^{v_a} c_a(s) ds,$$

$$\text{s.t.} \quad \sum_{k \in K_i} h_k = g_i, \quad \forall i \in I, \tag{13}$$

$$\sum_{i \in I} \sum_{k \in K_i} \delta_{ak} h_k = v_a, \quad \forall a \in A, \tag{14}$$

$$h_k \geq 0, \quad \forall k \in K_i, \forall i \in I. \tag{15}$$

Fisk (1988) was first to give a bilevel formulation of the OD-matrix estimation problem. She used a variational inequality formulation to express the equilibrium conditions and in this way she allowed general link cost functions, which must not be separable (see Section 2.4).

It is well-known that bilevel programming problems are in general non-convex; see for example Bard (1998). By using methods known today, at the best a local optimum solution is obtained.

Spiess (1990) suggested a heuristic approach to solve problem (12). It is an iterative procedure, in which $\gamma_1 = 0$ and \hat{g} is used as initial solution. In his approach, an approximate gradient of the objective function with respect to the OD-matrix is computed, under the assumption that the proportion matrix $P(g)$ is locally constant. Spiess' heuristic is efficient for large-scale applications and has been included in the software tool Emme/2 (1999). Doblas and Benitez (2005) have shown how the method could be made even more efficient by adding linear constraints bounding the possible changes of travel demands in the OD-pairs.

Florian and Chen (1995) reformulated the bilevel problem into a single level problem using the concept of marginal functions. They proposed to use a Gauss-Seidel type heuristic to solve the problem. Chen (1994) proposed an augmented Lagrangean approach, which can be shown to converge to a stationary point. This approach, however, requires that all used paths in each OD-pair are known beforehand, and it is thus applicable only to very small problem instances.

Yang et al. (1992), Yang (1995) and Yang and Yagar (1995) all use the bilevel formulation and propose heuristics, which iteratively solve the upper and lower level

problems. Information from the lower level problem is transferred by so called influence factors, which are defined by route proportions or explicit derivatives. The derivatives are computed using the sensitivity analysis by Tobin and Friesz (1988). Assuming complementarity conditions to hold and disregarding any topological dependencies, they get approximate values of the derivatives, which are acceptable for small to medium sized networks. However, for larger networks the topological dependencies are significantly greater. Also, since all these methods include matrix inversions they are computationally very demanding for large problem instances.

Maher and Zhang (1999) have developed an iterative method where a first order Taylor approximation is used to express the changes of the assignment map $P(g)$ with respect to the OD-matrix g . In a first step the assignment map is kept constant and a tentative OD-matrix is estimated with some of the techniques discussed in Section 3.5. The tentative OD-matrix is, however, not taken as it is, but is used to give a search direction from the present OD-matrix. By assigning the tentative OD-matrix to the network, we get an approximation of how the assignment map will change along the search line direction. Maher et al. (2001) have further developed the method to the case where a stochastic assignment is assumed.

The method proposed in Paper I of this thesis is based on the more general sensitivity analysis presented by Patriksson (2004). It is a further development of the method by Spiess (1990), where a second order approximation is used for the partial derivatives. Further, a solution scheme is proposed, where the mutual influence between the OD-pairs, which is considered, can be restricted to keep a good balance between the computational time needed for yielding the search direction and finding an optimal step length.

The algorithm developed by Codina and Barceló (2004) is an application of the general method for non-differentiable convex minimization, proposed by Wolfe (1975). It is based on subgradients and does not need any sensitivity information.

Neither the method proposed in Paper I of this thesis nor the method by Codina and Barceló (2004) is affected by topological dependencies, as the methods based on the sensitivity analysis by Tobin and Friesz (1988). Further, none of them involves matrix inversions and they therefore seem to be efficient also for larger networks.

The method proposed by Sherali et al. (1994), and further developed by Sherali et al. (2003), assumes a deterministic user equilibrium. It is based on error-free information on the travel times for all links in the network, and this information must be consistent with the link flow observations. In the method, it is assumed that an equilibrium link flow is observed, and thus that the set of used routes is known beforehand. This requirement makes a comparison to other methods unfair.

The methods proposed by Cascetta and Postorino (2001), Clegg et al. (2001), Maher et al. (2001), and Yang et al. (2001) differ from the other methods in the sense that they presume a stochastic user equilibrium.

4 Estimation of time-dependent OD-matrices

If the travel demand in the OD-matrix is assumed to vary over time, the OD-matrix is said to be time-dependent. The number of applications where a time-dependent OD-matrix is required has grown rapidly in the last decade, mainly as a result of the increasing computing possibilities and the new techniques for supplying interactive information via internet, variable message signs and so on. An accurately estimated time-dependent OD-matrix is a basis for the decisions in many *Intelligent Transportation Systems* (ITS). Lind (1997) gives an overview of ITS applications with special attention to Swedish conditions.

Time-dependent OD-matrices are used both for strategic and operational purposes. In the strategic area the computations are made off-line and the aim is to model the normal traffic situation as well as possible. Such OD-matrices are used for evaluating the time-dependent effect of different scenarios, for example for generating plans of actions in case of incidents. This type of ITS applications are sometimes referred to as *Advanced Traffic Management Systems* (ATMS).

The pre-calculated scenarios and plans of actions are also used as a default description of the traffic conditions in the operational management. For an accurate real-time model of the traffic, these values, of course, must be combined with instantaneous estimates. This type of operational models are used to produce travel time forecasts, which in turn are essential for different kind of route guidance systems. *Advanced Traveler Information Systems* (ATIS) is a commonly used term for such applications.

In the following, first a generic formulation of the estimation problem for the time-dependent case is given, analogous to the formulation of the time-independent case in Section 3.1. We will then discuss some of the algorithms which have been proposed to solve the time-dependent OD-matrix estimation problem, and, finally, we will present some of the methods used for real-time estimation. It should be pointed out that the models where time-dependent OD-matrices are used, in general consider networks smaller than those considered in the time-independent case. Especially, the methods for real-time estimation are mostly designed for very small networks, typically with a corridor structure.

Estimating time-dependent OD-matrices is relatively new research area and unlike the time-independent case, survey articles are hardly found. For some introduction, we refer to the overviews by Lindveld (2003) and Balakrishna (2006).

4.1 Problem description for the time-dependent case

A general formulation of the time-dependent OD-matrix estimation problem can be derived from the time-independent formulation, given as problem (5) in Section 3.1. In the time-dependent estimation problem we aim at finding an OD-matrix

$g = \{g_{it}\} \in \Omega$, where element g_{it} expresses the travel demand in OD-pair $i \in I$ leaving the origin node in time period $t \in T$. The assignment of the OD-matrix onto the links in the network is made according to the assignment mapping $P = \{p_{it}^{ar}\}$, $a \in A$, $r \in T$, $i \in I$, $t \in T$, where each element in the matrix is defined as the proportion of the travel demand g_{it} passing link $a \in A$ during time period $r \in T$. As for the time-independent case these proportions may depend of the demand.

When assigned to the network, the OD-matrix induces a flow $v = \{v_{ar}\}$, $a \in A$, $r \in T$, on the links in the network. We assume that observed flows, $\tilde{v} = \{\tilde{v}_{ar}\}$, are available for a subset of the links, $a \in \hat{A} \subseteq A$, in all time periods $r \in T$, although observations must not be performed in all time periods for all of the observed links. We could even consider a different discretization of time for the link flows than for the OD-matrix. In practice, however, the same set of time periods T is used for both the OD-matrix and the observed link flows.

Given a target OD-matrix $\hat{g} \in \Omega$ we can now formulate the generic time-dependent OD-matrix estimation problem as

$$\begin{aligned} \min_{g,v} \quad & F(g, v) = \gamma_1 F_1(g, \hat{g}) + \gamma_2 F_2(v, \tilde{v}), \\ \text{s.t.} \quad & \sum_{i \in I} \sum_{t \in T} p_{it}^{ar}(g) g_{it} = v_{ar}, \quad \forall a \in \hat{A}, r \in T, \\ & g \in \Omega. \end{aligned} \quad (16)$$

Deviation measures $F_1(g, \hat{g})$ and $F_2(v, \tilde{v})$ for the time-dependent formulation are chosen as for the time-independent case, see Section 3.2.

The set of feasible OD-matrices, Ω , is bounded analogously to the time-independent case. A new possibility is to add constraints restricting the maximum deviation of the time-aggregated OD-demand from the target matrix, i.e. the distance of $\sum_{t \in T} g_{it}$ from $\sum_{t \in T} \hat{g}_{it}$. Also for the time-dependent case Ω will remain convex and be easily handled.

4.2 Methods for time-dependent estimation

One of the first models for time-dependent OD-matrix estimation was proposed by Willumsen (1984). This model simply makes the assumption that the route choice ratio is fixed and independent of time, i.e. that $p_{it}^{ar} = p_{it}^{ar}(g)$ holds for all $a \in A$, $r \in T$, $i \in I$, $t \in T$. This is an extension to the proportional assignment assumption in the time-independent case, see Section 3.5. As in the time-independent case, it is assumed that the congestion level is kept throughout the estimation, meaning that route choice and flow propagation are constant.

Willumsen's model is used in the OD-matrix estimation procedure in the software tool Contram; see Contram (2002). In this implementation the proportional as-

segment mapping P is taken from an assignment of the target matrix \hat{g} , which is performed by simulation. If the OD-matrix to be estimated, g , is close to the target matrix, \hat{g} , the assignment mapping for the target matrix, $P(\hat{g})$, is probably a good approximation to the actual assignment mapping, $P(g)$. If, however, the target matrix \hat{g} is unreliable, then so is the assignment mapping $P(\hat{g})$. Especially if the network is congested and/or there are many alternative routes, the assignment mapping will be sensitive also to small changes in the travel demand.

Davis and Nihan (1991) uses a stochastic procedure for generating the assignment mapping, which is parameterized by the means and variances of the travel demand. They then develop a maximum likelihood estimator, which can be viewed as a development of the method proposed by Spiess (1987) for the static case. Davis (1993) extends the ideas to a general Markov model, for which it can be shown that consistent OD-matrix estimates can be derived from link flows, also under relatively weak conditions.

Bell et al. (1991) make assumptions on the travel time distribution and thereby they can account for different flow propagation in different time periods. This improvement is important for larger networks, where the assumption on equal travel times might be too rough. Hereby, the model becomes dynamic both in flow propagation and in route choice. However, none of these aspects is related directly to that congestion, which is actually given by the OD-matrix, but only to the time period. The relationship between the OD-matrix g and the assignment mapping P is based on statistics only.

Cascetta et al. (1993) develop a model for a general two-objective form of the problem. In their numerical tests a general least-square estimator has been used. The proportional assignment mapping,, P is expressed as a product of a time-dependent link-route incident mapping, $\Delta = \{\delta_{ar}^{kt}\}$, and a probability term, $\rho(k|t)$, expressing the probability that a traveler in OD-pair i , departing in time-period t , will choose route $k \in K_i$:

$$p_{it}^{ar} = \sum_{k \in K_i} \delta_{ar}^{kt} \rho(k|t). \quad (17)$$

This model is further developed by Tavana (2001), and Ashok and Ben-Akiva (2002). The latter authors also address a real-time formulation of the model (see Section 4.3).

Another statistical model is proposed by Hazelton (2000), who assumes P to be given by a Poisson distribution of the demand in the OD-matrix, in which the variation of route choice proportions is represented. The network is assumed to be uncongested in the sense that the demand has no influence neither to route choices nor to flow propagation. As objectives the maximum multivariate normal approximation of the likelihood is used. Some ideas for decreasing the complexity of the algorithm are given by Hazelton (2003).

In the model by Sherali and Park (2001) each time-dependent route flow, h_k^t , where $\sum_{k \in K_i} h_k^t = g_{it}$, is estimated individually. They develop a column generation procedure where the master problem is to find a non-negative route flow solution h , such that the objective in (16) is minimized. The column generation problem is to compute paths in a time-expanded network. Unfortunately, convergence cannot be guaranteed.

The model proposed by Lindveld (2003) is basically a time-dependent extension of the time-independent model proposed by Maher and Zhang (1999), for the case of deterministic user equilibrium, and further developed by Maher et al. (2001) for stochastic user equilibrium. The method has been successfully implemented for a small test network with a corridor structure, i.e. a network with no route choices.

Zhou and Mahmassani (2006) extend the generic formulation in (16) with a third objective capturing information from automatic vehicle identification data. This is used as a target for the proportion of the flow on a certain link in a certain time period, which passes another link in a succeeding time period. In the experiments, the software Dynasmart (2002) is used for the dynamic traffic assignment.

4.3 Methods for real-time estimation

A special case of the time-dependent OD-matrix estimation problem, introduced in Section 4.1, occurs when the estimation is performed instantaneously. We will refer to this as *real-time estimation* or *on-line estimation*. Such an estimate of the current travel demand in the network is a useful input for different kind of operational management, especially for computing travel time forecasts and producing route guidance information.

Most of the models proposed for real-time estimation can be used for off-line purposes as well. For a real-time estimator, however, the computing time available is very limited, and therefore the accuracy in the methods generally is lower. It should be noted, though, that some of the concepts presented for real-time estimation could be refined for more accurate computation, if the estimation is performed off-line.

For defining the real-time estimation problem, the set of time-periods must be specified as an ordered, possible infinite, set, $T = \{t_1, t_2, t_3, \dots\}$. An estimation is performed at a time τ , when the link flow observations are available only for the present and preceding time periods, i.e. $\tilde{v} = \{\tilde{v}_{ar}\}, a \in \tilde{A}, r \in T^\tau$, where $T^\tau = \{t_1, t_2, t_3, \dots, t_\tau\} \subseteq T$.

The model proposed by Cremer and Keller (1987) is one of the first of this type. In real-time they estimate the turning proportions at intersections utilizing the causal dependencies between the observed link flows. They allow for different route choices in different time periods. The turning proportions are expressed as functions time only, and the model contains no relations between the travel demand and the route choice. Further, it does not take the different flow propagations, and thus, the

differences in travel times, into account. Their work compares weighted least-square estimation, constrained optimization, simple recursive estimation, and a Kalman filtering approach. The results from Nihan and Davis (1987, 1989) and Sherali et al. (1997) are basically improvements of the computational techniques for the model.

The model concept is further considered by van der Zijpp and Hamerslag (1996), who use an improved Kalman filtering technique. Their method is developed for a corridor network, where only one possible route is available for every OD-pair. In the model proposed by van der Zijpp (1997), the estimation also incorporates automated vehicle identification data.

The models by Bell (1991), and Chang and Wu (1994) are more generalized in the sense that also varying travel times are taken into account. Bell uses statistical expressions for the travel times, comparable to those presented by Bell et al. (1991) for the off-line approach, which is discussed in Section 4.2 above. Chang and Wu compute the present travel demand recursively from the data collected from previous time intervals. Chang and Tao (1996) further extend the model to also incorporate signal settings. In the extension by Wu (1997) general network structures are considered.

Trying to restrict the computational effort required, Ashok and Ben-Akiva (1993, 2000) work with a state vector, only containing the (time varying) deviations from the target matrix \hat{g} . The authors also experiment with an aggregated formulation, estimating the deviations in departure rates per origin, rather than for a specific OD-pair. Though some accuracy is being lost, this second approach shows a promising performance. The reliability is improved by combining it with an off-line estimation, as is suggested in Ashok and Ben-Akiva (2002). (See also the discussion in Section 4.2.)

Camus et al. (1997) further develop the model by Ashok and Ben-Akiva (1993) to make forecasts of the future travel demand. Basically the idea is to propagate the flows from the origins and thereby predict future link flows, which replace the observed flows in the estimation problem.

In the algorithm proposed by Bierlaire and Crittin (2004) the state description by Ashok and Ben-Akiva is combined with a general technique for solving large-scale least-square problems. This iterative algorithm is shown to be more efficient than the Kalman filter approach suggested by Ashok and Ben-Akiva.

The approach proposed by Li and de Moor (2002) explores the effects of incomplete link flow observations. As objective they use a recursive generalized least-square estimator. The algorithm is comparable to that of Sherali et al. (1997).

Beside the well-defined models described above, different types of genetic algorithms and neural networks are widely implemented for handling various transportation problems in practice, for instance parameter estimation, traffic forecasting, traffic pattern analysis, and traffic control. Neural network estimators are given by, for example, Vythoulkas (1993), Dougherty (1995), Dougherty and Cobbett (1997), Kirby

et al. (1997), Barceló and Casas (1999), Mozolin et al. (2000), Jiao and Lu (2005), and Balakrishna et al. (2006). These methods are easy to implement and use. Only considered as black box tools, however, a fair comparison to the above-described methods, which rely on a traffic assignment model, cannot be performed. A drawback of the neural network approaches is their computational expense. Practical experiences from the Netherlands (Dougherty and Cobbett, 1997) have shown that though a neural network, which makes use of all available input information, can produce well-fitted predictions, such a strategy is not viable for real-time use.

5 The thesis

This thesis is devoted to the problem of estimating reliable OD-matrices from link flow information, provided by traffic counters. It basically covers three aspects of the OD-matrix estimation problem: the time-independent case, the time-dependent case, and the problem to collect the link flow observations such that the reliability of the estimated OD-matrix can be ensured.

5.1 Motivation

For most analyses in the field of planning and control of traffic there is a need of an accurate estimate of the underlying travel demand. Forecasts on queues, travel times, emissions, and accessibility, which in turn are used to predict the need of new-building, re-building, regulation, and information in the traffic, all rely on estimates of the travel demand. Especially the interest for time-dependent models grows for several reasons: traffic signal control, dynamic speed limits, incident detection and management of road tolls, the supply of information for different kind of route guidance systems, etc. Link flows, which are observed on some of the links in the modeled network, constitute an important data source for the estimation of OD-matrices. Compared to other available information sources, observed link flows supply cheap, reliable and easily updated information on the travel demand.

The time-independent OD-matrix estimation problem is well-known and has been studied in the literature for the last 25 years, at least. Due to congestion and the assumption on an equilibrium assignment, the estimation problem is in general hard to solve and until today only simplified solution procedures are available in commercial software. In particular, the latest advances in sensitivity analysis for traffic equilibria are not fully utilized by the currently available OD-matrix estimation procedures.

When a time-dimension is introduced into the OD-matrix estimation problem, not only the amount of travel demand, but also the departure time of every trip, has to be estimated. The development of accurate methods for OD-matrix estimation is limited by the absence of well-established and theoretically founded methods for time-dependent traffic assignment. Until today only assignment methods based on

simulation are applicable for large-scale networks, and it is a challenge to adopt consistent OD-matrix estimation procedures.

In the OD-matrix estimation problem, the observed link flows are the only factors that induce a change of the target OD-matrix. It is therefore important to measure link flows which cover the travel demand as well as possible. Link flow detectors are traditionally allocated manually to the network and their placement is often motivated by other reasons than to ensure the best possible quality of the estimated OD-matrix. The problem of allocating link flow detectors to the network with the pronounced purpose to estimate OD-matrices accurately, is surprisingly little studied in spite of its profound importance for the quality of the estimated OD-matrices.

5.2 Contribution

The thesis gives the following contributions to the research on OD-matrix estimation from link flow observations:

- An accurate algorithm for solving the bilevel formulation of the time-independent OD-matrix estimation problem. The algorithm can be easily implemented, and performs well for large-scale networks.
- An efficient algorithm for computing a second order approximation for how the link flows depend on the OD-matrix. The algorithm utilizes the latest advances in sensitivity analysis of time-independent traffic equilibria.
- Pre-adjustment schemes for improving the quality of a time-dependent OD-matrix, which is subsequently estimated with conventional techniques.
- A case study, in which these schemes are successfully applied to a network modeling parts of Gothenburg, Sweden.
- A heuristic for solving the time-dependent OD-matrix estimation problem, which is one of the most accurate algorithms being proposed for the generic problem formulation, without any special requirements on the network structure, assignment procedure, etc.
- Important insights in the complexity of the time-dependent OD-matrix estimation problem, with implications on the importance of reliable input data.
- A survey of models and methods proposed for the problem of allocating link flow detectors with the aim to estimate OD-matrices.
- An experimental environment where different strategies for link flow detector allocation can be analyzed and evaluated.
- An empirical evaluation of some of the established detector allocation strategies, with respect to the quality of the estimated OD-matrix.

- Outline of a new method for the detector placement problem, where impact to the estimated OD-matrix is emphasized.

5.3 Methodology

Methodologically, traffic modeling combines different academic areas. Traffic modeling methods utilize concepts from statistics, economics, automatic control, computer science, and civil engineering. The methods applied in this thesis mainly comes from the mathematical disciplines, and especially from the area of mathematical modeling and optimization.

5.4 Appended papers

Five papers are annotated to the thesis. The author of the thesis has contributed to the papers by a major involvement in the development and implementation of the solution methods, in the writing process, and in the analyses of the results.

Paper I: A Heuristic for the Bilevel Origin–Destination Matrix Estimation Problem

Co-authored with Jan T Lundgren.

This paper considers the time-independent case. The estimation problem is given a nonlinear bilevel formulation, where the lower level problem is to assign a given OD-matrix onto the network according to the user equilibrium principle. The problem is reformulated into a single-level problem, where the objective function contains link flows that are implicitly given by the assignment of the OD-matrix at hand. A descent heuristic, which is an adaptation of the well-known projected gradient method, is proposed as solution procedure.

When computing a search direction, the difficulty lies in the calculation of the Jacobian matrix, which represents the derivatives of the link flows with respect to a change in the OD-flows. We do this by solving a set of quadratic programs with linear constraints. If the objective function is differentiable at the current point, the Jacobian is uniquely determined and we obtain a gradient direction. Also if differentiability does not hold, the returned direction can be used heuristically for computing a good search direction. Numerical experiments, with both some well-known test networks and a larger network from Stockholm are presented. The results indicate that the solution approach is efficient for medium to large sized networks.

The content of Paper I has been presented at:

- The International Conference on Operation Research, Duisburg, Germany,

September 3–5, 2001.

- The 9th Meeting of the EURO Working Group on Transportation, Bari, Italy, June 10–13, 2002.

An earlier version of Paper I is included in the licentiate thesis (Peterson, 2003).

Paper I has been revised for probable publication in *Transportation Research B*.

Paper II: Methods for Pre-Adjusting Time-Dependent Origin–Destination Matrices — an Application to Gothenburg

Co-authored with Jan T Lundgren and Stellan Tengroth.

We here consider an application of OD-matrix estimation from Gothenburg, in which the initial time-dependent OD-matrix is compounded from information about the daily travel demand and its time distribution for each trip purpose. With conventional estimation techniques this detailed information is aggregated, before estimation with respect to the link flow observations is performed. The idea presented in the paper is to use pre-adjustment schemes, that utilize the structure of the OD-matrix, and thereby obtain a better initial target matrix for the subsequent estimation procedure, which is performed with a traditional methodology. Three schemes are proposed for pre-adjusting: the overall travel demand for each trip purpose, the aggregated time-distribution for the total travel demand, and the time distribution for each trip purpose. The schemes are based upon assumptions about how the observed link flows are distributed over different trip purposes and how the average travel time from the origins to the observation links can be approximated.

Numerical tests are presented for both a small test network and for the Gothenburg network, where more than one million OD-matrix elements are estimated. The results show that a significantly better agreement to the observed link flows is obtained by using the pre-adjustment schemes. We believe that the developed schemes, possibly with small modifications, are applicable to many other traffic models with similar characteristics.

The content of Paper II has been presented at:

- Transportforum, Linköping, Sweden, January 14–15, 2004.

A previous version of Paper II is published as:

- “Methods for pre-adjusting time-dependent origin–destination matrices”, in proceedings of the *10th World Congress and Exhibition on Intelligent Transport Systems and Services*, Madrid, Spain, November 16–20, 2003, Paper No. 2436.

Paper II is included in the licentiate thesis (Peterson, 2003).

Paper III: A Heuristic for the Estimation of Time-Dependent Origin–Destination Matrices from Traffic Counts

Co-authored with Jan T Lundgren and Clas Rydbergren.

This paper considers a generic formulation of the time-dependent OD-matrix estimation problem. Special attention is given to the assignment map, which describes how the travel demand is transferred onto route and link flows in the network, and its relationship to the level of congestion. We decompose the assignment map in two parts, handling the route choice (traffic assignment) and the traffic flow propagation (network loading), respectively, and discuss how these components are affected by an adjustment of the travel demand, and why the effects are more difficult to predict when some links are oversaturated in one or more time-periods. We suggest an algorithm which is an extension to previously proposed methods for the time-independent OD-matrix estimation problem, where the changes of the assignment map are approximated by a difference quotient between two assignment maps in the search direction.

The algorithm is implemented together with a mesoscopic tool for dynamic traffic assignment and verified for a small test network, Brunnsviken, in the northern part of Stockholm. By numerical experiments the importance of the parameter settings are illustrated.

The content of Paper III has been presented at:

- Nordic MPS '04, Norrköping, Sweden, October 21–23, 2004.
- The 10th Jubilee Meeting of the EURO Working Group on Transportation, Poznan, Poland, September 13–16, 2005.
- Transportforum, Linköping, Sweden, January 11–12, 2006.

An earlier version of Paper III was published in:

- *Advanced OR and AI Methods in Transportation, Proceedings of the 10th Jubilee Meeting of the EURO Working Group on Transportation, Poznan, Poland, 13–16 September 2005*, A. Jaszkiewicz et al. (eds.), Publishing House of Poznan University of Technology, Poznan, Poland, pp 242–246.

Paper IV: Allocation of Link Flow Detectors for Origin–Destination Matrix Estimation — a Comparative Study

Co-authored with Torbjörn Larsson and Jan T Lundgren.

This paper is devoted to the problem of allocating link flow detectors, with the aim to ensure the best possible quality of the OD-matrix to be estimated. To meet this goal, the “coverage” of the traffic should be maximized. Coverage of traffic can be defined in different ways, and especially the coverage of OD-pairs and travel demand (weighted OD-pairs) are frequently used. We develop an experimental framework, where the effect on the estimated OD-matrix induced by different allocation strategies can be evaluated. This framework is based on the assumption that a synthetic “true” OD-matrix is available, and hence, that the deviation from the estimated OD-matrix can be measured. The framework enables studies not only on how successful the allocation strategies are, but also on how differences in the implementations affect the result. As a side-effect we can also study the importance of how the assignment map is chosen.

Beside the literature survey and an experimental framework, which can be used for further similar studies, Paper IV also contains an empirical study with three traffic networks. The results indicate that it is crucial to consider the travel demand, and not only the number of OD-pairs, when covering the traffic. The choice of assignment map seems to be of less importance for the quality of the estimated OD-matrix.

Paper IV is under revision for possible publication in *Transportation Research B*.

Paper V: A Novel Model for Placement of Detectors for Origin–Destination Matrix Estimation

Co-authored with Torbjörn Larsson and Jan T Lundgren.

Paper V is based on the assumption that the most correct OD-matrix is estimated if consistent and complete link flow observations are available. Non-complete information is interpreted as a relaxation of this requirement. From this assumption, the problem to place link flow detectors, as to maximize the quality of the estimated OD-matrix, is expressed in terms of discarding those links, with the least importance for the estimation problem. We state the OD-matrix estimation problem by using a Lagrange-dual formulation, and show that the non-availability of a link flow observation is equivalent to fixing a dual variable to zero. This observation leads to an interesting max-min formulation of the detector placement problem. Two related measures, expressing how well the OD-matrix to be estimated is determined by the choice of detector links, are also discussed.

The content of Paper V is theoretical and its purpose is to analyze the relationship between OD-matrix estimation and link flow detector allocation.

5.5 Future research

The algorithms which are presented in the appended papers are all implementable for real-life applications, and may be included in commercial software for OD-matrix estimation. Some adaptations might be required and the computational efficiency can be improved.

The rapid development of applications of time-dependent traffic models, all of which require accurately estimated OD-matrices, is a challenge for the future. A lot of work remains to establish and validate procedures for time-dependent traffic assignment, for which sensitivity analyses can be derived and applied to improve a time-dependent OD-matrix estimation procedures.

Throughout the work in this thesis, link flow observations are assumed to be the only information collected for describing the traffic situation, which should be expressed in the OD-matrix to be estimated. Today, however, also observations of travel times and/or speeds are available. For a majority of the time-independent models this extra information is redundant, since reliable analytical relations exist between all three quantities. For a time-dependent model, however, it might be interesting to analyze how this information can be used to improve the quality of the estimation.

The collection of information about the current traffic situation is an interesting research area in itself. A natural continuation of the research presented in Paper IV and V is the development of new methods for allocating link flow detectors to the network, with the purpose to ensure the best possible quality of the OD-matrix to be estimated.

References

- Abrahamsson, T. (1998) “Estimation of origin–destination matrices using traffic counts — a literature survey”, Interim Report IR-98-021, International Institute for Applied Systems Analysis, Laxenburg, Austria.
- Ashok, K., and M.E. Ben-Akiva (1993) ‘Dynamic origin–destination matrix estimation and prediction for real-time traffic management systems’, in: *Transportation and Traffic Theory*, Proceedings of the 12th International Symposium on Transportation and Traffic Theory, Berkeley, California, 21–23 July, 1993, C.F. Daganzo (ed.), Elsevier, pp 465–484.
- Ashok, K., and M.E. Ben-Akiva (2000) “Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows”, *Transportation Science* 34, pp 21–36.
- Ashok, K., and M.E. Ben-Akiva (2002) “Estimation and prediction of time-dependent origin–destination flows with a stochastic mapping to path flows and link flows”,

Transportation Science 36, pp 184–198.

Balakrishna, R. (2006) “Off-line calibration of dynamic traffic assignment models”, PhD thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Boston, Massachusetts.

Balakrishna, R., Ben-Akiva, M., and H.N. Koutsopoulos (2006) “Time-dependent origin–destination estimation without assignment matrices”, in: *Methodological Issues in Transport Simulation*, Proceedings of the International Symposium of Transport Simulation (ISTS06), September 4th–6th, 2006, Lausanne, Switzerland, Paper 0059 .

Barceló, J. (1997) “A survey of some mathematical programming models in transportation”, *Sociedad de Estadística e Investigación Operativa Top* 5, pp 1–40.

Barceló, J., and J. Casas (1999) “The use of neural networks for short-term prediction of traffic demand”, in: *Transportation and Traffic Theory*, Proceedings of the 14th International Symposium on Transportation and Traffic Theory, Jerusalem, Israel, July 20–23, 1999, A. Ceder (ed.), Pergamon, pp 419–443.

Bard, J.F. (1998) *Practical Bilevel Optimization — Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Bates, J. (2000) “History of demand modelling”, in: *Handbook of Transport Modelling*, D.A. Hensher and K.J. Button (eds.), Elsevier, Oxford, England, pp 11–33.

Beckmann, M., McGuire, C.B., and C.B. Winsten (1956) *Studies in the Economics of Transportation*, Yale University Press, New Haven, Connecticut.

Bell, M.G.H. (1983) “The estimation of an origin–destination matrix from traffic counts”, *Transportation Science* 17, pp 198–217.

Bell, M.G.H. (1984) “Log-linear models for the estimation of origin–destination matrices from traffic counts: an approximation”, in: *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*, J. Volmüller and R. Hamerslag (eds.), VNU Science Press, Utrecht, The Netherlands, pp 451–470.

Bell, M.G.H. (1991) “The real time estimation of origin–destination flows in the presence of platoon dispersion”, *Transportation Research B* 25, pp 115–125.

Bell, M.G.H., and Y. Iida (1997) *Transportation Network Analysis*, Wiley, West Sussex, United Kingdom.

Bell, M., Inaudi, D., Lange, J., and M. Maher (1991) “Techniques for the dynamic estimation of O–D matrices in traffic networks”, in: *Advanced Telematics in Road Transport*, Proceedings of the DRIVE Conference, Brussels, Belgium, February 4–6, 1991, Vol. 2, Elsevier, pp 1040–1056.

Ben-Akiva, M. (1987) “Methods to combine different data sources and estimate

origin–destination matrices”, in: *Transportation and Traffic Theory*, N.H. Gartner and N.H.M. Wilson (eds.), Elsevier, pp 459–481.

Ben-Akiva, M., and S.R. Lerman (1985) *Discrete Choice Analysis — Theory and Application to Travel Demand*, MIT Press, Cambridge, Massachusetts.

Bianco, L., Confessore, G., and P. Reverberi (2001) “A network based model for traffic sensor location with implications on o/d matrix estimates”, *Transportation Science* 35, pp 50–60.

Bierlaire, M. (2002) “The total demand scale: a new measure of quality for static and dynamic origin–destination trip tables”, *Transportation Research B* 36, pp 837–850.

Bierlaire, M., and F. Crittin (2004) “An efficient algorithm for real-time estimation and prediction of dynamic od tables”, *Operations Research* 52, pp 116–127.

Bierlaire, M., and P.L. Toint (1995) “Meuse: an origin–destination matrix estimator that exploits structure”, *Transportation Research B* 29, pp 47–60.

Brenninger-Göthe, M., Jörnsten, K., and J.T. Lundgren (1989) “Estimation of origin–destination matrices from traffic counts using multiobjective programming formulations”, *Transportation Research B* 23, pp 257–269.

Camus, R., Cantarella, G.E., and D. Inaudi (1997) “Real-time estimation and prediction of origin–destination matrices per time slice”, *International Journal of Forecasting* 13, pp 13–19.

Carey, M., Hendrickson, C., and K. Siddharthan (1981) “A method for direct estimation of origin/destination trip matrices”, *Transportation Science* 15, pp 32–49.

Cascetta, E. (1984) “Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator”, *Transportation Research* 18, pp 289–299.

Cascetta, E., Inaudi, D., and G. Marquis (1993) “Dynamic estimators of origin–destination matrices using traffic counts”, *Transportation Science* 27, pp 363–373.

Cascetta, E., and M. Postorino (2001) “Fixed point approaches to the estimation of o/d matrices using traffic counts on congested networks”, *Transportation Science* 35, pp 134–147.

Chang, G.L., and X. Tao (1996) “Estimations of dynamic O-D distributions for urban networks”, in: *Transportation and Traffic Theory*, Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France, 1996, J.B. Lesort (ed.), Pergamon, pp 1–20.

Chang, G.L., and J. Wu (1994) “Recursive estimation of time-varying origin–destination flows from traffic counts in freeway corridors”, *Transportation Research B* 28, pp 141–160.

Chen, Y. (1994) "Bilevel programming problems: Analysis, algorithms and applications", PhD thesis, Publication No. 984, Centre de Recherche sur les Transports, Université de Montréal, Montreal, Canada.

Clegg, J., Smith, M., Xiang, Y., and R. Yarrow (2001) "Bilevel programming applied to optimising urban transportation", *Transportation Research B* 25, pp 41–70.

Codina, E., and J. Barceló (2004) "Adjustment of O–D trip matrices from observed volumes: an algorithmic approach based on conjugate directions", *European Journal of Operational Research* 155, pp 535–557.

Contram (2002) "CONTRAM8: User Manual for Contram 8 Version 1", Mott MacDonald Limited and TRL Limited, Doc Ref 26661/T&P/WIN/02A, Winchester, United Kingdom.

Cremer, M., and H. Keller (1987) "A new class of dynamic methods for the identification of origin–destination flows", *Transportation Research B* 21, pp 117–132.

Daganzo, C.F., and Y. Sheffi (1977) "On stochastic models of traffic assignment", *Transportation Science* 11, pp 253–274.

Davis, G.A. (1993) "A statistical theory for estimation of origin–destination parameters from time-series of traffic counts", in: *Transportation and Traffic Theory*, Proceedings of the 12th International Symposium on Transportation and Traffic Theory, Berkeley, California, 21–23 July, 1993, C.F. Daganzo (ed.), Elsevier, pp 441–463.

Davis, G.A., and N.L. Nihan (1991) "Stochastic process approach to the estimation of origin–destination parameters from time series of traffic counts", *Transportation Research Record* 1328, pp 36–42.

Doblas, J., and F.G. Benitez (2005) "An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix", *Transportation Research B* 39, pp 565–591.

Domencich, T., and D. McFadden (1975) *Urban Travel Demand: A Behavioral Analysis*, North Holland, Amsterdam, The Netherlands.

Dougherty, M. (1995) "A review of neural networks applied to transport", *Transportation Research C* 3, pp 247–260.

Dougherty, M.S., and M.R. Cobbett (1997) "Short-time inter-urban traffic forecasting using neural networks", *International Journal of Forecasting* 13, pp 21–31.

Dynasmart (2002) "DYNASMART-P User's guide", Version 0.930.0, Maryland Transportation Initiative, University of Maryland, College Park, Maryland.

Emme/2 (1999) "EMME/2 User's manual", Software Release 9, Inro Consultants Inc., Montreal, Canada.

- Erlander, S., Nguyen, S., and N. Stewart (1979) “On the calibration of the combined distribution-assignment model”, *Transportation Research B* 13, pp 259–267.
- Erlander, S., and N.F. Stewart (1990) *The Gravity Model in Transportation Analysis — Theory and Extentions*, VSP, Utrecht, The Netherlands.
- McFadden, D. (1973) “Conditional logit analysis of qualitative choice behavior”, in: *Frontiers in econometrics*, P. Zarembka (ed.), Academic Press, pp 105–142.
- Fisk, C.S. (1988) “On combining maximum entropy trip matrix estimation with user optimal assignment”, *International Transactions in Operational Research* 2, pp 165–179.
- Fisk, C.S., and D.E. Boyce (1983) “A note on trip matrix estimation from link traffic count data”, *Transportation Research B* 17, pp 245–250.
- Florian, M., and Y. Chen (1995) “A coordinate descent method for the bi-level O–D matrix adjustment problem”, *Transportation Research B* 22, pp 69–73.
- Frank, M., and P. Wolfe (1956) “An algorithm for quadratic programming”, *Naval Research Logistics Quarterly* 3, pp 95–110.
- Garber, N.J., and L.A. Hoel (1999) *Traffic and Highway Engineering*, 2nd edition, Brooks/Cole Publishing Company, Pacific Grove, California:
- Gur, Y., Turnquist, M., Schneider, M., and L. Leblanc (1980) “Estimation of an origin–destination trip table based on observed link volumes and turning movements”, Vol 1. Report RD-80/034, FHWA, U.S. Department of Transportation, Washington, District of Columbia.
- Han, S. (2003) “Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks”, *Transportation Research B* 37, pp 225–249.
- Hazelton, M.L. (2000) “Estimation of origin–destination matrices from link flows on uncongested networks”, *Transportation Research B* 34, pp 549–566.
- Hazelton, M.L. (2003) “Some comments on origin–destination matrix estimation”, *Transportation Research A* 37, pp 811–822.
- Jiao, P., and H. Lu (2005) “Dynamic origin–destination flows estimation for freeway corridors using genetic algorithm”, *Journal of the Eastern Asia Society for Transportation Studies* 6, pp 2702–2717.
- Jörnsten, K., and S. Nguyen (1979) “On the estimation of a trip matrix from network data”, Technical report LiTH-MAT-R-79-36, Department of Mathematics, Linköping University, Linköping, Sweden.
- Khisty, C.J., and B.K. Lall (1998) *Transportation Engineering — an Introduction*, 2nd edition, Prentice Hall, Upper Saddle River, New Jersey.

Kirby, H.R., Watson, S.M., and M.S. Dougherty (1997) “Should we use neural networks or statistical models for short-term motorway traffic forecasting?”, *International Journal of Forecasting* 13, pp 43–50.

Li, J., and B. de Moor (2002) “Dynamic identification of origin–destination matrices in the presence of incomplete observations”, *Transportation Research B* 36, pp 37–57.

Lind, G. (1997) “Strategic assessment of intelligent transport systems”, PhD thesis, TRITA-IP FR 97-29, Department of Infrastructure and Planning, Royal Institute of Technology, Stockholm, Sweden.

Lindveld, K. (2003) “Dynamic O-D matrix estimation: a behavioural approach”, PhD thesis, Trail Thesis Series, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands.

Lundgren J.T. (1989) *Optimization approaches to travel demand modelling*, PhD thesis, Linköping Studies in Science and Technology, Dissertations, No. 207, Department of Mathematics, Linköping University, Linköping, Sweden.

Maher, M.J., and X. Zhang (1999) “Algorithms for the solution of the congested trip matrix estimation problem”, in: *Transportation and Traffic Theory*, Proceedings of the 14th International Symposium on Transportation and Traffic Theory, Jerusalem, Israel, July 20–23, 1999, A. Ceder (ed.), Pergamon, pp 445–469.

Maher, M.J., Zhang, X., and D. van Vliet (2001) “A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows”, *Transportation Research B* 35, pp 23–40.

Merchant, D.K., and G.L. Nemhauser (1978a) “A model and an algorithm for the dynamic traffic assignment problems”, *Transportation Science* 12, pp 183–199.

Merchant, D.K. and G.L. Nemhauser (1978b) “Optimality conditions for a dynamic traffic assignment model”, *Transportation Science* 12, pp 200–207.

Mozolin, M., Thill, J.C., and E.L. Usery (2000) “Trip distribution forecasting with multilayer perceptron neural networks: a critical evaluation”, *Transportation Research B* 34, pp 53–73.

McNeil, S., and C. Hendrickson (1985) “A regression formulation of the matrix estimation problem”, *Transportation Science* 19, pp 278–292.

Nguyen, S. (1977) “Estimating an OD matrix from network data: A network equilibrium approach”, Publication No. 60, Centre de Recherche sur les Transports, Université de Montréal, Montreal, Canada.

Nie, Y., Zhang, H.M., and W.W. Recker (2005) “Inferring origin–destination trip matrices with a decoupled GLS path flow estimator”, *Transportation Research B* 39, pp 497–518.

- Nihan, N.L, and G.A. Davis (1987) “Recursive estimation of origin–destination matrices from input/output counts”, *Transportation Research B* 21, pp 149–163.
- Nihan, N.L, and G.A. Davis (1989) “Application of prediction-error minimization and maximum likelihood to estimate intersection O–D matrices from traffic counts”, *Transportation Science* 22, pp 77–90.
- Oppenheim, N. (1995) *Urban Travel Modelling — From Individual Choices to General Equilibrium*, John Wiley, New York, New York.
- Ortúzar, J. de D., and L.G. Willumsen (2001) *Modelling Transport*, 3rd Edition, Wiley, West Sussex, United Kingdom.
- Papacostas, C.S., and P.D. Prevedouros (2001) *Transportation Engineering and Planning*, 3rd edition, Prentice Hall, Upper Saddle River, New Jersey.
- Patriksson, M. (1994) *The Traffic Assignment Problem — Models and Methods*, VSP, Utrecht, The Netherlands.
- Patriksson, M. (2004) “Sensitivity analysis of traffic equilibria”, *Transportation Science* 38, pp 258–281.
- Peeta, S., and A.K. Ziliaskopoulos (2001) “Foundations of dynamic traffic assignment: the past, the present and the future”, *Networks and Spatial Economics* 1, pp 233–265.
- Peterson, A. (2003) “Origin–destination matrix estimation from traffic counts”, licentiate thesis, Linköpings Studies in Science and Technology. Theses, No. 1057, Department of Science and Technology, Linköpings universitet, Norrköping, Sweden.
- Sheffi, Y. (1985) *Urban Transportation Networks — Equilibrium Analysis with Mathematical Programming Methods*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Sherali, H.D., Arora, N., and A.G. Hobeika (1997) “Parameter optimization methods for estimating dynamic origin–destination trip-tables”, *Transportation Research B* 31B, pp 141–157.
- Sherali, H.D. and T. Park (2001) “Estimation of dynamic origin–destination trip tables for a general network”, *Transportation Research B* 35, pp 217–235.
- Sherali, H.D., Narayanan, A., and R. Sivanandan (2003) “Estimation of origin–destination trip-tables based on a partial set of traffic link volumes”, *Transportation Research B* 37, pp 815–836.
- Sherali, H.D., Sivanandan, R., and A.G. Hobeika (1994) “A linear programming approach for synthesizing origin–destination trip tables from link traffic volumes”, *Transportation Research B* 28, pp 213–233.
- Sjöstrand, H. (2001) *Passenger assessments of quality in local public transport —*

measurement, variability and planning implications, PhD Thesis, Bulletin 202, Department of Technology and Society, Lund Institute of Technology, Lund, Sweden.

Spiess, H. (1987) “A maximum likelihood model for estimating origin–destination matrices”, *Transportation Research B* 21, pp 395–412.

Spiess, H. (1990) “A gradient approach for the O–D matrix adjustment problem”, CRT Pub. No 693, Centre de Recherche sur les Transports, Universit de Montreal, Montreal, Canada.

Statistics Sweden (2007), <http://www.scb.se/>, download 2007-02-05

Tamin, O.Z., and L.G. Willumsen (1989) “Transport demand model estimation from traffic counts”, *Transportation* 16, pp 3–26.

Tavana, H. (2001) ”Internally-consistent estimation of dynamic network origin–destination flows from intelligent transportation systems data using bi-level optimization”, PhD thesis, Department of Civil Engineering, University of Texas at Austin, Austin, Texas.

Tobin, R.L., and T.L. Friesz (1988) “Sensitivity analysis for equilibrium network flow”, *Transportation Science* 22, pp 242–250.

Vythoulkas, P.C. (1993) “Alternative approaches to short term traffic forecasting for use in driver information systems”, in: *Transportation and Traffic Theory*, Proceedings of the 12th International Symposium on Transportation and Traffic Theory, Berkeley, California, 21–23 July, 1993, C.F. Daganzo (ed.), Elsevier, pp 485–506.

Wardrop, J.G. (1952) “Some theoretical aspects of road traffic research”, *Proceedings of the Institute of Civil Engineers*, Part II, Road Paper No. 36, pp 325–378.

Willumsen, L.G. (1981) “Simplified transport models based on traffic counts”, *Transportation* 10, pp 257–278.

Willumsen, L.G. (1984) “Estimating time-dependent trip matrices from traffic counts”, in: *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*, J. Volmüller and R. Hamerslag (eds.), VNU Science Press, Utrecht, The Netherlands, pp 397–411.

Wilson, A.G. (1974) *Urban and Regional Models in Geography and Planning*, John Wiley and Sons, Bristol, United Kingdom.

Wolfe, P. (1975) “A method of conjugate subgradients for minimizing nondifferentiable optimization”, *Mathematical Programming Study* 3, pp 145–173.

Wright, P.H., and N.J. Ashford (1998) *Transportation Engineering — Planning and Design*, 4th edition, John Wiley and Sons, New York, New York.

Wu, J. (1997) “A real-time origin–destination matrix updating algorithm for on-line

- applications”, *Transportation Research B* 31, pp 381–396.
- Yang, H. (1995) “Heuristic algorithms for the bilevel origin–destination matrix estimation problem”, *Transportation Research B* 29, pp 231–242.
- Yang, H., Meng, Q., and M.G.H. Bell (2001) “Simultaneous estimation of the origin–destination matrices and travel-cost coefficient for congested network in a stochastic user equilibrium”, *Transportation Science* 35, pp 107–123.
- Yang, H., Sasaki, T., Iida, Y., and Y. Asakura (1992) “Estimation of origin–destination matrices from link traffic counts on congested networks”, *Transportation Research B* 26, pp 417–434.
- Yang, H., and S. Yagar (1995) “Traffic assignment and signal control in saturated road networks”, *Transportation Research A* 29, pp 125–139.
- Zhou, X., and H.S. Mahmassani (2006) “Dynamic origin–destination demand estimation using automatic vehicle identification data”, *IEE Transactions on Intelligent Transportation Systems* 7, pp 105–114.
- van der Zijpp, N.J. (1997) “Dynamic origin–destination matrix estimation from traffic counts and automated vehicle identification data”, *Transportation Research Record* 1607, pp 87–94.
- van der Zijpp, N.J., and R. Hamerslag (1996) “Improved Kalman filtering approach for estimating origin–destination matrices for freeway corridors”, *Transportation Research Record* 1443, pp 54–64.
- van Zuylen, H. (1978) “The information minimising method: validity and applicability to transportation planning”, in: *New Developments in Modelling Travel Demand and Urban Systems*, G.R.H. Jansen et al. (eds.), Saxon, Farnborough, United Kingdom.
- van Zuylen, H.J., and D.M. Branston (1982) “Consistent link flow estimation from counts”, *Transportation Research B* 16, pp 473–476.
- van Zuylen, H., and L.G. Willumsen (1980) “The most likely trip matrix estimated from traffic counts”, *Transportation Research B* 14, pp 281–293.

Probabilistic Models For Mobile Phone Trajectory Estimation

by

Arvind Thiagarajan

S.M., Computer Science, Massachusetts Institute of Technology (2007)

B.Tech., Computer Science and Engineering, Indian Institute of Technology Madras
(2005)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

© Massachusetts Institute of Technology 2011. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
September 2nd, 2011

Certified by
Hari Balakrishnan
Professor of Computer Science and Engineering
Thesis Supervisor

Certified by
Samuel R. Madden
Associate Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Probabilistic Models For Mobile Phone Trajectory Estimation
by
Arvind Thiagarajan

Submitted to the Department of Electrical Engineering and Computer Science
on September 2nd, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

This dissertation is concerned with the problem of determining the track or trajectory of a mobile device — for example, a sequence of road segments on an outdoor map, or a sequence of rooms visited inside a building — in an *energy-efficient* and *accurate* manner.

GPS, the dominant positioning technology today, has two major limitations. First, it consumes significant power on mobile phones, making it impractical for continuous monitoring. Second, it does not work indoors. This dissertation develops two ways to address these limitations: (a) sub-sampling GPS to save energy, and (b) using alternatives to GPS such as WiFi localization, cellular localization, and inertial sensing (with the accelerometer and gyroscope) that consume less energy and work indoors. The key challenge is to match a sequence of *infrequent* (from sub-sampling) and *inaccurate* (from WiFi, cellular or inertial sensing) position samples to an accurate output trajectory.

This dissertation presents three systems, all using probabilistic models, to accomplish this matching. The first, *VTrack*, uses Hidden Markov Models to match noisy or sparsely sampled geographic (*lat, lon*) coordinates to a sequence of road segments on a map. We evaluate *VTrack* on 800 drive hours of GPS and WiFi localization data collected from 25 taxicabs in Boston. We find that *VTrack* tolerates significant noise and outages in location estimates, and saves energy, while providing accurate enough trajectories for applications like travel-time aware route planning.

CTrack improves on *VTrack* with a Markov Model that uses “soft” information in the form of raw WiFi or cellular signal strengths, rather than geographic coordinates. It also uses movement and turn “hints” from the accelerometer and compass to improve accuracy. We implement *CTrack* on Android phones, and evaluate it on cellular signal data from over 126 (1,074 miles) hours of driving data. *CTrack* can retrieve over 75% of a user’s drive accurately on average, even from highly inaccurate (175 metres raw position error) GSM data.

iTrack uses a particle filter to combine inertial sensing data from the accelerometer and gyroscope with WiFi signals and accurately track a mobile phone indoors. *iTrack* has been implemented on the iPhone, and can track a user to within less than a metre when walking with the phone in the hand or pants pocket, over 5× more accurately than existing WiFi localization approaches. *iTrack* also requires very little manual effort for training, unlike existing localization systems that require a user to visit hundreds or thousands of locations in a building and mark them on a map.

Thesis Supervisor: Hari Balakrishnan
Title: Professor of Computer Science and Engineering

Thesis Supervisor: Samuel R. Madden
Title: Associate Professor of Computer Science and Engineering

To my parents, to whom I owe everything.

Acknowledgments

First, I must thank both my advisors, Hari Balakrishnan and Sam Madden, without whose guidance, support and mentoring (each in their own unique way) this thesis would not have been possible.

Hari has been an inspiration from when I interned with him in my undergraduate days, through all of graduate school, and I have learned more about how to approach research from him than anyone else. His high standards, optimistic approach and exhortations to jump in and tackle the hardest and most meaningful problems have helped me become progressively more adventurous, risk-taking and confident in my research as well, and most importantly helped me have a lot of fun doing it. His excellent writing and presentation skills have also helped me learn a little about how to give talks and write papers. Last but not least, he has been a very supportive mentor, never hesitating to give help immediately when asked.

Sam has been an extremely helpful and encouraging mentor, and I can never thank him enough for his positive feedback, encouragement and constant support. He has always had the time and willingness to sit down and discuss any detail of a research problem, be it over person, phone, or e-mail, and provide much needed feedback, structure and guidance on concrete directions to pursue and baby steps to take towards solving a big problem. His guidance on FunctionDB and my Masters thesis, in particular, helped me a lot with getting started with research. What never ceases to amazes me is that I would often go to him for advice on a problem I'd spent a long time thinking about without success, and he'd instantly point me in the right direction after thinking for a minute or two. It was his vision and energy that helped transform VTrack from an algorithm in the lab to a real system that continuously processes data from the Cartel testbed and computes traffic delays. I also owe a lot to him for teaching me how to hack iPhone apps in the early stages of working on iTrack, which has been rewarding and fun in its own right.

I thank Seth Teller for serving on my thesis committee at incredibly short notice, reading and offering suggestions and comments. A lot of my work on indoor localization has been inspired by previous work he has done in this space. I thank Yoni Battat and Seth for their help with securing digital floorplans of the Stata Center for our iTrack project, and Sachi Hemachandra for allowing me to test drive their awesome robotic wheelchair.

Lenin Ravindranath has been the most incredible collaborator and lab-mate I could have asked for. His energy levels, propensity to generate new ideas every minute, and level of excitement about research are absolutely infectious, and hopefully this has rubbed off a little on me too. I would say it was his coming to MIT that reminded me that research can and should be fun, and helped shape my thesis topic at a time when I was still shopping around for ideas to work on. I have really enjoyed working with him and learning from him — both hacking and brainstorming, on a number of projects, including VTrack, CTrack and Code in the Air. Even the tiniest details of research and smallest of problems (like how to tweak an algorithm or design an experiment) turned into long and fun brainstorming sessions with him, and led to unexpected new directions to pursue.

Jakob Eriksson, my former office-mate, has been an awesome collaborator from the day he came to MIT. I admire Jakob's application-driven approach to research and his ability to quickly build useful, real systems at scale. It was his painstaking work that helped get PlanetTran taxicabs on board and build the Cartel testbed. The testbed has supplied me and countless other students with a rich vein of real data, without which this thesis would simply not exist. Later, I really enjoyed traveling to Chicago and working with him and his students James and Tomas on the Transitgenie bus tracking project, in a collegial, friendly and fun atmosphere. He has always been willing to help with advice on wireless networks, system building and the like, and we have had many fruitful and interesting discussions.

I thank Sivan Toledo and Katrina LaCurts for helping out with the experiments and writing for

the VTrack paper, and Bret Hull for access to Cartel data for the paper, as well as for patiently answering all my questions on Cartel. I also thank Sejoon Lim for his help with sharing data and thoughts, and getting me started with thinking about the map-matching problem.

Lewis Girod has been an extremely helpful resource on all things hardware related, and it has always been instructive to talk to him. I learn something new each time I go to his office. He was instrumental in getting me started off with research by mentoring me on the WaveScope project, and I am grateful to him for that.

I thank PlanetTran and Seth Riney for providing us with data and agreeing to run our software and devices in their cabs. Seth was always available and prompt to help us re-power and reboot the devices on-site when they failed or experienced errors.

I thank all the anonymous reviewers and shepherds of the papers I have co-authored over the years. Their comments and constructive suggestions have helped make this thesis stronger.

I thank Sheila Marian for her prompt help with all administrative issues — travel, reimbursements, visa paperwork, and many more. Cannot neglect to mention her awesome cakes.

Dorothy Curtis has been a life saver on innumerable occasions when my laptop or the server has gone down, offering prompt help and fixes. She has also been helpful whenever we needed to sort out issues with the Cartel database, free up disk space or buy new hardware.

I thank my office mates over the years at MIT: Emil Sit, Mike Walfish, Mythili Vutukuru, Lenin Ravindranath, Katrina LaCurts, Raluca Popa, Ramki Gummadi, Shuo Deng and Jakob Eriksson for their noise tolerance (particularly to my feverish keystrokes!), and many, many fun discussions. Mike, Jakob and Ramki in particular have given me plenty of high-level guidance on writing, presentation and picking problems to work on.

I thank all my other floor-mates and friends from CSAIL over the years — Adam Marcus, Arnab Bhattacharya, Dan Abadi, Dan Myers, Micah Brodsky, Alvin Cheung, Eugene Wu, Evan Jones, Eddie Nikolova, Vladimir Bychkovsky, Allen Miu, Stan Rost, Rahul Hariharan, George Huo, Jayashree Subramaniam, Yuan Mei, Bret Hull, Calvin Newport, Ramesh Chandra and others I've no doubt left out, for the many many engaging and fun discussions, help with practice talks, support and advice.

Paresh Malur and Tim Kaler have both been amazingly committed and fun undergrads to work with, and I have really enjoyed working, hacking and brainstorming with them on topics ranging from traffic prediction to smartphones.

I thank Michel Gorazcko, who was instrumental in helping me get set up, guiding me, and helping me learn the ropes of research when I first came to MIT as a wide-eyed undergrad intern seven years ago.

I thank my former internship mentors at Yahoo! Research — Utkarsh Srivastava, Brian Cooper, Adam Silberstein and Raghu Ramakrishnan for helping me get excited about research at a time when I was a bit unsure of what to work on. I enjoyed my experience at Yahoo! working on the PNUTS system a lot, thanks in no small part to the people there. I also thank Roopesh Ranjan, whom I first met during that internship, for being an awesome friend and sounding board over the years.

A special thanks to the baristas at the Forbes cafe in the Stata Center where I (and I imagine, many others like me) have faithfully consumed morning coffee over my years at MIT. Without you, no research would probably happen in this building.

I thank Krishna, Prabha, Jayku, Vivek, Mythili, Lavanya and Aditi for the awesome company, great home-cooked food, and many great memories during my stay in Ashdown. The cooking group, outings, and nightly discussions on random topics were, and will remain, some of the best times of my PhD experience at MIT. Vivek Jaiswal's incredibly tasty home-made *dal* and *fried rice* in particular sustained me for the better part of two years, I think. I also thank Aditi, Basky and

Shashi for the fun times talking music, and practising, performing and planning musical pieces for Diwali Nite, the Ashdown Concert and the like. I thank Aruna, Murali, Arvind Shankar, Karthik, and Anna for the fun times, outings, potlucks and ski trips during my last two years at MIT. I also thank my apartment mates (and friends) Vijay, Pranesh, Sharat and Ivan for the company and good times over the years.

I thank Krishna for introducing me to the incredible ocean that is Carnatic Music, which has become a big part of my life over the years. I can never be indebted enough to him for that. I thank my guru, Smt. Geetha Murali, for helping me get a real foothold in learning music and teaching me patiently in spite of my deficiencies. I thank Hari Arthanari for the many great musical memories — all the padams, Balu's concerts, and the *impromptu* Brindamma and Vishwa listening sessions we used to have in his car. I also thank Chintan Vaishnav for the Bade Ghulam Ali Khan and Amir Khan sessions. It would not be out of place here to thank the great masters of Carnatic Music who have (posthumously) given me innumerable hours of listening pleasure.

I thank Raju Mahodaya and Sharada Mahodaya for teaching me whatever little of Samskritam (Sanskrit) I know today, and Raju in particular for exhorting me to take up teaching Samskritam to learn it better. The Wednesday sessions of teaching and learning about this incredible language were a welcome break from the routine of graduate student life.

I thank my close friends, Vikram and Kamesh, for their support and amazing level of belief in me at all times, and for the many fun times over the years.

I am indebted to all my near and far family members. In particular my *Patti* (grandmother) Chajja, Sudhaman *Thatha*, my *Kollu Patti* (great-grandmother) Dhamma, and Jayamani and Jalaja *Patti* for their support, love, advice, never-shaking belief, and their constant prayers for me. Seethalakshmi *Patti* has always been razor sharp about making sure I was focused on finishing my PhD, as has Krishnamoorthy *Thatha*, and their contribution to my finishing is as big as anything else.

My wife Poorna's selfless love, affection, patience, home cooked meals, filter coffee, and constant support and belief in me while juggling her own exam preparations and volunteer work during my last year of PhD research at MIT is what has made this thesis possible. My in-laws' love, support and encouragement has also meant a lot to me during this time.

I owe everything in life that is good to my parents. Some debts are simply not possible to repay. *Appa* and *Amma* have been rock solid in their emotional support, optimism, constant encouragement and even detailed technical advice from thousands of miles away — through the worst of times, and the lowest of the lows. *Appa* has often been more excited and enthusiastic about my research than I have been. Both of them have had unshakeable belief in me. I dedicate this thesis to them.

Contents

1	Introduction	17
1.1	Why Is Trajectory Estimation Hard?	17
1.2	Contributions Of This Dissertation	22
1.3	VTrack: Map-Matching Noisy and Sparse Coordinates	23
1.4	CTrack: Accurate Tracks From Soft Information	25
1.5	iTrack: Accurate Indoor Tracks From Inertial Sensing and WiFi	26
1.6	Road Map	27
2	Energy Studies	29
2.1	Theoretical Factors Driving Energy Cost	29
2.2	Energy Experiments	34
2.3	Conclusion	38
3	Map-Matching With Markov Models	39
3.1	Background: Traffic Monitoring	42
3.2	Applications	43
3.3	<i>VTrack</i> Architecture	44
3.4	The <i>VTrack</i> Algorithm	45
3.5	Travel Time Estimation	54
3.6	Evaluation of <i>VTrack</i>	56
3.7	Revisiting the Energy Question	71
3.8	Related Work	74
3.9	Conclusion	74
4	Map-Matching With Soft Information	77
4.1	Why Cellular?	77
4.2	How CTrack Works	77
4.3	Why Soft Information Helps	78
4.4	<i>CTrack</i> Architecture	81
4.5	<i>CTrack</i> Algorithm	82
4.6	Sensor Hint Extraction	88
4.7	Evaluation	93
4.8	Related Work	103
4.9	Conclusion	104

5 Indoor Trajectory Mapping	105
5.1 The Training Challenge	105
5.2 <i>iTrack</i> And Contributions	106
5.3 Inertial Phone Sensors	107
5.4 Inertial Navigation	109
5.5 The <i>iTrack</i> System	115
5.6 Simplifying Training With <i>iTrack</i>	134
5.7 Implementation	135
5.8 Evaluation	138
5.9 Related Work	148
5.10 Conclusion	150
6 Conclusion	153
6.1 Future Work	154

List of Figures

1-1	Screenshots of two track-based applications for mobile phones.	18
1-2	Screenshot of Carweb, a web application that allows users to visualize and compute statistics for their commute paths.	19
1-3	WiFi and GSM localization are highly error-prone.	21
1-4	WiFi localization is error-prone indoors.	22
1-5	Upward sloping drift in user acceleration measured on an iPhone.	23
2-1	Energy consumption: GPS vs WiFi vs cellular on an Android phone.	38
3-1	iCartel application showing real-time delays and congestion information.	44
3-2	Web application showing traffic delays.	45
3-3	VTrack system architecture.	46
3-4	VTrack server.	47
3-5	Example illustrating an HMM.	49
3-6	The map-matching process used by <i>VTrack</i> . A raw location trace is gradually refined to produce a final, high quality street route. In this example, due to noise and outages, only the first three segments produced quality estimates.	50
3-7	Coverage map of our evaluation drives.	58
3-8	CDF of optimality gap when route planning using <i>VTrack</i> . Larger optimality gaps are worse.	61
3-9	CDF of errors in time estimation for individual segments on WiFi localization estimates.	63
3-10	Map matching errors for WiFi localization.	64
3-11	End-to-end time estimation accuracy using WiFi localization.	64
3-12	Spurious segment rates from map-matching for different sensors and sampling rates.	65
3-13	Success rate of hotspot detection with <i>VTrack</i>	66
3-14	False positive rate of hotspot detection with <i>VTrack</i>	66
3-15	Accuracy of <i>VTrack</i> vs nearest segment matching (denoted by NN).	68
3-16	Speed constraint improves map-matching precision.	69
3-17	<i>VTrack</i> 's transition probability is better than a partitioned transition probability.	70
3-18	Diagram showing optimal strategy as a function of power budget and GPS to WiFi energy cost ratio.	73
4-1	Example demonstrating the benefits of soft information for map-matching.	79
4-2	Geographic spread of exact matches. The dashed line shows the 80th percentile.	80
4-3	<i>CTrack</i> system architecture.	81
4-4	Steps in <i>CTrack</i> algorithm.	83

4-5	<i>CTrack</i> map-matching pipeline. Black lines are ground truth and red points/lines are obtained from cellular fingerprints.	84
4-6	Anomaly detection in accelerometer data.	91
4-7	Movement hint extraction from the accelerometer.	91
4-8	Turn hint extraction from the magnetic compass.	93
4-9	Coverage map of driving data set.	94
4-10	CDF of precision: <i>CTrack</i> is better than <i>Placelab</i> + <i>VTrack</i>	96
4-11	CDF of recall: comparison.	96
4-12	Precision with and without grid sequencing.	98
4-13	Sensor hints from the compass and accelerometer aid map-matching. Red points show ground truth and the black line is the matched trajectory.	99
4-14	Precision with and without sensor hints.	100
4-15	CDF showing precision/recall for hint extraction.	100
4-16	Precision/recall as a function of the amount of training data.	101
4-17	CDF of traversal counts for each road segment, with 40 hrs of training data.	102
5-1	What a smartphone accelerometer measures.	108
5-2	What a smartphone gyroscope measures.	109
5-3	Drift when integrating smartphone accelerometer data.	113
5-4	Angular velocity of an iPhone in a user's pocket when walking.	115
5-5	<i>iTrack</i> system architecture.	116
5-6	Illustration of steps in <i>iTrack</i>	117
5-7	Distinguishing phone-in-hand and phone-in-pocket using Euler angles.	120
5-8	Walking detection using fourier transforms.	121
5-9	Peaks and valleys in acceleration, from phone held in user's hand.	122
5-10	Variation in yaw within each stride, phone in pocket.	124
5-11	Example showing ambiguity in the output of the particle filter.	132
5-12	Knowledge of initial direction helps improve trajectory matching.	133
5-13	<i>iTrack</i> application for the iPhone.	136
5-14	Ground truth setup for evaluating <i>iTrack</i>	140
5-15	Absolute accuracy of <i>iTrack</i> compared to WiFi localization.	142
5-16	More WiFi training data reduces localization error.	143
5-17	Accuracy of <i>iTrack</i> for different amounts of seed data.	144
5-18	Knowing initial position and/or orientation improves accuracy.	145
5-19	Iterative training can improve accuracy of <i>iTrack</i>	146
5-20	Angle and stride length models do not affect localization error.	147

List of Tables

2.1	Energy experiments comparing GPS and WiFi localization on an iPhone 3G.	35
2.2	Battery lifetime: GPS vs accelerometer on the iPhone.	36
3.1	Linear interpolation is as good, or better than SP (shortest paths).	71
3.2	Strategies for different power budgets and GPS to WiFi energy cost ratios.	72
4.1	Windowing and smoothing improve median trajectory matching precision.	98
4.2	<i>CTrack</i> on cellular vs <i>VTrack</i> on 10% duty-cycled WiFi.	103
5.1	Step counts from pocket are more accurate than from hand.	123
5.2	Survival rate of <i>iTrack</i> for different amounts of seed data.	145
5.3	Survival rate of <i>iTrack</i> for different initialization configurations.	146
5.4	Survival rate of <i>iTrack</i> for different error models.	147
5.5	Worst case (99th) percentile localization errors of different approaches.	148

Chapter 1

Introduction

This dissertation is concerned with the problem of estimating the trajectory, or *sequence* of locations visited by a mobile phone, in an accurate and energy-efficient manner.

Two key trends — the proliferation of mobile smartphones, and the availability of a wide variety of location sensors on these phones, in particular global positioning technology (GPS) — have led to the emergence of many mobile applications that use trajectory estimation as a fundamental primitive.

For example, crowd-sourced traffic applications like iCartel [39] (Figure 1-1(b)) collect raw (latitude, longitude) location data from end user smartphones and match them to a sequence of road segments to obtain travel time estimates for individual roads. They then share these estimates with other users, who use them for applications like *traffic-aware route planning*, i.e., finding the best path to a destination given real-time traffic conditions and travel times.

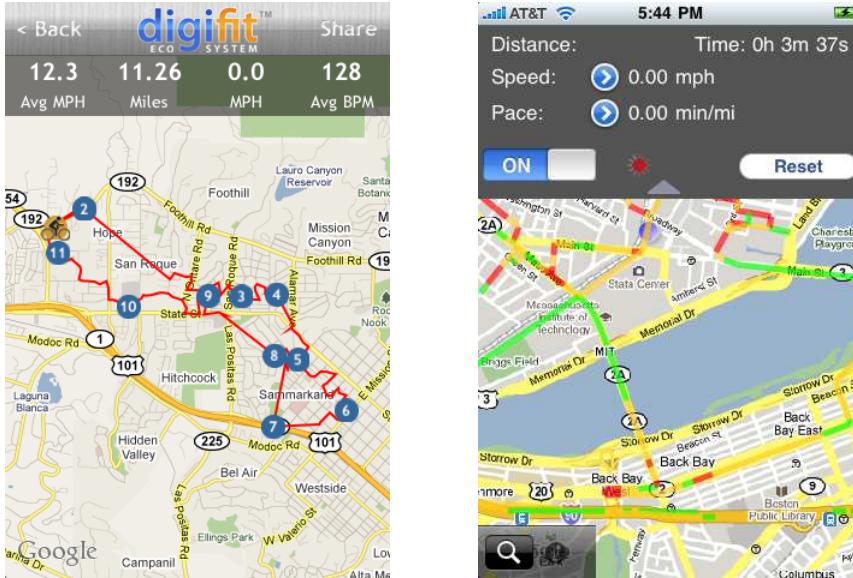
Personal fitness applications like Digitfit [27] (Figure 1-1(a)) and MapMyRun [59] allow smartphone users to record and view their biking or running tracks, calorie consumption and other fitness-related metrics on a map. In a similar vein, personalized commute applications like Carweb (Figure 1-2) allow smartphone users to visualize and optimize their commute paths.

Trajectory estimation is also useful in many indoor sensing and analytics applications. Today's indoor location systems primarily use custom hardware or dedicated sensor devices [62, 74, 30], but are poised to become a reality on commodity mobile smartphones in the near future. For example, researchers have trialed mobile sensing systems that find movement patterns of people indoors, and use them to improve workplace efficiency [18]. There exist commercial products that track the location of patients, doctors or medical equipment indoors to identify and rectify system inefficiencies and improve patient care — for example, the Massachusetts General Hospital has trialled such an analytics system from Radianse [74].

1.1 Why Is Trajectory Estimation Hard?

Trajectory estimation on mobile phones is a non-trivial problem because *today's phones have no known position sensor that is accurate everywhere and energy-efficient*. The position sensors available on phones today include:

- GPS, which uses the time difference of arrival between transmissions from multiple satellites orbiting the earth to calculate the (latitude, longitude) position of a mobile phone accurate to within a few metres.



(a) Digifit: record and view your running route.

(b) iCartel: download traffic delay information and contribute them to other users.

Figure 1-1: Screenshots of two track-based applications for mobile phones.

- WiFi and cellular localization, which use nearby wireless base stations and cellular towers for positioning. A phone looks up the set of nearby base stations or cell towers it can hear, and their signal strengths, in a pre-existing training database (built using GPS) to find its position.
- Inertial MEMS (micro-electro-mechanical) sensors such as accelerometers, gyroscopes and magnetic compasses, which measure the *change* in a mobile phone's position or orientation and can be used to indirectly compute its location.

However, all of the above sensors fail to meet one or the other of three key requirements of the applications mentioned earlier — *accuracy*, *energy-efficiency* and *wide availability*. As we show:

- GPS is accurate, but is not energy-efficient and is not available everywhere.
- WiFi and cellular localization are more energy-efficient than GPS and widely available, but are highly inaccurate.
- Inertial MEMS sensors are highly energy-efficient, but also inaccurate.

We elaborate on each of the above below.

Limitations of GPS. While GPS is accurate to within a few metres outdoors, it is energy-intensive and drains a mobile phone's battery when kept switched on. A common feature of all the applications mentioned earlier is that they need to *continuously* monitor the location of a mobile device to find its trajectory. This is impractical with GPS because it results in poor battery life. For example, from our own experience building and deploying the iCartel system [39], we found that end users' phones barely last for a few hours with GPS switched on all the time. This is unacceptably low and

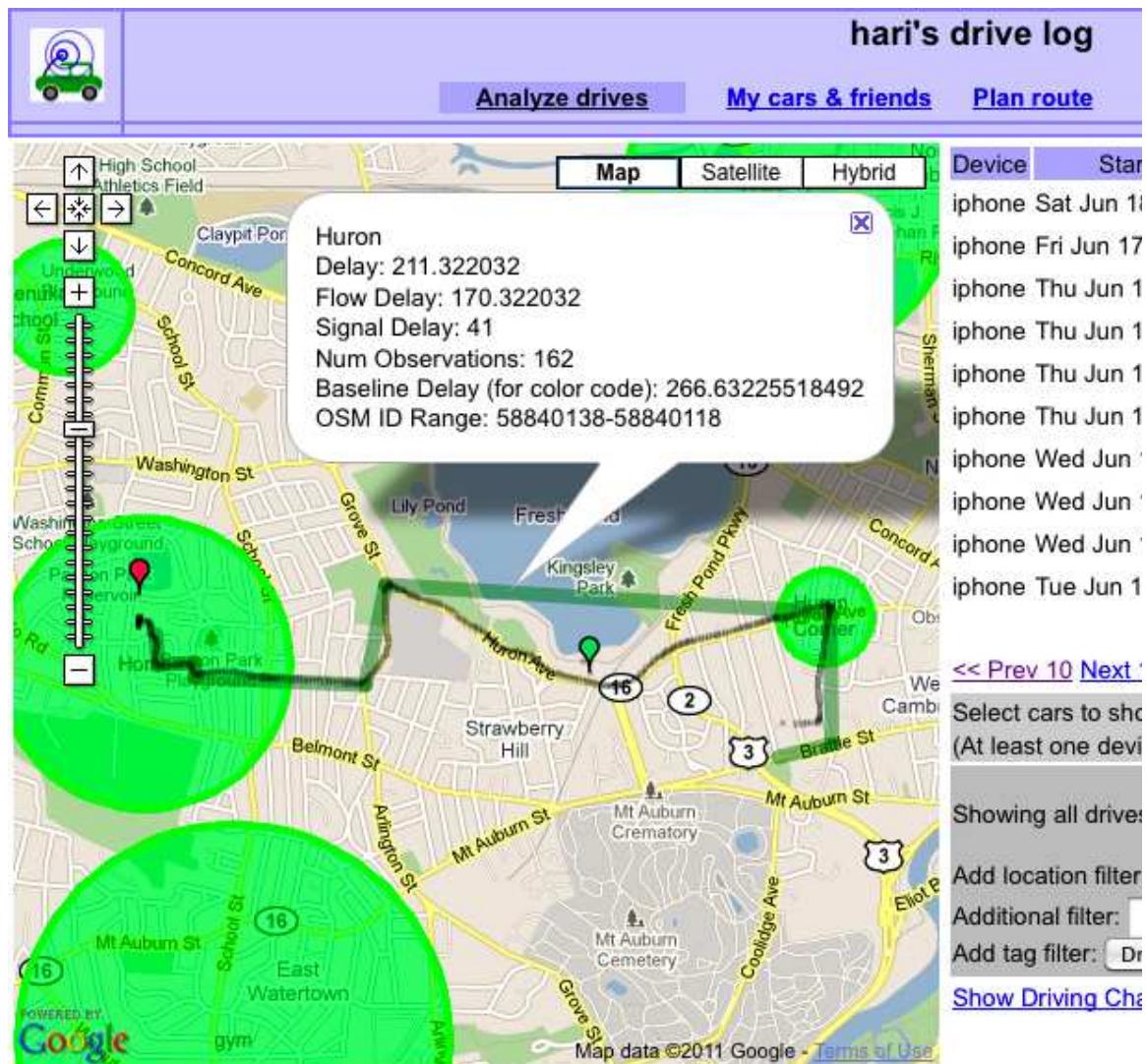


Figure 1-2: Screenshot of Carweb, a web application that allows users to visualize and compute statistics for their commute paths.

makes users reluctant to install the application on their phones. A similar concern exists for all of the outdoor tracking and fitness applications mentioned earlier.

We note that GPS energy consumption is a fundamental limitation that does not show a trend of reducing with improvements in underlying GPS hardware. This is because GPS satellites orbit 11,000 miles above the earth’s surface with a transmission power of approximately 50 W, resulting in only 2×10^{-11} mW/m² of effective radiated power (ERP) at a receiver on the earth’s surface. GPS receivers need sophisticated signal processing to successfully extract information from a signal with such low power (in contrast, the typical cell phone signal has an ERP of 10 mW/m² [35]). This processing is computationally-intensive and consumes significant energy on a mobile device.

The second fundamental limitation of GPS is that it does not work indoors, or in areas with significant obstructions that lower the GPS signal-to-noise ratio, such as downtown areas of cities with many tall skyscrapers. This makes GPS unusable for indoor trajectory estimation, and inaccurate (at times) for outdoor applications as well.

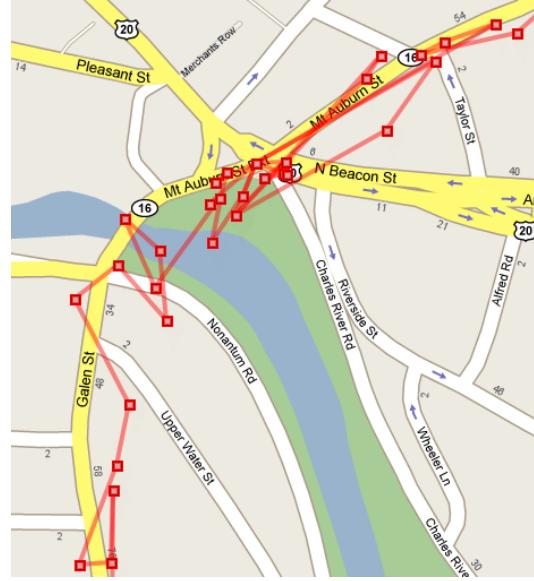
Limitations of WiFi and Cellular. Both WiFi and cellular localization consume less energy than GPS because of the higher ERP of WiFi and cellular signals, which are transmitted from much closer to the earth’s surface than GPS satellites. Cellular localization is particularly attractive because it consumes little or no *marginal energy*: a phone’s cellular receive circuits are typically always on anyway to receive calls.

The flip side is that these sensors are *much* less accurate than GPS, owing to the propagation of WiFi and cellular signals. We find that raw WiFi position estimates can have errors of up to 50-100 metres outdoors, and raw cellular position estimates have errors ranging from a few hundred metres to as much as a kilometre. Figure 1.1 shows examples of WiFi and cellular (GSM) localization estimates. In the figures, the red points are raw locations obtained by matching observed WiFi or cellular signals to their closest match in a pre-built training database. The actual roads traversed are shown in black. It is non-trivial to recover the true sequence of road segments from the raw position estimates, let alone recover accurate travel times for each individual segment.

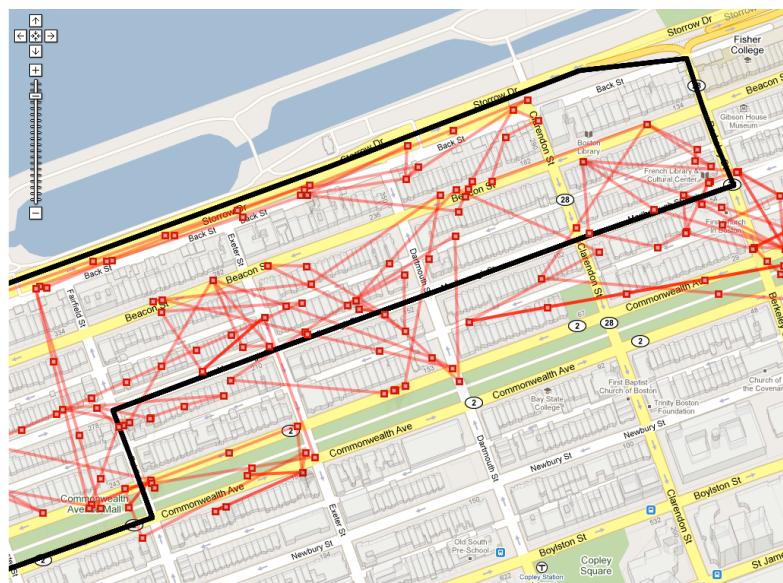
Both WiFi and cellular localization work indoors, where GPS does not work. However, they are also inaccurate indoors. Our experiments show a mean positioning error of 5-6 metres indoors when using WiFi localization, also corroborated by other studies on both WiFi and cellular localization [28, 8]. The real problem is not the 5-6 metre mean location error, but that a significant fraction of position estimates can have errors ranging from 10-20 metres. This is too high to be useful for applications that want to locate where a person is in a store, hospital or workplace, or find room-level location in a building. Figure 1.1 compares the true trajectory walked by a person indoors to the output of WiFi localization. WiFi can approximately find the outline of the trajectory walked, but is subject to significant errors at multiple points in the walk.

A second major limitation is that WiFi localization indoors requires extensive manual training effort to associate known locations to WiFi signal measurements. Outdoors, training is relatively easy because ground truth can be obtained from GPS. Building an indoor map of WiFi, on the other hand, is *much* more tedious and difficult to scale, because training requires dedicated personnel or volunteers to survey hundreds or thousands of locations in a building, and manually mark them on a map to associate them to WiFi signals at that location.

Limitations of MEMS sensors. MEMS sensors such as the accelerometer and the gyroscope are relatively energy-efficient, and they work indoors and outdoors. In theory, it should be possible to

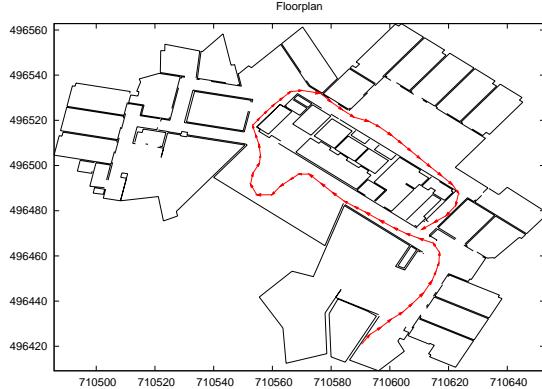


(a) WiFi localization estimates.

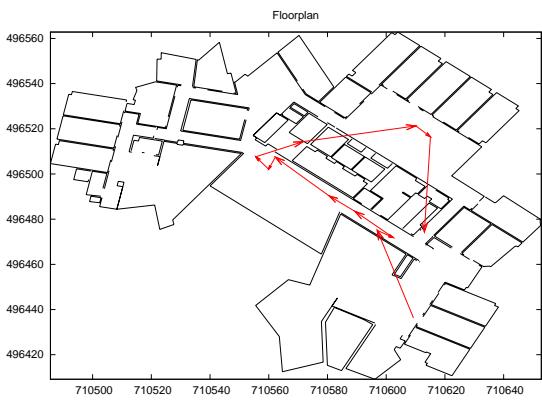


(b) GSM localization estimates.

Figure 1-3: WiFi and GSM localization are highly error-prone.



(a) True trajectory walked by a person indoors.



(b) Trajectory estimated from WiFi localization.

Figure 1-4: WiFi localization is error-prone indoors.

perform *dead reckoning* with inertial sensors by integrating angular velocity estimates from a gyroscope to obtain the orientation of the phone at any time instant, and using this orientation to subtract the effect of the earth’s gravity from measured acceleration. The resulting “user acceleration” when integrated twice should give the displacement, and hence location of the phone. However, the problem is that even a small measurement error by an MEMS sensor results in a large, rapidly growing *drift error* when the output of the sensor is used as input to higher-order integration, because the errors add up. The net effect is that the estimates from integration are unusable for localization. Figure 1.1 illustrates upward sloping drift in user acceleration estimated from accelerometer and gyroscope measurements on an iPhone.

1.2 Contributions Of This Dissertation

This dissertation describes systems that address all the above challenges, using measurements of:

- GPS
- WiFi signals
- Cellular signals

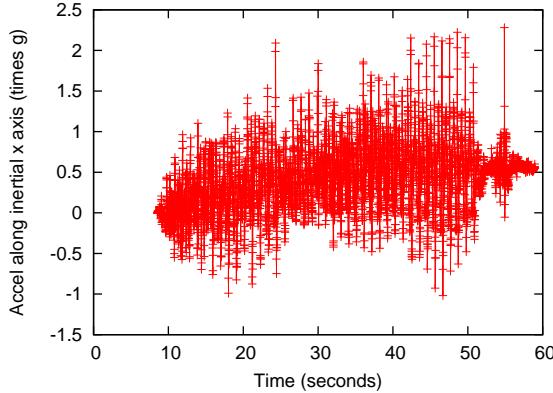


Figure 1-5: Upward sloping drift in user acceleration measured on an iPhone.

- Inertial sensors (accelerometer and gyroscope)

to estimate the trajectory of a mobile device accurately and energy-efficiently, indoors or outdoors. These systems develop and use two key techniques:

- Techniques that can extract accurate trajectories from both *sparsely sampled* and *noisy* location estimates. The ability to handle sparse input is important because it permits us to use *intermittently* sampled (sub-sampled) GPS, and thereby duty cycle the GPS when not in use to save a significant amount of energy. Handling noise permits us to use inaccurate WiFi and cellular localization estimates as input, and still estimate trajectories accurate at the level of individual road segments.
- Sensor fusion techniques that *combine* data from one or more of the above sensors — GPS, WiFi, cellular, and MEMS. While each individual sensor does not have all three desired properties of accuracy, energy efficiency and wide availability, sensor fusion enables us to achieve all the properties.

The systems we present all rely on a class of probabilistic model called a Hidden Markov Model (HMM). An HMM is a Markov process that models a sequence of noisy observations as coming from a sequence of (unknown) hidden states with a known probability distribution. It also models *transitions* between hidden states at each time step using a different distribution. Given an HMM and a sequence of noisy observations, a dynamic programming algorithm called Viterbi decoding can find the most likely sequence of hidden states corresponding to the observations.

While a HMM is a well-known tool to model and correct noise in input data, it is only as good as the underlying choice of hidden states, observations and probabilities used in it. We show in this dissertation that a careful choice of HMM is essential to handle sparsely sampled and noisy location data, and to perform effective sensor fusion. In the following sections, we describe the three systems we have developed: *VTrack*, *CTrack* and *iTrack*, and the key insights from each of the systems.

1.3 VTrack: Map-Matching Noisy and Sparse Coordinates

VTrack matches noisy position samples in the form of geographic (*lat, lon*) coordinates to a sequence of road segments on a map, and extracts travel time estimates for each output road segment.

VTrack uses an HMM where the observations are noisy coordinates and the hidden states are road segments. *VTrack* uses two insights that are critical to making the HMM work well:

- It uses a *speed constraint* that enforces the fact that a minimum amount of time must be spent on each road segment. This prevents the output of the HMM from “jumping around” to follow the noise in the input data. We show that the speed constraint is particularly important for matching the noisiest trajectories, reducing the error of map-matching (a term we define precisely in Chapter 3) by a factor of $3\times$ for the noisiest 10% of input trajectories.
- It *interpolates* its input location samples before feeding them to the HMM. This simplifies the design of the HMM because it can enforce an *adjacency constraint* on the output road segments. This helps ensure output a continuous sequence of road segments even if the input has outages or is intermittently sampled. Interestingly, we find that a simple linear interpolation strategy performs as well as more sophisticated strategies such as shortest path interpolation.

VTrack is optimized offline for each mobile phone hardware configuration it runs on: it uses offline measurements of GPS and WiFi energy costs on a mobile phone in conjunction with a given energy budget to determine whether to use GPS sub-sampling, or WiFi, or a combination of duty-cycled GPS and WiFi location samples on that phone hardware.

VTrack has been deployed as part of the Cartel [17] mobile sensing system at MIT. Our evaluation is over GPS and WiFi location data from a testbed of 25 taxicabs, and from mobile phone users who run the iCartel iPhone application and contribute crowd-sourced traffic data for the Boston area. It has been running live for nearly two years and has been used to map-match thousands of trajectories of raw position samples. The travel time estimates produced by *VTrack* are served to both users of the iCartel application and the Carweb web portal.

We evaluate the accuracy and energy-efficiency of *VTrack* on 800 hours of driving data from the Cartel testbed. We use *VTrack* to extract trajectories from GPS and WiFi location samples, and evaluate the accuracy of the trajectories produced by *VTrack* by comparing to ground truth GPS. In addition, we evaluate the usefulness of the output trajectories in the context of two applications: using travel times extracted from the trajectories for better route planning, and for detecting individual traffic “hotspots”. Our key findings are:

- Using a Hidden Markov Model for map-matching is robust to significant amounts of noise, producing trajectories with a median error of less than 10% on WiFi localization samples that have a median error of over 50 metres. Using an HMM is much more robust to noise in the input data than a simple strategy like matching each point to the closest segment on a road map, which breaks down when map-matching WiFi positioning or noisy GPS data.
- We find that *VTrack*’s interpolation technique is suitable for map-matching location data consisting of sub-sampled GPS. When GPS is available, sampling GPS periodically to save energy is a viable strategy. On our data, for up to 30-seconds duty cycling (i.e., turning on the GPS every 30 seconds to get a location sample , then turning it off for 30 seconds to save energy) sub-sampled GPS can produce high quality trajectories if processed with a HMM.
- Travel times from WiFi localization alone are accurate enough for route planning, even though estimates for individual road segments are poor. This is because groups of segments are typically traversed together, and our estimation scheme ensures that errors on adjacent or nearby segments cancel out.

- If a phone is also WiFi-equipped, the tradeoff between sampling GPS every k seconds, sampling WiFi or a hybrid strategy (*both* sub-sampled GPS and WiFi) depends on the energy costs of each sensor. Chapter 3 of this thesis explores these tradeoffs in detail.

1.4 CTrack: Accurate Tracks From Soft Information

As mentioned earlier, extracting tracks from cellular signals alone is attractive because the *marginal* energy consumption of cellular localization is close to zero on mobile phones, since the cellular radio is always on anyway. However, raw cellular position samples (such as in Figure 1-3(b)) have errors ranging from a few hundred metres to as much as a kilometre, making them challenging to match to a map. As we show, the one-pass HMM used in *VTrack* breaks down when run over data with so much noise.

The second component of this dissertation is *CTrack*, a system that makes it possible to extract accurate tracks from cellular signals. *CTrack* uses two novel insights to achieve this:

- Using *soft* information in the form of signal strengths from cellular (or WiFi) radios. *CTrack* uses a two-pass HMM that is very different from the HMM used in *VTrack*. It divides space into grid cells, and determines the most likely sequence of traversed grid cells corresponding to a sequence of cellular base station observations.

The *CTrack* HMM uses training data to build a probabilistic model of which base stations are seen from which grids, and with what signal strengths, in contrast to the simple model of a propagation error in a latitude/longitude coordinate that the *VTrack* HMM uses. This model is better than *VTrack* for cellular signal observations, because a given cellular “radio fingerprint” (consisting of cell towers and their signal strengths) is seen from a wide range of locations, sometimes spaced as much as hundreds of metres apart. Averaging this soft information to produce a single “hard” location estimate loses a lot of information.

- A probabilistic model for the HMM that uses *sensor hints* from MEMS sensors on the phone — the accelerometer to detect movements, and the compass or gyroscope to detect turns.

We have implemented *CTrack* on the Android smartphone platform, and evaluated it on nearly 126 hours of real drives (1,074 total miles) from 20 Android phones in the Boston area. We find that *CTrack* is good at identifying the sequence of road segments driven by a user, achieving 75% precision and 80% recall accuracy (these terms are defined and explained in Chapter 4 of this dissertation). This is significantly better than state-of-the-art cellular fingerprinting approaches like Placelab [57] applied to the same data, reducing the error of trajectory matching by a factor of $2.5\times$. We also measure the energy consumption of *CTrack* on Android and find that *CTrack* has a significantly better energy-accuracy trade-off than other candidate approaches, including sub-sampling GPS data, or using *VTrack*, reducing energy cost by a factor of $2.5\times$ for the same level of mapping accuracy. We also show that sensor hints can correct some common systematic errors that arise with cellular localization.

VTrack is still useful when the only data available is “hard” information in the form of coordinates, which is the case on some platforms. For example, on Apple’s iOS, WiFi and cellular location can only be accessed indirectly via a location API that provides geographic position samples.

1.5 iTrack: Accurate Indoor Tracks From Inertial Sensing and WiFi

Raw WiFi location samples indoors are subject to an average error of the order of 5-6 metres, and can occasionally experience errors of up to 10 or 20 metres. It is possible to use a Hidden Markov Model to extract tracks from this data (similar to *VTrack* or *CTrack*) but the extracted tracks still have limited accuracy. Secondly, as mentioned earlier, WiFi localization requires extensive and cumbersome manual training, either dedicated or crowd-sourced, before it can be deployed in a building.

iTrack is a system that addresses both these limitations by using a novel sensor fusion approach that combines data from inertial MEMS sensors with data from WiFi signals. While either of these sensors taken on its own is inaccurate, we show that *combining* data from these sensors using a carefully designed Markov Model can achieve high trajectory estimation accuracy. *iTrack* can accurately find the trajectory of a person walking steadily indoors with a smartphone in his/her hand or pants pocket to within less than a metre of error.

iTrack also significantly reduces the manual training effort required for WiFi localization by allowing tracks extracted from walks to be used to contribute training data for WiFi localization automatically, without requiring users to manually mark their location on a map.

iTrack uses two key novel ideas. The first key novelty of *iTrack* is how we avoid accumulation of drift error when dealing with data from MEMS sensors. *iTrack*'s approach works in four steps: *walking detection*, *step counting*, *shape extraction* and *map-matching*:

- Walking detection uses acceleration measurements to detect walking and distinguish it from other movements of a mobile phone such as talking, picking up and other normal use. A periodic up-and-down or sideways swaying motion is characteristic of walking, whether the phone is held in a users' hand, pants pocket, or bag, and causes accelerometer data to exhibit a periodic pattern with a peak on its Discrete Fourier Transform (DFT).
- Step counting uses the acceleration signal to accurately determine the number of steps walked using an iterative peak-finding algorithm.
- Shape extraction integrates angular velocity from the gyroscope to find *changes* in orientation (turns), and hence the *approximate shape* of a user's walk. The key insight is that using estimates of *changes* in orientation is more robust to drift error than using absolute orientation.
- Map-matching matches the extracted shape and size from the MEMS sensors to the contours of a floorplan (assumed to be known) using a *particle filter*. A particle filter is essentially a generalization of an HMM to a continuous state space, here the (x, y) coordinate of the phone on the floorplan. Unlike *VTrack* and *CTrack* (where the state space is discrete, consisting of road segments or grids on a map), the Viterbi algorithm cannot be used for a continuous state space. Instead, a particle filter simulates a large number of paths with similar shape and size to that determined from the MEMS sensors, and eliminates paths that cross a wall or obstacle in the floorplan to try to find a good match.

The second key novelty of *iTrack* is the actual Markov model used in the particle filter. While particle filters with a Gaussian error model have been previously used for indoor pedestrian tracking with foot mounted sensors [67, 29], we show that a *mixture model* for stride length and turn angle are preferable to a Gaussian error model in the mobile phone context. This is because a person's

walking pace typically varies significantly even within a given walk, and it is difficult to extract accurate length information for each stride from a mobile phone not rigidly mounted on a users' person.

The particle filter used in *iTrack* can optionally fuse WiFi signal information to improve the quality of the tracks it extracts. *iTrack* uses a small amount of “seed WiFi data” on each floor (typically 4-5 walks known to be on that floor) to identify the floor and building a person is on. The initial WiFi helps to initially extract tracks of adequate quality from inertial sensing.

As more WiFi data is collected, this sensor fusion has a feedback effect that enables *iTrack* to achieve the goal of indoor tracking with very little manual input. *iTrack* continuously adds tracks extracted from inertial sensing to a training database containing WiFi access points, their signal strengths, and locations in the floorplan they were seen from. As more tracks become available, *iTrack* iteratively uses the extracted WiFi information to *re-extract* tracks from previously seen walks, and rebuild the WiFi training database. This has the effect that the quality of extracted tracks improves, and further improves the quality of the WiFi training database, resulting in a positive feedback cycle.

Therefore, once the seed information has been populated for each floor of interest, the rest of the process is entirely automated and does not require manual input — dedicated operators simply walk around the floor to add more training data, and volunteers can contribute “crowd-sourcing” data easily by simply downloading and running a background phone application. Since crowd-sourcing is entirely automated and requires no intervention on behalf of the user beyond normal use of the application, this is a significant improvement over techniques requiring extensive manual input from a human to indicate where he/she is.

We evaluate *iTrack* on 50 walks collected on the 9th floor of the MIT CSAIL building in the Stata Center. We use tape markings made on the ground to accurately estimate ground truth for each trajectory. We find that with seed WiFi data from just 4 walks (that took only 5-10 minutes to collect), *iTrack* can extract accurate trajectory data from 80% (40 out of 50) of the walks (20% of the walks fail to produce any match). The walks extracted have a median error of ≈ 3.1 feet, corresponding to approximately one stride. This contrasts to a median error of over 18 feet if using just WiFi.

1.6 Road Map

The rest of this dissertation is organized as follows. Chapter 2 describes energy measurements of the different localization technologies and sensors used in this dissertation (GPS, WiFi, cellular and inertial sensors) on different phone platforms. These measurements motivate and drive the design of the algorithms subsequently described in the dissertation. Chapter 3 describes the *VTrack* system for map-matching a sequence of noisy geographic location samples to road segments, and extracting travel time information from these segments. Chapter 4 describes and evaluates *CTrack*, a system that can accurately map-match soft information from cellular (GSM) signals. Chapter 5 first provides background on how MEMS inertial sensing technology works, and then describes and evaluates *iTrack*. Chapter 6 concludes the dissertation and outlines directions for future work.

Chapter 2

Energy Studies

We return to the first problem mentioned in the introduction:

- *Given an energy budget and a requirement on accuracy, what is the best combination of sensor(s) and sampling rate(s) to use for trajectory mapping?*

We will show that the answer to this question depends on the specific mobile device hardware being used, and in particular on the energy costs of sampling different sensors. For this reason, for any given mobile device whose trajectory we want to map, it is necessary to first measure the energy costs of each sensor and use the measurements to drive an energy-efficient algorithm design.

In this chapter, we first present a discussion of the fundamental factors that drive the on-device energy cost of each of the sensors under consideration — GPS localization, WiFi localization, cellular localization and inertial/orientation sensors (accelerometer, magnetic compass and gyroscope). As part of the discussion, we survey a number of experimental studies by other researchers that measure the energy costs of these sensors on different mobile devices.

We then present three of our own experiments that measure the energy costs of these sensors when sampled at different rates on actual phone hardware. The first experiment measures the energy costs of GPS and WiFi localization on the iPhone 3G. The second experiment measures the energy cost of accelerometer sampling relative to the cost of GPS sampling on the iPhone 3G-S and iPhone 4. The third experiment measures the energy cost of GPS, WiFi localization, cellular localization, accelerometer and magnetic compass on an Android G1 phone.

The results of the energy experiments in this chapter drive the design of the energy-efficient trajectory mapping algorithms described in Chapters 3 and 4. We do this by building cost models for the energy consumption of each sensor, and using the cost models to choose the right parameters for the algorithm i.e., which sensor(s) to use and at what sampling rate(s).

2.1 Theoretical Factors Driving Energy Cost

2.1.1 GPS Energy Cost

The energy cost of GPS is rooted in the need for processing gain to acquire positioning signals from GPS satellites. On modern GPS chipsets, there are two distinct parts of this process which

have different energy consumption profiles. To understand the energy consumption requirements of GPS, it is first necessary to understand how a GPS receiver works.

To estimate its location, a GPS receiver on earth measures the transmission delay from multiple GPS satellites orbiting the earth's surface. The receiver uses the measured transmission delay to calculate its distance (also called *pseudo-range*) to each visible satellite. For each satellite, the receiver has prior knowledge of the precise orbit of that satellite, and hence knows that its location must lie on a sphere with that satellite's location as centre and the measured pseudo-range as radius. If the receiver can measure its pseudo-range for at least 3 (preferably 4) satellites, it can use *triangulation* to solve for the intersection of the spheres and determine its precise latitude, longitude, and altitude above the earth's surface.

The GPS triangulation process requires the following pre-requisites before it can work:

- At least 3 (preferably 4) visible GPS satellites i.e., satellites that the GPS receiver can hear and decode transmissions from.
- Knowledge of the precise orbits of the visible satellites so that their location can be calculated exactly. This is called *ephemeris* information.

When a GPS receiver first powers up, its first task is to determine a set of 3 or more visible satellites to triangulate from. The naive method of doing this (and the fallback in case quicker techniques fail) is brute-force search. Each satellite is assigned a unique binary sequence called a *Gold code*. The GPS signal is decoded after demodulation using modulo 2 addition of the Gold codes corresponding to each satellite. When performing brute-force search, a GPS receiver cycles through *all* the possible Gold codes until at least one of the satellite transmissions can be decoded. However, as might be expected, this process is extremely time-consuming. On older GPS receivers with only one reception channel, obtaining a brute-force satellite lock can take between 7.5 to 15 minutes. Modern GPS receivers can receive on up to 12 channels in parallel, significantly reducing the brute-force fix time. However, even on modern chipsets, obtaining a brute-force satellite lock remains a fairly time consuming process, requiring 3-5 minutes [34].

To reduce the time to get a fix, GPS receivers maintain a cache of visible satellites and their approximate coarse-grained locations, called an *almanac*, which helps narrow down the brute-force search. The GPS receiver uses a brute force search to obtain a lock the first time it is used. On first use, it downloads an almanac from the GPS satellite it is connected to. Once downloaded, the almanac information remains valid for a few months.

Once the almanac information is available and the GPS has narrowed down a set of 3 or more visible satellites, it must next download the orbital (ephemeris) information for these satellites before it can triangulate its position. Each GPS satellite transmits its ephemeris information every 30 seconds. The satellite link over which the transmission takes place has limited bandwidth, and the maximum possible transmission rate over the link is only 50 bytes per second. This means it can take up to 40 seconds to download the ephemeris information from each satellite. Older GPS receivers cannot download this information in parallel (since they can only receive on one channel) but newer chipsets can receive on up to 12 channels and hence can download the ephemeris information from 3 or 4 satellites in parallel. Thus, a modern GPS receiver can obtain a so-called “GPS lock” (i.e., knowledge of ephemeris information for at least 3 visible satellites), and compute a first GPS location estimate in about 40-45 seconds [34]. This process of downloading the ephemeris information

is called a *cold start* in GPS terminology, and must be repeated each time a GPS is powered on in a new location.

A further optimization modern GPS chipsets use is *warm start*. If a GPS chipset is turned off and then turned on again within a certain (short) time interval, and is at a location near where it was turned off, it is likely to be able to hear the same set of satellites it downloaded ephemeris information for previously. Hence, it can use ephemeris information saved from the previous cold start to reduce the time required to obtain a fix. The time to fix is still not instantaneous, because the receiver needs to verify that the ephemeris data is valid and that at least 3 satellites it has data for are still visible in the sky. A “warm start” on a modern GPS receiver takes between 6-15 seconds.

The simplest way we have developed to measure the time to a GPS warm start on a mobile device is to drive through a tunnel in an urban area (or other known region with poor GPS reception). The device will lose its GPS lock and re-acquire it when the vehicle emerges from the tunnel. It is easy to measure the time required for the device to re-acquire its GPS position estimate from the instant the vehicle emerges from a tunnel. Using this method, we estimated an average “warm start” time of 6 seconds on the iPhone 3G, and about 12 seconds on the Android G1.

Once a GPS chipset has downloaded the necessary ephemeris information, it switches to “continuous tracking mode”. In this mode, the main function of the GPS is to use periodic satellite transmissions to perform triangulation and continuously determine its position.

GPS Energy Requirements. The process of obtaining a GPS lock from scratch (“cold start”) is the most energy-intensive part of GPS operation, since it requires significant processing gain over a period of 40-50 seconds to decode and download ephemeris information from the relatively weak (-160 dB) GPS signal. This energy requirement is somewhat lower for warm start, but warm start still consumes energy to download and verify satellite orbital information during the 6-15 second initialization period. As might be expected, this energy consumption increases in conditions where the signal is weaker and requires more processing gain e.g., on a cloudy day without clear satellite visibility, or in areas with interference from other obstacles, such as downtown urban areas with tall skyscrapers. The power consumption of continuous tracking mode once a fix is obtained is typically lower, but still significant on a mobile device. Most GPS receivers report a power consumption requirement of about 50-75 mW in continuous tracking mode. In practice, when including the energy required to obtain a fix, we have found in our experiments that the average power consumption of GPS chipsets on mobile smartphones is much higher, closer to 350-400 mW. These results are also corroborated by studies by other researchers, presented in Section 2.1.5.

2.1.2 WiFi Localization

WiFi localization on mobile phones works by periodically scanning for nearby WiFi access points. The phone periodically looks up the set of access points it sees and their signal strengths (aka “WiFi signature”) in a local or remote database to determine its approximate location.

WiFi localization consumes energy because the WiFi radio must be on to periodically scan for nearby WiFi access points. Localizing with WiFi does not actually require associating to an access point or transmitting data over WiFi. Duty-cycling the WiFi radio incurs an initialization and shutdown energy cost on most phones [50]. Studies on a Nokia N95 mobile phone [45] and [50] have found the maximum power consumption of WiFi localization on a mobile phone to be of the order of 1 Watt, during the time period when the phone is actually scanning for WiFi access points.

Fortunately, a device using WiFi localization need not scan all the time. The *average* power consumption of WiFi localization over *all time* i.e., not just the time period when a scan is being

performed, is lower than the peak power consumption of 1 Watt, and is mainly driven by the time duration of each WiFi scan and the interval between successive scans. The authors of [50] found an average scan time of 0.7 seconds on an AT&T Tilt phone. We have found the number to be smaller on more modern iPhone and Android devices, closer to 0.5 seconds or even smaller. A reasonable assumption is that the maximum possible rate of WiFi sampling corresponds to obtaining a new WiFi localization estimate every 2 seconds. Typically, more frequent scanning than 0.5 Hz is not useful since more frequent scans tend to return cached results or result identical to the previous scan. Assuming a scan interval of 2 seconds and a mean scan duration of 0.5 seconds, we can derive a theoretical average power draw of approximately $\frac{1}{4} \times 1 = 250$ mW, smaller than 1 Hz GPS sampling (about 400 mW).

Our actual experiment with WiFi energy consumption measured the energy cost on an Android G1 phone with a request interval of approximately 2 seconds between WiFi scans. We found that battery life of the phone was close to 10 hours with WiFi scanning, compared to only 6 hours with 1 Hz GPS sampling. This corroborates the back-of-the-envelope theoretical calculation above, as well as the power consumption findings in [50, 45].

A second driver of energy cost with WiFi localization is the network transmission cost of looking up WiFi access points if using a remote server-side database to do the lookup. For example, each time an Android phone makes a “network location” lookup, it contacts a Google server with a list of the access points it can see nearby to determine its location. The request must be transmitted and a reply received from the server over the internet. This uses either WiFi or 2G/3G transmission and consumes energy.

Fortunately, in practice it is possible to store a locally relevant cache of the WiFi access point database on the phone, thus significantly reducing or eliminating the energy cost associated with the lookup request. Current implementations of WiFi localization all appear to use caching of WiFi tiles to reduce latency and energy consumption [41]. For this reason, we ignore the energy cost of network lookups in all of our subsequent analysis.

To summarize, the power consumption of WiFi localization is often 50-60% lower than GPS (assuming both sampled at the highest possible rate), with the exact savings depending on the mobile phone hardware. Previous work [77] corroborates these end-to-end energy consumption numbers on a Nokia N95, finding that WiFi localization can provide 2-3 times the battery life of GPS localization. The iPhone 3G, which we used for one of our energy experiments (§2.2.2), appears to be an exception to this approximate trend, with GPS being an order of magnitude more expensive than WiFi — perhaps owing to poor GPS power management.

2.1.3 Cellular Localization

Because phones continuously track cell towers as a part of their normal operation, it costs very little extra energy to keep the cellular radio switched on, and log the cell towers seen and their signal strengths. Hence, the marginal energy cost of cellular localization is small, and driven primarily by CPU load. On the Android OS, for example, a list of neighbouring towers and their signal strengths can be obtained by querying the `TelephonyManager` API. An application or service that provides/uses cellular localization periodically reads the list of neighbouring cell towers and their signal strengths using the API. Keeping such an application running continuously consumes a small amount of energy since the phone cannot put the CPU in low-power mode or frequency scaling mode while it is running. Processing a cell tower signature might require at most 100,000 instructions, which costs 5 nJ on a current generation 1 GHz Qualcomm Snapdragon processor.

In embedded (non-phone) applications that do not need the cellular radio on as part of their normal operation, it is possible to track only the signal quality and cell ID portions of the GSM protocol. This requires observing only the BCH slots of the GSM beacon channel, which are 4.6 ms long and are transmitted once per each 1.8 second cycle. A 10% GSM receiver duty cycle should be adequate to track the strongest towers. Assuming a GSM receiver uses 17 mA at 100% duty cycle, this represents an additional power consumption of 5 mW (1.7 mA @ 2.7 V) [11, 79].

Our experiments measuring battery lifetime with an Android G1 phone found that the phone lasted nearly 3 days when running a background application that periodically requested and wrote a list of neighbouring cell towers and their signal strengths to flash storage. This is close to the phone’s “stand by” lifetime when not running any applications.

2.1.4 Inertial Sensors: Accelerometer, Compass and Gyroscope

Accelerometers, magnetic compasses and gyroscopes on today’s mobile phones use low-energy MEMS (micro-electro-mechanical sensing) technology. These devices have a relatively low energy overhead. A significant portion of the overhead comes from CPU cost for processing inertial sensing data on the phone, rather than the sensor sampling cost. For example, ADXL 330 accelerometers use about 0.6 mW when sampling at the highest possible rate, and at 10 Hz can be idle about 90% of the time, suggesting a power overhead of around .06 mW for sampling the accelerometer [13]. The MicroMag3 compass uses about 1.5 mW when sampling at the maximum possible rate, suggesting a power consumption of .15 mW or less at 10 Hz [73].

Our experiments with battery life on the Android G1 have found that sampling the accelerometer and compass at 20 Hz add a negligible amount of energy consumption. In fact, running cellular localization, accelerometer and the compass all together yields a lifetime of over 60 hours, or about 10× the lifetime when sampling GPS at 1 Hz. A different experiment we performed on the iPhone 3G-S and iPhone 4 (also described later) showed a similarly insignificant energy overhead from sampling the accelerometer at 20 Hz on the iPhone.

2.1.5 Other Energy Studies and Discussion

In [45], the authors showed that Nokia N95 phones use about 370 mW of power when GPS is left on, versus 60 mW when idling. They also found that 1 Hz GPS sampling results in 9 hours of total battery life. Several other papers [82, 60, 32, 37, 77] suggest similar numbers for N95 phones (battery life in the 7–11 hour range) with regular GPS sampling. On a more recent AT&T Tilt phone [50], the authors found that 1 Hz GPS sampling used 400 mW, a single GPS fix costs 1.4-5.7 J of energy (depending on whether previously downloaded satellite ephemeris information is cached or not) and a WiFi scan consumed about 0.55 J of energy. These numbers are consistent with our discussion above.

2.1.6 Summary

We summarize the discussion above:

- The theoretical power consumption of cellular scanning plus sensors on a phone should be less than 5 mW.
- The power consumption of inertial sensors alone should be less than 1 mW, low enough that it does not reduce the phone’s overall lifetime even when in standby mode.

- In contrast, the best-case theoretical power consumption (as reported by datasheets) for GPS is 75 mW in “continuous tracking” mode when a fix is already acquired, but in practice is closer to 400 mW on most mobile phone chipsets, *when including the energy to periodically re-acquire GPS fixes*.
- WiFi scanning every second or two requires keeping the radio on, consuming 50-60% of the energy cost of GPS on some platforms. On other platforms like the iPhone 3G, WiFi localization is significantly cheaper than GPS.

2.2 Energy Experiments

2.2.1 Does GPS Sub-Sampling Save Energy?

Continuously sampling the GPS is energy-intensive and constitutes a dominant fraction of the energy cost for trajectory mapping. A natural question that arises is whether the strategy of *sub-sampling* i.e., querying the GPS for position estimates intermittently rather than at the maximum possible rate, can save energy.

Saving energy with sub-sampling depends on the ability of a phone to turn off, or *duty-cycle* its GPS unit when not in use. Unfortunately, the location sensing APIs of most current mobile smartphone platforms do not provide direct low-level control over the GPS sensor. For example, Apple’s iOS does not even provide direct control over whether GPS is used or not for determining location. Instead, mobile phone APIs such as iOS and Android allow only indirect control of the GPS via the operating system’s location API. A phone application can request location updates with a specified accuracy or at a specified time interval, and the operating system takes care of the low-level details of whether GPS is duty-cycled or not. Hence, the energy savings in practice from sub-sampling depend on both the chipset and the policy used by the operating system.

Our initial goal was to obtain an understanding of whether sub-sampling GPS can save energy in a *stand-alone* setting, independent of the quirks of the underlying operating system. To do this, we performed a simple experiment with a stand-alone GPS unit to measure how well current GPS sensors perform in a setting where we have direct control over them.

We used a Bluetooth GPS unit with an MTK GPS two-chip chipset (MTK MT3301 and MT3179). This device has a power switch. We measured the time for the device to power on and acquire a “warm fix” after it had previously obtained a “cold fix” and saved it. As mentioned earlier, this is the typical mode of operation when a GPS is duty-cycled frequently. We found that with a power-off time of anywhere from 5 seconds to 30 minutes, it took the GPS receiver about 6 seconds to power on, acquire a “warm fix” (i.e., verify all its cached satellite ephemeris data), and report the first reading over Bluetooth to a laptop computer.

Based on this experiment, it is possible to estimate the theoretical cost of GPS sub-sampling on a hypothetical mobile device that performs *perfect* GPS power management — by turning off the GPS immediately after a location sample is reported to the application, and turning it on 6 seconds before the application needs the next location sample. If such a hypothetical device were to acquire a GPS “warm fix” every $k > 6$ seconds, it would use approximately $\frac{6}{k}$ of the power of 1 Hz GPS sampling.

Our subsequent phone experiments, described next, measure the savings due to GPS sub-sampling on the iPhone 3G and the Android G1 phones. As we shall see, the two phones’ implementations

Location Mechanism	Sampling Period	Lifetime
None	-	7 h
GPS	1 sec	2 h 24 m
GPS	30 sec	2 h 27 m
GPS	2 min	2 h 44 m
WiFi	1 sec	6 h 30 m

Table 2.1: Energy experiments comparing GPS and WiFi localization on an iPhone 3G.

of GPS power management are very different. The iPhone 3G’s firmware/OS — at least as of the time we performed our experiment — appears to do poor GPS power management, resulting in negligible or no savings due to GPS sub-sampling. In contrast, the Android G1 seems to perform good GPS power management, and sub-sampling the GPS actually saves energy in accordance with a formula similar to the one mentioned above.

2.2.2 Experiment on iPhone 3G

Our first experiment on actual phone hardware measures the energy consumption of GPS and WiFi localization on the iPhone 3G. We did not test cellular localization on the iPhone since we do not know of a way to programmatically retrieve neighbouring cell towers on the iPhone. It appears to be possible to retrieve the connected cell tower using an undocumented API, and a list of nearby cell towers when the phone is put in “field test” mode, but no other information appears to be available during normal operation.

We wrote a simple iPhone application that repeatedly requests location estimates at either GPS or WiFi accuracy, with a user-specifiable periodicity, until the battery drains to a fixed level from a full charge (in our experiments we drain the battery to 20%). The iPhone 3G produces a localization accuracy estimate for each location it reports to the application. The “vertical accuracy” field of this estimate has a non-zero (valid) value when the phone uses GPS localization, and has an invalid value when the phone uses WiFi or cellular localization to obtain a fix.

Because the iPhone 3G does not permit background applications (unlike the newer iPhone 4 which does), we ran this application in the foreground with the phone’s screen turned on at minimum brightness. Therefore, we ran a control experiment and measured the total lifetime with the screen on at minimum brightness, but without requesting location estimates.

Our results are summarized in Table 2.2.2. On the iPhone 3G, GPS is extremely power hungry and reducing the sampling period does not reduce energy consumption appreciably. The reason appears to be that the iPhone OS leaves the GPS on for about a minute even when an application de-registers for position estimates, rather than duty-cycling it, perhaps to avoid the additional time delay incurred for a warm or cold start. Hence, sampling GPS every 30 seconds is as expensive as sampling once a second, and even obtaining one GPS sample every two minutes does not save a significant amount of power. In contrast, the iPhone seems to do a much better job of aggressively managing WiFi power consumption when no data is being sent and the radio is being used only for localization.

Estimating WiFi Cost. While GPS sub-sampling does not save a significant amount of energy on the iPhone 3G, WiFi localization does turn out to save a significant amount of energy over sampling GPS. We use the numbers in the table above to solve for the WiFi energy cost on the

Sensors Used	iPhone 3G-S	iPhone 4
No sensors	18.6 hr (1.7)	16.6 hr (0.8)
Accelerometer 1Hz	19.5 hr (1.4)	17.3 hr (1.8)
Accel. 20Hz	18.3 hr (3.1)	16.9 hr (0.8)
GPS	6.1 hr (0.6)	10.1 hr (0.3)

Table 2.2: Battery lifetime: GPS vs accelerometer on the iPhone.

iPhone as a fraction of the GPS@1 Hz sampling cost. Suppose the battery has capacity c Joules. The baseline lifetime, without any location estimates but with the screen on, is 7 hours (25,200 seconds). Therefore, the baseline (with screen on) power consumption $b = c/25200$ Watts. The lifetime with both GPS and the screen on is 8,640 seconds, so $g + b = c/8640$. Solving, we get $g = c/13147$. On doing a similar analysis for WiFi, we get $w = c/327360$ and $g/w = 24.9$ —that is, the cost per sample of GPS is $24.9 \times$ the cost per sample of WiFi. This also suggests that WiFi sampling is about 8 percent of the total power consumption when the phone is running continuously with the screen on (since $w/g = .08$), which means that a foreground application that samples WiFi is unlikely to dramatically alter the battery life of the iPhone 3G compared to a foreground application that does not. This is partly because WiFi is more energy-efficient, and partly because the additional energy consumption due to WiFi is drowned in the cost of keeping the screen switched on, even at minimum brightness.

2.2.3 Experiment on Multiple iPhones

The experiment in the previous section only measures GPS and WiFi energy cost on the iPhone 3G. We now describe the results of a second energy experiment on the iPhone platform that measures the energy cost of the accelerometer in comparison to GPS. This experiment was performed by the author of this dissertation in collaboration with other researchers for a project not part of this dissertation [5].

This experiment measures energy consumption of 1 Hz GPS sampling and accelerometer sampling at 1 Hz and 20 Hz on the iPhone 3G-S and the iPhone 4. As in the previous experiment, it was not possible to turn off the iPhone’s screen and at the same time keep the sensors running. Hence, as before, all the experiments were run with screen brightness set to a minimum

Table 2.2.3 reports the measured battery life for four sensing configurations:

- No sensors (i.e., “stand-by” lifetime).
- Accelerometer sampled at 1 Hz.
- Accelerometer sampled at 20 Hz.
- Continuous GPS sampling.

Each number reported in the table is the mean of 5 measurements. The standard deviation of each number is indicated in parentheses next to it.

The table shows that accelerometry has a negligible effect on energy consumption: the battery duration was similar for running the accelerometer at 1Hz or 20 Hz, and for discharging the battery with no sensors turned on. There was considerable variance between multiple runs of the experiment.

The phone with accelerometry enabled having *higher* lifetime than without (as seen from the table) is an artifact of this variance.

We note here that the lifetimes in this experiment are not directly comparable to the results for the iPhone 3G in the earlier experiment. The lifetimes in this experiment are higher because the phone models we use — iPhone 3G-S and iPhone 4 — are newer models with improved battery life. Moreover, the earlier experiment measures time discharge to 20% of maximum battery life, while this experiment measures time to complete battery discharge.

2.2.4 Experiment on Android G1

We also performed an experiment to quantify the energy consumption of each of the sensors of interest — GPS, WiFi localization, cellular localization, magnetic compass and the accelerometer on an Android G1 phone. For each sensor, we wrote an Android application to sample the sensor at some given frequency, and query the battery level indicator periodically. We charged the phone to 100%, configured the screen to turn off automatically when idle (this is the default behaviour), and started the application. We used the Android `TelephonyManager` API to retrieve neighbouring cell towers and their associated signal strength values.

Figure 2-1 shows the remaining battery life reported by the Android OS as a function of time for four sensor sampling configurations:

- GPS sampled every second.
- GPS sub-sampled every two minutes.
- WiFi scanning at the max possible rate, every 2 seconds.
- A configuration using three sensors: logging GSM cell towers every second, compass at 20 Hz, and accelerometer at 20 Hz.

The last configuration above measures the cost of cellular scanning bundled with the compass and accelerometer because the experiment was originally performed in the context of evaluating *CTrack*, the trajectory mapping algorithm discussed in Chapter 4 that uses cellular localization in conjunction with “hints” from the accelerometer and the compass for trajectory mapping.

We highlight three key points from Figure 2-1:

- **Cellular localization is extremely energy-efficient.** Even when also simultaneously sampling the compass and accelerometer, cellular (GSM) localization results in a saving of approximately 10× in battery life compared to sampling GPS every second, and over 6× compared to WiFi. The lifetime of over 60 hours is close to the stand by lifetime of the Android G1 (not shown in the figure) which is a little over 3 days.
- **GPS sub-sampling saves energy.** Unlike the iPhone 3G, the Android G1 seems to have a sensible GPS power management policy, appearing to duty-cycle the GPS when it is not in use. For example, we see that sub-sampling the GPS every 2 minutes saves a significant amount of battery life compared to sampling it once a second.

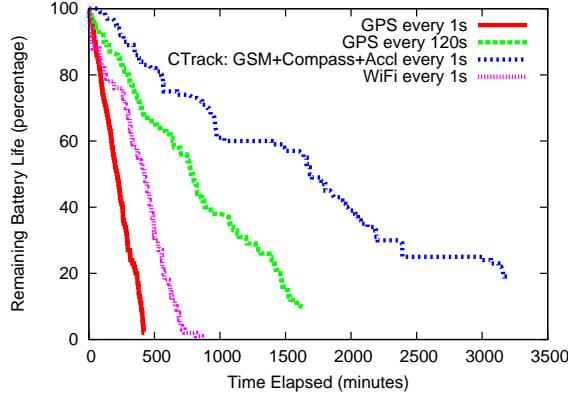


Figure 2-1: Energy consumption: GPS vs WiFi vs cellular on an Android phone.

- **WiFi saves some energy compared to GPS.** On the Android G1, WiFi localization is cheaper than GPS, with a battery lifetime of over 10 hours, compared to only 6 hours for GPS sampled at 1 Hz. This is significant, but not as much of a saving as in the iPhone 3G.

The battery drain curves in Figure 2-1 look irregular because the G1 phone does not estimate remaining battery life accurately. We performed a similar experiment on a Nexus One, a more recent Android phone model. The experiment on the Nexus One showed a very similar trend in terms of relative costs of the different sensors, but the battery drain curves looked like straight lines.

2.3 Conclusion

The energy consumption profiles of different sensors are different on the different phones we tested. For example, because the Android G1 permits background applications to run without the screen having to be kept switched on, WiFi sampling in the background on the G1 *does* significantly alter the G1's lifetime, unlike the iPhone 3G experiment where WiFi sampling cost was drowned out by the power consumption of the screen. Similarly, because the GPS power management policy is different on the iPhone 3G and the Android G1, the energy savings due to GPS sub-sampling are very different on these two platforms.

The main lesson is that an algorithm suitable for energy-efficient trajectory mapping on one platform may not be energy-efficient on a different platform, and vice versa. For example, an algorithm based on GPS sub-sampling may be suitable for the Android G1 phone but not for the iPhone 3G. Therefore, different platforms require different algorithms for energy-efficient trajectory mapping. The trajectory mapping strategy described in this dissertation adapts to different platforms using an *offline optimization* strategy, where a battery drain experiment similar to the ones presented in this chapter is performed for each distinct phone model or platform we want to optimize for. The experiment estimates the *relative* energy costs for each sensor (GPS, WiFi, or Cellular) on that platform. Chapter 3 discusses how the measured relative energy costs from such an offline experiment can be used as input to an *sensor selection framework* that decides what sensor(s) and sampling rate(s) to use on that platform to achieve a desired level of trajectory mapping accuracy.

Chapter 3

Map-Matching With Markov Models

The energy studies presented in Chapter 2 highlight two fundamentally different ways to reduce the energy requirement for trajectory mapping:

- Sub-sample GPS, i.e., obtain a GPS location sample every k seconds for some value of k .
- Use GPS alternatives: WiFi localization, cellular localization, and/or inertial sensing.

One can also imagine strategies that use a combination of the above approaches.

With sub-sampling, the trajectory mapping system needs to process a stream of *infrequent* position samples and map them to an accurate trajectory. If using GPS alternatives, the key challenge is to process a stream of *inaccurate* position samples and map them to an accurate trajectory, since WiFi and cellular localization are typically much less accurate than GPS.

This chapter presents *VTrack* [7], a trajectory mapping system for the outdoor setting. *VTrack* uses a Hidden Markov Model (HMM) to accurately match a stream of *infrequent and/or inaccurate* position samples to a sequence of road segments on a map.

VTrack is most suitable when raw localization data is already in the form of geographic (lat, lon) coordinates — either GPS coordinates obtained from sub-sampling, or coordinates obtained by looking up WiFi or cellular radio signals in a training database. If more detailed “soft” information such as WiFi or cellular base station sightings and their signal strengths is available, *CTrack*, the algorithm presented in the next chapter, is a better strategy than converting the soft information to coordinates and processing it with *VTrack*. However, *VTrack* is very useful in practice because soft information is not available in many situations. For example, Apple’s documented iPhone API does not provide low-level access to WiFi or cellular signal strengths — WiFi and cellular localization can only be accessed as (lat, lon) coordinates indirectly by querying a location API.

VTrack produces good trajectory estimates with coordinates from either GPS, sub-sampled GPS, or WiFi localization. It has been deployed as part of a *real-time traffic monitoring* system developed as part of the Cartel project [17], which uses smartphones as *traffic probes*, as do other systems [1, 55, 17, 71]. Unlike professional fleets and tracks which contribute to much of the currently available commercial traffic data, end users travel on the roads and at the times that are most useful to monitor for the purpose of reducing end-user commute duration. The idea is to obtain

timestamped position estimates from smartphones carrying GPS, WiFi or cellular radios, and deliver the estimates to a central server. The *VTrack* server processes position estimates to estimate *driving times on individual road segments of the road network*.

In the context of a traffic monitoring system, it is possible to give a precise characterization of the accuracy requirement for trajectory mapping. The end goal of *VTrack* is to produce an accurate sequence of road segments traversed by a vehicle, and accurate travel times for individual segments in the output path. In other words, it is important to be able to identify the road segment that each position sample came from accurately, and to estimate travel times accurately.

VTrack uses a probabilistic model, called a Hidden Markov Model, to model a vehicle trajectory over a detailed road map of the area, and performs *map-matching*, which associates each position sample with the most likely point on the road map, producing travel time estimates for each traversed road segment. The travel times are shared with other users of the system who would like to view or use the real-time traffic data.

We evaluate *VTrack* on a data set of GPS and WiFi location samples from nearly 800 hours of drives gathered from 25 cars. The data was gathered from two sources: an iPhone application, and from embedded in-car computers equipped with GPS and WiFi radios. We use a cleaned version of the GPS data without outliers and outages, sampled every second, to estimate the ground truth for these drives with high confidence, and use the ground truth in our evaluation.

The main question we investigate in the evaluation is how good the *end-to-end* accuracy of *VTrack* is. In particular, *how does the accuracy of travel time estimates depend on the sensor(s) being sampled and the sampling frequency?*

We answer this question by evaluating *VTrack* in the context of two applications that use the travel time estimates produced by the system. The first application reports *hotspots*, or roads with travel times far in excess of that expected from the speed limit. The second is a traffic-aware route planner that finds paths with the shortest expected travel times using the road segment travel times estimated by *VTrack*. Our key findings are:

- *VTrack*'s HMM-based map-matcher is robust to noise, producing trajectories with median error less than 10% when run over noisy WiFi localization data, as well as for simulated Gaussian noise with standard deviation up to 40 metres.
- *Travel times from WiFi localization alone are accurate enough for route planning, even though individual segment estimates may be poor.* When location samples are noisy, it is difficult to attribute a car's travel time to small stretches of road—for example, time estimates from WiFi for individual segments have a median error of 25%. However, somewhat counter-intuitively, using these times to find shortest paths works well—over 90% of shortest paths found using WiFi estimates have travel times within 15% of the true shortest path. This is because groups of segments are typically traversed together, and *VTrack*'s estimation scheme ensures that errors on adjacent or nearby segments “cancel out.” *Moreover, estimating real drive times actually matters—for congested scenarios, using just speed limits to plan paths yields paths that are up to 35% worse than optimal.*
- *Travel times estimated from WiFi localization alone cannot detect hotspots accurately, due to outages present in WiFi data.* We find that a hotspot detection algorithm based on *VTrack* misses many hotspots simply because of a lack of data on those segments. This problem is not as apparent in route planning, since in that scenario we are focused on choosing a path that

has a travel time closest to the shortest path, rather than worrying about particular segments. However, on the subset of segments for which we *do* have WiFi data, we are able to accurately detect more than 80% of hotspots, and flag fewer than 5% incorrectly.

- When GPS is available and free of outliers, sampling GPS periodically to save energy is a viable strategy for both applications. On our data, for up to $k = 30$ seconds (corresponding to roughly 2 or 3 road segments), sampling GPS every k seconds produces high-quality shortest paths, assuming GPS is available whenever it is sampled. If the device also has WiFi, the tradeoff between sampling GPS every k seconds, sampling WiFi or a hybrid strategy depends on the energy costs of each sensor, as mentioned in Chapter 2. Later in the chapter, we discuss this tradeoff and present a sensor selection framework for *VTrack* that makes this decision based on energy costs measured offline.

Using Hidden Markov Models for map matching is not a new idea [36, 44]. However, previous research has focused mainly on map matching frequently-sampled GPS data with low noise, and on qualitative studies of accuracy. A key contribution of our work on *VTrack* is a quantitative evaluation of the end-to-end quality of time estimates from noisy and sparsely sampled locations.

The rest of this chapter is organized as follows. Section 3.1 provides background on traffic monitoring, explaining the motivation and original context for *VTrack*. Section 3.2 describes two key applications that use the travel time estimates produced by *VTrack* and uses them to concretize the “how accurate does map-matching need to be” question. Section 3.3 describes the architecture of the *VTrack* trajectory mapping system. Section 3.4 formally states the map-matching problem and describes the *VTrack* algorithm. This section describes the Hidden Markov Model used to model a car’s trajectory and solve for the best match to a map. Section 3.5 explains how *VTrack* estimates travel times for individual segments from the output of trajectory matching.

Section 3.6 evaluates *VTrack* using two kinds of accuracy measurements:

- End-to-end accuracy measurements on the two real applications mentioned earlier, route planning and hotspot detection.
- Micro-benchmark evaluations of how robust *VTrack*’s Hidden Markov Model is to varying levels of simulated noise in its input data.

Section 3.7 uses the results of Section 3.6, mainly from the end-to-end applications, to revisit and provide some answers to a central question of this dissertation:

- What is the best strategy for energy-efficient trajectory mapping given an end-to-end accuracy requirement?

We answer the question of what sensor(s) and sampling rate(s) should be chosen as input to *VTrack* to achieve a given accuracy, given a set of energy cost measurements of sensors on the mobile device being used.

Section 3.8 describes related work to *VTrack*, and Section 3.9 concludes this chapter.

3.1 Background: Traffic Monitoring

Traffic congestion is a serious problem facing the road transportation infrastructure in many parts of the world. With close to a billion vehicles on the road today, and a doubling projected over the next decade [25], the excessive delays caused by congestion show no signs of abating. Already, according to much-cited data from the US Bureau of Transportation Statistics, 4.2 billion hours in 2007 were spent by drivers stuck in traffic on the nation’s highways alone [20], and this number has increased by between $3\times$ and $5\times$ in various cities over the past two decades. In addition, various surveys from news organizations show that in the developed world, dependence on roads and cars remains very high. For example 90% of US workers drive to work, the typical metro commuter spends on average 100 minutes driving per day, 33% of commuters in US cities are stuck in heavy traffic at least once a week, and worst-case delays are more than double the average delay [10, 9].

Real-time traffic information, either in the form of travel times or vehicle flow densities, can be used to alleviate congestion in a variety of ways: for example, by informing drivers of roads or intersections with large travel times (“hotspots”); by using travel time estimates in traffic-aware routing algorithms to find better paths with smaller expected time or smaller variance; by combining historic and real-time information to predict travel times in specific areas at particular times of day; by observing times on segments to improve operations (e.g., traffic light cycle control), plan infrastructure improvements, assess congestion pricing and tolling schemes, and so on.

An important step for all these tasks is the ability to *estimate travel times on segments or stretches of the road network*. Over the past few years, the idea of using *vehicles as probes* to collect travel time data for roads has become popular [1, 40, 71]. Here, vehicles equipped with GPS-equipped mobile devices or embedded computers, or individual commuters carrying mobile smartphones, log the current time and position periodically as they travel, sending this data to a server over some wireless network. This approach is better than flow-monitoring sensors, such as inductive loops, deployed on the roadside because vehicles can cover large areas more efficiently.

Estimating travel times on individual segments of the road network requires energy-efficient and accurate map-matching, which is the subject of the algorithms described in this chapter. In the crowd-sourced, mobile phone-based traffic monitoring context, keeping the GPS on all the time is not a viable strategy owing to energy concerns. If all users keep their phones powered while driving, then energy isn’t as much of a concern, but imposing that constraint is unreasonable barrier to large-scale deployment of a “crowd-sourced” traffic monitoring system. As we have discussed, energy concerns force the use of either GPS sub-sampling, or lower energy sensors such as WiFi or cellular radios, but these are accurate only to tens or hundreds of meters. Travel time estimation is a challenging problem in this context because *the closest road in the map to a position sample is often not the road that a vehicle actually drove on*.

We note that using smartphones as traffic probes raises some privacy concerns, which are out of scope for this dissertation; see Section 3.8 for a discussion of other work that addresses this concern.

To give a flavour of how *VTrack*’s estimates are actually used, the next section discusses two key applications that use travel time estimates: *hotspot detection* and *traffic-aware route planning*. This also helps gain a more precise understanding of how accurate the trajectories produced by *VTrack*, and the travel time estimates for these trajectories, need to be.

3.2 Applications

3.2.1 Detecting And Visualizing Hotspots

We define a “hotspot” to be a road segment on which the observed travel time exceeds the time that would be predicted by the speed limit by some threshold. Hotspots are not outliers in traffic data. They occur every day during rush hour, for example, when drivers are stuck in traffic. The goal of hotspot detection is to detect and display all the hotspots within a given geographic area which the user is viewing on a browser. This application can be used directly by users, who can see hotspots on their route and alter it to avoid them, or it can be used by route avoidance algorithms, to avoid segments that are frequently congested at a particular time.

For hotspot detection, the travel time estimates produced by a trajectory mapping algorithm need to be accurate enough to keep two metrics numerically low:

- The *miss rate*, defined to be the fraction of hotspot segments that the algorithm fails to report.
- The *false positive rate*, the fraction of segments reported as hotspots that actually aren’t.

If the trajectory mapping algorithm maps one or more input location estimate(s) to the wrong road segments, this may increase *both* the miss rate and the false positive rate for hotspot detection.

We give a simplified, but realistic, example (one that occurs quite often in the real data we have collected) to illustrate this point. Suppose the trajectory mapping algorithm needs to map a sequence of noisy WiFi location samples to one of two stretches of a freeway: A or B, with the segments lying on either side of a major exit on the freeway. Suppose in addition that what really happened was that segment A was congested — the vehicle being mapped spent a significant amount of time on segment A, *before* the exit, and sped up on segment B, immediately after traffic cleared at the exit. However, if the trajectory mapping system is confused by the error in the WiFi location samples and maps most or all of them to segment B, *after* the exit, then segment B will be flagged incorrectly as a hotspot instead of segment A. This causes an increase in the miss rate because the algorithm missed flagging segment A, and an increase in the false positive rate because the algorithm flagged segment B as a hotspot incorrectly. The consequence is that a route avoidance algorithm using the hotspot data would try to avoid segment B and perhaps route cars via segment A, which is ineffective and probably counter-productive (since the cars would get stuck on segment A in any case).

3.2.2 Traffic-Aware Route Planning

With the exception of hotspots, users are generally more concerned about their total travel time, rather than the time they spend on a particular road. Traffic-aware route planning that minimizes the expected time is one way for users to find a good route to a particular destination. Also, real-time route planning is useful because it allows users to be re-routed mid-drive.

Route planning using time estimates produced by *VTrack* has been deployed as part of the *iCartel* iPhone app as well as the *Carweb* personal driving portal, and has been running live for nearly two years now, providing route guidance to users of these applications. *iCartel* shows the user his/her position, nearby traffic, and offers a navigation service, while continuously sending position estimates back to a set of central server(s). The *Carweb* personalized driving portal is a website which shows users their own drives to/from their workplace and home, and also allows users to

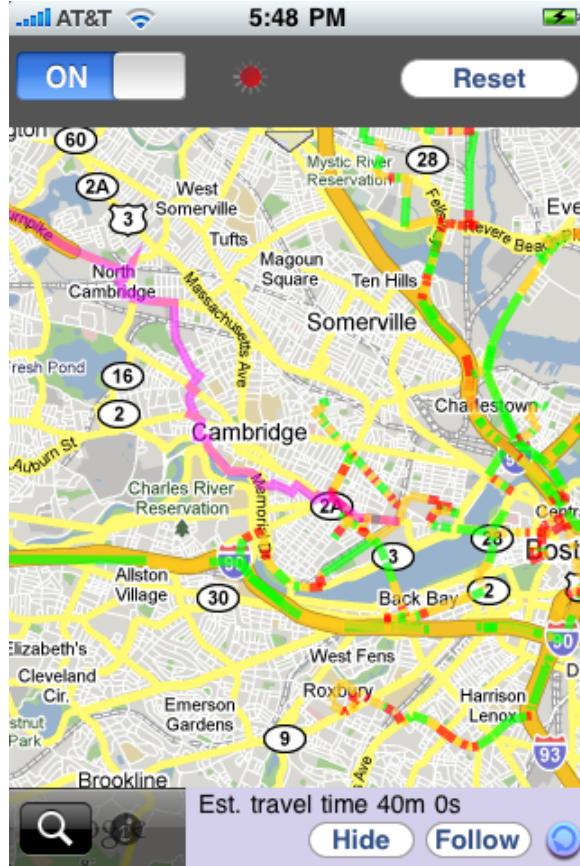


Figure 3-1: iCartel application showing real-time delays and congestion information.

view current traffic delays and receive route planning updates. Figure 3-1 shows a screenshot of the iCartel iPhone app and Figure 3-2 shows a screenshot from Carweb.

The goal of many route planning systems is to provide users with routes that minimize expected travel times. To meet this goal, *how accurate do the trajectories mapped need to be?* To analyze how accurate a set of travel time estimates are in the context of route planning, we use the travel time estimates to compute shortest paths in the road network — in terms of time — and study *how much worse* the shortest paths found are when compared to the true shortest paths, assuming the ground truth travel time is known. This is a realistic measure of how the quality of trajectory mapping estimates impacts the quality of end-to-end route planning. We believe that errors in the 10-15% range are acceptable — corresponding to at most a 3 to 5 minute estimation error on a 30 minute drive.

A second requirement is that the map-matching algorithm be efficient enough to run in real time as data arrives. Some existing map-matching algorithms run A*-style shortest path algorithms multiple times per point, which we found to be prohibitively expensive.

3.3 VTrack Architecture

Figure 3-3 shows the architecture of the VTrack system. Mobile smartphones report position data to the VTrack server periodically while driving. VTrack works with GPS and WiFi position estimates

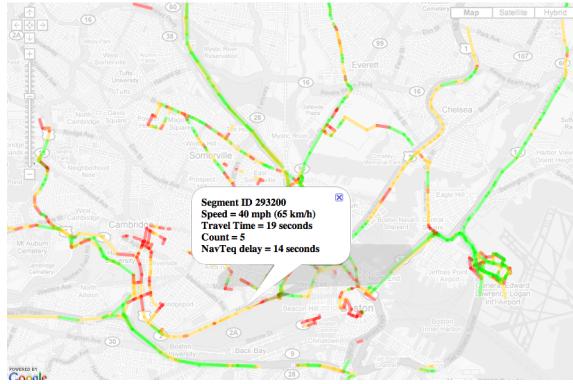


Figure 3-2: Web application showing traffic delays.

in the form of (lat, lon) coordinates at some (constant or variable) sampling interval. The server runs a Hidden Markov Model and travel-time estimation algorithm that uses these noisy position samples to identify the road segments on which a user is driving and estimate the travel time on these segments. The estimates are then used to identify hotspots and provide real-time route planning updates back to participating phones, as well as to a web service that provides traffic updates. The real-time estimates are also periodically logged to a database that stores historical delays. The historical travel time database is used in concert with real-time information to perform *traffic prediction*: i.e. predict what travel times might look like at a given time of day, or five minutes from now. Traffic prediction algorithms have been developed and are running live on the *VTrack* platform, but they are out of the scope of this dissertation.

Figure 3-4 shows the *VTrack* server in more detail. *VTrack* uses WiFi for position estimation as follows. Access point observations from WiFi in smartphones are converted into position estimates using a “centroid localization” algorithm that we shall describe shortly. This algorithm uses a training (also known as “war-driving”) database of GPS coordinates indicating where WiFi access points been observed from in previous drives. Positions from these sensors are fed in real-time to our estimation algorithm, which consists of two components: a *map-matcher*, which determines which roads are being driven on, and a *travel-time estimator*, which determines travel times for road segments from the map-matched trajectory.

3.4 The *VTrack* Algorithm

This section formally states the map-matching problem and then describes the Hidden Markov Model-based algorithm used in *VTrack*.

3.4.1 Problem Statement

The trajectory mapping problem in *VTrack* is the problem of matching input location samples to output road segments. Given the following as input:

- A known map of the road network that contains the geography of all road segments in an area of interest — examples include OpenStreetMap [68], and NAVTEQ [66]. While we have implemented *VTrack* on both map platforms, all the results in this dissertation use the NAVTEQ road map.

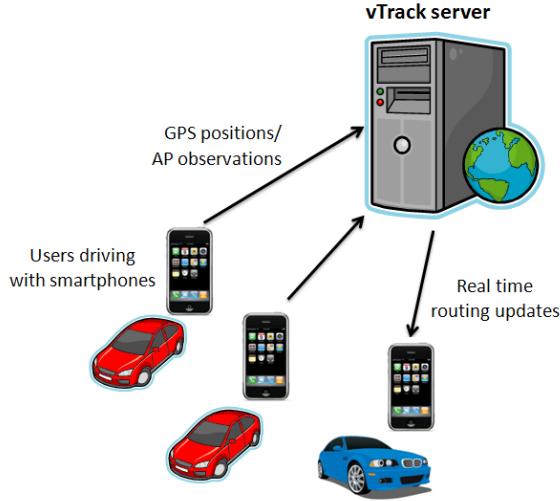


Figure 3-3: VTrack system architecture.

- A sequence of infrequent and/or inaccurate location coordinates with a possibly varying time interval between samples.

The goal is to produce the following as output:

- The most likely sequence of road segments traversed in the network. A given input point is matched to exactly one road segment. Multiple consecutive input points can be matched to the same road segment.

When our goal is to produce accurate travel time estimates from *sparse* input data, we need to modify the above definition slightly. If the location samples input to map-matching are very sparse (e.g., separated by minutes), then producing one road segment for each input location sample may not be sufficient to produce an unbroken sequence of road segments as output, especially if some of the road segments in the network are small segments.

In these cases, it is desirable to *interpolate* the output in some way to produce an unbroken sequence of road segments as output. Here, “unbroken” means each segment in the output sequence must be identical to, or adjacent to the next segment in the sequence. With this in mind, we use the following, slightly modified definition of “output” in *VTrack*:

- The most likely continuous sequence of road segments (guaranteed to be adjacent) traversed in the road network. A given input point is matched to exactly one road segment, but there can exist segments in the output not associated with any input point.

With the above modified definition, the system can (and often does) output *interpolated* road segments not corresponding to any particular input sample, especially if samples are spaced far apart.

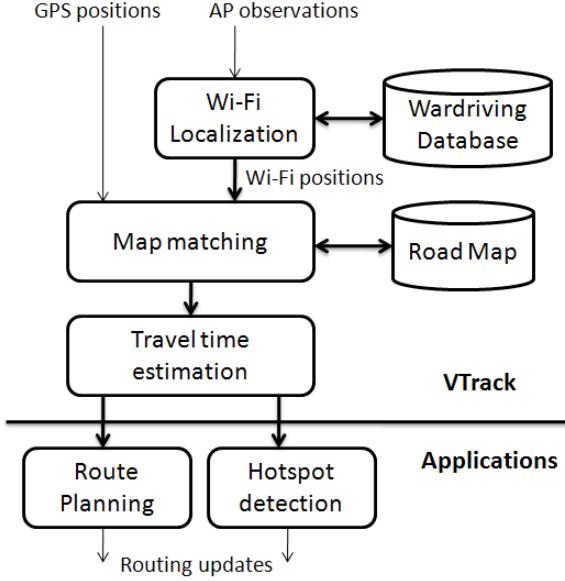


Figure 3-4: VTrack server.

3.4.2 Modeling Error In Location Samples

The input to *VTrack* is in the form of geographic (lat, lon) coordinates, irrespective of whether they are obtained from GPS, WiFi or cellular localization. *VTrack* also uses a simple *Gaussian* model of error in all of its input location samples irrespective of the localization technology used to compute them. It uses a different variance for the Gaussian varies based on which technology is used (smaller for GPS and larger for WiFi and cellular).

While the Gaussian model of error may be reasonably accurate for GPS [33], the error distribution of WiFi and cellular “location samples” depends on the procedure used to compute them. In the case of radio localization, the lowest-level raw data is usually in the form of a *radio signature* — a set of WiFi or cellular base station observations and their signal strengths. How this data was converted to geographic coordinates will determine what the errors in the coordinates are.

Our deployment of *VTrack* uses WiFi localization data from phones and from Linux-based access points re-purposed as in-car embedded devices. On the iPhone, WiFi position samples were obtained using Apple’s location API. On the embedded devices, WiFi scanning data was converted to coordinates using an averaging procedure that we call *centroid localization*. The procedure relies on a training database of access points and their signal strengths, and the ground truth GPS locations they were seen from. First, an approximate location is computed for each access point by finding the centroid of *all* GPS points in the training database from which that access point was ever seen. A given WiFi signature is converted to a location estimate by computing the “centroid of centroids”, i.e., the centroids of the individual access point centroid locations in the signature. The training database used for centroid localization was built over a period of over one year using 25 taxicabs part of the Cartel [17] testbed. Each taxicab in the testbed was equipped with an embedded Linux access point and a GPS device.

If there are multiple locations in the training database from which a WiFi signature can be seen, all far apart, then the averaging process in centroid localization is likely to introduce significant error. In particular, a single stray sighting of an access point from a far off location can introduce a large

error in the centroid estimate. Even if the Gaussian assumption no longer holds true, in practice, we find that in practice, *VTrack*'s Hidden Markov Model is able to correct errors in WiFi location estimates, and yields fairly accurate trajectories when matching location samples computed by the centroid algorithm.

3.4.3 Hidden Markov Models

In this section, we provide a brief introduction to Hidden Markov Models and explain how they apply to the map matching problem.

The simplest way to understand HMMs and how they apply to map-matching is the following. Suppose we first ignore the problem of infrequent location samples, and assume that location samples are frequent enough that it is sufficient to map each input location sample to an output road segment to produce an unbroken sequence of output segments. In this scenario, the simplest approach to map matching would be to map each position sample to the nearest road segment. However, as we show in Section 3.6.5, this scheme fails even in the presence of small amounts of noise in the input location estimates. What we want is not an approach that matches individual location samples, but an approach that matches a *sequence* of samples and exploits constraints on the transitions a moving vehicle or phone can make between the location measurements. This is exactly what a Hidden Markov Model enables.

A Hidden Markov Model (HMM) is a discrete-time *Markov process* with a discrete set of *hidden states*. The HMM takes as input a sequence of *observables*. Each observable is modeled as coming from an *emission probability distribution*, which is the probability distribution of the observable conditioned on the (unknown) hidden state. A HMM also permits *transitions* between its hidden states at each time step, governed by an adjacency graph that specifies which hidden states can transition to which other hidden state(s). These transitions are also governed by a different set of probability distributions, called *transition probabilities*.

In the context of *VTrack*:

- The hidden states are the (unknown) road segments a vehicle drove on that we want to identify.
- The observables are raw latitude and longitude position samples from GPS or WiFi.
- The emission probability for a given (segment, position) pair $\langle S, P \rangle$ pair represents the probability distribution of seeing position sample P given that the vehicle is on road segment S .
- The transition probability for a given pair of segments $\langle S_1, S_2 \rangle$ represents the probability of transitioning (driving) from segment S_1 to segment S_2 .

A Markov Model satisfies the following important independence property: the probability distribution of the observable at a given time t is conditionally independent of the distribution of past and/or future hidden states conditioned on the hidden state at time t . This makes it tractable to find the *most likely sequence* of hidden states corresponding to a given sequence of input observables. Due to conditional independence, the likelihood of any sequence of hidden states can be analytically shown to be equal to the product of emission and transition probabilities for that sequence. Hence, the most likely sequence of hidden states is simply the sequence that maximizes the product of emission and transition probabilities.

This most likely sequence of hidden states can be found exactly using an efficient *dynamic programming* technique called Viterbi decoding without having to exhaustively search all the possible paths through the HMM [4].

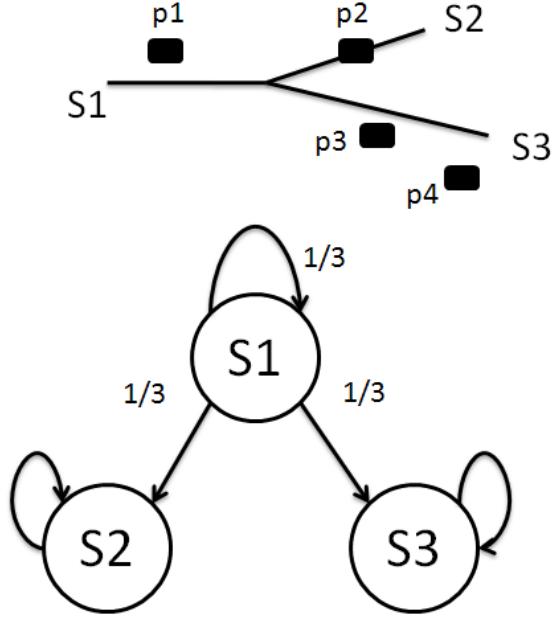


Figure 3-5: Example illustrating an HMM.

Figure 3-5 shows an example of an HMM where S1, S2, and S3 are road segments and p1, p2, p3, and p4 are position samples. There is an equal probability of transitioning from S1 to S1, S2, or S3. Because the emission probability density function is a decreasing function of distance from a road segment, assuming the transition constraints as shown in the state diagram, the maximum likelihood sequence of segments for the given sequence of position samples is S1, S3, S3, and S3. Hence, *although p2 is closer to S2*, the most probable hidden state of the point is S3 given the transition constraints.

As this simple example illustrates, a HMM makes sense for map-matching noisy location data because:

- It is robust to noisy position samples that lie closer to a *different* road segment than the one from which they were observed.
- It is able to guarantee a continuous (unbroken) output trajectory using transition constraints, unlike techniques that match individual samples to a sequence of output segments that may or may not be adjacent.

The HMM used in *VTrack* differs from previous HMM-based approaches to map-matching [36, 44] in two key ways:

- How it handles *gaps* and outages in the input data.
- How it models transition probabilities between road segments.

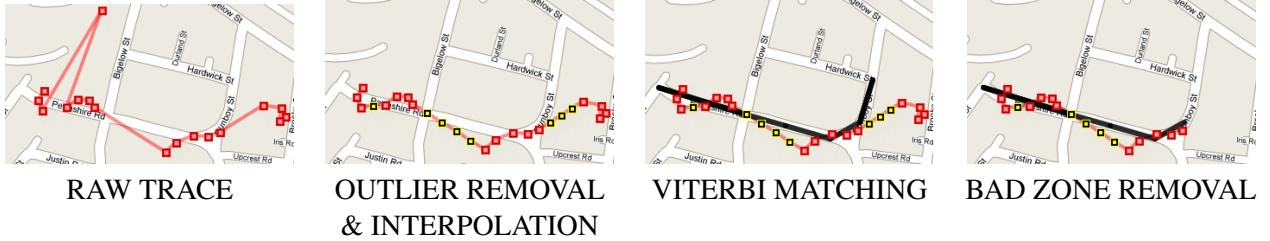


Figure 3-6: The map-matching process used by *VTrack*. A raw location trace is gradually refined to produce a final, high quality street route. In this example, due to noise and outages, only the first three segments produced quality estimates.

An additional difference is that *VTrack* uses *both* pre-processing and post-processing in addition to a HMM: it pre-processes input position samples to remove outliers and post-processes the HMM output to remove low-quality matches. This prevents it from producing inaccurate travel time estimates, as we shall explain below.

The following sections discuss the various stages of the *VTrack* algorithm: pre-processing, Viterbi decoding, and post-processing. Figure 3-6 provides a visual illustration of the stages of *VTrack*. We use this illustration as a running example through the sections that follow.

3.4.4 Pre-Processing and Outlier Removal

Prior to using a Hidden Markov Model, *VTrack* pre-processes its input location data to eliminate outliers. We say a location sample p is an *outlier* if it violates a speed constraint; that is, for some threshold speed $S_{outlier}$, the car would have had to travel faster than $S_{outlier}$ kmph to travel from the previous sample to p in the observed time between p and the previous sample. We choose $S_{outlier} = 400$ km per hour, well over twice the maximum expected vehicle speed on any road. This threshold is intentionally conservative and accommodates for errors in subsequent noisy GPS or WiFi samples.

After outlier removal, the next challenge is dealing with gaps in the input data. Gaps in input location data can occur due to two reasons:

- *Outages*, which occur when the data is missing. For example, GPS location data can be unavailable when the vehicle being tracked is in a tunnel or “urban canyon” with tall skyscrapers, and WiFi localization can experience outages when a mobile device is unable to find any access points in its vicinity that have been previously seen in the training database.
- *Sub-sampling*. If using a strategy such as GPS sub-sampling to save energy, the mobile device intentionally duty-cycles its GPS unit and only provides a location sample to *VTrack* sporadically, e.g., one GPS sample every 30 seconds or every minute.

VTrack uses a simple, computationally efficient, pre-processing scheme to deal with *all* gaps in its input location data — whether they be from outages or from sub-sampling. It uses *linear interpolation* to insert interpolated points in regions where location data is missing. The algorithm generates interpolated samples at a pre-determined interval along the line segment connecting the last observed point before the gap and the first following the gap, assuming a constant speed of travel along this line. The interpolated points are then fed to the Hidden Markov Model along with

the sampled points, and the HMM has the (relatively) easier task of matching each input location sample to a unique hidden state (road segment).

The pre-determined interpolation time interval used by *VTrack* is governed by the need to ensure the frequency of input to the HMM exceeds a threshold. Specifically, there is a danger that if the time between interpolated points is too high, the HMM cannot accurately match candidate paths that pass through small road segments. The input frequency must be high enough to ensure that even if driving through the *smallest* road segment in the road network database, at least one input point is associated with that segment. The smallest segment in the OpenStreetMap [68] and NAVTEQ road maps we use is roughly 30 metres. Assuming a maximum speed of 65 mph = 105 kmph = 29 m/s, the minimum frequency required is about once-a-second. Higher speeds than 105 kmph generally occur on freeways where segments are almost always longer than 30 metres. For this reason, *VTrack*'s linear interpolation scheme uses an interpolation interval of 1 second.

One limitation of linear interpolation is that it can fail in some situations where the interval of time between interpolated points is high. For example, suppose there are two points A and B in the road network with two candidate paths connecting them. One path (C) is curved while the other path (S) is straight. If *VTrack* is given a location sample at A, and then its next sample only at B, linear interpolation will always assume the vehicle drove on the straight path S even if the true path happened to be C. In this kind of scenario, a technique that computes the most likely path from A to B based on other external knowledge (e.g., an A* shortest path, or a path that drivers actually take most often in the road network) may work better than linear interpolation.

In practice, we have found that if using sub-sampled GPS, the *VTrack* HMM matches a straight-line interpolation to the *approximate shortest path* between the points being interpolated quite well for short gaps. This is because vehicles typically tend to follow shortest paths, *especially* at short time scales (a few seconds to a minute).

The outlier removal and interpolation steps used in *VTrack* are illustrated in the second panel of Figure 3-6 on an example trajectory. Interpolated points are shown in green.

We note that previous work that uses HMMs for map-matching either does not consider gaps [36], or deals with them by performing multiple computationally expensive shortest path calculations, or by invoking a route planner [44] to find the best transitions in the HMM. This is a computationally intensive process compared to linear interpolation.

3.4.5 Viterbi Decoding

Once outliers have been removed and interpolation has been applied, the Viterbi algorithm is used to determine the most likely sequence of road segments corresponding to the observed and interpolated points (third panel of Figure 3-6).

The output of *VTrack* depends quite critically on the choice of emission and transition probabilities for the HMM. We now explain our choice of emission and transition probabilities.

Emission Probabilities. The emission probabilities used in *VTrack* reflect the notion that it is more likely that a particular point was observed from a nearby road segment than a segment farther away, while not necessarily forcing the map-matcher to always output the closest segment to that point. Concretely, the emission probability density of segment i at position sample ℓ is $N(\text{dist}(i, \ell))$ where N is a Gaussian distribution with zero mean, and $\text{dist}(i, \ell)$ is the Euclidean distance between

i and ℓ . The standard deviation of N depends on the sensor that produced the sample. We use different standard deviations for GPS and for WiFi because they have different error distributions. For GPS samples, we use a standard deviation of 10 metres (as measured in previous studies of GPS accuracy), and for WiFi localization estimates, we use 50 metres (this number was obtained by computing the average error of WiFi localization over a subset of empirical data). The absolute value of the variances used does not matter (except for numerical precision) because the Viterbi decoder only uses the emission probabilities to *compare* the overall probabilities of paths, rather than using the absolute values of the probabilities.

Transition Probabilities. The transition probabilities in *VTrack* reflect the following three notions:

- For a given road segment, there is a probability that at the next location sample, the car will still be on that road segment.
- A car can only travel from the end of one road segment to the start of the next if it uses the same intersection (taking into account one-way streets). This constraint ensures that the output road segments form a continuous path through the road network.
- A car cannot travel unreasonably fast on any segment.

We define the transition probability p from a segment i at sample $t - 1$ to a segment j at sample t as follows:

- If $i = j$, $p = \epsilon$ (defined below).
- If j does not start where i ends, $p = 0$.
- If j starts where i ends, $p = \epsilon$ or 0 (this reflects the third notion, and is explained below).

An important point to note here is that *VTrack* sets *all* of the non-zero transition probabilities to a constant. The reason we do this is to avoid preference for routes with low-degree segments — routes with intersections without too many outgoing roads. An alternative approach, used by some other map matching algorithms such as [36], is to partition one unit of probability between all the segments that start at the end of i . This approach seems intuitive but results in overall higher transition probabilities at low-degree intersections than at high-degree intersections, which is *not* a faithful model of the underlying process. Section 3.6.7 of our evaluation shows that the partitioning approach has poorer map-matching accuracy than setting all the non-zero transition probabilities to a constant value.

Therefore, we want to ensure that the non-zero transition probabilities are constant across *all segments*, but at the same time sum to 1 for a given segment. To model this, we use a dummy “dead-end” state \emptyset and a dummy observable \perp . We set ϵ to be $\leq 1/(d_{\max}+1)$, where d_{\max} is the maximum number of segments that start at the end of the same segment (i.e., the maximum out-degree of a graph with intersections as vertices and road segments as edges). We set the transition probability from i at $t - 1$ to \emptyset at t so that the sum of the transition probabilities for segment i at t normalizes to 1. The transition probability from \emptyset at $t - 1$ to \emptyset at t is 1 — thus effectively assigning zero probability to all paths that transition to \emptyset .

The third item in the transition probability definition reflects the fact that it is desirable to prohibit cars from traveling unreasonably fast on any segment. If we are considering a transition from segment i to segment j , *VTrack* calculates the time it would have taken the car to travel from i to j , based on the times at which the positions were observed. If this time implies that the car would have had to travel at higher than a certain threshold speed, $S_{outlier}$, we conclude that this transition is impossible and assign it a probability of 0; otherwise, the transition is possible and we assign it a probability of ϵ . As mentioned before, we use a relatively relaxed value of 400 kmph for $S_{outlier}$ to avoid over-constraining the Hidden Markov Model in the presence of noisy position estimates. Section 3.6.6 of our evaluation shows that the speed constraint is essential to achieving good map-matching accuracy with noisy data.

The third panel in Figure 3-6 illustrates the most likely route found by the Viterbi decoder on our running example, in black. Note that the output is actually a sequence of \langle point, road segment \rangle pairs, where consecutive points can lie on the same road segment.

Performance Optimization. The running time of the Viterbi decoder is $O(mn)$ where m is the number of input location samples after interpolation, and n is the number of search states i.e. the number of candidate road segments that may lie on the path we are matching to. Including every segment in the road network in the search space is not feasible computationally, so *VTrack* uses a simple *geographic pruning* strategy to reduce the state space of the Hidden Markov Model. It does so by dividing the entire road network into relatively coarse-grained grids, which are a few hundred metres in size. We consider all the grids that include *any* point in the input trajectory being map-matched, and include all the segments in these grids in the search space. The grid size we use is approximately 500 metres, which is a sweet spot: using smaller grids sometimes fails to include relevant segments (in case of extreme localization errors in WiFi localization), and using a larger grid size increases computational complexity significantly if the road network is dense. The computational complexity increases with road density.

We have implemented *VTrack* in both C++ and Java. In practice, we have found both the implementations to be comfortably faster than real-time, map-matching hour-length drives within 2 minutes on a MacBook Pro with 2.33 GHz CPU and 3 GB RAM. Our live deployment of *VTrack* as part of Cartel [17] is able to map-match real-time data collected from an iPhone app and provide near real-time routing delays to end users, while running on data from the entire US road network.

3.4.6 Bad Zone Removal

The final step in *VTrack* is a *post-processing* step called “bad zone removal” whose goal is to remove zones where the output of map-matching is *known* to be bad or unreliable. Removing such zones is useful in the context of applications like travel time estimation for traffic monitoring. The idea is that if it is possible to use only data from segments of the output that are likely to be reliable, this improves the accuracy of travel time estimates and reduces the likelihood of using a wrong travel time estimate for routing.

Bad zone removal in *VTrack* uses a simple *confidence metric*: $d(P)$, the distance of each *observed* position sample P from the segment it is matched to in the output. The intuition is that if an observed position sample is too far from the segment it is matched to, it is extremely unlikely to be correctly matched. We use a conservative threshold of $d(P) \geq 100$ metres for bad zone removal (the 100 metres represents twice the approximate expected noise of 50 metres from WiFi localization estimates). Bad zone removal in *VTrack* works by eliminating entire zones around peaks in this “distance function” d . When the algorithm finds a candidate “bad zone” point P , it *backtracks* and

computes values of $d(x)$ for points x in the output preceding P . As long as $d(x)$ keeps decreasing, it continues tagging points, stopping only when $d(x)$ reaches a local minimum. The same strategy is also applied to points following P in the output.

The intuition behind this removal strategy is that when a position sample has high raw error, it is often likely that some samples preceding the sample have had increasing error, and vice-versa for following samples. This is the case with both GPS localization and WiFi localization. It is therefore prudent to eliminate the entire range of location samples from the output.

As we shall see in Chapter 4, bad zone removal works well mainly on GPS and WiFi localization data: no simple bad zone removal technique or confidence metric works effectively while map-matching cellular data, because the raw error of position samples is very high.

Bad zone removal is illustrated on our running example in the fourth panel of Figure 3-6.

3.5 Travel Time Estimation

The next step of *VTrack* is a *travel time estimator* that extracts travel times for individual segments from the output of map-matching.

The output of the map-matching step is a continuous sequence of segments in the road network. Each segment of the output has a corresponding (possibly empty) set of input point(s) matched to it. The traversal time $T(S)$ for any segment S in the output consists of three parts:

$$T(S) = T_{left}(S) + T_{matched}(S) + T_{right}(S)$$

$T_{left}(S)$ is the time between the (unobserved) entry point for S and the first observed point (in chronological order) matched to S . $T_{matched}(S)$ is the time between the first and last points matched to S . $T_{right}(S)$ is the time between the last point matched to S and the (unobserved) exit point from S .

As stated in Section 3.4.3, map matching adds interpolated points to ensure that each segment in the output has at least one corresponding input point. Hence, if map matching outputs a continuous sequence of segments, both $T_{left}(S)$ and $T_{right}(S)$ are upper-bounded by 1 second, and for segments that are not too small, $T_{matched}(S)$ is the main determinant of delay. To perform time estimation, we first assign $T_{matched}(S)$ to the segment S . We then compute the time interval between the first point matched to S and the last point matched to S_{prev} , the segment preceding S in the map match, and divide it equally¹ between $T_{right}(S_{prev})$ and $T_{left}(S)$, and similarly for the segment S_{next} following S .

Map matching does not always produce a continuous sequence of segments because bad zone detection removes low confidence matches from the output. We omit time estimates for segments in, immediately before, and immediately after a bad zone to avoid producing estimates known to be of low quality.

3.5.1 Travel Time Conservation

The strategy presented above, of partitioning the total travel time among individual segments, is only one possible way to estimate traffic on a road segment. An alternative, equally intuitive, approach

¹How we divide does not affect estimates significantly because the interval is bounded by 1 second.

to estimating traffic or congestion levels on individual segments might be to compute approximate vehicle *speeds* and associate the observed speeds to the map-matched segments. However, speeds computed from raw GPS data or even map-matched points produced by the HMM tend to be inaccurate, particularly on small segments where a small number of points (sometimes only one or two) are matched to the segment. More importantly, if not computed carefully, average speeds may not satisfy the following important *conservation principle*:

- The sum of estimated travel times for individual segments on *any* stretch of road segments S always equals the total observed travel time on S .

Note that the simple strategy of computing average speed on each traversed segment does not always satisfy the principle (it depends on how “speed” is defined).

The conservation principle is actually non-trivial: it turns out to be important in cases where the algorithm is confident of the travel time on a longer stretch, but unsure of the *distribution* of time on shorter sub-segments in that stretch. For example, given two road segments separated by a traffic light, we may not know precisely whether a car waited for a red light at the end of one segment, or crossed the intersection quickly but slowed down at the beginning of the next segment. However, as long as the *VTrack* travel time estimator conserves travel time, it will estimate the correct end-to-end travel time on any path that involves *both* segments — even if individual estimation errors on the two segments are high, they cancel out in aggregate. This cancellation of errors would not occur if a simple average speed estimator were used to compute travel time.

As our evaluation of *VTrack* in Section 3.6 shows, this “cancellation behaviour” actually occurs with travel times extracted from WiFi localization data: while time estimates on individual segments often have high errors, these errors tend to cancel out when the estimates are aggregated together for an application like end-to-end route planning.

The next section explains the types of estimation errors made by *VTrack*.

3.5.2 Estimation Errors

The main source of error in *VTrack*’s time estimates is inaccuracy in the map-matched output, which can occur for two reasons:

Outages during transition times. Transition times are times when a car is moving from one road segment to another. Without observed samples at these times, it is impossible to determine the travel time on each segment exactly. While map-matching can use interpolation to determine the correct sequence of segments, accurate time estimation for individual segments is harder than just finding the correct trajectory. For example, we cannot know whether a car waited for a red light at the end of one segment, or crossed the light quickly but slowed down at the beginning of the next segment.

Noisy position samples. Suppose that the location of a car is sampled just after it enters a short length segment, but a noisy sensor estimates the location of the car to be near the end of the segment. If this is the only sample matched to that segment, the location error is likely to translate into an extremely inaccurate time estimate. In particular, determining travel times for *small segments*, whose lengths are of the order of the standard deviation of the noise in the location data, is impossible with our approach. For example, consider the case of using GPS sampled every 60 seconds. In the absence of a sophisticated model for car speed, it is impossible to determine the travel time on a segment more accurately than to within 60 seconds, even if *all the location samples*

are perfectly accurate. Similarly, if a 100 metre error occurs in raw location data, it is impossible to estimate accurate travel time on a segment whose size is not much larger than this “resolution” of 100 meters.

Although *VTrack* is unable to estimate accurate times or speeds for individual small segments, Section 3.6 shows that the conservation principle is helpful here. Travel time estimation errors on adjacent or nearby segments tend to have *opposite* signs as long as the overall trajectory from trajectory mapping is correct. Because groups of segments (representing commonly traversed sub-paths) tend to be reused repeatedly in end-to-end routes, the *end-to-end travel time estimates* for routes produced by *VTrack* are highly accurate even with noisy or sub-sampled input data, and prove to be adequate for applications like route planning (Section 3.6).

3.6 Evaluation of *VTrack*

We have evaluated *VTrack* on a large data set of 800 drive hours of GPS and WiFi location estimates from real vehicular drives, obtained from the testbed of the MIT Cartel project [17].

This section is organized as follows. We first describe the Cartel testbed and the evaluation data set obtained from the testbed (Section 3.6.1). Since an important challenge in the evaluation was obtaining reasonable ground truth for the vehicular drives, this section also explains the procedure used to clean and obtain reasonable ground truth. We then overview the sensor sampling strategies that we compare in the evaluation (Section 3.6.2).

The first part of our evaluation is based on the two end-to-end applications mentioned earlier in this chapter:

- **Traffic aware routing**, whose goal is allow end users to compute shortest paths between a given source and destination using real-time traffic data collected from *VTrack*’s trajectory mapping system.
- **Hotspot detection**, whose goal is to detect road segments that are unusually congested or slow compared to their own historical mean, enabling users to manually avoid these hotspots.

These applications were built on *VTrack*. In each case, we use relevant *end-to-end metrics* from the application’s point of view to evaluate the quality of travel time estimates produced by *VTrack* for different sensor configurations with different GPS and WiFi sampling rates.

At a high level, we show that *VTrack* can achieve good accuracy for route planning using either WiFi localization or sparsely sampled GPS, saving energy in both cases (Section 3.6.3). We drill down into the results and show that this accuracy is achieved *in spite of large travel time estimation errors on individual segments*. We also show that sub-sampled GPS is effective up to certain sub-sampling frequencies for hotspot detection, but WiFi localization is less useful for hotspot detection owing to outages in WiFi data (Section 3.6.4).

The second part of our evaluation presents four micro-benchmarks that evaluate the impact of *VTrack*’s algorithm design choices on map-matching accuracy:

- We first (Section 3.6.5) address the question of how robust *VTrack* is to noise in input data. We show using simulations that *VTrack*’s HMM-based map matching algorithm is robust to significant amounts of simulated Gaussian noise, but breaks down beyond a certain levels of input noise, such as that found in cellular localization estimates.

- We next show that the speed constraint used in *VTrack* is essential to achieving good map-matching accuracy (Section 3.6.6).
- We then show that *VTrack*'s transition score improves map-matching accuracy over a simple partitioning approach (Section 3.6.7).
- Last, we evaluate the impact of the linear interpolation strategy used in *VTrack* and show that it compares favourably to more sophisticated strategies such as shortest-path interpolation in terms of accuracy on intermittently sampled input data (Section 3.6.8).

3.6.1 Data and Ground Truth

The Cartel Testbed. We use location data from the Cartel testbed [17], drawn from two sources:

- 25 taxicabs equipped with embedded computer running Linux with on-board GPS and WiFi, manufactured by Meraki [61]. The computer fits in a box, and draws power supply from the OBD connector on-board the vehicle.
- iPhone users running iCartel who contribute GPS data.

The Cartel testbed has been running for over 4 years as of the date of this thesis (since 2007). It has been used to collect and analyze a wide variety of sensor data — wireless network throughput, real-time traffic data, accelerometer data for detecting potholes [42], etc. Most of the taxicabs are operated by a Boston-area company called PlanetTran, who have been using Cartel's visualization interface to visualize the current location and historical drives taken by their taxicabs. 10 livery vehicles from JB Livery in South Boston have contributed data since 2010.

For the evaluation, we begin with a collection of 3998 drives collected from 25 of the PlanetTran taxicabs. GPS location data is sampled from an attached GPS unit that plugs into the Meraki computer in each car. The Meraki computer runs a WiFi driver which periodically scans for WiFi access points. The GPS and WiFi data are joined in time and uploaded back to the Cartel server either via any available WiFi uplink.

We first pre-processed the WiFi access point sightings to produce location estimates to feed to *VTrack*. The location estimates were computed using a “centroid of centroids” approach (as discussed in Section 3.4.2). Figure 3-7 shows the coverage map of our evaluation data, i.e., the set of distinct road segments on which our vehicles drove. The data set *after* cleaning (using a procedure described below) amounted to nearly 800 distinct hours of driving with both GPS and WiFi location estimates.

Ground Truth. Obtaining ground truth is a fundamental challenge in the evaluation of any map-matching and travel time estimation system. Approaches such as recording additional data in test drives are accurate but very time consuming, not scaling beyond relatively small amounts of test data. For example [48] makes video recordings to record road segment transitions. Another approach is to simply use GPS sampled at the highest possible rate as ground truth [57]. This works well a significant fraction of the time, but fails in regions where GPS is subject to errors and outages (e.g., near large buildings and in tunnels), making it inappropriate to use such data for ground truth.

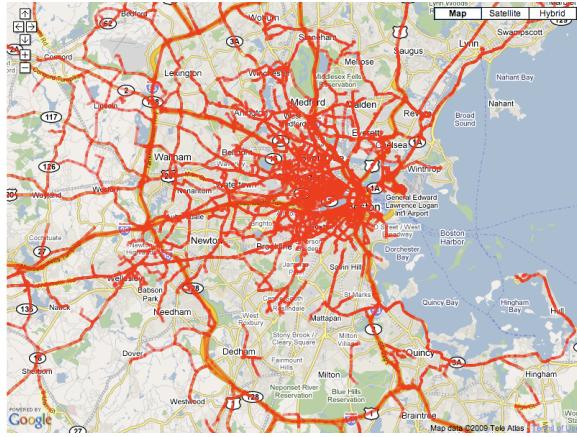


Figure 3-7: Coverage map of our evaluation drives.

As others have noted [44], it is difficult to get perfect ground truth for a large volume of data. We use GPS for ground truth, but rather than using it directly, we aggressively clean the data to produce ground truth with reasonable confidence for a subset of our drives. The goal of the cleaning is to *identify a subset of GPS data that can be treated as highly accurate with high confidence*.

The steps we use to clean raw GPS data are as follows:

- For each GPS point g in a drive, we consider the set of road segments S_g within a 15 meter radius of g (we picked 15 metres because it is the width of a typical road segment).
- We search the space of these segments to match the sequence of points g to a continuous sequence of segments X_g , such that each $X_g \in S_g$. Therefore, each GPS point is matched to one of its neighbours found in the previous step. We throw out all points g that cannot be matched this way (for example, if there are no segments within a 15 meter radius).
- We now look for outages of more than a certain duration (we used 10 seconds, approximately the median traversal time for a segment) and split the drive into multiple drives on either side of the outage, ignoring the duration of the outage.
- Finally, we project each g to the closest point on X_g to obtain a corresponding ground truth point g' .

The output traces from data cleaning, which we term *clean drives*, satisfy three key properties:

- No gap in the data exceeds 10 seconds.
- Each GPS point is matched to a segment at most 15 meters from it.
- The resulting segments form an unbroken drive.

These three properties taken together define a subset of the GPS data that can be treated as ground truth with high confidence. It is possible that systematic errors in GPS could cause this approach to fail, but we assume these are rare.

Limitations. The main limitation of the cleaning approach to ground truth is that using only clean GPS data (known to be within 15 metres of a road segment) may bias the comparison of GPS and WiFi localization. When using only clean GPS data, *GPS sub-sampling* will also yield accurate longitude/latitude coordinates, and is likely to do better on the subset of data we have selected compared to the entire data set. To avoid biasing our evaluation in favour of GPS and to realistically understand how it compares to WiFi, we perturb GPS samples with additive Gaussian noise before feeding them to the *VTrack* algorithm, as described below.

The advantage of the perturbation approach is that it is cheap because it only uses easy to collect GPS data, but realistic, because it restricts the evaluation to drives with near-perfect travel time data. Importantly, the evaluation does *not* test the system on these nearly-perfect samples, which would introduce bias by making GPS look artificially good, but instead tests on noisy versions of the samples to simulate real GPS behaviour.

The limitation of the perturbation approach is that it cannot model the impact of outliers in GPS, because there is no ground truth for regions where GPS fails. Hence all the results in this dissertation involving GPS are for the *outlier-free* case. For cases where GPS has significant outliers — *e.g.*, “urban canyons” or tunnels — using WiFi localization may be preferable. This is because it is possible to collect WiFi localization training data in urban canyons using expensive, high-accuracy dedicated GPS receivers that can decode GPS signals in urban canyons better than commodity mobile phone GPS chipsets.

Ground Truth Validation. We attempted to validate the above ground truth cleaning procedure on a small data sample (about 1 hour of driving) using a simple field drive in the Boston area. We used an Android phone equipped with GPS to record the location of the phone, and a phone application to mark the locations of turns. A human operator pressed a button whenever the vehicle crossed an intersection, signal or stop sign. We compared the travel times between successive turns *i.e.*, successive button presses, obtained from the application — which is as close to ground truth as human error would permit — to that obtained from the cleaned version of the GPS data obtained using the procedure described above. The average error in travel times from the cleaned version was 4.7% for one such 30 minute drive in Boston with approximately 30 segments, and 8.1% for another 30 minute drive in the Cambridge/Somerville area with approximately 50 road segments. Manual inspection of some of the errors revealed that a significant portion of the error was from human reaction time when marking turns, rather than from residual errors not filtered by the cleaning procedure. While this is a small sample, it gives us some confidence that the basic cleaning procedure is sound, and useful to establish ground truth.

Cleaning. We cleaned the raw data from the Cartel testbed using the procedure described above. In addition, we also omitted traces shorter than 2 kilometres, with fewer than 200 location samples (about 3 minutes) or fewer than 10 road segments. For the purpose of the travel time estimation applications, we also eliminated portions of traces where a car traveled slower than 2 km/h for 60 or more consecutive seconds, which we interpreted as parked. Typical traffic signals in the urban United States do not last more than this, but this number could be modified to accommodate longer signal wait times.

3.6.2 Strategies We Compare

We use the clean drives obtained using the ground truth procedure described above to obtain accurate ground truth travel estimates for each segment on each drive. We then evaluate the accuracy of travel

time estimates produced by four different sensor sampling strategies when using *VTrack*. The last of these is a *control strategy*. The four strategies are:

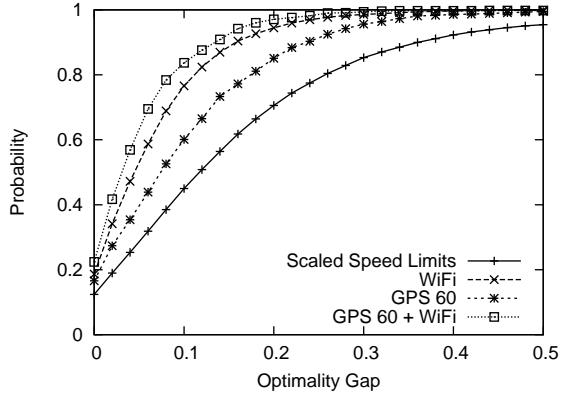
- **Continuous WiFi.** We use *VTrack* to map-match WiFi location data, compute travel times from the match, and compare to travel times obtained from the clean drives.
- **GPS every k seconds.** We sub-sample ground truth GPS data at intervals of k seconds, discarding $k - 1$ samples and use every k^{th} sample as input to the *VTrack* trajectory mapper. We add Gaussian noise with a standard deviation of approximately 7 meters to every k^{th} GPS sample (7 metres of Gaussian noise is acknowledged to be a reasonable model for GPS [33]).
- **GPS every k seconds + WiFi in between.** This is a combination of the sampling strategies above, where the *VTrack* trajectory mapper is given input samples *both* from sub-sampled GPS, and from WiFi location data in between GPS samples.
- **Scaled speed limits.** This approach is a “control” that does not use sensors. It instead uses speed limits from the NAVTEQ road database [66] to produce static travel time estimates for road segments. Since drivers do not drive exactly at the speed limit, directly using speed limits to compute travel times incurs a systematic bias. To alleviate this problem, we scaled all the speed limits by a constant factor k , and chose the best value of k — the value that pegs the mean difference between time estimates from speed limits and time estimates from ground truth to zero. We find the value of k to be close to 0.67, reflecting that drivers in our data drove at 67% of the speed limit on average. This strategy is intended to simulate the *theoretically best possible accuracy* that a strategy for route planning using only speed limits can achieve.

3.6.3 Route Planning

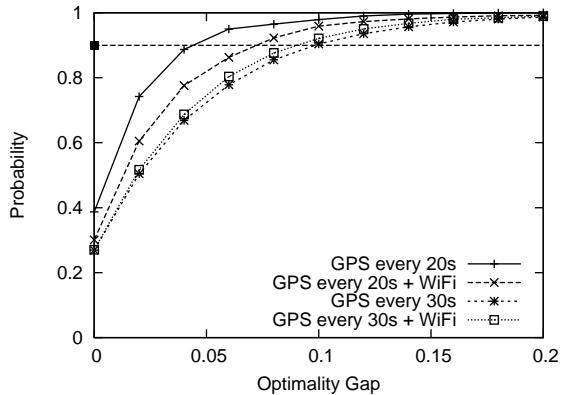
We evaluate the travel time estimates produced by *VTrack* for different combinations of GPS sampling rates and WiFi localization, in the context of route planning. In route planning, we are interested in finding the best end-to-end routes between source-destination pairs in the road network that optimize some metric. In this evaluation, we choose a popular metric: minimizing the expected drive time.

For each sensor setting, our evaluation takes as input a set of clean drives D_{gt} and a corresponding set of noisy or sub-sampled drives D_{noisy} — for example, using WiFi, or sub-sampled GPS, or a combination. We run *VTrack* on D_{noisy} to produce travel time estimates for each segment on those drives. Our goal is to understand how errors in these estimates from *VTrack* affect the quality of the shortest path estimate, i.e., *how much worse the paths the route planner finds using these inaccurate travel times are*, compared to the best paths it can find from ground truth travel times.

We consider the set of road segments for which we have a ground truth travel time estimate from at least one clean drive in D_{gt} . Call this set S_{gt} . We construct the *induced graph* G_{gt} of the entire road network on the set S_{gt} , i.e., the subset of the road network defined by the segments with ground truth and their end points. Note that G_{gt} may be disconnected: if so, we divide it into connected components. We pick a connected component at random and simulate a drive between a random source-destination pair within that component. In our simulation, we divided the graph into zones and picked the pair from different zones to ensure drives that have a minimum length of 2 kilometres. For each segment S in the graph, we pick two shortest path “weights”:



(a) GPS vs WiFi vs Scaled Speed Limits.



(b) GPS vs Hybrid (GPS + WiFi).

Figure 3-8: CDF of optimality gap when route planning using *VTrack*. Larger optimality gaps are worse.

- A ground truth weight for S picked from a clean drive D_{gt} in which S appears. D_{gt} is chosen at random from all clean drives containing S .
- An estimated weight, picked to be the *VTrack* travel time estimate for S from the noisy or sparse version of D_{gt} , D_{noisy} . This quantity is available whenever *VTrack* includes S in its estimated trajectory for D_{noisy} . If the *VTrack* trajectory for D_{noisy} omitted S , we fall back on estimating the travel time using the scaled speed limit.

We now run Dijkstra's algorithm on G_{gt} with the two different weight sets to find two different shortest paths, P_{gt} and P_{noisy} . To evaluate the quality of route planning, we compare the *ground truth times* for these two paths and compute the following “optimality gap” metric for each source-destination pair:

$$\text{Optimality Gap} = \frac{\text{Time}(P_{noisy}) - \text{Time}(P_{gt})}{\text{Time}(P_{gt})}$$

Note that the ground truth time for a path is always available because we use the graph induced by ground truth observations as the basis for route planning.

Figure 3-8(a) shows CDFs of the optimality gap across 10,000 randomly selected source-destination pairs for different combinations of sensors and sampling rates, as well as for a strawman which performs route planning using just scaled speed limits. As mentioned earlier in this chapter, we believe that an optimality gap of up to 10-15% is reasonable. This means that for an energy-saving strategy to be useful, it must produce a shortest path no worse than 35 minutes in travel time when the true shortest path takes 30 minutes in travel time.

Figure 3-8(b) shows the same CDF, but compares $GPS\ k$, the strategy of sampling GPS every k seconds and interpolating in between, to $GPS\ k + WiFi$, the strategy of using *both* sub-sampled GPS and WiFi location estimates in between.

Below, we state the most important conclusions from the results above:

- Travel times from WiFi localization alone are good enough for route planning. The 90th percentile of the optimality gap is 10-15% for WiFi, implying that 90% of simulated commutes found paths that were no worse than 10-15% compared to the optimal path, when using travel times estimated from WiFi localization alone.
- Travel times from GPS sampled every 30 seconds (or slightly more than 30 seconds) are good enough for high quality route planning. The 30-35 second value appears to reflect that interpolating raw GPS locations works very accurately for up to 3 or 4 missing road segments, but suffers more significant errors if interpolating over a long time scale.
- As we would expect (and hope), both GPS sub-sampling and WiFi localization significantly outperform the control strategy of using scaled speed limits for route planning. Using speed limits works reasonably well in the median, but incurs a significant tail of poor path predictions. All of these poor predictions in the tail correspond to scenarios of traffic congestion, which are precisely the important scenarios to evaluate a traffic delay estimation system.
- A hybrid strategy using $GPS\ 30 + WiFi$, i.e., both sub-sampled GPS every 30 seconds, and WiFi in between, improves performance over just sub-sampling GPS every 30 seconds ($GPS\ 30$) or over using only WiFi. However, as we shall see shortly, the gains are tempered by the significantly higher energy costs of using both sensors simultaneously.
- Using $GPS\ 20$ i.e., sub-sampling GPS every 20 seconds with interpolation in between, surprisingly outperforms $GPS\ 20 + WiFi$, i.e., sampling and using WiFi location estimates in between GPS samples. This is because outliers in WiFi hurt map-matching accuracy. In contrast, $GPS\ 30 + WiFi$ outperforms $GPS\ 30$. This suggests a “break-even” point where using WiFi is better than using interpolation of 30 seconds. The intuition is that *VTrack*’s map-matching interpolation technique works better than using WiFi localization over a time scale of 20 seconds or smaller. This time interval corresponds to one (or at most two) road segment(s). For the $GPS\ 20$ case, there is no inherent reason why adding WiFi information should hurt. We believe it should be possible to modify *VTrack* to filter out WiFi outliers and use only interpolated GPS information.
- For GPS sub-sampling intervals beyond 60 seconds, route planning starts to perform worse, with at least 10% of commutes finding paths at least 20-25% worse than optimal. This suggests that while sampling GPS less often than a minute might save significantly more energy, it is likely to yield medium-quality (reasonable but not very good) route predictions, at least for urban areas. Sampling GPS every minute corresponds to approximately half a kilometer in terms of distance at typical city driving speeds.

We now drill down into our results and show that, somewhat counter-intuitively, WiFi localization works adequately for route planning *in spite of large estimation errors on individual road segments* because *VTrack* correctly matches WiFi localization data to the correct trajectory most of the time.

Figure 3-9 shows a CDF of per-segment delay estimation errors, reported relative to ground truth travel time, on individual road segments when using only WiFi localization. WiFi localization has close to 25% median error and 50% mean error. However, the errors in the CDF are actually two-sided because *VTrack* when run on WiFi localization estimates often finds the *correct trajectories*. The main errors it makes are to mis-assign points on either side of segment boundaries. Hence, errors on groups of segments traversed on a given drive tend to cancel out, making end-to-end estimates more accurate than individual segment estimates. For example, the *VTrack* algorithm might distribute a travel time of 30 seconds as 10 seconds to segment *A* and 20 seconds to an adjacent segment *B* when the truth is actually 20 on *A* and 10 on *B*. However, if *A* and *B* are always traversed together, *this error does not affect the end-to-end travel time estimate*.

Figure 3-10 shows the accuracy of only trajectory mapping in isolation. This graph shows the CDFs for two metrics:

- *Point Error Rate (PER)* measures the frequency with which any given position sample in the input is matched to the wrong segment in the output of *VTrack*.
- *Segment Error Rate (SER)*, defined as the ratio $\frac{ED}{G}$, where *EG* is the edit distance between the map-matched output of *VTrack* and the ground truth trajectory, and *G* is the number of segments in the ground truth trajectory. *SER*, unlike *PER*, is not concerned with errors in matching individual points as long as the overall trajectory is correct.

We see from the figure that *VTrack* performs *significantly* better on *SER* than on *PER*.

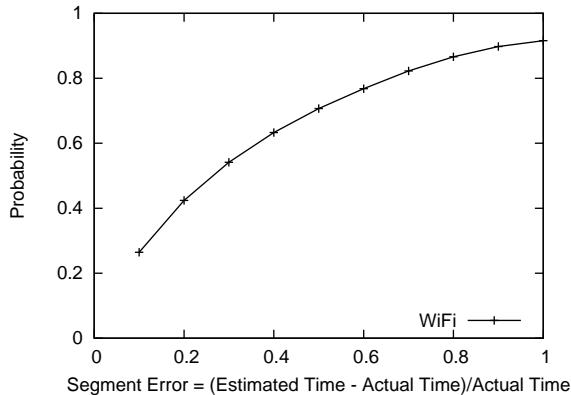


Figure 3-9: CDF of errors in time estimation for individual segments on WiFi localization estimates.

Figure 3-11 shows, in contrast, the errors in the actual estimated times for *end-to-end* routes found using WiFi localization. Here, we compare the estimated time for a path predicted using travel time estimates from *VTrack* with the ground truth time for the same path. The figure shows a plot of the CDF of the relative error in prediction. The graph shows that the end-to-end travel times predicted using WiFi localization are highly accurate, even at the 90th percentile, in contrast to times predicted using scaled speed limits.

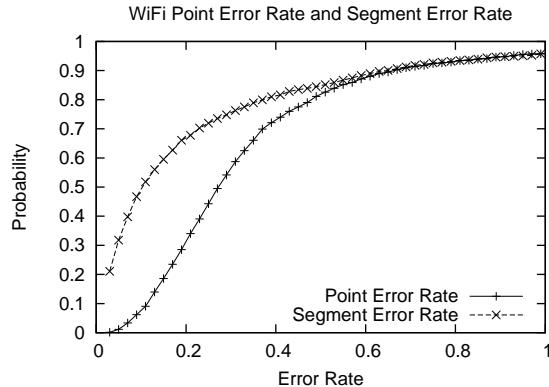


Figure 3-10: Map matching errors for WiFi localization.

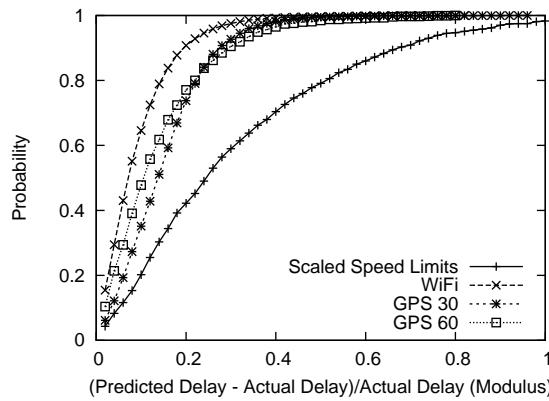


Figure 3-11: End-to-end time estimation accuracy using WiFi localization.

The results above validate the hypothesis that the trajectories produced by *VTrack* for WiFi localization data are often correct. It is the matching of points to segments within a trajectory that is more often wrong, which is why the strategy of following a *conservation principle* when calculating travel times (as discussed in Section 3.5.1) is a good idea.

Spurious Segments. The optimality gap metric presented above captures the impact of *both* incorrect predictions and missed segments in prediction. This is because *VTrack* falls back on a less accurate travel time, calculated from a scaled speed limit, when it misses a segment in its output. However, a limitation of the evaluation here is that it cannot directly capture the impact of *spurious segments*, segments produced by the map-matching algorithm that are absent in the original ground truth. We have found in separate experiments that the “error rate” for spurious segments i.e., number of spurious segments as a fraction of total number of segments in an output trajectory — is less than 15% for most of the sensor sampling strategies discussed in this evaluation — including WiFi alone, GPS sub-sampling more often than a minute, and combinations of GPS and WiFi (see Figure 3-12).

3.6.4 Hotspot Detection

This section *VTrack*’s time estimates in the context of hotspot detection, i.e., detecting which road segments are highly congested so that drivers can avoid them. We say a road segment is a “hotspot”

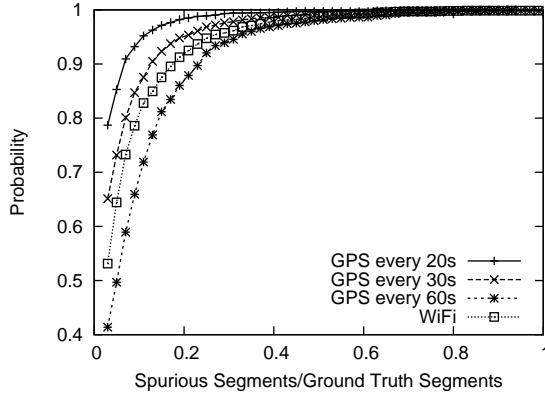


Figure 3-12: Spurious segment rates from map-matching for different sensors and sampling rates.

if the true travel time on a segment exceeds the travel time estimated from scaled speed limits by at least *threshold* seconds. The hotspots for a particular threshold value are the set of road segments which have high travel time based on the ground truth data. We evaluate *VTrack* by examining how many of those hotspots we can detect using the different sensor sampling strategies: *GPS k*, WiFi sampling, and *GPS k + WiFi* for different values of *k*.

An alternative definition of a hotspot would be *multiplicative*: a road segment in which the observed travel time is more than *threshold times* the travel time estimated with scaled speed limits. We found that this multiplicative definition is excessively biased by the length of the segment. Small segments have comparable variance in travel times to large segments. However, because they have a small travel time, they are more likely to be flagged as hotspots than large segments, which have a higher travel time. Therefore, we chose to use the first (additive) definition of “hotspot”, as it more accurately reflects the road segments that drivers truly view as hotspots.

To detect hotspots using a trace of WiFi, GPS+WiFi, or sub-sampled GPS data, we first map-match the data using *VTrack*. We then find the segments that meet the additive definition, i.e., have a travel time that exceeds the time estimated from scaled speed limits by some threshold. In addition to classifying each of these segments as a hotspot, we also classify the segments *adjacent* to each such segment as hotspots if they appear in the map-matched output. We use this strategy of flagging “groups” of segments as hotspots because, as we have seen earlier, the travel time estimation algorithm is not always certain as to precisely which road segment a high travel time should be attributed to, and almost always tends to be off by one segment, such as in a scenario with two segments on either side of a congested intersection.

To measure the accuracy of *VTrack*, we use two metrics:

- *Success rate*, defined as the fraction of hotspots found using ground truth travel time data that are detected from the output of *VTrack*.
- *False positive rate*, defined as the fraction of segments output by *VTrack* as hotspots, that are *not* hotspots in the ground truth.

A false positive amounts to suggesting that a driver avoid a segment that is not actually congested.

False positives and groups interact as follows. We record a false positive if we flag a group of segments as a hotspot, but that *group* is not a hotspot. We define a group as a hotspot if the total travel time on the group is more than $threshold \times number\ segments\ in\ group$ seconds above the travel time estimated by scaled speed limits.

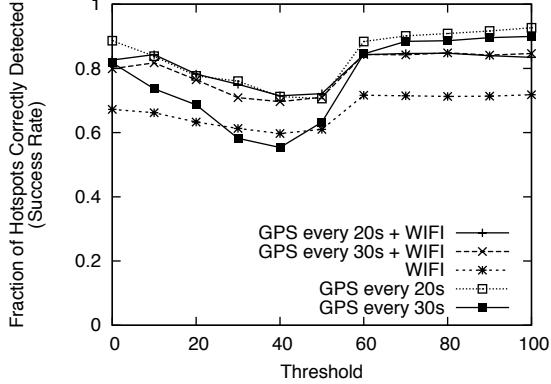


Figure 3-13: Success rate of hotspot detection with *VTrack*.

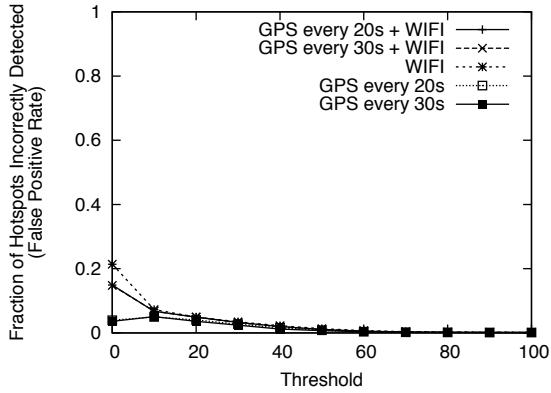


Figure 3-14: False positive rate of hotspot detection with *VTrack*.

Figure 3-13 shows the success rates when running *VTrack* on each sensor sampling strategy, and Figure 3-14 shows the false positive rates for each strategy.

There are a few interesting conclusions from these graphs:

- First, for the strategies involving GPS, the success rate is consistently above 0.8, and often around 0.9, implying that strategies using GPS sub-sampling can consistently detect between 80% and 90% of hotspots.
- The success rate for WiFi localization is much worse, frequently around 0.65, because WiFi localization experiences significant outages when there are no WiFi access point results returned in a scan. Hotspot detection cannot find a hotspot in regions for which it has no data. In contrast, GPS data has more complete coverage *even when sub-sampled*.
- If we restrict statistics to road segments where WiFi data does exist, WiFi localization has

a success rate comparable to all the GPS schemes. Hence rather than errors contributing to incorrect map matches, it is outages that affect the accuracy of hotspot detection.

- In all schemes, the false positive rate remains low. This is a desirable property as we do not want users to avoid road segments that are not actually congested.
- It is interesting to note that, with the exception of GPS every 30 seconds, *VTrack*'s success rate in every strategy remains relatively consistent across different values of the hotspot “threshold”. This indicates that our algorithm is fairly robust to a variety of applications that may have different requirements for what constitutes a hotspot.
- *VTrack*'s false positive rate also remains low for most threshold values, and for the most part only increases for small thresholds. This is due to the fact that, with a small threshold, *VTrack* is likely to flag many groups of segments as hotspots, and these groups may include some segments that do not have high travel time, but were included because they were adjacent to a segment with high travel time.
- We note that there is a dip in all of the strategies at around 40 seconds. At small thresholds, *VTrack* flags many segments as hotspots, and thus has both a high success rate and a relatively high false positive rate. As the threshold begins to increase, the algorithm starts to miss some hotspots, but the false positive rate decreases dramatically. This explains the portion of the graph before a threshold of 40.
- The second portion of the graph and the dip can be explained by examining the total number of hotspots. As the threshold increases, the number of hotspots naturally decreases. At a threshold of about 40 seconds, the rate of decrease slows, and from 60 seconds on, the number of hotspots remains fairly constant. This means that many of the road segments that are hotspots with a threshold of 60 are also hotspots with a threshold of 100; their observed time differs from their estimated time by over 100 seconds. As a result, *VTrack* does very well flagging hotspots at larger thresholds, since they are the more “obvious” hotspots, in some sense, and relatively resistant to small errors in travel time estimation.

Discussion of WiFi Outages. In our data, we found that WiFi localization has an outage rate of 42%, i.e., 42% of the time which we are trying to use WiFi localization, we do not get a WiFi scan result. This raises the question: how can a sensor that is so unreliable still perform well in some applications? In particular, we saw that although WiFi sensors did not work particularly well for hotspot detection, they did work well for route planning.

The reason for this is that in route planning, using the scaled speed limit estimates on segments where there is no WiFi data is generally sufficient to do reasonably well. Outages in WiFi tend to cause missed data points on segments that are relatively small in size. These are exactly the segments where scaled speed limit estimates are reasonable. Using scaled speed limit estimates on an *entire* path does not perform well because they cannot account for any variance or traffic congestion. However, using scaled speed limits as a back-up for segments missing WiFi localization *can* work well in certain cases.

In hotspot detection, on the other hand, we can never use the scaled speed limit estimates in place of actual location or map-matching data. After all, we define a hotspot as a road segment where the observed time estimate *differs* from the scaled speed limit estimates. This explains why WiFi localization alone is unsuitable for hotspot detection.

3.6.5 Robustness To Noise

In this section, we *simulate* the performance of *VTrack* using micro-benchmarks that test the algorithm for varying levels of noise in input data. To provide a reference strawman and to understand how much a Hidden Markov Model actually helps, we compare *VTrack* to the simple approach of simply matching each point to the nearest road segment in the road map. We demonstrate that *VTrack* is relatively robust to noise, much more so than nearest-segment matching.

In these micro-benchmark experiments, the input is once again “clean drives” cleaned using the procedure described in Section 3.6.1. Each location sample in each of these drives is labeled with the ground truth segment it came from.

To perform the micro-benchmarks, we generate a *perturbed* version of each cleaned drive by adding random zero-mean Gaussian noise with different standard deviations of 15, 40, and 70 meters to both the X and Y coordinates of each point in the drive. As a reference point, 40 meters noise roughly corresponds to the average standard deviation of WiFi localization data, though of course WiFi localization errors are generally non-Gaussian.

We compare the accuracy of *VTrack* and the simple nearest-segment matching strategy on each of the perturbed drives, in terms of the *point error rate* (PER) of each approach. As defined previously, PER is the fraction of points on a drive that are assigned to the wrong segment.

Figure 3-15 shows a CDF of the PER over all the perturbed drives, for each of the strategies under varying amounts of added noise.

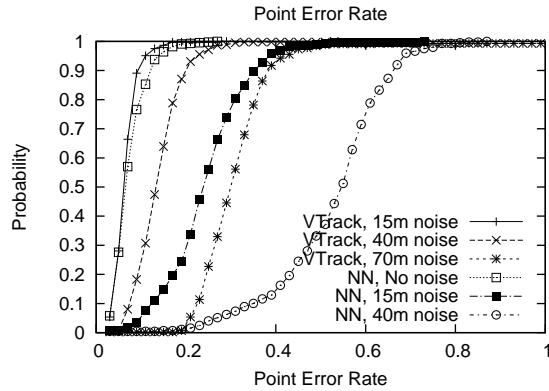


Figure 3-15: Accuracy of *VTrack* vs nearest segment matching (denoted by NN).

The results show that *VTrack* performs extremely well with 15 meter noise—about as well as nearest-neighbor matching with no noise. The median PER is less than 5%, and the 90th percentile PER is less than 8%. For 40 meter noise, these values are approximately 8% and 10%, not significantly worse.

Nearest segment matching performs much worse — even with just 15 meters noise, the median point error rate is almost 20%. Even with 70 meters of input noise, *VTrack* does better than nearest neighbor with 40 meters noise, achieving a median PER of about 20%.

In general, a Markov Model will be more accurate than nearest neighbor matching for the reasons described in Section 3.4.3; in particular, in cases like that in Figure 3-5, the HMM is able to correctly assign noisy point P_2 (which is nearest to segment S_2) to segment S_3 .

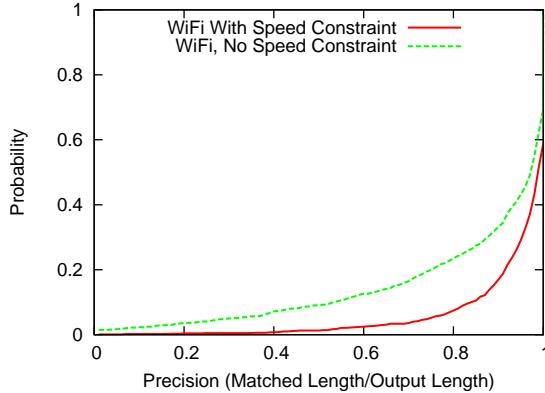


Figure 3-16: Speed constraint improves map-matching precision.

Beyond 70 meters noise (not shown in the figure), *VTrack* performs poorly, with a point error rate significantly higher than 20%. This suggests that the *VTrack* algorithm is not suitable to map match location estimates from cellular triangulation, which has errors of the order of hundreds of metres in urban areas.

In Chapter 4, we evaluate *VTrack* on cellular location data, and show that the *VTrack* approach of converting radio measurements (be they cellular or WiFi) into (latitude,longitude) coordinates is flawed when estimates are very noisy.

3.6.6 Impact Of Speed Constraint

In this section, we show that the speed constraint used in *VTrack* is critical to achieving good map-matching accuracy for the noisiest input drives.

We ran *VTrack* on WiFi localization data with and without the speed constraint enabled and compared the accuracy of map-matched outputs. For this evaluation, we used *precision*, a closely related metric to SER (segment error rate). Given a ground truth trajectory G and an output trajectory X , we match the road segments of G and X using a dynamic program that finds the closest alignment between the trajectories. We define the precision to be the fraction of the output trajectory X that is matched to some part of the ground truth G . A higher precision means better accuracy.

Figure 3-16 shows a CDF of the precision of map-matching with and without the speed constraint. The graph shows that while the speed constraint does not matter much in the median, it improves precision significantly — from 55% to 85% — for the worst 10% of input trajectories. The speed constraint is especially important for input data with significant amounts of noise because the Hidden Markov Model tends to “jump” from segment to segment following the noise without it.

3.6.7 Impact Of Transition Score

As mentioned earlier, *VTrack* assigns a transition probability equal to a small constant ϵ for all transitions out of a given intersection. The alternative would be to assign a transition probability of $\frac{1}{n}$ to each of n road segments leaving an intersection.

Figure 3-17 compares the two alternatives, showing a CDF of the map-matching precision. The figure shows that similar to the speed constraint, using an ϵ transition score is important to achieving good map-matching accuracy in the tail end of the CDF, i.e. for the noisiest of the input drives.

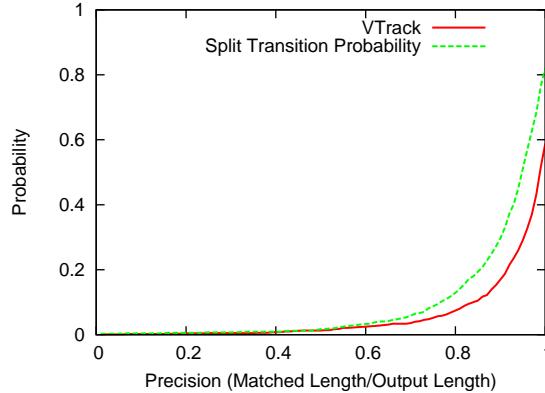


Figure 3-17: *VTrack*'s transition probability is better than a partitioned transition probability.

For the worst 10% of input trajectories, *VTrack*'s transition score improves precision over a simple partitioning strategy from 70% to 85%. This is because a simple partitioning strategy tends to overly bias in favour of paths containing intersections with low degree, i.e., with fewer major junctions.

3.6.8 Impact Of Interpolation Strategy

VTrack uses a simple linear interpolation strategy to deal with gaps in input data. One can also imagine alternative, more sophisticated interpolation strategies, such as computing the *shortest path* in the road network between intermittent location samples and using this path to “fill in” the sequence of road segments output by map-matching.

We performed a simple experiment to compare simple linear interpolation to shortest paths interpolation on sub-sampled GPS data. The shortest paths interpolation strategy interpolates between two location samples L_1 and L_2 as follows: it finds the closest segment S_1 in the road network to L_1 and the closest segment S_2 to L_2 . It then computes the shortest path P between S_1 and S_2 using Dijkstra’s algorithm, run with times from scaled speed limits used as weights, and fills in interpolated location samples along P at a frequency of at least one per second. The interpolated samples are fed together with the original samples to *VTrack*.

We use two metrics for evaluation: *precision*, as defined above, and *recall*. Given an output trajectory X and a matched region M between X and the ground truth trajectory G , the recall is defined as $\frac{\text{length}(M)}{\text{length}(G)}$, i.e., the fraction of the ground truth (in terms of length) recovered by the map-matching algorithm.

Table 3.6.8 shows the precision and recall of the two interpolation strategies for different frequencies of sub-sampling GPS. The table shows that, somewhat counter-intuitively, simple linear interpolation performs as well or better than shortest paths interpolation at all frequencies. This is likely because the shortest paths implementation uses nearest-segment matching to find the end points for shortest path computation, which is prone to error.

It may be possible to design a better shortest paths interpolation strategy. We chose to use linear interpolation in *VTrack* since it appears to perform at least as well as (or slightly better than) a simple shortest paths implementation, and has the added advantage of being simpler to implement and faster than shortest paths interpolation. Shortest path interpolation requires issuing multiple shortest path queries in each run of the map-matching algorithm.

GPS Sampled At	Precision		Recall	
	Linear	SP	Linear	SP
60 Hz	93.9%	91.3%	85.4%	84.6%
120 Hz	83.9%	74.8%	55%	57.9%
240 Hz	63.4%	60.5%	30.4%	34.0%

Table 3.1: Linear interpolation is as good, or better than SP (shortest paths).

3.7 Revisiting the Energy Question

Now that we are armed with:

- Energy measurements from a sampling of two different mobile platforms, and
- A characterization of how accurate trajectory mapping needs to be in real applications,

we are equipped to provide an answer to a central question of this dissertation: determining the most energy-efficient strategy for trajectory mapping on a given mobile device, and for a given application.

In this section, we combine the results presented above with the empirically determined model of energy costs for WiFi and GPS presented in Chapter 2. We compare three different strategies: GPS sub-sampled periodically, WiFi, and a hybrid strategy that combines the two. We show how the answer of which strategy to use depends on the following factors:

- The ratio of WiFi to GPS energy cost on the mobile device.
- The end-to-end application we care about (e.g., route planning vs hotspot detection).
- An “energy budget” that specifies the maximum energy that can be used (e.g., in the form of “the algorithm should not drain more than X% of the battery”).

Discussion Of Results

Based on Figure 3-8(a) presented earlier, we can infer that on devices where GPS is significantly more expensive than WiFi localization in terms of energy, WiFi sampling is the best strategy. The figure shows that WiFi localization is more accurate for route planning than GPS sampled every minute (*GPS 60*) and worse than GPS sampled every 30 seconds (*GPS 30*). It is approximately equivalent in accuracy to GPS sampled every 40 seconds (*GPS 40*). Hence, if *GPS 40* uses more energy than WiFi, WiFi should always be used in preference to GPS.

For example, on the iPhone 3G, it makes sense to use WiFi localization for route planning because GPS sub-sampling does not save energy on the iPhone (Chapter 2). However, even on a hypothetical iPhone with the ability to duty cycle GPS, it turns out WiFi localization would be preferable. Assuming GPS takes 6 seconds to acquire a “warm fix”, sampling GPS at frequencies up to two minutes per sample would use more energy than WiFi. Sampling GPS less often than every two minutes is *much* less accurate than WiFi — resulting in a higher optimality gap than WiFi), and sampling GPS more often than that on the iPhone 3G drains the battery quicker than using WiFi. Our prototype iPhone implementation using WiFi estimation would use about $\frac{1}{6}^{th}$ of the total charge of the

GPS Cost	Power Budget	Optimal Strategy
0.5	0.1	GPS 30
6	1	GPS 36
7	1	WiFi
24.9 (iPhone)	5	GPS 30
24.9 (iPhone)	3.5	GPS 60 + WiFi

Table 3.2: Strategies for different power budgets and GPS to WiFi energy cost ratios.

phone if it were run by a user for an hour a day with the screen on, which seems to be a reasonable level of battery consumption.

Similarly, Figure 3-13 shows that for hotspot detection, if WiFi sampling is very cheap such as on the iPhone it may be a good idea to *supplement* GPS sampling with WiFi localization. The figure shows that the *GPS 30 + WiFi* strategy yields a $5\times$ reduction in energy compared to GPS at 1 Hz, better than *GPS 20* (only a $3.3\times$ reduction) while having *equivalent accuracy*. Using WiFi alone for hotspot detection does not work owing to outages in the WiFi data, as we have discussed previously.

The results are quite different on the Android G1 phone where the relative cost of GPS is much lower than WiFi localization. On the G1, WiFi sampling is less than $2\times$ cheaper in terms of energy consumption than GPS, with the phone lasting a little over 10 hours with WiFi switched on, compared to a lifetime of 6 hours for GPS sampled at 1 Hz (Figure 2-1 in Chapter 2).

The G1 phone is also good at duty cycling GPS, acquiring a GPS “warm fix” in an average of 12 seconds each time the GPS is powered back on. Therefore, *GPS 24* is cheaper than or equivalent to WiFi localization in terms of energy consumption, while also being more accurate. On the Android G1 phone, GPS sub-sampling is consequently always a better strategy than WiFi localization.

Offline Optimization

Given a characterization of the energy costs of WiFi and GPS localization on *any* mobile device and a power budget, the evaluation results presented in this chapter make it possible to perform this kind of analysis to derive an optimal sampling strategy for that device. *VTrack* uses such an offline optimization strategy, first measuring the energy costs of GPS and WiFi sampling using a battery drain experiment (such as that presented in Chapter 2) and then using the cost values to decide the sensor(s) and sampling rate(s) to use.

For any device, we measure the power consumption of sampling GPS at the max possible rate (g) and WiFi (w) and determine the ratio g/w of the power consumed by GPS to that consumed by WiFi. Now, suppose we are given a target power budget p and the ratio g/w , it is possible to perform an analysis to determine the best sensor(s) and sampling strategy to use on that device to maximize accuracy for the target application (e.g., routing or hotspot detection) for that particular combination of p and g/w . The space of options we consider in this discussion are GPS every k seconds for some k , WiFi or a combination of GPS every k seconds with WiFi. It is possible to extend a similar analysis easily to other combinations of sensors, such as WiFi sub-sampling.

Table 3.7 shows the results of solving this optimization problem with route planning as the target application for some sample values of p and g , assuming the WiFi cost w is 1 unit.

For example, a power budget of $p = 2$ units in Table 3.7 means that *VTrack* is allowed to use at most twice the power consumed by WiFi.

Clearly, some settings of these parameters don't make sense — for example, $p < 1$ and $g > 1$ — so we only consider meaningful parameter settings. We see that there is a threshold of GPS cost g beyond which using WiFi is the best option, as one would expect. Also, given a small power budget on a phone like the iPhone where GPS is power-hungry, *VTrack* starts to use WiFi localization.

Figure 3-18 illustrates the solution to the optimization problem visually to make the solution space to the energy optimization clearer. The solution is presented visually as a function of the power budget $\frac{p}{w}$ and the GPS sampling cost $\frac{g}{w}$, both expressed as ratios to WiFi sampling cost. To read the graph easily, first look at the case when $p = w$, i.e., the part of the graph along the x axis. Here, it is impossible to use both GPS and WiFi together, and there is a clear choice between using *GPS k* and WiFi. For values of g below a certain threshold, *GPS k* is preferable and for values above, WiFi is preferable. Next, when $p \geq w$, *GPS + WiFi* is always preferable to just WiFi, because the additional GPS points that can be sampled with the extra power budget never hurt accuracy.

The next choice we consider is between *GPS k* for some sampling interval k , and *GPS k' + WiFi* at a higher sampling interval k' , where k' is chosen so that the energy cost of *GPS k' + WiFi* and *GPS k* are equal. First consider the case when $\frac{g}{w}$ is very large. In this case, the cost of WiFi is negligible, and for any k , *GPS k* and *GPS k' + WiFi* have approximately the same energy consumption, so it purely comes down to which one is better in terms of accuracy. From our empirical data, for approximately $k = 20$ (labeled $k_{critical}$ in the graph), using *VTrack*'s interpolation with GPS is preferable to map-matching both GPS and WiFi together. Hence, whenever the power budget exceeds approximately that of GPS 20 (the dotted line) it is preferable to use only sub-sampled GPS. The graph also shows that for lower values of g , or beyond a certain power budget, it is better to use the power to increase the GPS sub-sampling rate (i.e., to reduce k) than to use WiFi.

In the figure, the iPhone 3G corresponds to the region of the graph where $\frac{g}{w}$ is large, and hence using *GPS + WiFi* or WiFi is preferable to just sub-sampling GPS. The Android G1, on the other hand, falls in the region of the graph where $\frac{g}{w}$ is relatively low and it is preferable to always use GPS sub-sampling and never use WiFi localization.

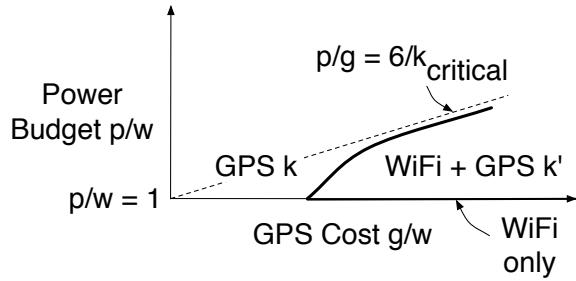


Figure 3-18: Diagram showing optimal strategy as a function of power budget and GPS to WiFi energy cost ratio.

Limitations of Analysis. We note that the accuracy results for routing and hotspot detection, and hence the optimal sensor sampling strategy to use, also depend on other factors not studied here. In a non-urban setting with lower road density, lower GPS sampling rates may be sufficient to estimate travel times accurately. There may be fewer or more WiFi hotspots in some areas, which would make WiFi a less or more viable strategy, respectively. At the opposite extreme, in very dense urban environments, WiFi localization might even be preferable to GPS. GPS in these settings may

perform worse than indicated by our results, because “urban canyon” effects from tall buildings impair its accuracy.

3.8 Related Work

As in the Cartel project, the Mobile Millennium project at UC Berkeley [1] built software to report traffic delays using mobile phones as probes. They focused on real-time traffic reporting. Claudel et al. [22, 21] develop a model for predicting flows and delays in the near future from traces of probe data. However, they assume GPS data and do not look into the effects of noisy data.

The NeriCell project [71] focused on monitoring road conditions and traffic using smartphones. Their goal was to combine data from GSM localization, GPS, and accelerometers to get a picture of road surface quality as well as traffic conditions, such as locations where users are braking aggressively. Their primary contributions are in processing accelerometer data, as well as power management. They do not provide algorithms for travel time estimation or map-matching.

Using Hidden Markov Models and Viterbi decoding for map-matching has been proposed by Hummel et al. [36] and Krumm et al. [44]. However, their work has mainly focused on map-matching GPS data with low noise. To the best of our knowledge, there have been no quantitative studies of accuracy. There has been some work in the database community on the map-matching problem [24, 76], which has largely focused on efficient algorithms for matching points to road segments when data is regularly sampled and not particularly noisy (comes from GPS or sensors without outages).

Gaonkar et al. [77] present the idea of a “micro-blog” where users can annotate locations and pick up other users’ annotations as they travel. They use an energy-efficient location sampling technique, but focus on providing approximate estimates of location, rather than performing map matching or estimating travel time.

Yoon et al. [48] use GPS data to classify road conditions as “good” or “bad,” both spatially and temporally (which reflect the steadiness and speed of traffic, respectively). They take frequently-sampled GPS points and calculate how a car’s delay is distributed over each segment. This “cumulative time-location” data is converted to spatio-temporal data and then classified. This work differs from ours in that it assumes the ability to get relatively accurate travel time estimates on individual segments. Their method would likely fail on noisy data such as WiFi localization estimates, because the cumulative time-location measurements would be incorrect.

Privacy is an important concern in location-based smartphone applications, but is out of the scope of this dissertation. For completeness, we refer the reader to some related work on privacy-preserving smartphone traffic monitoring here. Using approaches inspired by the notion of *k-anonymity* [52], Gruteser and Grunwald [54] show how to protect locational privacy using spatial and temporal cloaking. A number of recent works show how to protect locational privacy while collecting vehicular traffic data [14, 43, 53] and in GPS traces [15]. In addition, some recent papers [16, 15] have developed tools to quantify the degree of mixing of cars on a road needed to assure anonymity (notably the “time to confusion” metric). The virtual triplines scheme [16] proposes a way to determine when it is “safe” from the standpoint of privacy for a vehicle to report its position using such a quantification. Many of these techniques could be used in *VTrack*.

3.9 Conclusion

This chapter presented *VTrack*, a system for trajectory mapping on mobile phones that can accurately estimate road travel times from a sequence of inaccurate and/or infrequent position samples

using a map-matching algorithm based on Hidden Markov Models. *VTrack* uses a measurement-driven optimization strategy to choose the best location sensor(s) (GPS or WiFi localization) and sampling rate(s) to use, on any given mobile device.

We evaluated *VTrack* on two end-to-end traffic monitoring applications: route planning and hotspot detection. We presented a series of results that showed an approach based on Hidden Markov Models can tolerate significant noise and outages in location estimates, while still providing sufficient accuracy for end-to-end traffic monitoring applications.

We presented an analysis using the evaluation results in this chapter that provides some answers to the question of which sensor(s) and sampling rate(s) to use for a given energy budget and mobile phone platform. A key result was that GPS sampled once every 40 seconds was approximately equivalent in accuracy to WiFi localization, and both WiFi and GPS sampled at a frequency up to once a minute worked reasonably well for map-matching.

One limitation of *VTrack* and the Hidden Markov Model we identified is that it begins to break down for increasing amounts of noise in the input data (Section 3.6.5), and in particular does *not* work for map-matching cellular location estimates. Since cellular localization is virtually free in terms of energy consumption on most phones (Chapter 2) this leaves open the following question:

- *Can we accurately map-match highly inaccurate cellular location data? How?*

This question is the focus of the next chapter.

Chapter 4

Map-Matching With Soft Information

The *VTrack* system presented in the previous chapter matches noisy geographic coordinates from GPS or WiFi localization to a sequence of road segments. With WiFi or cellular localization, the raw data collected is usually in the form of base stations and their signal strengths. If this *soft* information is available, it is possible to use it to significantly improve accuracy over *VTrack*-like strategies that use only “hard” information, i.e., (lat, lon) coordinates. This chapter presents *CTrack* [6], a system that can use soft information from radios to perform more accurate map-matching than *VTrack*.

We implement and evaluate *CTrack* in the context of map-matching cellular signal data. Soft information is crucial to achieving reasonable accuracy for map-matching cellular signals. Cellular fingerprints can be seen from locations as far apart as hundreds of metres to a kilometre, even in densely populated urban areas with many cell towers. *Averaging* these locations using a process like centroid localization (described earlier) or fingerprinting [57] results in large errors in individual “location samples”. As we have seen earlier, *VTrack* produces very inaccurate output trajectories when its input location coordinates have more than 70-80 metres of error.

4.1 Why Cellular?

The major advantage of a trajectory mapping system that can use purely cellular signals is that on a mobile phone, the marginal energy consumed for trajectory mapping is nearly zero (Chapter 2).

A second reason why cellular is interesting is that a large number of phones today do not have GPS or WiFi on them—85% of phones shipped in 2009, and projected to remain over 50% for the next five years [19]. The users of these devices, a disproportionate number of whom are in developing regions, are largely being left out of the many new location-based applications.

4.2 How CTrack Works

CTrack incorporates soft information by using a *two-pass* Hidden Markov Model. The first pass matches raw radio tower and signal strength data to a sequence of *grid cells* in the area of interest. The second pass uses *VTrack* to match the sequence of grid cells output by the first pass to a sequence of road segments on the road map.

The second contribution of *CTrack* is that it augments radio fingerprints with “sensor hints” from *inertial sensors* on a smartphone: the magnetic compass and the accelerometer. Specifically, *CTrack* extracts two kinds of binary hints:

- A *movement hint* that indicates if the phone is moving or not at a given time instant.
- A *turn hint* that indicates if the phone is turning or not at a given time instant.

CTrack fuses these “sensor hints” into its Markov Model to improve the accuracy of trajectory mapping. We show in this chapter that using movement and turn hints helps correct some of the systematic errors that arise with cellular localization, while consuming almost no energy.

4.2.1 Summary Of Results

We have implemented *CTrack* on the Android smartphone platform, and evaluated it on nearly 125 drive hours of real cellular (GSM) signal data (1,074 total miles) from 20 Android phones in the Boston area. We find that:

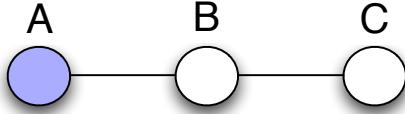
- *CTrack* is good at identifying the sequence of road segments driven by a user, achieving 75% “precision” and 80% “recall” accuracy (these terms are defined later in this chapter). This is significantly better than state-of-the-art cellular fingerprinting approaches such as Placelab’s algorithm [57] applied to the same data, reducing the error of trajectory matches by a factor of $2.5\times$.
- Although *CTrack* identifies the exact segment of travel incorrectly 25% of the time, trajectories produced by *CTrack* are on average only 45 metres away from the true trajectory. This implies that *CTrack* is more than adequate for applications like route visualization, or personalized drive monitoring applications that need to know which of k possible routes a user took — for example, to find the best of k different routes from a user’s home to workplace in terms of travel time [64]. In this respect, *CTrack* is $3.5\times$ better than map-matching “hard decision” coordinates from cellular fingerprints, which has a median error of 156 metres.
- *CTrack* on cellular signal data has a significantly better energy-accuracy trade-off than *GPS k*, the GPS sub-sampling strategy described in Chapter 3. For example, it reduces energy cost by a factor of $2.5\times$ compared to *GPS k* for the same level of accuracy on the Android platform.
- In absolute terms, *CTrack* with cellular signal data uses very little energy on a modern smartphone. Our experiments on both the Android G1 and Nexus One phones show that these phones sampling data for *CTrack* have lifetimes close to their standby lifetime when not sampling any sensors.

The rest of this chapter is organized as follows. Section 4.3 explains how and when soft information helps map-matching. Section 4.4 describes the architecture of the *CTrack* system. Section 4.5 describes the actual *CTrack* algorithm. Section 4.6 describes sensor hints. Section 4.7 evaluates *CTrack* on real driving data collected on the Android platform. Section 4.8 discusses related work to *CTrack*, and Section 4.9 concludes this chapter.

4.3 Why Soft Information Helps

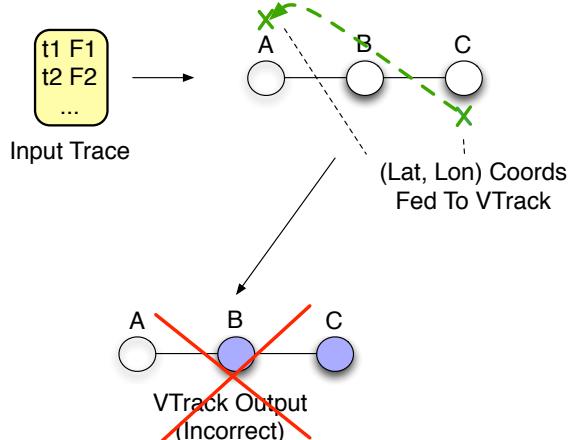
We use the term *fingerprint* to refer to a set of observed IDs of WiFi or cellular base stations and their associated received signal strength (RSSI) values at any time instant. *CTrack* uses training data to compute a detailed probabilistic model of which fingerprints are seen from which geographic

A sees F₁,F₂ C sees F₁



Ground Truth = A

(a) Ground truth: device at location A.



(b) What *VTrack* would do.

Figure 4-1: Example demonstrating the benefits of soft information for map-matching.

locations. It uses this probabilistic model to match a sequence of fingerprints to a sequence of grid cells on a map.

An approach that does not use soft data would instead convert each fingerprint to the closest or most likely geographic location, and run *VTrack* over the resulting locations. There exist different approaches to converting a radio fingerprint to a geographic coordinate. For example, Placelab and RADAR [57] identify the closest matching fingerprint in their training databases for a given WiFi or GSM fingerprint, and output its location. One can also use a centroid or averaging approach as described earlier. All of these “hard decision” approaches must reduce a fingerprint to a single geographic location, which *loses valuable information* and results in lower trajectory mapping accuracy.

To understand why, consider a simplified world where there are only three discrete possible locations: A, B, and C, all of which lie on a straight line, with B midway in between A and C (as shown in Figure 4-1(a)). A mobile device can only transition one location in one time step (it can go from A to B, or B to C, but not directly from A to C). Suppose further that the device whose trajectory we are trying to map lies at location A throughout the experiment (though the algorithm we want to design does not know this).

Suppose we are given two radio fingerprints F_1 and F_2 from the device at two instants of time t_1 and t_2 . Our goal is to determine the trajectory of the mobile device. The true answer is that the device stayed at location A at both time instants. Both A and C see fingerprint F_1 , but only A sees fingerprint F_2 (perhaps because they are equidistant from the same cell tower). The closest location

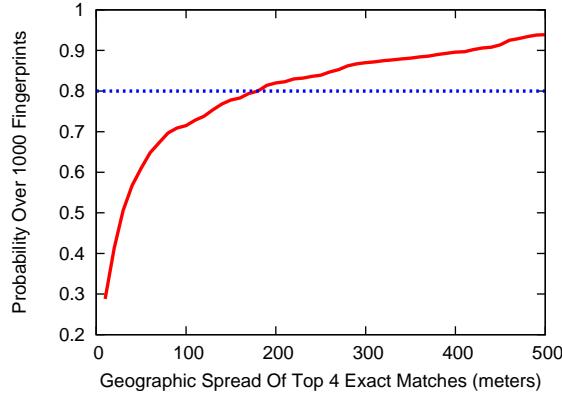


Figure 4-2: Geographic spread of exact matches. The dashed line shows the 80th percentile.

corresponding to fingerprint F_1 is C, while the closest location corresponding to fingerprint F_2 is A. In this situation, a centroid or closest-match algorithm that first converts the fingerprints to the closest matching location obtains CA as the sequence of geographic locations to map-match (as shown in Figure 4-1(b)). When such a sequence is input to an algorithm like *VTrack*, it is likely to output the wrong answer (perhaps CB as shown in the figure or BA, but unlikely to be AA).

What we really need to do is to *not* throw out the “soft” information that F_1 is also likely to be seen from A, which is what a hard decision approach effectively ends up doing. Since fingerprint F_1 is likely to be seen from *both* A and C, the trajectory AA is most likely *even though the closest location to F_1 is C*.

To quantify this intuition on real cellular signal data, we selected approximately 1000 cellular fingerprints at random from a large training data set collected in the Boston area from Android G1 phones (as described later in this chapter). For each fingerprint F , we found all the *exact matches* for F , i.e., locations F' with the exact same set of towers in the training data as F . We ordered the matches by similarity in signal strength, most similar first, and computed the geographic *diameter* of the top k matches for each fingerprint (using $k = 4$).

The figure shows that more than 20% of matching sets have a diameter exceeding 150 meters, and more than 10% have a diameter exceeding 400 meters. The centroid or closest match approaches result in large “conversion error” for fingerprints with such a large geographic spread of locations.

Hence, the right approach is to use the soft information by keeping track of *all* possible likely locations a fingerprint is seen from, and using a continuity constraint to *sequence* these locations. This is the key intuition behind the *CTrack* algorithm we present later in this chapter.

A good analogy to understand *CTrack* vs *VTrack* is that of *soft-decision decoding*, as opposed to *hard-decision decoding* in wireless communication [47]. In hard-decision decoding, a sequence of symbols received on the wireless channel is converted to the most likely bit — either 0 or 1 without retaining any soft information about the likelihood that the symbol was in error. Soft-decision decoding, in contrast, preserves this likelihood information and passes it to the error correcting decoder.

The rest of this chapter describes and evaluates our implementation of *CTrack* on cellular (GSM) signal data.

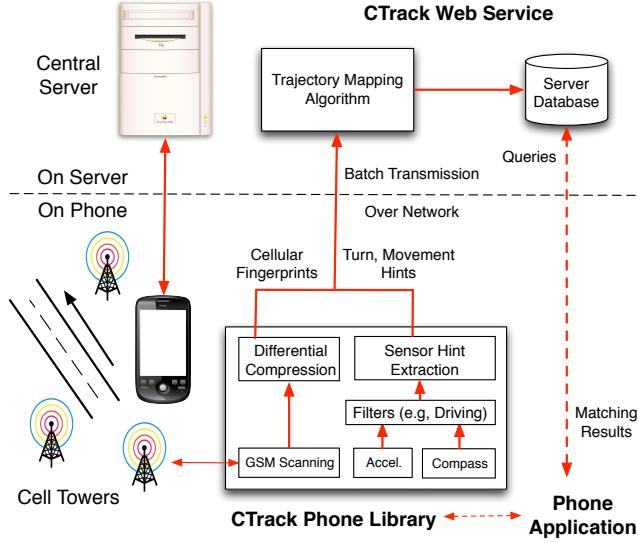


Figure 4-3: *CTrack* system architecture.

4.4 *CTrack* Architecture

Figure 4-3 shows the system architecture of *CTrack*. *CTrack* consists of two software components, the *CTrack Phone Library*, and the *CTrack Web Service*. The library collects, filters, and scans for GSM and sensor data on the phones, and transmits it periodically via any available wireless network (3G, WiFi, etc.) to the web service, which runs the trajectory mapping algorithm on batches of sensor data to produce map-matched trajectories. The mapping algorithm runs on the server to avoid storing complete copies of map data on the mobile device, and to provide a centralized database to which phone or web applications can connect to view and analyze matched tracks e.g., for visualizing road traffic or the path taken by a package or vehicle.

4.4.1 Phone Library

The phone library collects a list of neighbouring GSM towers sampled every 2-3 seconds approximately — because a cellular scan yields a new signature only once every 2-3 seconds. Optionally, if accelerometer, compass, or gyroscope sensors are available on the phone, the library also collects current values of *sensor hints* extracted from these inertial sensors. These sensor hints are binary values indicating if the phone is moving and/or turning. Section 4.6 describes how we extract sensor hints. The phone library uses the accelerometer to filter out portions where a user is stationary or walking, as described in [75, 5] and in Chapter 5 of this dissertation. This is to ensure we supply only relevant data to applications that want data only from moving vehicles. The library may also be configured to periodically collect GPS data for use in the training phase of *CTrack* from users who wish to contribute training data.

Since the goal of running *CTrack* on cellular data is to minimize energy consumption, it is important to minimize the energy consumption of transmitting cellular data over the network to the map-matching server. Our current implementation collects and transmits about 120 bytes/second of raw ASCII data on average. This quantity varies because the number of cell towers visible varies with location. In practice, it is possible to use two additional ideas to minimize the energy consumption of delivering sensor data over the network:

- Simple compression. Using gzip compression on our data set of 125 hours of test drives resulted in an average of just 11 bytes/second of data to be delivered.
- Batching the sensor data and uploading a batch every t seconds. At 11 bytes/sec, with even small batches, using a 3G uplink with an upload speed of 30 kBytes/s (typical of most current 3G networks in the US) results in very low 3G radio duty cycles—for example, setting t to 60 seconds results in the radio being awake only 0.03% of the time in theory. The duty cycle in practice is higher because the 3G radio does not immediately switch to a low power state after a transmission [63], but a large enough batching interval consumes a tiny amount of additional power. Once in 5 minute ($t = 300$) reporting is sufficient for most trajectory mapping applications (many of which use data from historical tracks in any case)—including recovering user tracks, traffic reporting, package tracking, and vehicular theft detection.

We chose not to run trajectory matching on the phone because it results in negligible bandwidth savings, while consuming extra CPU overhead and energy. For low data rates, the primary determinant of 3G or WiFi transmission energy is the transmitter duty cycle [63], making batch reports a good idea. However, we do extract sensor hints on the phone because the algorithms for hint extraction are simple and add negligible CPU overhead, while significantly reducing data rate. The raw data rate from sampling the accelerometer/compass without compression or hint extraction is about 1.3 MBytes/hour, which means that an application collecting this data from a user’s phone for two hours a day could easily rack up a substantial energy and bandwidth bill without on-phone computation.

4.4.2 Web Service

The *CTrack* web service receives GSM fingerprints and converts them into map-matched trajectories. These matched trajectories are written into a database. Optionally, for real-time applications, the user’s current segment can be sent directly back to the phone. A detailed description of the trajectory mapping algorithm is given in the next section.

4.5 *CTrack* Algorithm

The *CTrack* algorithm for map-matching a sequence of radio fingerprints differs from previous approaches in two key ways, as mentioned earlier. First, *CTrack* uses a two-pass algorithm which first matches cellular fingerprints to a sequence of spatial grids on a map, and then matches (lat, lon) coordinates from the grids to road segments. The goal of the two pass algorithm is to avoid the loss of information from reducing a fingerprint to a single geographic location.

Second, *CTrack* optionally fuses sensor hints from the accelerometer and the compass to improve trajectory matching accuracy. Specifically, we show that turn hints can help remove spurious turns and kinks from GSM-mapped trajectories, and movement hints can help remove loops, a common problem with GSM localization when a vehicle is stationary.

4.5.1 Algorithm Outline

CTrack takes as input:

- A series of radio fingerprints from a mobile device, one per second in our implementation. In our implementation for GSM localization, the Android OS gives us the cell ID and the RSSI of up to 6 neighboring towers in addition to the associated cell tower. Each RSSI value is an integer on a scale from 0 to 31, where higher means a higher signal-to-noise ratio.

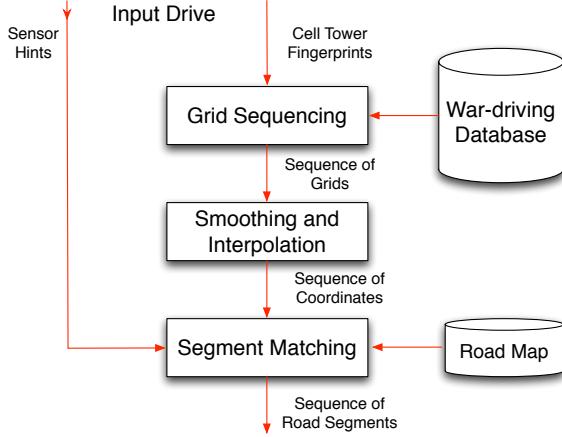


Figure 4-4: Steps in *CTrack* algorithm.

- If available, time series signals from accelerometer, and compass or gyroscope sampled at 20 Hz or higher. These are converted to “sensor hints” using on-phone processing as explained later in this chapter.
- A known map database that contains the geography of all road segments in the area of interest, similar to that used by *VTrack*. We use the NAVTEQ road map database for the experiments in this chapter.

The output of *CTrack* is the likely sequence of road segments traversed, one for each time instant in the input.

Figure 4-4 shows the different components and passes of the *CTrack* algorithm. *Grid Sequencing* uses a Hidden Markov Model (HMM) to determine a sequence of spatial grid cells corresponding to an input sequence of radio fingerprints. The output of grid sequencing is *smoothed*, *interpolated*, and fed to *Segment Matching*, which matches grid cells to a road map using a *different* HMM. This stage also uses the output of a *Hint Extraction* phase (not shown), which processes raw accelerometer and compass data to extract *movement hints* and *turn hints* at different time instants in the drive. Offline, the *Training* phase (not shown) builds a training database, which maps ground truth locations from GPS to observed base stations/cell towers and their RSSI values.

Before we go into the details of *CTrack*, we point the reader to Figure 4-5. This figure illustrates the *CTrack* algorithm visually with an example. We shall use this trajectory as a running example throughout this section to explain how *CTrack* works.

We begin by noting that the input “raw points” in Figure 4-5(a) are shown only to illustrate the extent of noise in the input data. They are not actually used by *CTrack*, which uses radio fingerprints as inputs to its first step. The “raw points” in Figure 4-5(a) were computed by using the Place-lab/RADAR fingerprinting algorithm [57], where a radio (here, cell tower) fingerprint is assigned a location equal to the centroid of the closest k fingerprints in the training database (we used $k = 4$). This approach is the state-of-the-art technique for computing a single closest match to a cellular fingerprint, and is shown here mainly to illustrate the extent of error in the raw location data without using any probabilistic model to sequence the data.

The next sections describe each stage of *CTrack*.

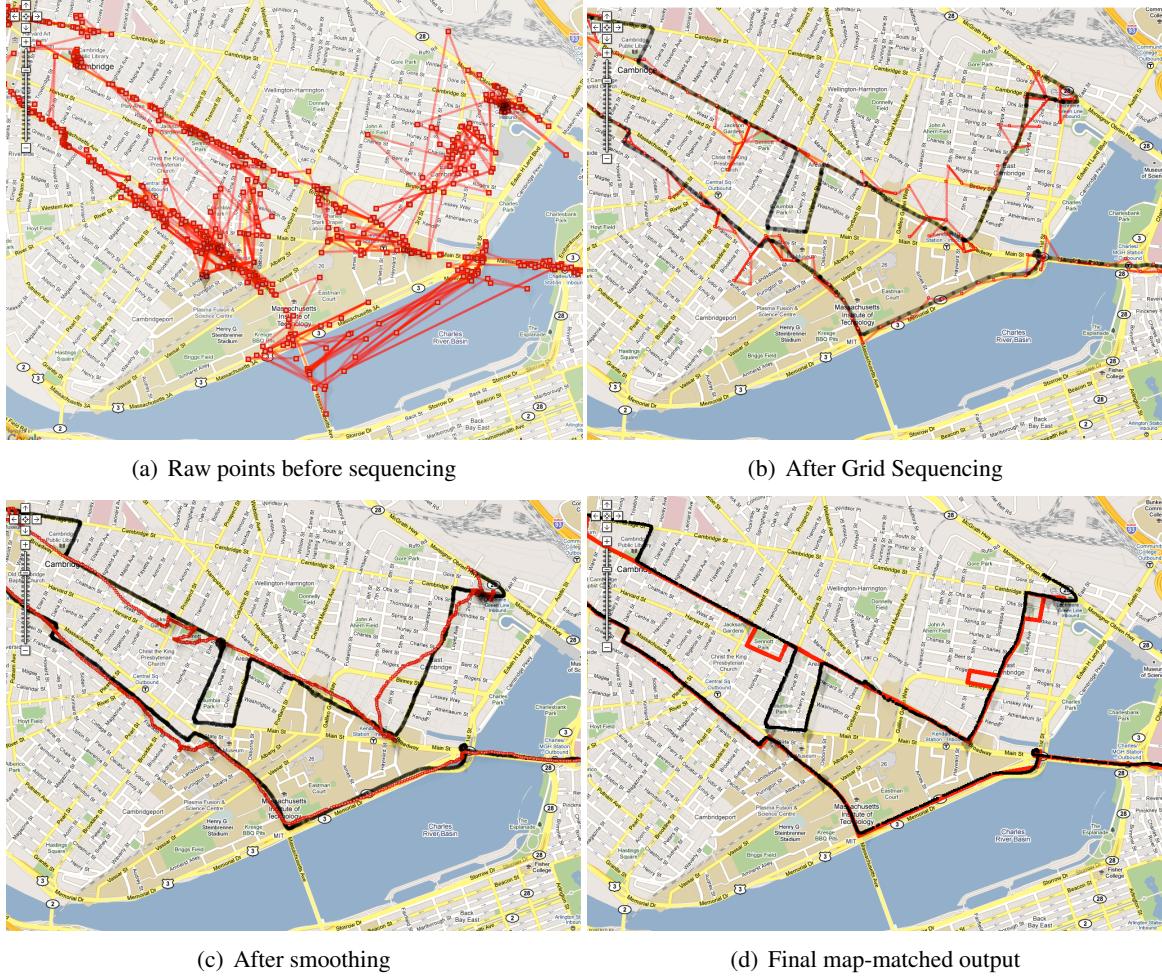


Figure 4-5: CTrack map-matching pipeline. Black lines are ground truth and red points/lines are obtained from cellular fingerprints.

4.5.2 Training

We divide the geographic area of interest into uniform square grid cells of fixed size g_s . We associate with each cell an ordered pair of positive integers (x, y) , where $(0, 0)$ represents the south-west corner of the area of interest. We use $g_s = 125$ meters. As we shall see, a smaller grid size results in higher accuracy if there is sufficient training data within each grid. However, too small a grid size can also result in lack of sufficient training data within each grid, as well as higher computational cost. Our grid size of 125 metres was chosen to balance running time and density of training data against the absolute algorithm precision (though we did find a reasonable range of grid sizes smaller and bigger than 125 metres to work reasonably well).

We train *CTrack* for the area of interest using software on mobile phones that logs a timestamped sequence of ground truth GPS locations and associated fingerprints. For each grid G in the road map, our training database stores F_G , the set of distinct fingerprints seen from some training point in G . We did not use special rules to decide which training paths to drive, but we tried to cover all the road segments within the area of interest at least once.

Training can be done out-of-band using an approach similar to the Skyhook [78] fleet, which maps

WiFi access points to ground truth GPS coordinates. Once the training database is built, it can be used to map-match or track any drive, and needs to be updated relatively infrequently. We can also collect new training data in-band from consenting participating phones that use the *CTrack* web service whenever the user has enabled GPS.

4.5.3 Grid Sequencing

CTrack's grid sequencing step uses a Hidden Markov Model to determine the sequence of grid cells corresponding to a timestamped sequence of fingerprints. For an overview of how Hidden Markov Models work, we refer the reader to Section 3.4.3 of Chapter 3.

In the HMM used for grid sequencing, the hidden states are grid cells and the observables are radio fingerprints. The emission score, $E(G, F)$ captures the likelihood of observing fingerprint F in cell G . The transition score, $T(G_1, G_2)$, captures the likelihood of transitioning from cell G_1 to G_2 in a single time step.

CTrack first pre-processes input fingerprints using the *windowing* technique described below. *CTrack* then uses Viterbi decoding on the HMM to find the maximum likelihood sequence of grid cells corresponding to the windowed version of the input sequence.

We now describe the four key aspects of the HMM used by *CTrack* for grid sequencing: windowing, hidden states, emission score, and transition score.

Windowing

Because it is common for a single cell tower scan to miss some of the towers near the current location, we group fingerprints into *windows* rather than use raw fingerprints captured once per second. We aggregate the fingerprints seen over W_{scan} seconds of scanning. to improve the reliability of the fingerprint. We chose $W_{\text{scan}} = 5$ seconds empirically: the phone typically sees all nearby cell towers within 3 scans, which takes about 5 seconds. In our evaluation, we show that windowing improves accuracy (Table 4.1 of this chapter).

Hidden States

The hidden states of our HMM are grid cells. Since there are an enormous number of grid cells in the entire road map, running the Viterbi dynamic programming algorithm over *all* the grid cells in the map is computationally intensive. It is necessary to intelligently identify a subset of grid cells that are relevant candidates for a given fingerprint and narrow down the search space of the Viterbi decoder. Recall that this was easy to do in *VTrack* where the radio fingerprints had already been converted to (lat, lon) coordinates. *VTrack* could hence use a simple geographic approach to prune the state space of its Hidden Markov Model. This is less straightforward in *CTrack*.

CTrack uses the following heuristic for pruning its search space. Given an observed fingerprint F , a grid cell G is a candidate hidden state for F if there is at least one training fingerprint in G that has at least one cell tower in common with F . Note that we might sometimes omit a valid possible hidden state G if the training data for G is sparse or non-existent. This is a problem because eliminating such a grid G will also eliminate all paths that transition through road segments in G , incorrectly pruning the search space.

To overcome this problem, *CTrack* uses a simple *wireless propagation model* to predict the set of towers seen from cells that contain no training data. The model simply computes the centroid and

diameter of the set of all geographic locations from which each tower/base station is seen in the training data. The model draws a “virtual circle” with this centre and diameter and assumes that all cells in the circle see the tower in question.

Thanks to the propagation model, it is possible for *CTrack* to retain some flexibility in producing trajectories that pass through the occasional grid with little or no training data, but are otherwise highly likely conditioned on the input fingerprints.

Emission Score

The emission score $E(F, G)$ is intended to be proportional to the likelihood that a fingerprint F is observed from grid cell G . A larger emission score means that a cell is a more likely match for the observed fingerprint. *CTrack*'s emission score uses the following heuristic. We find F_c , the closest fingerprint to F seen in training data for G . “Closest” is defined to be the value of F_c that maximizes a pairwise emission score $E_P(F, F_c)$. The pairwise score is inspired by RADAR [69]. It captures both the number of matching towers, M , between two fingerprints, and the Euclidean distance d_R in between the signal strength vectors of the matching towers:

$$E_P(F_1, F_2) = M \lambda_{match} + (d_R^{max} - d_R(F_1, F_2)) \quad (4.1)$$

where λ_{match} is a weighting parameter and $d_R^{max} = 32$ is the maximum possible RSSI distance between two fingerprints. A higher number of matching towers and a lower value of d_R both correspond to a higher emission score. The maximum value of the pairwise emission score is then *normalized* as described below, and assigned as the emission score for F .

As an example, consider the fingerprints $\{(ID=1, RSSI=3), (ID=2, RSSI=5)\}$ and $\{(ID=1, RSSI=6), (ID=2, RSSI=4), (ID=3, RSSI=10)\}$. The distance between them would be $2\lambda_{match} + (32 - \frac{\sqrt{(3-6)^2 + (5-4)^2}}{2})$. The weighting parameter affects how much weight is given to tower matches versus signal-strength matches: we chose $\lambda_{match} = 3$.

CTrack normalizes all emission scores to the range $(0, 1)$ to ensure that they are in the same range as the transition scores, which we discuss next.

Transition Scores

The transition score in *CTrack* is given by:

$$T(G_1, G_2) = \begin{cases} \frac{1}{d(G_1, G_2)} & , G_1 \neq G_2 \\ 1 & , G_1 = G_2 \end{cases}$$

where $d(G_1, G_2)$ is the Manhattan distance between grid cells G_1 and G_2 represented as ordered pairs (x_1, y_1) and (x_2, y_2) . This transition score is based on the intuition that, between successive time instants, the user either stayed in the same grid cell or moved to an adjacent grid cell. It is unlikely that jumps between non-adjacent cells occur, but we permit them with a small probability to handle gaps in input data. When implementing *CTrack* on cellular data, it turns out to be important to permit such a *lax* transition score because cellular data often has abrupt “jumps” or “transitions” and does not immediately react to user movement — a given fingerprint is seen for some time, and then a sudden jump takes place to a new fingerprint.

Figure 4-5(b) shows the output of the grid sequencing step for our running example. As we can see, sequencing removes a significant amount of noise from the input data. In our evaluation, we demonstrate that the sequencing step is critical (Figure 4-12).

4.5.4 Smoothing and Interpolation

The next stage of *CTrack* takes a grid sequence as input and converts it into a sequence of (lat, lon) coordinates that are then processed by the *Segment Matching* stage. Note that this conversion does not significantly impact trajectory mapping accuracy, unlike the conversion used in “hard-decision” map-matching, where the fingerprints are converted to coordinates *before* doing the map-matching.

We describe the steps involved in smoothing and interpolation below.

Centroid Computation

For effective segment matching, we want to pick the most representative coordinate in a grid cell. If the grid cell has training points, we use the centroid of the training points seen inside the grid cell. If not, we output the geometric centre of the grid. We use this “centroid heuristic” because in many cases, all the training points seen in a grid lie on a particular road or sequence of roads. In these cases, the centroid has the advantage that it will also lie on a road segment and can subsequently be map-matched easily.

Smoothing Filter

Typically, centroids from grid sequencing have high frequency noise in the form of back-and-forth transitions between grids. Figure 4-5(b) illustrates this problem on our running example. To mitigate this, we apply a smoothing low-pass filter with a sliding window of size W_{smooth} to the centroids calculated above. The filter computes and returns the centroid of centroids in each window. This filter helps us to accurately determine the overall direction of movement and filter out the high frequency noise. We chose the filter window size, $W_{smooth} = 10$ seconds, empirically.

Interpolation

Earlier, we windowed the input trace and grouped cellular scans over a longer window of W_{scan} seconds. As a result, the smoothing filter produces only one point every W_{scan} seconds. We linearly interpolate these points to obtain points sampled at a 1-second interval, and pass them as input to the *Segment Matching* step described in Section 4.5.5.

The reason for interpolation is that segment matching produces a continuous trajectory where each segment is mapped to at least one input point. As in *VTrack*, the minimum frequency of input to the segment matcher is one that ensures that even the smallest segment has at least one point, which is approximately once a second or more, much as in *VTrack*.

Figure 4-5(c) shows the example drive after smoothing and interpolation. This output is free of back-and-forth transitions and correctly fixes the direction of travel at each time instant. Table 4.1 in our evaluation quantifies the benefit of smoothing.

4.5.5 Segment Matching

Segment Matching in *CTrack* maps sequenced, smoothed grids from the previous stages to road segments on a map. It takes as input the sequence of points from the *Smoothing and Interpolation* phase, and binary (0/1) turn and movement hints from inertial sensors on the mobile phone, to determine the most likely sequence of segments traversed. We describe how movement and turn hints are extracted in Section 4.6.

For segment matching, *CTrack* uses a version of *VTrack* that is augmented to incorporate movement and turn hints as follows:

- The HMM hidden states are the set of possible triplets $\{S, H_M, H_T\}$, where S is a road segment, $H_M \in \{0, 1\}$ is the current movement hint, and $H_T \in \{0, 1\}$ is the current turn hint.
- The emission score of a point (lat, lon, H_M, H_T) from a state (S, H'_M, H'_T) is zero if $H_M \neq H'_M$ or $H_T \neq H'_T$. Otherwise, we make it Gaussian just like in *VTrack*, proportional to e^{-D^2} , where D is the distance of (lat, lon) from road segment S .
- The transition score between two triplets $\{S^1, H_M^1, H_T^1\}$ and $\{S^2, H_M^2, H_T^2\}$ is defined as follows. It is 0 if segments S^1 and S^2 are not adjacent, disallowing a transition between them. This restriction ensures that the output of matching is a continuous trajectory. For all other cases, the base transition score is 1. We multiply this transition score with a *movement penalty*, $\lambda_{movement}(0 < \lambda_{movement} < 1)$, if $H_M^1 = H_M^2 = 0$ and $S_1 \neq S_2$, to penalize transitions to a different road when the device is not moving. We also multiply with a turn penalty, $\lambda_{turn}(0 < \lambda_{turn} < 1)$ if the transition represents a turn, but the sensor hints report no turn. We used $\lambda_{movement} = 0.1$ and $\lambda_{turn} = 0.1$. Our algorithm is not very sensitive to these values, since the penalties are multiplied together and a small enough value suffices to correct incorrect turn/movement patterns.
- Similar to *VTrack*, the HMM here also includes a *speed constraint* that disallows transitions out of a segment if sufficient time has not been spent on that segment. The maximum permitted speed can be calibrated depending on whether we are tracking a user on foot or in a vehicle.

The output of the segment matching stage is a set of segments, one per fingerprint in the interpolated trace. The output for the running example is shown in Figure 4-5(d).

When running online as part of the *CTrack* web service, the segment matcher takes turn hints and sequenced grids as input in each iteration and returns the current segment to an application querying the web service.

Running Time

As with *VTrack*, the run-time complexity of the entire *CTrack* algorithm, including all stages, is $O(mn)$, where m is the number of input fingerprints and n is the number of search states — which is the larger of the number of grid cells and road segments on the map. Our Java implementation of *CTrack* on a MacBook Pro with 2.33 GHz CPU and 3 GB RAM map-matched an hour-long trace in approximately two minutes, approximately 30 times faster than real time. It is straightforward to reduce the run time by more aggressively pruning the search space.

4.6 Sensor Hint Extraction

An important component of *CTrack* is the “sensor hint extraction layer” that processes raw phone accelerometer readings to infer information about whether the phone being tracked is moving or not, and processes orientation sensor readings from a compass or a gyroscope to heuristically infer vehicular turns. These hints are transmitted along with the GSM fingerprint to the server for map matching.

In this section, we provide some background on inertial sensors included on commodity smartphones, and what inertial sensor data on a smartphone represents. We then explain how *CTrack*

extracts useful hints from this data — as we shall see, this is a non-trivial problem. For more detail on how inertial sensors work, we refer readers to a more detailed explanation, with illustrations, in Chapter 5 of this dissertation (Section 5.3).

4.6.1 Inertial Sensors

Many modern smartphones including the latest generation of iPhone and Android phones are equipped with a 3-axis MEMS accelerometer and a 3-axis MEMS magnetometer (magnetic compass). Some are also equipped with a 3-axis gyroscope, such as the iPhone 4 and the Android Nexus S.

An accelerometer is capable of measuring a quantity proportional to the force experienced by the phone along three axes in space. The reported raw force value includes the effect of the earth's gravity. If we subtract out the gravity vector, the accelerometer measurement yields the actual acceleration experienced by the phone along three axes.

Similarly, a magnetometer uses measurements of magnetic field to estimate the position of magnetic north, which it can then correct to obtain the approximate orientation of the phone with respect to true (or magnetic) North. In our experiments with phones, we have found that the magnetometer works best at measuring absolute orientation of a phone when the plane of the phone is parallel to the earth's surface. However, as we shall see, the magnetometer is good at estimating *change in orientation* even if the plane of the phone is not strictly parallel to the earth's surface — it works across a range of orientations.

A magnetometer can be inaccurate when there is a significant amount of metal or electromagnetic interference in the vicinity of a phone. For this reason, newer phones include a 3-axis gyroscope that also measures orientation, but is robust to electromagnetic interference. A gyroscope does not directly measure orientation, but instead measures angular velocity around three axes: two in the plane of the phone and one perpendicular to this plane. This angular velocity can be integrated to obtain the instantaneous orientation of the phone at any instant. Chapter 5 describes this procedure in more detail.

4.6.2 Challenges

Our goal is to extract useful information from the inertial sensors to aid trajectory matching. The idea is that movement and direction of movement of a person/vehicle can be approximately inferred from inertial sensor data. There are two major challenges that must be overcome for this idea to work:

- The *movement of a phone does not always correspond to a movement of the user/vehicle carrying the phone.*

For example, consider a user driving a vehicle with her phone in her pocket. As long as the phone is in her pocket, the accelerometer on the phone accurately reflects the movement of the vehicle because the position of the phone is (approximately) fixed with respect to the vehicle. However, suppose the user now takes the phone out of her pocket to pick up a phone call. This movement will register on all the inertial sensors (accelerometer or compass/gyroscope), but will *not* reflect actual motion of the vehicle — it could be stopped at a traffic light for all we know.

- The *absolute orientation of a phone may not correspond to absolute orientation on a road map.* If a phone is in a user's pocket, for example, the readings from the magnetic compass

reflect only the phone’s orientation with respect to magnetic north, which has little to do with the vehicle or user’s direction of movement.

The first challenge is to accurately filter out inertial sensing data that does not come from times where the phone’s position with respect to the user/vehicle being tracked is fixed. We refer to this process as “anomaly detection”, and it is dealt with by a special component of *CTrack* that aims to remove these anomalies from raw sensor data.

The second challenge drives the design of how sensor hints are used in *CTrack*. *CTrack* only uses orientation values to measure the *change in orientation* or direction of movement of a user or vehicle, and *not* the absolute orientation. This works reasonably well even if the phone is in an arbitrary orientation, be it in a users’ pocket or handbag, or the dashboard or coffee holder of a car, as long as anomaly detection filters out situations where this relative orientation/position changes abruptly.

Below we describe anomaly detection.

4.6.3 Anomaly Detection

“Anomaly detection” filters out periods when the user is lifting the phone, speaking on the phone, texting, waving the phone about, or otherwise using the phone.

We have found empirically in our experiments with the iPhone and the Android G1 and Nexus One phones that when driving with the phone at rest in a vehicle or in a pocket, the raw accelerometer magnitude tends to be smaller than 14 ms^{-2} . Hence, anomaly detection looks for spikes in the raw accelerometer magnitude that exceed a threshold of 14 ms^{-2} . Whenever we encounter such a spike, we ignore all accelerometer and compass data in the map-matching algorithm until the phone comes back to a state of rest (this can be detected using standard deviation of acceleration, as explained below). On more recent phones such as the iPhone 4, the in-built gyroscope gives the exact orientation of the phone which can be directly read to determine if the phone is on a flat surface/in a user’s pocket.

Figure 4-6 illustrates anomaly detection real data. The driver had the phone in his pocket while driving until $t = 1767$; at this point, he took the phone out of his pocket, causing a sudden spike. The driver replaced the phone in his pocket a while later. We cannot immediately infer that it is safe to begin using sensor hints, but we do know that the phone came to a state of complete rest after $t = 1880$; at this point, we can start using the accelerometer and compass hints once again.

Having filtered out “anomalous periods”, the hint extraction layer proceeds to process the left over periods, which we call “stable periods”, to extract movement and turn hints, as explained in the sections below.

4.6.4 Movement Hint Extraction

The logical way to use raw accelerometer data to infer changes in position is to *integrate* it. However, this is challenging, as we explain below.

Recall that acceleration, being the second derivative of position, needs to be integrated twice to obtain displacement, or change in position. Integrating once yields the change in velocity over some time period, and then integrating the velocity again yields the displacement. Any measurement on an MEMS sensor is subject to noise. A simple, and reasonably accurate model of sensor error is to assume the raw acceleration values have a random zero-mean Gaussian error with a standard

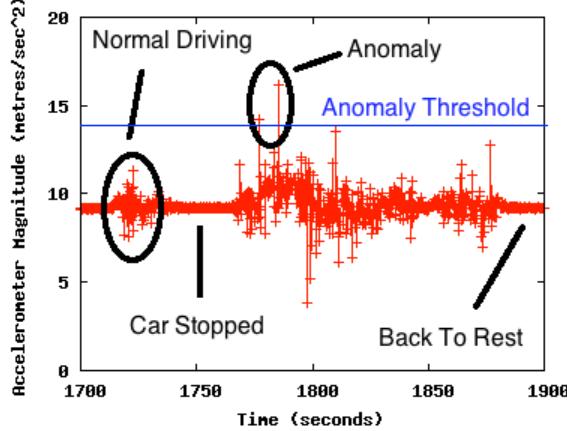


Figure 4-6: Anomaly detection in accelerometer data.

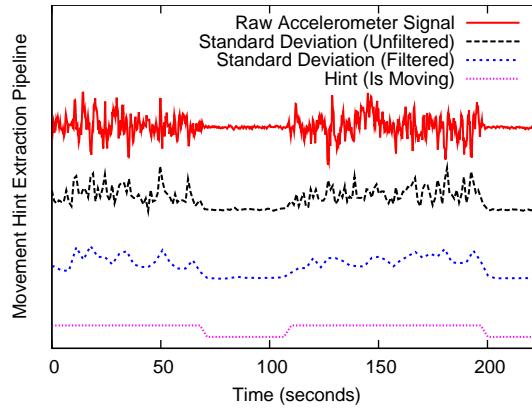


Figure 4-7: Movement hint extraction from the accelerometer.

deviation of N_{acc} . If this is the case, the mean *squared* error in velocity computed by integrating raw acceleration data will grow linearly with the number of observations integrated (or equivalently, linearly with time). This is analogous to the classic “drunkard’s random walk”, or Brownian motion in a straight line: if a person randomly moves 1 step backward or forward in each time step, the expected distance from the origin after N steps is proportional to \sqrt{N} .

The error in velocity is biased in one direction. Hence integrating the velocity with time causes the *drift*, or error in position to grow approximately as $O(t^{3/2})$, where t is time (or number of observations being integrated). This is a fast growth rate, and in practice quickly results in integration results from accelerometry being unusable.

Because accelerometer data is noisy and difficult to integrate accurately without accumulating significant error drift, *CTrack* instead extracts a simple “static” or “moving” (0/1 hint) from the accelerometer, rather than integrating the accelerometer data to compute velocities or processing it in a more complex way.

In contrast to integration, as we and other researchers have observed [75, 5] it is relatively easy to detect movement with an accelerometer: within a stable (spike-free) period, *the accelerometer shows a significantly higher variance while moving than when stationary*.

Accordingly, *CTrack* uses standard deviation to compute a boolean (true/false) movement hint for each time slot. We divide the data into one-second slots and compute the standard deviation of the 3-axis magnitude of the acceleration in each slot. Directly thresholding standard deviation sometimes results in spurious detections when the vehicle is static and the signal exhibits a short-lived outlier. To fix this, we apply an exponential weighted moving average (EWMA) filter to the stream of standard deviations to remove short-lived outliers. We then apply a threshold $\sigma_{movement}$, on the standard deviation to label each time slot as “static” or “moving”.

We used a subset of our driving data across multiple phones as training (where we know ground truth from GPS) to learn the optimal value of $\sigma_{movement}$. The best value turned out to be approximately 0.15 ms^{-2} , when the standard deviation is computed over a one-second window.

Figure 4-7 visually illustrates the steps of *CTrack*’s movement hint extraction algorithm on example accelerometer data from an Android G1 phone.

CTrack uses accelerometer data sampled at 20 Hz. A few (5-10) samples of accelerometer data are sufficient to detect movement using the technique described above. If it is sufficient to detect movement within a second, any sampling frequency higher than 5 Hz should work for movement hint extraction.

4.6.5 Turn Hint Extraction

As we have discussed earlier, the information provided by the orientation sensor on a smartphone depends on whether it comes from a magnetometer or gyroscope. A magnetic compass provides orientation of the plane of the phone with respect to geographic or magnetic north, while a gyroscope provides angular velocity about three axis which can be integrated to obtain accurate orientation of the phone with respect to a “world reference frame”. A magnetometer is less accurate than a gyroscope to obtain orientation because it often suffers significant electromagnetic interference, usually from nearby metallic objects.

However, at the time *CTrack* was originally implemented, the magnetometer was the only way to get orientation information on smartphones since no smartphone came with a gyroscope. This has since changed, with many newer smartphones including gyroscopes. For this reason, the rest of the discussion in this chapter is focused mainly on using magnetometer orientation information for turn hints. The basic algorithm we describe is applicable to gyroscopes as well, but because a gyroscope can provide more accurate information, it is likely that the accuracy of *CTrack* can be significantly improved if using a gyroscope, by extracting *much richer* turn hint information than is possible with a magnetometer owing to its accuracy limitations. We refer the interested reader to the *iTrack* algorithm for indoor trajectory mapping discussed in the next chapter, which extracts very rich sensor hint information from the phone gyroscope.

The orientation sensor APIs on most smartphone platforms provide orientation about three axes. We are most interested in the axis that provides the relative rotation of the phone about an axis parallel to gravity (called “yaw” in the iPhone 4 and Android conventions).

As mentioned earlier, because a phone can be in any orientation in a handbag or pocket, *CTrack* does *not* use absolute orientation information. The key observation used instead is that *irrespective of how the phone is situated*, a true change in orientation manifests as a *persistent, significant, and steep* change in the value of the orientation sensor.

With magnetic compass data, the main challenge is that the orientation reported is noisy owing to interference from nearby metal, or because the compass goes out of calibration for some other

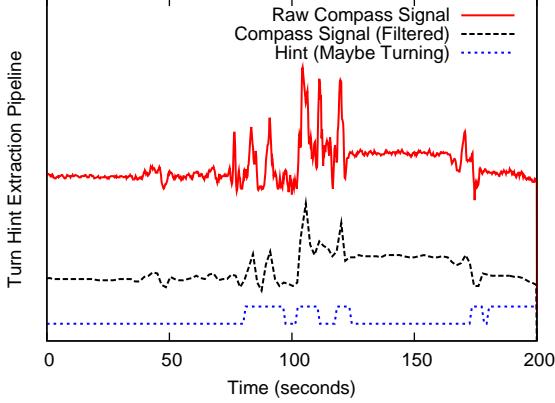


Figure 4-8: Turn hint extraction from the magnetic compass.

reason. *CTrack* solves this problem by applying a *median filter* with a 3-second window to the raw orientation values, which filters out non-persistent noise with considerable success. We note that the obvious approach of using a *mean*, or EWMA filter fails. A mean filter also successfully removes noise, but it tends to *blur* sharp transitions in the compass value, which is *exactly what we want to detect*.

After applying a median filter, *CTrack* finds transitions in the raw compass value with a magnitude exceeding at least 20 degrees and slope exceeding a minimum threshold (signifying a sharp change as opposed to a slow drift of the compass value). We fixed the slope threshold at 1.5 by experimentation.

Figure 4-8 illustrates a plot of the compass data with the sequence of processing steps required to generate a turn hint.

Limitation. We note that the above filtering approach is not perfect. In particular, it can (and often does) produce false positives. A true change in orientation can sometimes be produced by a phone sliding around within a pocket or a bag, or changing orientation for reasons other than the user or vehicle being tracked actually turning. However, we find that in practice, the filtering approach successfully eliminates *false negatives* — i.e., failing to detect a possible turn. For this reason, the segment matching HMM used in *CTrack* is designed specifically to penalize paths through the Viterbi decoder that take a turn when the reported sensor hint is “no turn”, but *not* vice-versa.

4.7 Evaluation

This section evaluates the *CTrack* algorithm. We show that the trajectory matches produced by *CTrack* on cellular (GSM) data are:

- Accurate enough to be useful for various tracking and positioning applications,
- Superior to sub-sampled GPS in terms of the accuracy-energy tradeoff, and
- Significantly better than using only “hard” information, by reducing each cellular fingerprint to the closest location before matching.

We investigate how much each of the four techniques used in *CTrack* — sequencing, windowing, smoothing, and sensor hints — contribute to the gains in trajectory mapping accuracy.

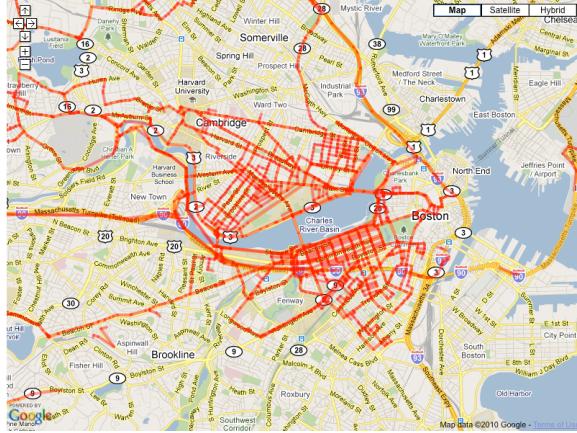


Figure 4-9: Coverage map of driving data set.

4.7.1 Method and Metrics

We evaluate *CTrack* on 126 hours of driving data in the Cambridge-Boston area, collected from 15 Android G1 phones and one Nexus One phone over a period of 4 months. We configured the *CTrack* phone library for the Android OS to continuously log the ground truth GPS location and the cell tower fingerprint every second, and the accelerometer and compass at 20 Hz. The data set covers 3,747 road segments, amounts to 1,718 km of driving, and 560 km of distinct road segments driven. The data set includes sightings of 857 distinct cell towers. Figure 4-9 shows a coverage map of the distinct road segments driven in our data set.

From 312 drives in all, we selected a subset of 53 drives verified manually to have high GPS accuracy as *test drives*, amounting to 109 distinct kilometres driven. We picked a limited subset as test drives to ensure each test drive was contained entirely within a small bounding box with dense training coverage. This is because evaluating the algorithm in areas of sparse coverage (which many of the other 259 drives venture into) could bias results in favour of *CTrack* by reducing the number of candidate paths to map-match to.

For each test drive, we perform *leave-one-out* evaluation of the map-matching algorithm: we train *CTrack* on all 311 drives excluding the test drive, and then map-match the test drive. We do this to ensure enough training data for each drive, and at the same time to keep the evaluation fair.

We compare *CTrack* to two other strategies in terms of energy and accuracy:

- *GPS k* gets one GPS sample every k minutes ($k = 2, 4$), interpolates, and map-matches it using *VTrack* (Chapter 3).
- *Placelab-VTrack* computes the best geographic location for each time instant using Placelab's fingerprinting technique [57], and matches the locations using *VTrack*.

We use three metrics in our evaluation of accuracy: *precision*, *recall*, and *geographic error*. The “precision” and “recall” we use are similar to the metrics used in the micro-benchmarks for *VTrack* in Chapter 3. They are similar to conventional precision and recall, but take the order of matched segments in the trajectory into account. We say that a subset of segments in a trajectory T_1 that also appears in trajectory T_2 are *aligned* if those segments appear in T_1 in the same order in which they

appear in T_2 . Given a ground truth sequence of segments G and an output sequence X to evaluate (produced by one of the algorithms), we run a dynamic program to find the maximum length of aligned segments between G and X . We define:

$$Precision = \frac{\text{Total length of aligned segments}}{\text{Total length of } X} \quad (4.2)$$

$$Recall = \frac{\text{Total length of aligned segments}}{\text{Total length of } G} \quad (4.3)$$

We estimate the ground truth sequencing of segments by map-matching GPS data sampled every second with *VTrack*, and manually fixing a few minor flaws in the results.

In one sense, for applications like traffic monitoring, we care more about precision than recall because we want to minimize incorrect output, but maintaining a minimum level of recall is important because it would be trivial (but useless) for an algorithm to output no segments and achieve perfect precision.

Geographic error. Precision and recall are relevant to applications that care about obtaining information at a segment-level, such as traffic monitoring. However, applications such as visualization or finding if the route a vehicle took was one of k pre-defined paths (e.g. [64]) do not need to know the exact road segments traversed, but may want to identify the broad contours of the route followed. In such applications, mistaking a road for a nearby parallel road may be acceptable in some cases.

To quantify this notion, we compute a third metric, *geographic error*, which captures the spatial distance between the ground truth and the matched output. We compute the maximum alignment between the ground truth trajectory G and output trajectory X using dynamic programming. This alignment matches each segment S of X to either the same segment S on G (if *CTrack* matched that segment correctly) or to a segment $S_{\text{wrong}} \in G$ (if matched incorrectly).

We define the *segment geographic error* to be the distance between S and S_{wrong} for incorrect segments, and 0 for correctly matched segments. The mean segment geographic error over all segments in X is the *overall geographic error*.

A small value of the geographic error implies that the algorithm finds a trajectory very close to the original trajectory in terms of distance. A small overall geographic error means that most segments are matched correctly (contributing zero error), or that incorrectly matched segments are close to corresponding segments on the ground truth (rather than completely off), or both.

4.7.2 Key Findings

The key findings of our evaluation are:

- *CTrack* on cellular data has 75% precision and 80% recall in both the mean and median, and a median geographic error of 44.7 meters. We discuss what these numbers mean in the context of real applications below.
- *CTrack* has $2.5 \times$ better precision and $3.5 \times$ smaller geographic error than *Placelab+VTrack*.

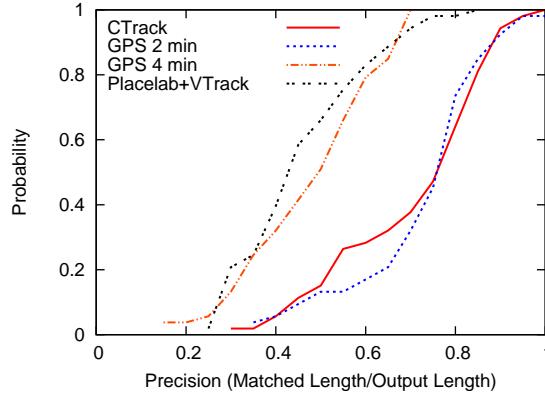


Figure 4-10: CDF of precision: *CTrack* is better than *Placelab + VTrack*.

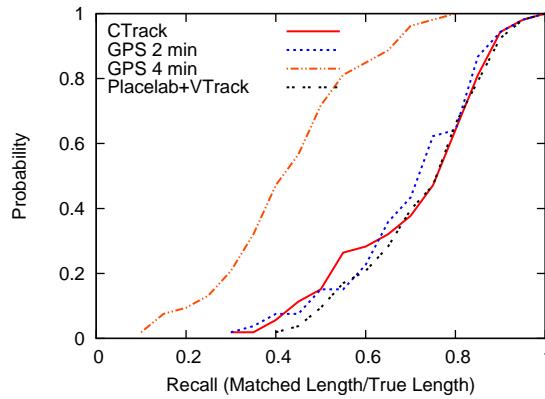


Figure 4-11: CDF of recall: comparison.

- *CTrack* is equivalent in precision to map-matching GPS sub-sampled every 2 minutes while consuming over $2.5 \times$ less energy. It also reduces error ($1 - \text{precision}$) by a factor of over $2 \times$ compared to sub-sampling GPS every 4 minutes, consuming a similar amount of energy. *CTrack* is $6 \times$ better than continuous WiFi sampling in terms of battery lifetime on the Android platform, with somewhat lower precision. Going by the “segment error rate” results from *VTrack*, WiFi has a precision of about 85%, 10% more than *CTrack*.
- The first step of *CTrack*, grid sequencing, is critical. Without grid sequencing, *CTrack* effectively reduces to computing a (lat, lon) estimate from the best fingerprint match, and ignoring a variety of location matches for that fingerprint. The median precision when not using sequencing is only 50%, similar to the accuracy of *VTrack* when run on cellular positioning data from Placelab. See Section 4.7.5 for more detail.
- We can extract movement and turn hints from raw sensor data with approximately 75% precision and recall. These hints improve accuracy by removing spurious “loops” and “turns” in the output. Using sensor hints improves the precision of trajectory mapping by 6% and the recall by 3%. See Section 4.7.6 for more detail.

4.7.3 Accuracy Results

The questions we aim to answer in this section are: is overall trajectory matching accuracy with cellular data and *CTrack* adequate for the kind of applications we are interested in? How does *CTrack* stack up against the other methods (in particular *VTrack* applied to cellular data, or to sub-sampled GPS data) in terms of accuracy and energy consumption?

Figure 4-10 shows a CDF of the map-matching precision for *CTrack*, *GPS k* (for $k = 2, 4$ minutes) and *Placelab+VTrack*. We highlight some important points from the figure:

- *CTrack* has a median precision of 75%, much higher than the both the energy-equivalent strategy of sub-sampling GPS every 4 minutes (48%), and *Placelab+VTrack* (42%).
- In effect, *CTrack* has over $2\times$ lower error ($1 - \text{precision}$) than sub-sampling GPS every 4 minutes, and over $2.5\times$ lower error than map-matching cellular localization estimates output by the *Placelab* method.
- Also, *CTrack* has equivalent precision to map-matching GPS sub-sampled every two minutes, while reducing energy consumption by approximately $2.5\times$ compared to this approach (Figure 2-1, Chapter 2).

Figure 4-11 shows a CDF of the recall. All the strategies except *GPS 4 min* are equivalent in terms of recall. Sub-sampling GPS every four minutes has poor recall (median only 41%) because a four-minute sampling interval misses significant turns in our input drives and finds the wrong path. The fact that *Placelab+VTrack* has identical recall shows that simple cellular localization and hard-decision map-matching do manage to recover a significant part of the input drive. However, converting cellular fingerprints to coordinates before map-matching them results in significant noise and long-lived outliers, and hence produces a large number of incorrect segments when map-matched directly, resulting in low precision.

4.7.4 What Does 75% Precision Mean?

To understand what 75% precision might mean in terms of a an actual application, we refer readers to the results from Chapter 3 on *VTrack*, which study the relationship between map-matching accuracy and the accuracy of two end-to-end applications: traffic delay monitoring and traffic hot-spot detection. In Chapter 3, we found that sub-sampled GPS and WiFi localization, which have median precision of the order of 85% (corresponding to a “Segment Error Rate” of 15% in the terminology used there) were usable for accurate traffic delay estimation. Our results for cellular (75%) are only somewhat worse, and while not directly comparable, they suggest a significant portion of travel time data from *CTrack* could be useful.

For applications such as route visualization, or those that aggregate statistics over paths (e.g., to compute histograms over which of k possible routes is taken), or those that simply show a user’s location on a map, getting most segments right with a low overall error is likely sufficient. The median geographic error when using *CTrack* is quite low—just 45 meters—suggesting *CTrack* would have sufficient accuracy for such applications. In contrast, the median geographic error of the *Place-lab+VTrack* approach is 156 meters, over $3.5\times$ worse than *CTrack*.

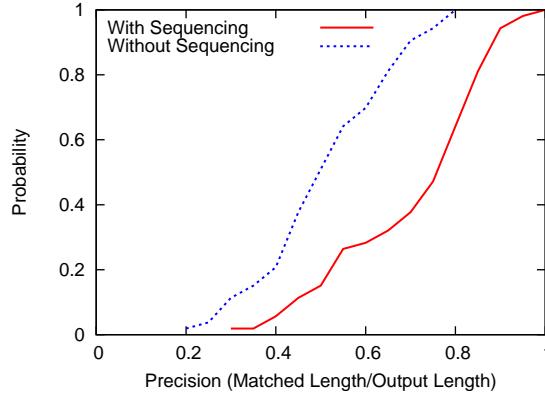


Figure 4-12: Precision with and without grid sequencing.

4.7.5 Benefit of Sequencing

This section elaborates on the key technical contribution of *CTrack*; the idea that the first pass of grid sequencing *before* converting fingerprints to geographic locations is crucial to achieving good matching accuracy. We provide experimental evidence supporting this idea. We also show that windowing and smoothing help improve matching accuracy, though to a lesser extent.

Figure 4-12 is a CDF that compares the precision of *CTrack* with and without the first pass of grid sequencing. This figure shows that sequencing is critical to achieving reasonable accuracy: without sequencing, the median precision drops from 75% to 50%. The reason is that running *CTrack* without sequencing amounts to reducing each fingerprint to its best match in the training database, and running *VTrack* on it.

Windowing and Smoothing Table 4.1 shows the precision and recall of *CTrack* with and without windowing and smoothing, two other heuristics used in *CTrack*. We see that each of these features improves the precision by approximately 10%, which is a noticeable amount. The recall does not improve because the algorithm without windowing/smoothing is good enough to identify most of the segments driven: the heuristics mainly help eliminate loops in the output, which arise from the back-and-forth grid transitions output by the first pass HMM.

	With		Without	
	Prec.	Recall	Prec.	Recall
Windowing	75.4%	80.3%	65.6%	82.3%
Smoothing	75.4%	80.3%	66.5%	82.5%

Table 4.1: Windowing and smoothing improve median trajectory matching precision.

4.7.6 Do Sensor Hints Help?

Figure 4-13 illustrates by example how turn hints extracted from the phone compass help in trajectory matching. Without using turn hints (Figure 4-13(a)), our algorithm finds the overall path quite accurately but includes several spurious turns and kinks, owing to errors in cellular localization. After including turn hints in the HMM, the false turns and kinks disappear (Figure 4-13(b)).

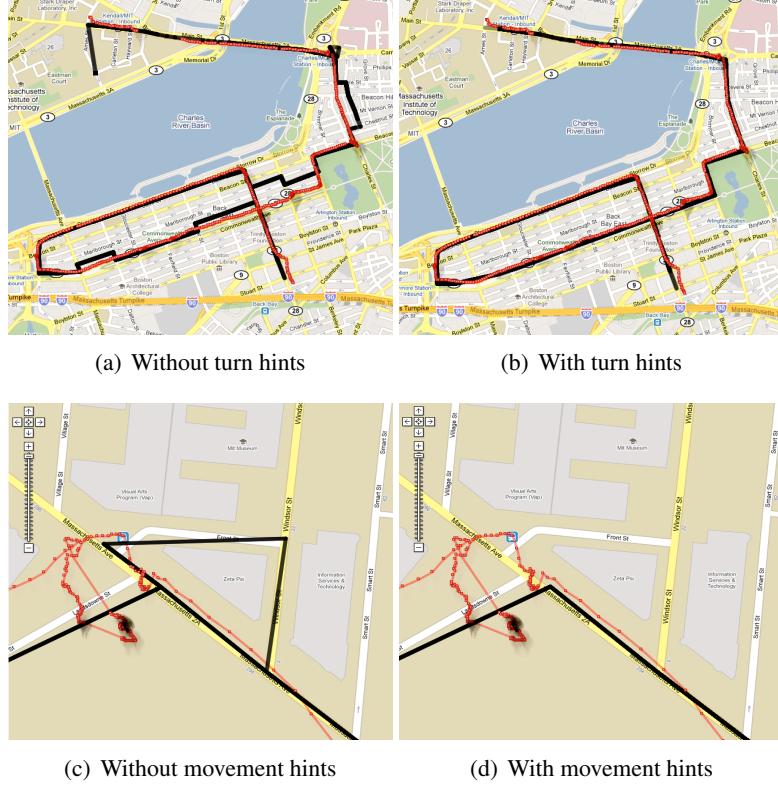


Figure 4-13: Sensor hints from the compass and accelerometer aid map-matching. Red points show ground truth and the black line is the matched trajectory.

In Figure 4-13(c), the driver stopped at a gas station to refuel, which can be seen from the cluster of ground-truth GPS points. Before using movement hints, errors from cellular localization were spread out, causing the map-matching to introduce a loop not present in the ground truth (Figure 4-13(c)). After incorporating movement hints, the speed constraint in *CTrack*'s HMM eliminates this loop because it detects that the car would not have had sufficient time to complete the loop (Figure 4-13(d)).

Limitation. We note a limitation of the movement hint. Stop detection works because the phone was placed on the dashboard: if it had been in the driver's pocket during refueling, the movement hints would not have helped had the driver gotten out of the car and been moving about, since that portion would have been filtered out.

Figure 4-14 is a CDF that compares the precision of *CTrack* with and without sensor hints (both movement and turn). This figure shows that sensor hints improve the median precision of matching by approximately 6%. While this may not seem huge, there exist several trajectories for which the hints do help significantly, suggesting that using them is a good idea when available. In our experience, the main benefit of the hints is in eliminating the several “kinks” and spurious turns in the matched trajectory, which the quantitative metrics don't adequately capture. We expect that more accurate turn hints, such as from a gyroscope (as opposed to a magnetometer) should yield greater benefit.

We now drill down into how accurately our approach is able to extract individual movement and turn

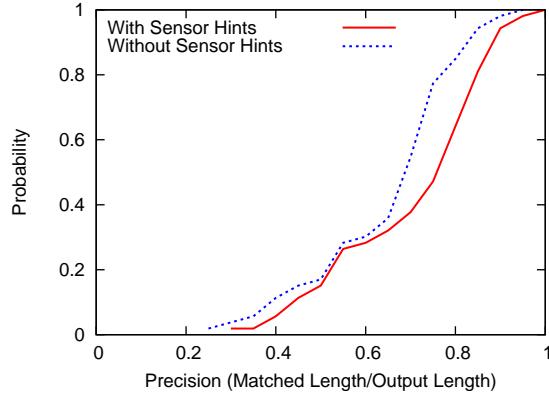


Figure 4-14: Precision with and without sensor hints.

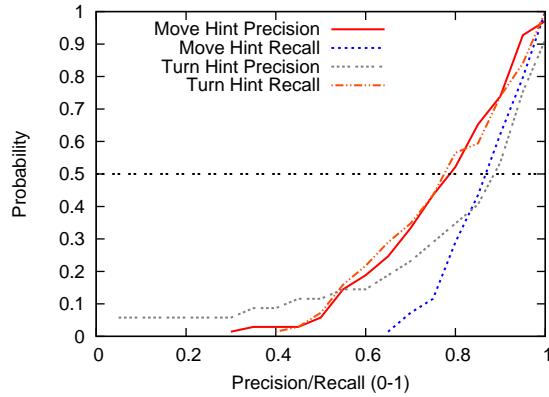


Figure 4-15: CDF showing precision/recall for hint extraction.

hints. We used the ground truth GPS for all of our test drives to compute the status of the device in each slot (i.e. moving or stopped, turning or not turning). We then computed precision and recall metrics, defined as follows. For motion hints, the precision is the fraction of stops we report that are actually stops, and the recall is the fraction of stopped time slots that we report as stopped. We compute the precision for stop detection because a movement hint of 0 is what impacts the transition score in our HMM (Section 4.5.5). Similarly, for turn hints, the precision is the fraction of “no-turn” zones we report that actually do not contain a turn, and the recall is the fraction of “no-turn” zones we detect as “no-turn” zones.

Figure 4-15 shows that the precision and recall of motion and turn hint extraction all exceed 75%. We do not claim that *CTrack*'s algorithms for hint extraction are the best possible; our broader point is to show reasonably accurate hint extraction is feasible and helps trajectory matching.

4.7.7 How Much Training Data?

An important aspect of deploying a cellular location system based on *CTrack* is that it requires effort to build a training database of cell towers by driving around the geographic area of interest. This section seeks to answer the question of *how much training data* is required to bootstrap a system like *CTrack*.

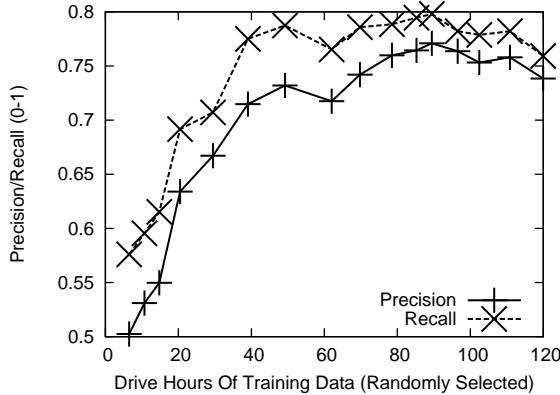


Figure 4-16: Precision/recall as a function of the amount of training data.

To quantify the amount of training data essential to achieving good trajectory mapping accuracy with *CTrack*, we picked a pool of test drives at random, amounting to 5% of our data set (8 hours of data), and designated the remaining 95% as the training pool. We picked subsets of the training pool of increasing size, i.e., first using fewer drives for training, then using more. In each run, the training subset was used to train *CTrack* and then evaluated on the test pool. Figure 4-16 shows the mean precision and recall of *CTrack* on the test pool as a function of the number of drive hours of training data used to train the system. The accuracy is poor for very small training pools, as expected, but encouragingly, it quickly increases as more training data is available. The algorithm performs almost as accurately with 40 hours of training data as with 120, suggesting that 40 hours of training is sufficient for our data set.

The 40-hour number, of course, is specific to the geographic area we covered in and around Boston, and to the test pool. To gain more general insight, we measure the *drive count* for each road segment in the test pool, defined as the number of times the segment is traversed by any drive in the training pool. Figure 4-17 shows the distribution of *test segment drive counts* corresponding to 40 hours of training data. While the mean drive count is approximately 3, this does not mean each road segment on the map needs to be driven thrice to achieve good accuracy. As the graph shows, about 60% of the test segments were not traversed even once in the training pool, but we can still map-match many of these segments correctly. The reason is that they lie in the same grid cell as some nearby segment that was driven in the training pool. This result is promising because it suggests that training does not have to cover every road segment on the map to achieve acceptable accuracy.

4.7.8 How Long Is Training Data Valid?

A related concern with both WiFi and cellular localization techniques that rely on building a training database is that as the set of access points or cell towers in an area changes, the training database slowly becomes invalid, requiring the training effort to be re-undertaken periodically (e.g., by driving through the area again).

We have not explicitly quantified the time for which training (wardriving) data remains valid. Our qualitative experience has been that it decays slowly and remains usable over months to years, at least for cellular data. For example, we have evaluated *CTrack* on driving data as recent as February 2011 using training data collected no later than July 2010, and some data collected as early as late 2009. We have found no marked degradation in accuracy.

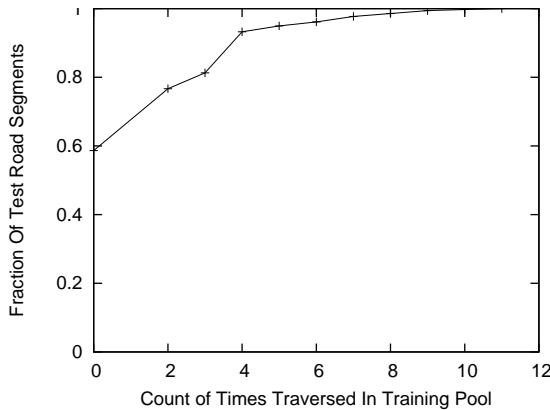


Figure 4-17: CDF of traversal counts for each road segment, with 40 hrs of training data.

4.7.9 A Confidence Metric?

For accuracy sensitive applications, a reasonable trade-off would be to use a *filtering* step to throw out tracks, or portions of tracks, where the *CTrack* algorithm has low confidence in its output. This has the effect of sacrificing some of the “recall” for substantially better “precision”. Recall the “bad zone detector” used in *VTrack*: can we develop something similar for *CTrack*?

We have spent some time investigating whether a confidence metric could be used to filter out drives on which *CTrack* does poorly. Overall, our results have been quite weak, unlike in the *VTrack* case. Our failure to find a confidence metric has led us to conjecture (though we do not have conclusive proof) that the problem of finding a good confidence metric for radio localization is as hard as localization itself. The rough intuition is that if it were possible for the algorithm to know exactly on which inputs it does poorly, it would not do as poorly on those inputs.

However, we have found two predictors that are *weakly* correlated with map-matching accuracy:

- The 90th percentile distance of smoothed grids from the segments they are matched to, *before* the second pass HMM used in *CTrack*.
- The mean difference (over all points P) in emission score between the segment that P is matched to in the output, and the segment closest to P .

The intuition in both cases is that a point far away from the road segment it is matched to, or closer to a different road segment, implies lower confidence in the match. When applying these confidence filters to our output drives, we currently improve the median precision from 75% to 86%, but lose substantially in terms of recall, whose median reduces from 80% to 35%. One interesting question for future work is whether machine learning techniques like *boosting* [81] could be used to combine these weak confidence predictors into a stronger predictor of confidence.

4.7.10 Comparison To WiFi Localization

As we have seen earlier, continuously sampling WiFi is much more energy-intensive than GSM (by over 6x on an Android G1 phone). Sub-sampling WiFi by duty-cycling the WiFi radio is one way to reduce this energy requirement, which makes sub-sampled WiFi an interesting alternative to using cellular localization with *CTrack*.

We ran a simple experiment to compare trajectory matching using *CTrack* on GSM fingerprints to trajectory matching with sub-sampled WiFi. Unfortunately, at the time of performing the experiments for *CTrack*, we did not have access to a training WiFi database. Hence, we only have access to “hard decision” WiFi information for this experiment. We used the Android network location API which scans for WiFi periodically and uses Google’s wardriving database to look up the location corresponding to a WiFi scan, as a proxy for WiFi localization.

In our experiment, we duty-cycled the WiFi radio manually every 90 seconds, and obtained a new “network location” fix each time the radio was switched on. We chose a value of 90 seconds based on a battery drain experiment similar to that in Section 2.2.4, which found that sub-sampling WiFi every 90 seconds is equivalent to *CTrack* in terms of battery life (the WiFi radio takes about 10 seconds on average to turn on and off, so this corresponds to a duty cycle of about 10% for WiFi). We also continuously sampled GPS for ground truth, and GSM for *CTrack*. We ran *CTrack* on the GSM data and applied *VTrack* [7] to map-match the WiFi data. While we do not have a large corpus of evaluation data for this experiment, Table 4.2 below reports the precision/recall results for three test drives in different areas of Boston, amounting to over an hour of driving data. The results are too small a sample set to be conclusive, but suggest that the approaches could be comparable in terms of accuracy.

We note that *CTrack* can also be applied to soft information from WiFi localization. We believe it is likely to improve accuracy over the hard decision *VTrack* approach, though we have not implemented or evaluated this.

CTrack		Sub-Sampled WiFi	
Prec.	Recall	Prec.	Recall
Drive 1	64.7%	68.6%	30.6%
Drive 2	57%	60.1%	42.8%
Drive 3	50.1%	65.8%	61.6%
			64.2%

Table 4.2: *CTrack* on cellular vs *VTrack* on 10% duty-cycled WiFi.

4.8 Related Work

Placelab performed a comprehensive study of GSM localization and used a fingerprinting scheme for cellular localization [57]. RADAR used a similar fingerprinting heuristic for indoor WiFi localizations [69], and the map-matching emission score used in *CTrack* is inspired by these methods. However, neither Placelab nor RADAR address the problem of trajectory matching, and are concerned more with the accuracy of individual localization estimates, rather than finding the optimal sequencing of estimates. As shown by the results in this dissertation, the use of soft information in the sequencing step is critical: applying a map-matching algorithm directly to Placelab-style location estimates results in significantly worse accuracy (by a factor of over 2×) compared to *CTrack*.

Letchner et al. [44] and our previous work on *VTrack*, presented earlier in this dissertation, also both use HMMs for map-matching. These algorithms do not use sensor hints. Moreover, as has been explained, these previous algorithms use and process (*lat*, *lon*) coordinates as input and use a Gaussian noise model for emissions, and are hence inaccurate for map-matching cellular fingerprints.

CompAcc [38] proposes to use smartphone compasses and accelerometers to find the best match for a walking trail by computing directional “path signatures” for these trails. They do not use cell towers. However, from our understanding, the paper uses absolute values of compass readings. This

approach did not work in our experiments, because the absolute orientation of a phone can be quite different depending on whether it is in a driver’s pocket, on a flat surface, or held in a person’s hand. For this reason, we chose to use boolean turn hints instead, which are more robust and can be accurately computed regardless of changes in the phone’s initial orientation or position.

For extracting motion hints and detecting walking and driving using the accelerometer, we use algorithms similar to those in [75, 5, 51].

Some previous papers [37, 45, 60] have proposed energy-efficient localization schemes that reduce reliance on continuously sampling GPS by using a more energy-efficient sensor, such as the accelerometer, to trigger sampling GPS. RAPS [45] also uses cell towers to “blacklist” areas where GPS accuracy is low and hence GPS should be switched off, to save energy. However, none of these papers address trajectory matching or propose a GPS-free, accurate solution for map-matching.

Skyhook [78] and Navizon [65] are two commercial providers for WiFi and Cellular localization, providing databases and APIs that allow programmers to submit WiFi access point(s) or cell tower(s) and look up the nearest location. However, to the best of our knowledge, they do not use any form of sequencing or map-matching, and focus on providing the best static localization estimate.

4.9 Conclusion

This chapter described *CTrack*, an algorithm that uses soft information to improve the accuracy of map-matching radio localization data.

Applying *CTrack* to cellular signal data yields an energy-efficient, GPS-free system for trajectory mapping using cellular tower fingerprints alone. On the Android platform, our *CTrack* implementation uses close to zero extra energy while achieving good mapping accuracy, making it a good way to distribute collaborative trajectory-based applications like traffic monitoring to a huge number of users without any associated energy consumption or battery drain concerns. A GPS-free approach to trajectory matching also opens up the possibility of providing more fine-grained location services on the world’s most popular, cheapest phones that do not have GPS, but that do have GSM connectivity.

CTrack shows the promise of using sensor hints from inertial sensors to improve localization in the absence of GPS, and shows how such hints can be integrated into a probabilistic model for trajectory mapping. While the impact of sensor hints in *CTrack* is relatively small, sensor hints prove to be much more powerful in the context of *indoor trajectory mapping*, which will be the subject of the next chapter of this dissertation. The *iTrack* indoor trajectory mapping system we present in Chapter 5 uses many of the sensor hint ideas from *CTrack*. *iTrack* uses a probabilistic model that *fuses* inertial sensor information with WiFi localization measurements indoors to reduce training effort and improve accuracy.

Chapter 5

Indoor Trajectory Mapping

Chapters 3 and 4 focused primarily on energy efficient trajectory mapping for mobile devices in an *outdoor* setting. There, the main technical challenge was the high energy consumption of GPS, which necessitated using lower-energy WiFi and cellular localization.

Indoors, the challenge is that GPS simply does not work. This is because the GPS signal-to-noise ratio is extremely low inside most buildings.

People spend most of their time indoors, creating a tremendous opportunity for fine-grained indoor positioning and tracking to enable an exciting range of new services and applications. Examples include locating missing objects at home, tracking or locating key personnel such as doctors in a hospital or employees in an office, navigation and search in large indoor spaces such as malls and museums, assisting visually impaired or disabled people navigating indoor spaces, and location analytics — understanding where people spend time and how they use an indoor space.

There exist techniques for accurate indoor localization to within a few centimetres that use *specialized hardware*, such as ultrasound and radio-ranging systems (Cricket [62] and Active Bat [3]), laser ranging (LIDAR) systems [56], and foot-mounted inertial sensors [29, 67]. However, these are often expensive to deploy and do not work on commodity mobile phones.

For this reason, we turn to WiFi localization [69, 58, 46, 70] and cellular localization [8], which we investigated in the outdoor context previously. These can both be used indoors to localize within a few metres, and hold promise because they work on commodity mobile phones. However, they face a significant deployment challenge, as we describe below.

5.1 The Training Challenge

The biggest challenge with indoor WiFi localization techniques today is the requirement of *extensive manual training* to associate radio fingerprints to ground truth locations in the building. Because there is no ready ground truth reference indoors (unlike GPS outdoors), the training procedure for WiFi localization is cumbersome. It requires dedicated personnel or volunteers to manually visit hundreds or even thousands of locations in a building, mark the location they are at accurately on a map, and collect training data at each location. This process is painful, expensive and time-consuming. [69, 58, 70, 46]. In our own experience, even just dragging the cursor on a smartphone touchscreen to the correct location, or panning a zoom window of a map to locate where to mark points is cumbersome to do for a few dozen points, let alone hundreds or thousands.

Manual annotation is additionally subject to human error which in turn can cause errors in localization. The OIL project at MIT [46], which relies on crowd-sourcing indoor location from volunteers, has documented many examples of such errors.

5.2 *iTrack* And Contributions

iTrack is a system we have built for fine-grained indoor trajectory mapping on mobile smartphones that *fuses* information from inertial sensors (accelerometer and gyroscope) with WiFi fingerprints.

iTrack makes two key contributions. First, *iTrack* can recover the trajectory of a mobile phone when held by a user walking steadily with phone in his/her hand or pants pocket highly accurately, to within *less than a metre*. To do this, it uses inertial sensors to detect periods of steady walking, and to extract the approximate shape and size of each user walk from the gyroscope and accelerometer respectively (we define “shape” and “size” precisely later). Each walk is matched to the contours of a building floorplan using a *particle filter*. A particle filter uses Monte Carlo simulation to explore a large number of paths with shape and size close to the measured shape and size extracted from the sensor data. The filter narrows down a candidate path by eliminating particles (paths) that cross a wall or obstacle in the floorplan. The filter also progressively fuses in WiFi signal strength information as it becomes available to improve the quality of tracks it extracts.

Second, *iTrack* provides a novel, simplified approach to training a WiFi localization system in an indoor space by simply *walking around the space to map it*, significantly reducing human effort compared to the manual annotation approach. Training proceeds in two stages:

- The first stage initializes the algorithm with “seed WiFi data” from a small number of walks where a volunteer or dedicated trainer specifies his/her starting position approximately on a map. Unlike full-fledged manual training, the trainer only specifies one point per walk on the map, and only needs to specify it approximately. About 5-10 walks (amounting to only 10 minutes of training) suffice to seed an entire floor of the MIT CSAIL building, which is approximately 2,500 square metres in size.
- Once seed data has been collected, *iTrack* crowd-sources more data from users of the indoor space. Users can contribute more training data by simply downloading an application on to their smartphones and running it in the background. This process is completely automated and can rapidly collect a large amount of data as users walk around. It requires *no extra effort or annotation from the user* beyond running the application on their phone.

Our training procedure is a huge improvement over previous techniques requiring point by point manual input from a human to indicate where he/she is, or requiring multiple devices to be carefully measured and placed on a dense grid covering the floorplan to collect training data [69, 58].

5.2.1 Implementation And Evaluation

We have implemented *iTrack* on the iPhone 4, which includes a 3-axis gyroscope. We have also tested the algorithms used by *iTrack* on gyroscope data from an Android Nexus S phone, but have not yet built a full-fledged Android application.

We evaluate *iTrack* on 50 test walks collected using an iPhone 4 on the 9th floor of the MIT Stata Center. We use tape markings made on the ground to accurately estimate ground truth for each

walk. We find that given only 4 walks of seed training data (which took ≈ 5 minutes to collect), *iTrack* can extract accurate trajectories from 80% (40 out of 50) of the walks. The walks extracted have a median error of approximately 3.1 feet, or less than a metre.

We also found that WiFi localization data learned from 97 walks crowd-sourced with *iTrack* could localize a user to within 5.4 metres on average on test data, comparable to the expected accuracy from manual training approaches.

5.2.2 Rest of this chapter

The rest of this chapter is organized as follows. Section 5.3 provides background on inertial phone sensors and the raw data they provide. In particular, we illustrate raw data collected from the accelerometer and gyroscope on the iPhone 4 and explain what the values mean. Section 5.4 provides an overview of traditional inertial navigation. This explains how to integrate accelerometer and gyroscope data to obtain a precise estimate of a device's current position and orientation. Section 5.4.2 explains why the conventional approach does not apply directly to mobile phones. We also explain why modifications to the conventional approach used to deal with pedestrian foot-mounted sensors do not work for mobile phones.

Section 5.5 describes the *iTrack* algorithm, and each stage involved extracting an accurate indoor trajectory from inertial sensing data. Section 5.6 describes how the tracks extracted from *iTrack* can be used for rapidly learning a WiFi training database for indoor localization, while being iteratively used to improve *iTrack*'s own estimates. Section 5.7 describes our iPhone implementation. Section 5.8 evaluates *iTrack*. Section 5.9 describes related work, and Section 5.10 concludes this chapter.

This chapter focuses mainly on strategies to improve the *accuracy* of indoor trajectory mapping. We do not investigate energy consumption, which would be an interesting direction for future exploration.

5.3 Inertial Phone Sensors

5.3.1 Accelerometer

Smartphones in the market as of 2011 are almost all equipped with a three-axis accelerometer, which can measure the effective acceleration along three axes in space. These accelerometers use MEMS (micro-electro-mechanical sensing) technology. An accelerometer cannot measure free-fall acceleration, but rather measures the *weight* experienced by a test mass lying in the frame of reference of the accelerometer device. For example, a phone held on a user's hand will measure an acceleration of g , the acceleration due to gravity because this is proportional to the weight experienced by the accelerometer inside.

An MEMS accelerometer in a phone today typically consists of a micro-scale cantilever beam with a *proof mass* suspended from it. Under the influence of external acceleration, the proof mass deviates from its normal position. The deviation can be measured to obtain an estimate of the net external acceleration applied to the accelerometer casing, along one particular axis. Modern phone accelerometers combine two such devices in one plane with one device out of plane to measure acceleration along three axes. The iPhone, for example, uses an LIS302DL three-axis MEMS accelerometer.

Figure 5-1 shows the data provided by the iPhone accelerometer, as documented in Apple's Core Motion API [23]. The device measures the acceleration a_x , a_y and a_z along three axes defined

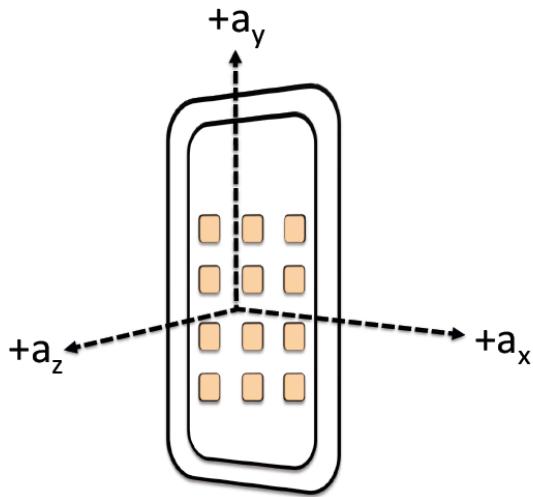


Figure 5-1: What a smartphone accelerometer measures.

with respect to the phone’s frame of reference. In the figure, a_z is the acceleration along an axis perpendicular to the plane of the phone, and a_x and a_y are accelerations along axes within the plane of the phone. The net acceleration due to gravity is *included* in the acceleration vector. Its components along the x, y and z axes depend on the absolute orientation of the phone, which cannot be determined using the accelerometer alone.

A typical smartphone accelerometer can sample accelerometer data at up to 50-100 Hz. Our iPhone implementation samples accelerometer data at 100 Hz.

5.3.2 Gyroscope

A gyroscope measures the angular velocity of rotation of an object about three axes in space. Modern MEMS gyroscopes also consist of a proof mass similar to MEMS accelerometers. Instead of being suspended from a cantilever as in the case of an accelerometer, the mass is made to vibrate continuously by applying a drive signal to a set of drive capacitor plates. When a user rotates the phone, the vibrating mass is displaced along all three axes by *Coriolis* forces. The MEMS gyroscope measures the deviation of the proof mass along all three axes to obtain an accurate estimate of the angular velocity of rotation.

Figure 5-2 shows the data provided by the iPhone gyroscope, as documented in Apple’s Core Motion API. The device measures the angular velocity ω_x , ω_y and ω_z about three axes defined with respect to the phone’s frame of reference. In the figure, ω_z is the angular velocity of rotation about an axis perpendicular to the plane of the phone, i.e., the angular velocity of rotation of the phone in the X-Y plane. Similarly, ω_x and ω_y represent the angular velocity of the phone’s rotation in the Y-Z and Z-X planes respectively.

A typical phone gyroscope can sample gyroscope data at 100 Hz or more. Our implementation of *iTrack* uses 100 Hz gyroscope data.

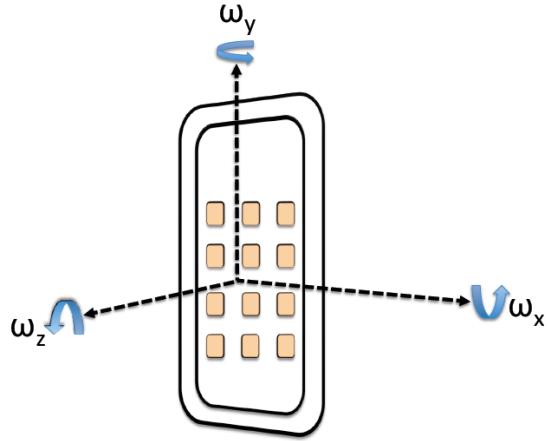


Figure 5-2: What a smartphone gyroscope measures.

5.4 Inertial Navigation

Conventional inertial navigation systems *combine* data from an accelerometer and a gyroscope to estimate the absolute position of a device, assuming its initial position, velocity and orientation are all known. These systems are in widespread use in aerospace systems, for tracking satellites (e.g., they are used to keep track of GPS satellites as they orbit the earth), industrial tracking applications and in robotics.

This section describes how inertial navigation works at a high level. We note from our discussion in the previous section that a gyroscope does not directly measure the absolute orientation of a phone. Rather, it measures angular velocity, which represents the rate of change of orientation. Given a known initial orientation, it is possible to integrate angular velocity measurements from a gyroscope to obtain absolute orientation at any time instant (described in Section 5.4.1, *Attitude Computation*).

By combining this orientation data with acceleration data from the accelerometer, it is possible to subtract out gravity, and indirectly compute the instantaneous *user acceleration* of the phone along three axes fixed in space. This in turn can be integrated twice to obtain an estimate of the user's position (described in Section 5.4.1, *Position Computation*).

The following section derives the equations used by inertial navigation systems to compute the position and orientation using integration.

5.4.1 Inertial Navigation Equations

The derivation presented in this section is based on material in Chapter 3 of *Strapdown Inertial Navigation Technology*, by David H. Titterton and John L. Weston [26].

We consider the problem of tracking the position and orientation of a device over time, given a sequence of instantaneous acceleration and angular velocity measurements of the form:

- $t, a_x(t), a_y(t), a_z(t)$ from the accelerometer.
- $t, \omega_x(t), \omega_y(t), \omega_z(t)$ from the gyroscope.

All the instantaneous measurements are with respect to the phone's frame of reference at any time instant t . Since the phone is continuously rotating, its frame of reference is also continuously changing. This means that the acceleration and angular velocity measurements need to be first converted to a fixed, or *inertial* reference frame before integrating them. The goal of integration is to find the absolute orientation, or *attitude* of the phone at any instant of time.

Attitude Computation

Assume a fixed (i.e., non-rotating) reference frame with three axes X, Y and Z in which the gravity vector points in the negative Z direction, and the X and Y axes are arbitrarily chosen. To represent the 3-D orientation of the phone at any instant of time, we use a representation called the *direction cosine* representation. The direction cosine matrix of the phone at time t , denoted by $C_{phone}(t)$, is a 3×3 matrix. The columns of this matrix represent unit vectors along the X, Y, and Z axes of the phone (as shown in Figure 5-1) as expressed in the fixed frame:

$$C_{phone}(t) = \begin{vmatrix} c_{11}(t) & c_{12}(t) & c_{13}(t) \\ c_{21}(t) & c_{22}(t) & c_{23}(t) \\ c_{31}(t) & c_{32}(t) & c_{33}(t) \end{vmatrix} \quad (5.1)$$

In the above matrix, the column vector $c_{11}(t)\hat{i} + c_{21}(t)\hat{j} + c_{31}(t)\hat{k}$ represents the unit vector along the phone's X axis (an axis parallel to the base of the phone, in the plane of the phone).

Given that the direction cosine matrix of the phone is $C_{phone}(t)$ at time t , any vector $\vec{v}_{phone}(t)$ at time t defined with respect to the phone's reference frame, can be transformed to the ground reference frame by pre-multiplying it with the direction cosine matrix:

$$\vec{v}_{ground}(t) = C_{phone}(t) \vec{v}_{phone}(t) \quad (5.2)$$

Our goal is to relate the instantaneous angular velocity from the gyroscope, $\vec{\omega}_{phone}(t)$ to the change in the direction cosine matrix. To do this, we write out an expression that yields the rate of change of the direction cosine matrix with time:

$$\dot{C}_{phone}(t) = \lim_{\delta t \rightarrow 0} \frac{C_{phone}(t + \delta t) - C_{phone}(t)}{\delta t} \quad (5.3)$$

The term $C_{phone}(t + \delta t)$ can be written as the product of two direction cosine matrices as follows:

$$C_{phone}(t + \delta t) = C_{phone}(t) A(t) \quad (5.4)$$

where $A(t)$ is a direction cosine matrix representing the (small) rotation of the phone's reference frame from time t to time $t + \delta t$.

For small times δt and hence small angles of rotation, the rotation $A(t)$ can be approximated as follows (for a proof, see [26]):

$$A(t) = [I + \delta t \Omega_{phone}(t)] \quad (5.5)$$

where I is the identity matrix, and $\Omega_{phone}(t)$ is the *skew symmetric matrix* form of the angular velocity vector $\vec{\omega}(t)$:

$$\Omega_{phone}(t) = \begin{vmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{vmatrix} \quad (5.6)$$

Substituting the expression for $A(t)$ above in the limit equation yields:

$$\dot{C}_{phone}(t) = C_{phone}(t) \Omega_{phone}(t) \quad (5.7)$$

where $\Omega_{phone}(t)$ is the skew-symmetric matrix representation of the angular velocity vector $\vec{\omega}(t)$.

This differential equation can be solved numerically to yield the instantaneous value of the direction cosine matrix, $C_{phone}(t)$ at any time instant t , and hence the instantaneous orientation of the phone at any time instant.

Position Computation

Given the instantaneous attitude direction cosine matrix $C_{phone}(t)$, we can use the following equation, analogous to equation 5.2 presented earlier to transform the phone accelerometer measurements to the ground reference frame:

$$\vec{a}_{ground}(t) = C_{phone}(t) \vec{a}_{phone}(t) \quad (5.8)$$

The effective user acceleration can then be computed by subtracting the effect of the earth's gravity, \vec{g} , from the acceleration vector expressed in the ground reference frame:

$$\vec{a}_{user}(t) = \vec{a}_{ground}(t) - \vec{g} \quad (5.9)$$

Suppose the phone has a known initial position $\vec{x}(0)$ and a known initial velocity $\vec{v}(0)$. The numerical integration equations to find velocity and position at each time instant are as follows:

$$\vec{v}(t + \delta t) = \vec{v}(t) + \vec{a}_{user}(t) \delta t \quad (5.10)$$

$$\vec{x}(t + \delta t) = \vec{x}(t) + \vec{v}(t) \delta t \quad (5.11)$$

In principle, one could use a numerical integration technique such as Simpsons' rule, or Runge-Kutta integration, to solve these equations and obtain the position of the mobile phone $\vec{x}(t)$ at any time instant t . However, as we shall see in the next section, this does not work in practice.

5.4.2 Challenge: Inertial Drift

Given a phone with a three-axis accelerometer and gyroscope, and a known initial position and velocity — for example, from a known time instant when the phone was static, it should in theory

be possible to directly solve the inertial navigation equations stated above to obtain the precise position of the phone at any time instant.

However, in practice, as with any real-world sensor, the acceleration and angular velocity measurements from inertial MEMS sensors are not perfectly accurate. As we shall see shortly, the measurement error is exacerbated when sensor data is used as input to higher-order integration, which is the case in the inertial navigation equations described above. This results in a problem known as *drift* in inertial navigation parlance. Drift is a phenomenon where the estimated trajectory of a device deviates from the true trajectory with a deviation that increases with time, often rapidly or non-linearly.

Both kinds of inertial measurements can experience both random noise and systematic bias, as well as more complex kinds of bias such as dynamically evolving bias. In this work, we focus on two types of errors: systematic errors and random noise. We describe the impact of both kinds of errors below.

Systematic Bias

The systematic bias of a sensor is a systematic error that is added to each measurement made by the sensor, and *always* has the same sign (positive or negative). It is easy to correct for because if the value of the bias is known, it can simply be subtracted from each sensor measurement.

It is easy to measure and correct for the systematic bias of both MEMS accelerometers and gyroscopes. A phone that is static — for example, placed stationary on a flat desk, should have a user acceleration vector that is zero. The long-term average of the measured user acceleration from the accelerometer during such a static period, after subtracting gravity, yields the bias of the accelerometer along all three axes.

Similarly, a phone at rest should have zero angular velocity. The long-term average of the angular velocity measured by the gyroscope when the phone is static yields the gyroscope bias.

Random Noise

Random noise is usually thermal in nature and does *not* always have the same sign. It is therefore harder to correct than systematic bias. Assuming all systematic bias has been corrected, the remaining error comes from an error distribution that has zero mean but a non-zero variance.

Recall the discussion in Section 4.6.4 of Chapter 4. Assuming raw acceleration values have a random zero-mean Gaussian distributed error with a standard deviation of N_{acc} , we showed in that discussion that the error in position, obtained by twice integrating the acceleration, grows approximately as $O(t^{3/2})$ with time t . In the inertial navigation case, a key point to note is that the instantaneous acceleration vector is obtained indirectly by transforming *measured* acceleration using the current orientation of the phone. The current orientation *itself is subject to error from integrating the raw gyroscope data*.

If the raw angular velocity $\vec{\omega}$ measured by the gyroscope is assumed to have zero-mean Gaussian noise with standard deviation N_{gyro} , the change in the direction cosine matrix at any time instant will also have Gaussian distributed error. Analogous to the walking drunkard discussed in Section 4.6.4, the orientation obtained by integrating (summing) gyroscope data will have mean squared error proportional to t , the time period of integration. This in turn means the transformed acceleration values from equation 5.8 have zero-mean error distributed with a variance proportional to \sqrt{t} .

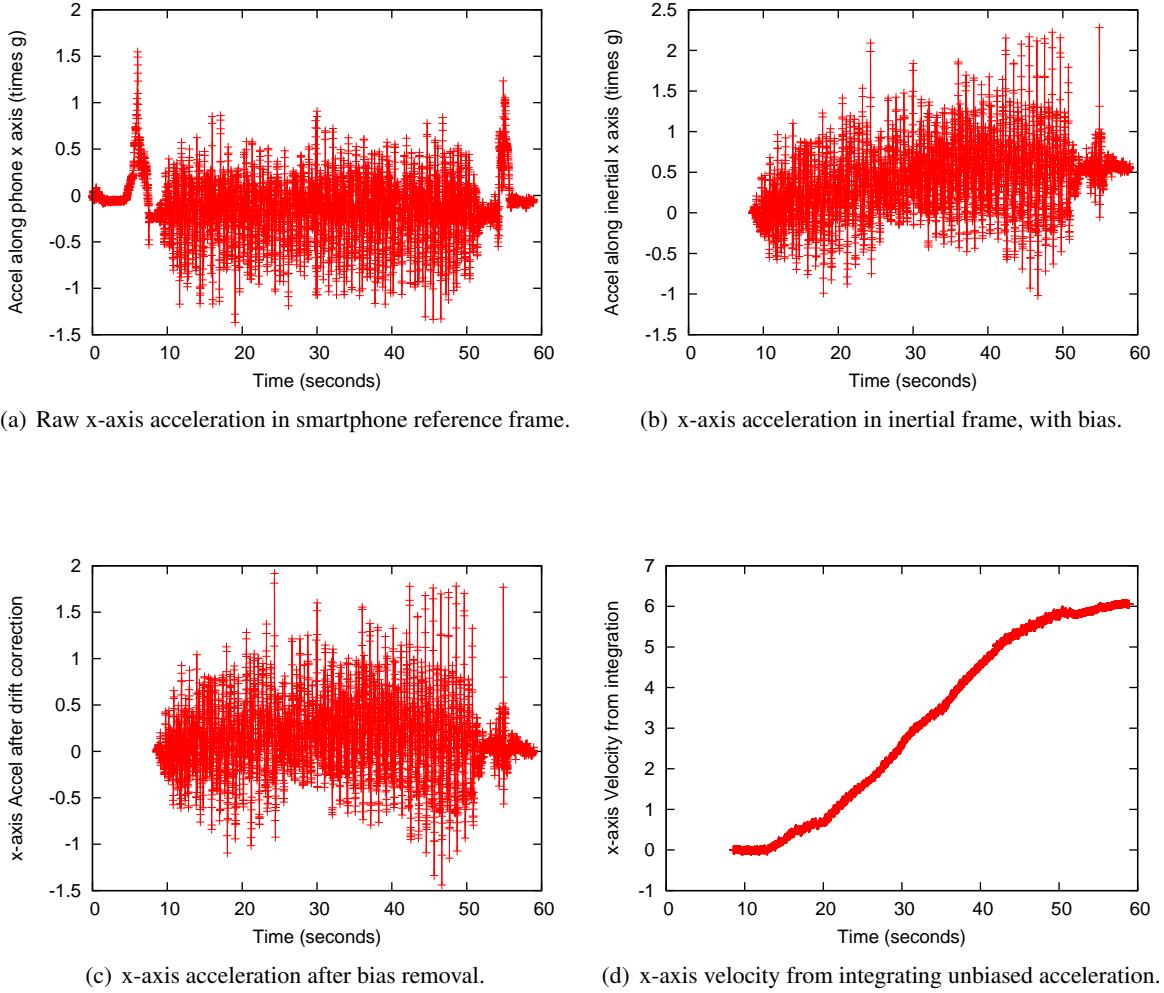


Figure 5-3: Drift when integrating smartphone accelerometer data.

If integrating the acceleration twice to obtain position, the expected error in position can be shown to grow as $O(t^{2.5})$, *even after correcting for systematic bias*. If we cannot correct for systematic bias, the growth rate of error is even higher, of the order of $O(t^3)$.

5.4.3 Correcting For Drift

Sensor drift is an extremely serious problem for most inertial navigation systems. In each application domain, inertial systems have evolved specialized tricks and techniques to periodically correct for drift and halt the growth of error in the estimated position. For example, GPS satellites periodically use knowledge about the direction of the earth's gravity vector to correct accumulated error in their direction cosine matrix, and thereby arrest some of the drift.

In the context of indoor localization, $O(t^{2.5})$ or $O(t^3)$ is too high a growth rate for position estimates from direct integration to be usable. Simply integrating raw accelerometer data results in estimates with so high an error that they are unusable for tracking a mobile device indoors.

Illustration

Figure 5-3 illustrates drift on real accelerometer data collected from an iPhone in a user’s pocket when walking indoors. The first pane, Figure 5-3(a) shows the measured raw value of acceleration along the X-axis. This value is always with respect to the phone’s reference frame. Since the phone is continuously changing orientation in the user’s pants pocket, this acceleration needs to be transformed to a global inertial frame using orientation from the gyroscope, as we have discussed earlier. In this particular trace, the user started in a stationary position and immediately started walking. The walking ends towards the very end of the trace.

The transformed acceleration obtained from equation 5.8 is shown in the second pane, Figure 5-3(b). As we can see, the transformed acceleration is already drifting upwards steadily and suffers significant bias. A big part of this drift is due to drift in the absolute device orientation computed by integrating the gyroscope using the equations described earlier. Even a small error in the computed direction cosine matrix, $C_{phone}(t)$, is exacerbated because the error in the matrix is multiplied by \vec{g} , the acceleration due to gravity, when transforming the acceleration to the ground frame using equation 5.8.

It is possible to correct this accumulated drift in acceleration somewhat by explicitly considering each walking period separately and factoring out a linearly growing drift error term, $d = \alpha t$ for some constant α . Figure 5-3(c) shows the result of such a linear drift correction term applied to the transformed acceleration, for the best possible value of the constant α .

However, even integrating this acceleration results in velocity drift owing to random noise. Figure 5-3(d) shows the estimated velocity of the mobile device along the X-axis obtained by numerically integrating the “unbiased” acceleration from Figure 5-3(c). As we see, the velocity quickly drifts upwards. The velocity drift is proportional to $O(\sqrt{t})$ because it comes from zero-mean random noise in the acceleration, as we discussed earlier. The velocities output by this procedure are themselves not accurate enough to be useful. A further level of integration to find displacement would produce even larger errors that grow faster, and would clearly be unusable for trajectory mapping.

Researchers have previously looked into the drift problem in the context of pedestrian indoor tracking using foot-mounted inertial sensors [67, 29]. These approaches deal with the problem of drift by using a technique called *zero-velocity updates* (ZUPTs). The idea is that the inertial sensor mounted on a person’s foot experiences zero velocity and zero angular velocity whenever the person’s foot is on the ground during his/her stride.

Graphs of angular velocities measured by such a foot-mounted gyroscope show a flat “plateau” region with zero or close to zero angular velocity, which can be easily detected and classified as a step event. Accordingly, the velocity estimated from inertial navigation can be corrected to zero at each such point, once for each stride. This correction can be back-propagated to correct the orientation matrix using a Kalman filter, as discussed in [67]. The authors of that work show that using this technique can reduce the growth in position drift to $O(t)$, and can accurately estimate the length of each individual stride.

Unfortunately, these existing approaches do not work for mobile phone tracking, because a phone is not mounted on a user’s foot. A phone is typically held in a user’s hand or pants pocket, or handbag, where it has some wiggle room. Hence, it never experiences zero angular velocity or zero velocity, or even velocity close to zero. In fact, integration drift error is compounded when inertial data comes from a mobile phone that can be swinging up and down in a users’ pants pocket or handbag, or from a phone oscillating gently up and down or sideways in a user’s hand when walking.

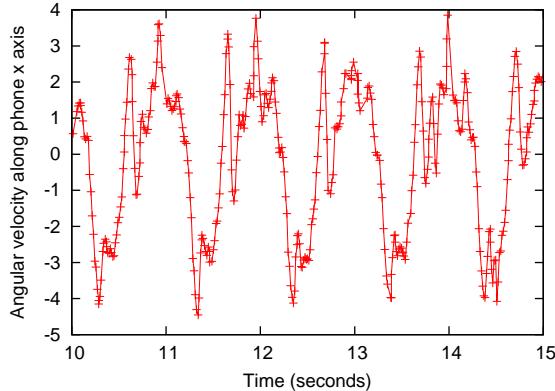


Figure 5-4: Angular velocity of an iPhone in a user’s pocket when walking.

Figure 5-4 shows raw angular velocity data from the iPhone gyroscope from the same trace used for illustration in Figure 5-3. As we see, the angular velocity does not experience any flat region close to zero. Rather, it varies continuously during all parts of the user’s stride.

Approach Used in *iTrack*

A key novel contribution of *iTrack* is how it deals with and corrects inertial sensor drift on a mobile phone. *iTrack* does not integrate acceleration data outright. *iTrack* instead uses acceleration data to specifically detect *steady walking*, and distinguish it from other movements such as one time movement, talking, picking up the phone and other phone use. Wherever it identifies a steady walking period, *iTrack* uses the accelerometer data to perform *step-counting* in the user’s walking periods, which, as we show in the next section, can be done more accurately than integrating noisy acceleration samples.

In contrast to the approach used for acceleration data, *iTrack* does integrate data from the gyroscope to estimate the approximate shape of a user’s walk. This does not suffer drift error because *iTrack* uses only *changes* in the gyroscope values rather than the absolute orientation information to determine a phone’s trajectory.

The shape and size of the walking trajectory (estimated from step count and gyroscope data) are used as a guide to find the best match to a known building floorplan.

5.5 The *iTrack* System

Figure 5-5 shows the architecture of *iTrack*. The system consists of a phone application and a server-side trajectory matching service. The phone application continuously samples WiFi fingerprints, gyroscope and accelerometer data at the maximum possible frequency (100 Hz for gyro and accelerometer on the iPhone). The accelerometer signal is used in conjunction with the gyroscope to detect periods of steady walking, using a *walking detector*. All the sensor data in walking periods is transmitted to the *iTrack* server, which matches it to a known building floorplan to find user trajectories. Currently, trajectory mapping is implemented as an offline process on the server. It would be interesting to explore performing trajectory mapping on the phone itself as part of future work.

In addition to sampling and sending sensor data back to the server for trajectory mapping, *iTrack* also uses observed WiFi fingerprints to continuously localize the phone. The *localization engine*

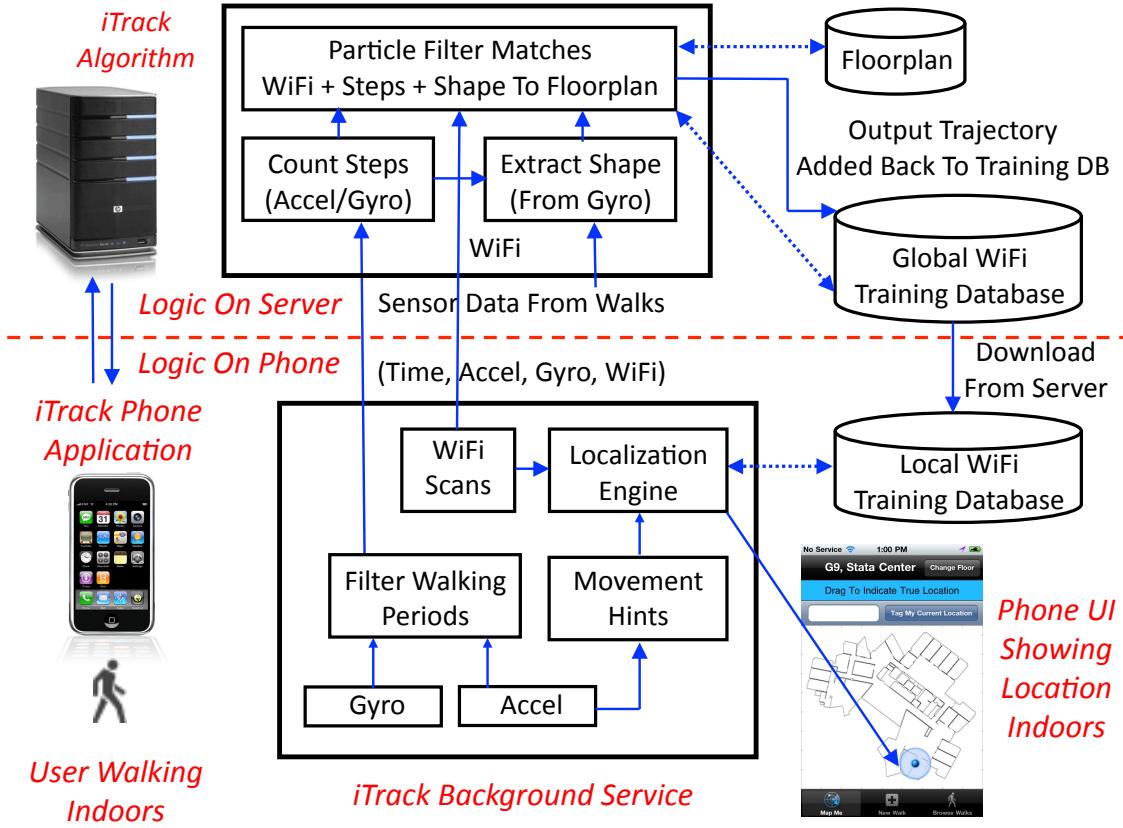


Figure 5-5: *iTrack* system architecture.

shown in the figure looks up fingerprints in a local WiFi training database to find the approximate location of the phone, and displays this location on an indoor map to the end user. The local training database is an on-phone cache of a global training database that is built up over time by *iTrack*. The localization engine also uses movement hints from the accelerometer to do better localization. It uses a simplified Viterbi algorithm similar to *CTrack*. We do not discuss the localization engine in this dissertation.

The task of the *iTrack server* is to extract accurate trajectories from WiFi fingerprints, accelerometer and gyroscope data in time periods when the user is known to be walking steadily with the phone either in his/her hand or pocket. The algorithm on the server consists of a *walking detector*, a *step counter*, a *shape extractor* and a *particle filter*. The step counter counts the number of steps in a walk using acceleration and/or gyroscope data. The shape extractor uses gyroscope data to estimate the approximate shape of the walk, i.e., where the significant turns in a user's trajectory lie.

The particle filter is an important component of the algorithm. Given a floorplan of the building, it uses Monte Carlo simulation to simulate paths with similar shape and size to the estimated shape and size, and find the most likely such path through the floorplan. The “size” of any stretch is obtained approximately from the step count using known bounds on human stride length. The Monte Carlo simulation helps eliminate candidate paths (or particles) that intersect walls or violate other constraints imposed by the floorplan. The particle filter also optionally fuses in previously collected WiFi training data to improve the accuracy of its output trajectories. Our filter is inspired by particle filters used in robotics, and to track pedestrians with foot-mounted sensors [67, 29].

However, as we shall see, there are some key differences from previous approaches to make the filter work with data from mobile phones.

The output trajectories produced by the particle filter are accurate to within less than a metre. In most cases, these trajectories can be joined with WiFi scanning data collected during the walk to generate new, accurate training data for WiFi localization. Training data is contributed back automatically to the global training database, which grows over time.

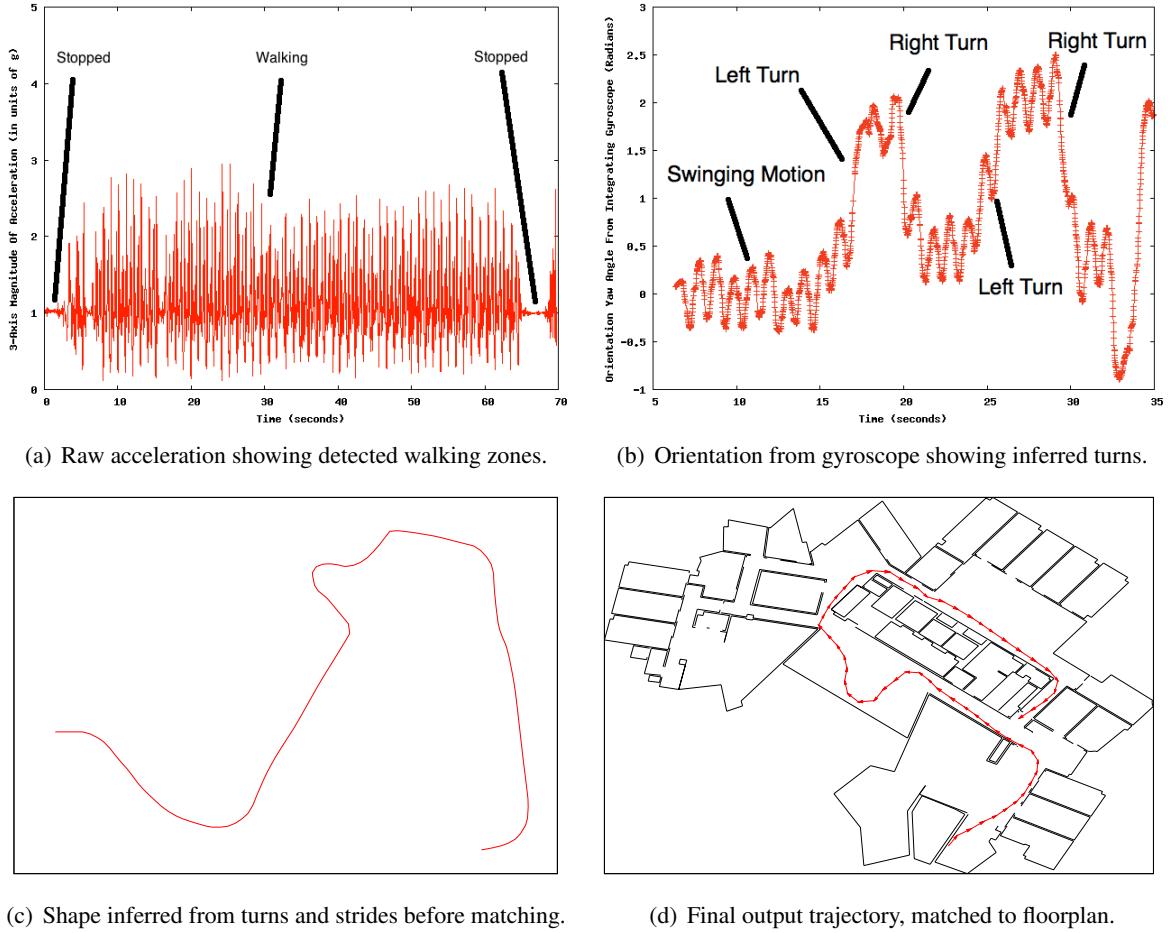


Figure 5-6: Illustration of steps in *iTrack*.

Figure 5-6 visually illustrates the steps to run *iTrack* on an example walk on the 9th floor of the MIT Computer Science department building. Figure 5-6(a) illustrates the first step, which identifies walking zones and extracts step counts from each walking zone in the accelerometer data. Figure 5-6(b) illustrates the process of shape extraction from gyroscope data. The figure shows the orientation computed by integrating gyroscope data, rather than raw angular velocities (which were shown earlier in Figure 5-4). As the plot shows, it is possible to infer the magnitude and direction of turns in a user's walk from this. Figure 5-6(c) shows an outline of the shape obtained by combining the turn information from the gyroscope with the step count information from the accelerometer. This shape is approximate and needs to be matched to the floorplan to find the most likely path taken by the user. The last step of *iTrack* is the particle filter, which takes the approximate shape and optionally, WiFi signal strength signatures as input, and matches the shape to a known building floorplan. The algorithm requires information about the locations of walls/obstacles (constraints)

and the absolute size of the floorplan to do this matching.

Figure 5-6(d) shows the output produced by the particle filter on our example walk. As we shall show, the trajectories produced by *iTrack* are accurate to within less than a metre.

The following sections describe the design of the individual components of *iTrack* in more detail: *walking detection*, *step counting*, *shape extraction* and *particle filtering*.

5.5.1 Walking Detection

The goal of walking detection is to identify zones where the user is walking steadily, so that we can extract trajectories for each of the walks made by the user. Walking detection takes raw acceleration from the accelerometer and angular velocity data from the gyroscope as input. It outputs a sequence of triplets $\langle t_{begin}, t_{end}, mode \rangle$, where t_{begin} and t_{end} represent the beginning and end of a time interval when the user is detected to be walking steadily, and $mode$ is an annotation for the interval that indicates if the phone was exclusively in the user's hand or pocket, or neither during the interval. Our current implementation of *iTrack* is able to extract accurate trajectory data from steady paced walks where the phone is in exactly one of the two poses: hand or pocket.

We model a phone as always being in one of three states: *stopped*, *moving* (but not walking steadily) and *steady walking*. Data from both stopped and steady walking zones is useful. In the stopped case, it can be used to learn how long a user spent at a particular location, or to learn what WiFi signatures occur at a particular location. In the steady walking case, it can be used to map the users' walk and/or collect WiFi training data for points in the walk.

In contrast, periods where the phone is moving, but the user is not walking steadily, usually represent the user using the phone in some way, such as to text, play a game, or make a call. These periods can also sometimes represent the user moving or walking slowly or irregularly, with frequent stops, so as to be undetectable as "steady walking" by the walking detector. We do not currently use the sensor data from these periods for mapping indoor trajectories because it is hard to distinguish slow or interrupted walks from random movements of the phone. Gyroscope data in such a period could reflect the user picking up or texting on the phone rather than changes in the orientation of the phone due to a user walking.

iTrack aims to accurately filter out and discard sensor data from two kinds of periods:

- Periods where a user is moving the phone in some way but not walking steadily.
- Periods where the phone changes position relative to its user, even in the midst of a steady walk (e.g., picking up a phone call).

We tackle this problem by dividing it into a number of simpler sub-problems that we describe in turn below.

Static vs Moving

The first sub-problem is identifying whether a phone is moving or at rest. *iTrack* uses the movement hint extractor developed for *CTrack* (described in Chapter 4, Section 4.6.4) to distinguish a phone at rest from a phone that is not.

Anomaly Detection

As discussed here and in the previous chapter on *CTrack*, it is essential to filter out “anomalies” where the phone is not at rest relative to its user to ensure accelerometer and gyroscope data accurately reflects turns and movements along a user’s walk. We use the same procedure described there (spike detection) to carry out anomaly detection. It should also be possible to use application state e.g., whether a call is active or what applications are running, to improve anomaly detection accuracy by filtering out some motions that don’t represent walking such as gaming, typing, or talking on the phone. We have not explored this in our current implementation, but believe it would be a useful extension to *iTrack*.

Hand vs Pocket

Assuming the pose of a phone is steady and anomalies or sudden transitions between poses have been filtered out, we present a procedure to determine the *pose* of the phone in any time window. We distinguish two specific poses: the phone being in the user’s hand, and the phone being in the user’s pocket. This problem is important because all the subsequent stages of *iTrack* (walking detection, step counting, and shape extraction) use different algorithms for the hand and pocket cases.

iTrack uses orientation computed by the gyroscope to distinguish a phone held flat in a user’s hand while walking from a phone held in a user’s pocket while walking. To do this, we use two key observations:

- The Euler *pitch* angle has magnitude close to zero when a phone is held flat in a user’s hand, but has non-zero magnitude (usually close to +1 or -1) when a phone is in a vertical orientation in a user’s pocket.
- The Euler *yaw* angle computed by integrating gyroscope data shows a marked, periodic up and down swing with high standard deviation in a user’s pocket when the user is walking, and a *much* lower standard deviation in a user’s hand when he/she is walking.

Recall from Section 5.4 that the output of integrating raw gyroscope data is a direction cosine matrix, $C_{phone}(t)$ at each time instant t . An alternative representation of 3-D orientation is the *Euler angle representation*: given any direction cosine or rotation matrix, it can be transformed to three Euler angles, called *roll*, *pitch* and *yaw*, often denoted by α , β , and γ respectively, that uniquely characterize a 3-D rotation in space [31].

It is possible to show that a change in yaw (γ) represents a rotation about the z-axis (parallel to gravity), which occurs whenever a user lifts her leg to take a stride. Similarly, a zero value of pitch ($\beta \approx 0$) represents a phone being flat in the x - y plane, most likely on a desk or on a user’s hand.

Figure 5-7 illustrates our observations. The panes of the figure plot the yaw angle γ and the pitch β obtained by solving the inertial navigation equation 5.7. The traces were collected by giving a user two iPhones, and asking him to walk with one iPhone in his hand and the other in his left pants pocket. Both yaw and pitch show a greater up-and-down swing with the phone in the pocket, but as it turns out, only yaw exhibits this swinging behaviour consistently across all the traces we have collected. Similarly, a pitch close to zero indicates a phone held flat, usually in a user’s hand if the phone is known to be moving.

iTrack windows its input trace into chunks of size a few seconds (over which this variance of yaw can be measured noticeably) and uses simple cutoffs on the values of two metrics to decide whether

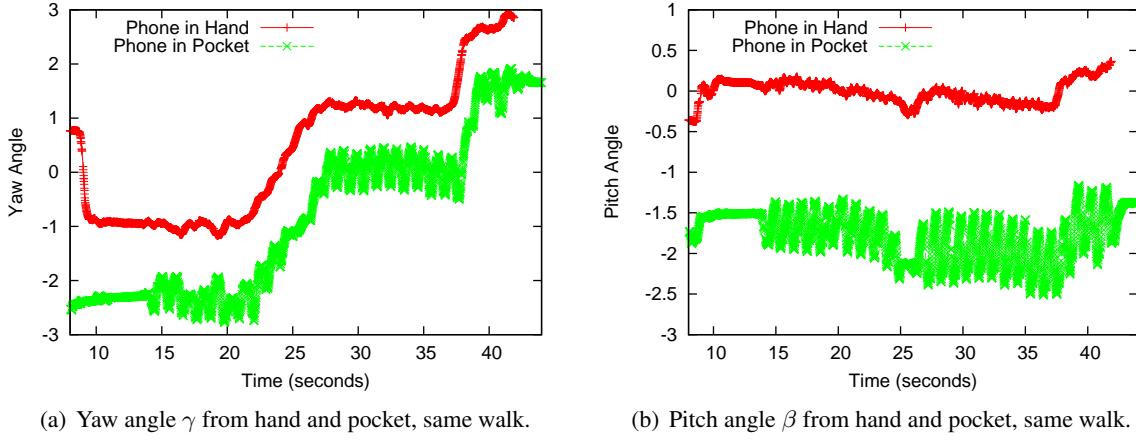


Figure 5-7: Distinguishing phone-in-hand and phone-in-pocket using Euler angles.

each window of data comes from a user’s hand or pocket. If the absolute value of pitch $|\beta| < 0.5$ or the median standard deviation of yaw $\sigma_\gamma < 0.1$, the window is determined to come from a user’s hand, otherwise it is determined to come from a user’s pocket. The values of these cutoffs were calibrated using a set of representative training walks.

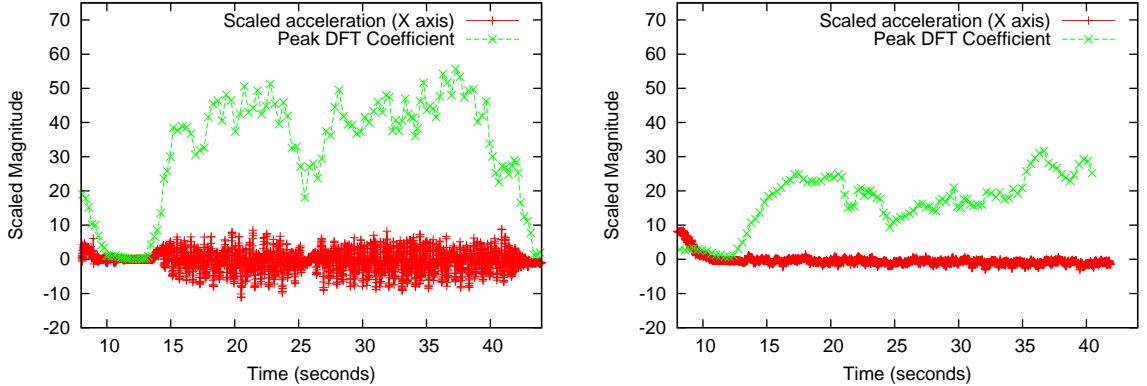
We note that it is possible that the yaw can have standard deviation $\sigma_{\text{gamma}} < 0.1$ in some time window if the phone is in the user’s pocket, but the user is not walking steadily. Since we only process data from steady walking, it is acceptable for the pose detector to mislabel such time windows as coming from a user’s hand. These time windows will be thrown out in any case in the next step (the walking detector) because they do not represent walking.

Walking vs Not Walking

Detecting walking is a trickier problem than detecting movement. To distinguish walking from periods of movement without walking, we use a *frequency domain* detection approach proposed in [75, 5], but modified to take into account the pose of the phone determined in the previous step (hand or pocket).

The key idea of frequency domain walking detection is that walking manifests as a specific kind of periodicity in the accelerometer signal. A phone experiences a periodic up-and-down swing in a user’s pants pocket or handbag when walking, or a gentle up-down or sideways motion even when a user walks with a phone in his/her hand. Both of these register a marked peak on a Discrete Fourier Transform (DFT). Such a peak is not found when a phone experiences a one-time (non-periodic) movement, such as a user picking up a phone call. It is possible to deliberately fool a walking filter based on an FFT by simply waving a phone up and down while at rest, but we believe this kind of adversarial behaviour should be rare in practice.

Figure 5-8 illustrates walking detection on data from an iPhone 4 accelerometer. The first pane, Figure 5-8(a) shows data from a trace collected with a phone in a user’s pocket when walking, and the second, Figure 5-8(b) shows data from a phone in a user’s hand when walking. The figures show the raw X-axis acceleration, as well as a plot of the peak DFT coefficient computed over a sliding window of 256 samples, as a function of time. The two magnitudes do not have the same units, but we have normalized their magnitudes to show them together on the same graph.



(a) X-Axis accel and peak DFT coefficient, phone in pocket. (b) X-Axis accel and peak DFT coefficient, phone in hand.

Figure 5-8: Walking detection using fourier transforms.

We can see that the walking zone is easy to visually identify in both situations, and corresponds to an increase in the magnitude of the peak DFT coefficient. Previous work in which the author of this dissertation was a co-author [5] quantitatively shows that the FFT peak power approach has high walking detection accuracy in terms of both precision (over 99%) and recall (over 97%).

The periodic motion is more pronounced in the pocket case because the phone swings up and down more regularly. It is less pronounced in the phone-in-hand case and corresponds to a smaller rise in the peak DFT coefficient. This suggests that the cutoff on DFT coefficient to identify a walking zone should be lower in the phone-in-hand case than the phone-in-pocket case.

To achieve this, we first *distinguish* the phone and hand cases explicitly as described in the previous section. We use a cutoff of 25 on the DFT coefficient for the phone-in-pocket case and a cutoff of 15 for the phone-in-hand case in our implementation of walking detection.

Limitations

Detecting walks when a phone in a user's hand is fundamentally unreliable, because a lower cutoff on DFT coefficient means a higher likelihood of *false positives*, where we identify a non-walking zone as a walking zone. Whether false positives are acceptable or not depends on the application. If using the walking data to crowd-source WiFi training in a building, we want a high cutoff to completely eliminate false positives and use only walks we are very confident about. If using the data mainly to understand a user's approximate movement pattern, false positives may be more tolerable. Fortunately, most false positives are short-lived and can be easily distinguished from longer walks by using a cutoff on walking time.

5.5.2 Step Counting

The step counting phase takes as input accelerometer and gyroscope data from walking zones detected by the walking detector. The output of step counting is a sequence of $\langle t_1, t_2 \rangle$ pairs for each walking zone indicating the beginning and end time of each stride in the walk. The total number of such pairs gives the number of strides in the walk.

The method used by *iTrack* for step-counting depends on whether the phone is in the user's pocket or held in her hand. In the pocket case, the periodic up-and-down swing of the phone is well-marked

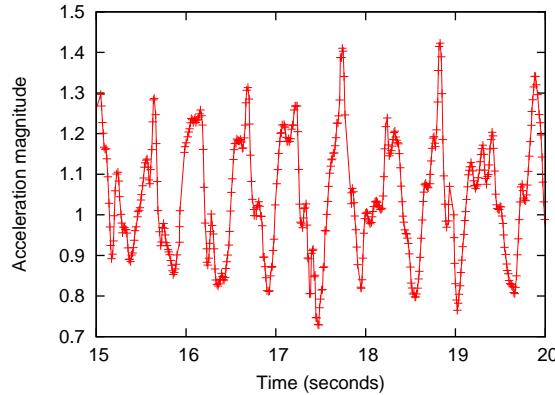


Figure 5-9: Peaks and valleys in acceleration, from phone held in user's hand.

on a plot of the yaw Euler angle, as shown in Figure 5-7(a) from the previous section. This swing can be used to be differentiate and count individual steps accurately.

In the phone in hand case, the individual steps are *not visible* on an orientation plot from the gyroscope, but the steps are somewhat discernible from a plot of raw acceleration. Figure 5-9 zooms in to a plot of the magnitude of acceleration over a smaller time scale. The *prominent* peaks and valleys in the acceleration signal approximately correspond to individual steps taken by the user.

In either case, *iTrack* needs to find the major peaks and valleys in a time domain signal — the yaw in the pocket case, and the acceleration magnitude in the hand case. Finding peaks and valleys turns out to be surprisingly challenging, because fluctuations in the value of yaw or acceleration magnitude commonly manifest as false peaks. Figure 5-9, for example, illustrates this problem: the acceleration magnitude often has multiple peaks per stride. It is non-trivial to design an algorithm to count the number of strides, even though the strides can be easily distinguished by eye.

One approach is to use a smoothing (low-pass) filter to eliminate some of the small high-frequency fluctuations. However, we have found that a low-pass filter ends up eliminating some genuine steps irrespective of how it is tuned, which is undesirable.

Hence, rather than using a low-pass filter, *iTrack* uses an iterative heuristic to find the relevant peaks and valleys. The heuristic we use operates in multiple passes. The first pass identifies *all* peaks and valleys (taking wrap-around of orientation into account if operating on yaw angle data). The peaks and valleys extracted in the first pass include fluctuations. The second pass examines the peaks and valleys found in the first pass, and *eliminates* a subset of valleys and peaks that are likely to be from fluctuations. Specifically, we eliminate peaks and valleys that satisfy the following condition(s):

- We eliminate any peak P that has a nearby *higher* peak P' (within some time interval) such that *there is no valley between P and P'* . The intuition is that such a peak is not the main peak in its stride with high probability.
- Analogous to the above, we eliminate any valley V that has a nearby lower valley V' (within some time interval) such that there is no peak between V and V' .

We repeat the above process iteratively. In practice, we have found that three passes are sufficient to remove all the fluctuations and retain only the major peaks and valleys, one per stride.

True Step Count	Estimated From Pocket	Estimated From Hand
78	80	90
32	34	30
29	30	36
49	50	56
64	66	68

Table 5.1: Step counts from pocket are more accurate than from hand.

How Well Does Step Counting Work?

Table 5.1 shows the accuracy of step counting on 5 walks of different lengths. Each walk was collected with two phones: one in the pants pocket and one held flat in the hand. The user counted out the number of steps loudly while walking to record the true number of steps walked.

The mean estimation error is 3.5% when the phone is in the pants pocket and over 11% when the phone is in the user’s hand. The peak-finding approach is extremely accurate for a phone in a user’s pants pocket, almost always finding the exact number of steps walked to within one or two. However, the approach is not as accurate for the phone-in-hand case since the up-and-down or sideways swings of the phone are not as well-marked and regular. For this reason, the estimated step count inevitably has some error in the phone-in-hand case.

However, the approximate step count even from the hand case is still accurate enough to provide useful information to the next step of *iTrack*, the particle filter. The particle filter in *iTrack* is a probabilistic model that assumes estimated stride counts have error distributed according to some distribution. As we shall see in Section 5.8.10, by using a carefully chosen *non-Gaussian* error distribution for stride length, the particle filter can correct for errors in stride count and still find the correct trajectory walked by the user. This is analogous to how the higher-layer Markov Model in *VTrack* (Chapter 3) deals with erroneous WiFi data obtained from a lower-layer centroid localization algorithm.

5.5.3 Shape Extraction

The goal of the shape extractor is to determine the shape of a user’s trajectory as precisely as possible from gyroscope data. The shape extractor takes raw gyroscope data as well as the output of step counting as input. The output of the shape extractor is a set of triplets of the form $\langle t_1, t_2, \delta\theta \rangle$ where t_1 and t_2 are the begin and end times of each stride (same as the output of step counting) and $\delta\theta$ is the *change* in walking direction from the previous stride to the current stride.

To understand how shape extraction works in *iTrack*, recall that the output of integrating the angular velocity from the gyroscope is a direction cosine matrix, $C_{phone}(t)$. The columns of this matrix represent unit vectors along the axes of the phone’s reference frame at time t . This matrix can be converted to a different representation in terms of three Euler angles: yaw (γ), pitch (β) and roll (α) [31]. Among these, the yaw represents rotation of the phone’s reference frame about the z axis (parallel to gravity), and is therefore the most relevant from the point of view of determining walking direction in the x - y plane.

When the phone is held in a user’s hand, shape extraction is easy (assuming, as before, that we have filtered out cases where the user uses the phone for a call or fiddles around with it) because the change in the walking direction corresponds to the change in the orientation of the phone in the x - y

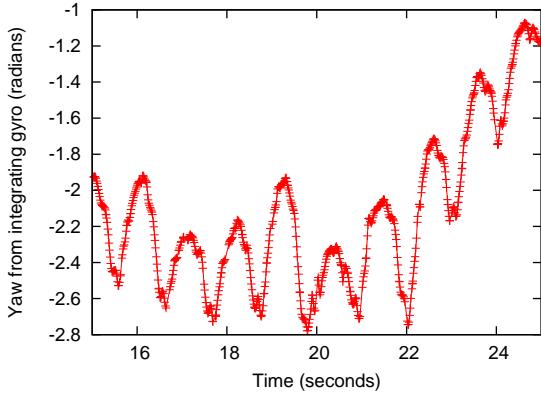


Figure 5-10: Variation in yaw within each stride, phone in pocket.

plane, which is the yaw. Therefore, we simply compute the change in yaw from each stride to the next to find the change in walking direction $\delta\theta$ from each stride to the next.

The shape extraction problem is challenging when a phone is in a user's pocket. This is because a phone in a pocket executes a complex rotation about all three axes in 3-D space when the user lifts her leg to start a stride or pivots on her leg while stepping forward with another leg. The relationship between the absolute yaw value of the phone and the direction of the user's walk is non-trivial. In fact, it is *not even fixed: the orientation of the phone relative to the walking direction varies continuously during a person's stride*. We refer the reader to Figure 5-10, which shows a zoomed in plot of yaw angle from integrating gyroscope data. The figure shows this up-and-down variation in yaw within each stride.

The figure, however, shows that the *baseline level* of the yaw (mid-point between peaks and valleys) tracks the turns made by a user quite well. In other words, although the 3-D orientation of the phone relative to user walking direction is unknown, and is never constant, the *change in the baseline level of yaw over multiple strides tracks the change in walking direction very precisely*. For example, observe the transition (increase) in the baseline yaw at approximately $t = 22$ seconds in the figure. This corresponds to a left turn when using the iPhone's convention for reporting angular velocity.

iTrack uses this observation to find the shape of a user's trajectory. We use the same peak finding algorithm discussed in the previous section for step counting. Having identified a sequence of peaks and valleys, the algorithm computes the midpoint of each $\langle \text{peak}, \text{valley} \rangle$ pair. It then computes the change from each midpoint to the next to find the change in walking direction $\delta\theta$ from each stride to the next.

How Well Does Shape Extraction Work?

We have seen that inertial drift is a major problem with integrating raw gyroscope data. *iTrack*'s shape extraction algorithm avoids the drift problem because it only estimates and uses *change* in orientation, rather than absolute orientation. Hence, the estimation error is confined to turn angles, and the error is upper-bounded by the relatively small integration drift error over the short time scale of each turn.

In practice, we have found that shape extraction is extremely accurate for a phone held flat in a user's hand. When a phone is in a user's pocket, however, we have found that raw gyroscope data

on the iPhone sometimes experiences outages lasting for a relatively long time (up to 1 second or more), which sometimes lead to significant errors in estimating turn angle.

It is difficult to provide a quantitative evaluation of shape extraction error because our ground truth approach relies on a human approximately following a marked path (Section 5.8.2), which is not accurate enough to measure the true turn angle precisely. Qualitatively, we have found that when a user is walking straight without changing his direction, *iTrack*'s shape extraction works perfectly and detects that the direction has not changed. When the user does make a significant turn, it is always detected, but the turn angle may sometimes be estimated incorrectly by integration.

This observation influences the design of the particle filter described in the next section. Previous approaches that use particle filters for inertial tracking have used a Gaussian model of error in orientation. As we have described earlier, this does not accurately reflect the underlying process. We instead use a model with a fixed probability of error when a user makes a significant turn, and a smaller probability of error when the user's walking direction is not changing significantly. We show that this model improves the *survival rate* of the particle filter.

5.5.4 Particle Filtering

The key step of *iTrack* is a particle filter algorithm that matches the approximate shape and size extracted by the shape extractor and step counter to a floorplan of the building to find the most likely trajectory taken by a user. The particle filter takes the following as input:

- A sequence of $\langle t_1, t_2, \delta\theta \rangle$ triplets indicating the beginning time, end time and change in direction for each stride, output by the shape extractor (see previous section).
- The geometry of the floor the user is known to be on (we discuss how we can find which floorplan to use later), consisting of the geometry of all walls and other constraints.

Optionally, the particle filter can also use the following inputs to improve accuracy:

- A sequence of observed WiFi signatures at times along the walk, of the form $\langle t, w \rangle$.
- A training database of WiFi signatures and known points in the floorplan they were observed from, of the form $\langle x, y, w \rangle$. These could have been learned over time from previous walks collected by the system.

The output of the particle filter is:

- The most likely trajectory taken by the user, as a sequence $\langle t, x, y \rangle$, if it succeeds in finding a feasible trajectory (it may not in all cases, as we shall see).

The key intuition behind the particle filter is that the walls and other obstacles in the floorplan *constrain* the set of possible trajectories the user could have taken. This, in combination with the constraint that the trajectory follows a particular shape and has (approximately) a particular length should in theory suffice to uniquely localize the user within the floorplan, and determine the exact trajectory.

However, this assumption is not true for short walks with few or no turns, and in floorplans with significant intra-plan symmetry. For example, if the floorplan has a box-like layout with identical corridors, knowing the shape and size of a walk does not suffice to uniquely determine its location. In such cases, *iTrack* can either use WiFi localization with its existing training database if available to disambiguate different trajectories, or it can uniquely identify the trajectory if the approximate starting point and/or direction of the walk is known.

We therefore use a hybrid approach, as discussed in the next section. We first collect a few “seed walks” with known start position (input manually on a map) to initialize the WiFi training database on each floor of the building we want to map. Subsequently, users carrying around phones can contribute data in the background without having to mark their locations on a map. We show in the evaluation that *iTrack* is able to use a small amount of seed WiFi training data, of the order of 4-5 walks to localize subsequent walks to within less than a metre. Collecting 4-5 walks of seed data with *iTrack* requires only 5-10 minutes of training effort on a $\approx 10,000$ sq. ft. floor.

How Does iTrack’s Particle Filter Differ From Previous Approaches?

The use of particle filters for indoor localization and tracking is not novel. Robots have used particle filters for localization and mapping (SLAM) indoors. However, they have usually relied on more accurate sensors such as laser range finders. More recently, Woodman and Harle [67] used a particle filtering technique similar to *iTrack* to track pedestrian movement indoors with foot-mounted inertial sensors. *iTrack*’s particle filter differs in three key respects:

- *iTrack*’s particle filter uses a mixture model for stride length to correct for errors in stride count estimation, which are quite common especially with phones held in a user’s hand. In contrast, [67] uses a Gaussian model for stride length. This works well for foot-mounted inertial sensors because the length of a pedestrian’s stride can be accurately estimated from such a sensor, but as we show, it *does not work well for phones*.
- *iTrack*’s particle filter uses WiFi measurements throughout the filter to improve tracking accuracy. Previous techniques have mainly used WiFi for initialization rather than throughout the course of the filter.
- Walks collected by *iTrack* are fed back to the training database to improve the quality of WiFi localization estimates. This means *iTrack* can be run iteratively with the same training data to further improve matching accuracy, and so on. This iterative process yields an improvement in accuracy and is a novel contribution of our system.

The rest of this section is organized as follows. We first explain what a particle filter is and how it works. We also explain how it relates to a Hidden Markov Model — which was used in *VTrack* and *CTrack* — and why, unlike in *VTrack* and *CTrack*, a particle filter is a more appropriate tool for *iTrack* than an HMM. We then describe the particle filter used in *iTrack*.

What is a Particle Filter?

A particle filter is a Monte Carlo simulation technique that can approximate the *posterior* probability distribution of a continuous-domain Markov process. The easiest way to understand particle filters is to understand how they relate to (and differ from) HMMs. As we have seen in Chapters 3 and 4, an HMM consists of a discrete set of underlying hidden states and observables. Each observable has

a probability of being emitted from each hidden state and there is a transition probability governing transitions between states. Given a sequence of observables, the Viterbi algorithm finds the most likely sequence of (discrete) hidden states.

HMMs naturally generalize to a *continuous* state space, where the hidden state at each time step is a continuous variable rather than one of a set of discrete possibilities. An indoor trajectory of a person walking through a building is an example of such a sequence of continuous-domain hidden states, where each unknown is a location coordinate $\langle x, y \rangle$ on the floorplan. In the continuous domain, one can no longer use the Viterbi algorithm to solve for the most likely sequence of hidden states, since there are an infinite number of possibilities. Finding the most likely trajectory is not analytically tractable in most cases.

Particle filters instead use Monte Carlo simulation to approximate the solution to this problem. The idea is to approximate the *posterior* probability distribution over the space of hidden states by a set of weighted *particles*. In the indoor trajectory mapping context, for example, each particle represents a candidate trajectory through the floorplan. At each step of the filter, each particle advances to a new hidden state (new location in the indoor positioning context), which is randomly sampled from a *proposal distribution*.

The weights of particles are updated after each simulation step to factor in the probability of seeing the observable (sensor data) at that time step from the newly simulated hidden state. Particles with extremely low or zero weight — in the indoor tracking context, paths that violate constraints in the floorplan, are discarded at the end of each step.

At the end of the simulation, what remains is a set of weighted particles that approximates the true posterior distribution over the sequence of hidden states. The most likely sequence of hidden states is simply chosen to be the particle with maximum weight.

Formalism

Given:

- A sequence of observations o_1, o_2, \dots, o_n .
- The probability density function of seeing an observable o from a given hidden state h , $P_{\text{emission}}(o|h)$.
- A density function governing transitions between hidden states, $P_{\text{transition}}(h_{i+1}|h_i)$.

The sequence of observations o_1, o_2, \dots, o_n corresponds to some true (unknown) sequence of hidden states h_1, h_2, \dots, h_n . The goal of a particle filter is to estimate the posterior probability distribution $P(h_1, h_2, \dots, h_n | o_1, o_2, \dots, o_n)$, and use this to find the most likely sequence of hidden states.

The particle filter approximates the posterior probability distribution at each step by a set of M particles with associated weights:

$$\langle p^1, w^1 \rangle, \langle p^2, w^2 \rangle, \dots \langle p^M, w^M \rangle. \quad (5.12)$$

The particles are initialized according to a prior probability distribution $P_{\text{prior}}(h_0)$ over the initial value of the hidden state h_0 .

At each step i , each particle p^j stores a sequence of hidden states up to step i : $h_1^j, h_2^j, \dots, h_{i-1}^j$. The next hidden state, h_i^j is sampled from a *proposal* distribution $P_{proposal}(h_i^j|h_{i-1}^j)$. The proposal distribution is usually the same as the transition probability distribution (though it need not be).

At each step i of the Monte Carlo simulation, the weight w^j of each particle j ($1 \leq j \leq M$) is updated using Bayes theorem:

$$w_i^j = w_{i-1}^j \times \frac{p(o_i|h_i^j) p(h_i^j|h_{i-1}^j)}{P_{proposal}(h_i^j|h_{0..i-1}^j, o_{0..i})} \quad (5.13)$$

If the proposal distribution used is the transition probability distribution (as is usually the case), this simplifies to:

$$w_i^j = \begin{cases} w_{i-1}^j \times p(o_i|h_i^j), & P_{proposal}(h_i^j|h_{0..i-1}^j, o_{0..i}) \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (5.14)$$

The above weight update rule can be interpreted intuitively as follows. If the newly simulated hidden state for a particle has zero probability — for example, if it violates a constraint, such as crossing a wall, then the particle’s weight should be set to zero, killing the particle. Otherwise, the particle’s weight should be multiplied by the emission probability for the new observation from the newly simulated hidden state. Unlike in the Viterbi decoding algorithm for a HMM, we do not multiply by the transition probability because it is already implicitly accounted for in the way we sample new hidden states from the proposal distribution.

Resampling. As the importance weights of a particle filter are updated according to equation 5.14, the importance weights of all but a few of the particles tend to approach zero or become zero. This happens in indoor trajectory mapping because most simulated particles start with the wrong initial position or orientation and quickly hit walls or violate other floorplan constraints. This leads to a problem called *degeneracy*: if only a small number of surviving particles are left behind after the first few steps of the simulation, they cannot satisfactorily explore the state space on subsequent steps. For this reason, most particle filters include a *resampling* step to ensure the number of surviving valid particles remains above a threshold.

We use a standard technique from the particle filtering literature called sequential importance resampling (SIR). SIR works by computing an estimate of the effective number of particles at the end of each step:

$$M_{eff} = \frac{1}{\sum_{j=1}^M (w_j^{norm})^2} \quad (5.15)$$

where w_j^{norm} is the normalized importance weight of particle j , obtained from the following equation:

$$w_j^{norm} = \frac{w_j}{\sum_{j=1}^M w_j} \quad (5.16)$$

If the effective number of particles M_{eff} is less than a given threshold M_{thresh} , then the filter performs resampling. Resampling works by re-drawing M particles from the current particle set

with probabilities proportional to their importance weights. The M new particles are used to replace the current set of particles, and all their importance weights are set to $\frac{1}{M}$.

Particle Filter Used In *iTrack*

We first describe the simplest version of the particle filter used in *iTrack* that does not take WiFi signal strength observations into account (we later show how we can modify the filter to do this). The basic filter is a slightly modified particle filter with the following components:

- The “steps” of the Monte Carlo simulation are individual strides made by the user, and extracted by the step counting algorithm described earlier.
- The unknown hidden state h_i at the end of step i is a triplet $\langle x_i, y_i, \theta_i \rangle$ where x_i and y_i are coordinates of the user’s location on the floorplan and θ_i is the *absolute* walking direction of the user ($\theta = 0$ represents due East if the floorplan is oriented N-S).
- The observable at each step o_i is $\delta\theta$, the change in angle from the previous stride to the next stride, obtained from the output of *iTrack*’s shape extractor.

Rather than using an explicit emission probability distribution for the observable $\delta\theta$, we fold the observable into the proposal distribution for the particle filter and set the emission probability to be equal to 1 always. The proposal distribution we use generates the new hidden state of the particle filter at step i given the old hidden state at step $i - 1$ using the following rules:

$$\theta_i = \theta_{i-1} + \delta\theta + E_{\delta\theta} \quad (5.17)$$

$$x_i = x_{i-1} + l \cos(\theta_i) \quad (5.18)$$

$$y_i = y_{i-1} + l \sin(\theta_i) \quad (5.19)$$

Importantly, particles where the transition from $\langle x_i, y_i \rangle$ to $\langle x_{i+1}, y_{i+1} \rangle$ crosses a wall or obstacle in the known floorplan are assigned zero weight and killed.

In the above rules:

- $E_{\delta\theta}$ is a perturbation error term that models the error in extracting turn angles from the gyroscope. This term is drawn from an *angle error distribution* we describe below.
- l is a stride length sampled from a *stride length distribution*.

The angle error distribution models the error in extracting a turn angle from the phone gyroscope. Each simulated particle is assigned a slightly different random turn angle which is a perturbation of the measured turn angle from the gyroscope.

Similarly, the stride length distribution models the uncertainty in how long each stride is. Each simulated particle uses a different random stride length at each time step. Some particles end up

modeling stretches of walk where the user takes short strides, and others explore paths with longer strides.

In a collective sense, the simulated uncertainty in turn angle and stride length help explore the space of possible trajectories thoroughly and discover the true trajectory of the end user, correcting for estimation errors made by the step counting or shape extraction phases. Particle trajectories that simulate the wrong stride lengths or angles end up crossing walls or violating constraints, and are assigned zero weight. Particles that do not violate the floorplan constraints are automatically given higher weight.

Resampling. In the *iTrack* case, some weights are zero and the other importance weights are all equal. In this case, the effective weight for resampling M_{eff} reduces to a simple count of the number of surviving particles. The resampling step reduces to:

- Discarding particles that have not survived.
- Making duplicate copies of the surviving particles randomly so that the total particle count remains at M .

Our particle filter implementation uses $M = 100,000$ and $M_{eff} = 10,000$. Both these values reflect a trade-off between how thoroughly we need to explore the state space, and how computationally expensive the particle filter ends up being. Larger values correspond to better exploration and higher accuracy, but slower running times, and vice versa. We describe more details of this tradeoff and our choice of parameters in Section 5.7.2.

Choice Of Error Models and Survival Rate

The particle filter as described above may not always produce a valid trajectory. Since the number of particles that can be simulated, M , is finite owing to computational constraints, if *all* the particles being simulated end up violating a floorplan constraint by crossing a wall or obstacle at the same time instant during the trajectory, the filter ends up producing no output at all.

The usual reason for a failure to survive is an error in the sensor measurements that is unlikely to occur under the model being used for angles or stride length. This produces an incorrect shape which cannot be corrected by the filter unless an order of magnitude more particles are simulated (which is computationally infeasible).

The choice of angle error distribution and stride length distribution are therefore important to ensure a good survival rate. Unlike previous approaches, *iTrack* uses non-Gaussian models for both angle error and stride length, which we show to have better survival rate than Gaussian models since they are more faithful to the underlying sensor data from a phone (Section 5.8).

Angle Error Model

As we have discussed earlier, the shape extractor usually does a good job of detecting when a person's walking direction is not changing much, and also of detecting turns. However, it sometimes computes the turn angle incorrectly owing to gaps in the gyroscope data that lead to errors in numerical integration. For this reason, rather than using a simple Gaussian model of error in $\delta\theta$, we use a mixture model that simulates two types of error in the angle:

- A Gaussian error with small variance σ_{small} that we add to each $\delta\theta$ sample from the gyroscope with probability p_1 .
- A Gaussian error with larger variance σ_{large} that we add *only to samples that represent significant turns*, with some probability p_2 . A significant turn is defined to be one that exceeds a certain cutoff ($\delta\theta \geq \theta_{turn}$).

The first type of error simulates small thermal or other noise, while the second captures integration errors that are made sometimes when computing the change in orientation over a significant turn.

We used $\sigma_{small} = \frac{\pi}{100}$, $\sigma_{large} = \frac{\pi}{10}$, and $\theta_{turn} = \frac{\pi}{10}$, calibrated from tuning experiments. A reasonably wide range of values appear to work for these constants.

We show in our evaluation (Section 5-20(b)) that using a mixture model improves the *survival rate* of our particle filter. Compared to a Gaussian model for angle that has a 68% survival rate, the mixture model we use has an 82% survival rate.

Stride Length Model

Why, the height of a man, in nine cases out of ten, can be told from the length of his stride.

- Sherlock Holmes to Dr. Watson, *A Study in Scarlet* (1887)

Similar to angle error, *iTrack* uses a mixture model for stride length rather than a simple Gaussian model. The reason for this is that a person's stride is usually multi-modal: the stride length varies depending on how fast he/she is walking. The stride length is usually longer when a person is walking faster and shorter when a person is walking slower. A Gaussian model for stride length cannot capture this because a sequence of short or long strides is always extremely unlikely (in fact, exponentially unlikely) in a Gaussian model, *irrespective of its standard deviation*.

For example, in situations where a person traverses a stretch or a corridor slowly using a sequence of short strides (quite common in our experience), a Gaussian model will inevitably fail to find even a single particle that survives the passage through the corridor.

For this reason, we use a mixture model which keeps track of the previous stride length, and either *transitions* to a different stride length on a new stride with a transition probability $P_t = 0.5$, or remains at the same stride length with probability $1 - P_t$. Each new stride length is drawn from a Gaussian with mean equal to approximately 3 feet (the average stride length of a human). As suggested by the quotation at the beginning of this section, the stride length model can be improved if the height of the user is known. However, we do not rely on this since this requires extra calibration.

We show in our evaluation that using a mixture model for stride length significantly improves the survival rate over using a Gaussian model from 64% to 82%.

Bootstrapping the Filter

We find in our evaluation that the simple particle filter (as described above) is inaccurate, producing the wrong output for 30% of its input trajectories. This happens because the initial position and orientation of the user are unknown, and there are multiple particles that have non-zero weight, all with different trajectories. With short or medium length walks, the same shape can be a valid path anywhere within a large floorplan owing to symmetry in the building plan. This is a serious problem

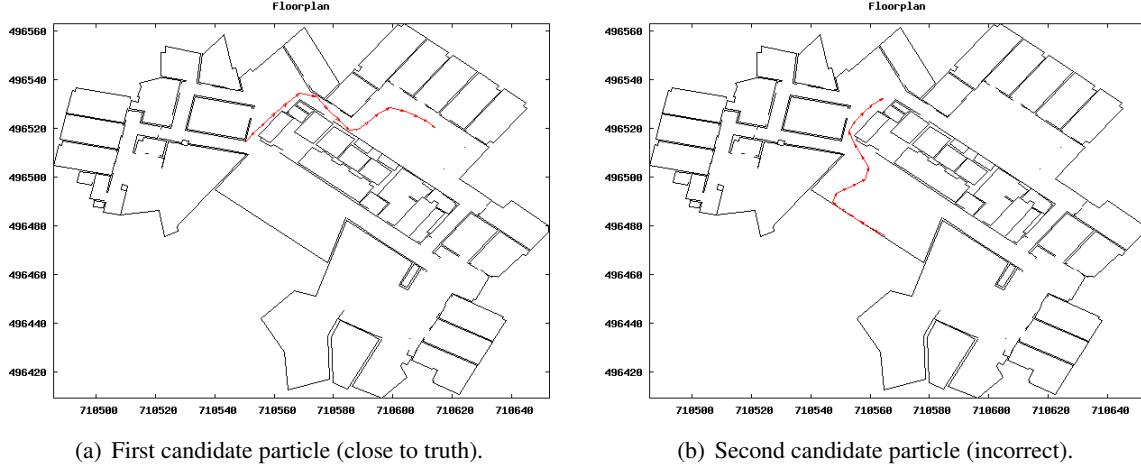


Figure 5-11: Example showing ambiguity in the output of the particle filter.

because in practice, many walks taken by a user are likely to be short or medium length walks e.g., a quick stroll to the lounge, a meeting room, or a vending machine in an office building, or a walk between close by sections of a store or museum.

Figure 5-11 illustrates the ambiguity in the output of the particle filter with an example from a short walking trace. The output of the filter consists of many particles with equal weight, of which the figure shows two. Figure 5-11(a) shows a particle that is close to the true trajectory walked by the user. Figure 5-11(b) shows a second particle whose shape is *identical* to the first particle, but is the incorrect output. Without knowledge of initial position or orientation or other external information, it is impossible to distinguish between the two traces from inertial sensing data alone.

iTrack uses two main techniques to eliminate ambiguity in matching short and medium length walks:

- If some “seed WiFi” information is available, it integrates WiFi localization probabilities into the particle filter to disambiguate output particles.
- It can use a *dedicated training phase* where the user explicitly inputs his/her initial position on a map of the floorplan, by marking it on a mobile device. However, unlike previous crowd-sourcing schemes [46, 70], it is sufficient for the user to mark the starting location — the rest of the walk can be deduced automatically using our particle filtering algorithm.

In practice, our system uses a combination of both the above techniques. For each floor of each building we want to cover, we use a dedicated training phase where a single person or group of people contribute walks to the system. The user marks initial position information explicitly on a map before he/she begins walking. We show a screenshot of our iPhone user interface for training in Section 5.7.

Once a small number of such “seed walks” have been collected spanning the entire floor, the WiFi data from the walks can be used to map subsequent walks accurately without knowing their initial position. When this happens, the dedicated training phase is complete. From here on, as we show in our evaluation, any walk the user takes with the phone in his/her hand or pocket can be tracked accurately irrespective of its length.

This is a significant improvement over previous crowd-sourcing techniques because it requires only a limited number of manual training points on each floor being mapped.

The next sections describe how *iTrack*'s particle filter incorporates initial position information and WiFi localization information, respectively.

Incorporating Initial Position

It is possible to incorporate approximate knowledge of initial position by constraining the prior distribution $P_{prior}(h_0)$. Suppose a user marks her initial location on the floorplan to be $\langle x_0, y_0 \rangle$. The particles used in our particle filter are initialized with values of h_0 randomly drawn from a uniform distribution within a circle with centre (x_0, y_0) and radius R . R is a radius of uncertainty to allow for some error in the user input. We use $R = 10$ feet in our implementation.

We show quantitatively in our evaluation that knowledge of initial position significantly improves the accuracy and recall of trajectory matching indoors.

Initial Orientation. In addition to incorporating initial position, we have implemented an improvement to *iTrack* that can use the magnetic compass to find initial orientation and use it to improve the accuracy of trajectory mapping. This is used in dedicated training mode when gathering seed WiFi data.

The training procedure requires users to hold a phone flat in their hand held forward *in the direction they are going to be walking in*. The system detects when the phone is flat in the person's hand using the techniques we have described earlier, and measures the absolute orientation of the phone relative to geographic north. This orientation is used to initialize the initial orientation θ_0 of particles in the prior. Orientation from the magnetic compass tends to be quite inaccurate, especially indoors near metallic objects. Hence, we sample the initial orientation of particles uniformly from a (relatively large) radius of uncertainty, $\theta_0 \pm \frac{\pi}{5}$, just as in the case of initial position.

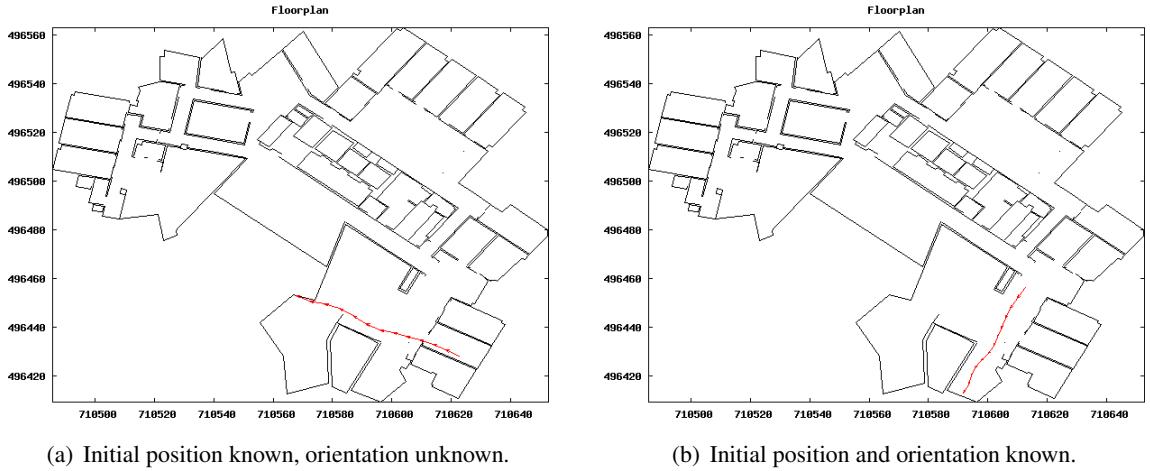


Figure 5-12: Knowledge of initial direction helps improve trajectory matching.

Figure 5-12 shows an example of a short walk on which knowledge of the initial orientation improves trajectory mapping. Figure 5-12(a) shows the output of our particle filter when given the approximate initial position (from WiFi localization), *but not the initial orientation*. Figure 5-12(b) shows the filter output when given both the initial position and walking direction (from compass).

This is also the true trajectory walked by the user. Without knowing the initial walking direction, we see that it is hard for the algorithm to prefer either of the trajectory in Figure 5-12(a) or Figure 5-12(b), because the initial location obtained from WiFi localization is approximate and cannot distinguish the two starting points.

Incorporating WiFi Information

If a seed WiFi training database is available, *iTrack* can use WiFi location information to significantly improve the accuracy of trajectory mapping. WiFi helps eliminate ambiguity between candidate trajectories in different parts of the floorplan space that have similar shape.

We incorporate WiFi location by modifying the particle filter to include an emission probability distribution for WiFi. Given a new observation o_i with a WiFi fingerprint F_i , we modify the particle weight update rule to:

$$w_i^j = w_{i-1}^j \times E(F_i|x_i^j, y_i^j) \quad (5.20)$$

Here, $E(F|x, y)$ is an emission score proportional to the probability of seeing wifi signature F from location (x, y) in the floorplan. We use the gridding approach proposed in *CTrack* to evaluate this emission score. We divide the floorplan area into equal-sized grids, and find the grid G containing coordinate (x, y) . We set $E(F|x, y)$ to be identically equal to $E(F|G)$, the the emission score in *CTrack* (Section 4.5.3, Chapter 4).

We also modify the SIR resampling step of the particle filter to incorporate the weights from WiFi observations and compute the true effective weight using equation 5.15.

We show in our evaluation that incorporating WiFi localization information significantly improves the accuracy of *iTrack*, and a small amount of seed WiFi information is in fact necessary for the system to work correctly.

5.6 Simplifying Training With *iTrack*

One of the objectives of *iTrack* is to significantly reduce the effort to gather training data in a new building or floorplan, and to enable users to easily contribute crowd-sourcing data with little effort. To achieve this, *iTrack* uses a novel *iterative* algorithm to maximize the accuracy of WiFi training data collected and to collect the highest quality training data from a set of user walks. The algorithm proceeds in three phases: *dedicated training*, and two passes of *crowd-sourcing*, as we explain.

5.6.1 Dedicated Training

In this phase, a dedicated training person or volunteer collects a few seed walks using a well-defined procedure to collect a small amount of “seed” WiFi training data for each floorplan.

Before each walk, if necessary, the person explicitly chooses which floorplan to train for from a set of nearby candidate floorplans, found using network location (such as obtained from the Skyhook or Google API) or last known GPS location. He then marks initial location on a map of the floor while holding the phone in the direction he intends to start walking. He puts the phone in his pocket, and starts walking. When done, he takes the phone out of his pocket and indicates that the walk is complete, upon which the walk is sent back to the *iTrack* server.

The seed walks should cover the important regions of the floorplan but do not have to comprehensively cover every location. For example, one person could collect the seed data for the 9th floor in the MIT Stata Center in a few minutes.

5.6.2 Crowd-sourcing Phase 1

End users download and install the *iTrack* application on their phones to view their approximate location on the map, and to contribute data. Importantly, once dedicated training is complete, *end users do not have to take any effort to contribute data — they simply need to run the *iTrack* application on their phones*. This is a big improvement over having to explicitly mark or indicate location on a map.

The system continuously runs in the background as shown in Figure 5-5, detecting walks made by the user. Each new walk is uploaded to the *iTrack* server.

Having collected a certain number of walks from users, the system runs an initial pass of the particle filter over all the walks, with the goal of extracting more WiFi data. This first pass of the filter uses the seed WiFi data from dedicated training.

A key optimization we use is to add WiFi data back from the first pass to the training database, and then *re-run *iTrack* over all the walks* using the augmented database. This constitutes Phase 2.

5.6.3 Crowd-sourcing Phase 2

Phase 2 of crowd-sourcing runs a second pass of the particle filter over *all the walks* collected by the system, using the augmented WiFi training database from the first pass as input. When matching a walk W in the second pass, we selectively use only the WiFi training data that did *not* come from the first pass of running *iTrack* over W . The idea of this “leave-out” technique is to avoid reinforcing errors made in the first pass. If the first pass produced a wrong trajectory for W , this incorrect data gets added to the WiFi training database and further increases the likelihood that the second pass will also produce the same (or very close by), but incorrect output. Our approach ensures this does not happen.

We show in our evaluation that using a second pass of iteration helps improve the accuracy of trajectory matching, especially in the tail (not so much in the median). Some of the walks that do poorest on the first pass improve substantially in the second pass. We use only two passes because we have not found additional passes of the algorithm to yield substantial benefit.

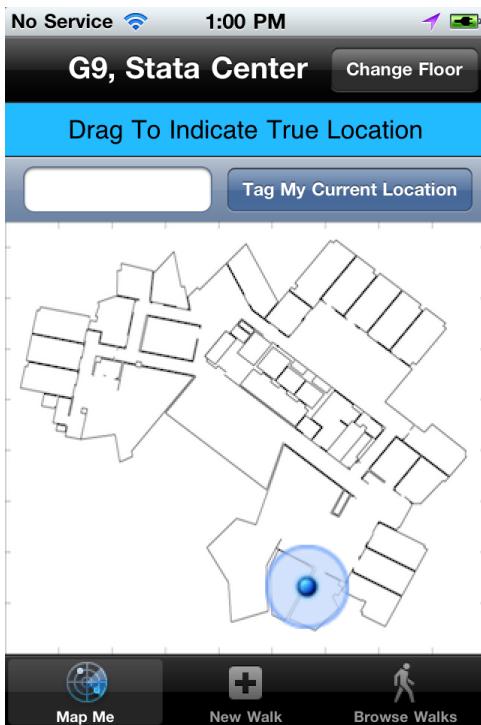
5.7 Implementation

We have implemented *iTrack* as an iPhone 4 application. The application samples the in-built accelerometer, gyroscope, and magnetic compass, and sends back data over the internet to a central server which runs a particle filter to perform trajectory matching, as shown in Figure 5-5. The application also makes it easy to carry out the training procedure described in the previous section.

5.7.1 iPhone App

The application, shown in Figure 5-13 has four tabs:

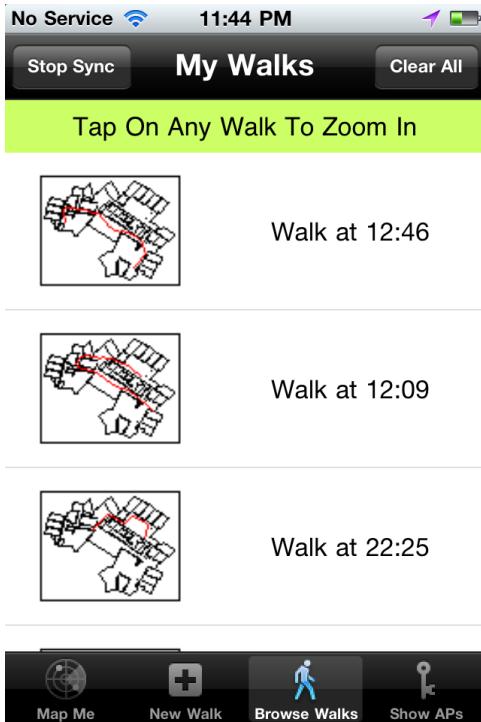
- A *Map Me* tab that shows the user their approximate (current) position on the floorplan using WiFi localization. Figure 5-13(a) shows a screenshot of this tab.
- A *New Walk* tab that allows a user to contribute a training walk in the dedicated training phase. Figure 5-13(b) shows a screenshot of this tab. It is almost identical to the *Map Me* tab, except that the blue dot in the centre can be moved around by the user to indicate her initial position to the system before walking. The user hits Start, puts the phone in her pocket, walks, and then hits Stop to send the data from the walk back to the central server.



(a) Tab showing current location.



(b) Tab to contribute training walks.



(c) Tab showing previous training walks.



(d) Tab showing nearby WiFi access points.

Figure 5-13: *iTrack* application for the iPhone.

- A *Browse Walks* tab that shows the user their previous walks. Each time the user collects a training walk, it shows up in this tab once the system has finished map-matching it. It also provides a UI to indicate whether or not the walk found was correct, to obtain user feedback. Figure 5-13(c) shows a screenshot of this tab.
- A *Show APs* tab shows nearby access points and their signal strengths. Figure 5-13(d) shows a screenshot.

5.7.2 Run Time of Map-Matching

The running time of *iTrack* is dominated by the run time of the particle filter since the steps to extract strides and angles are comparatively very fast, and linear in the size of the input sensor data.

The running time of the particle filter is upper-bounded by $O(MN)$, where M is the number of particles used for simulation and N is the number of strides in the user's walk. In practice, if M' is the threshold on effective weight for resampling, the running time is closer to $O(M'N)$ because approximately M' particles survive on average at each time step. It is possible to implement further optimizations that have been proposed in the particle filtering literature — for example, to track the convergence of the particle and adaptively reduce the number of particles simulated when the filter converges. We have not done so in our initial implementation.

We have implemented the *iTrack* particle filter in C++. The filter runs as a server side process that continuously listens for input from new walks. When a walk is complete on the phone side, the phone sends back inertial and WiFi data from the walk to the server, triggering the particle filter to run on the server side. We do not focus on server side scalability issues in our implementation (and this dissertation).

Our filter implementation uses $M = 100,000$ particles and a cutoff of $M' = 10,000$ to trigger SIR resampling. We have found this number to be necessary to achieve acceptable accuracy in our experiments. More particles give slightly more benefit, but incur excessive computational cost. As it stands, our C++ implementation runs approximately in real time on a 2.33 GHz Macbook Pro laptop with 3GB RAM, taking about a minute to match data from a minute of walking. While collecting training data, the trajectory matching can run in the background, so a user need not wait for it to complete before walking again.

5.7.3 Ideas For Performance Optimization

Taking a minute to match one minute of data is on the slow side if trying to scale *iTrack* to thousands of buildings or users. We are investigating the following performance optimizations in future work:

- Reducing the number of particles used, M , if the initial position and orientation are known more accurately (e.g. in steady state when a substantial WiFi training database has been built).
- Using known techniques from the particle filtering literature like adaptive resampling, which reduces the number of particles when the variance of the particle cloud falls below some threshold.
- Grouping together stretches of walk without significant turns, where the user walks straight, to reduce the effective number of steps of simulation. The most likely sub-trajectory within each such stride group can be solved for analytically. This is similar to Rao-Blackwellization, a state space factorization technique commonly used in the particle filtering literature [2].

- Parallelizing the particle filter by running each particle in parallel (though some coordination is required for periodic resampling).

5.8 Evaluation

We evaluate *iTrack* using a data set consisting of:

- 50 walks with known ground truth trajectories, collected on the 9th floor of the MIT Stata Center. Most of the walks were done with phone in the user's pocket.
- 97 walks without recorded ground truth information. These walks are mainly used to collect a dense data set of WiFi training data for comparison. These walks were collected from phones both in a user's hand and pocket.

We use tape markings made manually on the floor to determine precise ground truth for the walks. The markings are accurate to within less than a centimetre, but the accuracy of the ground truth itself depends on how faithfully the user followed the marked tape. We believe it should be well within a foot.

The key questions we answer in our evaluation are:

- What is the absolute accuracy of *iTrack* when mapping a person's trajectory indoors? How does this compare to WiFi localization without a gyroscope?
- How much seed WiFi training data is required for *iTrack* to work well? How much time and effort does it take to collect this seed data?
- How well does *iTrack* work in dedicated training mode when initial position and/or orientation are known?
- When used for crowd-sourcing, to what extent does *iTrack* help reduce manual training effort compared to marking points manually on a map?
- Are the specific models of angle and stride error used in *iTrack* better than using simple Gaussian models?

We do *not* answer the following (arguably important) questions in this dissertation:

- How much does the accuracy of *iTrack* depend on the layout, shape and size of the building floorplan? One would expect *iTrack* to perform better in floorplans with tighter corridors and more turns, and worse in open spaces where there is a lot of ambiguity, and a given shape can map to many different legal trajectories.
- How much does the accuracy of *iTrack* depend on the density and number of WiFi access points (relative to floorplan area)?

The above questions are interesting, and we plan to evaluate them going forward as we deploy *iTrack* on more floors and in different buildings.

5.8.1 Key Findings

We summarize the key findings of our evaluation below:

- *iTrack* is extremely accurate at mapping trajectories indoors given a small amount of seed WiFi training data collecting using our dedicated process. In our experiments, approximately 5 minutes worth of seed training data (4 training walks) resulted in a mean trajectory mapping error of 3.1 feet on our test data set. An error of 3.1 feet is 5-6× more accurate than using just WiFi localization in our building, which has a mean localization error of 5.4 metres.
- The recall (survival rate) of the algorithm is approximately 82%.
- The mixture models we use for stride length and angle error yield a significant improvement over the Gaussian models used in previous work in terms of recall. They improve the survival rate from 64% to 82% and 68% to 82% respectively.
- The dedicated training procedure we use is accurate and quick, with mapping errors smaller than 3 feet and a survival rate of 88% when using start position and walking direction information to initialize the particle filter.
- Using multiple passes of iteration for crowd-sourcing has significant accuracy benefits. The reduction in median error is not very big (from 3.5 feet to 3.1 feet), but the second pass reduces *tail* error significantly, correcting some of the worst mistakes in the first pass.

The rest of this evaluation is organized as follows. Section 5.8.2 describes the ground truth setup we use, and Section 5.8.3 describes the procedure we use for collecting walking data. Section 5.8.4 explains the evaluation metrics we use. The subsequent sections are devoted to answering the key questions posed earlier, starting with evaluating the absolute accuracy of *iTrack* (Section 5.8.5).

5.8.2 Ground Truth

Recall that we used a cleaned up form of GPS as ground truth when evaluating both *VTrack* and *CTrack*. Since GPS does not work indoors, indoor location ground truth is inherently harder to obtain and it is difficult to obtain a large amount of ground truth data.

We used a relatively “low-tech” approach to obtaining ground truth in which we manually placed tape markings on the 9th floor of the MIT Stata Center. The carpet is laid out in uniform square grids, making it relatively easy to place these markings at regular intervals. Figure 5-14(a) illustrates a section of the floor instrumented with these tape markings.

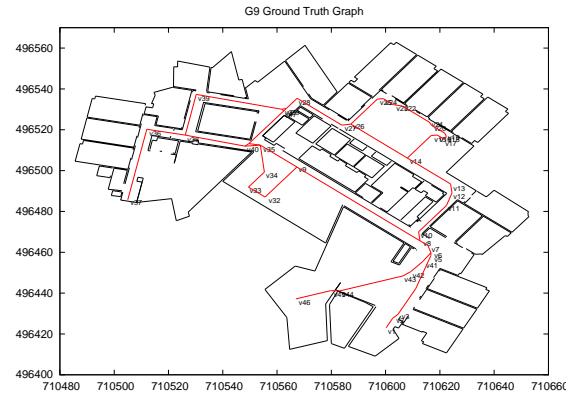
We placed about 150 such markings on the ground, of which 46 were designated as *vertices*. They satisfy the property that any of the markings lies on a line segment joining two vertices. Thus, all the markings lie on the edges of a *ground truth graph*. Figure 5-14(b) shows this graph on a floorplan of the building.

We manually measured the exact (x, y) coordinates of each of the 46 vertices of the graph using measuring tape to compute the distance from walls and/or other vertices in the graph, and known coordinates of the walls obtained from the floorplan architectural data.

The 50 “labeled” walks we used for evaluation were all collected by deliberately walking only along the contours defined by the markings. As soon as each walk was completed, a human being recorded



(a) Photograph of ground truth setup.



(b) Ground truth graph.

Figure 5-14: Ground truth setup for evaluating *iTrack*.

the sequence of vertices traversed in the walk. For each labeled walk, we are now guaranteed that the true trajectory must traverse exactly the sequence of vertices recorded in the ground truth log. Each walk has a steady pace and does not include stops or interruptions. However, some walks have sharp or sudden turns, including U-turns.

However, since we did not record any timing information for the walks, we do not know exactly *when* these vertices are crossed in the course of the walk. To recover this information very accurately, we use the (known) fact that the gyroscope data from the walks can precisely pinpoint the times of major turns in each walk. We ran a constrained version of our particle filter that *only explores particles that traverse the known sequence of vertices in the known order*. This helps align the turn times from the gyroscope exactly to the sequence of vertices and obtain accurate crossing times for each vertex. We manually verified that this was indeed the true crossing time for *each vertex in each of the 50 walks*. While this was a painstaking process that took about a day of repetitive work, we did this to avoid biasing the ground truth by mistakes made by the constrained particle filter. The result is a highly clean, reliable set of 50 ground truth walks, including timing information, that we can trust to be accurate.

5.8.3 Walk Data Collection

We collect each of the 50 ground truth walks using the following procedure. We go to a randomly chosen vertex in the ground truth graph (any vertex in the set v_1, v_2, \dots, v_{46} in Figure 5-14(b)). The user holds the phone outstretched in his palm pointed towards the initial direction of walking and marks his approximate location on the phone UI shown in Figure 5-13(b). He then hits a “start” button to start logging inertial sensor and WiFi data, and places the phone in his left or right pocket, or in his hand (most of our walks were collected from users’ pants pockets). He walks along the contours defined by the markings on the floor and finally stops at some end vertex. When done, the user takes the phone out of his pocket and hits a “Stop” button to send data back to the *iTrack* server.

While we recorded the starting position and orientation of each walk during data collection for convenience, some of our experiments do not use this information. We evaluate the performance of *iTrack* both with and without knowledge of the initial position and walking direction, as the following sections describe.

The next section defines the evaluation metrics we use throughout the rest of the evaluation.

5.8.4 Evaluation Metrics

We use two key metrics in our evaluation: *localization error* and *survival rate*.

Localization Error

Assume we have a ground truth trajectory G consisting of a sequence of the form:

$$\langle t_1^G, x_1^G, y_1^G \rangle, \langle t_2^G, x_2^G, y_2^G \rangle, \dots, \langle t_m^G, x_m^G, y_M^G \rangle,$$

and assume we are trying to evaluate a strategy that produces an output trajectory S :

$$\langle t_1^S, x_1^S, y_1^S \rangle, \langle t_2^S, x_2^S, y_2^S \rangle, \dots, \langle t_m^S, x_m^S, y_N^S \rangle,$$

Note that that it is possible that $N \neq M$.

We compute the localization error for each output location sample $\langle t_i^S, x_i^S, y_i^S \rangle$ by linearly interpolating the available ground truth locations to find the true location $\langle x^G, y^G \rangle$ at time t_i^S . The localization error in this sample is given by:

$$Err_i = \sqrt{(x_i^S - x^G)^2 + (y_i^S - y^G)^2} \quad (5.21)$$

The experiments in our evaluation show both mean values of the error Err_i over entire trajectories, as well as probability distributions of this error.

Survival Rate

As mentioned in Section 5.5.4, the particle filter in *iTrack* does not always succeed in producing an output trajectory on all of its runs. When errors in the input sensor data are too extreme to be corrected by the filter, it sometimes happens that all the particles being simulated hit walls and die, leaving no valid trajectory. If computational power or latency is not a constraint, it is possible to re-run the simulation with a larger number of particles (we normally use 100,000 particles) but this too may fail to produce valid output if the error in shape is too extreme.

For this reason, we measure a second metric: the *survival rate* of a strategy. This refers to the proportion of walks that produce some answer when run through our standard particle filter with 100,000 particles (chosen for its reasonable running time). A low survival rate indicates low recall accuracy. For example, a 50% survival rate would mean that 50% of the walks fail to produce any trajectory match when run through the particle filter.

5.8.5 Absolute Accuracy of *iTrack*

Figure 5-15 illustrates a CDF of the mean localization error over the test walks when using *iTrack*. The *iTrack* algorithm in this experiment was seeded with WiFi training data from 4 walks. The 4 walks were randomly selected from the set of 97 walks collected without ground truth, and did not overlap with the test set of 50 used for evaluation. The 4 walks amount to all of 150 WiFi samples, and required only 3-4 minutes of training time for one person to collect using the iPhone app. The particle filter in this experiment was not given the initial position or walking direction of the user.

The error shown in the graph was computed by finding the mean localization error over all the output location samples in a walk. The CDF for *iTrack* was computed across 41 of the 50 test walks (9 failed to produce valid output as described below). As a reference point for comparison, the figure

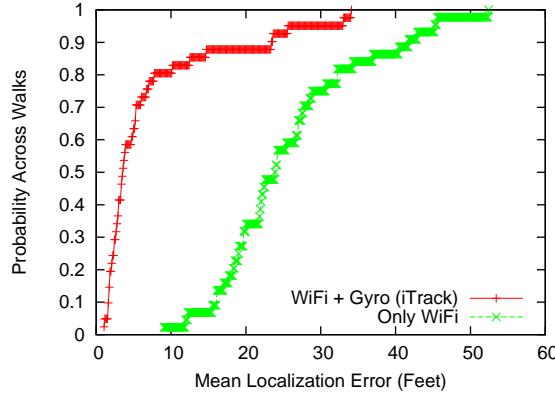


Figure 5-15: Absolute accuracy of *iTrack* compared to WiFi localization.

also shows a CDF of the mean localization error when using the seed WiFi data to find the user's track (with a *CTrack*-like algorithm).

As can be seen, using inertial data from the gyroscope results in significantly more accurate tracks than using WiFi localization for trajectory mapping. The median error with *iTrack* is 3.5 feet, 5-6× smaller than the median error with only WiFi, which is of the order of 5.4 metres. *iTrack* also yields significant benefits in the tail, reducing the 80th percentile error by a factor of over 3× and the 90th percentile error by over 1.5× compared to WiFi.

Survival Rate. We found that *iTrack* seeded with the 4 walks selected above had a survival rate of 82%, with 41 of the 50 test trajectories producing some output. 9 of the test walks failed to produce any output.

5.8.6 Impact On Training

In this section, we show that *iTrack* is nearly an order of magnitude faster and easier at building a WiFi training map for indoor localization than using manual training. It is important to be able to localize or track a phone using WiFi alone because not all devices have gyroscopes, and a phone is not always guaranteed to be fixed in a user's pocket or hand when being localized in real-life, in which case the gyroscope cannot be used to track it.

The benefits from *iTrack* over manual training are two-fold:

- A medium to long walk can cover a large number of points in one go, eliminating repeated requests to the user to mark her location on a UI.
- It is potentially less error-prone than manual training, since the accuracy of training coordinates does not depend on a user marking the location correctly on the UI.

Time And Effort Advantage

We performed a simple experiment to compare the time required to collect a single training walk in *iTrack* to collecting multiple points along the walk manually. The walk we performed took a total time of 41 seconds in *iTrack* — including hitting a “start” button on the phone UI, walking with the phone, and hitting a “stop” button to send back data after the walk. We found that WiFi

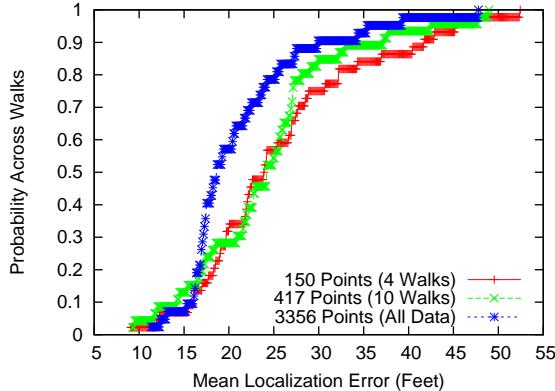


Figure 5-16: More WiFi training data reduces localization error.

scanning along the path collected 28 WiFi training points (this is smaller than the 41 second walk time because scans must be separated by more than one second to yield meaningful, non-cached results).

We also collected 28 training points along the walk using manual training. To do this, the user progressively advances to a new point on the walk by taking a stride or two. He/she then drags the blue circle shown in Figure 5-13(a) to the approximate position of the point manually, hits a “log” button and stops momentarily (to allow time for a WiFi scan to occur at that position). He/she then advances to the next point and repeats the procedure.

The total time for manual training on the walk we collected was 215 seconds, which is about $5.2 \times$ more than training with *iTrack*. The time savings are even more for longer walks. Moreover, the savings of $5.2 \times$ are just for the seed data, and do not even factor in *iTrack*’s powerful ability to *crowd-source* data in the background once seed data has been collected. If factoring in crowdsourcing, *iTrack* would be orders of magnitude quicker than manual training and make it possible to collect more data than ever possible with manual training.

Collecting additional data is important, because can significantly improve the accuracy of WiFi localization. To illustrate this, Figure 5-16 shows the accuracy of WiFi localization (implemented using a RADAR-like approach [69]) for different amounts of training data collected on our test floorplan. We see that there is a significant reduction in both the median and higher quantiles of localization error as more WiFi training data is available to the system.

Accuracy Advantage

Both we and researchers on the OIL project at MIT [46] have found manual training to be error prone, in addition to being slow. The OIL project has even evolved techniques to mitigate some of this error. We could not evaluate this benefit quantitatively because we did not have access to a large corpus of manual training input from users.

5.8.7 How Much Seed Data Do We Need?

The discussion in Section 5.8.5 shows that *iTrack* can produce high quality trajectories for most input walks when seeded with a small amount of training data. We performed a simple experiment to quantify the impact of seed training data and to understand how much seed data is really required to achieve good mapping accuracy.

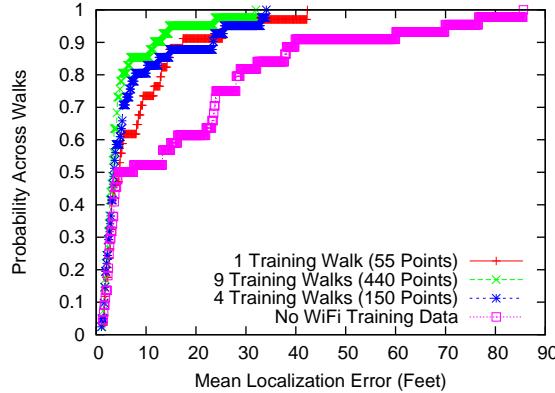


Figure 5-17: Accuracy of *iTrack* for different amounts of seed data.

Figure 5-17 shows the same CDF as shown in the absolute accuracy experiment in the previous subsection, but for varying amounts of seed WiFi training data — including the extreme of using no WiFi data at all and relying purely on inertial (accelerometer and gyroscope) data. As in the previous experiment, the seed data was drawn from one or more walks selected from the unlabeled pool of 97 walks.

The figure shows that it is essential to seed the algorithm with *some* WiFi data. Although the median error of *iTrack* without any seed WiFi data is close to the median error of the strategies with some seed WiFi data, the tail is considerably worse. The 90th percentile error without seed WiFi data can be as bad as 80 feet. Visual inspection also reveals that the output trajectories produced for these walks are completely or partially wrong.

We also see that the 80th and 90th percentile error drop dramatically even if adding seed WiFi data from just one walk (albeit one that covered a significant part of the floorplan). The errors drop further as more seed data is available. The graph indicates that 4-5 seed walks should be sufficient to ensure the median error is within 3.1 feet, which we believe to be good enough accuracy for many applications indoors (within room level or better). Using 9 training walks yields an improvement over 4 walks, but perhaps not enough to justify the extra manual training effort.

Table 5.2 shows the survival rate of the particle filter for the same scenarios shown in the CDF, as a function of amount of seed WiFi training data. Somewhat paradoxically, survival rate is highest when there is no seed WiFi data (88%). This can be explained by the fact that in the absence of any seed WiFi data, the system operates without any constraints whatsoever and so has greater leeway to find some particle that conforms to the measured shape, even if the shape has errors. Even one walk of training data significantly constraints the search space of the particle filter, and causes erroneous shapes to fail to survive the Monte Carlo simulation.

Interestingly, after the first walk, more training data improves the survival rate. We believe this is because more WiFi training data helps fix location more precisely, and allows more particles to explore a smaller area, making it more statistically likely that *some* particle survives the simulation.

The 4-5 number is obviously dependent on the particular floorplan we have evaluated on. This requirement is likely to vary depending on the floorplan geometry. It is likely to be more for a floorplan with larger area where 4-5 walks may not span all the major areas of a floor, and likely to be smaller in a smaller area where 2 or 3 walks may be enough.

Seed Training	Survival Rate
None	88%
1 Walk (55 Points)	68%
4 Walks (150 Points)	82%
9 Walks (440 Points)	82%

Table 5.2: Survival rate of *iTrack* for different amounts of seed data.

The next section measures the accuracy of *iTrack* when using our dedicated training procedure to collect the 4-5 seed WiFi walks.

5.8.8 How Accurate Is Dedicated Training?

As we have seen, *iTrack* requires a few seed walks to be collected as dedicated training data before it can crowd-source data without any user input. The walks used for dedicated training are all collected by asking the user to specify their approximate initial position. The system also infers approximate initial walking direction using the in-built phone compass as we have described earlier. Some of the initial positions may have human error but they are all approximately correct. In this experiment, the algorithm does not use any seed WiFi data in the training phase — though in practice, it could conceivably begin using its own WiFi data as soon as it matches a single walk.

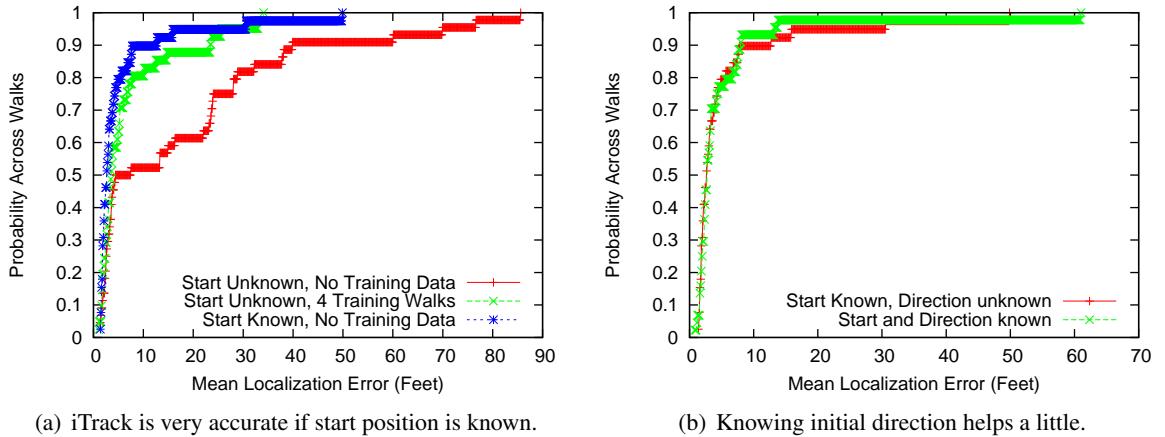


Figure 5-18: Knowing initial position and/or orientation improves accuracy.

Figure 5-18(a) shows a CDF of the mean localization error over our test walks when the algorithm knows the approximate initial position. The graph also shows two of the CDFs presented earlier as references for comparison — CDF of error when running the algorithm both without knowledge of initial position and without seed WiFi data, and CDF of error when running the algorithm with seed WiFi data, but without knowledge of initial position.

The figure shows that that median, 80th and 90th percentile of localization error are all within 1 metre, comparable or smaller than the case where we use 4 seed WiFi training walks. This is because knowledge of the initial position fixes the approximate region of the floorplan where the user has walked and eliminates the structural ambiguity owing to the same shape being permissible in different parts of the floorplan.

We now look at whether knowledge of initial walking direction is beneficial. This requires a little extra effort because a trainer needs to consciously remember to keep the phone flat in his/her hand

Algorithm Configuration	Survival Rate
Start Known, No Training	78%
Start+Angle Known, No Training	88%

Table 5.3: Survival rate of *iTrack* for different initialization configurations.

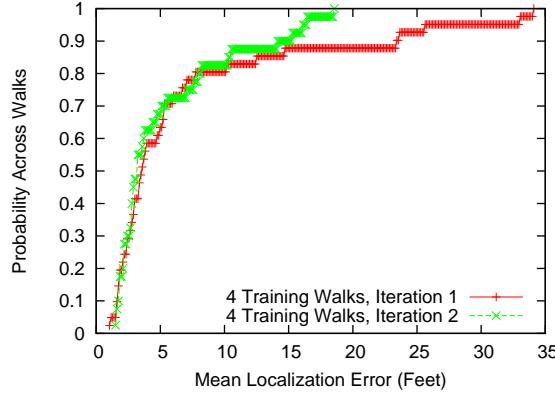


Figure 5-19: Iterative training can improve accuracy of *iTrack*.

towards the direction of walking while marking initial location on the map. Figure 5-18(b) shows a CDF of the mean localization error over the test walks with and without knowledge of the initial orientation. The orientation information helps overall. Interestingly, a close look at the graph will reveal that one of the walks in the tail is actually *hurt* by using initial orientation information: the magnetic compass happened to produce an incorrect heading on this walk, confusing the algorithm.

Survival Rates. Table 5.3 shows the survival rates with knowledge of initial position and orientation. This shows that with knowledge of initial position and orientation, the survival rate is high enough (close to 80%) to make training not too much a hassle. If the survival rate were lower, then collecting 4-5 valid seed walks would require walking much more than 4-5 times.

To summarize, the CDF of error shows that knowledge of initial orientation helps, but just the win in terms of lower error is not big enough in quantitative terms to justify complicating the training procedure. However, the improvement in survival rate is seen to be significant. Hence, we have adopted this optimization in our implementation of *iTrack*.

5.8.9 Impact Of Iteration

This section evaluates the benefit of using multiple passes of iteration to incorporate WiFi training data in the form of a feedback loop to improve the *iTrack* algorithm, as described in Section 5.6.

Figure 5-19 shows a CDF of the mean localization error across our test walks for the first pass and second pass of *iTrack*. The first pass uses only 4 seed WiFi walks to map the trajectory. The output of the first pass is incorporated into the training database and used in the second pass as feedback. The figure shows that while iteration does not improve the median accuracy by much, it helps significantly in the tail, helping correct some of the worst mistakes in mapping made by the first pass of *iTrack*.

The survival rate of the second pass was 80%, compared to a survival rate of 82% on the first iteration (one trajectory that survived the first pass — probably an incorrectly matched one — failed

Algorithm Configuration	Survival Rate
Regular <i>iTrack</i>	82%
Gaussian Stride Model	68%
Gaussian Angle Error Model	64%

Table 5.4: Survival rate of *iTrack* for different error models.

to survive the second pass). We did not find additional passes of the algorithm to yield substantial benefit.

5.8.10 Impact Of Non-Gaussian Models

In this section, we drill down into the choice of stride length and angle error models used in *iTrack*'s particle filter. We elaborate on one of the key contributions of *iTrack*, showing that the non-Gaussian models used by *iTrack* are a significant improvement over the Gaussian models used in previous work [67]. While the non-Gaussian models do not reduce the localization error of walks output by the system, they increase the survival rate and hence the recall of the system significantly, making it possible to map more walks correctly.

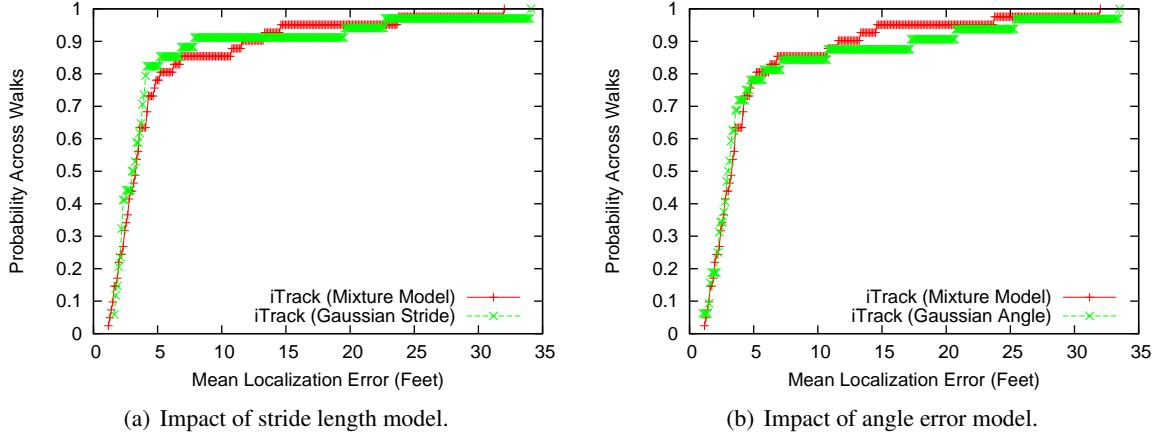


Figure 5-20: Angle and stride length models do not affect localization error.

Figure 5-20(a) shows a CDF of localization error using *iTrack*'s non-Gaussian stride model compared to a version of *iTrack* that uses a Gaussian stride model (assuming 9 training walks). Similarly, Figure 5-20(b) compares *iTrack*'s non-Gaussian angle model compared to a Gaussian model. The figures show that the choice of model has little or no influence on the localization error.

However, Table 5.4 shows that the choice of error model has a significant impact on the survival rate: the proportion of walks for which the particle filter finds one or more surviving valid particles. The non-Gaussian models we use both significantly increase the survival rate compared to a Gaussian model.

The explanation behind these results is that the walks that users perform tend to fall into one of two categories:

- Walks with little or no sensor error, whose true shape is almost exactly the observed shape from the gyroscope and whose true stride lengths match the average stride length of a human well (neither too slow nor fast).

Technique	99th Pctile. Error (metres)
WiFi Only	18.9
<i>iTrack</i> , Start Known	12.5
<i>iTrack</i> , 4 Seed Walks	15.6
<i>iTrack</i> , 9 Seed Walks	12.9
<i>iTrack</i> , 9 Seed Walks + Iteration	9.6

Table 5.5: Worst case (99th) percentile localization errors of different approaches.

- Walks with significant gyroscope sensor error, or portions with slow or fast walking, and hence a sequence of strides much smaller or longer than average.

The walks in the first category are easy to match and the particle filter always manage to find the true match irrespective of the choice of proposal distribution used to explore its state space. Even if the choice of proposal distribution is Gaussian, the deviations of both angle and stride from mean are small enough that some subset of particle(s) in the Monte Carlo simulation experience these deviations, and converge to a valid answer without violating a constraint in the floorplan.

The walks in the second category are the ones that are problematic for a Gaussian error model. Consider a walk where the user walked a sequence of steps that are much shorter than her average stride length to traverse a corridor or other constrained region of the floorplan. A Gaussian model for length will require a huge number of particles (exponential in the deviation from the mean) to encounter this possibility in a Monte Carlo simulation, and hence will never be able to converge to a valid path without always violating a wall or other constraint. The non-Gaussian model we use *does* explore such sequences and can direct the exploration of the particle filter (via periodic resampling) to the more fruitful parts of the search space that have a chance of satisfying the constraint and staying within the corridor.

A similar logic applies to errors in turn angle.

5.8.11 Robustness In The 99th Percentile

To gain a better understanding of the worst case errors with *iTrack*, we compared the *99th percentile error* of localization estimates for different configurations of *iTrack*, as well as for trajectory mapping using simple WiFi localization alone. Since we have only 50 walks, the number we report is the 99th percentile of point localization error across all the raw data points from the walks (rather than across the averaged errors over entire walks).

Table 5.5 shows the results. We see that with sufficient seed data and if using the iterative training procedure (Section 5.6), the 99th percentile error is smaller than 10 metres. This shows that *iTrack* is pretty robust even in the worst case.

5.9 Related Work

Indoor positioning and tracking are widely explored problems, with a large number of previous solution efforts spanning specialized ultrasound-based tracking systems, wireless and cellular location systems, inertial sensors, laser range finders, and many more.

The most closely related previous work to *iTrack* is a class of algorithms for pedestrian tracking with foot-mounted inertial sensors, such as the NavShoe [29], FootSLAM [72] and work from

the University of Cambridge using an XSens Mtx inertial measurement unit [67]. Like *iTrack*, all of these techniques use accelerometer and gyroscope data in conjunction with particle filtering to determine a user’s trajectory indoors. *iTrack* differs from this work in three main ways:

- First, these systems all use a specialized technique called zero-velocity updates (ZUPTs) to control the drift from inertial integration. This technique is peculiar to foot-mounted sensors which experience a point of zero velocity when a user’s foot rests on the ground during her stride, and *is not applicable to mobile phones*. *iTrack* uses a different approach to dealing with drift based on step counting, and finding peaks and valleys in gyroscope data.
- Second, most of these previous techniques assume a Gaussian error model for stride length and angle in the particle filter they use. While this may work for specialized foot mounted IMUs like the Xsens Mtx, it does not work for a mobile phone jiggling around in a user’s pocket. As we have shown, using a non-Gaussian model for stride length and angle error results in significantly improved recall for our particle filter.
- *iTrack* integrates WiFi localization information to improve its trajectory mapping accuracy. While some foot-tracking systems such as [67] have also proposed to use WiFi for initialization, it is not used throughout the mapping process, and in the iterative fashion that *iTrack* uses WiFi. As we have shown, a small amount of seed WiFi data is essential to achieving acceptable accuracy, and *iTrack*’s multi-pass iterative approach is effective at bootstrapping accurate indoor location from this data.

Inertial navigation itself is an old idea, having been used in robotics and in satellite tracking systems for well over four decades. For an excellent description of the core inertial navigation logic used in *iTrack* and a more rigorous derivation of the inertial navigation equations, see [26].

WiFi and cellular localization indoors is a well-studied topic. RADAR [69] was one of the pioneering systems in this field, proposing to locate a wireless receiver by finding the closest matching fingerprint(s) in a training database and computing their centroid. The Placelab project [57] used a similar approach indoors using a specialized GSM receiver than can receive *wide* GSM fingerprints. The Horus system [58] uses a probabilistic algorithm to improve the accuracy of WiFi localization compared to the RADAR or Placelab approaches. A number of commercial efforts and startups also exist offering indoor WiFi localization solutions including AlphaTrek [12], a commercial spin-off of Horus, and Ekahau [30]. *iTrack* achieves less a metre median tracking accuracy, similar to Horus [58], but with significantly simplified training compared to Horus.

iTrack falls into a class of indoor localization systems that aim to reduce training effort by using crowd-sourcing. Previous work in this area includes Redpin [70], Oil [46] and EZ [49]. Redpin and Oil both include user interfaces that allow a user to mark her location on an indoor map, and then scan for WiFi fingerprints to build a training database from user input. Oil includes an optimization to allow the user to indicate that he/she is static at a given location, so that it can obtain extended amounts of WiFi training data at the same location. This is a useful optimization, and we have automated it in *iTrack* by using the accelerometer to detect that a person is not moving.

The major difference between *iTrack* and previous WiFi crowd-sourcing systems such as Redpin and Oil is that *iTrack* *requires much less manual user input than these systems*. The training process on each floor requires a few controlled walks to collect seed training data, which can be done within 5-10 minutes. Thereafter, as we have shown, the system can crowd-source walks from users walking

around within the space automatically whenever the phone is in the appropriate position (hand or pocket), without any manual user input to indicate when they are walking, the starting point of the walk or any other information. This enables *iTrack* to gather a large amount of WiFi training data extremely quickly and easily compared to other crowd-sourcing systems.

The EZ crowd-sourcing system [49] also aims to crowd-source WiFi localization data from users walking around within a space without manual user input, and thereby alleviate the training problem for indoor localization. The system requires a small amount of seed training data in the form of GPS points obtained occasionally when a user walking the space is near a window or entrance to the building, and uses the rest of the (unlabeled) WiFi data only to establish constraints on wireless propagation. The system produces localization estimates that are less accurate than standard WiFi localization with more training data (e.g., RADAR or Horus), and as such, is likely to be less accurate for trajectory mapping than *iTrack*. Moreover, the availability of GPS locks indoors is rare. We have never been able to obtain a GPS lock inside our building, for example.

Ultrasound location systems such as Cricket [62], specialized radio ranging systems such as Active Bat [3] and laser range finder-based systems such as the MIT Intelligent Wheelchair [80] can measure the location of an object very precisely indoors, to within a few centimetres. The Cricket and Active Bat systems use the time difference of arrival between directed ultrasound or wireless signals to position the receiver, much as GPS works outdoors. Laser range finders use laser beams to measure their distance to the nearest obstacle in all directions, and use a particle filter to solve for the most likely trajectory of an object given a pre-built training map or a floorplan that is to scale. While all these systems are more accurate than (or comparable in accuracy to) *iTrack*, unlike *iTrack*, they require specialized, expensive hardware and do not work with commodity smartphones.

5.10 Conclusion

This chapter described *iTrack*, a system for easy and accurate trajectory mapping indoors where GPS does not work. *iTrack* uses a probabilistic particle filter model to fuse inertial sensors and WiFi on a mobile phone to localize a user’s trajectory to within less than a metre indoors. *iTrack* is easy to train compared to previous WiFi localization systems, requiring a small amount of seed WiFi data for each floor of interest that can be collected within 5-10 minutes using a simple training procedure, assuming pre-existing availability of a digital floorplan. Once the seed WiFi data is available, *iTrack* uses a novel technique to crowd-source walks from users walking around in the space, *without any manual user input* unlike previous crowd-sourcing approaches. We have implemented *iTrack* on the iPhone 4 and found that it can map a user’s walk to within a metre of accuracy, with minimal training effort.

The key novel technical contributions of *iTrack* we described were:

- The use of step-counting and peak finding to extract shape data without experiencing drift due to integration error.
- The use of non-Gaussian angle and stride length models to improve particle filter survival rate and hence recall, for constrained environments.
- The use of multiple passes of iteration by incorporating WiFi data from the first pass of running *iTrack* as training data for a second pass, to establish a positive feedback effect.

There is significant room for future work. The less than one metre tracking accuracy holds only for trajectories specifically known to be extracted when a user is walking steadily with her phone in her hand or pocket, or other similarly fixed configurations. *iTrack* is not currently robust to arbitrary phone use, or to real-world user movement with arbitrary stops and interruptions. This leaves open the following question for future work:

Can we continuously localize a user accurately to within less than a metre or better indoors when the phone is in an arbitrary orientation, or not experiencing a steady walking motion?

This is a challenging research problem because a phone experiencing typical use experiences complex movements — a user may flip it from hand to hand, play games with it, lift it from his pocket, or take a call. Future algorithms will need a generalized, robust way to detect and filter out, or even use inertial sensor data from these time periods without affecting localization accuracy.

Other important areas for future work include developing ways of running the particle filter on the phone, optimizing energy consumption and performance, and studying the accuracy of *iTrack*-like approaches for a wider variety of building layouts and WiFi access point densities.

Chapter 6

Conclusion

This dissertation has described three different, but closely related systems that can determine the trajectory of a mobile phone from sensor data:

- *VTrack*, which works over geographic coordinates extracted from sparsely sampled GPS and noisy WiFi localization.
- *CTrack*, which operates directly on cellular tower sightings and their signal strengths.
- *iTrack*, which fuses data from the accelerometer and gyroscope of a mobile phone with WiFi access point signatures indoors.

A common thread in all of these systems is the use of probabilistic models to extract *accurate information from inaccurate and/or infrequently sampled raw data*. In particular, the systems we have presented have used two kinds of (closely related) models:

- **Hidden Markov Models**, which model a trajectory as a sequence of hidden states and use dynamic programming to solve for the most likely sequence of states, and
- **Particle filters**, which also model a trajectory as a sequence of hidden states, but in a continuous state space. Since an exact solution is not tractable, unlike HMMs, they use Monte Carlo simulation to approximately solve for the most likely state sequence.

Probabilistic models have proved to be a powerful tool in all three of the systems and have been the key to recovering useful information from the raw data, which on its own is often wrong, or not accurate enough for the applications of interest.

We believe a “one-liner” take-away lesson from this dissertation research is that *a careful choice of underlying model is critical*: a machine learning tool such as a HMM or a particle filter is only as accurate as the underlying probabilistic model used in it. It is important to select the state space, observation space, emission probability distribution and transition probability distributions carefully to get these models to work in practice, and a simple or naive choice can often result in poor accuracy or even the model completely failing to work.

Examples of careful choice of model surface in all the systems we have presented in this dissertation, some of which we recap below:

- *VTrack*'s use of ϵ -transition probabilities along all outgoing segments from an intersection, rather than the simple solution of assigning a probability of $\frac{1}{n}$ to each of n outgoing segments. This choice was crucial to avoid biasing in favour of paths with fewer outgoing intersections, and was non-obvious before we discovered this bias.
- *CTrack*'s use of cell towers and their signal strengths as the observations of choice, rather than using geographic coordinates obtained from these fingerprints as the inputs to the Hidden Markov Model. As we showed, this choice reduced trajectory matching error by $2\times$.
- *iTrack*'s use of non-Gaussian models for stride length and angle error resulted in much higher survival rate and recall than using a Gaussian model, because a Gaussian proposal distribution could not model the common case where a *sequence of strides* are all shorter than average, or one user has shorter strides than another owing to height differences.

There are many more examples the reader can find if he/she reads this dissertation in more depth.

In each of these cases, rather than the algorithms used for the HMM or particle filter themselves being important, it was the *choice of model that was crucial*. The simple or straightforward choice of model resulted in poor accuracy in each of the three cases above.

6.1 Future Work

We conclude this dissertation with an outline of directions for future work.

6.1.1 VTrack and CTrack

We describe two directions that we think are interesting for future work:

- **Using Inertial Sensors:** Given that mobile smartphones on the market are increasingly equipped with accurate gyroscopes, it would be interesting to apply the lessons learned from *iTrack* to build an extremely accurate, yet energy-efficient outdoor tracking and navigation system. In particular, is it possible to integrate gyroscope data to obtain the approximate shape of a person's trajectory outdoors, and fuse with cellular localization, thereby significantly reducing the trajectory matching and point localization error of *CTrack* from the current levels of 75% and 45 metres respectively? We think this could be a good “fit of opposites” because cellular localization is good at a macro scale but poor at obtaining a trajectory right at a micro scale. In contrast, inertial sensors are good at a micro scale but not so good at localizing a device on a macro scale.
- **On-Phone Algorithms:** This dissertation has proposed map-matching algorithms that run on a server, mainly owing to computational constraints. Is it possible to come up with lightweight versions of the HMM algorithms that can be run continuously on a phone without hogging on-phone CPU and energy? Running map-matching on the phone would be useful for situations where there is no internet connectivity, or where the extra latency of communicating with the cloud is prohibitive, or where privacy is a significant concern.

6.1.2 iTrack and Indoor Localization

We believe that *iTrack*, for all its sophistication, barely scratches the surface of what is possible to make indoor localization more accurate and easy. We mentioned the following big technical challenge at the end of the last chapter:

Can we use inertial sensors to continuously localize a user accurately to within less than a metre or better indoors when the phone is in an arbitrary orientation?

The key challenge, as mentioned earlier, is to develop new robust algorithms to detect and filter out data from time periods where a user is using a phone unpredictably, such as flipping the phone around or taking a call.

Given the increasing accuracy and sophistication of MEMS inertial sensor technology, we believe it is a question of when, not if this will happen. Solving this research problem would enable much more fine-grained applications of indoor positioning at a wide scale simply not possible on mobile phones today — such as knowing where a consumer is in a store at rack level, where a tourist is in a museum at the level of an individual exhibit, or where a doctor is in a hospital at the level of a patient bed.

On that optimistic note, we conclude this dissertation.

Bibliography

- [1] The Mobile Millenium Project. <http://traffic.berkeley.edu>.
- [2] A. Doucet, J. de Freitas, K. Murphy and S. Russel. Rao-Blackwellized Particle Filtering For Dynamic Bayesian Networks. In *UAI*, 2000.
- [3] A. Harter, A. Hopper, P. Steggles, A. Ward and P. Webster. The Anatomy Of a Context-Aware Application. In *Wireless Networks*, 2002.
- [4] A. J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. In *IEEE Transactions on Information Theory*, 1967.
- [5] A. Thiagarajan, J. Biagioni, T. Gerlich and J. Eriksson. Cooperative Transit Tracking Using GPS-enabled Smartphones. In *Sensys*, 2010.
- [6] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden and L. Girod. Accurate, Low-Energy Trajectory Mapping For Mobile Devices. In *NSDI*, 2011.
- [7] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo and J. Eriksson. VTrack: Accurate, Energy-Aware Road Traffic Delay Estimation Using Mobile Phones. In *Sensys*, 2009.
- [8] A. Varshavsky, E. de Lara, J. Hightower, A. LaMarca and V. Otsason. GSM Indoor Localization. *Pervasive Mobile Computing*, 3:698–720, December 2007.
- [9] Traffic In the United States: ABC Survey. <http://abcnews.go.com/Technology/Traffic/story?id=485098&page=1>.
- [10] American Community Survey. http://www.census.gov/newsroom/releases/archives/american_community_survey_acs/cb07-cn06.html.
- [11] Analog Devices AD9864 Datasheet: GSM RF Front End and Digitizing Subsystem. http://www.analog.com/static/imported-files/data_sheets/AD9864.pdf.
- [12] AlphaTrek Inc. <http://www.alphatrek.com>.
- [13] Analog Devices, Inc. *ADXL330: Small, Low Power, 3-Axis +/-3 g iMEMS Accelerometer (Data Sheet)*, 2007. http://www.analog.com/static/imported-files/data_sheets/ADXL330.pdf.
- [14] B. Hoh, M. Gruteser, H. Xiong and A. Alrabady. Enhancing Security and Privacy in Traffic-monitoring Systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

- [15] B. Hoh, M. Gruteser, H. Xiong and A. Alraby. Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking. In *CCS*, 2007.
- [16] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J. Herrera, A. Bayen, M. Annavarapu and Q. Jacobson. Virtual Trip Lines for Distributed Privacy-Preserving Traffic Monitoring. In *Mobisys*, 2008.
- [17] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A.K. Miu, E. Shih, H. Balakrishnan and S. Madden. CarTel: A Distributed Mobile Sensor Computing System. In *Sensys*, 2006.
- [18] B. N. Waber, D. O. Oguin, T. Kim, A. Mohan, K. Ara and A. Pentland. Organizational Engineering Using Sociometric Badges. In *Netsci*, 2007.
- [19] GPS and Mobile Handsets. <http://www.berginsight.com/ReportPDF/ProductSheet/bi-gps4-ps.pdf>.
- [20] Bureau of Transportation Statistics. <http://www.bts.gov>.
- [21] C.G. Claudel, A. Hofleitner, N. Mignerey and A.M. Bayen. Guaranteed Bounds on Highway Travel Times using Probe and Fixed Data. In *88th TRB Annual Meeting Compendium of Papers*, 2009.
- [22] C.G. Claudel and A.M. Bayen. Guaranteed Bounds for Traffic Flow Parameters Estimation using Mixed Lagrangian-Eulerian Sensing. In *Allerton*, 2008.
- [23] Event Handling Guide for iOS: Motion Events. http://developer.apple.com/library/iOS/#documentation/EventHandling/Conceptual/EventHandlingiPhoneOS/MotionEvents/MotionEvents.html#/apple_ref/doc/uid/TP40009541-CH4-SW1.
- [24] D. Pfoser, S. Brakatsoulas, P. Brosch, M. Umlauf, N. Tryfona and G. Tsironis. Dynamic Travel Time Provision for Road Networks. In *International Conference on Advances in Geographic Information Systems*, 2008.
- [25] D. Sperling and D. Gordon. *Two Billion Cars: Driving Toward Sustainability*. Oxford University Press, 2009.
- [26] D. Titterton and J. Weston. *Strapdown Inertial Navigation Technology*. 2005.
- [27] Digifit. <http://www.digifit.com>.
- [28] D.J. Turner, S. Savage and A.C. Snoeren. On the Empirical Performance of Self-Calibrating WiFi Location Systems. In *IEEE Conference on Local Computer Networks (LCN)*, 2011.
- [29] E. Foxlin. Pedestrian Tracking with Shoe-mounted Inertial Sensors. In *IEEE Computer Graphics and Applications*, 2005.
- [30] Ekahau Inc. <http://www.ekahau.com>.
- [31] Euler Angles. <http://mathworld.wolfram.com/EulerAngles.html>.
- [32] F.B. Abdesslem, A. Phillips and T. Henderson. Less is more: Energy-efficient Mobile Sensing with SenseLess. In *Mobiheld*, 2009.

- [33] F.V. Diggelen. GPS Accuracy: Lies, Damn Lies, and Statistics. In *GPS World*, 1998.
- [34] Getting a fix with your Garmin. <http://www.gpsinformation.org/dale/gpsfix.htm>.
- [35] Information On Human Exposure To Radiofrequency Fields From Cellular and PCS Radio Transmitters. <http://www.fcc.gov/oet/rfsafety/cellpcs.html>.
- [36] B. Hummel. Map Matching for Vehicle Guidance. In *Dynamic and Mobile GIS: Investigating Space and Time*, 2006.
- [37] I. Constandache, I. S. Gaonkar, M. Sayler, R.R. Choudhury and L. Cox. EnLoc: Energy-Efficient Localization for Mobile Phones. In *INFOCOM*, 2009.
- [38] I. Constandache, R.R. Choudhury and I. Rhee. CompAcc: Using Mobile Phone Compasses and Accelerometers for Localization. In *INFOCOM*, 2010.
- [39] iCartel. <http://icartel.net/icartel-docs/>.
- [40] Inrix Inc. <http://www.inrix.com/>.
- [41] Apple Q&A On Location Data. <http://www.apple.com/pr/library/2011/04/27Apple-Q-A-on-Location-Data.html>.
- [42] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden and H. Balakrishnan. The Pothole Patrol: using a Mobile Sensor Network for Road Surface Monitoring. In *Mobisys*, 2008.
- [43] J. Krumm. Inference Attacks on Location Tracks. In *Pervasive*, 2007.
- [44] J. Krumm, J. Letchner and E. Horvitz. Map Matching with Travel Time Constraints. In *SAE World Congress*, 2007.
- [45] J. Paek, J. Kim and R. Govindan. Energy-Efficient Rate-adaptive GPS-based Positioning for Smartphones. In *Mobisys*, 2010.
- [46] J. Park, B. Charrow, J. Battat, D. Curtis, E. Minkov, J. Hicks, S. Teller and J. Ledlie. Growing an Organic Indoor Location System. In *Mobisys*, 2010.
- [47] J. Proakis. *Digital Communications, 4th Edition*. McGraw Hill, 2001.
- [48] J. Yoon, B. Noble and M. Liu. Surface Street Traffic Estimation. In *Mobisys*, 2007.
- [49] K. Chintalapudi, A.P. Iyer and V. Padmanabhan. Indoor Localization Without the Pain. In *Mobicom*, 2010.
- [50] K. Lin, A. Kansal, D. Lymberopoulos and F. Zhao. Energy-Accuracy Trade-off for Continuous Mobile Device Location. In *Mobisys*, 2010.
- [51] L. Ravindranath, C. Newport, H. Balakrishnan and S. Madden. Improving Wireless Network Performance Using Sensor Hints. In *NSDI*, 2011.
- [52] L. Sweeney. k-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 2002.

- [53] M. Gruteser and B. Hoh. On the Anonymity of Periodic Location Samples. In *Pervasive*, 2005.
- [54] M. Gruteser and D. Grunwald. Anonymous Usage of Location-based Services through Spatial and Temporal Cloaking. In *Mobisys*, 2003.
- [55] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard and R. West. PEIR, The Personal Environmental Impact Report, as a Platform for Participatory Sensing Systems Research, 2009.
- [56] M. Quigley, D. Stavens, A. Coates and S. Thrun. Sub-meter Indoor Localization in Unmodified Environments with Inexpensive Sensors. In *IEEE International Conference on Intelligent Robots and Systems*, 2010.
- [57] M. Y. Chen, T. Sohn, D. Chmelev, D. Haehnel, J. Hightower, J. Hughes, A. Lamarca, F. Potter, I. Smith and A. Varshavsky. Practical Metropolitan-scale Positioning for GSM Phones. In *UbiComp*, 2006.
- [58] M. Youssef and A. Agrawala. The Horus Location Determination System. *Wireless Networks*, 14:357–374, June 2008.
- [59] MapMyRun. <http://www.mapmyrun.com>.
- [60] M.B. Kjaergaard, J. Langdal, T. Godsk and T. Toftkjaer. EnTracked: Energy-Efficient Robust Position Tracking for Mobile Devices. In *Mobisys*, 2009.
- [61] Meraki Inc. <http://www.meraki.com>.
- [62] N. B. Priyantha, A. Chakraborty and H. Balakrishnan. The Cricket Location-Support System. In *Mobicom*.
- [63] N. Balasubramanian, A. Balasubramanian and A. Venkataramani. Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications. In *IMC*, 2009.
- [64] N. Malviya, S. Madden and A. Bhattacharya. A Continuous Query System For Dynamic Route Planning. In *ICDE*, 2011.
- [65] Navizon Inc. <http://www.navizon.com>.
- [66] NAVTEQ Data. <http://navteq.com/about/data.html>.
- [67] O. Woodman and R. Harle. Pedestrian Localisation For Indoor Environments. In *Ubicomp*, 2008.
- [68] OpenStreetMap. <http://www.openstreetmap.org>.
- [69] P. Bahl and V. Padmanabhan. RADAR: An In-Building RF-based User Location and Tracking System. In *INFOCOM*, 2000.
- [70] P. Bolliger. Redpin: Adaptive, Zero-Configuration Indoor Localization Through User Collaboration. In *1st ACM International Workshop On Mobile Entity Localization and Tracking in GPSless Environments*, 2008.

- [71] P. Mohan, V. Padmanabhan and R. Ramjee. Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones. In *Sensys*, 2008.
- [72] P. Robertson, M. Angermann and B. Krach. Simultaneous Localization and Mapping for Pedestrians using only Foot-Mounted Inertial Sensors. In *Ubicomp*, 2009.
- [73] PNI Corporation. *MicroMag3 3-Axis Magnetic Sensor Module*. <http://www.sparkfun.com/datasheets/Sensors/MicroMag3%20Data%20Sheet.pdf>.
- [74] Reduced Cost and Complexity Expand Application Potential for Healthcare. <http://www.radianse.com/press-tipping-062304.html>.
- [75] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using Mobile Phones to Determine Transportation Modes. *Transactions on Sensor Networks*, 6(2), 2010.
- [76] S. Brakatsoulas, D. Pfoser, R. Salas and C. Wenk. On Map-Matching Vehicle Tracking Data. In *VLDB*, 2005.
- [77] S. Gaonkar, J. Li, R.R. Choudhury, L. Cox and A. Schmidt. Micro-Blog: Sharing and Querying Content through Mobile Phones and Social Participation. In *Mobisys*, 2008.
- [78] Skyhook. <http://www.skyhookwireless.com>.
- [79] Telit GE865 Datasheet. <http://www.telit.com/module/infopool/download.php?id=1666>.
- [80] The MIT Intelligent Wheelchair Project. <http://rvsn.csail.mit.edu/wheelchair>.
- [81] Y. Freund and R.E. Schapire. A Decision Theoretic Generalization of Online Learning and an Application to Boosting. In *EuroCOLT*, 1995.
- [82] Y. Wang, J. Lin, M. Annavaram, Q. Jacobson, J. Hong, B. Krishnamachari and N. Sadeh. A Framework of Energy Efficient Mobile Sensing for Automatic User State Recognition. In *Mobisys*, 2009.



Estimation of origin-destination matrix from traffic counts: the state of the art

Sharminda Bera¹, K. V. Krishna Rao^{2*}

¹ Research Scholar, Civil Engineering Department, IIT Bombay, India 400076

² Prof., Civil Engineering Department, IIT Bombay, India 400076

Abstract

The estimation of up-to-date origin-destination matrix (ODM) from an obsolete trip data, using current available information is essential in transportation planning, traffic management and operations. Researchers from last 2 decades have explored various methods of estimating ODM using traffic count data. There are two categories of ODM; static and dynamic ODM. This paper presents studies on both the issues of static and dynamic ODM estimation, the reliability measures of the estimated matrix and also the issue of determining the set of traffic link count stations required to acquire maximum information to estimate a reliable matrix.

Keywords: Origin-Destination matrix; Static origin-destination matrix; Dynamic origin-destination matrix; Traffic count stations.

1. Introduction

In developing countries, changes in the land-use and economic state of affairs require momentous transportation planning. One of the most crucial requirements for the transportation planning is on arriving at the traffic pattern between various zones through Origin-Destination matrix (ODM) estimation. Traditional methods of estimating ODM are through large scale sampled surveys like home interview survey, roadside interview and license plate method conducted once in every 1-2 decades. But in situations of financial constraints these surveys become impossible to conduct. And by the time the survey data are collected and processed, the O-D data obtained become obsolete. In consequence, from 1970s many models have been proposed and widely applied for updating/estimating an old/sampled matrix using the current data of traffic counts collected on a set of links. The accuracy of these estimated matrix depends on

* Corresponding author: K. V. Krishna Rao (kvkrao@civil.iitb.ac.in)

the estimation model used, the input data errors, and on the set of links with collected traffic counts.

From the past studies, the ODM estimation models can be categorized as static and dynamic based on its application. In static methods the traffic flows are considered as time-independent and an average O-D demand is determined for long-time transportation planning and design purpose. Whereas from last two decades different dynamic approaches are proposed which are meant for short-term strategies like route guidance, traffic control on freeways, intersections etc. The assignment matrix which provides an approximate trip proportions based on the route choice behavior of the trip makers is the complicated part in the estimation problem. Another aspect, on which the reliability of the estimated ODM largely depends, is the optimum traffic counting locations. The traffic counts collected should provide as much traffic information as possible saving subsequent manpower requirement in data collection. Various rules are proposed in literature to select the traffic counting location points. This paper gives a vast study on the ODM estimation and the various related issues. It gives ideas about the various methodologies developed till date, the optimization algorithms used to solve the problem, convergence problems of those algorithms and most vital issue the reliability or the accuracy of the estimated ODM.

Though there are some softwares which can estimate ODM. But it is always wise to have coded algorithms which can be flexible with respect to the type of data available according to the study area. Still research is going on to estimate a reliable ODM efficiently. A comprehensive state-of-the-art review has been conveyed through this paper. Also the studies based on the optimum number of traffic counting locations required and their influences on estimated ODM are covered. This can help to acquire a good knowledge regarding the various developed models and also gives a direction for future research in this area.

The paper is organized as follows. Section 2 discusses about the static ODM estimation, its problem formulation and reviews the various methodologies developed till date. Discussions on dynamic and time-dependent ODM estimation are presented in Section 3. It encloses the models developed from past two decades. Section 4 gives a brief description of how the reliability of the estimated ODM is measured. The studies based on determining the traffic counting location points and its optimum number are presented in Section 5. Lastly Section 6 gives the final conclusions and the directions for future research.

2. Static ODM estimation from traffic counts

A static ODM estimation problem does not consider the time-dependent traffic flows and is assumed to represent a steady-state situation over a time period. The average traffic counts are collected for a longer duration to determine the average O-D trips. Let a transportation network is defined by W O-D pairs and a set of L links with $A \subseteq L$ as the subset of links with the observed traffic counts. Considering following notations:

t_w is the number of trips of O-D pair $w, w \in W$

p_w^a is the proportion of trips O-D pair $w \in W$ traversing link $a \in A$

v_a is the expected link flow for the link $a \in A$

The link flows, the trips between O-D pairs and the proportional matrix is related by the formulation given by eqn. (1).

$$\sum_{w \in W} p_w^a t_w = v_a \quad a \in A \quad (1)$$

This estimation problem is underspecified as A is less than W ODs and there is no unique solution. So, additional information (a prior or a target or a sampled matrix) is needed to determine a unique trip matrix.

2.1 Travel Demand Model Based Methods

Initially the researchers tried to relate the trip matrix as a function of models (like the gravity models) with related parameters. Some of the researchers like Robillard (1975), Hogberg (1976) used Gravity (GR) model based approaches and some (Tamin and Willumsen, 1989; Tamin et al, 2003) used Gravity-Opportunity (GO) based models for estimating ODM. These techniques require zonal data for calibrating the parameters of the demand models. The main drawback of the gravity model is that it cannot handle with accuracy external-external trips (refer Willumsen, 1981). The fundamental model for estimating the ODM based on traffic counts in this approach is given by combining eqn. (1) for trip purpose m with the standard gravity model,

$$v_a^m = \sum_m \sum_i \sum_j O_i^m D_j^m \cdot A_i^m \cdot B_j^m \cdot f_{ij}^m \cdot p_{ij}^{am} \quad (2)$$

where v_a^m gives the flows in particular link a for trip purpose m , O_i^m and D_j^m are the trips produced from zone i and attracted by zone j respectively for trip purpose m , f_{ij}^m here is the deterrence function, A_i^m and B_j^m are the balancing factors and p_{ij}^{am} is the proportion of trips traveling from zone i to zone j using link a for trip purpose m . Tamin and Willumsen (1989), for the calibration of unknown parameters, considered three estimation methods viz. non-linear-least-squares (NLLS), weighted-non-linear-least-squares (WNLLS) and maximum likelihood (ML). The GO models found to consume more time than GR model and does not guarantee the reliability of the estimated matrix.

2.2 Information Minimization (IM) and Entropy Maximization (EM) Approach

EM and IM techniques are used as model building tools in transportation, urban and regional planning context, after Wilson (1970) introduced the concept of Entropy in modeling. The entropy-maximizing procedure analyzes the available information to obtain a unique probability distribution. The number of micro-states $W\{t_w\}$ giving rise to meso-state t_w is given by,

$$W\{t_w\} = \frac{TN!}{\prod_w t_w!} \quad (3)$$

where TN is the total number of trips and t_w is the trips of O-D pair $w \in W$. With the information contained in the observed flows and with other available information, EM is used by Willumsen (1978) and IM approach by Van Zuylen (1978) (refer Van Zuylen and Willumsen, 1980) to estimate trip matrix. Van Zuylen and Willumsen (1980) have shown that both the methods are same except the unit of observation. Van Zuylen's model uses a counted vehicle and Willumsen's a trip. Both the models results in multi-proportional problem. But the convergence of the algorithm has not been proved. Van Zuylen and Branston (1982) further extended the study of Van Zuylen and Willumsen (1980) considering the inconsistency in traffic counts for the case when there is more than one count available on some or all links of the network. Unfortunately, IM and EM based models have the disadvantage of not considering the uncertainties in traffic counts and prior matrix which can be erroneous and can influence the output.

2.3 Statistical Approaches

Several models have been presented in order to estimate or to update ODM from traffic counts for the networks without congestion and with congestion effects via parametric estimation techniques like; Maximum Likelihood (ML), Generalized Least Squares (GLS) and Bayesian Inference (BI).

The method of **ML** is one of the oldest and most important in estimation theory. The ML estimates are the set of parameters which will generate the observed sample most often. In ODM estimation maximum likelihood approach maximizes the likelihood of observing the target ODM and the observed traffic counts conditional on the true trip matrix. The data consist of $\hat{\mathbf{V}} = [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_a]'$ representing a vector of a set of observed traffic counts, and a set of sampled O-D flows $\mathbf{N} = [n_1, n_2, \dots, n_w]'$. The likelihood of observing two sets of statistically independent data is expressed as

$$L(\mathbf{N}, \hat{\mathbf{V}} | \mathbf{T}) = L(\mathbf{N} | \mathbf{T}) \cdot L(\hat{\mathbf{V}} | \mathbf{T}) \quad (4)$$

The sampled O-D flows may be assumed to follow Multivariate normal distribution (MVN) or Poisson probability distribution. This is dependent on small sampling fractions α . Consider observed link counts on a set of A links and vector $\mathbf{T} = [t_1, t_2, \dots, t_w]'$ representing the populated trip matrix with elements t_w .

For the logarithm of the probability $L(\mathbf{N} / \mathbf{T})$ we have:

$$\ln L(\mathbf{N} / \mathbf{T}) = -\frac{1}{2}(\mathbf{N} - \alpha \mathbf{T})' \mathbf{Z}^{-1} (\mathbf{N} - \alpha \mathbf{T}) + const. \quad (MVN) \quad (5)$$

$$\ln L(\mathbf{N} / \mathbf{T}) = \sum_{w \in W} (-\alpha t_w + n_w \ln(\alpha t_w)) + const. \quad (Poisson) \quad (6)$$

where \mathbf{Z} is the covariance matrix of \mathbf{N} . If the observed traffic counts are also assumed to be generated either by a MVN distribution or Poisson probability distribution, then the similar expression for the probability $L(\hat{\mathbf{V}} / \mathbf{V}(\mathbf{T}))$ is obtained:

$$\ln L(\hat{\mathbf{V}} / \mathbf{V}(\mathbf{T})) = -\frac{1}{2}(\hat{\mathbf{V}} - \mathbf{V}(\mathbf{T}))' \mathbf{W}^{-1}(\hat{\mathbf{V}} - \mathbf{V}(\mathbf{T})) + const. \quad (MVN) \quad (7)$$

$$\ln L(\hat{\mathbf{V}} / \mathbf{V}(\mathbf{T})) = \sum_{a \in A} (\hat{v}_a \ln(V_a(\mathbf{T})) - V_a(\mathbf{T})) + const. \quad (Poisson) \quad (8)$$

where $V_a(\mathbf{T})$ denotes the flow volume on link $a \in A$, resulting from an assignment of \mathbf{T} and \mathbf{W} as the variance-covariance matrix. Hence ML function is the log-likelihood of the sum of eqns. (5) or (6) with (7) or (8).

Among a number of branches of regression analysis, the method of **GLS** estimation based on the well-known Gauss-Markov theory plays an essential role in many theoretical and practical aspects of statistical inference based model. The advantage of this approach is that no distributional assumptions are made on the sets of data and it allows the combination of the survey data relating directly to O-D movements with traffic count data, while considering the relative accuracy of these data (Bell, 1991a). Let vector $\hat{\mathbf{T}}$ denote the survey estimate of \mathbf{T} obtained from the grossing up the sample counts, independently of the sampling technique used. In GLS the following stochastic system of the equations in \mathbf{T} is considered:

$$\hat{\mathbf{T}} = \mathbf{T} + \boldsymbol{\eta} \quad (9)$$

$$\hat{\mathbf{V}} = \mathbf{V}(\mathbf{T}) + \boldsymbol{\varepsilon} \quad (10)$$

where $\boldsymbol{\eta}$ is the sampling error with a variance-covariance matrix \mathbf{Z} , and $\boldsymbol{\varepsilon}$ is the traffic count error with dispersion matrix \mathbf{W} . Thus the GLS estimator T_{GLS} of \mathbf{T} is obtained by solving:

$$T_{GLS} = \arg \min_{T \in S} (\hat{\mathbf{T}} - T)' \mathbf{Z}^{-1} (\hat{\mathbf{T}} - T) + (\hat{\mathbf{V}} - \mathbf{V}(T))' \mathbf{W}^{-1} (\hat{\mathbf{V}} - \mathbf{V}(T)) \quad (11)$$

S is the feasible set of \mathbf{T} . There the dispersion matrix \mathbf{Z} depends on the sampling estimator adopted.

BI method has also been applied in various transportation planning problems where prior beliefs are combined with the observations to produce the posterior beliefs. For further application and details of BI, refer Dey and Fricker (1994) paper where BI has been used for updating trip generation data. The BI approach considers the target ODM as a prior probability function $Pr(\mathbf{T})$ of the estimated ODM \mathbf{T} . If the observed traffic counts are considered as another source of information with a probability $L(\hat{\mathbf{V}} / \mathbf{T})$, then Bayes theorem provides a method for combining the two sources of information. The posterior probability $f(\mathbf{T} / \hat{\mathbf{V}})$ of observing \mathbf{T} conditional on the observed traffic counts is obtained as:

$$f(\mathbf{T} / \hat{\mathbf{V}}) \approx L(\hat{\mathbf{V}} / \mathbf{T}) \cdot Pr(\mathbf{T}) \quad (12)$$

2.3.1 Models without Considering Congestion Effects

The ODM estimation methods developed for networks with no congestion effects basically assume the route choice proportions and are independently determined outside the estimation process. For such networks Bayesians inference based approach has been first introduced by Maher (1983) for the ODM estimation. GLS estimator based approach has been studied by Cascetta (1984), Bell (1991a) etc. Bell (1991a) solved the GLS problem subject to inequality constraints and presented a simple algorithm but its application on real network has not been found in literature. Bierlaire and Toint (1995) proposed an ODM estimation method, called the Matrix Estimation Using Structure Explicitly (MEUSE), considering the information obtained from the parking surveys. Maximum likelihood based model is studied by Spiess (1987), Cascetta and Nguyen (1988), Hazelton (2000) etc. Two classical inference approaches; the ML and the GLS methods are derived and contrasted to the Bayesian method by Cascetta and Nguyen (1988). In all these studies, the link choice proportions used are estimated from the traffic assignment (TA) model and are assumed to be constant which may not estimate a dependable matrix. In consequence, Lo et al (1996) incorporated the randomness of the link choice proportions and discussed both Maximum likelihood and Bayesian approach for the estimation of the ODM by testing with a small network. Liu and Fricker (1996) introduced a stochastic logit model for calculating driver's route choice behaviour but it has certain drawback like all the link counts are considered to be known (further refer Yang et al, 2001). Lo et al (1999) extended the approach of Lo et al (1996) and developed a coordinate descent method using the partial linearization algorithm (PLA) for obtaining the optimum estimates and solved the new approach for large networks. Hazelton (2000) tested the performance of both multivariate normal (MVN) likelihood approximation and GLS techniques and found that the MVN method performed better. Hazelton (2001) studied the fundamental theoretical aspects of the ODM problem based on BI, defining the estimation, prediction and reconstruction problems as; a 'reconstruction' problem estimates the actual ODM occurring during the observation period, an expected number of O-D trips is obtained in 'estimation' problem and future O-D trips are obtained in 'prediction' problem. It has been shown that the estimation and reconstruction problems are different. There are some more studies (Hazelton, 2003; Van Aerde et al, 2003; Li, 2005, etc.) based on statistical approaches.

2.3.2 Models Considering Congestion Effects

Some authors included congestion effects in the estimation problem in which the dependence of the link costs, path choices and assignment fractions on link flows is considered. Equilibrium assignment approaches are particularly adopted for such cases. Nguyen (1977) first introduced the equilibrium based approach to estimate ODM through a mathematical programme (refer Leblanc and Farhangian, 1982). Also Yang et al (1994), Cascetta and Posterino (2001) and Yang et al (2001) solved the trip matrix estimation problem including congestion effects by considering different TA models. Cascetta and Posterino (2001) considered SUE assignment as a fixed-point problem whereas Yang et al (2001) considered the same problem of Liu and Fricker (1996) (developed without considering congestion effects) and proposed a non-linear optimization model (considering a weighted least square estimate) for the simultaneous

estimation of the ODM and travel-cost coefficient based on the logit-based SUE model. For solving this non-convex optimization problem a successive quadratic programming (SQP) method (which is a descent-feasible direction algorithm solving KKT solution) has been used. Further, Lo and Chan (2003) with SUE principle (multinomial logit model) estimated both the dispersion parameter θ in multinomial logit model and the trip matrix simultaneously using Quasi-Newton method. The objective function considered is a likelihood function. Most of the above developed models are found to be tested using small networks. And the models do not assure their applicability for large real size congested networks.

Combined Distribution and Assignment (CDA) Based Problem

CDA is a network equilibrium based approach combining distribution and assignment problem. The model by Erlander et al (1979) for the simultaneous prediction of the flows and the demands is modified and used for matrix estimation by some researchers. Fisk and Boyce (1983) introduced the link count data in calibrating the dispersion parameter in Erlander et al (1979) model. The observed link flows are assumed to be available for all the network links. Fisk (1989) examined that the problem of Fisk and Boyce (1983) is simplest to solve and recommended to use for the problems where link cost functions are separable. Kawakami and Hirobata (1992) extended the Fisk's model by including the mode choice behaviour and proposed an optimization problem in the form of a combined distribution, modal split, and TA method. An entropy constraint condition with respect to the traffic mode choice behaviour has been considered. The objective function is an entropy function for the traffic distribution. The constraints include a Beckmann-type user equilibrium for all traffic modes and the entropy constraint with respect to the traffic modes. An iterative convergent algorithm has been proposed and is applied for a road network in Nagoya (Japan) considering two categories of automobiles; large-sized trucks and buses, and small trucks and cars. The network consists of 16 zones, 154 nodes and 240 links. But while applying it to the road network, the estimated values found to be deviated from the observed value, might be because of generalized cost function.

Bi-level Programming Approach

Bi-level programming approach has been used for the problem of ODM estimation in case of congested network. In this approach the upper-level problem is the trip matrix estimation problem and the lower-level problem represents a network equilibrium assignment problem. Genetic Algorithm (GA, a probabilistic global searching method) can also be seen introduced to solve the bi-level programming models.

Spiess in 1990 formulated a convex minimization, gradient based model (method of steepest descent) which can be applicable for large real networks. This ODM adjustment problem, allows adjustments between the traffic flows from the assignment algorithm and counts. It has been implemented using EMME/2 transportation planning software. The convex minimization problem is the distance between the observed and the assigned volumes shown below,

$$\text{Min } F(T) = \sum_{a \in A} (v_a - \hat{v}_a)^2 \quad (13)$$

subject to

$$\mathbf{V} = \mathbf{P}(\mathbf{T}) \quad (14)$$

where v_a and \hat{v}_a are the estimated and the counted flows for the link a ($a \in A$) respectively. For the convexity of the problem a set of non-decreasing link cost functions on all links of the network are assumed to be obtained from equilibrium assignment externally. The approach estimates an approximate gradient of the objective function with respect to the O-D demands. The Spiess model can estimate non-zero values for O-D pairs with zero trips considered initially, which can be a matter of concern for the planners who wants to preserve the structure of the target ODM. Thus, Doblas and Benitez (2005) further extended the above Spiess (1990) study for cases when one need to preserve the structure of the target ODM. An efficient approach has been proposed to update trip matrix for large study areas with minimum stored information. A new formulation in addition to the above formulation (eqns. 13 and 14) has been proposed which incorporates constraints based on trip generation, trip attraction, total trips in the network and trips between the O-D pairs with their upper and the lower bounds decided according to the planner. The problem has been transformed to an augmented lagrangian function and optimized using Frank-Wolfe algorithm. The approach has been tested on a real large size network.

Yang et al (1992) through a heuristic technique solved the integrated problem of GLS technique/entropy maximization with equilibrium TA model in the form of a convex bi-level optimization problem. In this study, the GLS estimate needs a matrix inversion of the size equal to the number of O-D pairs. Such an inversion requirement can have computational problem for large networks. Yang (1995) extended the bi-level programming problem by including link flow interaction and developed a model with heuristic algorithms. Kim et al (2001) discussed the problem of the bi-level models developed by Yang et al (1992). The author discussed the problem of dependency on the target ODM of the bi-level model and proposed an alternative model using GA. The upper level is solved by a combined solution method with GA for ODM estimation. The mathematical proof of the solution being optimal has not been discussed. To circumvent the difficulties with the non-differentiabilities of upper level problem Codina et al (2006) presented two alternative algorithms for solving bi-level problems and tested on small networks. First is a hybrid scheme proximal point-steepest descent method (for upper level sub-problem) and considering elastic demand TA problem and second a simple bi-level program considering fixed point mapping using linear TA problem. An iterative column generation algorithm which converges into a local minimum under the concept of continuity in path cost function is formulated by Garcia and Verastegui (2008) and applied on small networks. Recently, Lundgren and Peterson (2008) developed a heuristic bi-level problem solving it by a descent algorithm and demonstrated the algorithm using a large size network for the city of Stockholm, with 964 links and 1642 O-D pairs.

Path Flow Estimation (PFE) based Models

Recently, models based on path flow estimator which determines ODM according to the solutions of path flows have been adopted. It is a single level mathematical program in which the interdependency between O-D trip table and route choice proportion

(congestion effect) is taken into account. The core component of PFE is a logit based path choice model, which interacts with link cost functions to produce a stochastic user equilibrium traffic pattern. Sherali et al (1994) proposed a linear path flow model employing user equilibrium based solution for reproducing the observed link flows (known for all links). The procedure utilizes shortest path network flow programming sub-problem and a column generation technique is applied to generate the paths out of alternate paths that will determine the optimal solution to the linear programming model. To avoid the path enumeration required in the model proposed by Sherali et al (1994), Nie and Lee (2002) solved the linear programming model considering an exogenous K-shortest-path for determining the equilibrium path flow pattern. Nie et al (2005) further extended the decoupled path flow estimator by Nie and Lee (2002) considering the generalized least squares framework in aspect of the limitations of the linear programming structure. Sherali et al (2003) enhanced the linear programming model of Sherali et al (1994) for situations where only a partial set of link volumes are available. This introduces nonlinear cost function because of the dependence of the link travel cost on link volumes and a fixed point solution is tried. Further tests using larger and real-size networks are required with these PFE based models for better assessment and efficiency checking of these models.

2.4 Multi-Vehicle ODM Estimation

In spite of single-vehicle information (aggregated information) some authors included additional information obtained from the survey of the individual vehicles types. The multiclass O-D estimation leads to eliminate the internal inconsistency of traffic flows among different vehicle classes. Baek et al (2004) used multiple-vehicle information for ODM estimation from traffic counts. The multi-vehicle ODM estimation method is given as:

$$\text{Min } F(t_w) = \frac{1}{2} \sum_c \sum_{a \in A} (v_a^c - \hat{v}_a^c)^2 + \gamma \frac{1}{2} \sum_c \sum_{w \in W} (t_w^c - \hat{t}_w^c)^2 \quad (15)$$

subject to

$$\mathbf{T} \geq 0 \quad (16a)$$

$$\mathbf{V} = \mathbf{P}(\mathbf{T}) \quad (16b)$$

where \mathbf{T} is the ODM with elements t_w^c as the trips for the vehicle type c between O-D pairs w , \hat{t}_w^c is the target O-D trips of vehicle type c between O-D pairs w , \mathbf{V} as the vector with elements v_a^c as traffic volume of the vehicle type c on the link a , \hat{v}_a^c is the observed volumes of vehicle type c on the link a , γ is the parameter reflecting the reliability of the target ODM and $\mathbf{P}(\mathbf{T})$ is the multi-vehicle traffic assignment map. For solving the non-convex problem, genetic algorithm has been used. The algorithm has been demonstrated only using a small network of 16 links with 8 O-D pairs and 9 nodes. It required more computing time due to involvement of genetic algorithm. An entropy maximization based ODM estimation of all vehicle classes with multiclass traffic counts observed directly from the network has been proposed by Wong et al (2005). A set of

multiplication factors for adjusting the demand matrix is used which may change because of changes in land use pattern or network configuration. The estimated matrix largely depends on the reliability of the input data.

2.5 Fuzzy Based Approach

Many non-linear real-life problems in the field of transportation planning and traffic control have been solved by fuzzy logic systems. Its application can be found in ODM estimation problem too. Fuzzy logic is used to model situations in which user making decisions are so complex that it is very hard to develop a mathematical model. Xu and Chan (1993a, b) estimated ODM with fuzzy weights (refer Teodorovic, 1999). Reddy and Chakroborty (1998) proposed a bi-level optimization approach on a multipath, fuzzy inference based flow dependent assignment algorithm for generating the route choice proportions which along with the observed link flows are then used in the ODM estimation. The proposed technique has been observed to give good O-D estimates irrespective of the flow pattern. Still not much study has been carried with fuzzy logic applications. The real-size network application of fuzzy logic systems should be further studied to ensure its efficiency.

2.6 Neural Network Based Approach

Considering the highly dynamic, large scale, complex and uncertain nature of many transportation systems, neural networks are recently considered as an efficient tool in solving numerous transportation problems. Gong (1998) developed the Hopfield Neural Network (HNN) model to estimate the urban ODM from link volumes, to promote the solving speed and the precision. Though there are quite a few studies carried out by researchers using Artificial Neural Networks (ANNs), but their application still need to be studied for real networks.

3. Dynamic/Time-Dependent ODM Estimation from Traffic Counts

Dynamic/time-dependent ODM estimation is crucial estimation problem for online/offline applications such as route guidance, dynamic traffic assignment (DTA) and freeway corridor control. Also used in various microscopic simulation based studies. Dynamic/time-dependent ODM estimation got much attention due to the development of Intelligent Transportation Systems (ITS). For dynamic (online) ODM estimation traffic counts are observed for short time say 15 min interval. Time-dependent (offline) ODM estimation considers time-series of traffic counts. Mostly the developed algorithms for both dynamic and time-dependent ODM estimation are applied for “closed” networks like intersections/interchanges and small freeways. The dynamic formulation of the matrix estimation process can be expressed as

$$f_{sh} = \sum_k \sum_w p_{sh}^{wk} t_{wk} \quad (17)$$

where f_{sh} is the flow crossing sensor s in time interval h , t_{wk} is the flow between O-D pair w that departed its origin during time interval k and p_{sh}^{wk} is an assignment parameter reflecting the proportion of the demand t_{wk} crossing sensor s in time interval h . The difference of dynamic formulation from static estimations lies in parameters k and h . And very few of them are developed considering large size network application. Classical techniques of dealing these systems are State-Space Modeling and Kalman filter which are discussed below.

To develop a **State-Space model**, a state is defined first. Once a state is defined, transition and measurement equations are specified. In dynamic systems, transition equations describe the evaluation of the state over time. Measurement equations on the other hand relate the unknown state to their observed indicators.

$$\text{Measurement Equation: } y_h = C_h x_h + e_h \quad (18)$$

$$\text{Transition Equation: } x_{h+1} = D_h x_h + z_h \quad (19)$$

Where x_h is the vector that represents the latent “true state” of the system during interval h , y_h is a vector of observations made in interval h , C_h and D_h are known matrices and, e_h and z_h are vectors of random errors.

Kalman filtering is an optimal state estimation process applied to a dynamic system that gives a linear unbiased and minimum error variance recursive algorithm. ODM estimation problem is formulated with a linear mapping from the state variable to the measurement variable. The formulation is stochastic which allows for measurement and state errors, and the initial state could be described stochastically. The transition equation for Kalman Filter is of the form

$$x_h = \phi_{h-1} x_{h-1} + \Gamma_{h-1} u_{h-1} + \Lambda_{h-1} q_{h-1} \quad (20)$$

where x_h is the state vector to be determined at the time h , ϕ_h is the matrix of autoregressive coefficients, u_h is a deterministic input, Γ_h the control gain, Λ_h the disturbance gain, and q_h is the disturbance input. The measurement equation is given by

$$y_h = H_{h-1} x_{h-1} + n_h \quad (21)$$

where n_h is the error (noise) term with zero mean and positive definite covariance matrix and the matrix H_h relates the state to the measurement of x_h . The filter computes the estimate of the state while minimizing the spread of the estimate error probability distribution function.

3.1 Dynamic or On-line ODM Estimation

The dynamic ODM estimation also stated as real-time ODM estimation, characterizes the time-variant trips between each origin and destinations. It is an essential input for DTA models and used for on-line identification of travel pattern for traffic controls systems. The dynamic models can be characterized as non-DTA based models and DTA

based models. The non-DTA based models like Kalman filter based models and Parameter optimization based models are basically applied for small networks like intersections, freeways etc where entry and exit flow information are available.

A state-space model with unknown O-D flows as the state vector has been first introduced by Okutani in 1987 (refer Kachroo et al, 1997). Ashok and Ben-Akiva (1993) (refer Sherali and Park, 2001) proposed a Kalman filtering approach to dynamically update an ODM. The O-D flow deviations from the prior estimates based on historical data are considered (for capturing the structural information) as the state-vectors in order to overcome inadequacy of autoregressive specification for O-D flows in Okutani's approach. Kachroo et al (1997) studied the applicability of Kalman filtering approaches for network ODM estimation from link traffic counts to explore the characteristics of the error terms in the underlying dynamic process of the O-D departures. The inconsistencies between the observed O-D flow patterns and Kalman Filter modeling assumptions is analyzed. It has been concluded that the noise is not a white Gaussian sequence. Ashok and Ben-Akiva (2000) further extended Ashok and Ben-Akiva's (1993) approach and presented a new formulation based on deviations of departure rates from origin and destination shares over time instead of destination flows.

Cremer and Keller (1987), Nihan and Davis (1987), Bell (1991b), Wu and Chang (1996), determined split parameters (averaged values) for input-output network relationships that is applicable for traffic flows at intersections or small freeway segments. Sherali et al (1997) developed a constrained optimization algorithm but with high computational cost. Li and Moor (1999) also proposed a recursive-based algorithm. The above approaches need all entry and exit information which is somewhat unrealistic. For the situations with incomplete traffic counts at some entrances and exits, Li and Moor (2002) formulated an optimization problem with linear equality constraints and non-negative inequality constraints. Van der Zijpp and Lindveld (2001) formulated a dynamic user optimal departure time and route choice (DUO-D&R) assignment problem which is used to estimate dynamic ODM with preferred departure times. There are some more studies (Van der Zijpp, 1997; Suzuki et al, 2000 etc.) on dynamic ODM estimation for intersections and freeways.

Due to disadvantages of Kalman filtering formulation (needs sufficient data and intensive matrix operations) Wu (1997) developed a real-time ODM updating algorithm based on a balancing method called multiplicative algebraic reconstruction technique (MART of Lamond and Stewart, 1981) considering entropy-maximization model. MART has been revised (RMART) by incorporating a normalization scheme, without giving any theoretical explanation of doing so. A diagonal searching technique is considered for improving the convergence speed. A numerical test has been carried out for checking the efficiency of the RMART with artificially generated database from some computer simulated problems. Zhou et al (2003) included the historical static information and ITS real-time link-level information to determine the dynamic O-D demand. The variation in day-to-day demand is studied by using multiday traffic counts. Nie and Zhang (2008) gives a brief review on dynamic ODM estimation algorithms and formulated a variational inequality problem determining the dynamic traffic assignment endogenously considering the user response to traffic congestion through a dispersion parameter θ . The problem finds out path flows denoted here as \mathbf{f} which satisfy the conditions given in eqn. (22) which are transformed into a variational inequality formulation.

$$\begin{cases} f_w^{jh}(c_w^{jh} - \theta d_w^{jh}) = 0 \\ c_w^{jh} \geq \theta d_w^{jh}, f_w^{jh} \geq 0 \end{cases} \quad (22)$$

Here d_w^{jh} denotes the path derivatives between O-D pairs w for assignment interval h and c_w^{jh} , the path travel time of path j during assignment interval h . A solution for the variational inequality formulation has been proposed using a space-time expanded network (STEN, refer the journal paper for further details) to generate paths. A column generation algorithm has been proposed to solve the dynamic matrix estimation problem iterating the two sub-problems; generating paths to construct a restricted VI problem and to find an optimum solution of the restricted problem. The results depend on the initial path flows and the convergence issues still need to be studied.

3.2 Time-Dependent or Off-Line ODM Estimation

This estimation is carried out off-line, given a time-series of link counts, travel times and prior O-D information. Chang and Tao (1999) proposed a model integrating the link constraints and intersection turning flows from available DTA model, for determining the time-varying O-D trips for intersections applying Kalman filtering approach. A parametric optimization approach for off-line processing purpose is developed by Sherali and Park (2001). The proposed model seeks path flows that compromise between the least cost O-D paths and those that provide a match for the observed link flows. But with the increase in O-D paths, this model seems to be difficult to solve. Ashok and Ben-Akiva (2002) introduced the stochasticity of the assignment matrix in estimating the time-dependent O-D flows from link volumes. A GLS based solution has been studied for minimizing the error criteria for each interval and is evaluated for a case study. Tsekeris and Stathopoulos (2003) coupled multi-proportional procedure (MPP) of Murchland (1977) and multiplicative algebraic reconstruction technique (MART) of Gordon et al (1970) respectively, with a quasi-DTA model and estimated dynamic trip departure rates and ODM over a series of successive time interval. Combining the algorithms a doubly iterative matrix adjustment procedure (DIMAP) has been proposed to obtain a consistency between the trip departure rates from each origin zone and the observed link flows. The simulation based quasi-dynamic model used is based on the instantaneous link travel cost definition. The effect of congestion is incorporated and the performance of the algorithms is studied using a real network. Estimation is carried out separately using simulated link flows and real traffic flows. The DIMAP has been compared with MART and RMART (Wu, 1997). While considering simulated link flows the DIMAP algorithm found to perform better than the case using real traffic flows. Recently, BI based parsimonious parameterized model for estimating time varying ODM with traffic counts (collected on daily basis) has been recommended by Hazelton (2008).

4. The Measure of reliability of the estimated ODM

Statistical measures

The outcome of a situation which is difficult to predict is generally measured through some statistical measures. Likewise in ODM estimation problem some statistical measures are used by the authors to verify the performance of their proposed algorithms. The statistical measures only can measure the closeness of the estimated values (trips and link flows) and their true values, if known. Following are the statistical measures mostly adopted:

Relative error (RE) %:

$$RE (\%) = \sqrt{\frac{1}{2} \sum_{w \in W} \left(\frac{t_w^* - t_w}{t_w} \right)^2} \times 100\% \quad (23)$$

Total Demand Deviation (TDD) %:

$$TDD (\%) = \frac{\left| \sum_{w \in W} t_w^* - \sum_{w \in W} t_w \right|}{\sum_{w \in W} t_w} \times 100\% \quad (24)$$

Mean absolute error (MAE) %:

$$MAE (\%) = \frac{\sum_{w \in W} |t_w^* - t_w|}{N} \times 100\% \quad (25)$$

Root Mean Square Error (RMSE) %:

$$RMSE (\%) = \frac{\sqrt{\frac{1}{N} \sum_{w \in W} (t_w^* - t_w)^2}}{\frac{1}{N} \sum_{w \in W} t_w} \times 100\% \quad (26)$$

where t_w^* is the true ODM and N is the number of O-D pairs. The TDD gives the quality of the estimated ODM. The RMSE percent error quantifies the total percentage error of the estimate. The mean percent error indicates the existence of consistent under-or-over-prediction in the estimate. Smaller values of these measures will indicate the high quality of the estimated ODM. But in situations when the true values are not known these statistical measures cannot be used.

Maximum Possible Relative Error (MPRE)

Yang et al (1991) through a simple quadratic programming problem introduced a concept of maximum possible relative error (MPRE) which represents the maximum possible relative deviation of the estimated ODM from the true one and can be used only when the route choice proportions are correctly specified and the traffic counts are error free. The reliability of the estimated ODM from traffic counts is measured as,

$$Re(t) = \frac{1}{1+E}, \quad E \geq 0 \quad (27)$$

where E is the measure of the error (average relative deviation) in the estimated ODM depending upon the relative deviation of the estimated O-D flows from the true ones for the O-D pairs.

Travel Demand Scale (TDS)

Based on statistical analysis the quality measure of both static and dynamic ODM models is proposed by Bierlaire (2002) by means of TDS which is independent of the estimation method and a priori matrix (say obtained from a previous study), but depends upon the network topology and route choice assumptions. The Travel Demand Scale is computed as,

$$TDS = \varphi_{\max} - \varphi_{\min} \quad (28)$$

where

$$\varphi_{\min} = \min_t \mathbf{T}' \mathbf{e} \quad (29)$$

and

$$\varphi_{\max} = \max_t \mathbf{T}' \mathbf{e} \quad (30)$$

subject to

$$\mathbf{P}\mathbf{T} = \hat{v}_a \quad (31a)$$

$$\mathbf{T} \geq 0 \quad (31b)$$

where φ_{\min} and φ_{\max} are minimum and maximum total level, \mathbf{e} is the vector only of ones and \mathbf{P} is the vector notation of the assignment matrix corresponding to the links where flow observations are available. The TDA value (for values refer the journal paper) helps to optimize the resources allocated during the surveys by identifying the nature of the additional information required. It finds out the unbounded O-D pairs (O-D pairs not captured by link flow data) so that surveys can be conducted to increase the quality of the corresponding entries in the a priori ODM. Thus it helps to assess the

level of investment necessary to collect data and build the a priori matrix. It is recommended to use in addition to the statistical measures.

5. Traffic Counting Location

For the ODM estimation, traffic counting or sampling survey data collection are carried out where a road or rail route crosses a cordon line and screen lines. The accuracy of the ODM estimated increases with the number of traffic counting stations adopted. But due to resource limitations, it may not be possible. Again, the traffic count at each location has different degree of influence to the ODM estimation. Hence it is necessary to determine the optimum number of counting stations and their locations on the network to intercept maximum O-D pairs. Yang and Zhou (1998) introduced four basic rules of locating traffic counting points based on the maximal possible relative error (MPRE) concept proposed by Yang et al (1991). Based on the *O-D covering rule* stated by Yang et al (1991), following are the rules proposed by Yang and Zhou (1998);

(1) *O-D covering rule*: At least one traffic counting point on the network must be located for observing trips between any O-D pair.

(2) *Maximal flow fraction rule*: The traffic counting points on a network must be located at the links between a particular O-D pair such that flow fraction ϕ_{aw} is as large as possible.

$$\phi_{aw} = \frac{P_{aw} t_w}{v_a} \quad (32)$$

where v_a is the link flow, $a \in A$ between the O-D pair $w \in W$ and p_{aw} is the proportion of trips used by link $a \in A$ for each O-D pair w .

(3) *Maximal flow-intercepting rule*: From the links to be observed, the chosen links should intercept as many flows as possible.

(4) *Link independence rule*: The traffic counts on all chosen links should be linearly independent.

For determining traffic count locations, the O-D covering rule and the link independence rule are treated as constraints and maximal flow fraction rule and maximal flow-intercepting rule are incorporated in objective function. Yim and Lam (1998) presented rules of maximum net O-D captured and maximum total O-D captured, and formulated in a linear programming model to determine the locations for ODM estimation. Bianco et al (2001) through proposed heuristic method solved the sensor location problem by proposing a two-stage procedure; determining the minimum number and location of counting points with known turning probabilities (assumed) and estimating the ODM with the resulting traffic flows. Yang et al (2003) studied the scheduling installation of traffic counting stations for long duration planning purpose and a Genetic Algorithm based sequential greedy algorithm has been proposed. Kim et al (2003) formulated two models, link-based and road-based model, to determine the location of the counting points on the link which minimizes the total cost. Three solution algorithm: Greedy Adding (GA) algorithm, Greedy Adding and Substituting

(GAS) algorithm and Branch and Bound (BB) algorithm are proposed and tested for a simple artificial network. Ehlert et al (2006) extended the problem of Chung (2001) (refer Ehlert et al, 2006) to optimize the additional counting locations assuming that some detectors are already installed increasing the O-D coverage. A software tool is developed based on mixed integer problem (MIP). With the budget restrictions for practical problems it is stated that the OD covering rule cannot always be satisfied i.e. the ODM estimation error measure MPRE cannot be applied. Gan et al (2005) studied both the traffic counting location and the error estimation measure in ODM estimation problem taking into consideration the route choice assumptions made in the TA models and the levels of traffic congestion on the networks. It has been noted that both MPRE and TDS measures are closely related to traffic counting locations. When the O-D covering rule is not satisfied by the link counting locations both the measures become infinite. Yang et al (2006) formulated an integer linear programming (ILP) problem using shortest path-based column generation procedure and branch-and-bound technique in determining screen line based traffic counting locations.

6. Conclusions

The basic goal of this paper is to explore the studies on one of the most promising topics for research which is the estimation of ODM using traffic counts on a set of links. The review shows the intricacy of the ODM estimation problem using traffic counts, the reason being the under-specification of the trip matrix estimation problem with less link count information than the number of unknowns. Till date both static and dynamic ODM estimation problems have been investigated by many researchers and models have been developed with different problem formulations, using different route choice decisions process and various solution algorithms.

The statistical approaches (ML, GLS and BI) have been mostly adopted by the researchers to solve the static problem considering congestion and without considering congestion effects. Both bi-level programming approach and simultaneous (single-level) optimization approaches have been studied in literature. Also Path flow estimation based algorithms are proposed assuming that all link costs are available, which may not be available in practical situations. Very few authors used fuzzy logic and neural network based approaches and their applicability need to be analyzed further. The review shows that most of the algorithms developed for static ODM have its own advantages and disadvantages and are implemented on small networks. However the most important consideration required is the applicability of the algorithms for real world networks which are large in size and highly congested. It is surprising to see a few realistic approaches in the literatures focused on large size network applications. Thus the developed algorithms need to be checked regarding their practical applicability for large size real networks.

The static ODM determined for long-time transportation planning and design purpose is easier to estimate as compared to dynamic ODM used for traffic management and operations for large networks because availability of real-time traffic information for all the O-D pairs required for dynamic trip matrix estimation is not possible. Compared to static ODM estimation, dynamic estimation based studies are few and mostly for intersections, freeways and small networks; as it is convenient to study the dynamic

state of these networks. Some authors tried to extend the study for large networks. But for practical applications dynamic ODM still is not much in use except for performing DTA on small scale networks.

To identify the reliability of the estimated ODM the statistical measures and the MPRE needs real trip values which are not always available in practical cases. The TDS though measures quality for both static and dynamic matrices but it does not serve alone the purpose of measuring the reliability. Some authors gave emphasis on finding the optimum number and location of the traffic counting points. Quite some rules have been proposed in the literature which can help in obtaining the optimum traffic counting locations and in receiving more information of travel pattern between O-D pairs.

As indicated earlier more studies and checking (mainly regarding computational difficulties) of the developed algorithms still need to be carried out especially for the case of its application for planning and designing purpose done for large size networks.

Acknowledgement

We wish to thank the referee for his useful comments.

References

- Ashok, K. and Ben-Akiva, M. E. (1993) "Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems", In: Daganzo, C. F. (Ed.), *Transportation and Traffic Theory*, Elsevier, Amsterdam.
- Ashok, K. and Ben-Akiva, M. E. (2000) "Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows", *Transportation Science* 34: 21-36.
- Ashok, K. and Ben-Akiva, M. E. (2002) "Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows", *Transportation Science* 36: 184-198.
- Baek, S., Kim, H. and Lim, Y. (2004) "Multiple-vehicle origin-destination matrix estimation from traffic counts using genetic algorithm", *Journal of Transportation Engineering*, ASCE, May/June: 339-347.
- Bell, M. (1991a) "The estimation of origin-destination matrices by constrained generalized least squares", *Transportation Research, Part B: Methodological* 25: 13-22.
- Bell, M. (1991b) "The real time estimation of origin-destination flows in the presence of platoon dispersion", *Transportation Research* 25B: 115-125.
- Bianco, L., Confessore, G. and Reverberi, P. (2001) "A network based model for traffic sensor location with implications on o/d matrix estimates", *Transportation Science* 35(1) : 49-60.
- Bierlaire, M. and Toint, Ph. L. (1995) "Meuse: an origin-destination matrix estimator that exploits structure", *Transportation Research, Part B: Methodological* 29: 47-60.
- Bierlaire, M. (2002) "The total demand scale: a new measure of quality for static and dynamic origin-destination trip tables", *Transportation Research, Part B: Methodological* 36: 837-850.
- Cascetta, E. (1984) "Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator", *Transportation Research, Part B: Methodological* 18: 289-299.
- Cascetta, E. and Nguyen, S. (1988) "A unified framework for estimating or updating origin/destination matrices from traffic counts", *Transportation Research, Part B: Methodological* 22: 437-455.
- Cascetta, E. and Postorino, M. N. (2001) "Fixed point approaches to the estimation of O/D matrices using traffic counts on congested networks", *Transportation Science* 35: 134-147.
- Chang, G. and Tao, X. (1999) "An integrated model for estimating time varying network origin-destination distributions", *Transportation Research, Part A: Policy and Practice* 33: 381-399.
- Chung, I. H. (2001) An optimum sampling framework for estimating trip matrices from day-to-day traffic counts, *Ph.D. Thesis*, University of Leeds.
- Codina, E., Garcia, R. and Marin, A. (2006) "New algorithm alternatives for the O-D matrix adjustment problem on traffic networks", *European Journal of Operational Research* 175: 1484-1500.

- Cremer, M. and Keller, H. (1987) "A new class of dynamic methods for the identification of origin-destination flows", *Transportation Research, Part B: Methodological* 21: 117-132.
- Dey, S. S. and Fricker, J. D. (1994) "Bayesian updating of trip generation data: combining national trip generation rates with local data", *Transportation* 21(4): 393-403.
- Doblas, J. and Benitez, F. G. (2005) "An approach to estimating and updating origin-destination matrices based upon traffic counts preserving the prior structure of a survey matrix", *Transportation Research, Part B: Methodological* 39: 565-591.
- Ehlert, M., Bell, M. G. H. and Grosso, S. (2006) "The optimization of traffic count locations in road networks", *Transportation Research, Part B: Methodological* 40: 460-479.
- Erlander, S., Nguyen S. and Stewart N. (1979) "On the calibration of the combined distribution/assignment model", *Transportation Research, Part B: Methodological* 13: 259-267.
- Fisk, C. S. and Boyce, D. E. (1983) "A note on trip matrix estimation from link traffic count data", *Transportation Research, Part B: Methodological* 17: 245-250.
- Fisk, C. S. (1989) "Trip matrix estimation from link traffic counts: the congested network case", *Transportation Research, Part B: Methodological* 23: 331-356.
- Gan, L., Yang, H. and Wong, S. C. (2005) "Traffic counting location and error bound in origin-destination matrix estimation problems", *Journal of Transportation Engineering*, ASCE/July: 524-534.
- Garcia, R. and Verastegui, D. (2008) "A column generation algorithm for the estimation of origin-destination matrices in congested traffic networks", *European Journal of Operational Research* 184: 860-878.
- Gong, Z. (1998) "Estimating the urban o-d matrix: a neural network approach," *European Journal of Operational Research* 106: 108-115.
- Gordon, R., Bender, R. and Hermann, G. T. (1970) "Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography", *Journal of Theoretical Biology*, Vol. 29: 471-481.
- Hazelton, M. L. (2000) "Estimation of origin-destination matrices from link flows on uncongested networks", *Transportation Research, Part B: Methodological* 34: 549-566.
- Hazelton, M. L. (2001) "Inference for origin-destination matrices: estimation, prediction and reconstruction", *Transportation Research, Part B: Methodological* 35: 667-676.
- Hazelton, M. L. (2003) "Some comments on origin-destination matrix estimation", *Transportation Research, Part A: Policy and Practice* 37: 811-822.
- Hazelton, M. L. (2008) "Statistical inference for time varying origin-destination matrices", *Transportation Research, Part B: Methodological* 42: 542-552.
- Hogberg, P. (1976) "Estimation of parameters in models for traffic prediction: a non-linear regression approach", *Transportation Research*, Vol. 10: 263-265.
- Kachroo, P., Ozbay, K. and Narayanan, A. (1997) "Investigating the use of Kalman filtering for dynamic origin-destination trip table estimation", Southeastcon'97, 'Engineering NewCentury', Proceedings, IEEE: 138-142.
- Kawakami, S. and Hirobata, Y. (1992) "Estimation of origin-destination matrices from link traffic counts considering the interaction of the traffic modes", *Regional Science Association International*, Vol. 71, No. 2: 139-151.
- Kim, H., Beak, S. and Lim, Y. (2001) "Origin-destination matrices estimated with a genetic algorithm from link traffic counts", *Transportation Research Record* 1771, Transportation Research Board, Washington, D.C.: 156-163.
- Kim, H. J., Chung, H. I. and Chung, S. Y. (2003) "Selection of the optimal traffic counting locations for estimating origin-destination trip matrix", *Journal of the Eastern Asia Society for Transportation Studies*, Vol. 5: 1353- 1365.
- Lamond, B. and Stewart, N. F. (1981) "Bregman's balancing method", *Transportation Research, Part B: Methodological* 15: 239-248.
- LeBlanc, L. J. and Farhangian, K. (1982) "Selection of a trip table which reproduces observed link flows", *Transportation Research, Part B: Methodological* 16: 83-88.
- Li, B. and Moor, B. D. (2002) "Dynamic identification of origin-destination matrices in the presence of incomplete observations", *Transportation Research, Part B: Methodological* 36: 37-57.
- Li, B. (2005) "Bayesian inference for O-D matrices of transport networks using the EM algorithm", *Technometrics* 47(4): 399-408.
- Liu, S. and Fricker, J. D. (1996) "Estimation of a trip table and the Θ parameter in a stochastic network", *Transportation Research, Part A: Policy and Practice* 30: 287-305.

- Lo, H., Zhang, N. and Lam, W. (1996) "Estimation of an origin-destination matrix with random link choice proportions: a statistical approach", *Transportation Research, Part B: Methodological* 30: 309-324.
- Lo, H. P., Zhang, N. and Lam, W. H. K. (1999) "Decomposition algorithm for statistical estimation of OD matrix with random link choice proportions from traffic counts", *Transportation Research, Part B: Methodological* 33: 369-385.
- Lo, H. P. and Chan, C. P. (2003) "Simultaneous estimation of an origin-destination matrix and link choice proportions using traffic counts", *Transportation Research, Part A: Policy and Practice* 37: 771-788.
- Lundgren, J. T. and Peterson, A. (2008) "A heuristic for the bilevel origin-destination matrix estimation problem", *Transportation Research, Part B: Methodological* 42: 339-354.
- Maher, M. J. (1983) "Inferences on trip matrices from observations on link volumes: a bayesian statistical approach", *Transportation Research, Part B: Methodological* 17: 435-447.
- Murchland, J. D. (1977) *The Multi-Proportional Problem*, Research Note JDM-263, University College London, London.
- Nguyen, S. (1977) "Estimating an OD matrix from network data: a network equilibrium approach", *Publication No. 87*, Center de recherché sur les transports, Université de Montréal, Montréal, Québec H3C 3J7.
- Nie, Y. and Lee, D. (2002) "Uncoupled method for equilibrium-based linear path flow estimator for origin-destination trip matrices", *Transportation Research Record* 1783, Transportation Research Board, Washington, D.C.: 72-79.
- Nie, Y., Zhang, H. M. and Recker, W. W. (2005) "Inferring origin-destination trip matrices with a decoupled GLS path flow estimator", *Transportation Research, Part B: Methodological* 39: 497-518.
- Nie, Y. and Zhang, H. M. (2008) "A variational inequality formulation for inferring dynamic origin-destination travel demands", *Transportation Research, Part B: Methodological* 42: 635-662.
- Nihan, N. L. and Davis, G. A. (1987) "Recursive estimation of origin-destination matrices from input/output counts", *Transportation Research, Part B: Methodological* 21: 149-163.
- Okutani, I. (1987) *The Kalman filtering approach in some transportation and traffic problems*, International symposium on transportation and traffic theory, N. H. Gartner. and N. H. M. Wilson (eds), Elveiser Science Publishing Company Inc.: 397-416.
- Reddy, K. H. and Chakroborty, P. (1998) "A fuzzy inference based assignment algorithm to estimate O-D matrix from link volume counts", *Comput., Environ. and Urban Systems*, Vol. 22, No. 5: 409-423.
- Robillard, P. (1975) "Estimating the O-D matrix from observed link volumes", *Transportation Research*, Vol. 9: 123-128.
- Sherali, H. D., Sivanandan, R. and Hobeika, A. G. (1994) "A linear programming approach for synthesizing origin-destination trip tables from link traffic volumes", *Transportation Research, Part B: Methodological* 28: 213-233.
- Sherali, H. D., Arora, N. and Hobeika, A. G. (1997) "Parameter optimization methods for estimating dynamic origin-destination trip-tables", *Transportation Research, Part B: Methodological* 31: 141-157.
- Sherali, H. D. and Park, T. (2001) "Estimation of dynamic origin-destination trip tables for a general network", *Transportation Research, Part B: Methodological* 35: 217-235.
- Sherali, H. D., Narayanan, A. and Sivanandan, R. (2003) "Estimation of origin-destination trip-tables based on a partial set of traffic link volumes", *Transportation Research, Part B: Methodological* 37: 815-836.
- Spiess, H. (1987) "A maximum-likelihood model for estimating origin-destination matrices", *Transportation Research, Part B: Methodological* 21: 395-412.
- Spiess, H. (1990) "A gradient approach for the O-D matrix adjustment problem", EMME/2 Support Center, CH-2558 Aegerten, Switzerland, <http://www.spiess.ch/emme2/demadj/demadj.html>.
- Suzuki, H., Nakatsuiji, T., Tanaboriboon, Y. and Takahashi, K. (2000) "Dynamic estimation of origin-destination travel time and flow on a long freeway corridor", *Transportation Research Record* 1739, Transportation Research Board, Washington, D.C.: 67-75.
- Tamin, O. Z. and Willumsen, L. G. (1989) "Transport demand model estimation from traffic counts", *Transportation* 16: 3-26.
- Tamin, O. Z., Hidayat, H. and Indriastuti, A. K. (2003) "The development of maximum-entropy (ME) and bayesian-inference (BI) estimation methods for calibrating transport demand models based on link volume information", *Proceedings of the Eastern Asia Society for Transportation Studies*, Vol. 4: 630-647.
- Teodorovic, D. (1999) "Fuzzy logic systems for transportation engineering: the state of the art", *Transportation Research, Part A: Policy and Practice* 33: 337-364.

- Tsekeris, T. and Stathopoulos, A. (2003) "Real-time dynamic origin-destination matrix adjustment with simulation and actual link flows in urban networks", *Transportation Research Record* 1857, Transportation Research Board, Washington, D.C.: 117-127.
- Van Aerde, M., Rakha, H. and Paramahamsam, H. (2003) "Estimation of origin destination matrices", *Transportation Research Record* 1831, Transportation Research Board, Washington, D.C.: 122-130.
- Van der Zijpp, N. J. (1997) "Dynamic origin-destination matrix estimation from traffic counts and automated vehicle information data", *Transportation Research Record* 1607, Transportation Research Board, Washington, D.C.: 87-94.
- Van der Zijpp, N. J. and Lindveld, C. D. R. (2001) "Estimation of origin-destination demand for dynamic assignment with simultaneous route and departure time choice", *Transportation Research Record* 1771, Transportation Research Board, Washington, D.C.: 75-82.
- Van Zuylen, J. H. (1978) "The information minimizing method: validity and applicability to transport planning", In: *New Developments in Modelling Travel Demand and Urban Systems* (edited by G. R. M. Jansen et al).
- Van Zuylen, J. H. and Willumsen L. G. (1980) "The most likely trip matrix estimated from traffic counts", *Transportation Research, Part B: Methodological* 14: 281-293.
- Van Zuylen, J. H. and Branston, D. M. (1982) "Consistent link flow estimation from counts", *Transportation Research, Part B: Methodological* 16: 473-476.
- Willumsen, L. G. (1978) "Estimation of O-D matrix from traffic counts: a review", *Working Paper* 99, Institute for Transport Studies, University of Leeds.
- Willumsen, L. G. (1981) "Simplified transport models based on traffic counts", *Transportation* 10: 257-278.
- Wilson, A. G. (1970) *Entropy in Urban and Regional Modeling*, Methuen, Inc., New York.
- Wong, S. C., Tong, C. O., Wong, K. I., Lam, W. H. K., Lo, H. K., Yang, H. and Lo, H. P. (2005) "Estimation of multiclass origin-destination matrices from traffic counts", *Journal of Urban Planning and Development*, ASCE, March: 19-29.
- Wu, J. and Chang, G. (1996) "Estimation of time-varying origin-destination distributions with dynamic screenline flow", *Transportation Research, Part B: Methodological* 30: 277-290.
- Wu, J. (1997) "A real-time origin-destination matrix updating algorithm for on-line applications", *Transportation Research, Part B: Methodological* 31: 381-396.
- Xu, W. and Chan, Y. (1993a) "Estimating an origin-destination matrix with fuzzy weights. Part 1: Methodology", *Transportation Planning and Technology* 17: 127-144.
- Xu, W. and Chan, Y. (1993b) "Estimating an origin-destination matrix with fuzzy weights. Part 2: Case Studies", *Transportation Planning and Technology* 17: 145-164.
- Yang, H., Iida Y. and Sasaki T. (1991) "An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts", *Transportation Research, Part B: Methodological* 25: 351-363.
- Yang, H., Sasaki, T., Iida, Y. and Asakura, Y. (1992) "Estimation of origin-destination matrices from link traffic counts on congested networks", *Transportation Research, Part B: Methodological* 26: 417-433.
- Yang, H., Sasaki, T. and Iida, Y. (1994) "The Equilibrium-based origin-destination matrix estimation problem", *Transportation Research, Part B: Methodological* 28: 23-33.
- Yang, H. (1995) "Heuristic algorithms for the bi-level origin-destination matrix estimation problem", *Transportation Research, Part B: Methodological* 29: 231-242.
- Yang, H. and Zhou, J. (1998) "Optimal traffic counting locations for origin-destination matrix estimation", *Transportation Research, Part B: Methodological* 32: 109-126.
- Yang, H., Meng, Q. and Bell, M. G. H. (2001) "Simultaneous estimation of the origin-destination matrices and travel-cost coefficient for congested networks in a stochastic user equilibrium", *Transportation Science* 35: 107-123.
- Yang, C., Chootinan, P. and Chen, A. (2003) "Traffic counting location planning using genetic algorithm", *Journal of the Eastern Asia Society for Transportation Studies*, Vol.5: 898-913.
- Yang, H., Yang, C. and Gan, L. (2006) "Models and algorithms for the screen line-based traffic-counting location problems", *Computers & Operations Research* 33: 836-858.
- Yim, P. K. N. and Lam, W. H. K. (1998) "Evaluation of count location selection methods for estimation of O-D matrices", *Journal of Transportation Engineering*, ASCE, July/August: 376-383.
- Zhou, X., Qin, X. and Mahamassani, H. S. (2003) "Dynamic origin-destination demand estimation with multiday link traffic counts for planning applications", *Transportation Research Record* 1831, Transportation Research Board, Washington, D.C.: 30-38.

Unravelling daily human mobility motifs

Christian M. Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda and Marta C. González

J. R. Soc. Interface 2013 **10**, 20130246, published 8 May 2013

Supplementary data

"Data Supplement"

<http://rsif.royalsocietypublishing.org/content/suppl/2013/05/02/rsif.2013.0246.DC1.htm>

References

[This article cites 40 articles, 10 of which can be accessed free](#)

<http://rsif.royalsocietypublishing.org/content/10/84/20130246.full.html#ref-list-1>

Subject collections

Articles on similar topics can be found in the following collections

[mathematical physics](#) (40 articles)

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](#)



Research

Cite this article: Schneider CM, Belik V, Couronné T, Smoreda Z, González MC. 2013 Unravelling daily human mobility motifs. *J R Soc Interface* 10: 20130246.
<http://dx.doi.org/10.1098/rsif.2013.0246>

Received: 18 March 2013

Accepted: 15 April 2013

Subject Areas:

mathematical physics

Keywords:

networks, mobile phone,
human dynamics, motifs

Author for correspondence:

Christian M. Schneider
e-mail: schnechr@mit.edu

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsif.2013.0246> or via <http://rsif.royalsocietypublishing.org>.

Unravelling daily human mobility motifs

Christian M. Schneider¹, Vitaly Belik^{1,2}, Thomas Couronné³,
Zbigniew Smoreda³ and Marta C. González^{1,4}

¹Department of Civil and Environmental Engineering, and ⁴Engineering Systems Division, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

²Max Planck Institute for Dynamics and Self-Organization, Am Fassberg 17, 37077 Göttingen, Germany

³Sociology and Economics of Networks and Services Department, Orange Labs, 38 rue du Général Leclerc, 92794 Issy les Moulineaux, France

Human mobility is differentiated by time scales. While the mechanism for long time scales has been studied, the underlying mechanism on the daily scale is still unrevealed. Here, we uncover the mechanism responsible for the daily mobility patterns by analysing the temporal and spatial trajectories of thousands of persons as individual networks. Using the concept of motifs from network theory, we find only 17 unique networks are present in daily mobility and they follow simple rules. These networks, called here motifs, are sufficient to capture up to 90 per cent of the population in surveys and mobile phone datasets for different countries. Each individual exhibits a characteristic motif, which seems to be stable over several months. Consequently, daily human mobility can be reproduced by an analytically tractable framework for Markov chains by modelling periods of high-frequency trips followed by periods of lower activity as the key ingredient.

1. Introduction

Our modern society and the environment are shaped by people's mobility patterns at different scales. Long-time and long-distance trips consist generally of rare and infrequent events such as international flights or movements between cities. By contrast, short-time trips mostly consist of intracity travels such as commuting to work or grocery shopping. These trips exhibit high regularity, typically following the daily circadian rhythm. Studies of human mobility at large scales, motivated by understanding the global spreading of epidemics [1–6], have unravelled interesting properties of the underlying mobility patterns.

Nowadays, large-scale human mobility patterns are described by three widely accepted indicators: the trip distance distribution $p(r)$, the radius of gyration $r_g(t)$ and the number of visited locations $S(t)$ over time [7–9]. The trip distance distribution of the entire population follows a power law $p(r) \sim r^{-\beta}$ with $\beta \approx 1.59$ [7].

Individual trajectories can be extracted from mobile phone data [10–13]. This enables the study of the area an individual visits which is characterized by the radius of gyration $r_g(t)$ [8]. This individual r_g can be understood as the characteristic distance an individual travels during a given time-period t . The distribution of the radius of gyration reveals heterogeneity in the population; most individuals travel within a short radius, but others cover long distances on a regular basis. Thus, each individual follows $p(r)$ within his or her characteristic distance $r_g(t)$. The distribution $p(r_g)$ within the population yields the power law observed in the aggregated trip distance distribution $p(r)$.

The frequent return to previously visited locations is captured by the number of visited places over time $S(t)$. This value grows sublinearly as $S(t) \sim t^\mu$ with $\mu = 0.6$ capturing individuals' tendency for revisiting locations [9]. These three measures contain the basic ingredients to describe the individual trajectories, in which frequent travels occur between a limited number of places, with less frequent trips to new places outside an individual radius. Such behaviour for large time scales can be reproduced by an exploration

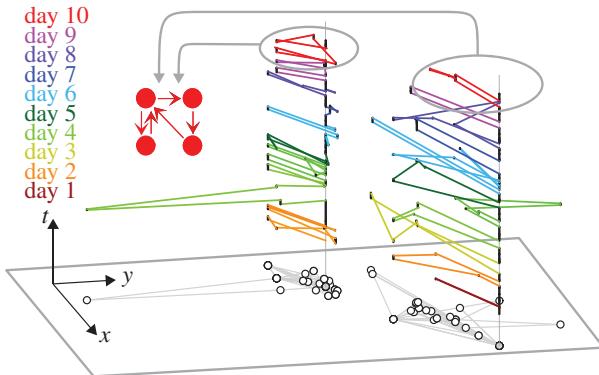


Figure 1. Decomposition of the mobility profile over 10 days into daily mobility patterns for two anonymous mobile phone users. The home location of each user is highlighted and connected over the entire observation period with a grey line. While the entire mobility profiles (black circles and grey lines in the xy -plane) are rather diverse, the individual daily profiles (brown to red from bottom to top for different days) share common features. The aggregated networks consist of $N = 16$ (22) nodes and $M = 37$ (43) edges with an average degree of $\langle k \rangle = 2M/N = 4.6$ (3.9). By contrast, the daily average number of nodes is $\langle N \rangle = 4.4 \pm 1.8$ (3.9 ± 1.3), and the average number of edges is $\langle M \rangle = 5.3 \pm 2.8$ (4.2 ± 2.2). The left user prefers commuting to one place and visits the other locations during a single tour, whereas the right user prefers to visit the daily locations during a single tour. On the last day, both users visit not only four locations, but also share the same daily profile consisting of two tours with one and two destinations, respectively.

and preferential return model with the displacement distribution as an input [9] which can be used to model epidemic spreading on the airline network [14].

However, the current model is designed to capture the long-term mobility behaviour. For example, the number of visited locations $S(t)$ does not show a robust scaling exponent μ , for $t < 24$ h [9]. Additionally, the radius of gyration stabilizes only after a few months of observation [8]. These indications suggest different underlying mechanisms for modelling mobility at the intercity and the intracity scale.

Current studies at the daily, intracity scale focus on forecasting traffic demand and on predicting human decisions based on optimizing a score function or a utility function. Such modelling approaches assume that each individual human tries to minimize his/her effort depending on socio-economic characteristics [15]. Therefore, agent-based models have been deployed, usually based on detailed data from travel surveys [16–23].

In this study, we investigate the common underlying mechanisms for daily human mobility patterns by combining the advantages of different large-scale data sources. In each dataset, we observe ubiquitous daily mobility patterns, which we statistically reproduce with an analytical model. Because the generated patterns of our model are only sensitive to the presence or absence of periods of activity followed by periods of inactivity, it implies that humans' daily trips follow a universal law.

2. Human mobility patterns

Human mobility is characterized by a sequence of visited locations and the trips among them. As an example, we show in figure 1 the aggregated mobility profile of two

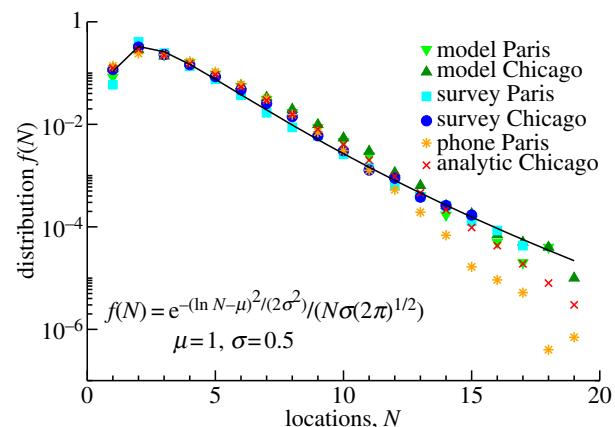


Figure 2. Daily human mobility patterns seem to follow a universal law. The daily number of visited locations can be approximated with a log-normal distribution $f(N) = \exp(-(\ln N - \mu)^2 / (2\sigma^2)) / (\sqrt{2\pi}\sigma N)$ with $\mu = 1$ and $\sigma = 0.5$. The distributions extracted from activity and travel surveys as well as from mobile phone billing data show similar behaviour. Moreover, the distributions of our perturbation model (see §3 and figure 6 for details) generated both analytically and numerically have the same shape. The broad distribution shows that although most of the people visit less than five locations, a small fraction behave significantly differently because people report visits up to 17 different places within a day in our surveys. Note that due to the mobile phone data limitations, the tail of the corresponding distribution is below the other datasets.

users and their corresponding daily profiles for a 10 day observational period. The time-dependent trajectories for different days are coloured from brown (first day) to red (10th day) from bottom to top. The black circles and grey lines in the xy -plane are the projection of the daily trajectories. Both the daily and the aggregated profiles can be described as directed networks, in which nodes represent the visited locations and directed edges stand for trips between them. To classify these networks on a daily basis, we further discard any additional information about the purpose of the activity, the travel time and the activity duration as well as the distances and the number of trips between the visited locations, consequently neither the nodes nor the edges are weighted. Only the trip direction is incorporated by the direction of the edge as highlighted for the last day in figure 1.

We first investigate the distribution of the number of different visited locations, which is the size distribution of the daily networks. As shown in figure 2, the size distributions $f(N)$ of the networks are similar for all datasets (see §3 for more detail). The shape of the observed distributions $f(N)$ can be approximated by a log-normal distribution

$$f(N) \sim \frac{e^{-(\ln N - \mu)^2 / (2\sigma^2)}}{\sigma\sqrt{2\pi}N}, \quad (2.1)$$

with the parameters $\mu = 1 \pm 0.1$ and $\sigma = 0.5 \pm 0.1$. The average number of locations $\langle N \rangle \approx 3$ is small; hence, most people visit only a few locations. In fact, 90 per cent of the population visit less than seven locations on a daily basis. All three datasets follow the same distribution despite the difference between the cities and if the dataset is a travel survey or phone data.

To further study the observed daily mobility patterns, the number of different daily networks is investigated. These networks reveal whether people prefer to visit different locations in a single round tour before returning to the starting

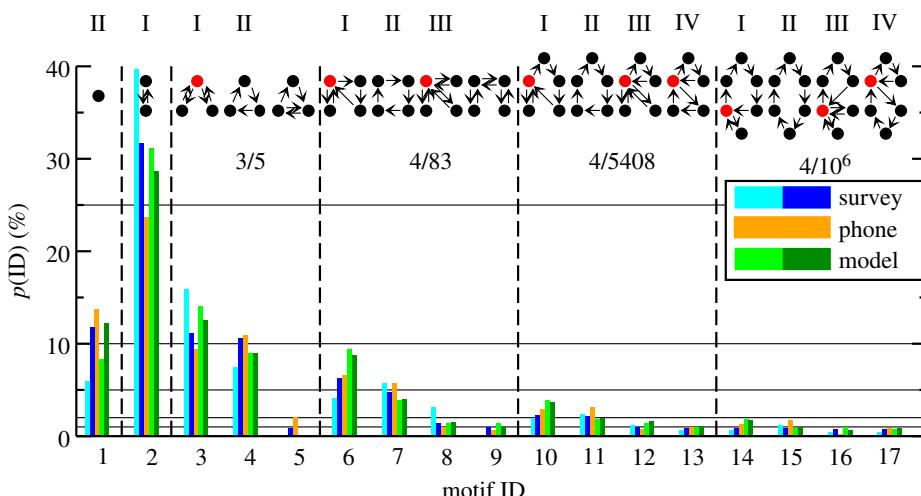


Figure 3. Possible daily mobility patterns are limited, because up to 90% of the identified daily mobility networks can be described with only 17 different motifs. The probability $p(ID)$ to find one of these 17 motifs in the surveys (cyan, Paris; blue, Chicago), the phone data (orange, Paris), and the model (light green, Paris; dark green, Chicago) is presented. The motifs are grouped according to their size separated by dashed lines. For each group, the fraction of observed over feasible motifs N_o/N_f is shown and the central nodes are highlighted. Most motifs can be classified by four rules: (I) motifs of size N consist of a tour with only one stop and another tour with $N - 2$ stops. (II) Motifs of size N consist of only a single tour with N stops. (III) Motifs of size N consist of two tours with one stop and another tour with $N - 3$ stops. (IV) Motifs of size N consist of a tour with two stops and another tour with $N - 3$ stops. Despite the fact that the number of workers is significantly different in both cities, the rank and the probability to find a specific motif exhibit similar behaviour.

location, or if they prefer to return to their starting location before visiting another location. In fact, for a given network size N , N_p edge combinations exist:

$$N_p(N) = 2^{N^2-N}. \quad (2.2)$$

Because we are interested in networks that picture human daily trips, the number of reasonable networks can be significantly reduced mainly due to two constraints: the need for sleep, and the consistency of trips. The need for sleep imposes that the trips start and finish at the same location, most likely at home. The consistency ensures that each of the N locations is visited at least once. These two conditions imply that for $N > 1$ all nodes have at least one ingoing and one outgoing edge. By counting the number of feasible daily networks that fulfil these two constraints, we obtain a large number N_f increasing rapidly with the number of locations ($N_f(1) = 1$, $N_f(2) = 1$, $N_f(3) = 5$, $N_f(4) = 83$, $N_f(5) = 5048$, $N_f(6) = 1\,047\,008$). Nevertheless, up to 90 per cent of the measured trips can be described with only 17 different daily networks for the surveys and the mobile phone data.

We call these 17 daily networks motifs in analogy to motifs in complex networks [24]. Many systems represented as networks consist of various subnetworks, either topological or temporal [25]. If these subnetworks occur more often than in randomized versions of the entire network, these subnetworks are called motifs. Because randomized versions of the mobility networks are not feasible, we call motifs the daily networks which are found on average more often than 0.5 per cent in the datasets (see the electronic supplementary material for further networks). Consequently, nearly the entire aggregated mobility network of a population can be constructed with these motifs.

In figure 3, the motifs obtained from Chicago and Paris surveys, mobile phone data from Paris, and our proposed model are compared. They are ordered by their size and their frequency of occurrence. Although the data sources cover different cities from different countries, the frequencies to observe a specific motif behave similarly. We can suppose

that the extracted motifs are general daily mobility characteristics that can be further used to model and simulate urban activity. The most common motif (ID 2) consists of two visited locations and two trips among them, followed by a motif with only a single location (ID 1). The next likely motifs are three locations with four trips, all starting and ending at the same location (ID 3), or with one round trip (ID 4). Interestingly, in none of the datasets is a motif with size N and more than $N + 2$ trips observed.

All motifs have at most one ‘central’ location, defined as a node with more than two directed edges, except the motifs with ID 5 and ID 9. This central node is the origin for a tour $T(x)$, a trip visiting x other locations before returning to the origin with $x < N$. The presence of a unique central node ensures that the edges of the motifs belong to exactly one tour. Hence, multiple trips along the same directed edge are suppressed, and the entire motifs are composed of a single Eulerian cycle: it is possible to visit all edges exactly once and this path ends at the starting node.

The motifs can be classified by four rules:

- (I) $T(1)$ and $T(N - 2)$
- (II) $T(N - 1)$
- (III) $T(1)$, $T(1)$ and $T(N - 3)$
- (IV) $T(2)$ and $T(N - 3)$

The rule that describes each motif is written on the top of figure 3. If a rule leads to a motif with a tour $T(x)$ visiting a negative number of nodes $x < 0$, then the motif is forbidden. By contrast, if a rule leads to a tour visiting no nodes, then only this tour $T(0)$ is ignored. For a given number of locations N , the likelihood of observing a motif is related to the rule number; thus the most likely motif can be described with the first rule. For $N \leq 6$, the upper limit of daily tours is three; thus the larger the size of the motif the more trips within a tour. Furthermore, we have found that the most common daily networks with more than six locations also follow these rules (see the electronic supplementary material).

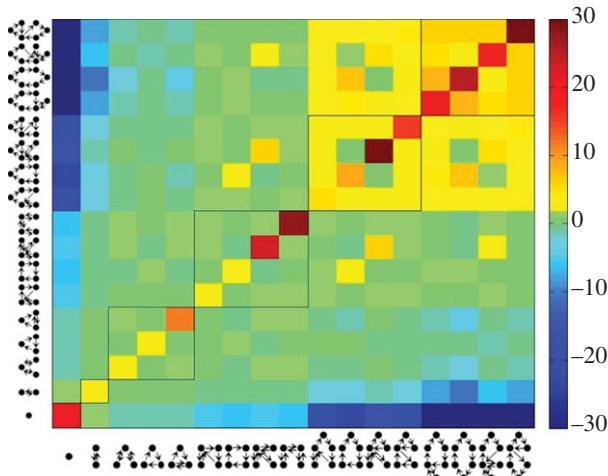


Figure 4. Daily human mobility patterns are stable over several months. The values, calculated by equation (2.3), show how more or less likely a motif is found during the observation period of six months under the condition that the individual has a given motif on another day. Positive values (yellow to red colours) indicate that these motifs are more likely than expected and negative values (cyan to blue colours) that these motifs are suppressed. The probability to find the same daily motif during another day is significantly larger compared with the randomized dataset. Additionally, active users, which visit more than four locations per day, seem to be active over time, whereas inactive users remain inactive. The emerging patterns of transitions between active motifs could be explained by the similarity of motifs. While transitions between motifs of group II are preferred, transitions between groups II and III are suppressed, because the number of tours is most different. As a guide to the eye, motifs with the same number of locations are marked with boxes.

Previous results on human predictability [13], as shown in the trajectories in figure 1, suggest that each individual has a typical daily motif; thus, the observed motifs are similar over several days. To verify this stability, the correlations between motifs of individual users are studied based on phone data, because our surveys provide only information for up to 2 days. The observed sequence is compared with the sequence of an average user based on the distribution from figure 3. In figure 4, the correlations are shown:

$$C_{ij} = \begin{cases} \frac{N(i)N(j)}{N_r(i)N_r(j)} - 1, & \text{for } \frac{N(i)N(j)}{N_r(i)N_r(j)} > 1 \\ 1 - \frac{N_r(i)N_r(j)}{N(i)N(j)}, & \text{for } \frac{N(i)N(j)}{N_r(i)N_r(j)} \leq 1, \end{cases} \quad (2.3)$$

with the observed $N(i)$ and average $N_r(i)$ number of motifs with ID i . First, the highest correlation of each motif is the self-correlation C_{ii} which is usually 10–30 times more likely than expected by selecting individual motifs according to the observed distribution. Second, the likelihood to find a motif with similar number of visited places with small variations (± 2 locations) behaves like the average, but for higher differences, the probability is significantly suppressed. Additionally, active users $N > 4$ seem to be active during the entire observational period, because they have significantly higher probability to visit any motif with $N > 4$. Interestingly, within the blocks of motifs with size four, five and six some correlations are suppressed or enhanced. We observe that the correlations are enhanced if both motifs follow the same rule with different number of visited locations N , for example visiting all locations within one tour. By contrast, the correlations

are suppressed if the motifs are less similar, i.e. if the number of tours differs by more than one. In fact, this is observed for motifs created according to rules (II) and (III).

In general, motifs may not be unique, because a person may repeat a tour several times within a day. However, the repetition of tours is uncommon; thus, an edge corresponds to exactly one trip. In the survey data, the observed motifs without multiple trips are sufficient to reproduce over 95 per cent of the travel behaviour correctly and we observe that tours $T(x)$ with $x > 1$ are performed only once during a day (for details see the electronic supplementary material).

These observations imply that each person has a characteristic daily motif although the visited locations can change. Thus, a user has a personal number of preferred places on a daily basis, which are most likely visited in a specific sequence given by its characteristic motif.

3. Perturbation-based model

It is surprising that nearly the entire population can be described with a few unique daily motifs. To understand this observation, we study the time spent at certain locations as well as the time between the starting time of an activity and the next activity of the same kind.

From both surveys, the frequency of staying at a place for a particular time period is extracted for three groups of activities, home, work and other, as shown in figure 5a. The time spent for working and staying at home is relatively flat distributed with some characteristic durations of 3.5 and 8.6 at work and 14 h at home. By contrast, the probability of an activity at another place decreases with its duration. This staying-time distribution has no characteristic duration, suggesting that the location changes are not distributed evenly over time, but in groups interspersed with periods of inactivity. To support this observation, we study the time between two similar activities, shown in figure 5b. While the time based on home and work is governed by the daily routines, the time between other locations follows a broad distribution. Such short inter-event time dynamics has been reported in specific human activities such as Web browsing, printing patterns, e-mail and phone communication [26–37], but it has not been incorporated in models of human mobility. Inspired by these observations, we developed a perturbation-based model, to reproduce not only the observed daily motifs, but also their frequency of occurrence.

In the following, the model for a non-working (NW) agent is explained and additional, minor features for working (W) agents are described in the electronic supplementary material. Accounting for the difference of home and other locations, the model assumes a fixed activity at home and any number of flexible activities elsewhere (shopping, recreation, etc.). Agents prefer staying at home and perform other activities as a kind of perturbation only; thus they return home after finishing a flexible activity, if they have no other flexible activity scheduled. On the other hand, when people are already perturbed, it is more likely that they perform another flexible activity afterwards (e.g. after having dinner in the city, visiting a nearby bar).

In the model, the day is divided into $K = 48$ 30-min intervals. The actual number of discrete time slots is insignificant as long as it is larger than the maximal number of visited locations $K > 20$. For each of these time slots, the agent

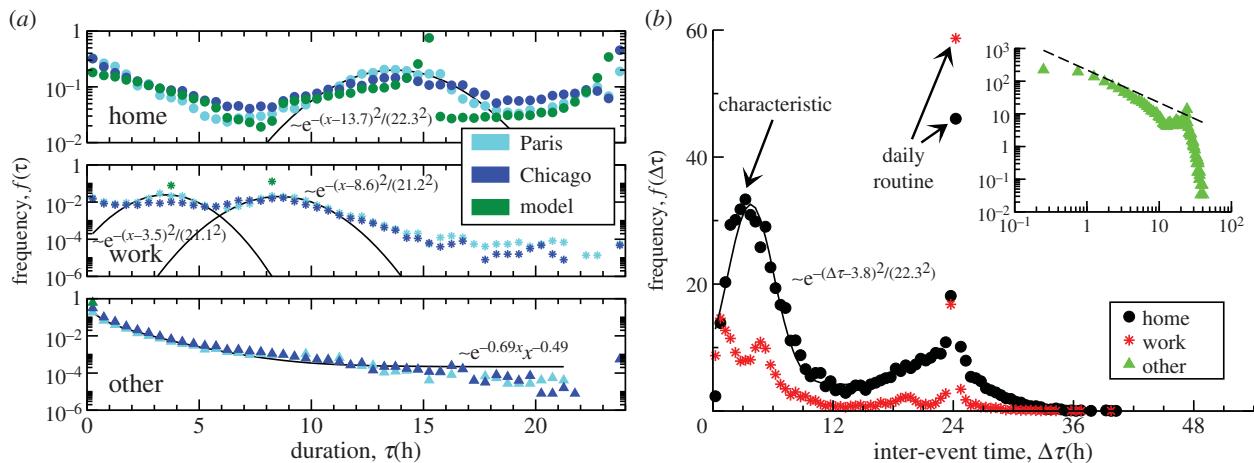


Figure 5. Fundamental differences between home/work and other locations. (a) The duration spent at either home or work is relatively flat distributed with peaks around characteristic time spans of 14 h at home as well as 3.5 and 8.6 h at work. By contrast, the time spent at other activities is broadly distributed. For a guide to the eye, Gaussian distributions are fitted around the characteristic durations for home/work locations and a power law with an exponential cut-off is fitted for other locations. Our model captures these main characteristics. (b) The frequency of observing an inter-event time τ between the beginning of two similar activities, if another location has been visited in between. For the home and work location, daily routines dominate the distribution with additional characteristic times. By contrast, the distribution for other locations exhibits a broad distribution dominated by short inter-event times with a suppressed daily routine. For a guide to the eye, the characteristic inter-event time between home location is approximated by a Gaussian distribution and in the inset a power law with exponent -1 is included.

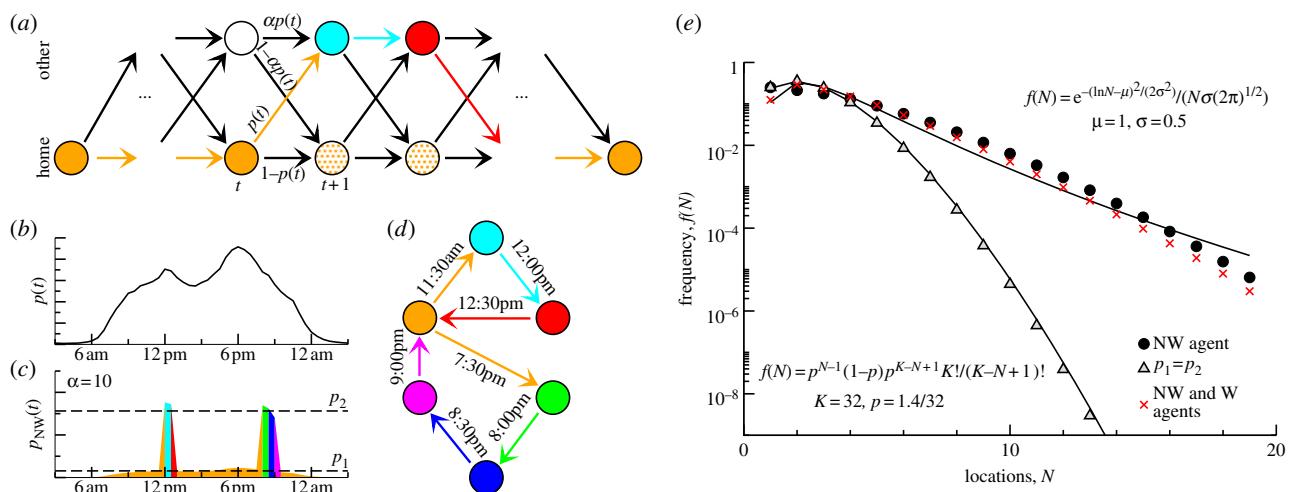


Figure 6. The introduced model is illustrated for a non-working agent. (a) The possible trajectories of an agent are shown. The agent starts the day at home and finishes it at home. At a given time t , depending on the actual location of the agent, the probability to be at home at time $t + 1$ is either $1 - p(t)$ or $1 - \alpha p(t)$ with a parameter α and with probability $p(t)$ and $\alpha p(t)$ the agent travels to another location. The filled circles and the coloured path is an exemplary trajectory. The time-dependent probabilities can be related to the circadian rhythm of activity, shown in (b). (c) The location-dependent probabilities for the exemplary agent with $\alpha = 10$ are shown. The time-dependent probabilities can be approximated by only two values, p_1 and p_2 for being at home and being at another location. With this approximation, the model can be solved analytically. (d) The exemplary trajectory is converted into the corresponding motif with six locations and seven trips among them. (e) The daily number of visited locations obtained from the analytical model under three different conditions is shown. While the removal of workers does not change the tail of the log-normal distribution, $\alpha = 1$ leads to a binomial distribution; thus periods of activities are key for the observed behaviour. Note that the absolute difference between analytical and numerical model is less than 0.01.

receives a task with the corresponding time-dependent probability $p_{NW}(t)$, and assigns it to the next free time slot. Initially, all time slots are free, besides a 9 h sleeping period during night. Because most tasks occur and are executed during daytime, we use the simple assumption that the probability to receive a task is related to the circadian daily rhythm. This rhythm is approximated by the normalized phone activity $p(t)$ of the entire population as shown in figure 6b: $\int_0^{24h} p(t)dt = \gamma_{NW}$ with the parameter γ_{NW} (see the electronic supplementary material). The most important ingredient for modelling the observed motifs from surveys and phone data is the assumption that after receiving

a task $p_{NW}(t) = p(t)$, the probability to get another task $p_{NW}(t + 1) = \alpha p(t + 1)$ for the next time slot is significantly higher, $\alpha > 1$, as shown in figure 6c. For the sake of simplicity, we increase the probability by one order of magnitude $\alpha = 10$. This ensures that the inter-event time distribution of flexible activities is dominated by short times as observed in figure 5b and generates the daily tours. In figure 6a–d, an example of modelling a NW agent is shown. The peaks in $p_{NW}(t)$ in figure 6c correspond to activities outside home.

Note that the model has no assumptions about the locations of the individual tasks, their number or the number of trips. Only the average number of different visited

locations is controlled by the parameter γ_{NW} , and the fraction of working and NW agents is preset. However, this is sufficient to reproduce the overall behaviour of the data as shown in figures 2, 3 and 5. Additionally, the model also reproduces the fraction of trips between home, work and other locations with an absolute error of at most 2 per cent (see the electronic supplementary material).

The model can be treated analytically by mapping it on a coin flipping or independent non-identical Bernoulli trials problem, with the reasonable assumption of only two different probabilities $p_1 = \langle p(t) \rangle$ and $p_2 = 10p_1$ instead of a time-dependent variable (figure 6c). A person, having K free time slots, flips a coin to change the location in the next slot. A success H leads to stay at home or return home, whereas failures T lead to the exploration of new locations. The coin flipping occurs with different probabilities dependent on the current state:

$$H \xrightarrow{p_1} T,$$

$$H \xrightarrow{1-p_1} H,$$

$$T \xrightarrow{p_2} T$$

$$T \xrightarrow{1-p_2} H.$$

and

By applying the modified finite Markov chain embedding technique [38] for independent non-identical Bernoulli trials, the probability for the number N of locations visited during a day or equivalently the number of successes $P(N)$ after K Bernoulli trials can be written as

$$P(N) = \xi_0 \left(\prod_{t=1}^K \Lambda_t \right) U'(C_N), \quad (3.1)$$

with ξ_0 being an initial condition vector in the state space of the corresponding Markov chain, Λ_t the transition probability matrix, and $U'(C_N)$ a transposed vector corresponding to the subspace with N successes (for details see the electronic supplementary material). As one can see in figure 2, this simple coin flipping model can reproduce the empirical findings very well.

To confirm that the assumption $\alpha \gg 1$ is the key to get the broad distribution of the number of daily visited locations, we show in figure 6e the analytical results for three different models: one with two kinds of agents, one with only non-workers and one with only one probability $\alpha = 1$. While the presence of two kinds of agents has a minor impact on the overall motifs and their size distribution, the removal of the perturbation ($p_2 = p_1$) changes the results from an approximately log-normal size distribution, to a binomial size distribution. Moreover, not only the motif distribution changes, but different motifs which are not present in the surveys, mostly star-like ones, emerge. Therefore, the ‘perturbed’ behaviour $p_2 = 10p_1$ is the crucial ingredient to reproduce daily mobility.

4. Final remarks

Advances in transforming large data into meaningful information are essential to improve our understanding of socio-technical systems. In our study, we contribute to this end by analysing networks of daily trips obtained from individuals' surveys and anonymized mobile phone data. We found that both travel surveys and phone traces from two different

cities reveal the same set of ubiquitous networks that we called motifs. We can suppose that these motifs are general human mobility characteristics that can be further used to model and simulate urban activity. Besides, we found that perturbed states with periods of high activity followed by periods of low activity is the indispensable ingredient to correctly reproduce those motifs. We remark that owing to the limited observation period of at most 2 days in our survey, the question whether a heavy tail occurs in the inter-event time distribution in figure 5b remains open.

Our model successfully reproduces the frequency of visiting different locations and the occurrence rate of the motifs, but it is designed for a single day and therefore it does not incorporate the correlations of motifs between different days. The model captures main characteristics of the duration spent at home by assuming fixed duration for the other activities. The model's inter-event time distributions share some common features with the data, but owing to the duration differences as well as the restriction to a single day it cannot accurately reproduce the observed distributions (see the electronic supplementary material).

The future avenues for related research are diverse. Understanding daily routines promises a better assessment of planning and control, which is the core interest of urban and epidemiological applications. Our findings reduce the dimensionality of choices in agent-based modelling helping to enhance current urban simulators (<http://www.matsim.org/>, <http://code.google.com/p/transims/>). In epidemic spreading, usually only up to three locations, daily visited by a host, are considered in modelling contagious dynamics [39–42]. Thus, our presented insights can straightforwardly extend mobility in current epidemiological models.

5. Material and methods

To identify motifs, we use three different datasets: a survey and mobile phone billing data from Paris and a survey from Chicago (<http://www.cmap.illinois.gov/travel-tracker-survey>). In the surveys, 23 764 and 23 429 weekdays of people were selected in such a way that the data are representative for the entire population of Chicago and Paris, respectively. In the Chicago survey, each participant answered a questionnaire with his/her activity information for one or two entire days, containing the following information: weekday, duration, location, reason for and mode of trip. With this information, it is possible to reproduce the entire daily activity patterns of the anonymous individuals. The Paris survey has the same information, but instead of geographical locations only the trip lengths are provided. Because weekday and weekend behaviour can be rather different, we focus in this study only on weekdays.

From phone billing data of millions of mobile phone users, the extraction of relevant information needs preprocessing. The phone company provides information about the incoming and outgoing calls and short-message services. Thus, we have locations of the operating towers, time of the events and user identification numbers. With this information, we reconstruct daily mobility networks of the users during a six month period. The main challenge is converting call information into the corresponding mobility profile of a user. Therefore, only the 39 820 most active users are investigated according the following scheme (the rules are visualized in the electronic supplementary material, figures S1 and S2):

- the day, starting at 03.00, is divided into 48, 30-min slots for each of the 154 days;

- to remove towers which are only used during travel, all towers which are less frequently visited than a certain threshold are ignored; in this study, less than 0.5 per cent during the entire observational period;
- to eliminate signal transitions between neighbouring towers, these towers are merged for one day, if more than three back and forward transitions between them are recorded during a single day;
- to remove towers used during travel on daily basis, records are taken into account only if the next records have the same tower location;
- to identify an activity location, only the most frequently observed location during each time slot is assigned as an activity location for this time slot;
- a day is discarded, if less than a certain number (in this case eight) of time slots exhibit location information. Too small a number would favour smaller motifs, whereas too large a threshold would exclude too many individuals. The results are stable for different threshold values;
- to overcome the small number of night calls, the location which is visited most frequently during all nights between 24.00 and 06.00 of a single user is assigned as the user's home location; in our survey this assumption correctly identifies over 98 per cent of the home locations for a single day. User starts and finishes its day at home, if the user has no other information in the corresponding night-time slots at 03.00 and 03.30; and
- based on the activity locations for each time slot, the motifs shown in figure 3 are constructed for weekdays only.

We have published C++ code of our proposed model, the algorithms how to identify motifs and simulated data to test all algorithms on our website at <http://humnetlab.mit.edu/downloads>.

V.B. gratefully acknowledges the financial support by the Volkswagen Foundation. This work was funded by New England UTC Year 23 grant, awards from NEC Corporation Fund, the Solomon Buchsbaum Research Fund.

References

1. Anderson RM, May RM. 1992 *Infectious diseases in humans*. Oxford, UK: Oxford University Press.
2. Lloyd AL, May RM. 2001 How viruses spread among computers and people. *Science* **292**, 1316–1317. (doi:10.1126/science.1061076)
3. Hufnagel L, Brockmann D, Geisel T. 1992 Forecast and control of epidemics in a globalized world. *Proc. Natl Acad. Sci. USA* **101**, 15124–15129. (doi:10.1073/pnas.0308344101)
4. Colizza V, Pastor-Satorras R, Vespignani A. 2007 Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.* **3**, 276–282. (doi:10.1038/nphys560)
5. Balcan D, Colizza V, Goncalves B, Hu H, Ramasco JJ, Vespignani A. 2009 Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci. USA* **106**, 21484–21489. (doi:10.1073/pnas.0906910106)
6. Belik V, Geisel T, Brockmann D. 2011 Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X* **1**, 011001. (doi:10.1103/PhysRevX.1.011001)
7. Brockmann D, Hufnagel L, Geisel T. 2006 The scaling laws of human travel. *Nature* **439**, 462–465. (doi:10.1038/nature04292)
8. González MC, Hidalgo CA, Barabási A-L. 2008 Understanding individual human mobility patterns. *Nature* **453**, 779–782. (doi:10.1038/nature06958)
9. Song C, Koren T, Wang P, Barabási A-L. 2010 Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823. (doi:10.1038/nphys1760)
10. Eagle N, Pentland A. 2006 Reality mining: sensing complex social systems. *Pers. Ubiquit. Comput.* **10**, 255–268. (doi:10.1007/s00779-005-0046-3)
11. Onnela J, Saramaki J, Hyvonen J, Szabo G, Lazer D, Kaski K, Kertész J, Barabási A-L. 2007 Structure and tie strengths in mobile communication networks. *Proc. Natl Acad. Sci. USA* **104**, 7332–7336. (doi:10.1073/pnas.0610245104)
12. Eagle N, Pentland A, Lazer D. 2009 Inferring friendship network structure by using mobile phone data. *Proc. Natl Acad. Sci. USA* **106**, 15274–15278. (doi:10.1073/pnas.0900282106)
13. Song C, Qu Z, Blumm N, Barabási A-L. 2010 Limits of predictability in human mobility. *Science* **327**, 1018–1021. (doi:10.1126/science.1177170)
14. Nicolaides C, Cueto-Felgueroso L, González MC, Juanes RA. 2012 Metric of influential spreading during contagion dynamics through the air transportation network. *PLoS ONE* **7**, e40961. (doi:10.1371/journal.pone.0040961)
15. Zipf GK. 1949 *Human behaviour and the principle of least effort: an introduction to human ecology*. Cambridge, MA: Addison-Wesley.
16. Beckman RJ, Baggerly KA, McKay MD. 1996 Creating synthetic baseline populations. *Transp. Res. A* **30**, 415–429. (doi:10.1016/0965-8564(96)00004-3)
17. Ben-Akiva ME, Bowman JL. 1998 Integration of an activity-based model system and a residential location model. *Urban Stud.* **35**, 1231–1253. (doi:10.1080/004209984529)
18. Arentze T, Hofman F, van Mourik H, Timmermans H. 2000 ALBATROSS: multiagent, rule-based model of activity pattern decisions. *Transp. Res. Rec.* **1706**, 136–144. (doi:10.3141/1706-16)
19. Kitamura R, Chen C, Pendyala RM, Narayanan R. 2000 Microsimulation of activity-travel patterns for travel demand forecasting. *Transportation* **27**, 25–51. (doi:10.1023/A:1005259324588)
20. Keuleers B, Wets G, Arentze T, Timmermans H. 2001 Association rules in identification of spatial-temporal patterns in multiday activity diary data. *Transp. Res. Rec.* **1752**, 32–37. (doi:10.3141/1752-05)
21. Axhausen KW, Zimmermann A, Schönfelder S, Rindsfüser G, Haupt T. 2002 Observing the rhythms of daily life: a six-week travel diary. *Transportation* **29**, 95–124. (doi:10.1023/A:1014247822322)
22. Schlich R, Axhausen KW. 2003 Habitual travel behaviour: evidence from a six-week travel diary. *Transportation* **30**, 13–36. (doi:10.1023/A:1021230507071)
23. Charypar D, Nagel K. 2005 Generating complete all-day activity plans with genetic algorithms. *Transportation* **32**, 369–397. (doi:10.1007/s11116-004-8287-y)
24. Alon U. 2007 Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**, 450–461. (doi:10.1038/nrg2102)
25. Kovanen L, Karsai M, Kaski K, Kertész J, Saramäki J. 2011 Temporal motifs in time-dependent networks. *J. Stat. Mech.* P11005. (doi:10.1088/1742-5468/2011/11/P11005)
26. Eckmann J, Moses E, Sergi D. 2004 Entropy of dialogues creates coherent structures in e-mail traffic. *Proc. Natl Acad. Sci. USA* **101**, 14333–14337. (doi:10.1073/pnas.0404572101)
27. Barabási A-L. 2005 The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211. (doi:10.1038/nature03459)
28. Harder U, Paczuski M. 2006 Correlated dynamics in human printing behavior. *Physica A* **361**, 329–336. (doi:10.1016/j.physa.2005.06.079)
29. Candia J, González MC, Wang P, Schoenhar T, Madey G, Barabási A-L. 2008 Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A* **41**, 224015. (doi:10.1088/1751-8113/41/22/224015)
30. Malmgren RD, Stouffer DB, Motter AE, Amaral LA. 2008 A Poissonian explanation for heavy tails in e-mail communication. *Proc. Natl Acad. Sci. USA* **105**, 18153–18158. (doi:10.1073/pnas.0800332105)
31. Iribarren JL, Moro E. 2009 Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.* **103**, 038702. (doi:10.1103/PhysRevLett.103.038702)
32. Miller KJ, Sorensen LB, Ojemann JG, den Nijs M. 2009 Power-law scaling in the brain surface electric

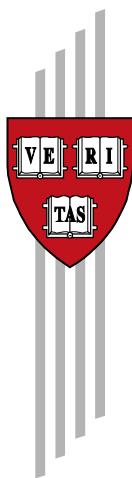
- potential. *PLoS Comput. Biol.* **5**, e1000609. (doi:10.1371/journal.pcbi.1000609)
33. Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA. 2009 Scaling laws of human interaction activity. *Proc. Natl Acad. Sci. USA* **106**, 12640–12645. (doi:10.1073/pnas.0902667106)
34. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A. 2010 Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.* **105**, 158701. (doi:10.1103/PhysRevLett.105.158701)
35. Hennig H, Fleischmann R, Fredebohm A, Hagemayer Y, Nagler J, Witt A, Theis F, Geisel T. 2011 The nature and perception of fluctuations in human musical rhythms. *PLoS ONE* **6**, e26457. (doi:10.1371/journal.pone.0026457)
36. Jo H-H, Karsai M, Kertész J, Kaski K. 2012 Circadian pattern and burstiness in mobile phone communication. *New J. Phys.* **14**, 013055. (doi:10.1088/1367-2630/14/1/013055)
37. Karsai M, Kaski K, Barabási A-L, Kertész J. 2012 Universal features of correlated bursty behavior. *Sci. Rep.* **2**, 397. (doi:10.1038/srep00397)
38. Fu JC, Koutras MV. 1994 Distribution theory of runs: a Markov chain approach. *J. Am. Stat. Assoc.* **89**, 1050–1058. (doi:10.1080/01621459.1994.10476841)
39. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS. 2006 Strategies for mitigating an influenza pandemic. *Nature* **442**, 448–452. (doi:10.1038/nature04795)
40. Viboud C, Bjornstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. 2006 Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–451. (doi:10.1126/science.1125237)
41. Ciofi degli Atti ML, Merler S, Rizzo C, Ajelli M, Massari M, Manfredi P, Furlanello C, Tomba GS, Iannelli M. 2008 Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios. *PLoS ONE* **3**, e1790. (doi:10.1371/journal.pone.0001790)
42. Ajelli M, Gonçalves B, Balcan D, Colizza V, Hu H, Ramasco JJ, Merler S, Vespignani A. 2010 Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. *BMC Infect. Dis.* **10**, 190. (doi:10.1186/1471-2334-10-190)

Evidence That Calls-Based and Mobility Networks Are Isomorphic

Michele Coscia and Ricardo Hausmann

CID Working Paper No. 309
December 2015

© Copyright 2015 Coscia, Michele; Hausmann, Ricardo;
and the President and Fellows of Harvard College



Working Papers

Center for International Development
at Harvard University

Evidence that Calls-based and Mobility Networks are Isomorphic

Michele Coscia^{1,*} and Ricardo Hausmann¹

1 Center for International Development, Harvard University, Cambridge MA, USA

* E-mail: Corresponding Author michele_coscia@hks.harvard.edu

Abstract

Social relations involve both face-to-face interaction as well as telecommunications. We can observe the geography of phone calls and of the mobility of cell phones in space. These two phenomena can be described as networks of connections between different points in space. We use a dataset that includes billions of phone calls made in Colombia during a six-month period. We draw the two networks and find that the social network resembles a higher order aggregation of the mobility network and that both are isomorphic except for a higher spatial decay coefficient of the mobility network relative to the social network: when we discount distance effects on the social connections with the same decay observed for mobility connections, the two networks are virtually indistinguishable.

Introduction

Telecommunications was supposed to bring forth the death of distance in social relations. New technologies permit people to communicate instantaneously around the globe. Yet, many social and economic characteristics and behaviors remain strongly affected by distance, such as trade [1], investment [2], research and development, and knowledge spillovers. This literature has looked at the impact of distance on the patterns of patent citation [3], of R&D and patent output [4,5], of R&D and productivity [6], and on the sales of subsidiaries of multinational corporations [7].

New kinds of data have allowed the study of human behavior in space. These methods include tracking the actual movements of people across the territory, by looking at how dollar bills travel across the US [8], by looking at the social connections they establish [9,10], or by directly observing human mobility [11]. Previous research has shown that on-line social media [12] and Twitter [13] also are strongly affected by distance, a fact that has often been interpreted as a consequence of homophily, meaning that people tend to connect with people similar to themselves [14]. But homophily must be itself endogenous to some other forms of human interaction that are affected by distance. Predicting mobility and social ties has had many applications in computer science [15–18], especially using temporal patterns [19]. Applications include, for instance, the impact of social influence on usage of city areas [20,21], ride sharing services [22,23], the impact of epidemics [24–26] and the study of mobility motifs [27]. Researchers have been able to relate (and predict) relevant facts about the economy by looking at mobility and social media usage [28,29].

In this paper we advance this research by studying the relationship between the social network as captured by cell phone calls and the mobility networks as captured by cell phones movements. We observe more than two billion calls made by around seven million phone numbers in Colombia during a six-month period. We do not have access to any phone numbers or identity of the phone users, and no personal data has been shared by the telecommunication operators. Each phone number is encrypted and anonymized and we preserve the users' privacy by aggregating our data at the municipal level. Each phone is associated with the municipality where it spends most of the time. We are then able to reconstruct the network of social relations across Colombian municipalities by looking at the phone calls between municipalities. At the same time, we are able to track the movements of each phone, since we know the cellphone tower with which it connects to initiate its calls. We use this data to determine the network of human flows across municipalities. Our aim is to study the relationship between these two

networks. To do so, we group municipalities into coherent modules or clusters: municipalities in the same module have strong relations to each other and weak relations to municipalities outside the module. We then compare the calls-based modules with the mobility modules.

As in many other areas of human activity, we also find that social connections based on phone calls decay strongly with distance, even though the cost of the calls themselves are unaffected by distance. This means that the intensity of the underlying social interactions that make people call each other do decay with distance. We find that calls-based modules are geographically compact.

Our main finding in this paper is that the social network captured by phone calls is a higher-order aggregation of human mobility. When we compare the calls-based modules with the mobility modules, we see that social modules are larger and include many neighboring mobility modules. Each mobility cluster can be assigned to a parent social cluster and very few municipalities escape this hierarchical order – a number that is an order of magnitude lower than random expectation. To the best of our knowledge the intimate and hierarchical relationship between the calls-based and the mobility networks is a new result. We push the result further by rescaling social connections as if they were influenced by distance to the same degree as mobility connections. When performing this operation, we observe a remarkable correlation between the rescaled social network and the mobility network. We present this observation as the first evidence that social and mobility networks are isomorphic, which is a stronger results than the one so far discussed in the literature, where the two types of networks are just considered correlated. This better understanding of the relationship between telecommunications-based and face-to-face social life could improve our current applications that rely on human mobility, for example geographical marketing studies [30], and the investigation of the ties between economic development and mobility [31].

Mobility clusters in Colombia have already been delineated and studied recently [32], using transportation-based commuting data and focusing exclusively on the most populous areas of Colombia. We use these results to test the robustness of our subdivision of the Colombian territory. Our results are in agreement with this independent study, implying that our data is capturing a robust pattern.

Results

Network Topology

From cellphone usage data we are able to create a calls-based social network S of Colombia. In this network we aggregate the data at the municipality level, connecting two municipalities if there is a significant number of calls between them. For details about the municipality aggregation, edge creation and significance threshold, see the Method section. A similar procedure is employed to build the mobility network M . In this case, too, we aggregate at the municipality level. Municipalities are connected if we observe a significant number of trips flowing from one municipality to the other. Both networks are asymmetric, i.e. the strength of the connection from municipality m_1 to municipality m_2 is not the same of the one from m_2 to m_1 . This asymmetry is significant in both the mobility and in the calls-based network. In the mobility network we are tracking movements from one place to another. Thus, we expect to find popular places attracting visitors more than other places, in line with classical geographic theories like the central place theory [33]. In the social network, we also expect an asymmetry due to the nature of the data: some people are initiators of social relationships and some are social attractors. The asymmetric nature of some social relations has been studied multiple times in the literature [34–36].

Figures 1 and 2 respectively depict the resulting social and mobility networks. We calculate some topological features of both networks and report the results in Table 1. We expect to find a strong effect of distance on the mobility network and to find smaller effects on the phone calls network. After all, mobility requires significant energy and time to move our bodies across the country, whether by road or plane. Phone calls, on the other hand, enable people to talk instantaneously with anybody in the country and costs are not affected by distance. We find that calls decay less strongly with distance but they still

conform to compact topological features.

The networks are built selecting the most significant edges, using a threshold which can be interpreted as the p-value of the considered connections (for a full explanation about why this is the case, we refer to the original paper proposing this thresholding technique [37]). In the calls-based network we impose a much stricter significance threshold (0.0001, while the mobility threshold equals to 0.01). Even with a much stricter significance threshold, we still obtain more social edges (9,639) than mobility edges (6,614). Note that the different thresholds are chosen in order to minimize the amount of noisy edges. Since there are more candidate edges in the social network, we need to impose a stricter threshold.

There is a pronounced degree disassortativity and lower reciprocity in the social network. Degree disassortativity means that low degree nodes tend to connect to high degree nodes, in a hub-and-spoke logic. These hubs, in turn, do not reciprocate connections, thus there are low levels of bidirectional links. These facts suggest that there is a large difference between in and out degree in the calls network. In fact, the in and out degree distributions in Figure 3 show that they are very different in the calls network, while the distributions vastly overlap in the mobility network. These facts point to the existence of social aggregators: the in-degree distribution looks scale free, while the out-degree does not, and this kind of differential scaling has been observed in directed social networks. It is usually interpreted as a limited bandwidth effect: the number of people a person can follow is bounded, but in principle a single superstar could be followed by everybody [34, 35]. On the other hand, the mobility network has only a few medium scale hubs (e.g. big cities like Bogota and Medellin). Neither the in- nor the out- degree distributions in the mobility network are scale free and they overlap to a larger extent. Another hint that calls span more freely across the territory is given by the average path length in the network, which is lower in the calls-based network (Table 1).

The analysis of simple topological properties of the calls-based and mobility networks seems to suggest a significant difference between them. We now turn to a more advanced analysis of network clusters which aims to show the common properties of the two structures.

Network Clusters

While the cost of a national phone call does not change with distance, the social network based on calls does decay with distance, thus exhibiting spatial clustering. For this paper we define network clusters (or communities) as a set of nodes that are densely connected to each other and sparsely connected with the rest of the network. Network clusters in the calls-based social network are one measure of the intensity of social interactions between municipalities. The mobility-based network uncovers relationships that require face-to-face interaction. The algorithm used to detect clusters is Infomap [38] and it calculates the optimal number of clusters to minimize the codelength of the representation. Codelength is an information theoretic concept: it calculates the number of bits required to encode all nodes in the network, given their affiliation to the identified communities. Higher codelength means that more bits are required, implying that separating nodes into a few clusters is not enough because clusters are not cleanly separated. Note that codelength is dependent on the number of nodes in the network: more nodes require more bits to be encoded. But since in our case the two networks have the same number of nodes, the comparison of the codelengths is meaningful. Table 1 reports the codelengths for both networks, showing that, as expected, the calls network has a higher codelength. Comparing the two networks, it is clear that the calls-based clusters are larger in the sense that they include more nodes than the mobility clusters (in Figures 1 and 2 nodes are color-coded depending on the community to which they belong). As a consequence, there are fewer phone-call clusters (22) than mobility clusters (81).

We interpret this result as a reflection of the fact that calls are less affected by distance than mobility connections creating a higher degree of interconnectivity across more distant municipalities. A quantitative estimation of this is the codelength measure.

In Figure 4 we display the territorial distribution of the calls-based and mobility clusters in Colombia. Note that white municipalities are excluded because they possess no cellphone towers. In these municipi-

palities cellphone connection signal is either very poor or non-existent (see Figure 5). We exclude these areas.

The analysis of the cluster maps in Figure 4 reveals that the calls-based social structure is influenced by geographical distance. If distance did not affect the calls-based social network, its clusters would not be spatially contiguous and we would observe enclaves and long-range structures. Instead, we see that the social clusters are highly compact, with very few and small exceptions. This is confirmation of previous works on the relationship between social ties and distance [12, 14].

We extend these previous results by noticing the overlapping relationship between the calls-based social clusters and the mobility clusters. By visual inspection, one can perceive that the smaller mobility clusters appear to be included in the larger social clusters. Visual inspection is confirmed by the overlap analysis. We calculate the degree of overlap of all mobility clusters with all social clusters, by counting the fraction of the nodes included in a cluster that are also included in another cluster. If a mobility cluster contains 10 nodes and 9 of them are also included in calls-based social cluster, then the corresponding overlap is equal to .9. We then associate each mobility cluster to a “parent” social cluster, that is the social cluster containing most of the child cluster’s nodes. Finally, we estimate the degree of mismatch by counting all nodes that are in a mobility cluster but are not in the corresponding parent social cluster. The observed mismatch is equal to 9.73%, meaning that 9.73% of nodes in the mobility clusters are not present in their parent social cluster.

We test the significance of this observation through a null model. In the null model we generate 22 random social clusters and 81 random mobility clusters. Each random cluster contains the same number of nodes of its corresponding real-world cluster, but its members are chosen randomly. We calculate the mismatch ratio using the same procedure described above. The average mismatch ratio observed in the null models is around 72%. We ran 10,000 iterations of the null model and Figure 6 reports the resulting mismatch distribution. Given the distribution’s average and standard deviation we can conclude that the observed mismatch ratio carries $p \sim 0$. We obtain a very comparable expectation and standard deviation from the null model if we randomize only the social clusters, keeping the actual mobility clusters fixed.

Note that the result could be explained by noticing that the mobility connections are a subset of the social connections. While this is true (80% of mobility edges are also in the social network), it is not a complete explanation of our results. The mobility edges need to be a very specific subset of the social edges. In fact, if we draw random social edge subsets of equal size to the mobility edges, we do not obtain comparable mobility clusters.

Our interpretation of these results is that there is a significant overlap between the calls-based network and the mobility network in Colombia but calls-based relationships are less influenced by distance than relationships that involve mobility. This can be shown by creating a rescaled version of the calls-based network where we rescale the weights by the expected distance-based decay of the mobility network. Given two municipalities m_1 and m_2 , $S(m_1, m_2)$ and $M(m_1, m_2)$ are the weights of their connections in the social and mobility networks, respectively; and $\delta(m_1, m_2)$ is their spatial distance. We can rescale S as follows:

$$S'(m_1, m_2) = S(m_1, m_2) \times f(\delta(m_1, m_2)),$$

where $f(\delta(m_1, m_2))$ is the expected mismatch between social and mobility links at the given distance. The details on the formulation of f are reported in the Methods section. In practice this equation normalizes the weights in S' by scaling S weights as if distance affected them as much as it affects M weights. S' is what S would look like if phone calls were affected by distance as much as mobility (M).

If mobility clusters are really distance-bounded calls clusters, then S' clusters should match perfectly with M clusters. We calculate S' clusters following the same procedure used for M clusters. We obtain 90 clusters, which is closer to the number of M clusters (81) than to the number of S clusters (22). We compare the S' clusters with M clusters by calculating the Normalized Mutual Information (NMI), which is a standard way to compare partitions. NMI evaluates the agreement of two partitions and it

equals 1 for perfectly aligned clusterings and 0 for random clustering. In this test, NMI equals 0.89. This remarkable matching between the partitions confirms that rescaling social connections to take into account the effect of distance generates a structure comparable to the one observed by tracking human mobility. This confirms the hierarchical relationship between calls-based and mobility ties. Calls-based clusters can be considered the parents of several mobility clusters. Telecommunications does allow for spatially more extended connections than face-to-face interactions by reducing the impact of distance. However, this impact is attenuated, not eliminated. As a consequence, the calls-based network appears as a higher-order aggregation of the mobility network.

To check that our results are not caused by a particular feature of our community discovery algorithm but by the data itself, we apply our algorithm to a different dataset. In particular, as mentioned above, Duranton identified spatial modules using mobility data based on commuting data [32]. We consider the subdivision of the Colombian territory made by Duranton. This study detected shorter-range clusters, with only 170 municipalities belonging to clusters larger than one municipality. These Duranton clusters are almost completely included in our mobility clusters. We assigned each Duranton cluster to a parent mobility cluster with the same logic we applied before for finding the parent phone-call cluster of a mobility cluster. In this case, we found a mismatch for only seven municipalities out of 170. This means that our mobility clusters are a hierarchical level above the Duranton clusters and they almost completely include them.

We validate our results against two possible edge thresholding criticisms. First: to build the social and mobility networks of Colombia, we had to filter out many connections between cities, with a harsh threshold. We did so because the sparser networks are less noisy and easier to visually inspect and interpret. We check whether our results are influenced by this choice, and whether our results would be very different if we used the entire set of connections with no thresholding. We reject this criticism by calculating phone-call and mobility clusters on the full set of connections. We cannot use a community discovery algorithm, because the un-thresholded networks are almost fully connected. Thus, we run the matrix clustering algorithm k-Means instead. We fix k to 22 and 81 for the social and mobility network, respectively. To compare the agreement between matrix and network clusters we use NMI scores, as done previously for S' clusters. For the social and mobility clusters we obtained NMI values equal to 0.66 and 0.81, respectively. Given that we obtain a strong alignment for the filtered and non-filtered networks, we can conclude that thresholding did not significantly impact our results.

The second criticism involves a multiple hypothesis testing argument. When selecting edges with $p < 0.01$, we are ensuring that the expected number of noisy edges is 1%. However, we could have been stricter and apply a Holm-Bonferroni correction [39] to ensure that the probability of having one noisy edge is 1%. We show that both strategies yield comparable results. We create both a social and a mobility network using the Holm-Bonferroni correction fixing the actual p-value at 0.01, for both networks. We then calculate the network clusters of these robust networks and we compare them with the network clusters we discussed so far. We obtain a NMI score of 0.86 for the social network and of 0.95 for the mobility network. This proves that our threshold choice did not have a significant influence on the results. In particular, the reason lies in the fact that the Infomap algorithm is designed to handle a certain amount of noisy edges, as many analytic techniques dealing with complex networks are.

Note that the first criticism stated that our threshold was too strict, while the second proposed that it was not strict enough. By addressing both criticisms, we showed the full independence of our results from the threshold choice.

Discussion

In this paper we analyzed the calls-based and mobility networks in Colombia. The calls-based network is obtained by connecting each Colombian municipality with another municipality if we observe a significant number of phone calls flowing between them. The mobility network is built by observing the physical

flow of phones, as measured by the towers through which they connect. We find that the phone-call based network has a structure that is isomorphic to that of the mobility network with connections that decay less strongly with distance.

We confirm that, as with many other facets of social and economic life, the frequency of phone-call and travel decay with distance. Our novel contribution is to show the relationship between the phone-call and mobility borders: mobility borders encompass smaller areas and they are mostly included in a parent phone-call border. This is evidence of the fact that the social relations that are expressed through phone-calls occur in a space that is a higher-order aggregation of the face-to-face relations. Further, we rescale the phone-call network, so that connections decay as strongly as the mobility network. This operation also reveals the very strong similarity of the two structures, expressed in the fact that we obtain the same network clusters.

We consider these results as highly suggestive. The hierarchical relationship between the phone-call and the mobility networks, and their postulated identity, is worth further investigation. In particular, for this paper we limited ourselves to establishing it. Future work could explore the causal mechanism behind this similarity. Which social relations require face-to-face interactions and which can be carried through telecommunications? To what degree are these two forms of communication complements and to what degree are they substitutes? Do telecommunications-based relationships decay over time unless reinforced by lower-frequency face-to-face relations? Could we envision a future where more efficient transportation modes or more realistic telecommunication devices would make social and mobility clusters equivalent?

These networks can also be used to explore the diffusion of different forms of economic and social behavior and activity over space. What are the phenomena that diffuse through face-to-face connections vis-a-vis those that can tolerate longer distance interactions through telecommunications? What are the likely effects of improvements in physical connectivity vis-a-vis telecommunications in the diffusion of the different social phenomena of interest in different areas?

Materials and Methods

This paper is based on data obtained from telecommunications operators in Colombia. The data includes the metadata of all phone call records initiated by a cellphone card issued by one of these operators in Colombia. For each phone call, we have the following information:

- Encrypted and anonymized ID of the caller, consistent over the dataset (the same ID is always associated to the same cellphone card);
- Encrypted and anonymized ID of the callee, again consistent over the dataset. Note that while source phone numbers are all part of the operators networks, target phone numbers can be from any company and even from outside the country;
- Date and time of the call: the moment in time when the call started. The granularity of the data is at the second, for instance one call started on December 1st, 2013 at 6:07:41 PM;
- ID of phone tower: to which phone tower the source cellphone connected to initiate the call. This ID can be crossed with metadata we have about the cellphone tower to pinpoint the location of the caller when he initiated the call;
- Length of call: how many seconds the call lasted. We drop this information as it is of no interest for this study.

Note that the encrypted and anonymized phone numbers are provided by the telecommunication operators such that it is impossible to identify any individual.

We are unable to share the raw data used for this study for two reasons: there are privacy concerns of potentially identifiable individuals and the complete dataset is of prohibitive size ($\sim 250\text{GB}$). However, we can share the minimum data set necessary to replicate the analysis (which has been included as Supporting Information).

From our data, we see that the operators which provided the data clear on average approximately ten million calls every day, with a clear weekly pattern, depicted in Figure 7 (left). We can see that the market is pretty stable, without noticeable long term variations. As we already reported in Figure 5, cellphone coverage in Colombia is very dishomogeneous across its municipalities. Figure 7 (right) reports the number of antennas per municipality (*municipios*).

Both the social and the mobility networks discussed in the paper are generated with the same procedure. We keep only IDs which originated and received at least six calls during the observation period. In this way we can drop foreign phones and all special phone numbers that are likely to be not associated with an actual person (e.g. call centers). It is important to note that, after this cleaning phase, we do not follow any individual phone number. The networks are aggregated at the level of the municipality.

In the case of the mobility network, we connect two municipalities if a caller has been observed traveling from one municipality to the other. For each number, we generate the history of the calls it made, that is a time sequence of municipalities it connected to. We know the time and space difference between two consecutive calls. To add an edge between two municipalities, the phone number has to have made two consecutive calls in these municipalities.

In the case of the social network, we first have to associate a phone number to its base location. This is a solved problem: in [40] authors are able to associate a phone with its home and work location. In our case study, we feel that there is no reason to prefer either the home or the work location for a phone. We should associate a cellphone with the area in which it spends most of its activities. For each phone, we calculate the number of calls initiated during our entire observation period and we associate the phone's location with the tower to which it connects most frequently. Then, we know which other phones each phone called. We count each call as an edge between the two favorite municipalities of the phones involved.

In both cases, we have a directed weighted tower to tower network. We firstly aggregate this network at the municipality level, merging together all towers that are located in the same municipality, and aggregating their edges accordingly. Given the amount of data, the raw network contains vast amounts of noise, and we end up with an almost complete graph where every municipality connects to every other tower. To select the most significant edges, we firstly calculate the maximum spanning tree of the network (using the Kruskal algorithm), to ensure that all municipalities are part of the main network component. Then, we add edges by applying a technique designed to evaluate the statistical significance of an edge's observed weight [37]. For each edge, the technique returns a significance score which is equivalent to the p-value of the edge.

To detect network clusters, we rely on the large community discovery literature. We chose the Infomap algorithm [38] for two reasons: it performs well [41] and it returns disjoint clusters. Disjoint clusters are preferred because we do not want to deal with overlapping communities, where municipalities could belong to different clusters. We expect large cities to be across clusters, but we do not think that allowing them to span across communities is going to provide any useful insight for this work. On the other hand, it will make it harder to test the overlap between social and mobility clusters.

Finally, we need to define the f function, whose role is to rescale the weights of the social network as if they were affected by geographical distance as the weights of the mobility network. To do so, we need to know the ratio of weights in the mobility network M over the weights in the social network S at any given distance. We know the distance between each municipality, which is calculated as a straight line between their centers of mass, using the Haversine formula. For instance, if municipalities m_1 and m_2 are 70 kilometers apart, and their mobility and social weights are $M(m_1, m_2) = 120$ and $S(m_1, m_2) = 90$, then we associate a 70 km distance with the ratio $120/90 = 1.3$. When we perform this operation for all

observed links, we obtain the scatter plot depicted in Figure 8 (left). The color of the points encodes the number of observations that are associated with the same ratio at the same distance. The best fit found is a truncated power law of the form:

$$f(x) = \alpha x^\beta e^{-x/\gamma},$$

with $\alpha = 3567$, $\beta = -1.79$, and $\gamma = 97$. This is the f function we use to scale S down to S' .

Note that $S' = S \times f(x)$ and $f(x) = M/S$ do not necessarily entail $S' = M$, making our validation circular. If we generate random S and M weights, keeping the observed distance fixed, and derive their new $f(x)$ relationship, we do not obtain $S' = M$. If we calculate the correlation coefficient of the edge weights of the randomly generated M and S' (as derived from the random S), we obtain a distribution of correlation coefficients with average 0.25 and standard deviation of 0.13. Figure 8 (right) depicts the resulting distribution from 3,000 null models. The actual S' and M return a correlation score of 0.86 (represented by the black band in the figure).

Acknowledgements

The authors thank Frank Neffke, Clara Vandeweerdt and Renaud Lambotte for useful discussions and comments. We thank Marta C. Gonzales, Serdar Collack, Jameson Toole and Bradley Sturt for their help in gathering, cleaning and hosting the data. We thank the telecommunications operators for making available to us the data on which this paper is based, and opensignal.com for allowing us to use one of their maps. The Committee on the Use of Human Subjects at the authors' affiliation institution approved the usage of the data made in this paper, certifying that the rights of all subjects whose data have been examined in the study have not been violated.

References

1. Bergstrand JH (1985) The gravity equation in international trade: some microeconomic foundations and empirical evidence. *The review of economics and statistics* : 474–481.
2. Kleinert J, Toubal F (2010) Gravity for fdi. *Review of International Economics* 18: 1–13.
3. Jaffe AB, Trajtenberg M, Henderson R (1992) Geographic localization of knowledge spillovers as evidenced by patent citations. Technical report, National Bureau of Economic Research.
4. Branstetter LG (2001) Are knowledge spillovers international or intranational in scope?: Microeconometric evidence from the us and japan. *Journal of International Economics* 53: 53–79.
5. Bottazzi L, Peri G (2003) Innovation and spillovers in regions: Evidence from european patent data. *European Economic Review* 47: 687–710.
6. Keller W (2002) Trade and the transmission of technology. *Journal of Economic growth* 7: 5–24.
7. Keller W, Yeaple SR (2009) Multinational enterprises, international trade, and productivity growth: firm-level evidence from the united states. *The Review of Economics and Statistics* 91: 821–831.
8. Thiemann C, Theis F, Grady D, Brune R, Brockmann D (2010) The structure of borders in a small world. *PloS one* 5: e15422.
9. Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, et al. (2010) Redrawing the map of great britain from a network of human interactions. *PloS one* 5: e14248.

10. Lambiotte R, Blondel VD, de Kerchove C, Huens E, Prieur C, et al. (2008) Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications* 387: 5317–5325.
11. Rinzivillo S, Mainardi S, Pezzoni F, Coscia M, Pedreschi D, et al. (2012) Discovering the geographical borders of human mobility. *KI-Künstliche Intelligenz* 26: 253–260.
12. Scellato S, Mascolo C, Musolesi M, Latora V (2010) Distance matters: geo-social metrics for online social networks. In: Proceedings of the 3rd conference on Online social networks. pp. 8–8.
13. Takhteyev Y, Gruzd A, Wellman B (2012) Geography of twitter networks. *Social networks* 34: 73–81.
14. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual review of sociology* : 415–444.
15. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1082–1090.
16. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1100–1108.
17. De Domenico M, Lima A, Musolesi M (2013) Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* 9: 798–807.
18. Toole JL, Herrera-Yaque C, Schneider CM, González MC (2015) Coupling human mobility and social ties. *Journal of The Royal Society Interface* 12: 20141128.
19. Miritello G, Lara R, Moro E (2013) Time allocation in social networks: correlation between social structure and human communication dynamics. In: *Temporal Networks*, Springer. pp. 175–190.
20. Musolesi M, Mascolo C (2007) Designing mobility models based on social network theory. *ACM SIGMOBILE Mobile Computing and Communications Review* 11: 59–70.
21. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. *PloS one* 7: e37027.
22. Shmueli E, Mazeh I, Radaelli L, Pentland AS, Altshuler Y (2015) Ride sharing: A network perspective. In: *Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer. pp. 434–439.
23. Santi P, Resta G, Szell M, Sobolevsky S, Strogatz SH, et al. (2014) Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences* 111: 13290–13294.
24. Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, et al. (2011) Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one* 6: e16591.
25. Belik V, Geisel T, Brockmann D (2011) Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X* 1: 011001.
26. Halloran ME, Vespignani A, Bharti N, Feldstein LR, Alexander K, et al. (2014) Ebola: mobility data. *Science (New York, NY)* 346: 433.

27. Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* 10: 20130246.
28. Toole JL, Lin YR, Muehlegger E, Shoag D, Gonzalez MC, et al. (2015) Tracking employment shocks using mobile phone data. arXiv preprint arXiv:150506791 .
29. Llorente A, Cebrian M, Moro E, et al. (2014) Social media fingerprints of unemployment. arXiv preprint arXiv:14113140 .
30. Pennacchioli D, Coscia M, Rinzivillo S, Pedreschi D, Giannotti F (2013) Explaining the product range effect in purchase data. In: Big Data, 2013 IEEE International Conference on. IEEE, pp. 648–656.
31. Amini A, Kung K, Kang C, Sobolevsky S, Ratti C (2014) The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science* 3: 1–20.
32. Duranton G (2013) Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. Processed, Wharton School, University of Pennsylvania .
33. Berry BJ, Allen P, et al. (1961) Central place studies. a bibliography of theory and applications. *Central place studies A bibliography of theory and applications* .
34. Szell M, Lambiotte R, Thurner S (2010) Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* 107: 13636–13641.
35. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. ACM, pp. 591–600.
36. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th international conference on World wide web. ACM, pp. 721–730.
37. Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences* 106: 6483–6488.
38. Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one* 6: e18209.
39. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* : 65–70.
40. Toole JL, Colak S, Alhasoun F, Evsukoff A, Gonzalez MC (2014) The path most travelled: Mining road usage patterns from massive call data. arXiv preprint arXiv:14030636 .
41. Coscia M, Giannotti F, Pedreschi D (2011) A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4: 512–546.

Figure Legends

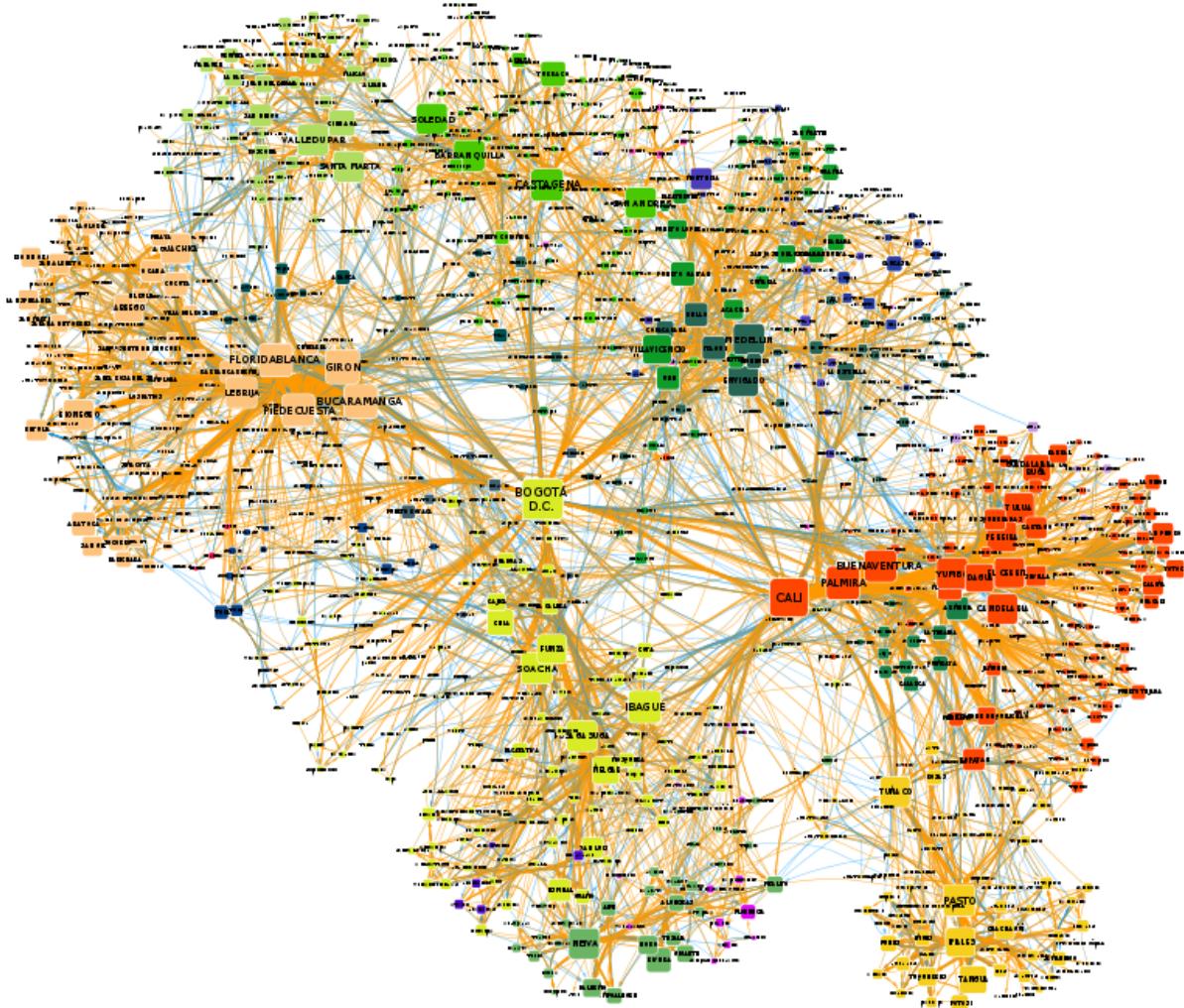


Figure 1. The social network of Colombia. The graph representing the social relationships in Colombia across municipalities. Each node is a municipality and directed links connect two municipalities if people from one municipality have a significant amount of social relations with people living in the other municipality. Node size is proportional to indegree, and node color indicates the node community, as detected by Infomap. Link size and transparency is proportional to its significance, as is its color: orange links are very significant, blue links are less strong.

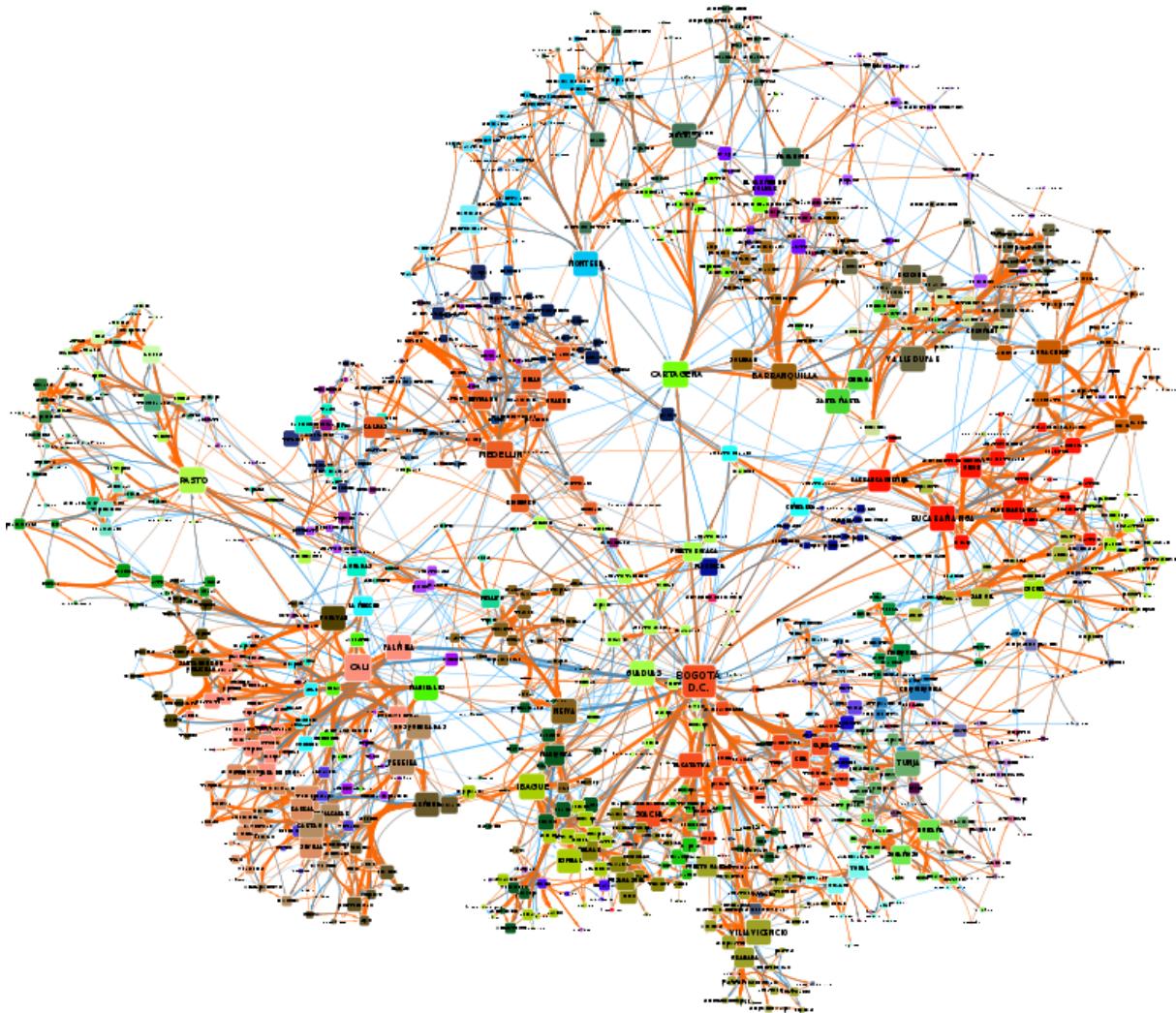


Figure 2. The mobility network of Colombia. The graph representing the mobility relationships in Colombia across municipalities. Each node is a municipality and directed links connect two municipalities if we observe a significant amount of trips flowing from one municipality to another. Graphical elements are defined similarly to the social network presented above, and we refer to Figure 1's caption for their discussion.

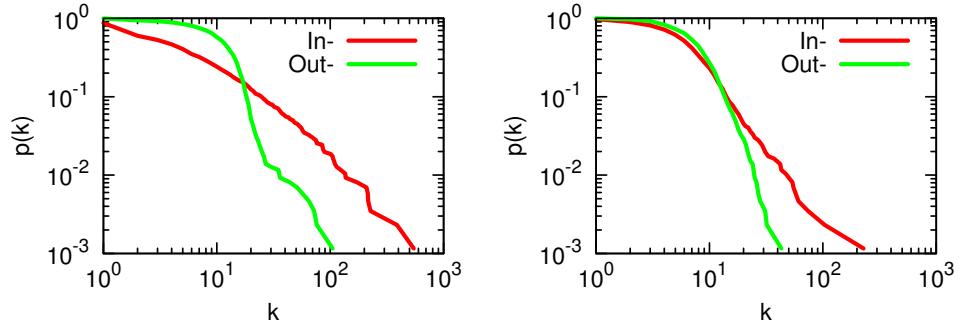


Figure 3. The social (left) and mobility (right) degree distributions. Cumulative indegree (red) and outdegree (green) distributions for the social and mobility networks. The plots report the probability that a node in each network has a given degree, or higher. For instance, in the social network there is roughly a 1% probability ($p(k) = 10^{-2}$) that a node has an indegree of 100 ($k = 10^2$) or more.

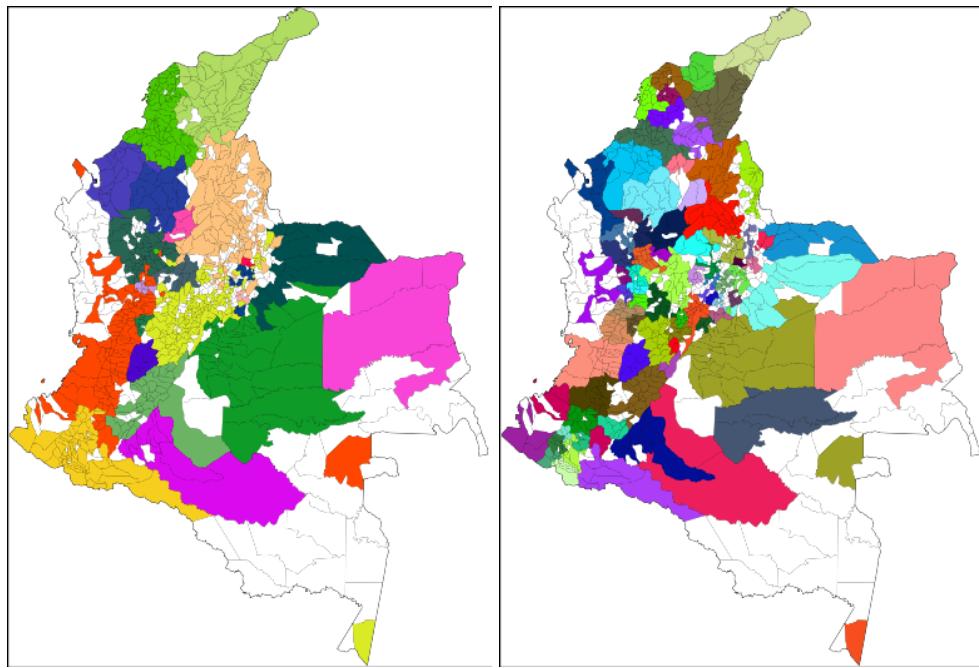


Figure 4. The social (left) and mobility (right) clusters on Colombia's territory. A geographical visualization of the network clusters computed on the social and mobility networks. Each municipality area is colored with its corresponding cluster. The color palette is the same used for Figures 1 and 2, so a node's color in those figures corresponds to its municipality color in these figures.



Figure 5. Colombia cellphone signal strength. The heatmap represents the signal strength of the cellphone network across the territory of Colombia. Red areas have a strong signal, blue areas a weak signal, and uncolored map areas have no signal. Image courtesy of opensignal.com, which granted us permission to use it under a CC BY 3.0 license.

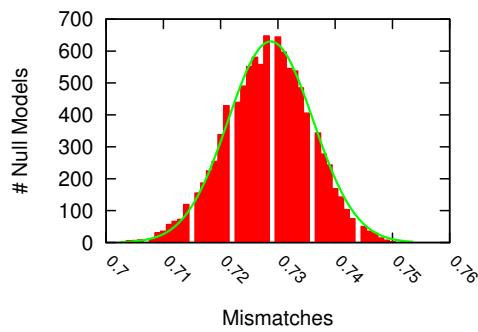


Figure 6. Null model mismatch distribution. We calculate null social and mobility community and we test their degree of overlap. The figure reports the number of null models (Y axis) scoring each overlap value interval (X axis). The visible white bands are caused by gaps in the possible overlap values that we can obtain, since the denominator is lower than 1,000 (it is equal to the number of nodes, $n = 863$). Green line fits the most likely probability distribution, which is a Gaussian with $\mu = 0.7287$ and $\sigma = 0.0074$.

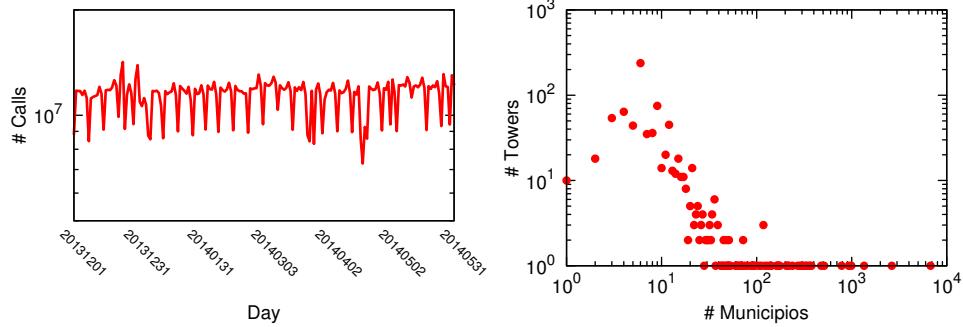


Figure 7. Statistics on Colombia cellphone usage. (Left) Number of calls (Y axis) per day (X axis) in our dataset. We cover a time window starting from December 1st, 2013 and ending in May 31st, 2014. The M-shaped pattern is typical of weekly human activities. Deviations are usually represented by national holidays and special occasions, such as New Year's Day. (Right) Distribution of number of cellphone towers per *municipio* in Colombia. *Municipios* with no towers are removed from the plot.

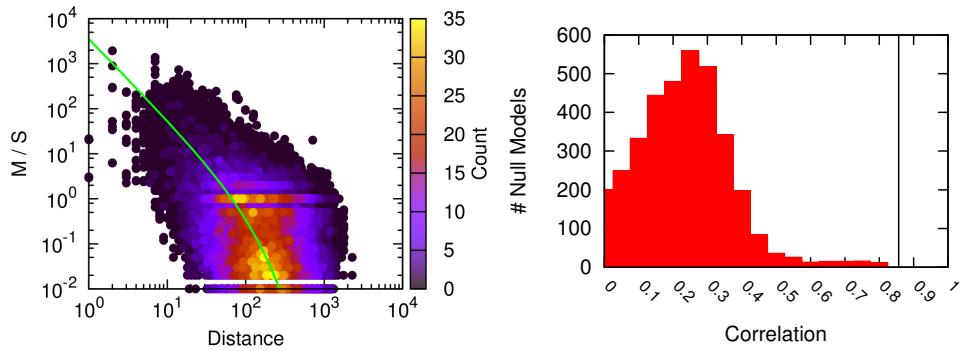


Figure 8. Mobility-Social weight ratio against distance. (Left) In the y axis we report the ratio of the weight of links connecting two municipalities at a given distance (x axis) in the mobility network over the weight of the same link in the social network. The green line represents our model function f . (Right) Distribution of edge weights correlation coefficients for our distance rescaling null models.

Tables

Measure	Social	Mobility
# Nodes	863	863
# Edges	9,639	6,614
Avg Degree	22.34	15.33
Avg Path Length	2.88	4.46
Degree Assort.	-0.142	0.004
Reciprocity	19.87%	37.30%
Codelength	5.156	4.151

Table 1. Network statistics of the social and mobility networks. We calculate the following topological features of the social and mobility networks: number of nodes (“# Nodes” or n), number of edges (“# Edges” or e), average degree (“Avg Degree”, $\frac{2e}{n}$), average path length (“Avg Path Length”, the number of edges needed to be crossed to go from a random node of the network to another), degree assortativity (“Degree Assort.”, correlation coefficient of the degrees of nodes connected by an edge), reciprocity (fraction of directed edges going in both directions), and codelength (number of bits required to encode the network given the communities calculated by Infomap, the lower the more well-separated are the communities).

Supporting Information

S1. Image sharing authorisation. The document with which OpenSignal granted us the rights of using Figure 5.

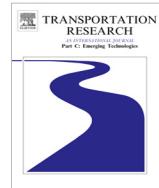
S2. IRB Documentation. The document with which the Committee on the Use of Human Subjects at the authors' affiliation institution approved the usage of the data made in this paper, certifying that the rights of all subjects whose data have been examined in the study have not been violated.

S3. Dataset Used. The minimal dataset necessary for the replication of the main results included in the paper. The zip file contains data and code used in the experiments. A readme file provides instructions.



Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

The path most traveled: Travel demand estimation using big data resources

Jameson L. Toole^{a,1}, Serdar Colak^{b,*1}, Bradley Sturt^a, Lauren P. Alexander^b, Alexandre Evsukoff^c, Marta C. González^{a,b}

^aEngineering Systems Division, MIT, Cambridge, MA 02139, United States

^bDepartment of Civil and Environmental Engineering, MIT, Cambridge, MA 02139, United States

^cCOPPE/Federal University of Rio de Janeiro, Brazil

ARTICLE INFO

Article history:

Received 1 June 2014

Received in revised form 19 April 2015

Accepted 21 April 2015

Available online xxxx

Keywords:

Mobility

Location based services

Congestion

Road networks

Mobile phone data

ABSTRACT

Rapid urbanization is placing increasing stress on already burdened transportation infrastructure. Ubiquitous mobile computing and the massive data it generates presents new opportunities to measure the demand for this infrastructure, diagnose problems, and plan for the future. However, before these benefits can be realized, methods and models must be updated to integrate these new data sources into existing urban and transportation planning frameworks for estimating travel demand and infrastructure usage. While recent work has made great progress extracting valid and useful measurements from new data resources, few present end-to-end solutions that transform and integrate raw, massive data into estimates of travel demand and infrastructure performance. Here we present a flexible, modular, and computationally efficient software system to fill this gap. Our system estimates multiple aspects of travel demand using call detail records (CDRs) from mobile phones in conjunction with open- and crowdsourced geospatial data, census records, and surveys. We bring together numerous existing and new algorithms to generate representative origin-destination matrices, route trips through road networks constructed using open and crowd-sourced data repositories, and perform analytics on the system's output. We also present an online, interactive visualization platform to communicate these results to researchers, policy makers, and the public. We demonstrate the flexibility of this system by performing analyses on multiple cities around the globe. We hope this work will serve as unified and comprehensive guide to integrating new big data resources into customary transportation demand modeling.

© 2015 Published by Elsevier Ltd.

1. Introduction

The accelerating growth of cities has made the estimation of travel demand and the performance of transportation infrastructure a critical task for transportation and urban planners. To meet these challenges in the past, methods such as the widely used four-step model and more recent activity based models were developed to make use of available data computational resources. These models combine meticulous methods of statistical sampling in local (Daganzo, 1980; Smith, 1979) and national household travel surveys (Stopher and Greaves, 2007; Richardson et al., 1995) to process and infer trip

* Corresponding author.

E-mail address: serdarc@mit.edu (S. Colak).

¹ These authors contributed equally to this work.

information between areas of a city. The estimates they produce are critically important for understanding the use of transportation infrastructure and planning for its future (Van Zuylen and Willumsen, 1980; Spiess, 1987; Maher, 1983; Lo et al., 1996; Hazelton, 2003, 2001, 2000; Lu et al., 2013; Cascetta, 1984; Bell, 1991).

While the surveys that provide the empirical foundation for these models offer a combination of highly detailed travel logs for carefully selected representative population samples, they are expensive to administer and participate in. As a result, the time between surveys range from 5 to 10 years in even the most developed cities. The rise of ubiquitous mobile computing has lead to a dramatic increase in new, *big data* resources that capture the movement of vehicles and people in near real time and promise solutions to some of these deficiencies. With these new opportunities, however, come new challenges of estimation, integration, and validation with existing models. While these data are available nearly instantaneously and provide large, long running, samples at low cost, they often lack important contextual demographic information due to privacy reasons, lack resolution to infer choices of mode, and have their own noise and biases that must be accounted for. Despite these issues, their use for urban and transportation planning has the potential to radically decrease the time in-between updated surveys, increase survey coverage, and reduce data acquisition costs. In order to realize these benefits, a number of challenges must be overcome to integrate new data sources into traditional modeling and estimation tools.

Analyzed on its own, data generated by the pervasive use of cellular phones has offered insights into abstract characteristics of human mobility patterns. Recent work has found that individuals are predictable, unique, and slow to explore new places (González et al., 2008; Brockmann et al., 2006; de Montjoye et al., 2013; Song et al., 2010a,b; Candia et al., 2008; Calabrese et al., 2013). The availability of similar data nearly anywhere in the world has facilitated comparative studies that show many of these properties hold across the globe despite differences in culture, socioeconomic variables, and geography. The benefits of this data have been realized in various contexts such as daily mobility motifs (Schneider et al., 2013; Sevtsuk and Ratti, 2010), disease spreading (Belik et al., 2011; Wesolowski et al., 2012) and population movement (Lu et al., 2012). While these works have laid an important foundation, there still is a need to integrate these data into transportation planning frameworks.

To make these new data useful for urban planning, we must clarify their biases and build on the progress made by transportation demand modeling even in the face of limited data resources. We must combine this domain knowledge with new algorithms and metrics to better understand travel behaviors and the performance of city infrastructure and we must update technologies to accommodate the computational requirements of processing massive geospatial data sets. Individual survey tracking and stay extraction (Asakura and Hato, 2004), OD-estimation and validation (Caceres et al., 2007; Nie et al., 2005; Wang et al., 2012; Iqbal et al., 2014), traffic speed estimation (Bar-Gera, 2007; Zhan et al., 2013), and activity modeling (Phithakkitnukoon et al., 2010; Reades et al., 2009) have all been explored using new massive, passively collected data. However, these studies generally present alternatives for only a few steps in traditional four-step or activity based models for estimating travel demand or fail to compare outputs to travel demand estimates from other sources. Moreover, many methods offered to date lack portability from one city to many with minimal additional data collection or calibration required.

Here we fill this gap with a modular, efficient computational system that performs many aspects of travel demand estimation billions of geo-tagged data points as an input. We review and integrate new and existing algorithms to produce validated origin-destination matrices and road usage patterns. We begin by outlining the system architecture in Section 2.1. In Section 2.3 we explain our methods of extracting, cleaning, and storing road network information from a variety of sources. We discuss recent advances in OD creation from mobile phone data in Section 3.1 and implement a simple, parallel incremental traffic assignment algorithm for these trips in Section 3.2. We present comparisons of these results to estimates from traditional survey methods in Section 4.1. Finally, in Sections 4.2, 4.3, 4.4 we present a variety of measurements that can be made with the proposed system as well as an online, interactive visualization for conveying these results to researchers, policy makers, and the public. To demonstrate the flexibility of the system, we perform these analyses for five metro regions spanning countries and cultures: Boston and San Francisco, USA, Lisbon and Porto, Portugal, and Rio de Janeiro, Brazil.

1.1. Description of data

Travel surveys are typically administered by state or regional planning organizations and are integrated with public data such as census tracts and the demographic characteristics of their residents, made available by city, state, and federal agencies. New data sources, however, come from new providers. Large telecommunications companies, private applications, and network providers collect and store enormous quantities of data on users of their products and services, presenting computational challenges for storing and analyzing them. Billions of phone calls must be processed, data from open- and crowd-sourced repositories must be parsed, and results must be made more accessible to individuals that generated them. At the same time, it is critical that measurements from these new sources are statistically representative and corrected for biases inherent in new data. This process requires integration of new pervasive data with reliable (though less extensive) traditional data sources such as the census or travel surveys. We combine the following data sets to illustrate the capabilities of the system architecture here proposed:

1. *Call Detail Records (CDRs)*: At least three weeks of call detail records from mobile phone use across each subject city. The data includes the timestamp and the location for every phone call (and in some cases SMS) made by all users of a particular carrier. The spatial granularity of the data varies between cell tower level where calls are mapped to towers

and triangulated geographical coordinate pairs where each call has a unique pair of coordinates accurate to within a few hundred meters. Market shares associated with the carriers that provide the data also vary. Personal information is anonymized through the use of hashed identification strings. For reference, 6 weeks of CDR data from the Boston area containing roughly 1 billion calls made by 1.6 million unique users consumes roughly 70 gigabytes of disk space in its raw format. In cities with longer observation periods, data size quickly becomes a performance issue.

2. *Census data*: At the census tract (or equivalent) scale, we obtain the population and vehicle usage rate of residents in that area. For US cities, the American Community Survey provides this data on the level of census tracts (each containing roughly 5000 people). Census data is obtained for Brazil through IBGE (Instituto Brasileiro de Geografia e Estatística) and for Portugal through the Instituto de Nacional de Estatística. All cities analyzed in this work have varying spatial resolutions of the census information.
3. *Road networks*: For many cities in the US, detailed road networks are made available by local or state transportation authorities. These GIS shapefiles generally contain road characteristics such as speed limits, road capacities, number of lanes, and classifications. Often, however, these properties are incomplete or missing entirely. Moreover, as such road inventories are expensive to compile and maintain, they simply do not exist for many cities in the world. In this case, we turn to OpenStreetMaps (OSM), an open source community dedicated to mapping the world through community contributions. For cities where a detailed road network cannot be obtained, we parse OSM files and infer required road characteristics to build realistic and routable networks. At this time, the entirety of the OSM database contains roughly 4 terabytes of geographic features related to roads, buildings, points of interest, and more.
4. *Survey and model comparisons*: Wherever possible, we obtain the most recent travel demand model or survey from a particular city and compare the results to those output by our methods. In Boston, we use the 2011 Massachusetts Household Travel Survey (MHTS) and upscale trips according to standard procedures, in San Francisco, the 2000 Bay Area Transportation Survey (BATS), in Rio de Janeiro, a recent transportation model output provided by the local government, and in Lisbon, the most recent estimates from the MIT-Portugal UrbanSim LUT model that uses the 1994 Lisbon transportation survey as input (Ferreira et al., 2010). We found no recent travel survey or model for Porto.

Table 1 compiles descriptive statistics for these data sources for each city we explore in the latter sections of this paper.

2. System architecture and implementation

2.1. Architecture

The system architecture to integrate the data sources above must be flexible enough to handle different regions of the globe which may have different data availability and quality and efficient enough to analyze massive amounts of data in a reasonable amount of time. The proposed system must also be modular, so that components can be updated easily as new technologies and algorithms become available. To meet these requirements, we choose an object-oriented approach with loose schema requirements. A final object is to make results accessible to a range of end users via online, interactive visualization. To satisfy these constraints, we propose the system architecture depicted in Fig. 1.

2.2. Parsing, standardizing, and filtering user data

One of the biggest challenges in parsing and analyzing travel survey data is the incredible variety in data schema, collection, and reporting practices. Each planning organization typically constructs its own set of data codes and definitions and provides data in unique formats. This makes it very difficult to compare surveys done in different cities. Call detail records, on the other hand, are typically available for many cities from the same provider and in the same format, and in most cases, translating between the formats of different carriers is simply a matter of shuffling columns. The first component of our system is a simple architecture to convert all CDR data to a standard format that can be expected by the rest of the components.

Given the size of these data sets and the rapidly evolving schema requirements of new models, choosing the proper data structure is critical. Google's open source Protocol Buffer library² is an ideal choice as they provide fast serialization for speed and space efficient file storage as well as flexible schemas that can be changed without compromising backwards compatibility. These structures were designed to serve some of the largest databases in the world and are more than enough for our task.

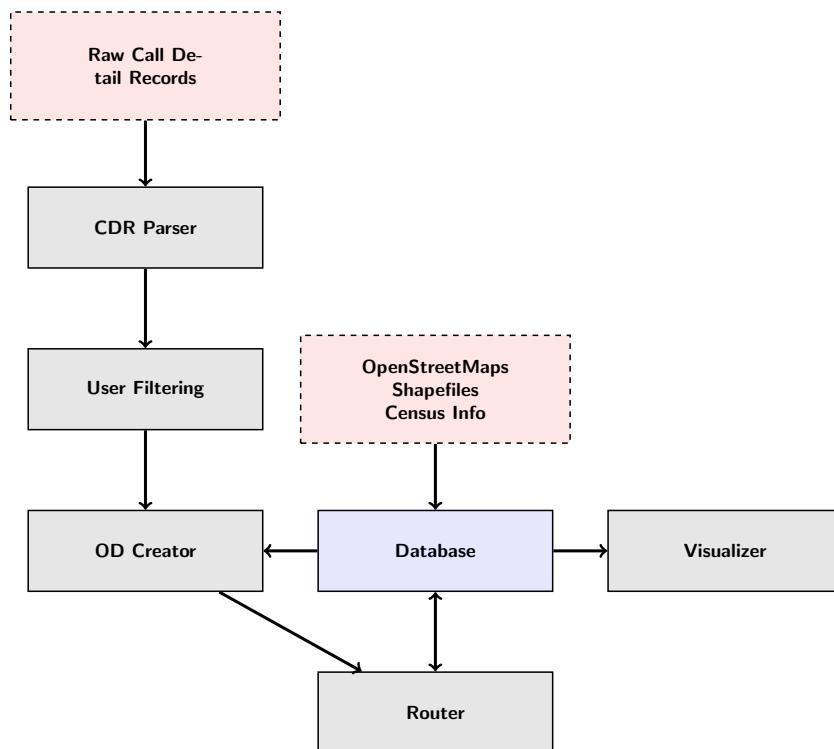
We take a user centric approach to CDR data. We define a *user_data* protocol buffer message that will form the core data structure for our custom User class in an object-oriented programming model. Each User object can be assigned a number of attributes such as the number of calls they make, their home and work locations, and mobility characteristics such as the average time between calls or the average distance traveled on each trip. More sophisticated methods can compute the number and distribution of their trips and even expand them based on census information. We define similar structures and classes for OD matrices, trips, and census data. The serialization routines built into the protocol buffer library ensures that storage of raw data is efficient. To analyze a new city, the user only needs to write two simple routines, one to parse a single

² Google Protocol Buffers <https://developers.google.com/protocol-buffers/>.

Table 1

A comparison of the extent of the data involved in the analysis of the subject cities.

City	Boston	SF Bay	Rio	Lisbon	Porto
Population (mil.)	4.5	7.15	12.6	2.8	1.7
Area (1000 km ²)	4.6	18.1	4.5	2.9	2.0
# of Users (mil.)	1.65	0.43	2.19	0.56	0.47
# of Calls (mil.)	905	429	1045	50	33
# of cell towers	N/A	892	1421	743	335
# of Edges (ths.)	21.8	24.3	22.7	28.1	15.1
# of Nodes (ths.)	9.6	11.3	22.1	16.1	8.6
# of Tracts	732	1139	729	295	272

**Fig. 1.** A flowchart of the system architecture.

line of the CDR file and populate relevant user attributes and one to populate census data objects. Standardizing the CDR data format in this way makes it very easy to compare the output of our estimation models across different cities.

2.3. Creating and storing geographic data

A relational database is used to store road network and census information for every city in a standard format. Given the current cost of computing resources, these systems provide adequate performance for storing static GIS and census data and have convenient, mature interfaces for easy access. We also use this database to store aggregated results from our estimates so that they can be made available to interactive web APIs and visualization platforms. We use a Postgres and the open source spatial extension PostGIS to store and manipulate census and road network data.

While census tract or TAZ (Traffic Analysis Zone) polygons and demographic information are stored in this database, it is computationally inefficient to perform point-in-polygon calculations for each user or call record in our CDR dataset. To dramatically speed these computations, we rasterize polygons into a small pixel grid, where pixel values is a unique identifier for the census tract covering that pixel. This raster is then used as a look-up table to convert the latitude and longitude of calls into census tract IDs. The rasterization introduces some error along the borders of tracts, but these errors are minimized by making pixel sizes much smaller than the size of the raster and resolution of the location estimates of calls (between 10 m and 100 m).

While the platform supports road networks supplied by local municipalities in the form of shapefiles, we have implemented a parser to construct routable road networks from OpenStreetMap (OSM) data due to its global availability. Transportation networks in OSM are defined by *node* and *way* elements. Nodes represent points in space that can refer to anything from a shop to a road intersection, while ways contain a list of references to nodes that are chained together to form a line. In our context, relevant ways are those used by cars and relevant nodes are intersections within the road network. Ways and nodes may also contain a number of tags to denote attributes such as “number of lanes” or “speed limit”. Many roads, however, do not include the whole set of attributes necessary for accurate routing. For example, city roads often lack speed limit information required to estimate the time cost, which in turn is used to find shortest paths based on total travel time. To infer this missing data, our system supports the creation of user-defined mappings between highway types and road properties. For example, ways tagged as “motorways” are generally major highways and have a speed-limit of 55 mph in the Boston area. They tend to have 3 lanes in each direction. “Residential” roads, on the other hand, have a speed-limit of 25 mph and 1 lane in each direction. Each road segment is also given a capacity based on formulas suggested by the US Federal Highway Administration. Using these mappings, we parse the OSM xml data to create a routable, directed road graph with all properties required to estimate realistic costs driving down any given road.

We implement two additional cleaning steps to improve efficiency. The first filters out irrelevant residential roads. These small local roads are filtered from our network, as they are not central to the congestion problem, yet tend to increase computation time significantly. Finally, in OSM data, a node object can refer to many things, for example an actual intersection or simply a vertex on a curve used to draw a turn. The latter case results in a network node with only one incoming and one outgoing edge (assuming U-turns are not allowed). These nodes are superficial and increase network size and routing algorithm run times needlessly. We simplify networks by removing these nodes from the network and only connecting true intersections, keeping the geographic coordinates of the nodes so that link costs still reflect actual geographic length of roads rather than straight line distances between start and end points. The parsed and cleaned edges are then loaded into the Postgres database, preserving attributes and geometry. Pseudo-code of the algorithm to parse and simplify OSM networks can be found in [Algorithm 1 in the supplementary materials](#).

3. Estimating origin–destination matrices

The following sections review algorithms for transforming billions of geo-tagged data points into validated origin destination matrices and assigning these flows to transportation infrastructure. Some of these algorithms are important for their deviation from traditional approaches and some are important for their computational efficiency, a requirement when faced with such massive data sets.

3.1. Measuring flow

Current methods to estimate the flow of people or vehicles from place to place in a city generally fall into two categories: four-step or activity based approaches. The former class of models breaks the process into a sequence of four steps from which it earns its name. The first three steps in a four-step model – trip generation, distribution, and mode choice – are designed to estimate origin–destination matrices containing the number of trips from place to place within a city. Traditional modeling approaches use data from travel surveys possibly combined with land use and point of interest information to generate estimates of trip production and attraction for locations. These trips are then distributed from their point of origin to destinations across the city using gravity or radiation models. Modes of transit are assigned using models estimated from survey data and information on the transit infrastructure. More recent activity based models approach travel demand from an individual level. Assuming that travel demand is created by the need to fulfill activities, these models use similar survey data to estimate utility curves for travels and predict behaviors using probit or logit models based on these preferences.

While new data sources such as CDRs do not provide the same detailed demographic and contextual information about individuals or trips, they do provide an opportunity to measure travel more directly. With billions of data points, high spatio, temporal resolution, and long observation periods, passive data collected by mobile devices provide unparalleled scale of observation. New methods to estimate travel demand must balance trade offs between small, but complete data for a short period of time and large, but incomplete data over a longer period of time. In both cases noise and biases must be carefully dealt with to produce valid measurements. In this section we adapt and integrate previous works that have tackled parts of this problem into a full implementation of travel demand estimation for cities.

Mobile phones offer good, but imperfect measurements of geographic position. The coordinates of a mobile phone event are either recorded as the location of a nearby tower through which the event was routed or as a triangulation based signal strength from multiple towers. This creates uncertainties of a few hundred meters in estimates of a user's location. Moreover, observations are only recorded when an individual uses his or her device, resulting in heterogeneous sampling frequencies between users and at different times for a given user. While sampling rates and data density are increasing rapidly with rising penetration rates and usage, these issues present statistical challenges.

Initial methods by Wang et al. construct *transient origin–destination matrices* by simply counting a trip for pair of consecutive calls made within the same hour from two different towers. However, this method lead to an abundance of short trips

and provided a very biased view of movement. Instead, mobile phone trajectories must be de-noised to remove spurious points or calls made in the middle of routes rather than origins or destinations. To extract meaningful locations, termed as *stays*, algorithms have been developed to smooth out this noise and control for these biases. Jiang et al. provide a thorough review of these techniques in [Jiang et al. \(2013\)](#) and we adapt the stay point algorithm originally described by [Zheng and Xie \(2011\)](#).

Given a user's trajectory of spatiotemporal points $P = \{p_1(x_1, y_1, t_1), \dots, p_n(x_n, y_n, t_n)\}$, the goal is to discover meaningful locations at which a user repeatedly stays for a significant amount of time. The algorithm begins by considering each call in a time ordered sequence. Two consecutive (p_i, p_{i+1}) points are considered to form the start of a *candidate set* of points at the same semantic location if the distance between them is less than a threshold $\Delta r_{i,i+1} < \delta$. Subsequent points are added to this candidate set if they also meet this criteria, e.g. p_{i+2} is added if $\Delta r_{i+1,i+2} < \delta$. The result is a candidate set $S = \{p_s(x_s, y_s, t_s), \dots, p_t(x_t, y_t, t_t)\}$ containing a number of consecutive calls. A candidate set is considered to represent a single *candidate stay* if time between the first and the last observation in the subsequence S are separated by a time greater than a threshold $\Delta t_{m,n} > \tau$. The geographic location of a candidate stay is set to be at the centroid of points in S . Due to noise in locations and daily call frequencies, multiple candidate stays that are actually the same place may be estimated at a slightly different geographic coordinate on different observation days. To account for this, a final agglomerative clustering algorithm is used to consolidate candidate stays to a single semantic location regardless of the temporal sequence of individual calls. Though many agglomerative clustering algorithms exist, we implement a simple, efficient grid based approach by assigning each filtered location to a grid cell and then defining a final *stay point* as the centroid of all filtered locations in each cell. A final pass through the original calls assigns any call within a distance δ from a stay point to that stay point regardless whether or not a consecutive call was recorded from that location. This algorithm removes noisy or spurious outliers from the data set while preserving as much information on visits as possible. It may also be run on both triangulated and tower-based CDR data. In the latter case it removes noise associated with calls from the same location being routed through different nearby towers due to environmental factors. Pseudo-code can be found in [Algorithms 2–5 in the SI](#).

With de-noised trajectories of stay points, the next step is to infer contextual information about each location. [Alexander et al. \(2015\)](#) and [Çolak et al. \(2015\)](#) improve on methods by [Wang et al. \(2012\)](#) and [Iqbal et al. \(2014\)](#) by using visit frequencies and temporal data to infer contextual information such as a location's function or trip purpose. A user's *home* location is defined as the stay point they are observed at most frequently between the hours of 8 pm and 7 am on weeknights. Their *work* location is defined as the stay point other than home that a user visits the most between the hours of 7 am and 8 pm on weekdays. Because many individuals do not work, we leave the work location blank if the candidate location is not visited more than once per week or if the location is less than 500 m from their home location. All remaining non-home or work stay points are designated as *other*.

Daily trips are estimated from filtered users by analyzing consecutive observations at different stay points during a given time window. They begin by defining an *effective day* as a period between 3 am one morning and 3 am on the next consecutive morning. This definition is used to minimize the number of trips that are prematurely ended due to the assumption that users start and end each day at home. A home-based work (HBW) trip is counted if a user is observed to travel between home and work, a non-home based (NHB) trip is counted if a user moves between two non-home stay points, and a home-based other (HBO) trip is counted if a user is observed moving between their home location and a location labeled as *other*.

Though a user must have traveled between two different observed stay points at some in time, we do not know the precise departure time. We assign a random departure time based on the conditional probability that user departed during an hour between the time they were last observed at the origin and the time they were first observed at the destination. This conditional probability function for departure time can be derived from surveys such as the National Household Travel Survey or estimated empirically using observed call frequencies of all users over the course of the day. Alexander et al. show that this method produces CDR trip departure time distributions in line with multiple surveys for the Boston region. Having assigned departure times and purposes to each trip, we can construct trips made by a given user. Generally, we are interested in trips between geographic areas such as towns or census tracts so here we convert origin and destination points to IDs of the tract or zone they are in. The result is a vector of trips between locations in the city for each user in our data set.

While a trip represents an observation of movement of at least one person between two locations, we expand these trip counts to represent all individuals in a city. Expansion is a critical step in models relying on survey data where the sample sizes are typically less than 1% of the population. Here we generally have hundreds of thousands of users in our sample, but must still be careful to control for differences in market share and usage rates across a city. We first scale trips based on how often an individual uses their phone. For each user, we calculate the average number of trips made during a given time window by dividing the number of trips counted by the number of days that user was observed making a call. This step effectively measures the average number of trips a user makes between two locations on a day given that they are observed in our data set.

Due to differences in daily usage of mobile phones among the population, not every user makes enough calls on a typical day to infer their movement patterns. For this reason, we must filter out users that do make enough calls. This step requires trade-offs between sample size and amount of data we have on each selected user. Because we will eventually be routing these trips through the transportation network, it is important to correctly estimate the total number of trips taken as well as the distribution of trips across the city. In practice, we find that filtering out users who we measure to make fewer than 2.5

trips per day leaves a large sample size of active users and results in valid estimates of trip tables and OD matrices as shown in subsequent sections. Those implementing these methods may find that different filtering criteria produce samples suited for different tasks.

We then expand the average trip counts of filtered users to account for market penetration rates. As with survey participants, the ratio of cell phone users to the population is not uniform within the region. Each user is assigned a home census tract and expansion factors are computed for each tract by measuring the ratio of the number of users assigned there and the reported population. In cities such as Boston, these expansion factors tend to be less than 10, but can be higher in places with lower market share. They are generally much lower than surveys which may only choose two or three individuals to represent hundreds or thousands in an area. Each user's typical daily trip volumes are then multiplied by the expansion factor corresponding to their home tract and the now represent the movements of some fraction of the tracts population.

Finally, we may wish to consider only trips via a certain mode, e.g. vehicle trips. Though CDR data does not provide resolution required to measure mode choice, vehicle trips can be approximated by weighting person trips by vehicle usage rates in the home census tract of users. In this way, full OD matrices for vehicle or person trips are computed by summing the expanded trip volume computed for all users between all pairs of census tracts. We also construct partial OD matrices containing only trips of a certain purpose during a certain time window. Due to the relative consistency of CDR data around the world, we can adopt this same OD creation procedure in all cities. Pseudo-code to generate OD matrices has been adapted from [Alexander et al. \(2015\)](#) and [Çolak et al. \(2015\)](#) and can be found in [Algorithms 6 and 7 in the SI](#). The results from this method are compared to the output of traditional models where applicable. Trip tables and correlations plots can be found below in Section 4.1.

3.2. Trip assignment

Having estimated OD flows, our next task is to efficiently assign these trips to transportation infrastructure, in this case a road network ([Bast et al., 2007](#)). The first step takes tract to tract OD matrices and distributes trips among nodes, or intersections. A trip originating in a census tract is assigned uniformly at random to an intersection in that tract and to an intersection within its destination tract. This distributes flows such as not to create artificial congestion points and reflects general uncertainty in the exact origin of trips. Other approaches, however, may consist of using abstract centroid nodes unique to each tract and connect to a number of other intersections within that tract using what's referred to as centroid connectors. With intersection to intersection flows, the next task is to assign traffic to routes.

Traffic assignment is another mature domain that has been studied extensively by urban and transportation planners. Static non-equilibrium models approaches consist of treating all users as homogenous agents who make route choices prior to departure based on some heuristic related to current traffic conditions (e.g. the path that minimizes travel time). Incremental Traffic Assignment (ITA) is a variant of these static non-equilibrium assignment models that assigns batches of trips serially and updates costs between increments, as an improvement over the simplest all-or-nothing assignment methods. However, it is known that dynamic equilibrium models are more realistic in assigning trips as outcomes are closer to the Wardrop principles ([Wardrop, 1952](#)), or Nash Equilibria, where drivers seek paths that minimize their travel time and in the final traffic conditions, no driver has an incentive to change their route. To take a step further from static models, Dynamic Traffic Assignment (DTA) ([Merchant and Nemhauser, 1978](#)) models take an iterative and temporally more coherent approach. The addition of these complexities help model traffic flow at finer granularity, enabling road segments to have different conditions within themselves and consequently the representation of phenomena like congestion spill-back, FIFO principle, and others ([Çolak et al., 2013](#)).

Our system is modular so that it may implement any number of traffic assignment algorithms. Here, however, we take a simple ITA approach, as it is computationally efficient for many trip pairs in detailed road networks and allows us to keep track of each vehicle as it is routed through the network. We develop a set of tools to perform large scale routing and traffic assignment using parallelization for speedups. First, the parsed and optimized road network is loaded into a graph object. In our implementation, we use the Boost Graph Library for its flexibility and efficiency. We can then compute shortest paths based on a user defined cost (in this case travel time on road segments). We choose the A* algorithm among the wide range of shortest path algorithms, as it's widely used in routing on geographic networks for its flexibility and efficiency. The A* algorithm implements a *best-first-search* using a specified heuristic function to explore more promising paths first. The euclidian distance between nodes provides an intuitive heuristic that ensures optimal solutions are found. While this algorithm provides the same results as Dijkstra's algorithm, we find that it becomes more efficient to compute paths one by one for sparse OD matrices.

On most city roads, free-flow speeds are rarely achieved due to congestion. As a result, traffic patterns may significantly change the time costs associated with using a particular route. To address this, we implement an Incremental Traffic Assignment (ITA) algorithm ([Ortúzar and Willumsen, 1994](#)). A simplified schematic explaining the procedure can be seen in [Fig. 2](#). This algorithm assigns trips in a series of increments and updates the costs of edges in the network based on the number of vehicles that were previously assigned to that road between increments. For example, the first increment assigns 40% of trips for each pair assuming each driver experiences free-flow speeds. The travel time cost associated with every road segment is then adjusted based on how many drivers were assigned to that road and the total number of cars a road can accommodate in unit time. The next 30% of drivers are then routed in the updated conditions. This process is repeated until all users have been assigned a route. The shortcoming of this method is that once a driver has been assigned

Full OD		Increment 1 width=0.7		Increment 2 width=0.3	
(o,d)	flow	(o,d)	flow	(o,d)	flow
(1,2)	1000	(1,2)	700	(1,2)	300
(1,3)	100	(1,3)	70	(1,3)	30
(2,3)	250	(2,3)	175	(2,3)	75
(3,2)	100				
(4,3)	1000				
(5,4)	500				

Update Cost	Batch 1	Batch 2	Batch 1	Batch 2	Update Cost
(o,d)	flow	(o,d)	flow	(o,d)	flow
(1,2)	700				
(1,3)	70				
(2,3)	175				
(3,2)	70				
(4,3)	700				
(5,4)	350				

Fig. 2. Our efficient implementation of the incremental traffic assignment (ITA) model. A sample OD matrix is divided into two increments and then split into two independent batches each.

a route it does not change, and consequently the approach does not converge to Wardrop's equilibrium even for very small increment sizes. Yet we use it here due ease of implementation and the fact that it is still insightful for the purposes of demonstrating the implementation of a modular data-driven travel demand model. Future work will explore the use of newer methods.

Relating travel performance to traffic conditions has been a long standing problem in transportation. Many different characterizations exist, ranging from conical volume-delay functions to more complex approaches (Branston, 1976; Spiess, 1990; Akcelik, 1991). One of the most simplistic and common metrics used in determining the travel time associated with a specific flow level is the ratio between the number of cars actually using a road (volume) and its maximum flow capacity (volume-over-capacity or V/C). At low V/C , drivers enjoy large spaces between cars and can safely travel at free-flow speeds. As roads become congested and V/C increases, drivers are forced to slow down to insure they have adequate time to react. Based on the volume-over-capacity (V/C) for each road, costs are updated according to Eq. (1), where $\alpha = 0.15$, $\beta = 4$ are used per guidelines set by the Bureau of Public Roads.³

$$t_{current} = t_{freeflow} \cdot (1 + \alpha(V/C)^\beta) \quad (1)$$

Though increments must be routed in serial, all routes discovered within an increment are independent. To speed up the routing process, we divide all trips in an increment into batches and send these batches to different threads for parallel computation. Because the road network remains fixed in each increment, we only need to store a single graph object shared by all threads. When a shortest path is found, we walk that path and increment counts of the number of vehicles that were assigned to each road and sum the counts from all batches after the increment has finished. We also keep track of the origin and destination census tracts of the assigned vehicles in a bipartite graph for later analysis. After all trips have been routed, we compute final V/C ratios and other metrics of each segment and update these values in the database so they can be used for other applications or visualization. Pseudo code for this ITA procedure can be found in [Algorithm 8 in the SI](#).

4. Results

In the following sections we demonstrate the range of outputs provided by our system. We first report trip tables and compare origin–destination matrices produced by our system to available estimates made using travel surveys. We then report road network performance as well as characteristics of road usage patterns enabled by the construction of a bipartite road usage network.

4.1. Trip tables and survey comparison

In order to understand when and where these new data will be effective and how the results differ from traditional approaches, we compare the output of our system to previous travel surveys wherever possible. In four of the cities studied, we find estimates of travel demand from surveys: the 2011 Massachusetts Household Travel Survey (MHTS) in Boston, the 2000 Bay Area Travel Survey (BATS) in San Francisco, a 2013 transportation plan in Rio de Janeiro, and estimates from a 2012 LUT model in Lisbon (Ferreira et al., 2010). While these surveys do not always produce all estimates we are able to generate with our system, we make comparisons wherever possible.

Trip tables report the total number of trips of a given purpose or during a given time of day for a city and represent the total load placed on transportation infrastructure. In [Table 2](#), we report trip tables for each city in this study. We find close agreement with trip tables estimated using CDR data and surveys in Boston and the San Francisco Bay Area and less agreement in Rio de Janeiro. We note, however, that the 3.74 million person trips estimated for Rio is far too low given the population of the region and highlights the difficulty in finding reliable planning resources in many areas. Finally, we note that in

³ Travel Demand Modeling with TransCAD 5.0, User's Guide <http://www.caliper.com/PDFs/TravelDemandModelingBrochure.pdf>.

Table 2

Trip tables estimates. Where possible, our results are compared to estimates made using travel surveys. For each city, we report the number of person trips in millions for a given purpose or time. Trip purposes include: home-based work (HBW), home-based other (HBO), and non-home-based (NHB). Trip periods include: 7–10 am (AM), 10 am–4 pm (MD), 4–7 pm (PM), and the rest of the day (RD). We note that the exact boundaries of the surveys do not exactly coincide with those used in our estimation so direct comparisons are not exact. In general, trip magnitudes align closely, with the exception of Rio de Janeiro, where the survey results report far too few trips, illustrating the difficulty of obtaining sensible measurements via certain techniques. No comparisons could be found for Porto.

City	HBW	HBO	NHB	AM	MD	PM	RD	Total
Boston	5.76	8.99	6.72	3.71	7.68	5.75	4.33	21.47
MHTS	3.22	12.83	9.49	5.32	8.87	8.20	3.15	25.54
SF Bay	4.07	10.05	7.04	4.47	7.81	5.35	3.53	21.16
BATS	4.60	11.54	4.66	4.18	6.90	4.22	3.00	20.80
Rio	9.92	17.17	11.46	7.71	14.09	10.47	6.29	38.55
Survey	2.06	–	–	1.31	1.19	1.24	–	3.74
Lisbon	1.08	2.01	1.21	0.79	1.67	1.26	0.58	4.30
Survey ^a	0.61	–	–	–	–	–	–	–
Porto	0.49	0.87	0.46	0.32	0.70	0.54	0.27	1.83
Survey	–	–	–	–	–	–	–	–

^a Note that the Lisbon Survey only contains estimates of vehicle trips in millions.

Lisbon, the survey results represent vehicle trips only, while we report person trips. When adjusting for mode car ownership rates in Portugal, our numbers align more closely. We were unable to find a survey or model for comparison in Porto.

In addition to trip tables, it is also necessary to compare the distribution of trips from place to place around the city. In order to make this comparison, the area unit of analysis for the survey and our model must be aligned. Given the resolution of mobile phone data, our system is designed to create ODs at the census tract (or equivalent) level while many surveys aggregate to larger traffic analysis zones or super districts. For comparison, we aggregate the OD matrices from CDRs to the coarser grained resolution provided by the survey and compare results. Fig. 3 shows correlation histograms comparing OD matrices at the largest spatial aggregation available produced by our methods and those produced by traditional methods. In general we find very high correlations in Boston, San Francisco, and Rio, with lower correlations in Lisbon. Lisbon, however, has the smallest units of aggregation and these results demonstrate the limitations of these comparisons at very high spatial resolutions. We hope future work explores how these correlations relate to the modifiable area unit problem. Finally, there is significant uncertainty in all models and we hope future works will explore this uncertainty further.

4.2. Road network analysis

The first output of this procedure is volume, congestion (volume-over-capacity), and travel times for all road segments. Using the outcomes of our analyses, we calculated the distributions of volumes on roads, along with V/Cs in Fig. 4. Interestingly, the results suggest qualitatively similarly distributed volumes and V/Cs for our five subject cities. Moreover, our findings are consistent with general congestion studies that identify Rio de Janeiro as one of the most congested cities in the world and the San Francisco Bay Area not far behind. Smaller cities such as Boston and Porto have fewer problems with congestion.

4.3. Bipartite road usage graph

In addition to measuring physical network properties of roads, the system architecture enables detailed analysis of individual road segments and neighborhoods within a city. Though the transient OD matrices constructed by Wang et al. (2012) correlate poorly with OD matrices developed by the methods above and traditional surveys, their work highlights new metrics of road usage patterns that can be measured via these new data sources. To this end we create a bi-partite usage graph. Every time a route between two location is assigned, we traverse the path and keep a record of how many trips from each driver source (census tract) used each road. This record is then used to construct a bipartite graph containing two types of nodes: road segments and driver sources, as shown in Fig. 5. Roads are connected to driver sources that contribute traffic to that segment and census tracts are connected to roads that are used by people who live here.

$$k_s^{road} = \sum_o A_{o \rightarrow s}, \quad k_o^{source} = \sum_s A_{o \rightarrow s} \quad (2)$$

$$A_{o \rightarrow s} = \begin{cases} 1, & \text{if vehicles from tract } o \text{ use road } s \\ 0, & \text{otherwise.} \end{cases}$$

We then examine the degree distributions of roads and census tracts using Eq. (2) in this bipartite graph to reveal patterns of road usage in Fig. 6. The number of roads used by residents of a given location is much more consistent between different cities and appears less affected by the size of the road network. On the other hand, the number of driver sources contributing traffic to a given road segment is broadly distributed, suggesting that most roads are *local* in that they serve only a few

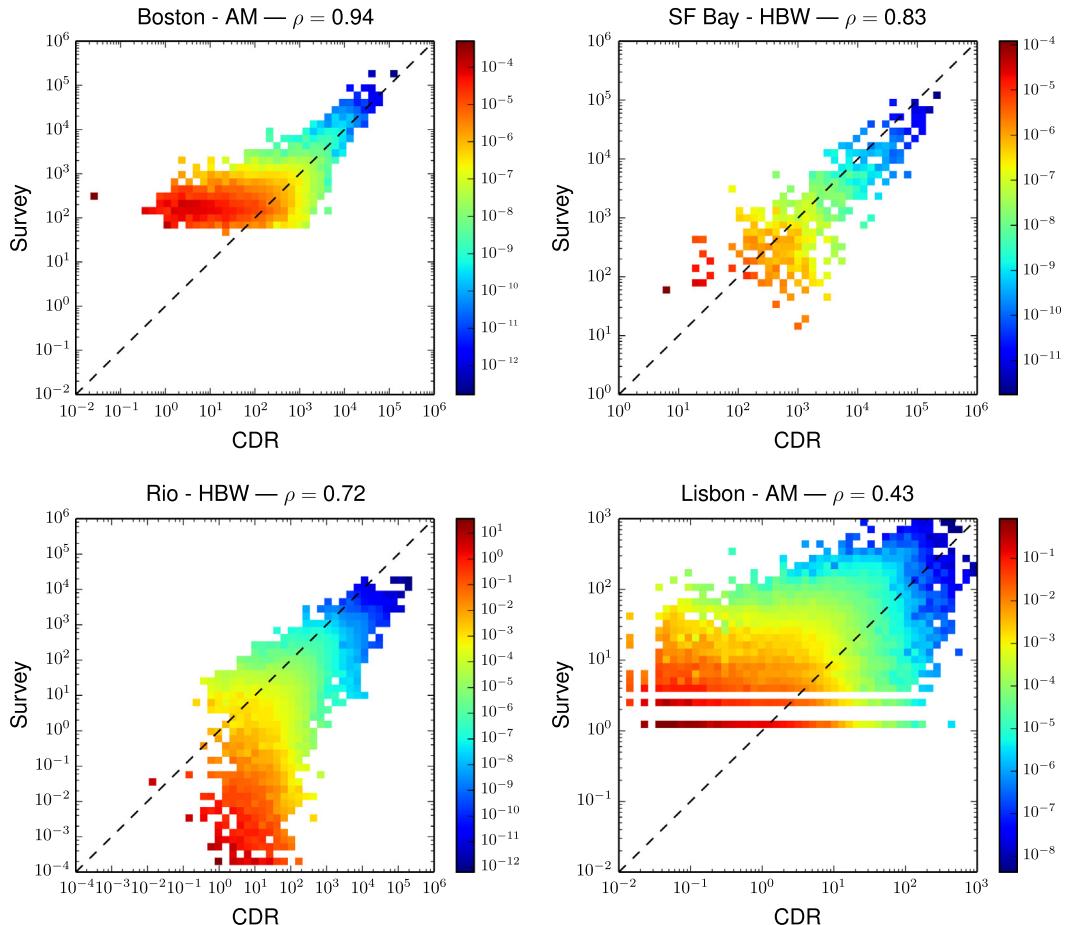


Fig. 3. Correlations between OD matrices produced by our system and those derived from travel surveys at the largest spatial aggregation of the two models. In Boston, this is town-to-town, in San Francisco, MTC superdistrict-to-super district, in Rio, census superdistrict-to-superdistrict, and in Lisbon, freguesia-to-freguesia. The larger of these area units (e.g. towns in Boston), the better our correlations, while correlations at the smallest aggregates (e.g. freguesias in Portugal), correlations are lower. However, more work must be done to understand uncertainties in estimates provided by both models.

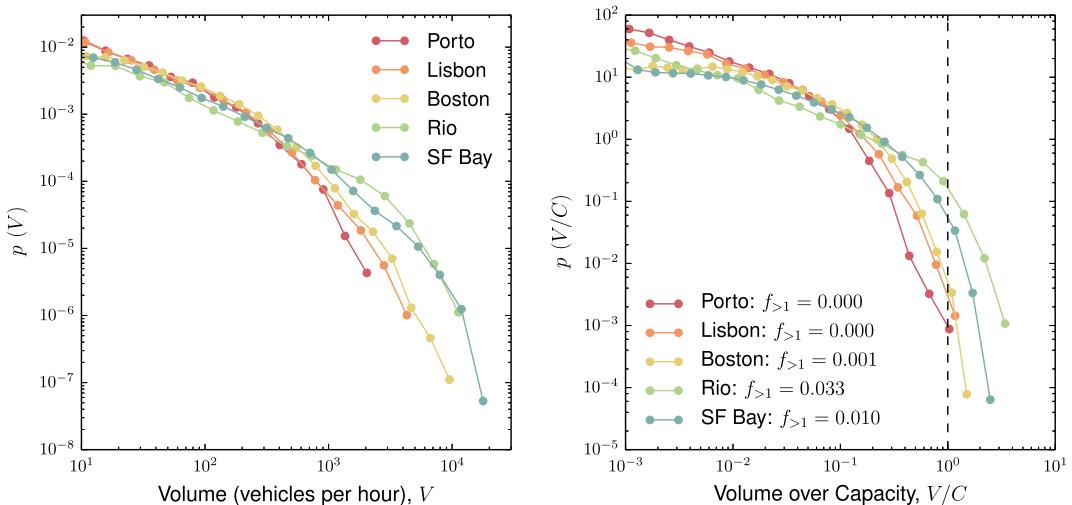


Fig. 4. Distributions of travel volume assigned to a road and the volume-over-capacity (V/C) ratio for the five cities. The values presented in the legend refers to the fraction of road segments with $V/C > 1$.

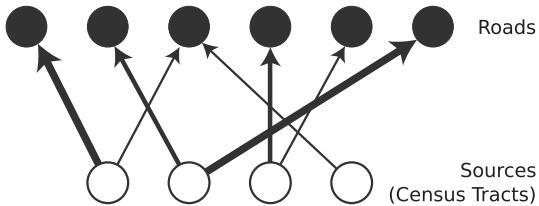


Fig. 5. A graphical representation of the bipartite network of roads and sources (census tracts), with edge sizes mapping the number of users using the connected road in their individual routes.

locations, while a few roads in the tail of the distribution are used for large fractions of the population. While this result is intuitive given that highways are designed for just this purpose, we hope future work explores the relationship between this bipartite usage graph and road network topology further.

An example of such an application was proposed by Wang et al. to classify road segments based on the relationship between topological and demand based metrics. Comparing the topological properties of roads in the physical network to the bipartite usage graph provides insights into their role in the transportation system. Edge betweenness centrality (Newman, 2005) captures the importance of a road by counting how many shortest paths between any two locations σ_{OD} must pass through that edge $\sigma_{OD}(e)$ (Eq. (3)). While this measure captures some aspects of importance, it treats all potential paths as equally likely and tends to be biased towards geographically central links. The degree of a road in the bipartite usage graph reflects the number of locations in the city that actually rely on that road because trips were assigned there from



Fig. 6. Distributions of k_{road} and k_{source} for the five cities.

actual travel demand. With these two metrics, betweenness centrality and a roads degree in the usage network, we can classify the role of a road in the cities transportation network.

$$bc_s = \sum_{o,d} \frac{\sigma_{OD}(s)}{\sigma_{OD}} \quad (3)$$

A simple classifier divides the betweenness usage degree space into four quadrants surrounding the point representing the 75th percentile for betweenness centrality and usage degree. Roads with betweenness and usage degree above the 75th percentile are both physical connectors and are used by large portions of the region. These roads tend to be bridges or urban rings. Roads with low betweenness, but high usage degree are attractors, receiving a higher proportion of trips than would be expected assuming uniform demand. Roads with high betweenness and low usage are physical connectors and serve an important purpose geographically, but may not be utilized by actual demand. Other roads, with low betweenness and low usage are local roads and primarily serve populations living and working nearby. Fig. 7 shows each road according to this classification using data from the ODs calculated via mobile phones.

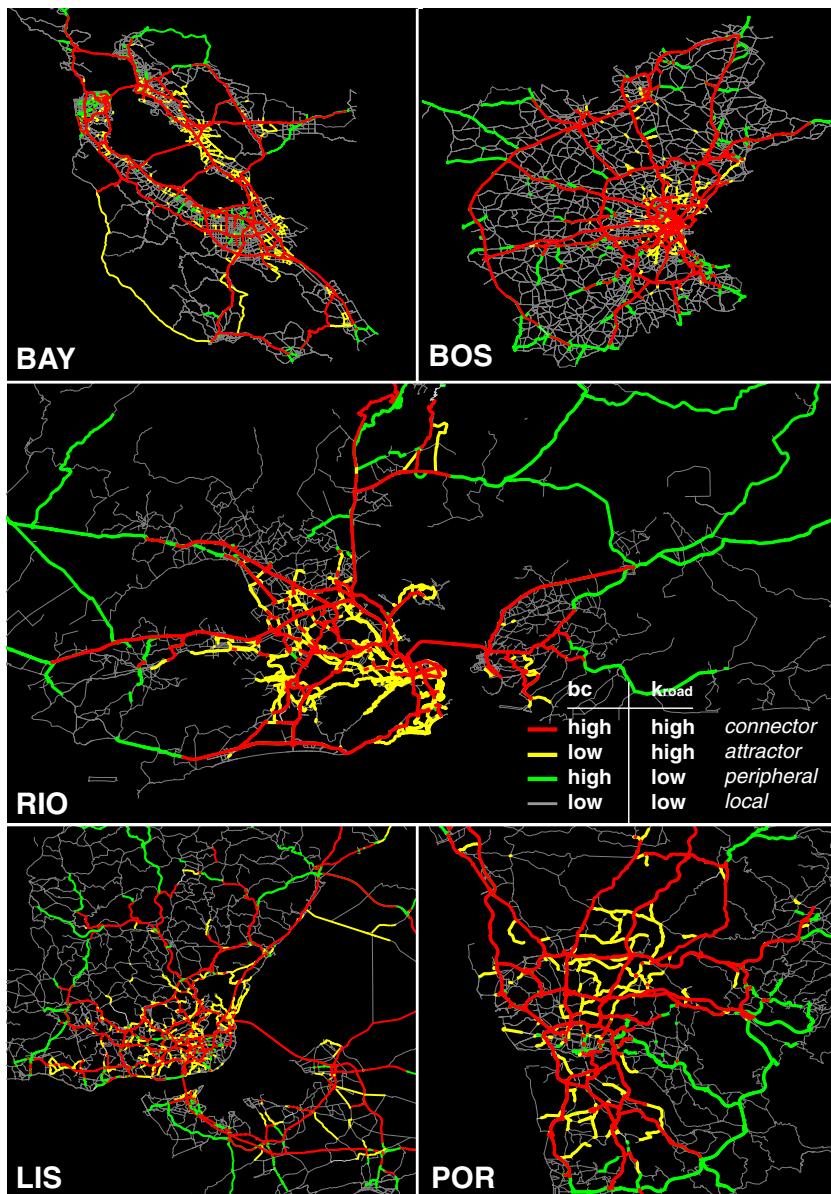


Fig. 7. Maps depicting the proposed road classification, summarized in the legend, for the five subject cities.

Finally, this bipartite framework of analysis allows us to augment visualizations of congestion maps in two ways. The first focuses on a single road segment. For example, when we identify a segment of a highway that becomes highly congested with traffic jams each day, we can easily query the bipartite graph to obtain a list of census tracts where drivers sitting in that traffic jam are coming from and where they are going to. The census tract nodes can also be given attributes from containing any demographic data a user wishes. With this information, it is possible to identify leverage points where policy makers can offer alternatives to these individuals or even power applications such as car sharing, by notifying drivers that others sharing the same road may be going to and from the same places. Moreover, businesses considering products or services based on who may be driving by or near different locations may find value in these detailed breakdowns.

Rather than selecting a road segment node, we may also select a single census tract, and check its neighbors to construct a list of all roads used by individuals moving to or from that location. For example, for a given neighborhood in a city we can identify all major arteries that serve that local population. This information provides a detailed look at a central location based on how much road usage it induces. Moreover, geographic accessibility, critical to many socio-economic outcomes, can now be measured in locations that were previously understudied.



Fig. 8. Two screen images from the visualization platform. (a) The trip producing (red) and trip attracting (blue) census tracts using Cambridge St., crossing the Charles River in Boston. (b) Roads used by trips generated at the census tract including MIT. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.4. Visualization

To help make these results accessible to consumers and policymakers, we build an interactive web visualization to explore road usage patterns in each city. Most GIS platforms can connect directly PostGIS databases to visualize and analyze road networks with our estimated usage characteristics. While these platforms are preferred by advanced users familiar with GIS data, they are opaque to many consumers who may benefit from more detailed information on road usage. A simple API is implemented to query the database and generate standard GeoJSON objects containing geographic information on roads as well as computed metrics such as level of service. We also implement queries to answer questions such as “What are all the census tracts used by drivers on a particular road?” or “What are all roads used by a given location in the city?”. These data are then parsed and displayed on interactive maps using any of the available online mapping APIs and D3js allowing users, with functionality that enables one to select individual roads and areas. Two screen images of this system is shown in Fig. 8.

5. Discussion and limitations

This paper has presented a full implantation of a travel demand model that uses new, big data resources as input. We have presented a system that combines and improved upon many disparate advanced in recent years to produce fast, accurate, and inexpensive travel demand estimates. We began by outlining methods to extract meaningful locations from noisy call detail records and estimate origin–destination matrices by counting trips between these places. Normalized and scaled trips counts are compared to estimates made using survey data in both trip tables and at the OD pair level. These flows are then assigned to road networks constructed from OpenStreetMap data using an incremental traffic assignment algorithm. As routes are assigned, a number of metrics on road usage are measured and stored.

While these results show great progress in making big data useful for transportation engineering, there are still limitations inherent in this data and our model. Specifically, we highlight three areas that are ripe for further study.

1. We have shown the level of aggregation applied to OD matrices can affect the correlation observed between model outputs. This is a standard manifestation of the modifiable area unit problem and a more detailed exploration may indicate which levels of analyses were better suited for different data sources. Moreover, a more detailed analysis of uncertainty in model estimates may make it easier to assess their correlation and validity.
2. Our traffic assignment algorithm is efficient, but simple. In the future, a stochastic dynamic user equilibrium assignment methods should be explored and compared. Moreover, route choice modeling may be significantly improved by the availability of high resolution GPS trajectories of drivers. We believe our system's modular design makes it easy to incorporate these new models.
3. Our mode choice model remains simple and will likely require more sophistication for modeling trips not taken in private vehicles. This, combined with improvements in route choice, may make it possible to estimate multi-modal trip demand, as public transportation, bike lanes, and even water transportation networks are included in OpenStreetMap data.

We hope future work will address these and improve further on the methods presented here.

6. Conclusions

Transportation engineers and urban planners have a rich history estimating flows of people within cities and mapping this flow onto transportation infrastructure. However, these efforts are often constrained by limited data resources. The rise of ubiquitous mobile sensors has generated a wealth of new data on human mobility, but new tools must be developed to integrate these data and insights into traditional transportation modeling approaches. To this end, we have demonstrated a full implementation of a travel demand model utilizing mobile phone data as an input. We presented algorithms to generate routable road networks from open source data repositories, generate validated OD matrices and trip tables from CDR data, and route these trips through road networks using a paralleled ITA algorithm. We have demonstrated a number of possible analyses that can be performed on the output of this system including network performance and classification measurements and an online, interactive visualization platform.

As more data becomes available in the form of calls, gps traces, or real time traffic monitoring systems, we are excited at the prospect of updating and improving these systems further.

Acknowledgments

This work was partially funded by the BMW-MIT collaboration under the supervision of PI Mark Leach,⁴ the World Bank-HuMNet collaboration agreement under the supervision of PI Shomik Mehndiratta⁵ and the Center for Complex

⁴ mark.leach@bmw.de.

⁵ smehndiratta@worldbank.org.

Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris.⁶ We thank Pu Wang for technical support, Shan Jiang for her help obtaining LUT model results for Lisbon, Nelson F.F. Ebecken for support with data, the Rio de Janeiro State Agency (FAPERJ) for the grant on this project, and the Rio City Hall for the support and the data they have provided. Our work was also supported, in part, by the UPS Center for Transportation and Logistics Graduate Research Fellowship awarded to Serdar Çolak and by the National Science Foundation Graduate Research Fellowship awarded to Jameson L. Toole. Lauren P. Alexander and Bradley Sturt are supported by the Austrian Institute of Technology and the MIT-Smart program, respectively.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.trc.2015.04.022>.

References

- Akcakil, R., 1991. Travel time functions for transport planning purposes: Davidson's function, its time dependent form and alternative travel time function. *Aust. Road Res.* 21 (3).
- Alexander, L.P., Jiang, S., Murga, M., González, M.C., 2015. Validation of origin–destination trips by purpose and time of day inferred from mobile phone data. *Transport. Res. Part C: Emerg. Technol.*
- Asakura, Y., Hato, E., 2004. Tracking survey for individual travel behaviour using mobile communication instruments. *Transport. Res. Part C: Emerg. Technol.* 12 (3), 273–291.
- Bar-Gera, H., 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from israel. *Transport. Res. Part C: Emerg. Technol.* 15 (6), 380–391.
- Bast, H., Funke, S., Sanders, P., Schultes, D., 2007. Fast routing in road networks with transit nodes. *Science* 316 (5824), pp. 566–566.
- Belik, V., Geisel, T., Brockmann, D., 2011. Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X* 1 (1), 011001.
- Bell, M.G., 1991. The estimation of origin–destination matrices by constrained generalised least squares. *Transport. Res. Part B: Methodol.* 25 (1), 13–22.
- Branston, D., 1976. Link capacity functions: a review. *Transport. Res.* 10 (4), 223–236.
- Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. *Nature* 439 (7075), 462–465.
- Caceres, N., Wiedberg, J., Benitez, F., 2007. Deriving origin destination data from a mobile phone network. *Intell. Transp. Syst. IET* 1 (1), 15–26.
- Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr., J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transport. Res. Part C: Emerg. Technol.* 26, 301–313.
- Canclia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A: Math. Theor.* 41 (22), 224015.
- Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. *Transport. Res. Part B: Methodol.* 18 (4), 289–299.
- Colak, S., Schneider, C.M., Wang, P., González, M.C., 2013. On the role of spatial dynamics and topology on network flows. *New J. Phys.* 15 (11), 113037.
- Çolak, S., Alexander, L.P., Alvim, B.G., Mehndirettta, S.R., González, M.C., 2015. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transport. Res. Rec. J. Transport. Res. Board* 2183 (1), 85–93.
- Daganzo, C.F., 1980. Optimal sampling strategies for statistical models with discrete dependent variables. *Transport. Sci.* 14 (4), 324–345.
- de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* 3.
- Ferreira, J., Diao, M., Zhu, Y., Li, W., Jiang, S., 2010. Information infrastructure for research collaboration in land use, transportation, and environmental planning. *Transport. Res. Rec. J. Transport. Res. Board* 2183 (1), 85–93.
- González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. *Nature* 453 (7196), 779–782.
- Hazelton, M.L., 2000. Estimation of origin–destination matrices from link flows on uncongested networks. *Transport. Res. Part B: Methodol.* 34 (7), 549–566.
- Hazelton, M.L., 2001. Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transport. Res. Part B: Methodol.* 35 (7), 667–676.
- Hazelton, M.L., 2003. Some comments on origin–destination matrix estimation. *Transport. Res. Part A: Policy Pract.* 37 (10), 811–822.
- Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. *Transport. Res. Part C: Emerg. Technol.* 40, 63–74.
- Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr., J., Fazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM, p. 2.
- Lo, H., Zhang, N., Lam, W.H., 1996. Estimation of an origin–destination matrix with random link choice proportions: a statistical approach. *Transport. Res. Part B: Methodol.* 30 (4), 309–324.
- Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proc. Nat. Acad. Sci.* 109 (29), 11576–11581.
- Lu, C.-C., Zhou, X., Zhang, K., 2013. Dynamic origin–destination demand flow estimation under congested traffic conditions. *Transport. Res. Part C: Emerg. Technol.* 34, 16–37.
- Maher, M., 1983. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. *Transport. Res. Part B: Methodol.* 17 (6), 435–447.
- Merchant, D.K., Nemhauser, G.L., 1978. A model and an algorithm for the dynamic traffic assignment problems. *Transport. Sci.* 12 (3), 183–199.
- Newman, M.E.J., 2005. A measure of betweenness centrality based on random walks. *Soc. Netw.* 27 (1), 39–54.
- Nie, Y., Zhang, H., Recker, W., 2005. Inferring origin–destination trip matrices with a decoupled GLS path flow estimator. *Transport. Res. Part B: Methodol.* 39 (6), 497–518.
- Ortúzar, J.D., Willumsen, L.G., 1994. Modelling Transport. John Wiley & Sons, Chichester, England.
- Phithakkittnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C., 2010. Activity-aware map: identifying human daily activity pattern using mobile phone data. *Hum. Behav. Underst.*, 14–25
- Reades, J., Calabrese, F., Ratti, C., 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environ. Plan. B: Plan. Des.* 36 (5), 824–836.
- Richardson, A.J., Ampt, E.S., Meyburg, A.H., 1995. Survey Methods for Transport Planning. Eucalyptus Press, Melbourne.
- Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unravelling daily human mobility motifs. *J. Roy. Soc. Interface* 10 (84), 20130246.
- Sevtsuk, A., Ratti, C., 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *J. Urb. Technol.* 17 (1), 41–60.
- Smith, M.E., 1979. Design of small-sample home-interview travel surveys. *Transport. Res. Rec.* 701, 29–35.
- Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010a. Limits of predictability in human mobility. *Science* 327 (5968), 1018–1021.
- Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010b. Modelling the scaling properties of human mobility. *Nat. Phys.* 6 (10), 818–823.

⁶ anas@mit.edu.

- Spiess, H., 1987. A maximum likelihood model for estimating origin–destination matrices. *Transport. Res. Part B: Methodol.* 21 (5), 395–412.
- Spiess, H., 1990. Technical note – conical volume-delay functions. *Transport. Sci.* 24 (2), 153–158.
- Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: where are we going? *Transport. Res. Part A: Policy Pract.* 41 (5), 367–381.
- Van Zuylen, H.J., Willumsen, L.G., 1980. The most likely trip matrix estimated from traffic counts. *Transport. Res. Part B: Methodol.* 14 (3), 281–293.
- Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. *Sci. Rep.* 2 (1001). <http://dx.doi.org/10.1038/srep01001>.
- Wardrop, J.G., 1952. Road paper. some theoretical aspects of road traffic research. *ICE Proceedings: Engineering Divisions*, vol. 1. Thomas Telford, pp. 325–362.
- Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O., 2012. Quantifying the impact of human mobility on malaria. *Science* 338 (6104), 267–270.
- Zhan, X., Hasan, S., Ukkusuri, S.V., Kamga, C., 2013. Urban link travel time estimation using large-scale taxi data with partial information. *Transport. Res. Part C: Emer. Technol.* 33, 37–49.
- Zheng, Y., Xie, X., 2011. Learning travel recommendations from user-generated GPS traces. *ACM Trans. Intell. Syst. Technol. (TIST)* 2 (1), 2.

Urban Attractors: Discovering Patterns in Regions of Attraction in Cities

May Alhazzani¹, Fahad Alhasoun², Zeyad Alawwad¹, Marta C. González^{2,3},

1 Center for Complex Engineering Systems, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

2 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

3 Center for Advanced Urbanism, Massachusetts Institute of Technology, Cambridge, MA 02139.

* martag@mit.edu

Abstract

Understanding the dynamics by which urban areas attract visitors is significant for urban development in cities. In addition, identifying services that relate to highly attractive districts is useful to make policies regarding the placement of such places. Thus, we present a framework for classifying districts in cities by their attractiveness to visitors, and relating Points of Interests (POIs) types to districts' attraction patterns. We used Origin-Destination matrices (ODs) mined from cell phone data that capture the flow of trips between each pair of places in Riyadh, Saudi Arabia. We define the attraction profile for a place based on three main statistical features: The amount of visitors a place received, the distribution of distance traveled by visitors on the road network, and the spatial spread of where visitors come from. We use a hierarchical clustering algorithm to classify all places in the city by their features of attraction. We detect three types of Urban Attractors in Riyadh during the morning period: **Global** which are significant places in the city, **Downtown** which is the central business district and **Residential** attractors. In addition, we uncover what makes these places different in terms of attraction patterns. We used a statistical significance testing approach to rigorously quantify the relationship between Points of Interests (POIs) types (services) and the 3 patterns of Urban Attractors we detected. The proposed framework can be used for detecting the attraction patterns given by type of services related to each pattern. This is a critical piece of information to inform trip distribution models.

Introduction

Understanding how different places in the city influence human mobility is significant for urban and transportation planning. A pressing need, in complex and congested cities is maintaining a robust transportation infrastructure. Understanding the patterns by which places in the city attract visitors is essential for planning and modifying the transportation system. Specifically, new data sources can help us to better predict the relation between population and trips attraction. This issue is of a particular significance to the city of Riyadh, Saudi Arabia where the largest metro project is being developed and promised to start running in 2019 [1, 2]. Moreover, understanding how different types of places affects the flow of trips in the city differently helps to inform

decisions and policies related to placing and modifying concentration of services. For example, how an industrial area influences the flow of trips during different times of the day; or where to place new business stores for higher profitability [3].

Today, with the ubiquity and pervasiveness of technology, data generated from mobile phones enable data analysts to better understand the behavior of individuals across many dimensions including their mobility patterns [4,5]. An interesting area is how patterns of human mobility are affected by different places in the city. The standard approach on categorizing urban areas classifies regions by their functionality and land use (i.e. commercial, educational, ...etc.). Recent works consider the human mobility aspects to classify regions. For example, Yuan *et al.* [6] proposed a topic modeling approach to classify districts into functional zones according to people's socioeconomic activities mined from taxi and public transport traces and points of interests (POIs) data. Pan *et al.* [7] proposed a land use classification approach based on the social functions of districts also analyzed from GPS taxi traces where districts witness change of land use class dynamically. Toole *et al.* [8] analyzed cell phone data to test cell phone activity patterns to classify for land use types. Less is known about how phone data can help to classify urban regions based on how attractive they are to different origins.

In this work, we present a novel computational framework for classifying urban places by their attraction patterns. We define attraction profiles in terms of statistical features of incoming trips on a given time window. Different places in the city attract visitors differently. Some places like universities and hospitals attract a large amount of visitors who come from all over the city and travel long distances to go there. On the other hand, some places in the city that provide local services such as restaurants, schools, and small clinics only attract few people from nearby areas. We aim to automatically identify patterns of attraction of places based on three main dimensions: how many visitors a place receives, where visitors are coming from, and how long visitors are traveling to reach that place. Secondly, we further show what makes a region attraction behavior. To accomplish that, we used statistical significance testing to automatically relate the decomposition of POI types (services) and the discovered attraction patterns. This information is useful to relate type of businesses and attraction profiles.

The main contribution to this work is as follows:

- We present a computational framework for detecting attraction patterns and further relating POI types to each pattern of attraction.
- We classify attraction patterns via the spatial dispersion of trip origins, the distribution of distances traveled by visitors through the road network and the total number of trips.
- We quantify the significance of POI types in a region using a statistical significance testing approach, which performs well in the context of phone and POI data.

Related work

Multiple studies used human mobility behavior to classify urban areas. A recent study investigated the relationship between land use and mobility [9]. The authors showed that purposes of people's trips are strongly correlated with the land use of the trip's origin and destination. Recently, the availability of dynamic sources of data allowed for dynamic segmentation of the city according to human mobility behavior. Some studies combined human mobility with land use or POIs data to segment districts in urban areas according to their functions or use. The type of data used to capture human

mobility behavior varies between individual GPS traces [10, 11], taxi pick up/drop off locations as in [7, 12], Call Detail Records (CDRs) as in [2, 8], social media check ins as in [13–15], and bus smart card data as in [16]. This work differs from previous studies, by being the first to classify the urban regions through attraction profiles.

Survey travel data has been used to detect the centers (significant places) of a city [17, 18]. A recent study proposed a method for measuring the centrality of locations that incorporates the number of people attracted to the location and the diversity of activities in which visitors engage [17]. The proposed method was tested on survey travel data in Singapore to identify the functional centers and track their significance over time. A similar approach focused on analyzing the aggregate behavior of the population to predict highly attractive events such as the times square during new years count down in New York [19]. Our method is based on validated Origin Destination (ODs) matrices mined from massive cell phone data that captures human mobility. More significantly, our approach incorporate not just the amount of people a place attracts, but also on where do they come from and the road distance they traveled.

Network analysis methods were used to detect hotspots based on flow patterns between locations [2, 20]. A recent paper [2] used ODs matrices extracted from cell phone data to identify the signature of mobility behavior as 4 main types of movements within the city: between hotspots, to hotspots, originating at hotspots and random flows. They showed how different cities have different mobility signatures. Additionally, a recent study used Taxi drop off/pick up traces in Shanghai to create a network of flow between places. They applied community detection to extract sub regions and analyze the interaction between and within sub regions. They found that urban sub-regions have larger internal interactions, while suburban centers are more significant on local traffic. This work made the breakdown of flow patterns instead of the impact of the place in attracting visitors, which is our aim in this paper. Researchers adapted modeling approaches from Natural Language Processing (NLP) in identifying functional zones in urban areas [6, 21]. One study applied a Latent Dirichlet Allocation (LDA) model on Foursquare check-ins to detect local geographic topics that indicate the potential and intrinsic relations among the locations in accordance with users' trajectories. Additionally, a recent study used LDA and POIs to detect functional zones [6]. Our work is different where we aim to analyze the attraction behavior of a place using measures that has not been used in any of the previous work.

Urban Attractors Framework

Fig.?? shows the general structure of the process of analyzing attraction patterns in cities with the input datasets and the outputs. The first step in the process is to extract trips information from Call Detail Records (CDRs) of cell phones using the validated origin destination extraction algorithm implemented in [22]. We use the ODs as a data source for estimating human mobility, where it provides the amount of trip from each pair of origin and destination. From the ODs, we mine three statistical features that quantify how attractive a place is: the number of trips a place receives, the spatial dispersion of the origins of all incoming trips, and the distance distribution visitors traveled to visit the place on the road network. We use these attraction features to classify all regions in the city according to their attraction behavior. Finally, using a statistical significance testing approach we relate each type of POIs that are significantly concentrated in each types of attractors identified. In the following sections we explain the process and its output in detail.

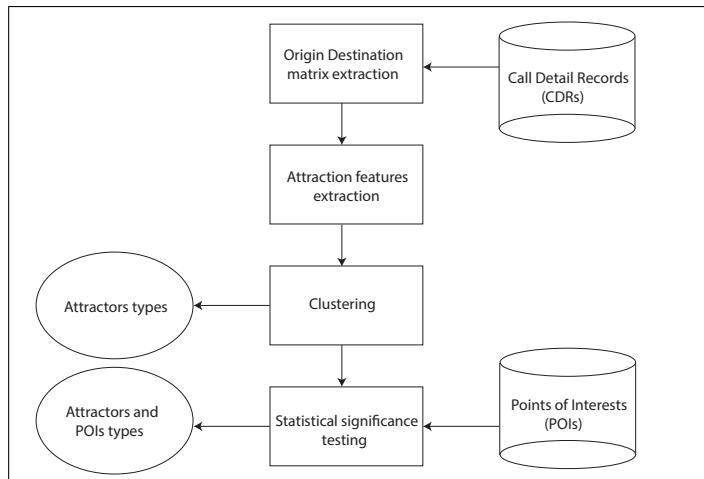


Fig 1. Urban Attractors framework

Origin Destination Matrix Extraction

The aim of this process is to extract the Origin-Destination matrices (ODs), that provides the number of trips between each pair of locations in the city for a specified time window of a typical weekday. Methods of estimating ODs range from traditional methods to more modern ones. Traditional methods include running surveys within cities and estimating the flows between locations of the city from the feedback of those surveys. Such methods consume longer periods of time and are inaccurate at times. They usually span smaller population sample sizes and thus are more prone to biases. Recent research in the domain of ubiquitous computing provided alternative methodologies for estimating more accurate ODs from user generated datasets like cell phone data. The methods proposed in [22, 23] uses mobile phone location traces (i.e. CDRs) to estimate the flows of people between areas in the city. The large scale of the cell phone data provide sufficient sample sizes and more accurate information compared to traditional methods. In this paper, we use state of the art methods of extracting OD matrices for the city of Riyadh between each pair of traffic analysis zones (TAZes) as shown in Fig.2.

Our primary source of data is one month (December 2012) of Call Detail Records (CDRs) of anonymous mobile phone users in the city of Riyadh, Saudi Arabia. Within the CDRs, each record contains an anonymized user ID of the caller and receiver, the type of communication (i.e., SMS, MMS, call, data etc), the cell tower ID facilitating the service, the duration, and a time stamp of the phone activity. Each cell tower ID is spatially mapped to its latitude and longitude where each Voronoi cell in Fig.2 correspond to a tower. The CDRs data contains more than 3 million unique users, which is a representative sample of Riyadh's population. Thus, the CDRs provide a proxy for tracking human mobility behavior in the city. However, computational steps are needed to extract clean trajectories from the CDRs.

The computational steps we take to transform raw mobile phone data (CDRs) to ODs are summarized as follows. First, we estimate transition probabilities between labeled locations such as home, work or other. Next, we filter users by number of records per day such that these locations are labeled with enough confidence (details of the method in [22]). A trip is marked by probable departures and arrivals to such locations within a specified time window. Then we scale from users to total population using census data. The output of this process is the OD matrix that indicates the

number of individuals traveling between every possible pair of locations between 7 am and 10 am on a typical weekday.

The spatial scale we used for locations is based on Traffic Analysis Zones, which is the official segmentation used in transportation planning. Conventionally segmenting the city into TAZes are based on census block information such as population per hour, where zones tend to be smaller in denser areas and larger in areas of low density. The TAZ based segmentation is more flexible and useful in analyzing places attraction patterns than using other segmentations such as neighborhood based or spatially uniform segmentation. Thus, we define our OD matrix T by aggregating cell phone towers on 1492 TAZes. The elements of the matrix are the number of trips between each pair of TAZes (i, j) in the city.

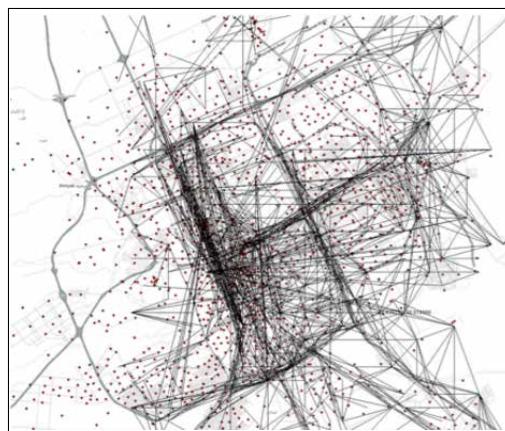


Fig 2. The ODs in Riaydh during the morning period. Each line represents a trip from a source to a destination.

Attraction Features

We aim to identify different patterns of attraction through statistical features of the inflow to each place in the city. The first feature is the total amount inflow a place receives. As the more visitors a place receives, the more attractive that place is. Additionally, a place is more attractive if it attracts people from various places in the city. Some places only attract people nearby which makes them local in terms of from where they attract people. On the other hand, some place attracts people from all over the city such as universities and hospitals. Thus, the second feature we measure is how spatially dispersed the origins are. Finally, we quantify the distances traveled to visit the place on the road network. In the following sections, we detail the calculations of each feature.

Inflow

The amount of visitors a place receives is the strongest indicator of how attractive the place is. This feature measures the attraction force of a location , where locations that have high inflow (number of visitors) are major attractors in the city. Fig.3 shows the distribution of the number of TAZes according to their inflow amount. The majority of TAZes have small to moderate inflow. However, there are few TAZes that have a very large inflow (colored red), which makes them highly significant. The inflow magnitude

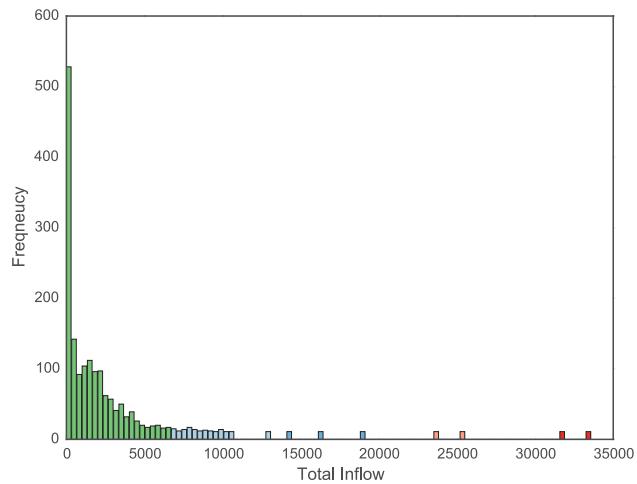


Fig 3. The distribution of total inflow received by TAZ's in Riaydh. the majority of places receive small to medium number of visitors. Few places receive very high inflow (colored red),which makes them highly attractive.

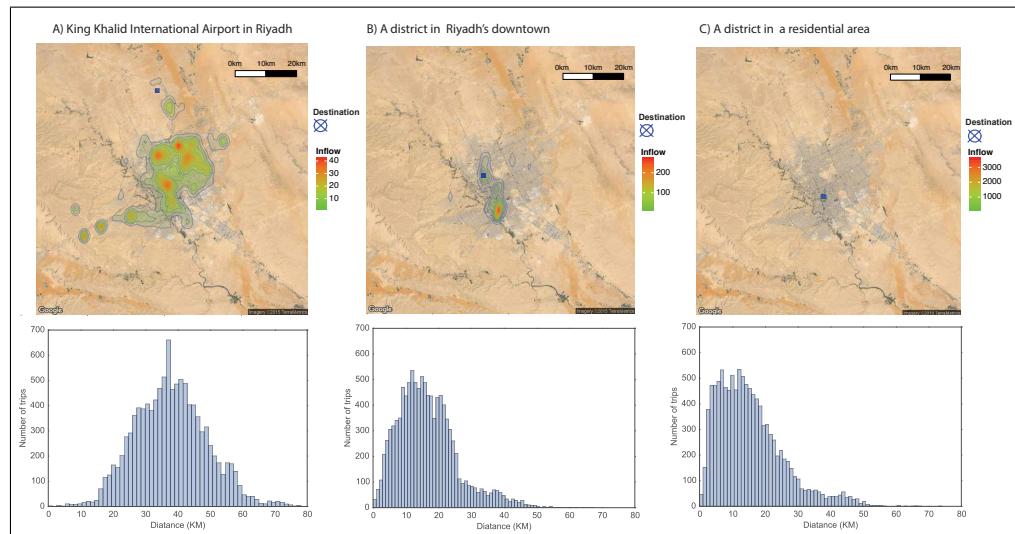


Fig 4. Spatial dispersion and distance distributions of three examples of types of attractors (marked by black dots). (A) The international airport in Riaydh. (B) A place in the downtown area. (C) A place in a residential area. The top row shows heatmaps of the origins of the inflow, where the color differs corresponds by the amount of trips from that location. The bottom row is the distribution of road distance traveled by visitors of the selected place.

of a TAZ i is simply calculated from the OD matrix as follows:

$$Inflow_i = \sum_{j=1}^n T_{ji} \quad (1)$$

Where n is the total number TAZes, and T_{ji} is the number of trips from TAZ j to TAZ i .

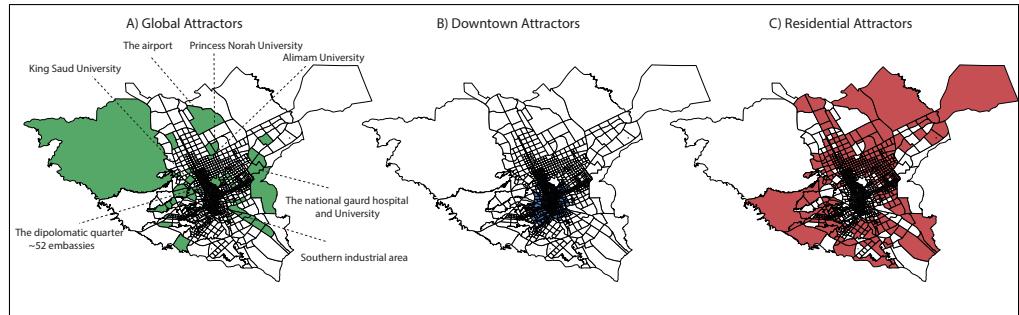


Fig 5. Types of attractors. (A) Global attractors, which are places with large number of visitors from all over the city traveling long distances to visit them. These include unique places in the city like the airport, large universities, hospitals, industrial areas, and the diplomatic quarter as annotated on the figure. (B) Attractors with large number of visitors due to their central location (downtown) they are more accessible and thus the distance visitors travel to reach them are shorter. (C) Less attractive places, that are located on the outer (residential) areas of the city.

Spatial dispersion

Another way to measure the popularity of a place is to measure the spatial dispersion of the origins it attracts. The spatial dispersion quantifies how spatially dispersed the locations of the origins of trips are in relation to the center of mass of all origins. A place is more attractive if it attracts visitors from various and spread-out places in the city. Major attractors tend to attract people from all over the city (large spatial dispersion), while less significant attractors only attract people nearby (presenting small spatial dispersion).

We measure the spatial dispersion of visitors by calculating a weighted standard distance deviation, which is a standard method used to measure the statistical dispersion of spatial data [24]. Mathematically, the weighted spatial dispersion (SD) for a TAZ i is defined as follows:

$$SD_i = \sqrt{\frac{\sum_{i=1}^n w_i (X_i - \bar{X}_c)^2 + \sum_{i=1}^n w_i (Y_i - \bar{Y}_c)^2}{\sum_{i=1}^n w_i}} \quad (2)$$

Where n is the total number TAZes. X_i and Y_i are the spatial coordinates of the origin of a trip i . w_i is the amount of inflow from source TAZ i . \bar{X}_c, \bar{Y}_c are the coordinates of the spatial center of mass of all origins of all the incoming flow calculated as follows:

$$\bar{X}_c = \frac{\sum_{i=1}^n w_i \cdot X_i}{\sum_{i=1}^n w_i}, \bar{Y}_c = \frac{\sum_{i=1}^n w_i \cdot Y_i}{\sum_{i=1}^n w_i} \quad (3)$$

Fig.4 shows three examples of places with different attraction behavior. The top row shows the heat maps of the inflow sources and their concentration. The destination TAZ is labeled with a target sign on the maps. Example A shows the heat map of the international airport in Riyadh city, where the heat is spread all over the city, which indicates strong attraction. Example B is a TAZ in the downtown area, where the inflow sources are moderately spread. Example C is in a TAZ in a residential area, where it only attracts visitors nearby with small spatial dispersion.

Distance distribution

Another characteristic that defines the attraction patterns is the distribution of distances traveled by all the trips the destination receives. The trip distance from each

source to the centroid of the attractor were calculated on the road network of Riyadh by using the Dijkstra shortest path algorithm [25] to find the optimal routes between all of the origin-destination pairs. This provides a more accurate estimation than the Euclidean or the Manhattan distances, as it accounts for the variation in the geometry of the road network.

The bottom row in Fig.4 shows the distance distribution of all trips a TAZ received. In Fig.4 A the distance distribution to the airport is unique, with long mean distance (around 40 km.) and a shift in the distribution due to the distant location of the airport in the very far north of the city. On the other hand, Fig.4 B has intermediate length of the mean distance with a distribution tail that corresponds to the long distance traveled by some visitors to reach downtown. Finally, in Fig.4 C, most of the trips correspond to short distances. Clearly, for different types of attractors the distance distributions differ. Thus, we select the mean and the standard deviation of the distribution as the features to distinguish attraction behaviors.

Clustering

To discover common patterns of inflow within cities, regions are clustered using the attraction features discussed in the previous section. We used a Hierarchical Agglomerative Clustering (HAC) approach to categorize all 1492 TAZes in Riyadh based on their attraction features. HAC classifies objects, where each object is represented as a vector of features that describe that object, based on specified similarity metrics. Here, a vector x_i represent the attraction features that describe TAZ i as follows:

$$x_i = [inflow_i, SD_i, \mu_i, \sigma_i] \quad (4)$$

Where $inflow_i$ is the inflow magnitude of TAZ i , SD_i is the spatial dispersion of the inflow sources for TAZ i , μ_i is the mean of the distances traveled to TAZ i , and σ_i is the standard deviation of the traveled distance distribution.

HAC starts by assigning each single object to a separate cluster, and sequentially merge the most similar clusters until it results in one cluster. Thus, HAC requires defining how to merge clusters and how to measure the distance between them. For merging clusters, We used complete-linkage algorithm, which merges two clusters based on their most dissimilar objects as follows:

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y) \quad (5)$$

Where $d(x, y)$ is the distance between two objects $x \in X$ and $y \in Y$, and X and Y are the 2 sets of clusters. Complete-linkage algorithm is conservative when merging clusters, thus it tends to find very compact clusters, which fits our objective in finding closely related attraction patterns. For measuring the distance between clusters' objects $d(x, y)$, we use correlation based distance metric defined as follows:

$$d(x, y) = 1 - \frac{(x - \bar{x}) \cdot (y - \bar{y})}{\|(x - \bar{x})\|_2 \|(y - \bar{y})\|_2} \quad (6)$$

Where \bar{x} and \bar{y} are the mean of the elements of vector x and y correspondingly, and $(x - \bar{x}) \cdot (y - \bar{y})$ is the dot product of the vectors $(x - \bar{x})$ and $(y - \bar{y})$. Correlation distance works well for finding unbalanced clusters sizes as we expect to have small number of places behaving very uniquely as strong attractors and larger number of places that are not as attractive. Additionally, the correlation score can correct for any scaling within a feature, while the final score is still being tabulated. Thus, different features that use different scales can still be used.

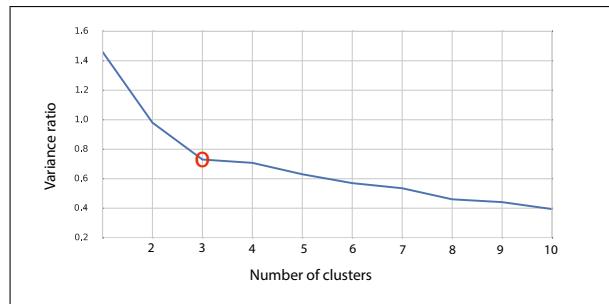


Fig 6. The ratio of the within-cluster variance to the total variance for each possible choice of K (number of clusters). The variance decreases as the data is split into more clusters until it stops decreasing significantly at $k = 3$ (marked by a red circle).

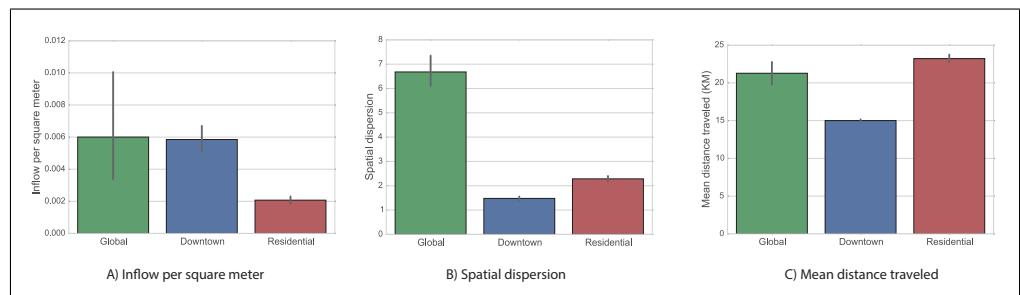


Fig 7. The attraction features of the three detected types of attractors clusters. (A) shows the total inflow, which is high for both global and downtown attractors but extremely low for residential attractors. (B) shows the spatial dispersion of origins of visitors, where it's significantly high in global attractors. (C) shows the mean of the distance traveled by visitors, where downtown attractors exhibit smaller distance mean due to its central location and thus higher accessibility to most visitors

HAC provides a hierarchy structure of the classified regions in the city. To determine the number of clusters k that best divide the data, we calculate the ratio of the between-cluster variance to the total variance for each possible k from 1 to 10. The variance drops as k increases, until it stops decreasing significantly. We select the k that correspond to the point where the variance stops decreasing significantly, which is $k = 3$ in our case as shown in Fig.6. The classification process over all TAZes in the city of Riyadh finds three types of attractors that have distinct features. The following section extends these findings to further interpret the results.

Attractor Types

Locations in Riyadh are classified into three main types of attractors based on distinguishable attraction of trips. Fig.5 shows the 3 types of attractors classes detected, where the polygons are the TAZes. The global attractors are the ones that have significant influence on the whole city, hence their name. Unlike the remaining clusters, the locations of these places seem to be random around the city. The second detected type is the downtown attractors, which play a significant influence, after global, to attract trips. These are mostly clustered in the downtown area of the city. Finally, the residential attractors, are the least influential attractors in the morning period of typical weekdays. They are mostly located on the outer places of the city.

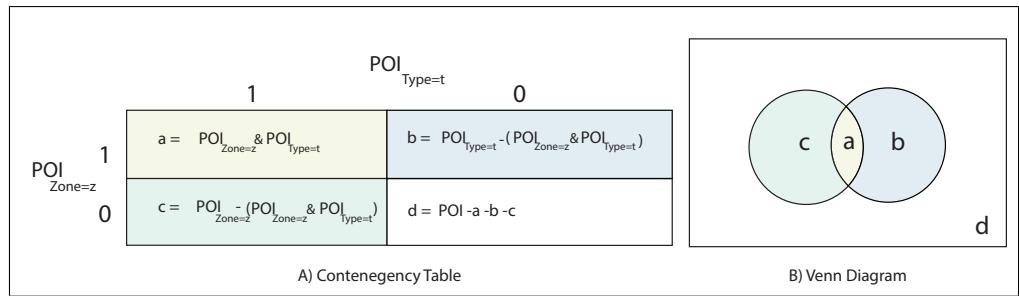


Fig 8. The contingency table and the corresponding Venn diagram for statistical significance testing method.

Global Attractors

The most distinguishable feature of global attractors is the large spatial dispersion of the incoming flows, as visitors come from all over the city to visit these places, as shown in Fig.7 A. Additionally, the amount of visitors they attract is the largest as shown in Fig.9 B, where we use inflow per square meters due to the unbalanced sizes of the TAZes. Moreover, the mean distances traveled by visitors to these locations is extremely high, which makes these places highly attractive and unique. We call these places global attractors because they strongly influence human mobility over the whole city. Global attractors always offer some unique *services* that makes them distinguished from other regions where visitors only find such services in those regions. Significant places in the city like the airport, major universities, and hospitals, that occupy TAZes on their own and are easy to identify from the map. Fig.5 A shows annotation of these major places in the city.

Downtown Attractors

The second type of attractors is the downtown attractors, shown Fig.7 B. They contain places that are mostly clustered in the central business district. These are TAZes that have relatively high inflow. However, because of their central location in the city, visitors from all over the city have short routes to access these places. They have smaller average distances compared to the other two types as shown in Fig.5 C. For the same reason, the dispersion of the origins from the center of mass of inflows is also small as shown in Fig.5 C. The significant feature of these places is that they attract great number of visitors and are accessible.

Residential Attractors

Residential attractors shown in Fig.5 C attract the smallest number of visitors. As they are located in the outer sides of the city, the visitors to these places travel long distance on average to reach them as shown in Fig.7 C. For the same reason, the dispersion of the small number of visitors is larger than the downtown attractors as shown in Fig.7 B.

Attractors and POIs

We aim to further uncover what may cause the trips to the different types of attractors discovered. Thus, we relate the classified TAZes to the composition of services offered in these TAZes. We used official Points of Interests (POIs) data that contains all places in Riyadh city, around 12,000, offered by Arriyadh Development Authority (ADA), which

is the official entity managing all urban planning tasks in the city. POIs are classified into 23 subtypes of services offered in the city such as restaurants, schools, hospitals ... etc. In the following section, We explain the methodology by which we quantify the relationship between different types of POIs and attractors types .

Statistical significance testing

We aim to relate each type of POIs to the different attractors types to explore what makes different places have different attraction profiles. Using only counts of POIs per attractor type is not enough, because the distribution of types of POIs is unbalance. For example, restaurants are very frequent and spatially distributed, whereas universities are much fewer and are present in few districts. Thus, to identify which type of POIs are significantly concentrated in each type of attractors, we use a statistical significance testing approach. The statistical significance testing method measures the probability of observing the amount of a POI type in a spatial zone. It factors in the amount of all POIs in the spatial zone in addition to the amount of POIs with that type tag in the whole city.

We used Fisher's Exact Test (FET) to relate each type of POIs to the different attractor types. We selected FET because it works for small observations and calculates the exact probabilities rather than approximations such as in Chi square test. FET aims to test the dependency between two categorical variables given the observed data. It takes a contingency table as an input, which represent the relationship between two categorical variables in terms of their frequency distribution and their overlap. In our context, the first variable $POI_{Type=t}$ is the number of POIs that belong to a specific type t , and the second variable $POI_{Zone=z}$ is the number of POIs in a spatial zone z . We aim to test the significance of the overlap $POI_{Zone=z} \& POI_{Type=t}$,which is the amount of POI t in zone z , between these two variables. Fig.8 shows the contingency table and the corresponding Venn diagram representation. In Fig. 8, a represents the overlap, which is the amount of POIs of type t that are located in zone z , b represents the amount of POIs of type t that are not located in zone z , c is the amount of other POI types located in zone z , and d represents the number of all the rest of POIs in the city.

The objective is to test whether the amount of POI type t is significantly concentrated in a spatial zone z . FET quantifies the probability (p-value) of observing that amount of POI type t or larger in zone z by chance. The smallest the p-value is, the strongest the concentration of type t in zone z is. FET calculates the p-value p as follows:

$$p = \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{a!b!c!d!n!} \quad (7)$$

Where n is the total amount of all POIs in the city. FET incorporates two crucial factors when measuring the significance. First, the significance depends on how many other POI types in zone z as a measure of purity. If a zone has a large amount of POIs of type t ,but it also has a lot of other POIs types, that makes POI of type t less significance due to this impurity in the decomposition of all POIs in that zone. Second, the significance depends on the amount of POI of type t that are not in the tested zone z as a measure of rarity. If there are large number of POIs of type t elsewhere, that makes type t insignificance in the tested zone. These two features makes FET superior to trivial methods like calculating the percentages of POIs types in spatial zones.

We find that the proposed method can capture how each type of attractor has a different composition of POI types as shown in Fig. 9. We discuss the attractors types and the significant POIs related to them in the following sections.

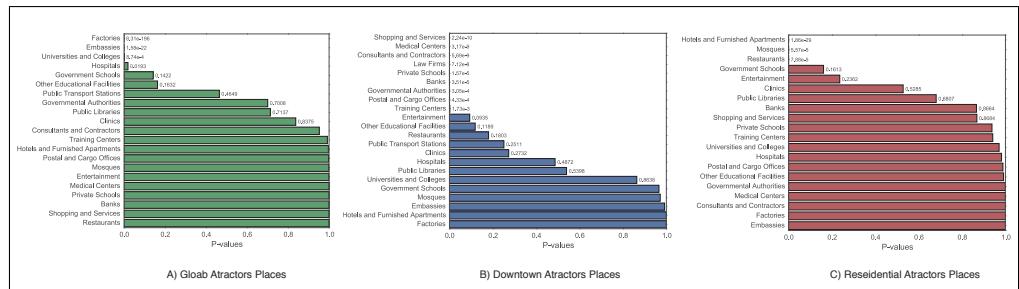


Fig 9. The types of POIs ordered by their statistical significance in the corresponding attractor's type. (A) The p-values of finding each of the POI types in TAZes classified as global attractors, where POIs types are ranked by their significance showing that factories, embassies, universities, and hospitals are type of services significantly located in global attractors. (B) POI types significantly located in the downtown attractors, which mostly includes business types of services. (C) POI types significantly present in the residential attractors that includes services and amenities such as furnished apartments, mosques (places of worship for Muslims), and public schools.

Global Attractors

Fig. 9 A shows each type of POIs ordered by their p-values, which represent how much they are concentrated in global attractors. Factories, embassies, universities, and hospitals are the top POI types that attract massive amounts of people coming from all over the city. The city has three major universities attracting a student body of 40k each contributing significantly to the observed global attraction. The city of Riyadh also has a major industrial area in the south to where a major number of factory workers commute. In addition, Riyadh city has the diplomatic quarter district, which hosts over 50 embassies attracting workers and visitors from all over the city to process documents. Hospitals are also expected to cause global attraction as visitors come from all over the city.

Downtown Attractors

Fig.9 B shows the types of POIs mostly located in the downtown attractors in a descendant order by significance. The first observation is that we witness a large number of POI types with strong significance (p-values <0.01) compared to the other two types of attractors. That is due to the richness in quantity and variation of types of POIs in this area. Thus, we expect larger number of POI types to be significant in that area. The common theme of significant POI types in this attractor is businesses, typical in downtown regions.

Residential Attractors

Fig.9 C shows the types of POIs concentrated in the residential attractors in a descendant order. Most significant types of POIs are services needed in residential areas like furnished apartments, mosques (worship places for Muslims), restaurants, including small restaurants and fast food places, and public schools. These types of places are not unique, so each residential neighborhood has its own share of these places to serve the population living nearby.

Conclusions and future directions

We present a novel computational framework to discover different attraction patterns in cities. We proposed 3 dimensions to define attraction of urban zones: total number of incoming trips, the spatial dispersion of the origins of trips, and the distribution of distances traveled by visitors to reach that district. Further, we present a method for understanding the relation between the decomposition of the types of POIs in a spatial zone and its attraction behavior. We applied the method and discuss the results in the city of Riyadh, the capital of Saudi Arabia.

The results of implementing the discussed modules mine data from mobile phones to provide a coherent understanding of the dynamics of the interaction between the flows of people to a district and types of services (POIs) that are located in that district. We detect three attraction patterns in the city of Riyadh according to the morning mobility dynamics. Global attractors, receive large share of the visitors traveling longer distances and coming from all over the city. These attractors have places of interest that are the destination of large student bodies, factory workers, hospital associates, and embassies. The second type of attractor is that of the downtown area, which receives high inflow of people from smaller distances and spatial dispersion due to its central location in the city that makes them accessible. The most significant POIs types located in the downtown attractors are business based places like firms, shopping and service places. The least significant trip attraction is to the residential areas in the morning hours, where the amount of inflow is the lowest. Residential attractors contain common POIs that serve inhabitants such as apartments, mosques, and schools.

Several interesting directions can follow this work. One is developing a predictive model of inflows and trip distributions, taking into account set of POIs, distance between origin and destinations and their population densities. Another possible direction is to compare these patterns among various cities to learn about best urban plans that reduce traffic.

Acknowledgments

The research was supported in part by grants from the Center for Complex Engineering Systems at KACST and MIT, and the U.S. Department of Transportation's University Transportation Centers Program.

References

1. Authority AD. King Abdulaziz Project for Riyadh Public Transport; 2016. Accessed: 2016-03-30. http://www.ada.gov.sa/ADA_e/DocumentShow_e/?url=/res/ADA/En/Projects/RiyadhMetro/index.html.
2. Louail T, Lenormand M, Picornell M, Cantú OG, Herranz R, Frias-Martinez E, et al. Uncovering the spatial structure of mobility networks. *Nature Communications*. 2015;6.
3. Karamshuk D, Noulas A, Scellato S, Nicosia V, Mascolo C. Geo-spotting: mining online location-based services for optimal retail store placement. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2013. p. 793–801.
4. Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2014;5(3):38.

5. Blondel VD, Decuyper A, Krings G. A survey of results on mobile phone datasets analysis. *EPJ Data Science.* 2015;4(1):1–55.
6. Yuan NJ, Zheng Y, Xie X, Wang Y, Zheng K, Xiong H. Discovering urban functional zones using latent activity trajectories. *Knowledge and Data Engineering, IEEE Transactions on.* 2015;27(3):712–725.
7. Pan G, Qi G, Wu Z, Zhang D, Li S. Land-use classification using taxi GPS traces. *Intelligent Transportation Systems, IEEE Transactions on.* 2013;14(1):113–123.
8. Toole JL, Ulm M, González MC, Bauer D. Inferring land use from mobile phone activity. In: *Proceedings of the ACM SIGKDD international workshop on urban computing.* ACM; 2012. p. 1–8.
9. Lee M, Holme P. Relating land use and human intra-city mobility. *PloS one.* 2015;10(10):e0140152.
10. Zheng Y, Zhang L, Xie X, Ma WY. Mining interesting locations and travel sequences from GPS trajectories. In: *Proceedings of the 18th international conference on World wide web.* ACM; 2009. p. 791–800.
11. Fan Z, Song X, Shibasaki R. Cityspectrum: A non-negative tensor factorization approach. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM; 2014. p. 213–223.
12. Liu X, Gong L, Gong Y, Liu Y. Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography.* 2015;43:78–90.
13. Zhan X, Ukkusuri SV, Zhu F. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics.* 2014;14(3-4):647–667.
14. Long X, Jin L, Joshi J. Exploring trajectory-driven local geographic topics in foursquare. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing.* ACM; 2012. p. 927–934.
15. Bassolas A, Lenormand M, Tugores A, Gonçalves B, Ramasco JJ. Touristic site attractiveness seen through Twitter. *EPJ Data Science.* 2016;5(1):1–9.
16. Han H, Yu X, Long Y. Discovering functional zones using bus smart card data and points of interest in Beijing. *arXiv preprint arXiv:150303131.* 2015;
17. Zhong C, Schläpfer M, Arisona SM, Batty M, Ratti C, Schmitt G. Revealing centrality in the spatial structure of cities from human activity patterns. *Urban Studies.* 2015;p. 0042098015601599.
18. De Nadai M, Staiano J, Larcher R, Sebe N, Quercia D, Lepri B. The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective. *arXiv preprint arXiv:160304012.* 2016;
19. Fan Z, Song X, Shibasaki R, Adachi R. CityMomentum: an online approach for crowd behavior prediction at a citywide level. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM; 2015. p. 559–569.
20. Wu L, Leung H, Jiang H, Zheng H, Ma L. Incorporating Human Movement Behavior into the Analysis of Spatially Distributed Infrastructure. *PloS one.* 2016;11(1).

21. Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2012. p. 186–194.
22. Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, González MC. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*. 2015;58:162–177.
23. Toole JL, Colak S, Alhasoun F, Evsukoff A, Gonzalez MC. The path most travelled: mining road usage patterns from massive call data. arXiv preprint arXiv:14030636. 2014;.
24. Mitchell A. The ESRI guide to GIS analysis, Volume 2: Spatial Measurements and Statistics. Redlands. CA: Esri Press; 2005.
25. Dijkstra EW. A note on two problems in connexion with graphs. *Numerische mathematik*. 1959;1(1):269–271.

Location-Based Services: A Road Towards Situation Awareness

Renato Filjar[†], Gordan Jezic*, Maja Matijasevic*

([†]Ericsson Nikola Tesla, Zagreb; *University of Zagreb)

(Email: renato.filjar@ericsson.com)

With the widespread use of mobile devices and increased demand for mobile services, Location-Based Services (LBS) represent a promising addition to service offerings of network operators as well as third-party service providers. Based on long-term research in LBS, our group has proposed a generic Enhanced LBS Reference Model (ELRM), which describes the concept, the architecture and the functionalities of the LBS. In addition, an evolutionary information process has been identified within the LBS, that represents knowledge maturity from position awareness to situation awareness. Both the ELRM and the information evolution process in LBS are presented in this article and illustrated by a case study within the framework of the 3GPP-standardised IP Multimedia Subsystem (IMS). This case-study emphasises the opportunities for navigation- and LBS-related solutions development provided by modern telecommunication technologies.

KEY WORDS

1. LBS model. 2. Location landscape. 3. Situation awareness. 4. Telecommunication network.

1. INTRODUCTION. The modern definition of location-based services (LBS) presents them as a group of emerging telecommunication services that successfully and purposefully merge ubiquitous position determination, mobile data communications and position-related content, with a defined level of quality of service (QoS). A considerable development has been made in recent years with the aim to define the general LBS reference model that would be able to present all entities required for implementation of a location-based service, and their interrelations. Our group has developed the unique LBS reference model which successfully presents the process of obtaining the best positioning estimate and communicates it with the LBS applications (Filjar *et al.*, 2004). However, in search of a more general reference model of LBS as an information and telecommunication service, a considerable enhancement of this model was to be developed. As a result, the Enhanced LBS Reference Model (ELRM) has been identified (Filjar, Bušić, 2007). Further evaluation of the ELRM has led to identification of an evolutionary process through which the “raw” *position* estimate matures to *location*. This evolution paves the road towards practical implementation of *situation awareness*.

This paper addresses improved ELRM concept, architecture, and functionalities, and describes location information evolutionary process. Practical implementation

that can provide situation awareness is discussed in the case study utilising the Internet Protocol (IP) Multimedia Subsystem (IMS), a key element of the next-generation converged telecommunication network architecture.

2. TERMINOLOGY. Before dissemination of the proposed ELMR, the authors have found it of utmost importance to distinguish the meanings of two terms frequently interchangeably used in association with the LBS: *position* and *location*. In view of later discussion, a definition of both terms will be given here.

As used in this article, a *position* is assumed to be a quantitative representation of the place of the physical object in a given spatial co-ordinate system. One of the most common means of expressing position, at least in terms of navigation, is through latitude, longitude and height above the sea level, with, for instance, the WGS84 system chosen as a reference. A *location* is to be assumed a position enriched with additional *information* elements referring to the relationships between the physical object, e.g. a building, (represented with its position) and the other objects in its surrounding area. In practical deployment and as an example, location of a physical object can be expressed by an address or the name of the building and with relation to other navigation- and orientation-relevant objects in the neighbourhood (for instance, building A, entrance B, to the north from the car park C). Evidently, the position of the physical object is embedded in location, and can be extracted from it by applying the appropriate procedures.

In view of the above-mentioned definitions, the process of position and location determination can be defined, as follows. Position determination simply means the process of determination of the co-ordinates of the physical place in a given reference system, where a certain physical object resides at the moment. For example, systems like Global Positioning System (GPS), Galileo, Glonass, Global Navigation Satellite System (GNSS), (Filjar, 2003), VHF Omni-directional Radio-range (VOR), Instrument Landing System (ILS) represent various positioning systems. On the other hand, location determination means the process of enhancement or enrichment of the position with the information about the object itself as well as the spatial relationship with the other objects residing in the neighbourhood. Every location-based service comprises the evolution from position to location, as it will be presented in the rest of this article.

So far, the LBS have been considered a mechanism for provision of location awareness, which can be defined as the ability to identify the static environment around the position given and spatial relations between nearby physical objects (buildings, roads, rivers, coastline etc.). Here a much broader concept is advocated which calls for the provision of situation awareness. The term refers not only to the ability to identify the static spatial relationship with surrounding physical objects, but also to provide near-real time information about events, temporal status of nearby objects and their expected or announced actions – all of that in order to allow decision making for a user's goals and objectives, both at present and in the near future. Observed in this context, situation awareness is a feat related to the user of a certain location-based service, rather than an issue related to a telecommunication network providing the location-based service in question. (Endsley, 2000) has defined situation awareness as:

“...the perception of elements in the environment within the volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.”

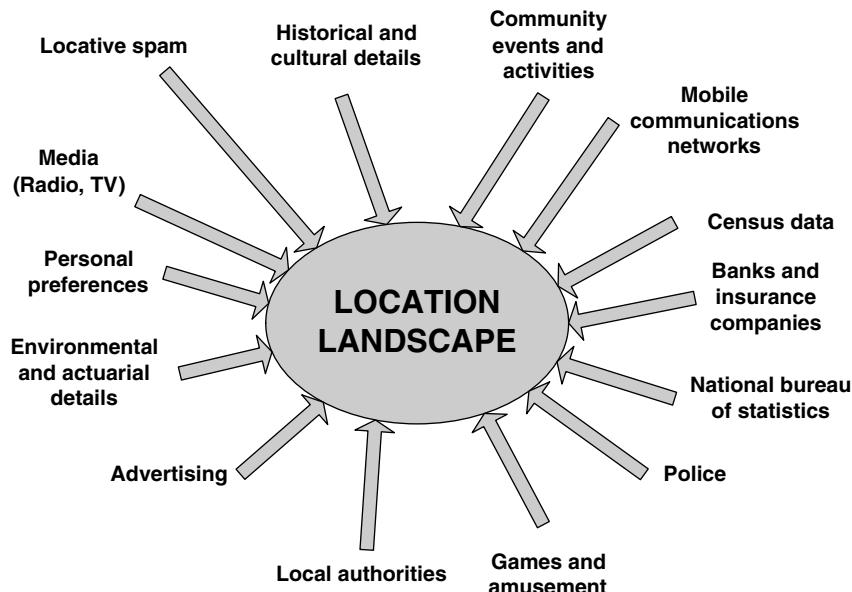


Figure 1. Various location landscape thematic layers.

Although the term situation awareness usually refers to military-related actions, its meaning is much broader, as seen from the definition above. This justifies the utilisation of the term in description of location-based services.

3. LOCATION LANDSCAPE CONCEPT. As the information services managing location-related content, LBS can be considered a part of the geoscope concept (Erle *et al*, 2005). So far, the living environment of every human has been considered in terms of the physical landscape. Following this line of development, LBS were concentrated on the *nearest-physical-object* types of solutions, greatly limiting the actual space of opportunities. The geoscope concept extends the LBS focus to the concept of the *nearest location-related information*. With this extension, the information space around the service initiators becomes multi-dimensional. Instead of focusing on the nearest physical objects, LBS starts to consider a wide number of layers in information space. A possible set of examples of these layers is presented in Figure 1.

The groups of location landscape layers contain location-related data in various forms. Thus, the location landscape can consist of database information, multimedia streaming content, and even location-related IT services. Census data, information about community events and activities, historical and cultural details, environmental and actuarial details, data from the national office of statistics and advertisements can be held in distributed databases and accessed using the appropriate searching algorithms. Multimedia content (photos, audio and video) can be geo-tagged, and thus made available for the content search using the appropriate tools and methods. Banks and insurance companies, as well as telecommunication operators and tourist agencies can offer location-related services designed according the needs and offers of the local

environment. Not to be neglected, location-related games and entertainment can be arranged in order to suit local tradition, physical landscape, or combined with some other groups of location landscape layers in order to provide the added value service. Finally, whilst it may not be welcome from the recipient, it would be possible to send location-related spam and advertisements in order to target visitors at certain premises.

4. LOCATION LANDSCAPE AND LBS. Extension to the *nearest location-related information*-type of service creates a multi-layered information space (called the *location landscape*, Erle *et al.*, 2005) that is to be explored according to the requirements of a particular LBS service. In that space, layers present the subsets of the location landscape that are to be deployed according to requirements of the LBS service. Location landscape consists of location-related content. The modern approach in classification of location-related content sees it in a 4-dimensional information space, with three dimensions related to spatial coordinates of position and the fourth related to time (age of data). The content may be considered as either static or dynamic (Tookey, 2007). Static location-related content is related to position only, and invariable to time (at least for a reasonable amount of time). On the other hand, dynamic location-related content is characterised by both spatial and time variability. Typical examples of static and dynamic content are the spatial distribution of buildings in a village and weather conditions in the observed area, respectively. Even a small community can provide a rich location landscape, as shown in Figure 2.

The formation of location landscape is essential for deployment of the geoscope concept. It is not, however, essential to form a centralised entity (a database, for instance) containing the whole location landscape information (Waller *et al.*, 2007). Rather, it is much more feasible to maintain an index of addresses where the most accurate and relevant location landscape layers are stored (list of related Internet sites, for example). The introduction of the geoscope concept brings an inevitable impact on the telecommunication network that supports the LBS. Apparently positive, this impact calls for a better integration of different telecommunication networks that a mobile user can use. Furthermore, the location landscape-based location services interweave a number of information systems that store location landscape layers (Waller *et al.*, 2007, Pottebaum and Torchia, 2006).

As an example of a targeted advanced LBS service, an investor could ask for all relevant information that can be applied in order to present a clear picture of where and when to invest money in property acquisition for a selected local geographical area. This could ask for a several location landscape layers to be deployed on the basic physical layers, for instance:

- spatial distribution of current prices for properties
- spatial distribution of household income
- spatial distribution of communal infrastructure (water supply, electricity lines, gas pipelines etc.).

On the other hand, an exemplary LBS service for local residents and tourists might collect the following location landscape layers in order to provide an information service:

- local road traffic conditions
- list of local community events
- tourist information (accommodation, restaurants, attractions).



Figure 2. An example of location landscape created around the small community of Baška, Krk Island, Croatia.

5. ENHANCED LBS REFERENCE MODEL. An initial LBS reference model, introduced by our team earlier (Filjar *et al.*, 2004), defined the process of managing position-related information. Focused on the generic positioning process based on two lower layers (Basic and Advanced Positioning), the LBS reference model left location context management to the upper layer (Application Layer). Considering the LBS as a synergy between ubiquitous positioning, mobile communications and location landscape (location-related content), the information traffic and location-related content management have been inevitably brought into the focus of the Enhanced LBS Reference Model (ELRM) (Filjar, Bušić, 2007). After the identification of an evolutionary information process within the LBS, the ELRM is improved in order to cover the whole situation awareness concept. The ELRM is briefly presented in this section.

5.1. ELRM description. The ELRM is established as a four-layer model as presented in Figure 3:

- Basic Positioning Layer (BPL)
- Advanced Positioning Layer (APL)
- Location Landscape Aggregation Layer (LLAL)
- Application Layer (AL)

The Basic Positioning Layer (BPL) consists of basic positioning methods. The BPL presents the raw positioning measurements related to the current user position, based

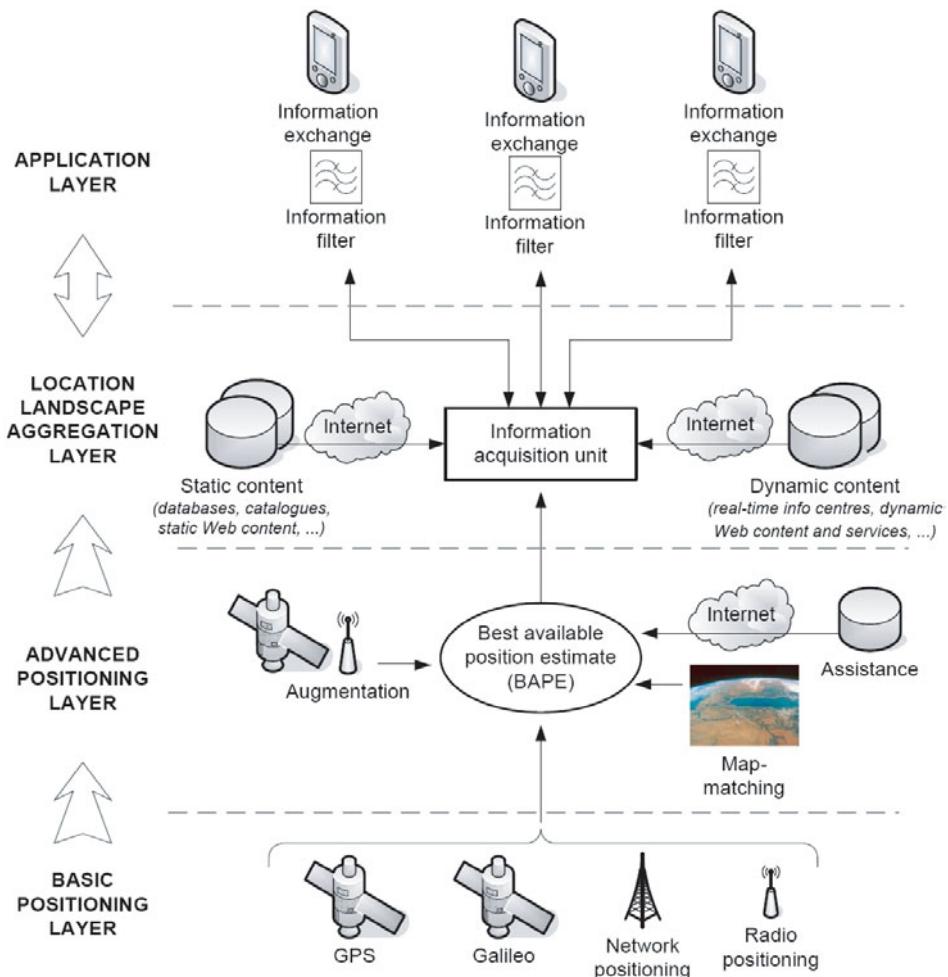


Figure 3. Enhanced LBS Reference Model (ELRM).

on standard (non-enhanced) positioning methods (satellite positioning, network positioning, radio positioning techniques etc.). For instance, a non-augmented GPS receiver that provides either pseudorange measurements or initial position estimate should be considered a constituent of the BPL.

The Advanced Positioning Layer (APL) consists of various methods for positioning assistance and augmentation to basic positioning. The APL yields the best available position estimate (BAPE) of the rover (mobile user), based on the raw positioning measurements, the positioning enhancements (positioning assistance, positioning augmentation etc.) and the appropriate positioning sensor fusion algorithm (Kalman filter, particle filter, neural network etc.). In this context, the BAPE should be considered the most likely place where a mobile subject (rover) should be found at the moment of location-based service initiation.

The Location Landscape Aggregation Layer (LLAL) creates a location landscape around the obtained BAPE, using the available location-related information from

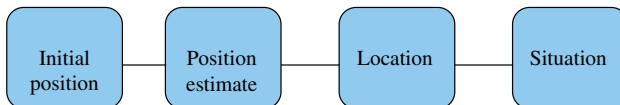


Figure 4. Information evolution in ELRM.

trusted and reliable sources, thus applying the previously mentioned geoscope concept. The location-related data may be obtained from, for instance, trusted and reliable Internet resources, such as GIS and statistical data offered by governmental agencies. Additionally, both internal and external sources are used for collecting information related to the dimension of time for every object, event, or action, when available. Location landscape is established in such a way that the sources of location-related information in both spatial- and time-dimensions are identified and the location-related content is made ready for use by the next ELRM layer.

The Application Layer (AL) selects the subset of location landscape content created at the previous layer and necessary for provision of requested location-based service, shapes it according to the specific user and service profiles requirements, and provides a framework for information exchange (information presentation, communication) with the service user. The AL applies the information filters as the entities capable of selecting the location information required for particular location-based service provision from a previously established location landscape around the BAPE. As a result, the LBS application will provide content and service with selected LBS information extracted from the location landscape created around the estimated position of the mobile user.

5.2. Information evolution in ELRM. An information evolution taking place within the Enhanced LBS Reference Model can be identified, as presented in Figure 4. Starting with the *initial position* estimate as the substance of every LBS, the process first enriches the initial position estimate with the augmentation, assistance and sensor fusion in order for the best available position estimate to be obtained. Then, the best available position estimate initiates the location landscape formation, as the necessary prerequisite for evolution from position to location, as defined in Section 2. In regard to the evolutionary path from initial position estimate to *location*, the latter should be considered a subset of the location landscape, which is determined by the user's (rover's) best available position estimate. In the final stage of LBS information evolution, *location* is enriched by dynamic location-related content, allowing for making decisions in regard of both present location and expected short-term activities of surrounding objects, thus forming a *situation*.

6. IMPACT OF THE ELM ON LBS DEVELOPMENT. The ELM and the LBS information evolution model together form the generic functional description of any location-based telecommunication service, as presented in Figure 5.

Components of the Basic and the Advanced Positioning Layers allow for performance of the advanced, best available position estimation. This is usually conducted within the access and core telecommunication networks. On the other side, location-related content management is a task usually performed by an application server. It creates a *location landscape* around the estimated position of mobile user, based on static and dynamic location-related content available for a place in question.

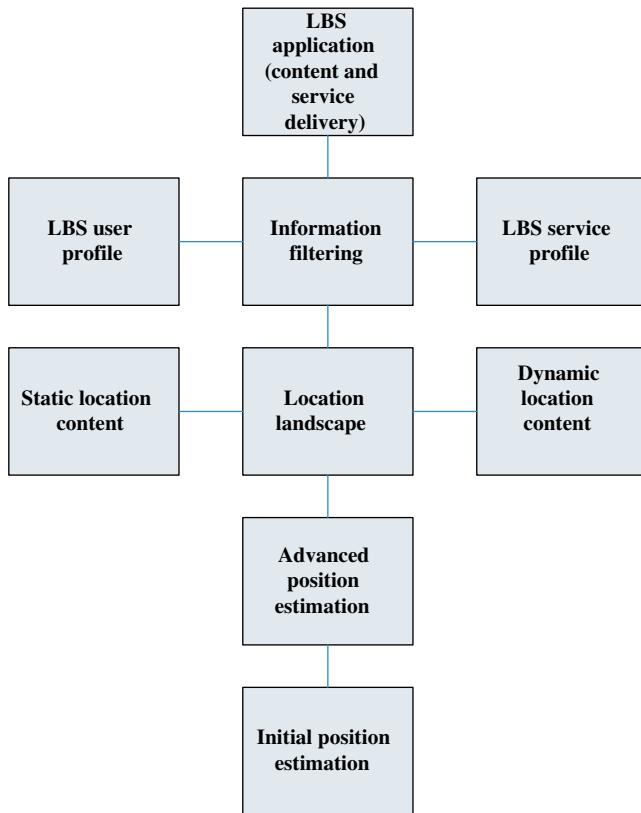


Figure 5. General functional description of a location based service.

Both static and dynamic location-related contents are provided by trusted and reliable sources, usually outside the telecommunication network in support of the invoked LBS service. Information filtering, needed for extraction of the sub-set of the location landscape related to the invoked LBS service is conducted based on the preferences of both the user and the LBS in question (expressed as user and service profiles, respectively). A selected sub-set of location landscape is then provided in the form of suitably converted content or service related to the position of the mobile user.

7. CASE STUDY. In this section, we present a possible implementation of the proposed model within the framework of the IP Multimedia Subsystem (IMS), the standardized next generation network architecture specified by the 3rd Generation Partnership Project (3GPP) and adopted by other standards organisations, such as Open Mobile Alliance (OMA), European Telecommunications Standards Institute TIPHON (Telecommunications and Internet Protocol Harmonisation over Networks) and SPAN (Services and Protocols for Advanced Networks) (ETSI TISPAN) and 3rd Generation Partnership Project (3GPP2) (Camarillo, 2004). The IMS architecture is usually depicted as having three layers, as shown in Figure 6: the Service Layer, the Control Layer, and the Connectivity Layer. The IMS uses the Session Initiation

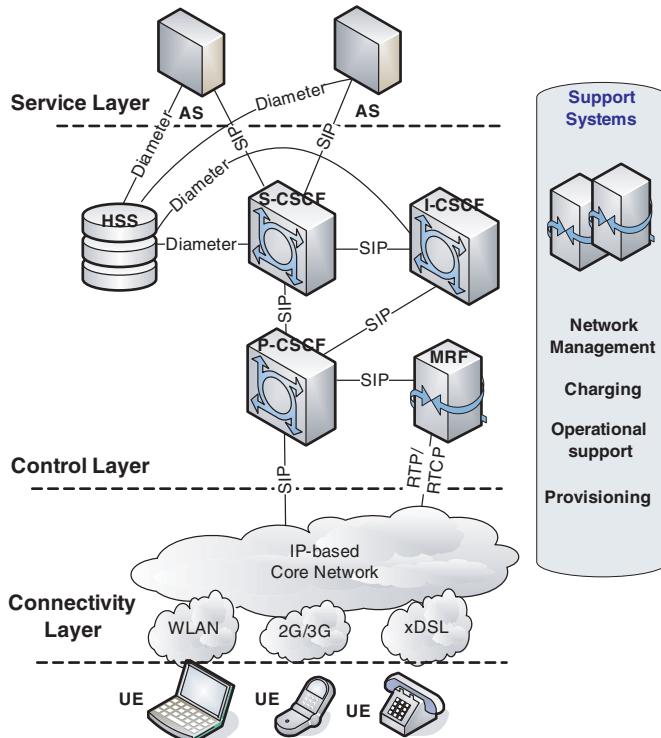


Figure 6. Simplified view of the IMS architecture.

Protocol (SIP), initially developed by the Internet Engineering Task Force (IETF), as the main signalling protocol.

The Service Layer comprises application and content servers which host and execute user services. The IMS allows for generic and common capabilities, implemented as services in SIP Application Servers (AS), to be reused as building blocks across multiple applications and services. A capability that may be utilised to provide a service to the end user, by itself or in conjunction with others, is called the *service enabler*. Some standardized enablers in IMS include presence, group list management, instant messaging, and call control. The Control Layer comprises databases and network control servers for managing call or session set-up, modification, and release. The key IMS entity in the Control Layer is the Call Session Control Function (CSCF). The CSCF is a SIP server responsible for session control and processing of signalling traffic. It controls a single session between itself and the User Equipment (UE). The CSCF plays three distinct “roles”: the Proxy Call Session Control Function (P-CSCF), Serving Call Session Control Function (S-CSCF), and the Interrogating Call Session Control Function (I-CSCF). For the purpose of this discussion, we limit our scope to the S-CSCF, which is responsible for session establishment, modification, and release. It is the function that registers the users and controls the process of services provisioning, although services themselves may reside on separate application servers. It performs routing of SIP requests, number/name translation, provides charging information to support systems, and maintains session timers. It

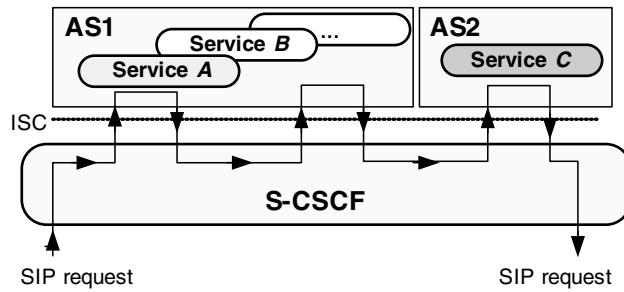


Figure 7. Composition of services in IMS.

also interrogates the main database containing user related information, the Home Subscriber Server (HSS), to retrieve authorisation, service triggering information, and user profile. The interfaces between the CSCFs and the HSS, as well as those between ASs and HSS are based on Diameter, the IETF's Authentication, Authorization and Accounting (AAA) framework for IP-based networks. The Connectivity Layer comprises routers and switches for the IP Core and access networks. Various access networks are envisioned, for example, 3GPP's General Packet Radio Service (GPRS) and Universal Mobile Telecommunications System (UMTS) Radio Access Network (RAN), 3GPP2's Code Division Multiple Access 2000 (CDMA2000) RAN, Wireless Local Area Networks (LANs), and various fixed Digital Subscriber Line (DSL) options. The User Equipment (UE), represents networked devices, such as personal computers (PCs), mobile phones, fixed phones etc., which connect to the network at this layer, thus allowing the IMS user to register and access various services offered by, or via, the IMS.

7.1. How the services are composed. Some benefits commonly associated with the adoption of an IMS-based infrastructure include flexible introduction of new services and service integration. Focusing on location-based services, we adopt the point of view that IMS should not only host services but also mediate and add value to 3rd party services (Gourraud, 2007). To achieve that, a service delivery environment is needed which allows the reuse of common enablers and resources. The important feature of IMS services is that they may be used as building blocks to build more complex services. The functional elements of IMS involved in service delivery are illustrated in Figure 7.

The S-CSCF is used for session control and orchestration, while the service logic is implemented within the ASs. Upon registration, the service profile linked to the user is downloaded from HSS to S-CSCF. Service profile includes service-triggering information in the form of prioritised Initial Filter Criteria (iFC). Each iFC contains information on a particular service which needs to be invoked when the particular triggering conditions (Service Point Triggers) are met. When a user issues a SIP request, the S-CSCF will route the request to the appropriate AS based on triggering information in service profile and invoke (zero or more) services, in sequence based on their priority order.

7.2. Adding a Location Enabler. Information about the user's position is a critical piece of information needed for any LBS. With the provisioning of that information considered as a generic and reusable network-provided enabling technology, in our previous work (Mosmondor, 2006) we presented the idea and prototype

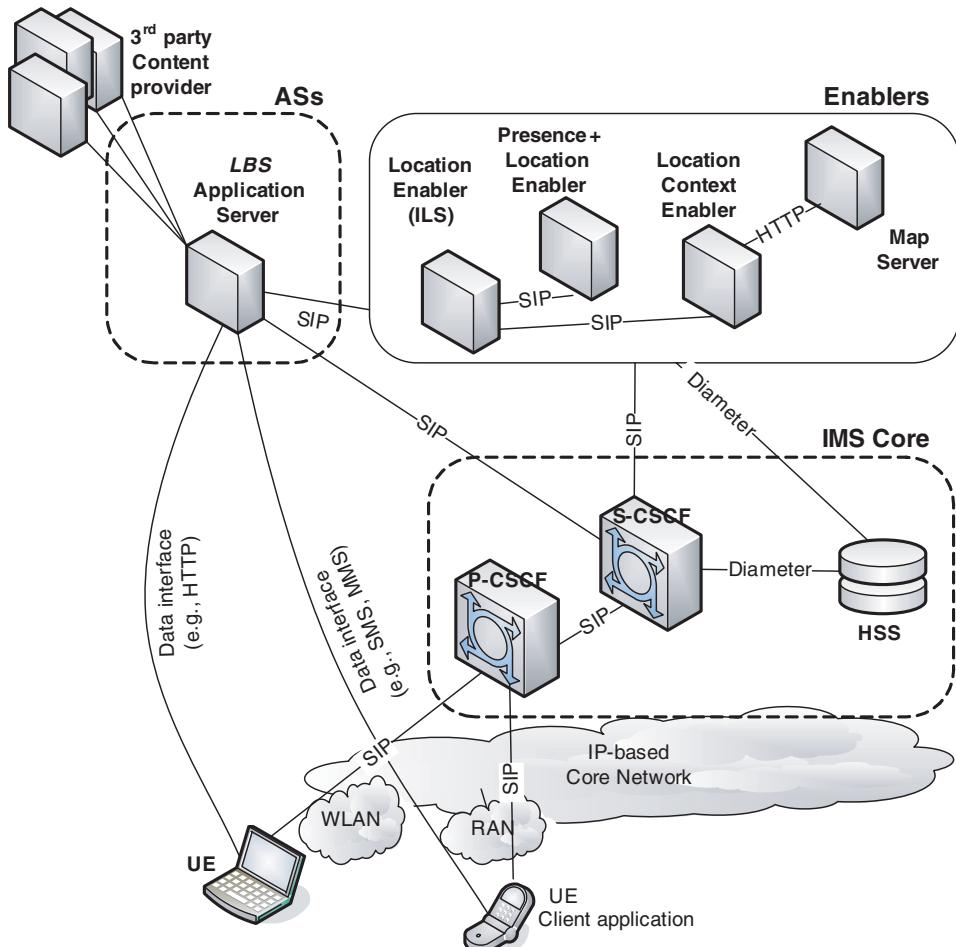


Figure 8. System architecture.

implementation of a service enabler called IMS Location Server (ILS). ILS is located in the IMS Service Layer. From the point of view of service logic in the IMS, the ILS acts as proxy towards various positioning systems, such as satellite and terrestrial radio positioning systems, as well as systems which utilize mobile communication network signals for position determination. It is important to note that, by providing a uniform (SIP) interface for retrieving the user position information, the ILS provides the means of making this information available to other IMS application servers, thus acting as a “location enabler”. With reference to the ELRM, the ILS comprises the functionality of the APL and the BPL. Its role in the process of LBS provisioning is to provide the BAPE of the user requesting the service to other IMS entities.

7.3. Other IMS Application Servers for LBS delivery. System architecture for LBS delivery in IMS is shown in Figure 8. The following components are involved:

- standard Presence Enabler
- Presence + Location Enabler

- Location Context Enabler
- Map Server
- LBS Application Server
- location-related content providers (3rd party content providers).

The Presence Enabler implements the presence service. In general, the presence service allows a user to be informed about the attainability, availability, and willingness to communicate for a given user on the network (e.g., “available”, “unavailable”, “offline”). When used as an enabler, presence can be invoked by an application that requires information on the status of a user.

The Presence + Location (PresLoc) Enabler is based on the standard Presence Enabler, enhanced with location information. The PresLoc Enabler uses the ILS for retrieving the user position data and sends this information to a subscribed IMS entity together with the requested presence status.

The Location Context Enabler (LCE) is an enabler which allows the user to define a list of personal landmarks. The LCE stores the list of landmarks and makes them available to the user when subscribing to a particular service. The user may also add new entries to the list, as well as modify and remove the existing locations. The location can be visualized on a map retrieved from the Map Server (via Hypertext Transfer Protocol, HTTP). The LCE may also store semantic interpretation of selected locations (e.g., main train station) in a generic presentation format, from which it could be reused for building more complex location-aware services.

The LBS AS has a SIP interface to the IMS Core and an (any standard) data interface towards external content providers. (For practical purposes, HTTP, or File Transfer Protocol, FTP, could be used.) Examples of location-related content providers include various contributors to the location landscape, as shown in Figure 2. The LBS AS performs matching, filtering, and adaptation of location-related content based on user preferences and terminal capabilities. For scalability reasons, the service provided by the LBS AS is based on the loosely coupled messaging model called *publish/subscribe*, shown in Figure 9. The “*publish/subscribe*” is an asynchronous messaging paradigm, in which the messages “published” by senders (“publishers”) are delivered only to those receivers (“subscribers”) who explicitly express their interest in a particular topic (subject) or content by subscribing to it. It is important to note that this model is multipoint-to-multipoint and anonymous – the publishers do not need to specify, or even know who the receivers are, and vice versa. Extensions to the basic model have been proposed to accommodate mobility (Podnar, 2004). Publish/subscribe systems may be classified as either topic-based (See Figure 10) or content-based (See Figure 11), depending on the subscriber’s ability to tailor the subscription to his or her interests. The topic-based subscription corresponds to a named logical channel, where channel names are fixed and assigned beforehand. Like in a common radio channel, a user subscribed to one or more topics receives all information published on any matching topic channel. The content-based subscription system, on the other hand, allows the subscriber to filter out the content of interest based on multiple filtering criteria and rules.

The LBS AS includes two components:

- Publish/Subscribe Component (PSC), and,
- Content Delivery Component (CDC).

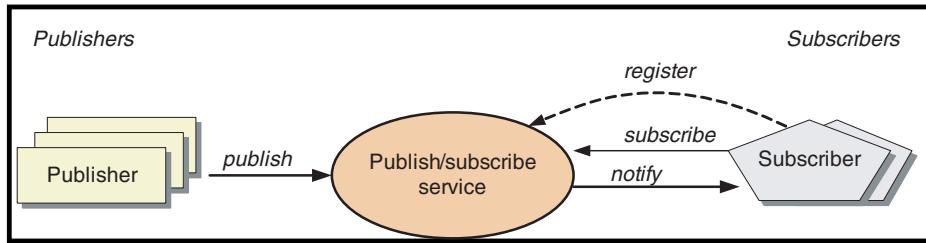


Figure 9. Publish-subscribe service.

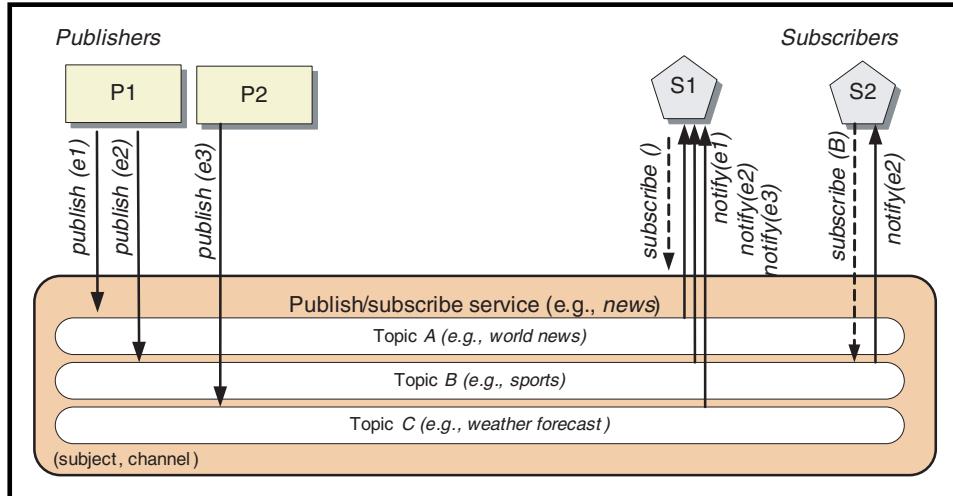


Figure 10. Topic-based subscription.

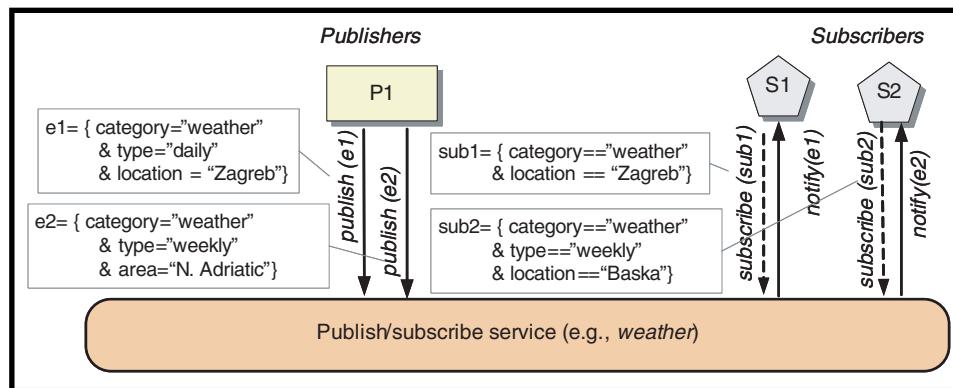


Figure 11. Content-based subscription.

The PSC receives and processes the user request for a content (e.g., service subscription, service request), while the CDC is responsible for adapting the content and delivering the result to the user. With reference to the ELRM, the LBS AS is located

in the AL and the PSC and CDC correspond to *Information filter* and *Information exchange* functionality, respectively. The PSC is implemented using the publish/subscribe mechanism extended to support location-based subscriptions and publications of location-related content (Devlic and Jezic, 2005). In the context of LBS, the “publishers” are information providers, who publish the content related to a specific geographical position (thus creating the location landscape), and the “subscribers” are registered IMS users, who declare interest in receiving the information referring to a given location. The goal of the PSC is to match publications with subscriptions according to the topic (or content), and the associated location.

Due to the heterogeneity of mobile terminals, it may also be necessary to adapt the form of the content to the requirements of the mobile terminal in use. The CDC performs content adaptation and alters the service behaviour according to the preferences explicitly expressed by the user and the capabilities of the terminal (e.g., amount of memory, processing capabilities, screen resolution, software characteristics, communication status). Some standardised formats for resources, terminal capabilities, and user preferences description include the Resource Description Framework (RDF) specified by the World Wide Web Consortium (W3C), User Agent Profile (UAProf), presence attributes specified by the Open Mobile Alliance (OMA), and PIDF-LO, specified by the IETF GEOPRIV working group for encoding location information and privacy policies. The adaptation and matching could also be performed so as to take into account not just the end-points, like the end-user and the application server, but the current situation in the telecommunication network as well. A possible way to achieve this could be by introducing a network Quality of Service (QoS) matching and optimization function, or a “QoS enabler” as has been envisioned in our previous work (Skorin-Kapov *et al.*, 2007).

7.4. LBS delivery. The delivery method is based on user profile, in which a user defines his/her location-based and non-location based subscription preferences. The preferences include, for example, device capabilities, preferred delivery method (SMS/MMS/e-mail), and list of topics of interest. Two types of LBS subscriptions are considered, one based on “current location” at any given moment, and the other, linked to a fixed “landmark” specified by the user. (It may be noted that the non-location based subscriptions are also possible, but are not considered of interest for the purpose of this discussion).

We first describe the LBS delivery process in case of a *current-location based subscription*. The user sends a SIP subscription request, which is routed by the S-CSCF (based on the IMS user profile retrieved from the HSS) to the PSC. The PSC sends the request to the ILS to locate the user and, once it receives the BAPE of the user, it creates a current-location-based subscription. As publishers publish the content over various topic “channels”, the PSC receives notifications based on its ongoing subscriptions. If the location/related information matches the current position to which the subscription is linked, the PSC will instruct the CDC to adapt the content to the subscriber’s terminal’s capabilities and deliver it via a preferred delivery method (SMS, MMS or e-mail). To take into account the user movements, the PSC can periodically query the ILS about the location, or, the ILS can periodically send location-push to the PSC, containing the new BAPE of the user. Due to scalability reasons, however, the frequency of such position updates must remain relatively low, say, refreshed no more than once every few minutes.

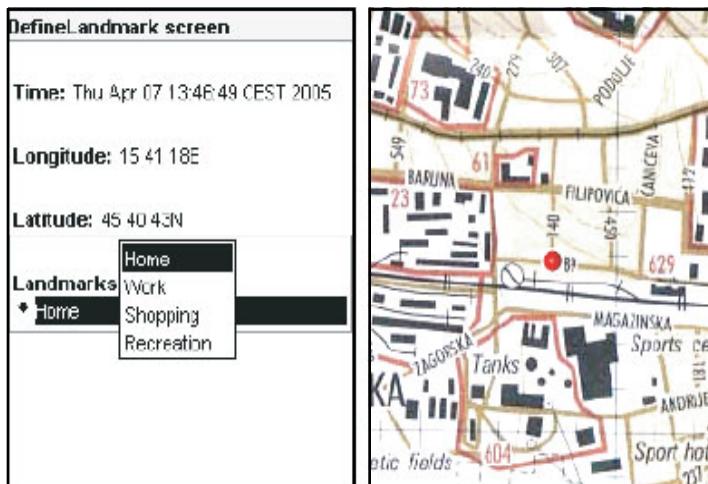


Figure 12. Landmark declaration.

In the case of a *landmark-based subscription*, it is assumed that the subscriber has a previously defined list of landmarks and that they are stored in the user profile and downloaded to Context Enabler during IMS registration. The PSC compares the location of the landmark specified in the subscription to that of the published location-related content. If the locations match, the CDC delivers the (adapted) content to the subscriber.

8. EXAMPLE. John, an IMS user, follows a routine on a typical working day: he leaves home to go to work in the morning, then in late afternoon on the way home he stops to eat or shop for groceries. During the day he either walks, or takes the tram, or drives the car to reach the target destination. John also owns a nice cottage in a fishing village at the coast, where, weather permitting, he likes to spend weekends. The location based service portal has a graphical user interface which displays an interactive city map, with some pre-defined areas (city's boroughs) and local landmarks. The location-related content offered includes various tourist information (pubs, restaurants, hotels), current weather, and current traffic conditions, provided by the national tourist association, the weather service, and the traffic information centre, respectively. In addition to predefined landmarks, John's personal landmarks include "Home", "Work", "Recreation", and "Cottage". Each landmark is assigned a position in the form of geographic coordinates and the landmarks can further be used for specifying landmark-based subscriptions (See Figure 12). John's user data, service data, IMS terminal capabilities, and various user preferences and service subscriptions are stored in the HSS, in his user profile. His preferred delivery method for all LBS is MMS (containing the map with info-elements overlay).

8.1. Scenario 1. John is driving to the office in the morning. Remembering that there is road construction work going on at one section along his usual route, he wants to know the traffic conditions, preferably before he reaches that point. By using his mobile phone, he subscribes to traffic status in the part of town of interest, by

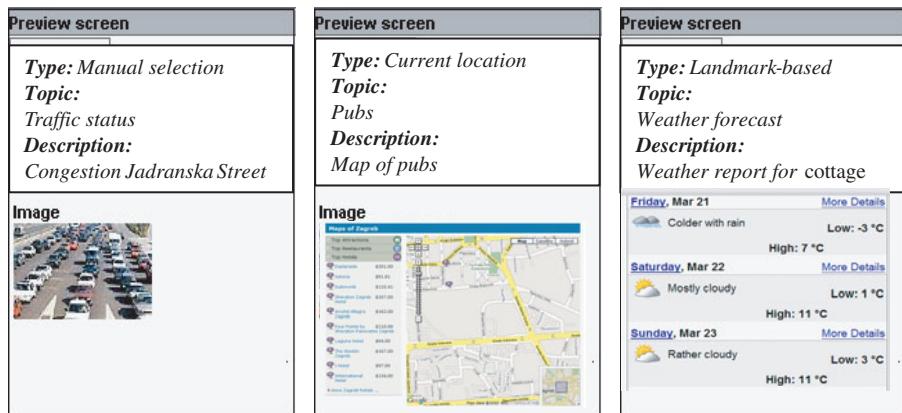


Figure 13. Information delivered. (Left) Scenario 1. Traffic congestion; (Centre) Scenario 2. Nearby Pubs; (Right) Scenario 3. Weather at cottage.

selecting one or more districts from a list. He also selects MMS as a preferred delivery method. The PSC is subscribed to the current traffic information provider and once the content is published, the PSC initiates the delivery of a traffic report to all subscribers with a matching subscription. In the course of delivery, the CDC receives the content and checks John's preferred delivery method from his profile, finds out it is MMS, and sends the traffic info as an MMS to John's mobile phone. The MMS content shows congested streets around the road construction site (Figure 13 (Left)), and John decides to take an alternative route to the office. (Note that the Figure 13 illustrations are not actual screenshots, they are simply constructed to indicate what the user interface might look like).

8.2. *Scenario 2.* John finishes his work for the day and decides to walk to one of his favourite nearby pubs, only to find it closed due to a private party. He decides to look for another pub "close to" his current location. He subscribes to the location-related content under the topic "pubs" within the tourist info section, by using a current-location-based subscription type. For this type of subscription, according to his preferences, "current location" means "within an approximately 500-metre radius from current position." The PSC retrieves John's location from the ILS. It also retrieves a list of pubs from the local tourist association, filters it based on distance from John's current location and initiates the delivery of the result via CDC. The CDC receives the content from PSC, reads John's preferences stored in his user profile and sends the map with pubs marked via MMS to John (Figure 13 (Centre)).

8.3. *Scenario 3.* The weekend is approaching and John considers spending it in his cottage at the coast. The critical issue, though, is the weather, since the heating system in the cottage is rather basic. John accesses the LBS weather service and subscribes to the location-related content (weather forecast) with a landmark-based subscription. John selects "Cottage" as a landmark at which he wishes to receive weather information, and submits his subscription to the PSC. Once the weather information is refreshed, the PSC compares the location with the specified landmark's position, and sends the weather report to the CDC to deliver it to John (Figure 13 (Right)).

9. CONCLUSION. Location-based services are very complex, involving various inputs and a special information evolution process. Only a generic and detailed description of location-related evolution process offers the identification of proper LBS functionalities and the architecture in support of LBS provisioning. This article provides the most generic description of location-related information evolution within the LBS, allowing for further research in identification of LBS functionalities and architecture components identification. Additionally, this article presents a case-study which utilises the IMS as one of the latest telecommunication technologies, with the aim of attracting the interest for utilisation of telecommunication networks in navigation-related solutions. Presented ELRM and findings related to LBS information evolution will be explored in further research in design and development of navigation systems and solutions that utilise telecommunication networks.

ACKNOWLEDGMENT

The authors acknowledge the support of research projects “New Architectures and Protocols in Converged Telecommunication Networks” (071-362027-2329) and “Content Delivery and Mobility of Users and Services in New Generation Networks” (036-0362027-1639) funded by the Ministry of Science, Education and Sports of the Republic of Croatia, and projects A-STORM and FAME of Ericsson Nikola Tesla, Croatia.

REFERENCES

- Beatty, C. (2002). Location-Based Services: Navigation for the Masses, At Last! *The Journal of Navigation*, **55**: 241–248.
- Camarillo, G., and M. Garcia-Martin (2004). The 3G IP Multimedia Subsystem: Merging the Internet and the Cellular Worlds. John Wiley & Sons.
- Devlic, A., G. Jezic (2005). Location-Dependent Information Services using User Profile Matching”. Proc. of ConTEL 2005 Conference, pp 327–335. Zagreb, Croatia.
- Erle, S., R. Gibson, J. Walsh. (2005). Mapping Hacks: Tips & Tools for Electronic Cartography. O'Reilly, Sebastopol, CA.
- Filjar, R., L. Basic. (2007). Enhanced LBS Reference Model. *Proc. of NAV07 Conference* (on CD), 8 pages. Westminster, London, UK.
- Filjar, R., S. Dešić, K. Tržec. (2004). LBS Reference Model. *Proc. of NAV04 Conference* (on CD), 8 pages. Westminster, London, UK.
- Filjar, R. (2003). Satellite Positioning as the Foundation of LBS Development. *Revista del Instituto de Navegación de España*, **19**: 4–20.
- Gourraud, C. (2007). Using IMS as a Service Framework. *IEEE Vehicular Technology Mag*, **2**(1): 4–11.
- Mosmondor, M., L. Skorin-Kapov, R. Filjar, M. Matijasevic. (2006). Conveying and Handling Location Information in the IP Multimedia Subsystem. *Journal of Communications Software and Systems*, **2**: 313–322.
- Podnar, I., I. Lovrek (2004). Supporting Mobility with Persistent Notifications in Publish/Subscribe Systems. *Proc. of the 26th International Conference on Software Engineering, 3rd Int. Workshop on Distributed Event-Based Systems*, pp 80–85. Edinburgh, Scotland, UK.
- Pottebaum, T., M. Torchia. (2006). Moving from Enterprise Location Data to Location Intelligence. *Directions Magazine*, (available at: <http://tinyurl.com/3xy4ml>, accessed on 23 January, 2008).
- Skorin-Kapov, S., M. Mosmondor, O. Dobrijevic and M. Matijasevic (2007), Application-Level QoS Negotiation and Signaling for Advanced Multimedia Services in the IMS. *IEEE Communications Magazine*, **45**(7): 108–116.
- Tooley, B. R. (2007). What Are the Drivers We See that Makes Us Believe Location Intelligence is Ripe for Exploitation in Our Business. *Proc. of NAV07 Conference* (on CD), 8 pages. Westminster, London, UK.
- Waller, A., J. Lewis, G. Jones, R. Craddock. (2007). Secure Situation Awareness using Web Based Mashups. *Proc. of NAV07 Conference* (on CD), 8 pages. Westminster, London, UK.

On the Analysis and Visualisation of Anonymised Call Detail Records

Varun J, Sunil H.S, Vasisht R
SJB Research Foundation
Uttarahalli Road, Kengeri,
Bangalore, India

Email: (varun.iyengari,hs.sunilrao,vasisht.raghavendra)@gmail.com

Srinidhi Saragur, Abhijit Lele
SJB Research Foundation
Uttarahalli Road, Kengeri,
Bangalore, India

Email: ssrinidhi@sjbrf.org, abhijit.mlele@gmail.com

Abstract— Global mobile traffic is estimated to grow at a compounded annual rate of 40%. In order to keep telecom operators profitable, network planning and optimisation, and personalized traffic plans are critical to success. In order to do this, telecom operators need to analyse the data generated from their telecom network and derive information and knowledge that will assist them in network planning and developing personalised traffic plans.

With this as the background, in this paper we analyse the *Call Detail Records (CDR)* provided by *Orange Telecom* as a part of the *Data for Development (D4D)* challenge. The data analysis of the CDR records is carried out using *Hadoop* and *Hive* framework. This paper focuses on analysing the data from telecom network optimisation and socio-economic perspectives. We utilise a visualisation tool to represent the derived information in a human understandable format. Based on visual inspection of the derived knowledge from the CDRs, we provide our recommendations for network optimisation.

I. INTRODUCTION

Telecommunications operating companies generate a large number of data records from switching systems. Data items are produced for every telephone call through a telephone network. *Call Detail Record (CDR)* is the fingerprint of how many seconds and at what time a customer is using a telephone and the associated infrastructure in terms of *Base Stations* used to process the call. Therefore, CDR displays a map of telephone customers behavior. The knowledge of customer behaviour obtained by analyzing CDRs has multiple applications. In [1] the authors deduce social attributes from calling behaviour. Mobile customer clustering based on CDR records from the perspective of marketing campaigns is discussed in [2]. In [3], the authors propose a method to derive transportation patterns based on CDR records. Considerable research is being carried out not only on socially relevant data, but also patterns that can be used to enhance the overall service parameters of telecom operations.

While methods that analyse CDR records and derive knowledge from it are important, it is equally important to visualise data and the knowledge derived from it in a human understandable format. This gives a quick reference point to the stakeholders in a manner that they can relate to. In this paper we share our experiences in visualization of the knowledge derived from CDR records and to some extent provide our observations on and interpretation of the derived knowledge.

The rest of the paper is organized as follows. Section II describes the CDR data set. The framework of *Hadoop* [4] and *Hive* [5] along with mapreduce [6] are discussed in Section III. Our approach to analysing the data sets to derive information from them and then visualising the data sets is discussed in Section IV. We briefly attempt to provide analysis of the knowledge derived from these data sets in Section V. We finally conclude with future work in Section VI.

II. CDR DATA

In this paper we use the Orange Telecom [7] *Data for Development D4D* data set. D4D is an open data challenge, encouraging research teams around the world to use four datasets of anonymous call patterns of Orange Telecom's Ivory Coast subsidiary, to help address society development questions in novel ways. The data sets are based on anonymized Call Detail Records extracted from Oranges customer base, covering the months of December 2011 to April 2012. Four datasets are provided by *Orange* in *Tabulation Separated Values (TSV)* plain text format.

- 1) **Antenna-to-Antenna:** Antennas are uniquely identified by an antenna identifier (A_{id}) and a geographic location. The data set aggregate hour by hour the duration of calls between any pair of antennas. The data set is represented by the following tuple (date_hour TIMESTAMP, originating_antenna INTEGER, terminating_antenna INTEGER, number_voice_calls INTEGER, duration_voice_calls INTEGER).
- 2) **Individual Trajectories High Spatial Resolution Data:** This dataset contains the trajectories of 50000 randomly sampled individuals for the entire observation period but with reduced spatial resolution. The data set is represented by the following tuple (antenna_identifier INTEGER, longitude FLOAT, latitude FLOAT), where antenna_identifier represents the antenna anonymized by an identifier and (user_identifier INTEGER, connection_datetime TIMESTAMP, antenna_identifier INTEGER).

3) Individual Trajectories Long Term Data:

In this data set, the trajectories of 50000 randomly selected individuals are provided for the entire observation period but with reduced spatial resolution. The data set is represented by the following tuple (subpref_identifier INTEGER, longitude FLOAT, latitude FLOAT), where subpref_identifier represents the subprefecture anonymized by an identifier and (user_identifier INTEGER, connection_datetime TIMESTAMP, subpref_identifier INTEGER).

- 4) **Communication Sub Graphs:** The dataset contains the communication sub graphs for 5000 randomly selected individuals identified by user_identifier. The data set is represented by the following tuple (Source_user_identifier INTEGER, destination_user_identifier INTEGER)

III. APPROACH TO CDR DATA ANALYSIS

As mentioned in section II, the CDR data is for a duration of 6 months with about twenty million records; hence, traditional methods of data analysis may not work. With this in mind we used the *Apache Hadoop* framework as a base line.

A. Framework for CDR Data Analysis

Apache Hadoop [4] is a framework for running applications on a large cluster built of commodity hardware. The Hadoop framework transparently provides applications for both reliability and data motion. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. In conjunction with *Hadoop*, we used *Hive* [5]. Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems. The overall framework of *Hadoop* in conjunction with *Hive* is shown in Figure 1. The major components of the framework are

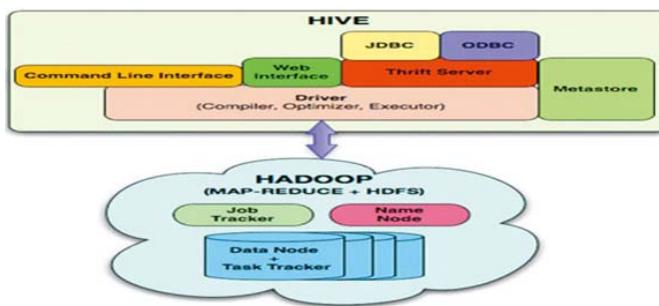


Fig. 1. Framework for CDR Data Analysis

the *Metastore* that stores all the structure information of the various tables and partitions in the warehouse and the *Execution Engine* that executes the execution plan created by the compiler. Considering the large data set *Mapreduce* technique is used to manage the complexity. For the sake of

brevity, details of *Mapreduce* technique are out of scope of this paper, but for the sake of completeness, the *Mapreduce* technique takes a set of data and converts it into another set of data, where individual elements are broken down into key-value pairs. Complex operations can then be performed on these distributed key-value pairs.

B. Methodology for Analysis and Visualization

The methodology used to analyse the data using *Hadoop* is shown in Figure 2. As a first step, the data is loaded to *Hadoop*.

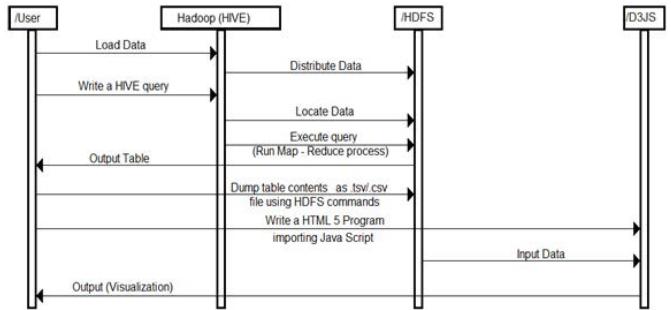


Fig. 2. Methodology used for Analysing Data

Hadoop distributes the data across *HDFS* in all the slave nodes. *Hive* query is used to fetch the required data. *Hive* compiles the query, locates data and executes the query. It returns the output data in a table. The table is stored in the local system in the *Hadoop* cluster. To visualise this stored data, *HTML 5* along with *java scripts* based on *D3JS* are used. These *java scripts* operate on the output of the query and provide the final visualisation.

IV. STATISTICAL ANALYSIS AND VISUALISATION

Based on the methodology discussed in Section III the CDR data was analysed and an appropriate visualisation provided. Since the focus of this paper is on visualization, we target four themes for data analysis.

A. Geographical Understanding

The CDR records are analysed to find the geographical location of antennas in terms of latitude and longitude. This information is correlated and overlaid with an actual *Geographical Information System* using *Openmaps* [8]. The resulting visualisation is given in Figure 3. As a second step, the sub-prefecture data was then overlayed on the knowledge acquired from antenna locations, resulting in an approximate zonal coverage of a set of antennas as shown in Figure 4.

B. Call Patterns

The most important parameter when monitoring call patterns is the *Call Density*. *Call Density* is defined as the number of calls originating from a *Base Station* within a cellular network. From a telecom operator perspective, the knowledge about the call density and specifically the correlation of the call density to the geography is an important call pattern. This call pattern helps the telecom operator to tune the cellular network

ANTENNAS IN IVORY COAST
Antenna ID 822

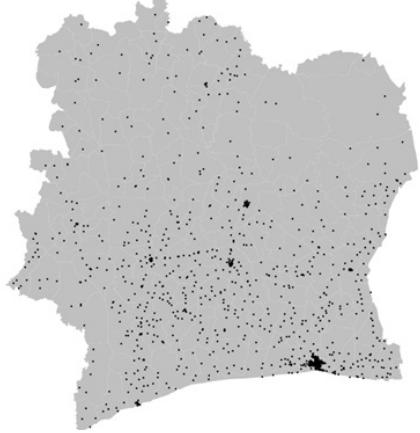


Fig. 3. Geographical Spread of Antennas

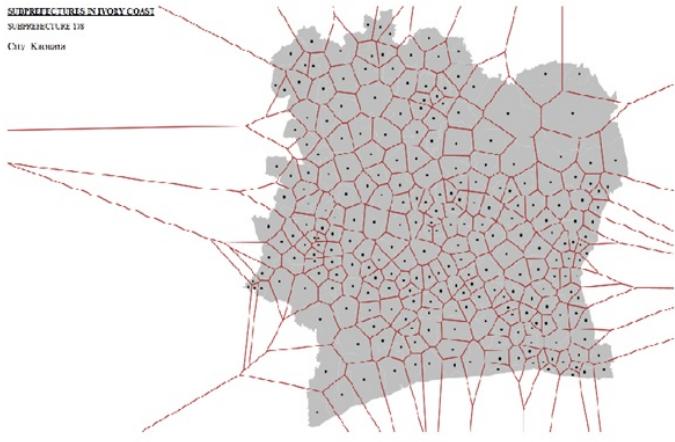


Fig. 4. Sub-prefecture based zone coverage

to minimise call drop rates and thus maximise revenue. Figure 5¹ and Figure 6 give a snap shot of the call density per sub-prefecture. The darker the shade higher the call density. Such a visualisation gives the telecom operator a quick mental picture of the call patterns. Such a visualisation when combined with actual data can assist the telecom operator to optimise the cellular network for optimal performance.

C. Market Analytics

Operating a cellular network is a CAPEX and OPEX intensive business. Telecom operators are continuously striving to optimise both the CAPEX / OPEX cost. One of the factors that influences the opex cost and determines the return on investment (profits) the telecom company can make is the *Active Air Time*. *Active Air Time* (AAT) is the aggregate time in which the air-waves are used. In other words AAT is the total number of calls made in a day, and monitored across all days / months of the years. In order to give the telecom operator

¹We thankfully acknowledge Paradigma Labs for the Visualisation concept and code.

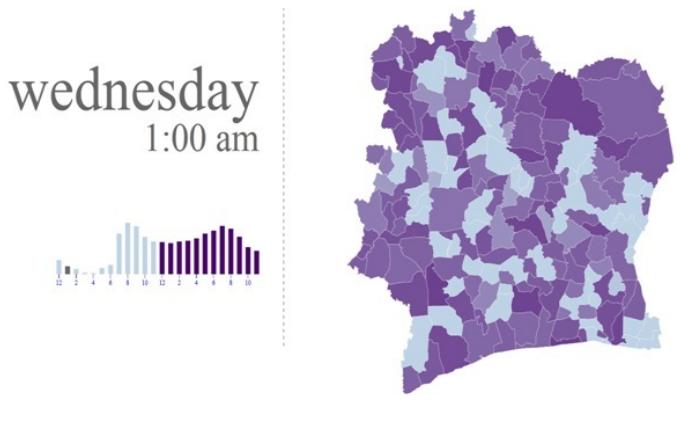


Fig. 5. Sub-prefecture based call density on Wed at 1:00 am

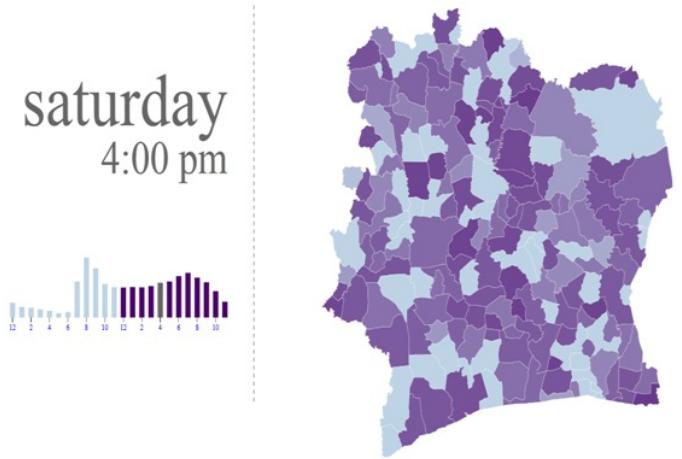


Fig. 6. Sub-prefecture based call density on Saturday at 4:00 pm

a pictorial view of this data, CDR records were analysed and daily aggregated call information plotted as shown in Figure 7 and Figure 8 respectively. While the value of visualizing the same data in two different views might be debatable, from a telecom operators' perspective it provides two different sets of information. Figure 7 gives a time series perspective of the data and assists the telecom operator to get a visual perception of the number of calls made across different days in a month. On the other hand Figure 8 gives the telecom operator an easy visual reference to compare the number of calls made on the same day across different months. These inputs are important to fine tune the market strategy of the telecom operator.

D. Sentiment Analysis

One important business objective that every telecom operator tries to achieve is *personalized services*. The personalisation can be in terms of customized tariff plans, personalised greetings and such. Sentiment analysis plays an important role in personalisation. While a lot more can be derived from CDR records from a sentiment analysis perspective, as a first

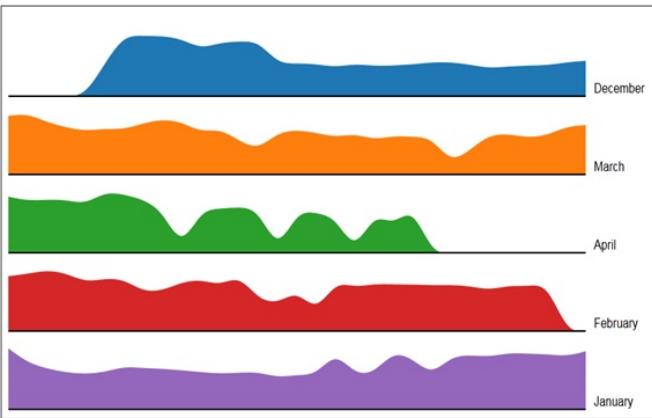


Fig. 7. Snapshot of Call Density variation across months as a Time series

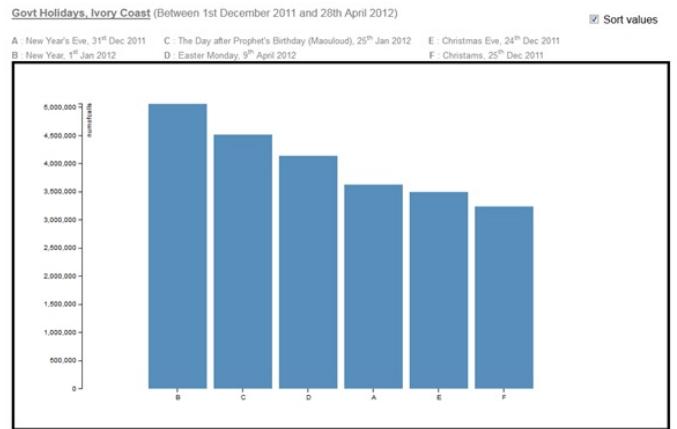


Fig. 9. Aggregated Calls around festive period

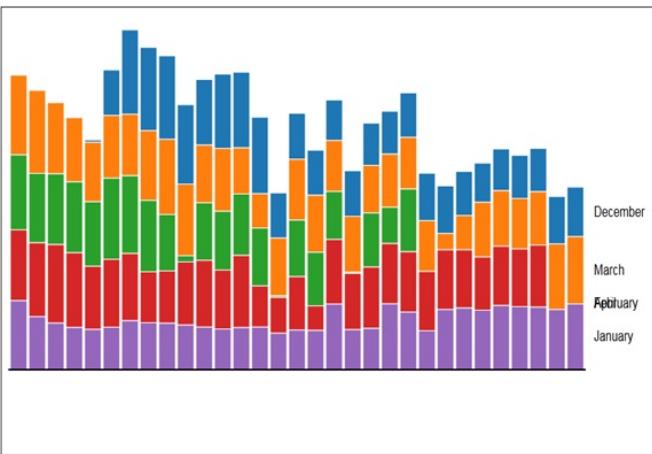


Fig. 8. Snapshot of Call Density variation across months as bar chart

step we have tried to analyse the call patterns around festive periods. Figure 9 shows a pictorial view of the number of calls around the festive period. Observe that on New Year's day the total number of calls is more than on New Year's eve. One interpretation of this data point is possibly indicative of a positive outlook of people of Ivory Coast. This interpretation will be discussed in detail in the following section when we correlate these data sets.

V. INTERPRETATION

Section IV discussed the visual representation of some of the information derived from the CDR. In this section we attempt to provide an interpretation to the data with possible applications of the derived knowledge. While more work needs to be done to have concrete recommendations based on the derived knowledge, this is the first step in that direction. While there can be numerous interpretations of the derived knowledge, we focus on two key aspects *viz.*, *Telecom Operator perspective* and *Socio-Economic perspective*.

A. A Telecom Operator Perspective

From Figure 3 observe that the antenna density is more towards the south east of *Ivory Coast*. Correlating this data with Figure 5 and Figure 6 one would expect a higher call density, but a combination of light and dark shades indicates that this is not the case. Our experiments with the CDRs show a consistent low call density in the region. It cannot be conclusively said whether this observation is true all year long since the CDRs are only for five months. However, based on the available data, it can be inferred that some amount of network planning to optimise antenna density to match call density is advisable².

Another interesting observation is the correlation between the antenna density on the north east of *Ivory Coast* (Figure 3) and the call density (Figure 6). While the antenna density is low, the call density is relatively high. This is likely to result in higher call drop rate, since there isn't enough network capacity available. Based on this inference, increasing the network capacity by increasing the number of antennas and in turn base stations is advisable.

The other observation is the correlation between the sub-prefectures and the antenna density from Figure 3 and Figure 4 respectively. Observe that the sub-prefectures are smaller towards the southern region of *Ivory Coast* and grow to be larger towards the northern region. However, the antenna density decreases as we travel north. While we are still in the process of deriving the call drop rate across different sub-prefectures, such a decreasing antenna density as the sub-prefecture size grows larger is non-intuitive.

From Figure 8 if we compare the call density across different months, and days of the month, it is interesting to note that the call density towards the middle of the month is lower in *January* than in other months. On the same lines note that call density in the second week of *December* is higher than other months, but reduces in the following weeks. Based on

²These are not recommendations, but only observations based on visual inspection

this observation, we are attempting to find the mobility pattern of mobile phone users and to determine whether the mobility pattern influences this call density pattern.

B. Socio-Economic Perspective

From Figure 9 observe that the call density is considerably higher on *New Year* day as compared to any other festival days. This is an indication of the importance of the New Year in the lifestyle of the people of *Ivory Coast*. What is worth noting is that aggregated call density on New Year's day is about 30% higher than any other normal day. On an average the call density on festival days is about 10-15% higher than normal. While it might be premature to infer conclusively, this seems to indicate a strong emotional bias in the people of *Ivory Coast*. We say this because the economic condition of the people of *Ivory Coast* is not very good, but at the same time during festive periods they do not mind spending on calls. We intend to correlate this data with region-wise socio-economic conditions of the people to infer the telecom spending index and then extrapolate this data to determine region-wise socio-economic conditions. For the sake of completeness and to provide a visual reference related to the geography of *Ivory Coast*, Figure 10 is a map of *Ivory Coast*. Based on our literature survey of the socio-

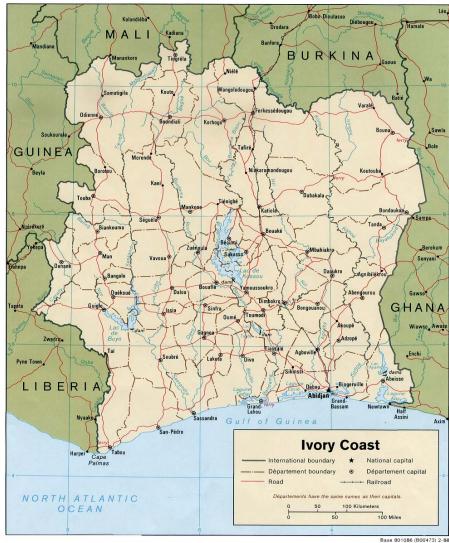


Fig. 10. Map of Ivory Coast

economic conditions of the people of *Ivory Coast* [9], the southern part of *Ivory Coast* (Figure 10) is the the economic capital of the country, but the political capital of the country is *Yamoussoukro* which is in the central region. Comparing Figure 10 and Figure 3 observe that the antenna density is lower in the central region where the capital is located than in the business district of *Abidjan* towards the southern region. Also observe from Figure 5 and Figure 6 that the average call density is higher in the business districts than around the captial of the country. This is indicative of the fact that more of the socio-economic growth is focused in and around the

business districts than around the capital. It will be interesting to determine the mobility profiles of users and check whether there is a distinct movement towards the business districts.

Note that all the above observations were made possible only because of the sophisticated visualisation techniques used to represent the derived information. When correlated with the actual inferred data, this can potentially be a handy tool for the telecom operators to optimise their cellular networks.

VI. CONCLUSION AND FUTURE WORK

This paper analyses anonymised CDRs obtained from *Orange Telecom* for a duration of five months, one of the telecom operators in *Ivory Coast*³. Four data sets as discussed in section II provided by *Orange Telecom* are analysed from a telecom operator and socio-economic perspective. The analysis has been implemented on a Hadoop and Hive framework. This paper proposed a visualisation tool using HTML 5 to visualise the information derived from CDR records.

Information and knowledge related to antenna density, sub-prefectures and call density patterns are derived by analysing CDR records. In this paper we attempt to correlate the derived information and provide observations from a telecom operator perspective and socio-economic perspective. One key observation, from a telecom operator perspective, is that the antenna density does not match with the call density patterns, and there is considerable scope for improvement. Another key observation, from a socio-economic perspective, is that the call density is saturated more towards the business districts than towards the political capital of the country.

While these are just preliminary investigations, in the future we intend to derive the mobility profile of the users and correlate it with the call density and antenna density patterns to provide concrete network optimisation recommendations to the telecom operator. In the process, our larger goal is to propose and develop a tool for CDR record analysis which is configurable and scalable.

REFERENCES

- [1] Chen Zhou et. all, Activity Recognition from Call Detail Record: Relation Between Mobile Behavior Pattern And Social Attribute Using Hierarchical Conditional Random Fields, *2010 IEEE International Conference on Green Computing and Communications & 2010 IEEE International Conference on Cyber, Physical and Social Computing*, June, 2010.
- [2] Qining LIN et. all, Mobile Customer Clustering Based On Call Detail Records For Marketing Campaigns, *International Conference on Management and Service Science* , 2009.
- [3] Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, Carlo Ratti, Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records, *13 International Conference on Intelligent Transportation Systems*, 2010
- [4] Tom White, *Hadoop A Definitive Guide*, O'reilly Press
- [5] Introduction to Hive, *Cloudera Inc.*
- [6] Mapreduce Tutorial, *Apache Software Foundation Tutorial*
- [7] <http://www.orange.com>
- [8] <http://www.openstreetmaps.org>
- [9] <http://www.gouv.ci/Main.php>

³We wish to thank Orange Telecom for providing anonymised CDRs on their Ivory Coast subscriber base for a duration of five months as part of the D4D challenge

Network Visualization with `ggplot2`

by Sam Tyner, François Briatte and Heike Hofmann

Abstract This paper explores three different approaches to visualize networks by building on the grammar of graphics framework implemented in the `ggplot2` package. The goal of each approach is to provide the user with the ability to apply the flexibility of `ggplot2` to the visualization of network data, including through the mapping of network attributes to specific plot aesthetics. By incorporating networks in the `ggplot2` framework, these approaches (1) allow users to enhance networks with additional information on edges and nodes, (2) give access to the strengths of `ggplot2`, such as layers and facets, and (3) convert network data objects to the more familiar data frames.

Introduction

There are many kinds of networks, and networks are extensively studied across many disciplines (Watts, 2004). For instance, social network analysis is a longstanding and prominent sub-field of sociology, and the study of biological networks, such as protein-protein interaction networks or metabolic networks, is a notable sub-field of biology (Prell, 2011; Junker and Schreiber, 2008). In addition, the ubiquity of social media platforms, like Facebook, Twitter, and LinkedIn, has brought the concepts of networks out of academia and into the mainstream. Though these disciplines and the many others that study networks are themselves very different and specialized, they can all benefit from good network visualization tools.

Many R packages already exist to manipulate network objects, such as `igraph` by Csardi and Nepusz (2006), `sna` by Butts (2014), and `network` by Butts et al. (2014) (Butts, 2008, see also). Each one of these packages were developed with a focus of analyzing network data and not necessarily for rendering visualizations of networks. Though these packages do have network visualization capabilities, visualization was not intended as their primary purpose. This is by no means a critique or an inherently negative aspect of these packages: they are all hugely important tools for network analysis that we have relied on heavily in our own work. We have found, however, that visualizing network data in these packages requires a lot of extra work if one is accustomed to working with more common data structures such as vectors, data frames, or arrays. The visualization tools in these packages require detailed knowledge of each one of them and their syntax in order to build meaningful network visualizations with them. This is obviously not a problem if the user is very familiar with network structures and has already spent time working with network data. If, however, the user is new to network data or is more comfortable working with the aforementioned common data structures, they could find the learning curve for these packages burdensome.

The packages described in this paper have, by contrast, have one primary purpose: to create beautiful network visualizations by providing a wrapper of existing network layout capabilities (see for example the `statnet` suite of packages by Handcock et al. (2008)) to the popular `ggplot2` package (Wickham, 2016). And so, our focus here is not on adding to the analysis of network data or to the field of graph drawing, (cf. Tamassia, 2013) but rather it is on implementing existing graph drawing capabilities in the `ggplot2` framework, using the common data frame structure. The `ggplot2` package is hugely popular, and many other packages and tools interface with it in order to better visualize a wide variety of data types. By creating a `ggplot2` implementation, we hope to place network visualization within a large, active community of data visualization enthusiasts, bringing new eyes and potentially new innovations to the field of network visualization. With our approaches, we have two primary audiences in mind. The first audience is made up of frequent users of network structures and those who are fluent in the language of packages such as `network` or `igraph`. This audience will find that two of our three approaches (`ggnet2` and `ggnetwork`) directly incorporate the network structures and functions with which they are familiar with into the less familiar visualization paradigm of `ggplot2` (Briatte, 2016). The second audience, targeted by `geomnet`, consists of those users who are not familiar with network structures, but are familiar with data manipulation and tidying, and who happen to find themselves examining some data that can be expressed as a network (Tyner and Hofmann, 2016a). For this audience, we do the heavy network lifting internally, while also relying on their familiarity with `ggplot2` externally.

The `ggplot2` package was designed as an implementation of the ‘grammar of graphics’ proposed by Wilkinson (1999), and it has become extremely popular among R users.¹

¹In order to give an indication of how large the user base of `ggplot2` is, we looked at its usage statistics from January 1, 2016 to December 31, 2016 (see <http://cran-logs.rstudio.com/>). Over this period, the `ggplot2` package was downloaded over 3.2 million times from CRAN, which amounts to almost 9,000 downloads per day. Almost 800 R packages import or depend on `ggplot2`.

Because the syntax implemented in the **ggplot2** package is extendable to different kinds of visualizations, many packages have built additional functionality on top of the **ggplot2** framework. Examples include the **ggmap** package by Kahle and Wickham (2013) for spatial visualization, the **ggfortify** package for visualizing statistical models (see Horikoshi and Tang (2016), Tang et al. (2016)), the package **GGally** by Schloerke et al. (2016), which encompasses various complementary visualization techniques to **ggplot2**, and the **ggbio** and **gtree** Bioconductor packages by Yin et al. (2012) and Yu et al. (2017), which both provide visualizations for biological data. These packages have expanded the utility of **ggplot2**, likely resulting in an increase of its user base. We hope to appeal to this user base and potentially add to it by applying the benefits of the grammar of graphics implemented in **ggplot2** to network visualization.

Our efforts rely upon recent changes to **ggplot2**, which allow users to more easily extend the package through additional geometries or ‘geoms’.²

In the remainder of this paper, we present three different approaches to network visualization through **ggplot2** wrappers. The first is a function, `ggnet2` from the **GGally** package, that acts as a wrapper around a network object to create a **ggplot2** graph. The second is a package, **geomnet**, that combines all network pieces (nodes, edges, and labels) into a single geom and is intended to look the most like other **ggplot2** geoms in use. The final is another package, **ggnetwork**, that performs some data manipulation and aliases other geoms in order to layer the different network aspects one on top of the other. The section [Brief introduction to networks](#) introduces the basic terminology of networks and illustrates their ubiquity in natural and social life. The next section [Three implementations of network visualizations](#) then discusses the structure and capabilities of each of the three approaches that we offer. The section [Examples](#) extends that discussion through several examples ranging from simple to complex networks, for which we provide the code corresponding to each approach alongside its graphical result. We follow with some considerations of runtime behavior in plotting networks in the section [Some considerations of speed](#) before closing with a discussion.

Brief introduction to networks

In its essence, a *network* is simply a set of vertices connected in pairs by a set of edges (Newman, 2010). Throughout this paper, we also use the term *node* to refer to vertices, as well as the terms *ties* or *relationships* to refer to edges, depending on context. The two sets of graphical objects that make up a network visualization, points and segments between them, have been used to examine a huge variety and quantity of information across many different fields of study. For instance, networks of scientific collaboration, a food web of marine animals, and American college football games are all covered in a paper on community detection in networks by Girvan and Newman (2002). Additionally, Buldyrev et al. (2010) study node failure in interdependent networks like power grids. Social networks such as links between television and film actors found on <http://www.imdb.com> and neural networks, like the completely mapped neural network of the *C. elegans* worm are also extensively studied (Watts and Strogatz, 1998).

These examples show that networks can vary widely in scope and complexity: the smallest connected network is simply one edge between two vertices, while one of the most commonly used and most complex networks, the world wide web, has billions of vertices (Web pages) and billions of edges (hyperlinks) connecting them. Additionally, the edges in a network can be directed or undirected: *directed* edges represent an ordering of vertices, like a relationship extending from one vertex to another, where switching the direction would change the structure of the network. The World Wide Web is an example of a directed network because hyperlinks connect one Web page to another, but not necessarily the other way around. *Undirected* edges are simply connections between vertices where order does not matter. Co-authorship networks are examples of undirected networks, where nodes are authors and they are connected by an edge if they have written an academic publication together.

As a reference example, we turn to a specific instance of a social network. A *social network* is a network that everyone is a part of in one way or another, whether through friends, family, or other human interactions. We do not necessarily refer here to social media like Facebook or LinkedIn, but rather to the connections we form with other people. To demonstrate the functionality of our tools for plotting networks, we have chosen an example of a social network from the popular television show *Mad Men*. This network, which was compiled by Chang (2013) and made available in **gcookbook** (Chang, 2012), consists of 52 vertices and 87 edges. Each vertex represents a character on the show, and there is an edge between every two characters who have had a romantic relationship.

²Version 2.1.0, released 1 March 2016. See <https://cran.r-project.org/web/packages/ggplot2/news.html> for the full list of changes in **ggplot2** 2.1.0, as well as the new package vignette, “Extending ggplot2”, which explains how the internal `ggproto` system of object-oriented programming can be used to create new geoms.

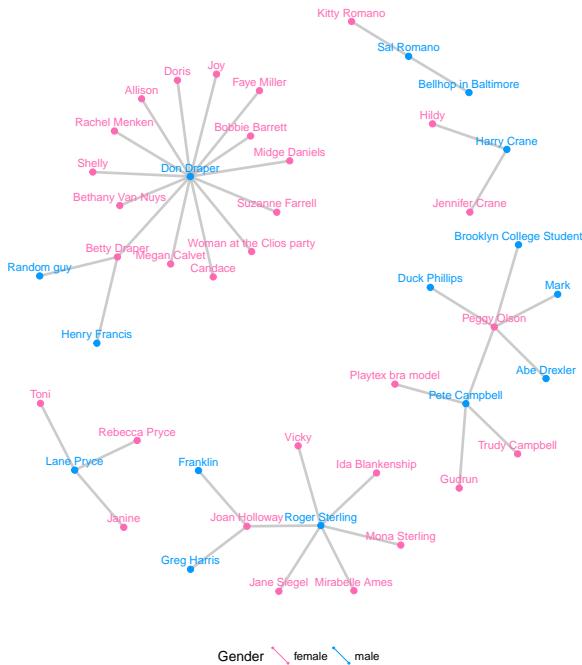


Figure 1: Graph of the characters in the show Mad Men who are linked by a romantic relationship.

Figure 1 is a visualization of this network. In the plot, we can see one central character who has many more relationships than any other character. This vertex represents the main character of the show, Don Draper, who is quite the “ladies’ man.” Networks like this one, no matter how simple or complex, are everywhere, and we hope to provide the curious reader with a straightforward way to visualize any network they choose.

Coloring the vertices or edges in a graph is a quick way to visualize grouping and helps with pattern or cluster detection. The vertices in a network and the edges between them compose the structure of a network, and being able to visually discover patterns among them is a key part of network analysis. Viewing multiple layouts of the same network can also help reveal patterns or clusters that would not be discovered when only viewing one layout or analyzing only its underlying adjacency matrix.

Three implementations of network visualizations

We present two basic approaches to using the **ggplot2** framework for network visualization. First, we implement network visualizations by providing a wrapper function, **ggnet2** for the user to visualize a network using **ggplot2** elements (Schloerke et al., 2016). Second, we implement network visualizations using layering in **ggplot2**. For the second approach, we have two ways of creating a network visualization. The first, **geomnet**, wraps all network structures, including vertices, edges, and vertex labels into a single geom. The second, **ggnetwork**, implements each of these structural components in an independent geom and layers them to create the visualization (Briatte, 2016). In each package, our goal is to provide users with a way to map network properties to aesthetic properties of graphs that is familiar to them and straightforward to implement. Each package has a slightly different approach to accomplish this goal, and we will discuss all of these approaches in this section. For each implementation, we also provide the code necessary to create Figure 1, and describe the arguments used. We conclude the section with a side-by-side comparison of the features available in all three implementations in Table 1.

ggnet2

The **ggnet2** function is a part of the **GGally** package, a suite of functions developed to extend the plotting capabilities of **ggplot2** (Schloerke et al., 2016). A detailed description of the **ggnet2** function is available from within the package as a vignette. Some example code to recreate Figure 1 using **ggnet2** is presented below.

```

library(GGally)
library(network)
# make the data available
data(madmen, package = 'geomnet')
# data step for both ggnet2 and ggnetwork
# create undirected network
mm.net <- network(madmen$edges[, 1:2], directed = FALSE)
mm.net # glance at network object
## Network attributes:
##   vertices = 45
##   directed = FALSE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
##   total edges= 39
##   missing edges= 0
##   non-missing edges= 39
##
## Vertex attribute names:
##   vertex.names
##
## No edge attributes
# create node attribute (gender)
rownames(madmen$vertices) <- madmen$vertices$label
mm.net %v% "gender" <- as.character(
  madmen$vertices[ network.vertex.names(mm.net), "Gender"]
)
# gender color palette
mm.col <- c("female" = "#ff69b4", "male" = "#0099ff")
# create plot for ggnet2
set.seed(10052016)
ggnet2(mm.net, color = mm.col[ mm.net %v% "gender" ],
       labelon = TRUE, label.color = mm.col[ mm.net %v% "gender" ],
       size = 2, vjust = -0.6, mode = "kamadakawai", label.size = 3)

```

The `ggnet2` function offers a large range of network visualization functionality in a single function call. Although its result is a `ggplot2` object that can be further styled with `ggplot2` scales and themes, the syntax of the `ggnet2` function is designed to be easily understood by the users, who may not be familiar with `ggplot2` objects. The aesthetics relating to the nodes are controlled by arguments such as `node.alpha` or `node.color`, while those relating to the edges are controlled by arguments starting with 'edge'. Additionally, as seen in the code above, the usual `ggplot2` arguments like `color` can be used without the prefix to map node attributes to aesthetic values. The arguments with the `node.` prefix are aliased versions for readability of the code. Thus, while `ggnet2` applies the grammar of graphics to network objects, the function itself still works very much like the plotting functions of the `igraph` and `network` packages: a long series of arguments is used to control every possible aspect of how the network should be visualized.

The `ggnet2` function takes a single network object as input. This initial object might be an object of class "network" from the `network` package (with the exception of hypergraphs or multiplex graphs), or any data structure that can be coerced to an object of that class via functions in the `network` package, such as an incidence matrix, an adjacency matrix, or an edge list. Additionally, if the `intergraph` package (Bojanowski, 2015) is installed, the function also accepts a network object of class "igraph". Internally, the function converts the network object to two data frames: one for edges and another one for nodes. It then passes them to `ggplot2`. Each of the two data frames contain the information required by `ggplot2` to plot segments and points respectively, such as a shape for the points (nodes) and a line type for the segments (edges). The final result returned to the user is a plot with a minimum of two layers, or more if there are edge and/or node labels.

The `mode` argument of `ggnet2` controls how the nodes of the network are to be positioned in the plot returned by the function. This argument can take any of the layout values supported by the `gplot.layout` function of the `sna` package, and defaults to 'fruchtermanreingold', which places the nodes through the force-directed layout algorithm by Fruchterman and Reingold (1991). In the example presented above, the Kamada-Kawai layout is used by adding '`mode = "kamadakawai"`' to the

function call. Many other possible layouts and their parameters can also be passed to `ggnet2` through the `layout.par` argument. For a list of possible layouts and their arguments, see `?sna::gplot.layout`.

Other arguments passed to the `ggnet2` function offer extensive control over the aesthetics of the plot that it returns, including the addition of edge and/or node labels and their respective aesthetics. Arguments such as `node.shape` or `edge.lty`, which control the shape of the nodes and the line type of the edges, respectively, can take a single global value, a vector of global values, or the name of an edge or vertex attribute to be used as an aesthetic mapping. This feature is used to change the size of the nodes and the node labels by including ‘`size = 2`’ and ‘`label.size = 3`’ in the function call.

This last functionality builds on one of the strengths of the “network” class, which can store information on network edges and nodes as attributes that are then accessible to the user through the `%e%` and `%v%` operators respectively.³ Usage examples of these operators can be seen above. The attribute of gender is assigned to nodes, which in turn is accessed to color the nodes and node labels by gender. If the `ggnet2` function is given the `node.alpha = "importance"` argument, it will interpret it as an attempt to map the vertex attribute called ‘`importance`’ to the transparency level of the nodes. This works exactly like the command `net %v% "importance"`, which returns the vertex attribute ‘`importance`’ of the “network” object `net`. This functionality allows the `ggnet2` function to work in a similar fashion to `ggplot2` mappings of aesthetics within the `aes` operator.

The `ggnet2` function also provides a few network-specific options, such as sizing the nodes as a function of their unweighted degree, or using the primary and secondary modes of a bipartite network as an aesthetic mapping for the nodes.

All in all, the `ggnet2` function combines two different kinds of processes: it translates a network object into a data frame suitable for plotting with `ggplot2`, and it applies network-related aesthetic operations to that data frame, such as coloring the edges in function of the color of the nodes that they connect.

geomnet

```
# also loads ggplot2
library(geomnet)

# data step: join the edge and node data with a fortify call
MMnet <- fortify(as.edgedf(madmen$edges), madmen$vertices)
# create plot
set.seed(10052016)
ggplot(data = MMnet, aes(from_id = from_id, to_id = to_id)) +
  geom_net(aes(colour = Gender), layout.alg = "kamadaKawai",
           size = 2, labelon = TRUE, vjust = -0.6, ecolour = "grey60",
           directed = FALSE, fontsize = 3, ealpha = 0.5) +
  scale_colour_manual(values = c("#FF69B4", "#0099ff")) +
  xlim(c(-0.05, 1.05)) +
  theme_net() +
  theme(legend.position = "bottom")
```

Data structure

The package `geomnet` implements network visualization in a single `ggplot2` layer. A stable version is available on CRAN, with a development version available at <https://github.com/sctyner/geomnet>. The package has two main functions: `stat_net`, which performs all of the calculations, and `geom_net`, which renders the plot. It also contains the secondary functions `geom_circle` and `theme_net`, which assist, respectively, in drawing self-referencing edges and removing axes and other background elements from the plots. The approach in `geomnet` is similar to the implementation of other, native `ggplot2` geoms, such as `geom_smooth`. When using `geom_smooth`, the user does not need to know about any of the internals of the loess function, and similarly, when using `geomnet`, the user is not expected to know about the internals of the layout algorithm, just the name of the algorithm they’d like to use. On the other hand, if users are comfortable with network analysis, the entire body of layout methods provided by the `sna` package is available to them through the parameters `layout.alg` and `layout.par`.

In network analysis there are usually two sources of information: one data set consisting of a description of the nodes, represented as the vertices in the network and vertex attributes, and another data set detailing the relationship between these nodes, i.e. it consists of the edge list and any additional edge attributes. The minimum amount of information needed is a vector of all vertex labels and

³See Butts et al. (2014, p. 22-24). The equivalent operators in the `igraph` package are called `E` and `V`.

a two column data frame that encodes the edge list of the network. In order for this geometry to work, these two data sets need to be combined into a single data frame. For this, we implemented several new `fortify` methods for producing the correct data structure from different S3 objects that encode network information. Supported classes are "network" from the `sna` and `network` packages, "igraph" from the `igraph` package, "adjmat", and "edgedf". The last two are new classes introduced in `geomnet` that are identical to the "matrix" and "data.frame" classes, respectively. We created these new classes and the functions `as.adjmat()` and `as.edgedf()` so that network data in adjacency matrix and edgelist (data frame) formats can have their own `fortify` functions, separate from the very generic "matrix" and "data.frame" classes. These `fortify` functions combine the edge and the node information using a full join. A full join is used because generally, there will be some vertices that are sinks in the network because they only show up in the 'to' column, and so we accommodate for these by adding artificial edges in the data set that have missing information for the 'to' column. The user may also pass two data frames to the function, e.g. `'data = edge_data'` and `'vertices = vertex_data'`, but we recommend using the `fortify` methods whenever possible.

A usage example of the `fortify.edgedf` method is presented in the code above with the creation of the `MMnet` data set. Two dataframes, `madmen$edges` and `madmen$vertices` are joined to create the required data. The first few rows of these data sets and their merged result are below.

```
head(as.edgedf(madmen$edges), 3)
##      from_id      to_id
## 1 Betty Draper Henry Francis
## 2 Betty Draper Random guy
## 3 Don Draper Allison
head(madmen$vertices, 3)
##             label Gender
## Betty Draper Betty Draper female
## Don Draper   Don Draper male
## Harry Crane Harry Crane male
head(fortify(as.edgedf(madmen$edges), madmen$vertices), 3)
##      from_id      to_id Gender
## 1 Betty Draper Henry Francis female
## 2 Betty Draper Random guy female
## 3 Don Draper Allison male
```

The formal requirements of `stat_net` are two columns, called `from_id` and `to_id`. During this routine, columns `x`, `y` and `xend`, `yend` are calculated and used as a required input for `geom_net`.

Other variables may also be included for each edge, such as the edge weight, in-degree, out-degree or grouping variable.

Parameters and aesthetics

Parameters that are currently implemented in `geom_net` are:

- **layout:** the `layout.alg` parameter takes a character value corresponding to the possible network layouts in the `sna` package that are available within the `gplot.layout.*()` family of functions. The default layout algorithm used is the Kamada-Kawai layout, a force-directed layout for undirected networks ([Kamada and Kawai, 1989](#)).
- **vertices:** any of `ggplot2`'s aesthetics relating to points: `colour`, `size`, `shape`, `alpha`, `x`, and `y` are available and used for specifying the appearance of nodes in the network. For example '`aes(colour = Gender)`' is used above to color the nodes and node labels according to the gender of each character.
- **edges:** for edges we distinguish between two different sets of aesthetics: aesthetics that only relate to line attributes, such as `linewidth` and `linetype`, and aesthetics that are also used by the point `geom`. The former can be used in the same way as they are used in `geom_segment`, while

the latter, like alpha or colour, for instance, are used for vertices unless separately specified. Instead, use the parameters `ecolour` or `ealpha`, which are only applied to the edges. If the group variable is specified, a new variable, called `samegroup` is added during the layout process. This variable is TRUE, if an edge is between two vertices of the same group, and FALSE otherwise. If `samegroup` is TRUE, the corresponding edge will be colored using the same color as the vertices it connects. If the edge is between vertices of a different group, the default grey shade is used for the edge.

The parameter `curvature` is set to zero by default, but if specified, leads to curved edges using the newly implemented `ggplot2` geom `geom_curve` instead of the regular `geom_segment`. Note that the edge specific aesthetics that overwrite node aesthetics are currently considered as ‘as.is’ values: they do not get a legend and are not scaled within the `ggplot2` framework. This is done to avoid any clashes between node and edge scales.

self-referencing vertices: some networks contain self references, i.e. an edge has the same vertex id in its from and to columns. If the parameter `selfloops` is set to TRUE, a circle is drawn using the new `geom_circle` next to the vertex to represent this self reference.

- **arrow:** whenever the parameter `directed` is set from its default state to TRUE, arrows are drawn from the ‘from’ to the ‘to’ node, with tips pointing towards the ‘to’ node. By default, arrows have an absolute size of 10 points. The entire structure of the arrow can be changed by passing an `arrow` object from the `grid` package to the `arrow` argument. If the user doesn’t wish to change the whole arrow object, the parameters `arrowsize` and `arrowgap` are also available. The `arrowsize` argument is of a positive numeric value that is used as a multiple of the original arrow size, i.e. `arrowsize = 2` shows arrow tips at twice their original size. The parameter `arrowgap` can be used to avoid overplotting of the arrow tips by the nodes, `arrowgap` specifies a proportion by which the edge should be shrunk with default of 0.05. A value of 0.5 will result in edges drawn only half way from the ‘from’ node to the ‘to’ node.
- **labels:** the `labelon` argument is a logical parameter, which when set to TRUE will label the nodes with their IDs, as is in Figure 1. The `aes` option `label` can also be used to label nodes, in which case the nodes are labeled with the value corresponding to their respective values of the provided variable. If `colour` is specified for the nodes, the same values are used for the labels, unless `labelcolour` is specified. If `fontsize` is specified, it changes the label size to that value in points. Other parameter values, such as `vjust` and `hjust` help in adjusting labels relative to the nodes. The parameters work in the same fashion as in native `ggplot2` geoms. Additionally, the label can be drawn by using `geom_text` (the default) or using the new `geom_label` in `ggplot2` by adding ‘`labelgeom = "label"`’ to the arguments in `geom_net`. Finally, with the help of the package `ggrepel` by [Slowikowski \(2016\)](#) we have implemented the logical `repel` argument, which when true, uses `geom_text_repel` or `geom_label_repel` to plot the labels instead of `geom_text` or `geom_label`, respectively. Using `repel` can be extremely useful when the networks are dense or the labels are long, as in Figure 1, helping to solve a common problem with many network visualizations.

ggnetwork

ggnetwork is a small R package that mimics the behavior of `geomnet` by defining several geoms to achieve similar results.

```
# create plot for ggnetwork. uses same data created for ggnet2 function
library(ggnetwork)
set.seed(10052016)
ggplot(data = ggnetwork(mm.net, layout = "kamadakawai"),
       aes(x, y, xend = xend, yend = yend)) +
  geom_edges(color = "grey50") + # draw edge layer
  geom_nodes(aes(colour = gender), size = 2) + # draw node layer
  geom_nodetext(aes(colour = gender, label = vertex.names),
                size = 3, vjust = -0.6) + # draw node label layer
  scale_colour_manual(values = mm.col) +
  xlim(c(-0.05, 1.05)) +
  theme_blank() +
  theme(legend.position = "bottom")
```

The approach taken by the `ggnetwork` package is to alias some of the native geoms of the `ggplot2` package. An aliased geom is simply a variant of an already existing one. The `ggplot2` package contains

several examples of aliased geoms, such as `geom_histogram`, which is a variant of `geom_bar` see (see Wickham, 2016, p. 67, Table 4.6).

Following that logic, the **ggnetwork** package adds four aliased geometries to **ggplot2**:

- `geom_nodes`, an alias to `geom_point`;
- `geom_edges`, an alias to either `geom_segment` or `geom_curve`;
- `geom_nodetext`, an alias to `geom_text`; and
- `geom_edgetext`, an alias to `geom_label`.

The four geoms are used to plot nodes, edges, node labels and edge labels, respectively. Two of the geoms that they alias, `geom_curve` and `geom_label`, are part of the new geometries introduced in **ggplot2** version 2.1.0. All four geoms behave exactly like those that they alias, and take exactly the same arguments. The only exception to that rule is the special case of `geom_edges`, which accepts both the arguments of `geom_segment` and those of `geom_curve`; if its curvature argument is set to anything but 0 (the default), then `geom_edges` behaves exactly like `geom_curve`; otherwise, it behaves exactly like `geom_segment`. Three of the four available aliased geoms are used above to create the visualization of the Mad Men relationship network.

Just like the `ggnetwork` function, the **ggnetwork** package takes a single network object as input. This can be an object of class "network", some data structure coercible to that class, or an object of class "igraph" when the **intergraph** package is installed. This object is passed to the 'workhorse' function of the package, which is also called `ggnetwork` to create a data frame, and then to the `data` argument of `ggplot()`.

Internally, the `ggnetwork` function starts by computing the `x` and `y` coordinates of all nodes in the network with respect to its `layout` argument, which defaults to the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold, 1991). It then extracts the edge list of the network, to which it adds the coordinates of the sender and receiver nodes as well as all edge-level attributes. The result is a data frame with as many rows as there are edges in the network, and where the `x`, `y`, `xend` and `yend` hold the coordinates of the network edges.

At that stage, the `ggnetwork` function, like the **geomnet** package, performs a left-join of that augmented edge list with the vertex-level attributes of the 'from' nodes. It also adds one self-loop per node, in order to ensure that every node is plotted even when their degree is zero—that is, even if the node is not connected to any other node of the network, and is therefore absent from the edge list. The data frame created by this process contains one row per edge as well as one additional row per node, and features all edge-level and vertex-level attributes initially present in the network.⁴

The `ggnetwork` function also accepts the arguments `arrow.gap` and `by`. Like in **geomnet**, `arrow.gap` slightly shortens the edges of directed networks in order to avoid overplotting edge arrows and nodes. The argument `by` is intended for use with plot facets. Passing an edge attribute as a grouping variable to the `by` argument will cause `ggnetwork` to return a data frame in which each node appears as many times as there are unique values of that edge attribute, using the same coordinates for all occurrences. When that same edge attribute is also passed to either `facet_wrap` or `facet_grid`, each edge of the network will show in only one panel of the plot, and all nodes will appear in each of the panels at the same position. This makes the panels of the plot comparable to each other, and allows the user to visualize the network structure as a function of a specific edge attribute, like a temporal attribute.

Examples

In this section, we demonstrate some of the current capabilities of `ggnetwork`, **geomnet**, and **ggnetwork** in a series of side by side examples. While the output is nearly identical for each method of network visualization, the code and implementations differ across the three methods. For each of these examples, we present the code necessary to produce the network visualization in each of the three packages, and discuss each application in detail.

For the following examples we will be loading all three packages under comparison. In practice, only one of these packages would be needed to visualize a network in the **ggplot2** framework:

```
library(ggplot2)
library(GGally)
```

⁴One limitation of this process is that it requires some reserved variable names (`x`, `y`, `xend` and `yend`), which should not also be present as edge-level or vertex-level attributes (otherwise the function will simply break). Similarly, if an edge attribute and a vertex attribute have the same name, like 'na', which the **network** package defines as an attribute for both edges and vertices in order to flag missing data, `ggnetwork` will rename them to 'na.x' (for the edge-level attribute) and 'na.y' (for the vertex-level attribute).

Functionality	ggnet2 (GGally)	geom_net (geomnet)	geom_nodes, geom_edges, etc (ggnetwork)
Data	object of class "network" or object easily converted to that class (i.e. incidence or adjacency matrices, edge list) or object of class "igraph"	a fortified "network", "igraph", "edgedf", or "adjmat" object OR one edge data frame and one node data frame to be merged internally	same as ggnet2
Naming conventions	node._, edge._, label._, edge.label._ for alpha, color, etc.	arguments identical to ggplot2 with exception of ecolor, ealpha	same as ggplot2
Layout package & default	sna , Fruchterman-Reingold	sna , Kamada-Kawai	sna , Fruchterman-Reingold
Aesthetic mappings to variables	all alpha, color, shape, size for nodes, edges, labels	colour, size, shape, x, y, linetype, linewidth, label, group, fontsize	same as ggplot2
Arrows	directed = TRUE, arrow.size, gap	arrowsize, gap, arrow = arrow() like ggplot2	specify arrows in geom_edge like in code-geom_segment, arrow.gap
Theme or palette changes	done in the function with arguments like ..legend, ..palette, etc. and adding ggplot2 elements	adding ggplot2 elements	adding ggplot2 elements
Creating small multiples	created separately, use grid.arrange from gridExtra	add group argument to fortify() and use facet_*(*) from ggplot2	use by argument in ggnetwork() and facet_* () from ggplot2
Edge labelling?	Yes	No	Yes
Draw self-loops?	No	Yes	No

Table 1: Comparing the three different package side-by-side.

```
library(geomnet)
library(ggnetwork)
```

Blood donation

We begin with a very simple example that most should be familiar with: blood donation. In this directed network, there are eight vertices and 27 edges. The vertices represent the eight different blood types in humans that are most important for donation: the ABO blood types A, B, AB, and O, combined with the RhD positive (+) and negative (-) types. The edges are directed: a person whose blood type is that of a *from* vertex can donate blood to a person whose blood type is that of a corresponding *to* vertex. This network is shown in Figure 2. The code to produce each one of the networks is shown above Figure 2. We take advantage of each approach's ability to assign identity values to the aesthetic values. The color is changed to a dark red, the size of the nodes is changed to be large enough to accomodate the blood type label, which we also change the color of, and we use the directed and arrow arguments of each implementation to show the precise blood donation relationships. Additionally, we change the node layout to circle, and the placement of the labels with the *hjust* and *vjust* options.

```
# make data accessible
data(blood, package = "geomnet")

# plot with ggnet2 (Figure 5a)
set.seed(12252016)
ggnet2(network(blood$edges[, 1:2], directed=TRUE),
       mode = "circle", size = 15, label = TRUE,
       arrow.size = 10, arrow.gap = 0.05, vjust = 0.5,
       node.color = "darkred", label.color = "grey80")

head(blood$edges,3) # glance at the data
##   from    to group_to
## 1 AB- AB+     same
## 2 AB- AB-     same
## 3 AB+ AB+     same
# plot with geomnet (Figure 5b)
set.seed(12252016)
ggplot(data = blood$edges, aes(from_id = from, to_id = to)) +
  geom_net(colour = "darkred", layout.alg = "circle", labelon = TRUE, size = 15,
           directed = TRUE, vjust = 0.5, labelcolour = "grey80",
           arrowsize = 1.5, linewidth = 0.5, arrowgap = 0.05,
           selfloops = TRUE, ecolour = "grey40") +
  theme_net()

# plot with ggnetwork (Figure 5c)
set.seed(12252016)
ggplot(ggnetwork(network(blood$edges[, 1:2]),
                 layout = "circle", arrow.gap = 0.05),
       aes(x, y, xend = xend, yend = yend)) +
  geom_edges(color = "grey50",
             arrow = arrow(length = unit(10, "pt"), type = "closed")) +
  geom_nodes(size = 15, color = "darkred") +
  geom_nodetext(aes(label = vertex.names), color = "grey80") +
  theme_blank()
```

In this example every vertex has a self-reference, as blood between two people of matching ABO and RhD type can always be exchanged. The **geomnet** approach shows these self-references as circles looping back to the vertex, which is controlled by using the parameter setting *selfloops* = TRUE.

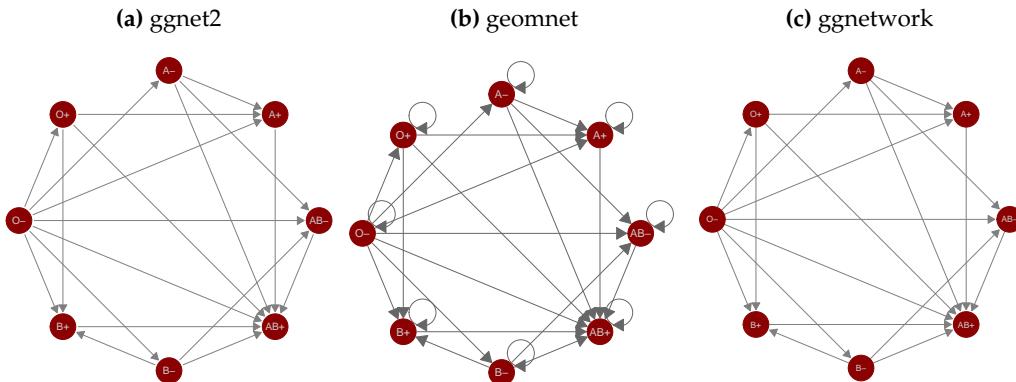


Figure 2: Network of blood donation possibilities in humans by ABO and RhD blood types.

colour and size aesthetics in Figure 2 are set to identity values to change the size and color of all vertices. We have also used the layout and label arguments to change the default Kamada-Kawai layout to a circle layout and to print labels for each of the blood types. The circle layout places blood types of the same ABO type next to each other and spreads the vertices out far enough to distinguish between the various “in” and “out” types. We can tell clearly from this plot that the O-type is the universal donor: it has an out-degree of seven and an in-degree of zero. Additionally, we can see that the AB+ type is the universal recipient, with an in-degree of seven and an out-degree of zero. Anyone looking at this plot can quickly determine which type(s) of blood they can receive and which type(s) can receive their blood.

Email network

The email network comes from the 2014 VAST Challenge (Cook et al., 2014). It is a directed network of emails between company employees with 55 vertices and 9,063 edges. Each vertex represents an employee of the company, and each edge represents an email sent from one employee to another. The arrow of the directed edge points to the recipient of the email. If an email has multiple recipients, multiple edges, one for each recipient, are included in the network. The network contains two business weeks of emails across the entire company. In order to better visualize the structure of the communication network between employees, emails that were sent out to all employees are removed. A glimpse of the data objects used is below.

```
em.net # ggnet2 and ggnetwork
## Network attributes:
##   vertices = 55
##   directed = TRUE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
##   total edges= 4743
##   missing edges= 0
##   non-missing edges= 4743
##
##   Vertex attribute names:
##     curr_empl_type vertex.names
##
##   Edge attribute names not shown
emailnet[1,c(1:2,7,21)] # geomnet
##                                     from_id
## 1 Ada.Campo-Corrente@gastech.com.kronos
##                                     to_id day
## 1 Ingrid.Barranco@gastech.com.kronos 10
##   CurrentEmploymentType
## 1                         Executive
```

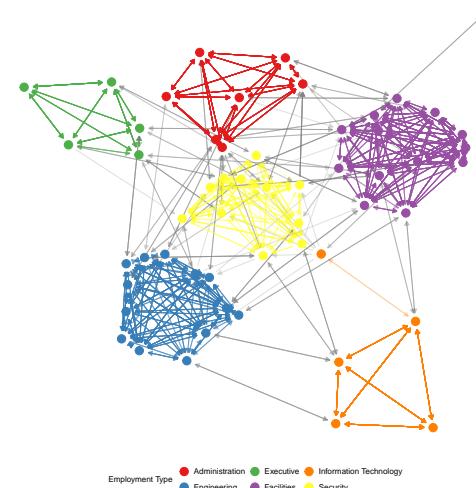
Emails taken by themselves form an event network, i.e. edges do not have any temporal duration.

(a) ggnet2

```
# make data accessible
data(email, package = 'geomnet')

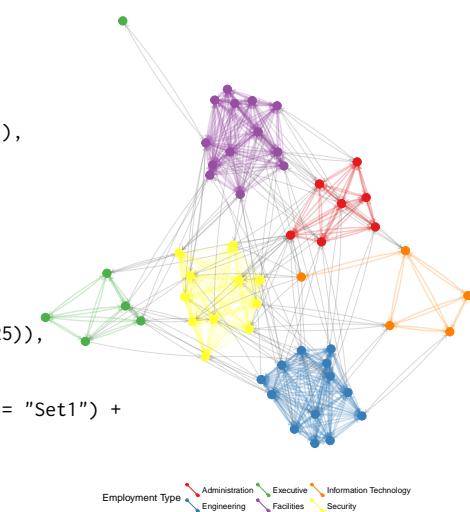
# create node attribute data
em.cet <- as.character(
  email$nodes$CurrentEmploymentType)
names(em.cet) = email$nodes$label

# remove the emails sent to all employees
edges <- subset(email$edges, nrecipients < 54)
# create network
em.net <- edges[, c("From", "to") ]
em.net <- network(em.net, directed = TRUE)
# create employee type node attribute
em.net %v% "curr_empl_type" <-
  em.cet[ network.vertex.names(em.net) ]
set.seed(10312016)
ggnet2(em.net, color = "curr_empl_type",
  size = 4, palette = "Set1", arrow.gap = 0,
  arrow.size = 5, edge.alpha = 0.25,
  mode = "fruchtermanreingold",
  edge.color = c("color", "grey50"),
  color.legend = "Employment Type" +
  theme.legend.position = "bottom")}
```



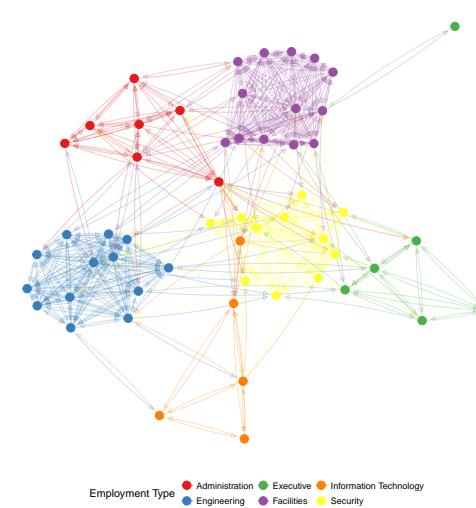
(b) geomnet

```
# data step for the geomnet plot
email$edges <- email$edges[, c(1,5,2:4,6:9)]
emailnet <- fortify(
  as.edgedf(subset(email$edges, nrecipients < 54)),
  email$nodes)
set.seed(10312016)
ggplot(data = emailnet,
  aes(from_id = from_id, to_id = to_id)) +
  geom_net(layout.alg = "fruchtermanreingold",
    aes(colour = CurrentEmploymentType,
      group = CurrentEmploymentType,
      linewidth = 3 * (...samegroup.. / 8 + .125),
      ealpha = 0.25, size = 4, curvature = 0.05,
      directed = TRUE, arrowsize = 0.5) +
  scale_colour_brewer("Employment Type", palette = "Set1") +
  theme_net() +
  theme.legend.position = "bottom")
```



(c) ggnetwork

```
# use em.net created in ggnet2step
set.seed(10312016)
ggplot(ggnetwork(em.net, arrow.gap = 0.02,
  layout = "fruchtermanreingold"),
  aes(x, y, xend = xend, yend = yend)) +
  geom_edges(
    aes(color = curr_empl_type),
    alpha = 0.25,
    arrow = arrow(length = unit(5, "pt"),
      type = "closed"),
    curvature = 0.05) +
  geom_nodes(aes(color = curr_empl_type),
    size = 4) +
  scale_color_brewer("Employment Type",
    palette = "Set1") +
  theme_blank() +
  theme.legend.position = "bottom")
```

**Figure 3:** Email network within a company over a two week period.

Here, however, we can think of emails as observable expressions of the underlying, unobservable, relationship between employees. We can think of this network as a dynamic temporal network, i.e. this network has the potential to change over time. The `ndtv` package by Bender-deMoll (2016) allows the analysis of such networks and provides impressive animations of the underlying dynamics. Here, we are using two static approaches to visualize the network: first, we aggregate emails across the whole time frame (shown in Figure 3), then we aggregate emails by day and use small multiples to allow a comparison of day-to-day behavior (shown in Figure 4).

For all of the email examples, we have colored the vertices by the variable `CurrentEmploymentType`, which contains the department in the company of which each employee is a part of. There are six distinct clusters in this network which almost perfectly correspond to the six different types of employees in this company: administration, engineering, executive, facilities, information technology, and security. Other features in the code include using alpha arguments to change the transparency of the edges, curvature arguments to show mutual communication as two edges instead of one edge with two arrowheads, and the addition of `ggplot2` functions like `scale_colour_brewer` and `theme` to customize the colors of the nodes and their corresponding legend.

In Figure 3 we can clearly see the varying densities of communications within departments and the more sparse communication between employees in different departments. We also see that one of the executives only communicates with employees in Facilities, while one of the IT employees frequently communicates with security employees.

A comparison of the results of `ggnet`, `geomnet` and `ggnetwork` reveals some of the more subtle differences between the implementations:

- In the `ggnet2` implementation, the opacity of the edges between employees in the same cluster is higher than it is for the edges between employees in different clusters. This is due to the fact that the email network does not make use of edge weights: instead, every email between two employees is represented by an edge, resulting in edge overplotting. The `edge.alpha` argument has been set to a value smaller than one, therefore multiple emails between two employees create more opaque edges between them. Multiple emails are also taken into account in the `geomnet` package. When there is more than one edge connecting two vertices, the `stat_net` function adds a weight variable to the edge list, which is passed automatically to the layout algorithms and taken into account during layout. This is thanks to the `sna` package, which supports the use of weights in its edge list. In addition to taking weights into account in the layout, we can also make use of them in the visualization. `geomnet` allows to access all of the internal variables created in the visualization process, such as coordinates `..x..`, `..y..` and edge weights `..weight...`. Note the use of the `ggplot2` notation `...` for internal variables.
- In the first two layouts of Figure 3, edges between employees who share the same employment type are given the color of that employment type, while edges between employees belonging to different types are plotted in grey. This feature is particularly useful to visualize the amount of within-group connectedness in a network. By contrast, in the last layout, edges are colored according to the sender's employment type, because the `ggnetwork` package does not support coloring edges as a function of node-level attributes.
- Finally, in the last two layouts of Figure 3, the curvature argument has been set to 0.05, resulting in slightly curved edges in both plots. This feature, which takes advantage of the `geom_curve` geometry released in `ggplot2` 2.1.0, makes it possible to visualize which edges correspond to reciprocal connections; in an email communication network, as one might expect, most edges fall into that category.

To give some insight into how the relations between employees change over time, we facet the network by day: each panel in Figure 4 shows email networks associated with each day of the work week. The code for these visualizations is below. The different approaches create small multiples in different ways. The `ggnet2` approach requires that the network be separated, each plot created individually, then placed together using the `grid.arrange` function from the `gridExtra` package (Auguie, 2016). The `geomnet` approach uses the `facet_*` family of functions just as they are used in `ggplot2`, and the `ggnetwork` approach uses the `by` argument in the `ggnetwork` function in combination with the `facet_*` functions. We present the full code for each of these approaches below.

First, the code for the `ggnet2` approach, which results in Figure 4(a):

```
# data preparation. first, remove emails sent to all employees
em.day <- subset(email$edges, nrecipients < 54)[, c("From", "to", "day") ]
# for small multiples by day, create one element in a list per day
# (10 days, 10 elements in the list em.day)
em.day <- lapply(unique(em.day$day),
                 function(x) subset(em.day[, 1:2], day == x))
```

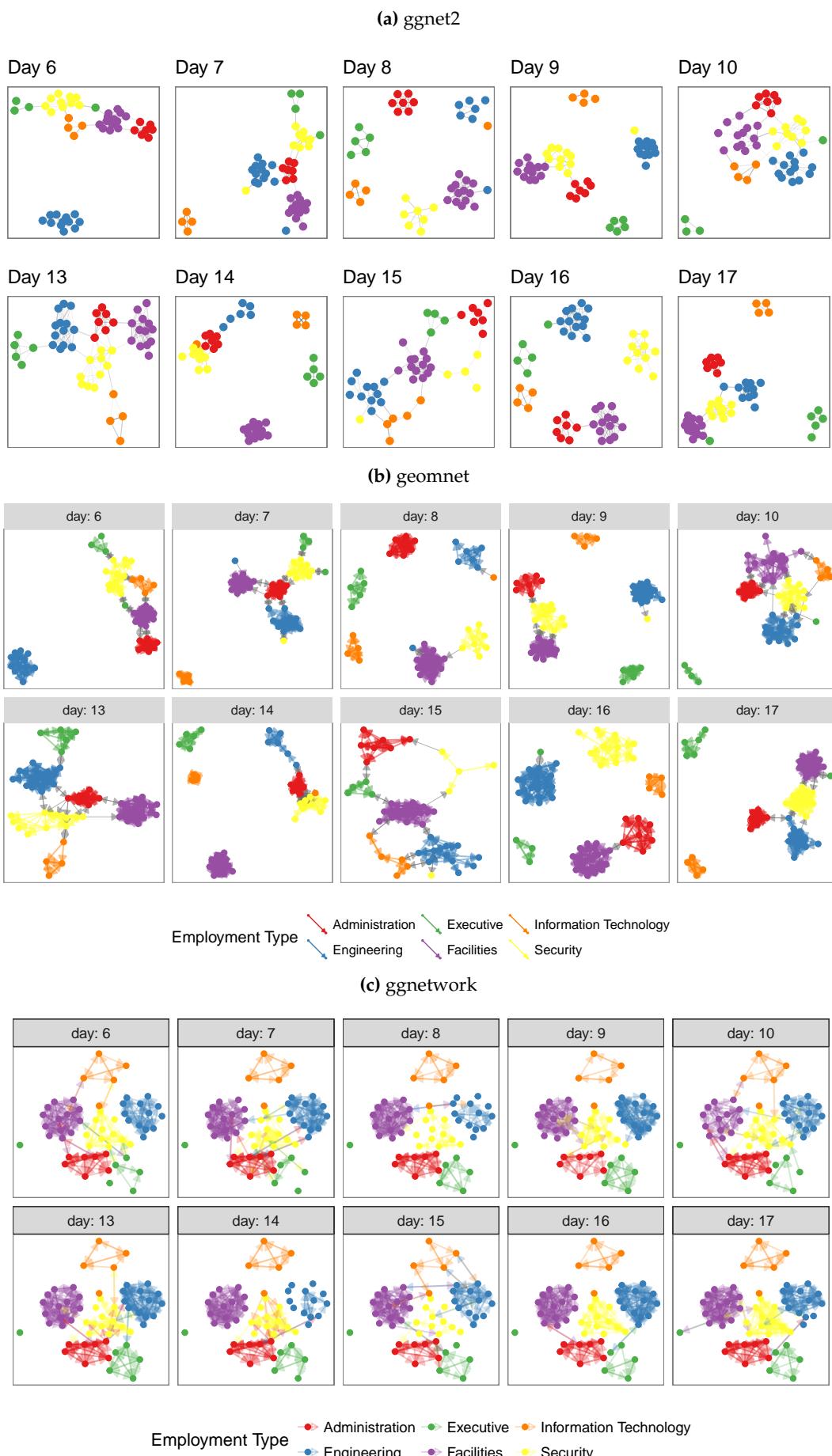


Figure 4: The same email network as in Figure 3 faceted by day of the week.

```

# make the list of edgelists a list of network objects for plotting with ggnet2
em.day <- lapply(em.day, network, directed = TRUE)
# create vertex (employee type) and network (day) attributes for each element in list
for (i in 1:length(em.day)) {
  em.day[[ i ]] %v% "curr_empl_type" <-
    em.cet[ network.vertex.names(em.day[[ i ]]) ]
  em.day[[ i ]] %n% "day" <- unique(email$edges$day)[ i ]
}

# plot ggnet2
# first, make an empty list containing slots for the 10 days (one plot per day)
g <- list(length(em.day))
set.seed(7042016)
# create a ggnet2 plot for each element in the list of networks
for (i in 1:length(em.day)) {
  g[[ i ]] <- ggnet2(em.day[[ i ]], size = 2,
                     color = "curr_empl_type",
                     palette = "Set1", arrow.size = 0,
                     arrow.gap = 0.01, edge.alpha = 0.1,
                     legend.position = "none",
                     mode = "kamadakawai") +
    ggtitle(paste("Day", em.day[[ i ]] %n% "day")) +
    theme(panel.border = element_rect(color = "grey50", fill = NA),
          aspect.ratio = 1)
}
# arrange all of the network plots into one plot window
gridExtra::grid.arrange(grobs = g, nrow = 2)

```

Second, the code for the **geomnet** approach, which results in Figure 4(b):

```

# data step: use the fortify.edgedf group argument to
# combine the edge and node data and allow all nodes to
# show up on all days. Also, remove emails sent to all
# employees
emailnet <- fortify(as.edgedf(subset(email$edges, nrecipients < 54)), email$nodes, group = "day")

# creating the plot
set.seed(7042016)
ggplot(data = emailnet, aes(from_id = from, to_id = to_id)) +
  geom_net(layout.alg = "kamadakawai", singletons = FALSE,
           aes(colour = CurrentEmploymentType,
               group = CurrentEmploymentType,
               linewidth = 2 * (...samegroup.. / 8 + .125)),
           arrowsize = .5,
           directed = TRUE, fiteach = TRUE, ealpha = 0.5, size = 1.5, na.rm = FALSE) +
  scale_colour_brewer("Employment Type", palette = "Set1") +
  theme_net() +
  facet_wrap(~day, nrow = 2, labeller = "label_both") +
  theme(legend.position = "bottom",
        panel.border = element_rect(fill = NA, colour = "grey60"),
        plot.margin = unit(c(0, 0, 0, 0), "mm"))

```

Finally, the code for the **ggnetwork** approach, which results in Figure 4(c):

```

# create the network and aesthetics
# first, remove emails sent to all employees
edges <- subset(email$edges, nrecipients < 54)
edges <- edges[, c("From", "to", "day")]
# Create network class object for plotting with ggnetwork
em.net <- network(edges[, 1:2])
# assign edge attributes (day)
set.edge.attribute(em.net, "day", edges[, 3])

```

```

# assign vertex attributes (employee type)
em.net %v% "curr_empl_type" <- em.cet[ network.vertex.names(em.net) ]

# create the plot
set.seed(7042016)
ggplot(ggnetwork(em.net, arrow.gap = 0.02, by = "day",
                  layout = "kamadakawai"),
       aes(x, y, xend = xend, yend = yend)) +
  geom_edges(
    aes(color = curr_empl_type),
    alpha = 0.25,
    arrow = arrow(length = unit(5, "pt"), type = "closed")) +
  geom_nodes(aes(color = curr_empl_type), size = 1.5) +
  scale_color_brewer("Employment Type", palette = "Set1") +
  facet_wrap(~day, nrow = 2, labeller = "label_both") +
  theme_facet(legend.position = "bottom")

```

Note the two key differences in the visualizations of Figure 4: whether singletons (isolated nodes) are plotted (as in the **ggnetwork** method), and whether *one* layout is used across all panels (as for the **ggnetwork** example) or whether individual layouts are fit to each of the subsets (as for the **ggnet2** and the **geomnet** examples). Plotting isolated nodes in **geomnet** is possible by setting `singletons = TRUE`, and it would be possible in **ggnet2** by including all nodes in the creation of the list of networks. Using the same layout for plotting small multiples in **geomnet** is controlled by the argument `fiteach`. By default, `fiteach = TRUE`, but `fiteach = FALSE` results in all panels sharing the same layout. Having the same layout in each panel makes seeing specific differences in ties between nodes easier, while having a different layout in each panel emphasizes the overall structural differences between the sub-networks. It would be interesting to be able to have a hybrid of these two approaches, but at the moment this is beyond the capability of any of the methods. Through the faceting it becomes obvious that there are several days where one or more of the departments does not communicate with any of the other departments. There are only two days, day 13 and day 15, without any isolated department communications. Faceting is one of the major benefits of implementing tools for network visualization in **ggplot2**. Faceting allows the user to quickly separate dense networks into smaller sub-networks for easy visual comparison and analyses, a feature that the other network visualization tools do not have.

ggplot2 theme elements

This example comes from the `theme()` help page in the **ggplot2** documentation (Wickham, 2016). It is a directed network which shows the structure of the inheritance of theme options in the construction of a **ggplot2** plot. There are 53 vertices and 36 edges in this network. Each vertex represents one possible theme option. There is an arrow from one theme option to another if the element represented by the 'to' vertex inherits its values from the 'from' vertex. For example, the `axis.ticks.x` option inherits its value from the `axis.ticks` value, which in turn inherits its value from the `line` option. Thus, setting the `line` option to a value such as `element_blank()` sets the entire inheritance tree to `element_blank()`, and no lines appear anywhere on the plot background.

Code and plots of the inheritance structure are shown in Figure 5. A glimpse of the data is below.

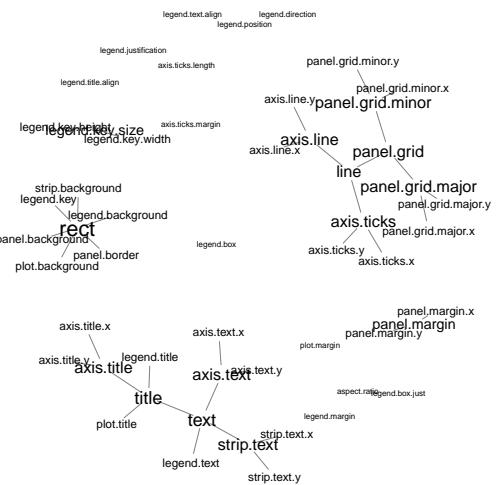
```

te.net
## Network attributes:
##   vertices = 53
##   directed = TRUE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
##   total edges= 48
##     missing edges= 0
##     non-missing edges= 48
##
##   Vertex attribute names:
##     size vertex.names
##
##   No edge attributes

```

```
# make data accessible
data(theme_elements, package = "geomnet")

# create network object
te.net <- network(theme_elements$edges)
# assign node attribut (size based on node degree)
te.net %v% "size" <-
  sqrt(10 * (sna::degree(te.net) + 1))
set.seed(3272016)
ggnet2(te.net, label = TRUE, color = "white",
       label.size = "size", layout.exp = 0.15,
       mode = "fruchtermanreingold")
```



```
# data step: merge nodes and edges and
# introduce a degree-out variable
# data step: merge nodes and edges and
# introduce a degree-out variable
TENet <- fortify(
  as.edgedf(theme_elements$edges[,c(2,1)],
            theme_elements$vertices)
TENet <- TENet %>%
  group_by(from_id) %>%
  mutate(degree = sqrt(10 * n() + 1))

# create plot:
set.seed(3272016)
ggplot(data = TENet,
       aes(from_id = from_id, to_id = to_id)) +
  geom_net(layout.alg = "fruchtermanreingold",
           aes(fontsize = degree), directed = TRUE,
           labelon = TRUE, size = 1, labelcolour = 'black',
           ecolour = "grey70", arrowsize = 0.5,
           linewidth = 0.5, repel = TRUE) +
  theme_net() +
  xlim(c(-0.05, 1.05))
```

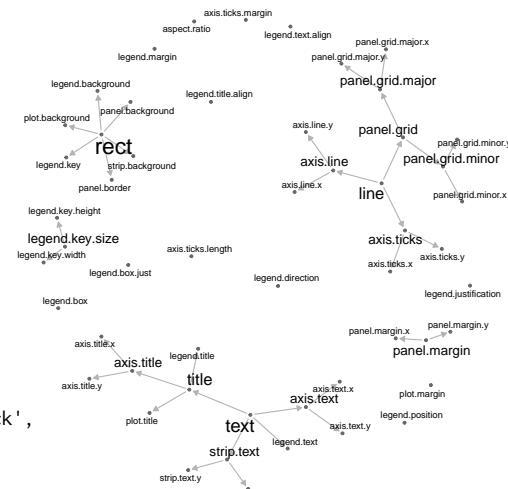


Figure 5: Inheritance structure of `ggplot2` theme elements. This is a recreation of the graph found at <http://docs.ggplot2.org/current/theme.html>.

(c)

```

set.seed(3272016)
# use network created in ggnet2 data step
ggplot(ggnetwork(te.net,
                  layout = "fruchtermanreingold",
                  aes(x, y, xend = xend, yend = yend)) +
  geom_edges() +
  geom_nodes(size = 12, color = "white") +
  geom_nodetext(
    aes(size = size, label = vertex.names)) +
  scale_size_continuous(range = c(4, 8)) +
  guides(size = FALSE) +
  theme_blank()

```

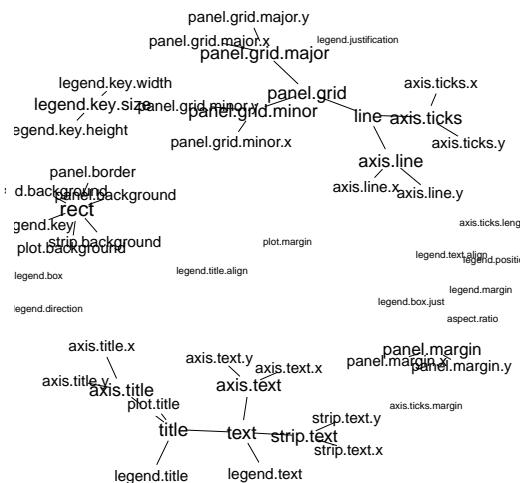


Figure 5: (continued) Inheritance structure of `ggplot2` theme elements. This is a recreation of the graph found at <http://docs.ggplot2.org/current/theme.html>.

```
head(TNet)
## Source: local data frame [6 x 3]
## Groups: from_id [2]
##
##   from_id      to_id degree
##   <fctr>      <fctr> <dbl>
## 1   text      title 6.403124
## 2   text legend.text 6.403124
## 3   text    axis.text 6.403124
## 4   text    strip.text 6.403124
## 5   line    axis.line 5.567764
## 6   line    axis.ticks 5.567764
```

Note the various ways the packages adjust the size of the labels to correspond to the outdegree of the nodes, including the use of the `scale_size_continuous` function in Figure 5(c). In each of these plots, it is easy to quickly determine parent-child relationships, and to assess which theme elements are unrelated to all others. Nodes with the most children are the `rect`, `text`, and `line` elements, so we made their labels larger in order to emphasize their importance. In each case, the label size is a function of the out degree of the vertices.

College football

This next example comes from M.E.J. Newman's network data web page ([Girvan and Newman, 2002](#)). It is an undirected network consisting of all regular season college football games played between Division I schools in Fall of 2000. There are 115 vertices and 613 edges: each vertex represents a school, and an edge represents a game played between two schools. There is an additional variable in the vertex data frame corresponding to the conference each team belongs to, and there is an additional variable in the edge data frame that is equal to one if the game occurred between teams in the same conference or zero if the game occurred between teams in different conferences. We take a look at the data used in the plots below.

```
fb.net
## Network attributes
##  vertices = 115
##  directed = TRUE
##  hyper = FALSE
##  loops = FALSE
##  multiple = FALSE
```

```

##  bipartite = FALSE
##  total edges= 613
##    missing edges= 0
##    non-missing edges= 613
##
##  Vertex attribute names:
##    conf vertex.names
##
##  Edge attribute names:
##    same.conf
head(ftnet)
##   from_id          to_id same.conf      value
## 1 AirForce    NevadaLasVegas      1 Mountain West
## 2 Akron        MiamiOhio       1 Mid-American
## 3 Akron    VirginiaTech       0 Mid-American
## 4 Akron         Buffalo       1 Mid-American
## 5 Akron  BowlingGreenState      1 Mid-American
## 6 Akron           Kent       1 Mid-American
## schools
## 1
## 2
## 3
## 4
## 5
## 6

```

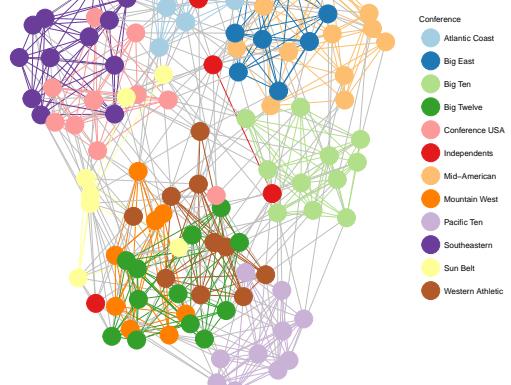
The network of football games is given in Figure 6. Here, the linetype aesthetic corresponds to games that occur between teams in the same conference or different conferences.

```

#make data accessible
data(football, package = 'geomnet')
rownames(football$vertices) <-
  football$vertices$label
# create network
fb.net <- network(football$edges[, 1:2],
                  directed = TRUE)
# create node attribute
# (what conference is team in?)
fb.net %v% "conf" <-
  football$vertices[
    network.vertex.names(fb.net), "value"
  ]
# create edge attribute
# (between teams in same conference?)
set.edge.attribute(
  fb.net, "same.conf",
  football$edges$same.conf)
set.seed(5232011)
ggnet2(fb.net, mode = "fruchtermanreingold",
       color = "conf", palette = "Paired",
       color.legend = "Conference",
       edge.color = c("color", "grey75"))

```

(a) ggnet2



These lines are dotted and solid, respectively. We have also assigned a different color to each conference, so that the vertices and their labels are colored according to their conference. Additionally, in the first two implementations, the edges between two teams in the same conference share that conference color, while edges between teams in different conferences are a default gray color. This coloring and changing of the line types make the structure of the game network easier to view. Additionally, we use the label aesthetic in Figure 6(b) to label only a few schools that are of interest to us. This is the conference consisting of Navy, Notre Dame, Utah State, Central Florida, and Connecticut, which is spread out, whereas most other conferences' teams are all very close to each other because they play within conference much more than they play out of conference. At the time, these five schools were all independents and did not have a home conference. Without the coloring capability, we would not have been able to pick out that difference as easily.

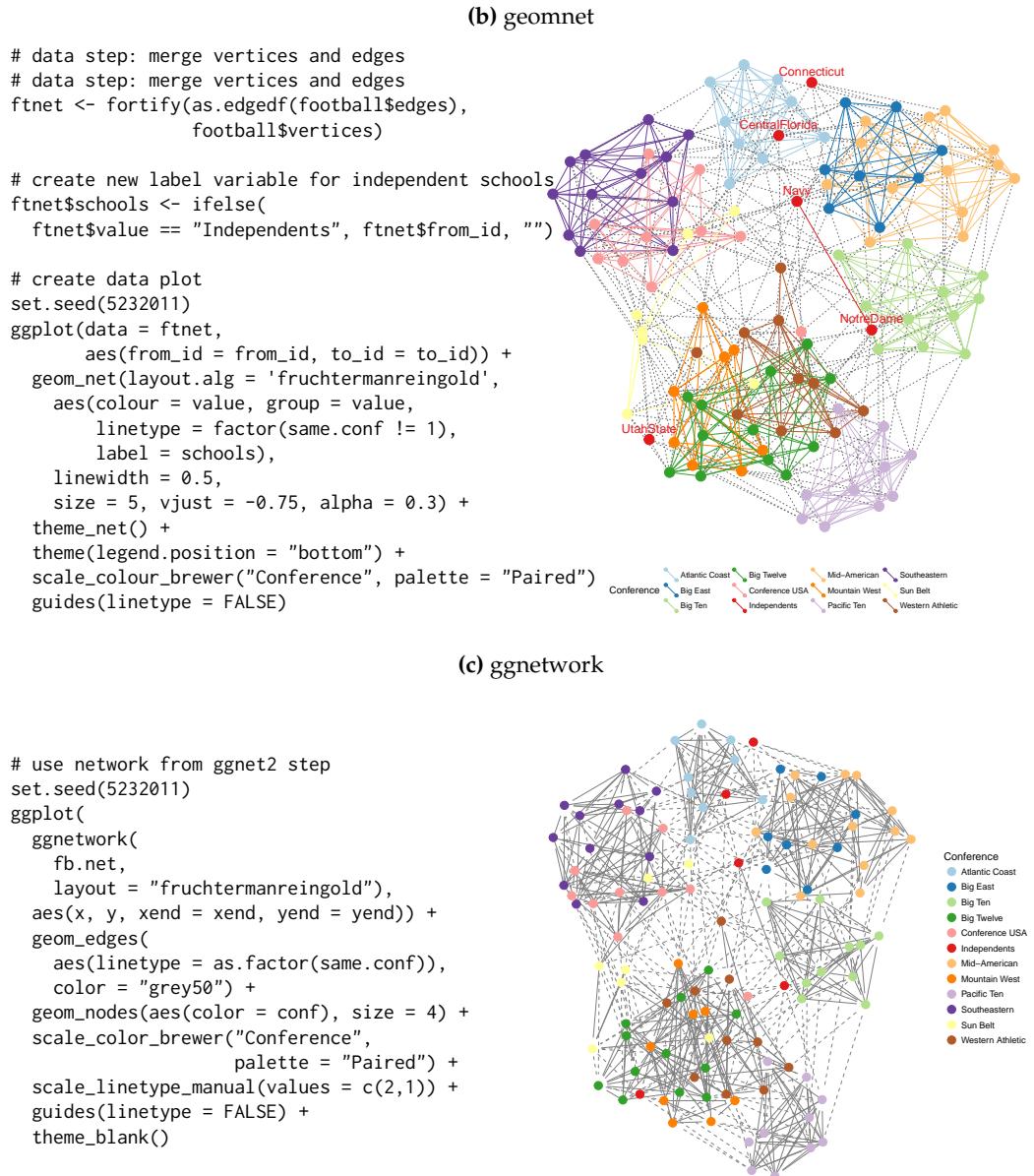


Figure 6: (continued) The network of regular season Division I college football games in the season of fall 2000. The vertices and their labels are colored by conference.

Southern women

Bipartite (or ‘two-mode’) networks are networks with two different kinds of nodes and where all ties are formed between these two kinds. Affiliation networks, which represent the ties between individuals and the groups to which they belong, are examples of such networks (see [Newman, 2010](#), p. 53-54 and p. 123-127).

One of the classic examples for a two-mode network is the network of 18 Southern women attending 14 social events as collected by [Davis et al. \(1941\)](#) and published e.g. as part of the `tnet` package ([Opsahl, 2009](#)). In this data, a woman is linked by an edge to an event if she attended it. One of the questions for these type of networks is gain insight in the interplay between the two different sets of nodes.

The data for the example of the Southern women is reported as edge list in form of ‘lady X attending event Y’. With a bit of data preparation as detailed below, we can visualize the graph as shown in Figure 7. In creating the plots, we use the shape and colour aesthetics to map the two different modes to two different shapes and colours.

```
# access the data and rename it for convenience
library(tnet)

data(tnet)
elist <- data.frame(Davis.Southern.women.2mode)
names(elist) <- c("Lady", "Event")
```

The edge list for the Southern women’s data consists of women attending events:

```
head(elist,4)
##   Lady Event
## 1     1     1
## 2     1     2
## 3     1     3
## 4     1     4
```

In order to distinguish between nodes from different types, we have to add an additional identifier element, so that we can tell the ‘first’ woman *L1* apart from the first event, *E1*.

```
elist$Lady <- paste("L", elist$Lady, sep="")
elist$Event <- paste("E", elist$Event, sep="")

davis <- elist
names(davis) <- c("from", "to")
davis <- rbind(davis, data.frame(from=davis$to, to=davis$from))
davis$type <- factor(c(rep("Lady", nrow(elist)), rep("Event", nrow(elist))))
```

The two different types of nodes are shown by different shapes and colors. We see the familiar relationship between events and groups of women attending these events. Women attending the same events then form a tighter knit subset, while events are also thought of as more similar, if they are attended by the same women. This defines the cluster of events E1 through E5, which are only attended by women 1 through 9, while events E6 through E9 are attended by (almost) everybody making them the core group of events.

Bike sharing in Washington D.C.

The data shows trips taken with bikes from the bike share company Capital Bikeshare⁵ during the second quarter of 2015. While this bike sharing company is located in the heart of Washington D.C. the company offers a set of bike stations just outside of Washington in Rockville, MD and north of it. Each station is shown as a vertex, and edges between stations indicate that at least five trips were taken between these two stations; the wider the line, the more trips have been taken between stations. In order to reflect distance between stations, we use as an additional restriction that the fastest trip was at most ten minutes long. Figure 8 shows four renderings of this data. The first is a geographically true

⁵<https://secure.capitalbikeshare.com/>

representation of the area overlaid by lines between bike stations, the other three are networks drawn with **geomnet**, **ggnet2**, and **ggnetwork**, respectively. The code for these renderings is shown below:

```
# make data accessible
data(bikes, package = 'geomnet')
# data step for geomnet
tripnet <- fortify(as.edgedf(bikes$trips), bikes$stations[,c(2,1,3:5)])
# create variable to identify Metro Stations
tripnet$Metro = FALSE
idx <- grep("Metro", tripnet$from_id)
tripnet$Metro[idx] <- TRUE

# plot the bike sharing network shown in Figure 7b
set.seed(1232016)
ggplot(aes(from_id = from_id, to_id = to_id), data = tripnet) +
  geom_net(aes(linewidth = n / 15, colour = Metro),
            labelon = TRUE, repel = TRUE) +
  theme_net() +
  xlim(c(-0.1, 1.1)) +
  scale_colour_manual("Metro Station", values = c("grey40", "darkorange")) +
  theme(legend.position = "bottom")

# data preparation for ggnet2 and ggnetwork
bikes.net <- network(bikes$trips[, 1:2 ], directed = FALSE)
# create edge attribute (number of trips)
network:::set.edge.attribute(bikes.net, "n", bikes$trips[, 3 ] / 15)
# create vertex attribute for Metro Station
bikes.net %v% "station" <- grepl("Metro", network.vertex.names(bikes.net))
bikes.net %v% "station" <- 1 + as.integer(bikes.net %v% "station")
rownames(bikes$stations) <- bikes$stations$name
# create node attributes (coordinates)
bikes.net %v% "lon" <-
  bikes$stations[ network.vertex.names(bikes.net), "long" ]
bikes.net %v% "lat" <-
  bikes$stations[ network.vertex.names(bikes.net), "lat" ]
bikes.col <- c("grey40", "darkorange")

# Non-geographic placement
set.seed(1232016)
ggnet2(bikes.net, mode = "fruchtermanreingold", size = 4, label = TRUE,
       vjust = -0.5, edge.size = "n", layout.exp = 1.1,
       color = bikes.col[ bikes.net %v% "station" ],
       label.color = bikes.col[ bikes.net %v% "station" ])

# Non-geographic placement. Use data from ggnet2 step.
set.seed(1232016)
ggplot(data = ggnetwork(bikes.net, layout = "fruchtermanreingold"),
       aes(x, y, xend = xend, yend = yend)) +
  geom_edges(aes(size = n), color = "grey40") +
  geom_nodes(aes(color = factor(station)), size = 4) +
  geom_nodetext(aes(label = vertex.names, color = factor(station)),
                vjust = -0.5) +
  scale_size_continuous("Trips", breaks = c(2, 4, 6), labels = c(30, 60, 90)) +
  scale_colour_manual("Metro station", labels = c("FALSE", "TRUE"),
                     values = c("grey40", "darkorange")) +
  theme_blank() +
  theme(legend.position = "bottom", legend.box = "horizontal")
```

To plot the geographically correct bike network layout in **geomnet**, we use the ‘`layout.alg = NULL`’ option and provide the latitude and longitude coordinates of the bike stations from the company’s data. A glance of the data that we used in the examples is shown below.

```
bikes.net
## Network attributes:
##  vertices = 20
##  directed = FALSE
##  hyper = FALSE
##  loops = FALSE
##  multiple = FALSE
##  bipartite = FALSE
##  total edges= 53
##  missing edges= 0
##  non-missing edges= 53
##
##  Vertex attribute names:
##    lat lon station vertex.names
##
##  Edge attribute names:
##    n
head(tripnet[,-c(4:5,8)])
##          from_id
## 1 Broschart & Blackwell Rd
## 2 Crabbs Branch Way & Calhoun Pl
## 3 Crabbs Branch Way & Calhoun Pl
## 4 Crabbs Branch Way & Calhoun Pl
## 5 Crabbs Branch Way & Calhoun Pl
## 6 Crabbs Branch Way & Calhoun Pl
##          to_id  n      lat      long
## 1      <NA> NA 39.10210 -77.20032
## 2 Crabbs Branch Way & Redland Rd 11 39.10771 -77.15207
## 3 Needwood Rd & Eagles Head Ct 14 39.10771 -77.15207
## 4      Rockville Metro East 51 39.10771 -77.15207
## 5      Rockville Metro West  8 39.10771 -77.15207
## 6 Shady Grove Metro West 36 39.10771 -77.15207
##  Metro
## 1 FALSE
## 2 FALSE
## 3 FALSE
## 4 FALSE
## 5 FALSE
## 6 FALSE
```

Because all three approaches result in the same picture, we only show one of these in Figure 8a. The code for creating the map is given here:

```
library(ggmap)
metro_map <- get_map(location = c(left = -77.22257, bottom = 39.05721,
                                   right = -77.11271, top = 39.14247))

# geomnet: overlay bike sharing network on geographic map
ggmap(metro_map) +
  geom_net(data = tripnet, layout.alg = NULL, labelon = TRUE,
           vjust = -0.5, ealpha = 0.5,
           aes(from_id = from_id,
               to_id = to_id,
               x = long, y = lat,
               linewidth = n / 15,
```

```

    colour = Metro)) +
  scale_colour_manual("Metro Station", values = c("grey40", "darkorange")) +
  theme_net() %>% replace% theme(aspect.ratio=NULL, legend.position = "bottom") +
  coord_map()

```

We can also make use of the option ‘`layout.alg = NULL`’ whenever we do not want to use an in-built layout algorithm but make use of a user-defined custom layout. In this case, the coordinates of the layout have to be created outside of the visualization and *x* and *y* coordinates have to be made available instead.

Some considerations of speed

In our examples thus far, we have focused on rather small social or relationship networks and one larger communication network. Now we present an example of a biological network, which comes from Jeong et al. (2001). It is the complete protein-protein interaction network in the yeast species *S. cerevisiae*. There are 2,113 proteins that make up the vertices of this network, with a total of 4480 edges between them. These edges represent “direct physical interactions” between any two proteins (Jeong et al., 2001, p. 42), resulting in a relatively large network. When these interactions and their associated proteins are plotted using the Fruchterman-Reingold layout algorithm, the runtime is extremely long, about 9.5 minutes for 50,000 iterations through the algorithm. The resulting layout is shown in Figure 9. When testing the three approaches with the larger network, we decided to use a random layout to save time. Despite its size, each one of the approaches in the **ggplot2** framework can be drawn in a few hundred milliseconds.

Another benefit that emerges from using **ggplot2** for network visualization is the speed at which it can plot fairly large networks. In order to assess the speed gain procured by our three approaches, we ran two separate tests, both of which designate **ggplot2**-based approaches as faster than the plotting functionality offered in the **network** package. They also show the **ggplot2** approaches to be largely on par with the speed provided by the **igraph** package. We first investigate average random layout plotting time of the protein network

shown in Figure 9, and then consider average plotting times of increasingly larger random networks. Note that in all tests, default package settings were used. The code to create benchmark results for both of these situations is provided in the vignette of the package **ggCompNet** (Tyner and Hofmann, 2016b). See the Supplementary Material section at the end of this paper for more information.

We plotted the protein interaction network of Figure 9 100 times using the **network** and **igraph** packages, and compared their run times to 100 runs each of the three visualization approaches introduced in this paper. The results are shown in Figure 10. We can see that on average, the **ggplot2** framework provides a two to three-fold increase in speed over the **network** package, and that **geomnet** and **ggnetwork** are faster than package **igraph**. The three **ggplot2** approaches also have considerably less variability in time than the **network** package. Despite the large number of vertices, the protein interaction network has a relatively small number of edges (4480 out of over 2.2 million theoretically possible connections resulting in an edge probability of just over 0.0020). Next, we examine networks with a higher edge probability.

The second test relies on random undirected networks in which the probability of an edge between two nodes was set to $p = 0.2$. We generated 100 of these networks at network sizes from 25 to 250 nodes, using increments of 25.

Figure 11 summarizes the results of these benchmarks using a convenience sample of machines accessible to the authors, including authors’ hardware and additional results from friends’ and colleagues’ machines. Network sizes are plotted horizontally, execution times of 100 runs under each visualization approach are plotted on the *y*-axis. Each panel shows a different machine as indicated by the facet label. Note that each panel is scaled separately to account for differences in the overall speed of these machines. What these plots indicate is that we have surprisingly large variability in relative run times across different machines. However, the results support some general findings. The **network** plotting routine is by far the slowest across all machines, while the **igraph** plotting is generally among the fastest. Our three approaches generally feature in between **igraph** and **network** with **ggnetwork** being as fast or faster than **igraph** plotting, followed by **ggnetwork** and **geomnet**, which is generally the slowest among the three. These differences become more pronounced as the size of the network increases.

Although speed was not the main rationale for our inquiry into **ggplot2**-based approaches to network visualization, a speed-based comparison shows a clear advantage of these approaches over

the plotting function included in the **network** package, which very quickly becomes much slower as network size increases.

Summary and discussion

At first glance, the three visualization approaches may seem nearly identical. However, each one brings unique strengths to the visualization of networks. Out of our three approaches, **ggnetwork** is most flexible and allows for a re-ordering of layers to emphasize one over the other. The flexibility is useful but does require the user to specify every single part of the network visualization. The **geomnet** implementation most closely aligns with the existing **ggplot2** paradigm because it provides a single layer that can be added to other **ggplot2** layers. **ggnet2** requires the user to know the least about the **ggplot2** framework, while resulting in a valid and extensible **ggplot2** object. Many features of the packages would not have been possible, or would have at least been difficult to implement, in prior versions of **ggplot2**. The increased flexibility of the current development version as well as the added geoms `geom_curve` and `geom_label` provided us with a strong, yet flexible, foundation for network visualization. Our approaches also benefit from the speed of **ggplot2**, making network visualization more efficient than the existing framework of **network** for a lot of the benchmark examples.

All three approaches rely on the package **sna** for layouts. This allows the user to access the many layout algorithms available for networks, and in the event that new layouts are implemented in **sna**, our packages will accommodate them seamlessly. A larger range of layouts is available through **igraph**, and can be implemented into our packages by setting the respective layout arguments to `NULL` and passing `x`, `y` coordinates calculated from **igraph**. There are some notable differences between the packages, such as in the parameters used for specific layout algorithms, e.g. **igraph** allows the use of weights for Fruchterman-Reingold placement, even though it is unclear from the original article how these are supposed to affect the layout. In all three approaches, it is feasible to tap into **igraph**'s functionality in a future version so that the user does not need to calculate the layout separately. Additional future work will explore the implementation of other network data structures, such as the `networkDynamic` class from **statnet**, which would benefit from the faceting capabilities of our implementations. This work will likely incorporate the `fortify` approach of **ggnetwork** and `geomnet::fortify.network()` for converting network data structures to a **ggplot2**-friendly format.

We have found that none of our approaches is unequivocally the best. We can, however, provide some guidance as to which approach is best for which type of user. The main differences between the three methods are in the way that network information is passed into the functions. For **ggnet2** and **ggnetwork**, data management and attribute handling is done through network operators on nodes and edges, while the **geomnet** approach does not require any knowledge of networks or existing network analysis packages from the user. This likely affects the user base of each package. We think that users who are well-versed with networks will find **ggnet2** and **ggnetwork** more intuitive to use than **geomnet**. These users might be looking to **ggplot2** as another avenue to create high-quality visualizations that tap into **ggplot2** advantages such as facetting and, for **ggnetwork**, layering. Users who are already familiar with **ggplot2** and some of the other **tidyverse** packages (see [Wickham \(2017\)](#)), and who find themselves dealing with network data will likely be more attracted to the **geomnet** implementation of network plotting. The data management skills needed for using **geomnet** are basic: some familiarity with the split-apply-combine paradigm, in the form of familiarity with **plyr** or **dplyr**, would be sufficient in order to make full use of the features of `geom_net` ([Wickham, 2011](#)). All in all, the three approaches we have presented here provide a wealth of resources to users of all skill sets who are looking to create beautiful network visualizations.

On a personal level we discovered that the collaboration on this paper has helped us to improve upon our initial versions of each of these packages. For instance, the edge coloring in the **ggnet2** function was designed so that edges between two vertices in the same group were colored with that group's vertex color. This inspired an implementation of it in **geomnet** through the traditional **ggplot2** group operator. During the process of writing the paper the authors collaborated on a solution for the problem of nodes being plotted on top of arrow tips. This solution was implemented in the **geomnet** `arrow.gap` parameter, which allows to re-track the tip of an arrow on a directed edge, and was also added to **ggnetwork**. In addition, the implementation of a **ggplot2** geom for networks within **geomnet** inspired the creation of the aliased geoms of the **ggnetwork** package.

Finally, curious users may be interested in how these three packages can fit together and replicate each other, since they are in fact so similar. Thanks to the flexibility inherent to **ggnetwork**, it is possible to write wrapper functions around **ggnetwork** functions in order to recreate the behavior and functionality of **ggnet2** and **geomnet**. Simple examples of such wrapper functions, called **ggnetwork2** and **geom_network**, respectively are shown below.

```
library(ggnetwork)
```

```
# mimics geom_net behavior
geom_network <- function(edge.param, node.param) {
  edge_ly <- do.call(geom_edges, edge.param)
  node_ly <- do.call(geom_nodes, node.param)
  list(edge_ly, node_ly)
}
# mimics ggnet2 behavoir
ggnetwork2 <- function() { ggplot() + geom_network() }
```

Similarly, **geomnet** can mimic the the behavior of **ggnet2**, as shown below.

```
library(geomnet)
geomnet2 <- function(net) {
  ggplot(data = fortify(net),
         aes(from_id = from_id, to_id = to_id)) +
    geom_net()
}
```

Mimicking **ggnetwork** with **geomnet** requires a little bit more work because the native data input for **geomnet** is a "data.frame" object fortified with **geomnet** methods, not a "network" object. Instead, the internal **ggplot2** function **ggplot_build** allows a plot created with **geomnet** function calls to be recreated with **ggnetwork**-like syntax. An example of using a **geomnet** plot to create a similar plot in the style of **ggnetwork** follows to reproduce Figure 2(c).

```
library(geomnet)
library(ggnetwork)
library(dplyr)
# a ggnetwork-like creation using a geomnet plot
data("blood")
# first, create the geomnet plot to access the data later
geomnetplot <- ggplot(data = blood$edges, aes(from_id = from, to_id =
                                             to)) +
  geom_net(layout.alg = "circle", selfloops = TRUE) +
  theme_net()
# get the data
dat <- ggplot_build(geomnetplot)$data[[1]]
# ggnetwork-like construction for re-creating network shown in Figure 5
ggplot(data = dat, aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_segment(arrows = arrow(type = 'closed'), colour = 'grey40') +
  geom_point(size = 10, colour = 'darkred') +
  geom_text(aes(label = from), colour = 'grey80', size = 4) +
  geom_circle() +
  theme_blank() + theme(aspect.ratio = 1)
```

Supplementary Material

Software: **ggnetwork** 0.5.1 and **geomnet** 0.2.0 were used to create the visualizations. **ggnet2** is part of **GGally** 1.3.0.

Reproducibility: All the code used in the examples is available as a vignette in the CRAN package **ggCompNet**. There are two vignettes: one for the speed comparisons and one for the visualizations provided in the Examples section. The package also provide our speed test data for creating Figure 11. We created this package to accompany this paper with the hope that interested users will compare these methods on their own systems and against their own code. Finally, all of the data we use in the examples, with the exception of the bipartite network example, is included as a part of the **geomnet** package.

Acknowledgements

The authors would like to thank the reviewers for their thoughtful input to and thorough reviews of our manuscript. We would also like to thank the editor of The R Journal for his enduring patience.

Bibliography

- B. Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2016. URL <https://CRAN.R-project.org/package=gridExtra>. R package version 2.2.1. [p³⁹]
- S. Bender-deMoll. *Ndtv: Network Dynamic Temporal Visualizations*, 2016. URL <https://CRAN.R-project.org/package=ndtv>. R package version 0.10.0. [p³⁹]
- M. Bojanowski. *Intergraph: Coercion Routines for Network Data Objects*, 2015. URL <http://mbojan.github.io/intergraph>. R package version 2.0-2. [p³⁰]
- F. Briatte. *Ggnetwork: Geometries to Plot Networks with 'ggplot2'*, 2016. URL <https://github.com/briatte/ggnetwork>. R package version 0.5.1. [p^{27, 29}]
- S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010. [p²⁸]
- C. T. Butts. network: a Package for Managing Relational Data in R. *Journal of Statistical Software*, 24(2), 2008. [p²⁷]
- C. T. Butts. *Sna: Tools for Social Network Analysis*, 2014. URL <http://CRAN.R-project.org/package=sna>. R package version 2.3-2. [p²⁷]
- C. T. Butts, M. S. Handcock, and D. R. Hunter. *Network: Classes for Relational Data*. Irvine, CA, 2014. URL <http://statnet.org/>. R package version 1.10.2. [p^{27, 31}]
- W. Chang. *Gcookbook: Data for "R Graphics Cookbook"*, 2012. URL <https://CRAN.R-project.org/package=gcookbook>. R package version 1.0. [p²⁸]
- W. Chang. *R Graphics Cookbook*. O'Reilly, Sebastopol, CA, 2013. ISBN 978-1449316952. [p²⁸]
- K. Cook, G. Grinstein, and M. Whiting. VAST Challenge 2014. <http://hcil2.cs.umd.edu/newwarepository/benchmarks.php>, 2014. [p³⁷]
- G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*:1695, 2006. URL <http://igraph.org>. [p²⁷]
- A. Davis, B. B. Gardner, and M. R. Gardner. *Deep South: A Social Anthropological Study of Caste and Class*. The University of Chicago Press, Chicago, IL, 1941. [p⁴⁷]
- T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991. [p^{30, 34}]
- M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002. [p^{28, 44}]
- M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris. Statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11, 2008. URL <http://www.jstatsoft.org/v24/i01>. [p²⁷]
- M. Horikoshi and Y. Tang. *Ggfortify: Data Visualization Tools for Statistical Analysis Results*, 2016. URL <http://CRAN.R-project.org/package=ggfortify>. R package version 0.4.1. [p²⁸]
- H. Jeong, S. P. M. A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001. [p⁵⁰]
- B. H. Junker and F. Schreiber. *Analysis of Biological Networks*. Wiley Series in Bioinformatics. John Wiley & Sons, 2008. ISBN 9780470253465. URL <https://books.google.com/books?id=2DloLXaXSNgC>. [p²⁷]
- D. Kahle and H. Wickham. Ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013. URL <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>. [p²⁸]
- T. Kamada and S. Kawai. An Algorithm for Drawing General Undirected Graphs. *Information Processing Letters*, 31(1):7–15, 1989. [p³²]
- M. E. J. Newman. *Networks : An Introduction*. Oxford University Press, Oxford New York, 2010. ISBN 978-0199206650. [p^{28, 47}]
- T. Opsahl. *Structure and Evolution of Weighted Networks*. University of London (Queen Mary College), London, UK, 2009. URL <http://toreopsahl.com/publications/thesis/>. [p⁴⁷]

- C. Prell. *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2011. ISBN 9781446290132. URL <https://books.google.com/books?id=wZYQAgAAQBAJ>. [p27]
- B. Schloerke, J. Crowley, D. Cook, H. Hofmann, H. Wickham, F. Briatte, M. Marbach, and E. Thoen. *GGally: Extension to Ggplot2.*, 2016. R package version 1.3.0. [p28, 29]
- K. Slowikowski. *Ggrepel: Repulsive Text and Label Geoms for 'ggplot2'*, 2016. URL <https://CRAN.R-project.org/package=ggrepel>. R package version 0.5. [p33]
- R. Tamassia, editor. *Handbook of Graph Drawing and Visualization*. CRC Press, 2013. [p27]
- Y. Tang, M. Horikoshi, and W. Li. Ggfortify: Unified interface to visualize statistical result of popular r packages. *The R Journal*, 2016. URL <http://CRAN.R-project.org/package=ggfortify>. [p28]
- S. Tyner and H. Hofmann. *Geomnet: Network Visualization in the 'ggplot2' Framework*, 2016a. URL <http://github.com/sctyner/geomnet>. R package version 0.2.0. [p27]
- S. Tyner and H. Hofmann. *ggCompNet: Compare Timing of Network Visualizations*, 2016b. URL <https://CRAN.R-project.org/package=ggCompNet>. R package version 0.1.0. [p50]
- D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. [p28]
- D. J. Watts. The "New" Science of Networks. *Annual Review of Sociology*, 30:243–270, 2004. [p27]
- H. Wickham. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1): 1–29, 2011. URL <http://www.jstatsoft.org/v40/i01/>. [p51]
- H. Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, 2016. ISBN 978-3-319-24277-4. URL <http://ggplot2.org>. [p27, 34, 42]
- H. Wickham. *Tidyverse: Easily Install and Load 'tidyverse' Packages*, 2017. URL <https://CRAN.R-project.org/package=tidyverse>. R package version 1.1.1. [p51]
- L. Wilkinson. *The Grammar of Graphics*. Springer-Verlag, New York, 1999. [p27]
- T. Yin, D. Cook, and M. Lawrence. Ggbio: An R package for extending the grammar of graphics for genomic data. *Genome Biology*, 13(8):R77, 2012. [p28]
- G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017. ISSN 2041-210X. URL <https://doi.org/10.1111/2041-210x.12628>. [p28]

Samantha Tyner
Department of Statistics and Statistical Laboratory
Iowa State University
United States
sctyner@mail.iastate.edu

François Briatte
European School of Political Sciences
Catholic University of Lille
France
francois.briatte@univ-catholille.fr

Heike Hofmann
Department of Statistics and Statistical Laboratory
Iowa State University
United States
hofmann@mail.iastate.edu

```
# Southern women network in ggnet2
# create affiliation matrix
bip = xtabs(~Event+Lady, data=elist)

# weighted bipartite network
bip = network(bip,
               matrix.type = "bipartite",
               ignore.eval = FALSE,
               names.eval = "weights")

# detect and color the mode
set.seed(8262013)
ggnet2(bip, color = "mode", palette = "Set2",
       shape = "mode", mode = "kamadakawai",
       size = 15, label = TRUE) +
  theme(legend.position="bottom")
```

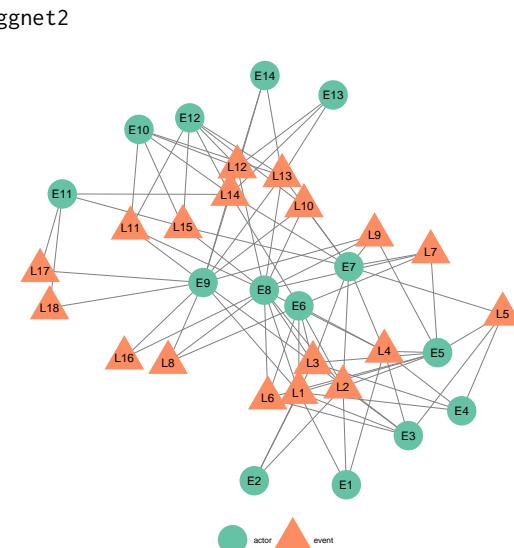


Figure 7: Graph of the Southern women data. Women are represented as orange triangles, events as green circles.

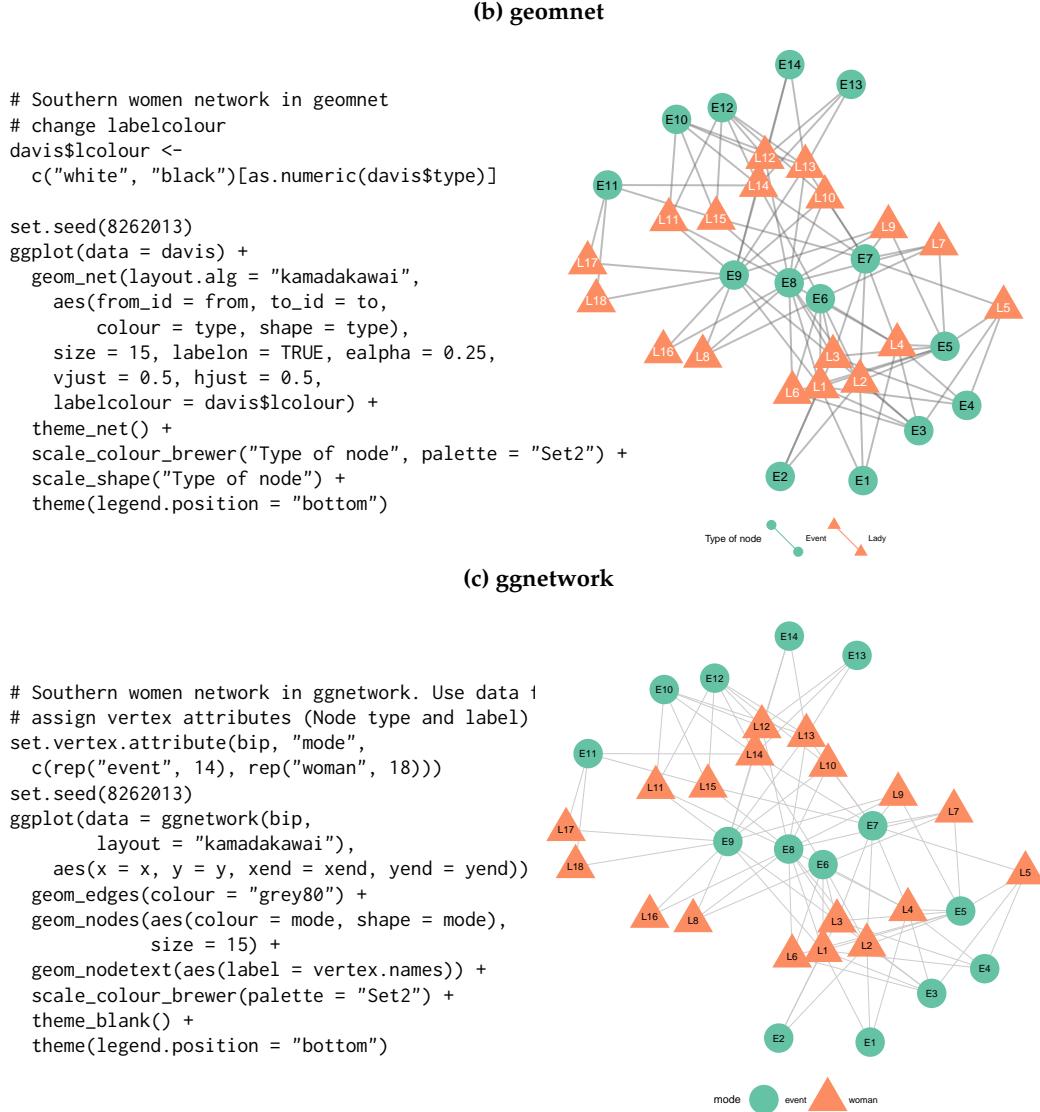


Figure 7: Graph of the Southern women data. Women are represented as orange triangles, events as green circles.

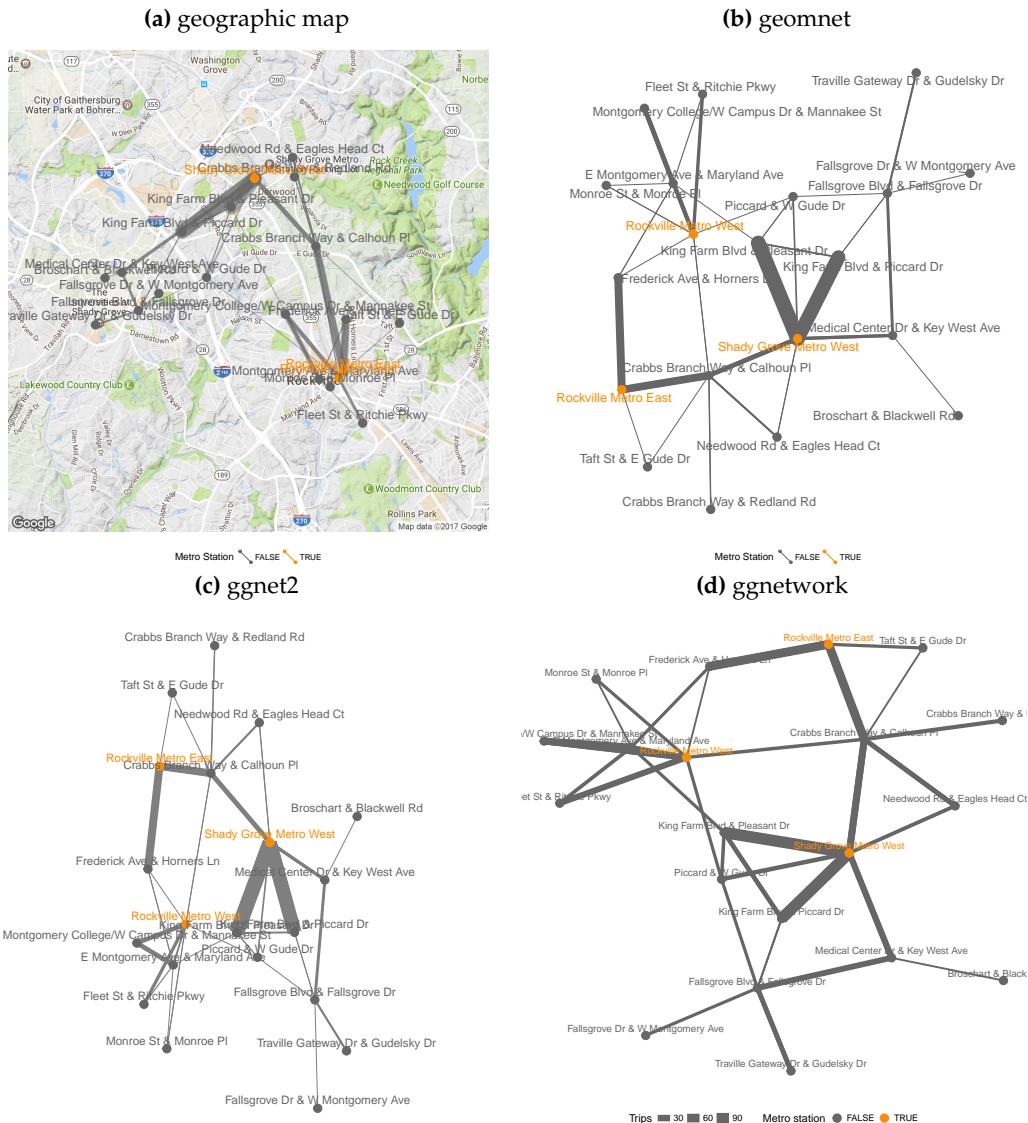


Figure 8: Network of bike trips using a geographically true representation (top left) overlaid on a satellite map, a Kamada-Kawai layout in **geomnet** (top right), a Fruchterman-Reingold layout in **ggnet2** (bottom left) and **ggnetwork** (bottom right). Metro stations are shown in orange. In both the Kamada-Kawai and the Fruchterman-Reingold layouts, metro stations take a much more central position than in the geographically true representation.

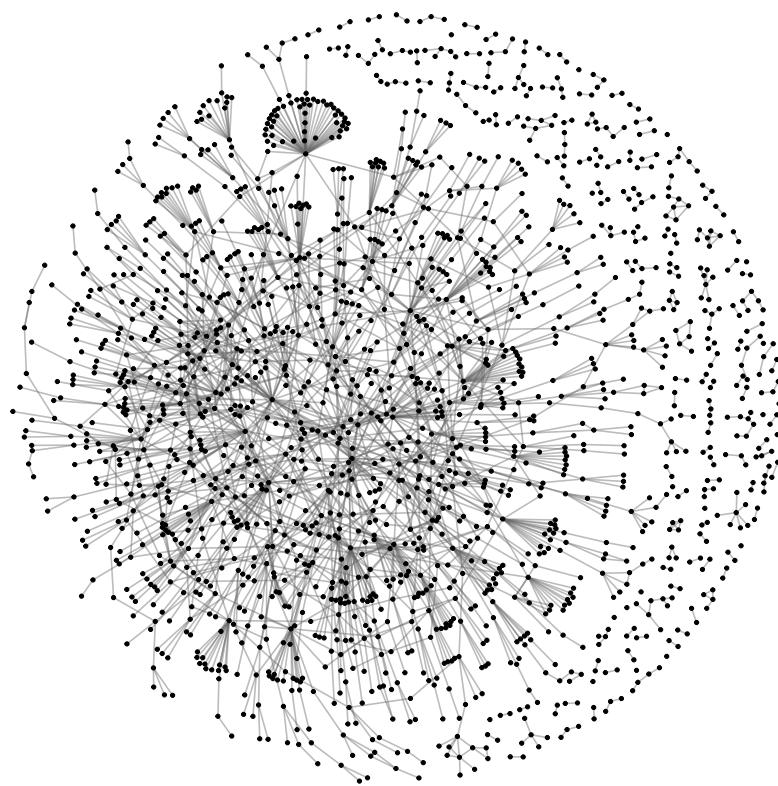


Figure 9: Protein-protein interaction network in *S. cerevisiae*. A Fruchterman-Reingold algorithm allowed to run for 50,000 iterations produced the coordinates for the nodes.

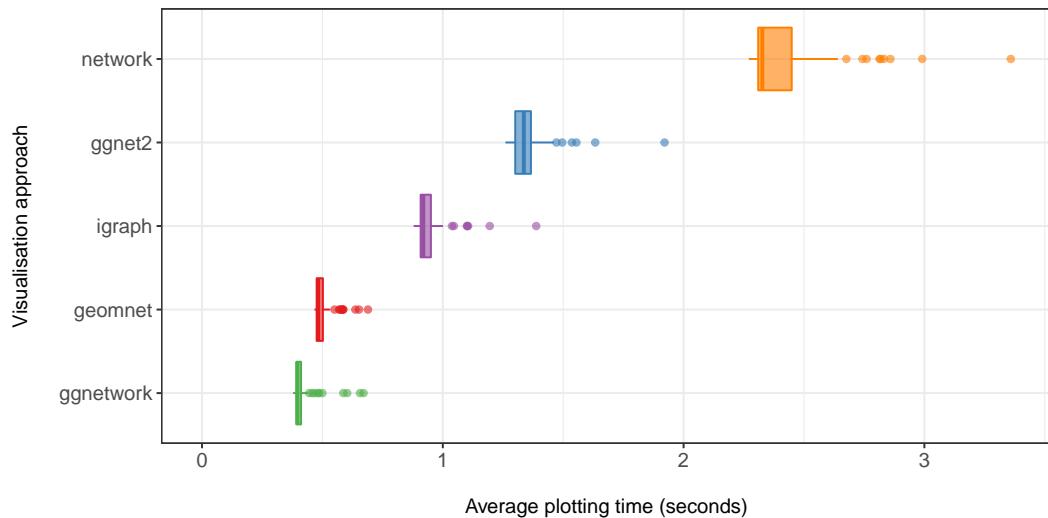


Figure 10: Comparison of the times needed for calculating and rendering the previously discussed protein interaction network in the three **ggplot2** approaches and the standard plotting routines of the **network** and **igraph** packages based on 100 evaluations each.

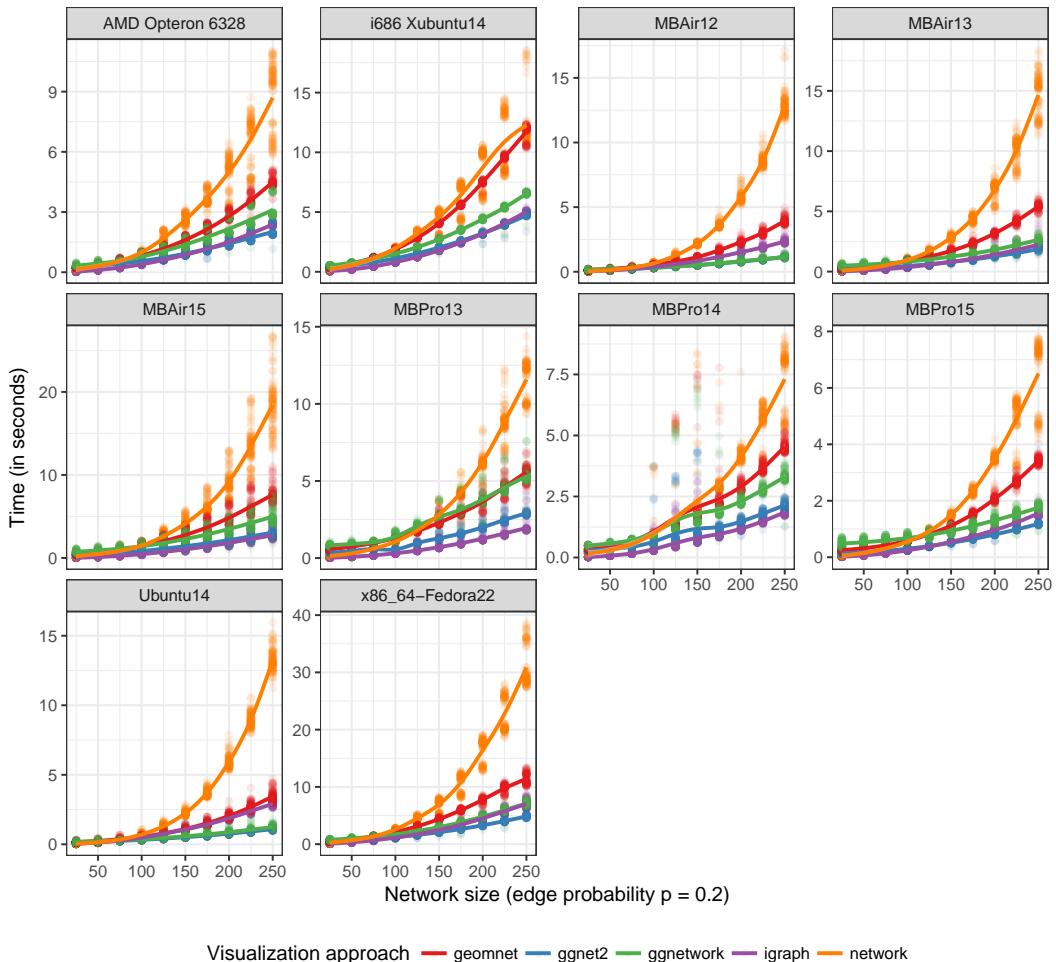


Figure 11: Plotting times of random undirected networks of different sizes under each of the available visualization approaches using their default settings. Note that each panel is scaled independently to highlight relative differences in the visualization approaches rather than speed of different hardware.

A COMPARISON OF METHODS FOR DYNAMIC ORIGIN-DESTINATION MATRIX ESTIMATION

Nanne van der Zijpp

Faculty of Civil Engineering, Delft University of Technology

P.O.Box 5048 2600 GA Delft NL

Email zijpp@ct.tudelft.nl

Abstract: An overview of existing dynamic OD-estimation methods is given, and a statistical technique to estimate parameters in dynamic travel demand models based on Bayesian inference is proposed. Using both synthesized and empirical data the new technique is compared to traditional ones such as the Kalman filter and least squares. This shows the new technique results in more accurate estimates of OD-split proportions. Surprisingly however, this does not result by default in more accurate predictions of link flow volumes.

Key-words: Dynamic OD-matrix estimation; Kalman filter; Bayesian Inference

1. INTRODUCTION

On-line prediction and control of freeway traffic usually requires estimates of time varying travel demand. Time varying travel demand is summarized in dynamic origin-destination (OD) matrices, and the estimation of these has been an active area of research over the last two decades. The estimation of OD-matrices from traffic counts may be considered as the reversal of another well-known problem, that of traffic assignment. In general, the unique reversal of traffic assignment is not possible, as many OD-matrices match a given set of traffic counts. However, there are various possibilities to define a measure of plausibility over the space of OD-matrices. By doing so, a 'best' dynamic OD-matrix is implied.

For dynamic matrices it is plausible to base such a measure on the similarity between the matrices corresponding to consecutive periods, expressing that traffic emerges as a result of slowly evolving travel demand. The methods that are central to the present paper are based on a distance measure that applies to the structure of the OD-matrix, as proposed in e.g. (Cremer and Keller, 1981; Nihan and Davis, 1987). This structure is expressed by the OD-split-matrix which is obtained from an OD-matrix by dividing each OD-matrix cell-value by its corresponding row-total. The resulting cell-values are referred to as split

proportions and express for each network entry the fraction of traffic that is destined for each network exit. As a result split proportions are non-negative and smaller than unity. Moreover, the split proportions within one row of the OD-split matrix, i.e. those corresponding to one network entry, add up to unity.

The various approaches that have been presented over the years may be categorized depending on the traffic characteristics that are taken into account, the data-sources that are used and the estimation procedures that are employed.

With respect to traffic characteristics, the problem may include the estimation of route choice proportions, travel time (see e.g. Chang and Wu, 1994), and travel time dispersion (see e.g. Bell, 1991). With respect to traffic data, these may consist of time-series of traffic counts only (see Cremer and Keller, 1991, 1987; Nihan and Davis, 1987; VanderZijpp and Hamerslag, 1994), or may be supplemented with historical data or with observations from probe-vehicles or license-plate readers (see VanderZijpp, 1996, 1997). Thirdly, the estimation procedure that is employed may be based on neural networks, least squares, maximum likelihood, Kalman filter techniques, or Bayesian inference.

The present paper attempts to isolate the influence of

the estimation procedure that is used, and presents a comparison between least squares, constrained least squares, the Kalman filter and a Bayesian updating algorithm that was presented in (VanderZijpp, 1996).

The latter Bayesian estimator was developed especially to deal with the non-negativity constraints that apply to the split proportions. Due to these constraints the probability distribution, of which the estimated split proportions should reflect the mean, have a typical asymmetric shape. Therefore, it is hypothesised that a significant improvement over the traditional methods can be obtained if instead of the point that maximizes the probability, the centre of the probability mass can be computed.

The Bayesian method is compared to other methods, such as least squares, constrained least squares and the Kalman filter, on the basis of both synthesized and empirical data. The latter data were extracted from the traffic monitoring system on the Amsterdam beltway, but unfortunately are not accompanied by a directly observed dynamic OD-matrix. The evaluation on the basis of empirical data gives rise to the interesting methodological issue of how to compare different methods in absence of a directly observed OD-table.

2. PROBLEM DEFINITION

The problem is to determine a dynamic OD matrix for a transport network. It is assumed that time series of observations of all entering volumes and a subset of the internal link volumes are available. The following symbols will be used, sizes of vectors and matrices are implicit in their definitions:

m, n, h	Number of entries, number of exits, and number of link volume observations (excluding entry volumes).
i, j, k	Indices corresponding to entries, exits, and link volume observations
$q(t)$	Vector of entry volume observations in period t , $t=1,2,\dots$
$f_{ij}(t)$	Number of trips for entry-exit (EE) pair $i-j$ with departure period t , $i=1,\dots,m$, $j=1,\dots,n$.
$y(t)$	Vector of link volume observations in period t . Element k of this vector contains the number of trips with departure period t that traverse location k .
τ	Path-link observation incidence map. $\tau_{ijk}=1$ if route $i-j$ uses link k and zero otherwise, $k=1,2,\dots,h$

3. CLASSIFICATION OF DYNAMIC OD-ESTIMATION METHODS

Methods described in literature for estimating dynamic OD-matrices can be arranged in a number of

ways, depending on their underlying model assumptions, the estimation techniques used, data used and details of their implementation.

3.1 Partitioning of the time-axis.

Data are usually available in time series of observations. They may be aggregated in time, e.g. to one or five minutes. Often the time-span to which a data record refers must be inferred from a time-stamp that accompanies each record.

Variables used in models corresponding to these observations need not necessarily use the same partitioning of the time axis. For practical reasons one may wish to convert observed data to a new time coordinate system. Usually this is done by using a regular grid that applies to all locations in the network. However, there are some practical advantages attached to using a *moving time coordinate system* (MTCS), in which the partitioning of the time-axis depends on the location. In a MTCS the boundaries between periods are given by time-space trajectories (Van Der Zijpp, 1996). This means that vehicles largely travel within one time-zone, and consequently only contribute to observations assigned to that period.

3.2 Dealing with travel time

The fact that vehicles need some time to travel through a network can be dealt with in a number of ways. A first possibility is to ignore travel times. This limits the applicability of models to small networks, as the ratio travel time / period length must not exceed a certain number. A second possibility is to take travel times into account, but to assume that they are known, for example from a traffic monitoring system. A third, yet unexplored, possibility would be to consider the link travel times as an extra set of unknowns to the problem. In the present paper we use the second approach, i.e. we assume that travel times are known and do not attempt to estimate them.

3.3 Dealing with travel time dispersion

Further refinements are obtained by taking travel-time dispersion into account. This can be done either by assuming a certain distribution of travel times, see e.g. (Bell, 1991b) or by introducing extra unknowns and (Chang and Wu, 1994). Such approaches require the estimation of extra parameters and hence do not lead to a reduced error of estimation by default. We do not consider them in the present paper.

3.4 Dealing with route choice

Also the estimation of route choice proportions introduces extra unknown parameters and may be avoided, for example by using a route choice model based on (estimated) travel times. In the present paper we avoid this issue by confining ourselves to corridors rather than networks.

3.5 Data requirements

Yet another way of subdividing dynamic OD-estimation methods is based on the input data that are used. These may be prior OD-matrices, time series of traffic counts or trajectory counts based on probe vehicles. Again we confine ourselves to a simple case where only time series of traffic counts are used.

3.6 Model assumptions

As discussed in the introduction, the problem of estimating dynamic OD-matrices from time-series of traffic counts is under-specified and extra assumptions are needed to define a unique solution. Imposing a model of travel demand is one option, but requires a certain level of temporal and spatial aggregation. This conflicts with the requirement of dynamic output. The remaining options are tracking OD-cell values and tracking OD-split proportions. From a methodological point of view the latter option is preferred: Due to the variation of entry volumes consecutive observations represent independent linear combinations of split proportions. This is not the case when OD-cell values are considered, as was done e.g. by (Ashok and Ben-Akiva, 1993). As stated in the introduction, the present paper deals with estimating split-proportions. Summarising, the model that is used in the paper is defined by the following equations:

$$E[f_{ij}(t)] = q_i(t) b_{ij}(t) \quad (1)$$

$$b_{ij}(t) = b_{ij}(t) + v_{ij}(t) \quad (2)$$

Where $b(t)$ denotes the vector of split proportions and $v(t)$ denotes a vector of small (zero mean) increments to these proportions.

4. OVERVIEW OF ESTIMATION METHODS

4.1 Least squares (DLS)

The simplest way to estimate OD matrices is by using (discounted) least squares:

$$\hat{b}(t) = \underset{b}{\operatorname{argmin}} J(b, t) \quad (3)$$

where the target J is defined by:

$$J(b, t) = \sum_{k=1}^t \lambda^{t-k} \|y(k) - H(k)b\|^2 \quad (4)$$

and the non-zero elements of measurement matrix H are given by:

$$H_{k, (i-1)n+j} = \tau_{ijk} \quad \forall ijk \quad (5)$$

The parameter λ , $0 \leq \lambda \leq 1$, determines the weight that is put on older observations. The advantage of the DLS method over other methods that are presented in the paper is that it also can be applied if not exit volumes are observed and its ease of implementation.

4.2 Discounted constrained least squares (DCLS)

The DCLS estimate for the split probabilities (*Nihan and Davis, 1987; Cremer and Keller, 1987*) is given by:

$$\hat{b}(t) = \underset{\substack{b \\ 0 \leq b \leq 1 \\ \pi' b = 1}}{\operatorname{argmin}} J(b, t) \quad (6)$$

where the non-zero elements of matrix π are given by:

$$\pi_{i, (i-1)n+j} = 1 \quad \forall ij \quad (7)$$

The advantage of the DCLS method is its superior performance relative to the DLS method.

4.3 Kalman filter (KF)

The Kalman filter is a recursive procedure that under certain assumptions produces unbiased minimum variance estimates. Although in general these assumptions will not be completely satisfied, the procedure has many advantages such as its flexibility and its ease of implementation.

To use the Kalman filter two additional matrices need to be specified: a first one (S_t) corresponding to the covariance matrix of the component $v(t)$ in the random walk equation $b(t+1) = b(t) + v(t)$, and a second one (R_t) corresponding to the error term $w(t)$ in the measurement equation $y(t) = H(t) b(t) + w(t)$. Guidelines how to choose S_t and R_t are given in (Vander Zijpp, 1996). The KF update equations are given by:

$$\begin{aligned} \hat{b}(t) &= \hat{b}(t-1) + K_t [y(t) - H(t)\hat{b}(t-1)] \\ K_t &= \Sigma_b(t) H'(t) [H(t)\Sigma_b(t)H'(t) + R_t]^{-1} \\ \Sigma_b(t+1) &= \Sigma_b(t) + S_t - \Sigma_b(t)H'(t) \dots \\ &[H(t)\Sigma_b(t)H'(t) + R_t]^{-1} H(t)\Sigma_b(t) \end{aligned} \quad (8)$$

4.4 Bayesian updating (BU)

A traditional interpretation of the Kalman filter is the Bayesian one. Here it is assumed that a multivariate normal (MVN) prior distribution of $b(t)$ is given and is updated with the information contained in the MVN distributed measurements $y(t)$. In this interpretation the vector $\hat{b}(t)$ and matrix $\Sigma_b(t)$ are the mean and covariance of the MVN distribution of $b(t)$.

However, another Bayesian interpretation is also possible. Assume that the prior distribution is not MVN but *truncated* MVN, i.e. has a shape identical to MVN, but is confined to the interval $b(t) \in [0,1]$, see Fig. 1. For this case it can also be shown that performing a Bayesian update with an MVN distributed observation $y(t)$ results in a TMVN posterior distribution (see VanderZijpp and Hamerslag, 1994). Moreover, the resulting distribution is still characterized by the parameters given by the recursion (8).

In this interpretation the vector $\hat{b}(t)$ is no longer the mean and therefore not the best possible point estimate for $b(t)$. In fact there is no tractable expression to compute the mean as this would require evaluating a high dimensional integral. Instead a practical approach is to compute the average of a large set of numbers sampled from the truncated MVN distribution.

5. EXPERIMENTS

To give an impression of the performance of the methods described in this paper a number of experiments were done. A first series of experiments was done using synthesized data. In a second series of experiments empirical data from the Amsterdam beltway were used.

5.1 Experiments with synthesized data

Test-data. In a first series of experiments test-data

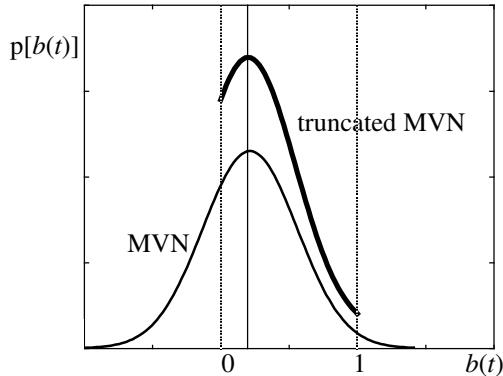


Fig. 1. A normal distribution (below) and its corresponding truncated distribution.

are generated according to six preset specifications. The test-data comprise entry-exit (EE) flows as well as synthesized traffic counts. The generation of the test-data involves the following steps:

1. *Generation of the split probabilities.* The split probabilities are generated by the random walk model $b(t+1) = b(t) + v(t)$ and initialized with a randomly generated vector $b(0)$. The following specification for the covariance matrix of $v(t)$ (S_t) is used:

$$S_t = \sigma_b^2 I \quad (9)$$

2. *Generation of the entry flow volumes.* The entry volumes are sampled from a normal distribution of which mean and variance equal a parameter \bar{q} .
3. *Generation and assignment of the EE-flows.* The EE-flows are sampled from the multinomial distribution, using the probabilities generated under step -1-.
4. *Generation of the entry- and link volume observation errors.* To the generated entry-flows and link-flows an zero mean error is added with covariance matrices given respectively:

$$\Phi = \sigma_q^2 I \quad (10)$$

$$\Theta = \sigma_y^2 I \quad (11)$$

In order to reduce random effects, for every set of parameters described in table 1, ten independent data-sets are generated. Each estimation method will be applied to all of these sets, after which the errors of estimation will be averaged. Each set consist of 48 periods and was generated bearing in mind a period length of ten minutes. Network 1 represents the default configuration. Each of the specifications 2-6 differs only in one parameter from this configuration.

Table 1: Network specifications
-values of parameters-

	Network					
	1	2	3	4	5	6
m	4	4	4	4	4	6
n	4	4	4	4	4	6
σ_b^2	10^{-4}	10^{-2}	0	10^{-4}	10^{-4}	10^{-4}
\bar{q}	100	100	100	100	100	100
σ_q^2	100	100	100	10	100	100
σ_y^2	100	100	100	100	10	100

Estimation methods. The methods that are compared are divided in the following categories:

1. The Least squares method (LS) see equation (3). After some experimenting, the discounting param-

- eter λ was set to $1 - \sigma_b^2$.
2. The discounted constrained least squares method (DCLS), see equation (6).
 3. The Kalman filter(KF) method, see equation (8).In accordance with the way the test-data are generated the matrix S_t is set to $\sigma_b^2 I$. The observation error covariance matrix R_t is set to a diagonal matrix with the average observed values on its diagonal:
 4. $R_t = \text{diag}(\bar{y})$
 5. The Bayesian Updating method (BU), see again equation (8), and section 4.4.

Evaluation criterion. The following measure is used as an evaluation criterion:

$$\text{RMSE}(t) = \sqrt{\frac{1}{N} \sum_{i,j} (q_i(t) \bar{b}_{ij}(t) - f_{ij}(t))^2} \quad (12)$$

where N represent the number of connected EE-pairs. For each period and network specification in table 1 the measure is averaged over the number of datasets that are generated. The averages over the last 40 periods of the measure are summarized in table 2.

Results. The numbers in table 2 relate to the specifications in table 1. It is clear that the best performance is obtained from the Bayesian updating method followed by the Kalman method, discounted least squares method and finally the least squares method.

For the first network specification the error of estimation is plotted as a function of the number of periods in figure 2. The main difference between the performance of the methods is the number of periods needed to reach a certain level of accuracy. In this respect the advantages of the BU methods are evident. This also makes the Bayesian updating the preferred method under less favourable circumstances, where the network may involve a larger number of EE-pairs, the measurements contain larger errors, or the rate of change in the split probabilities might be higher.

Table 2: EE-Flow Errors
- average over 10 computations (trips/period)-

	network					
	1	2	3	4	5	6
LS	21.49	30.09	21.98	22.28	16.98	25.59
FCLS	19.34	25.22	19.88	18.86	16.25	20.96
KF	17.97	20.07	18.85	16.71	14.74	18.70
BU	14.28	14.70	15.11	12.58	13.00	14.09

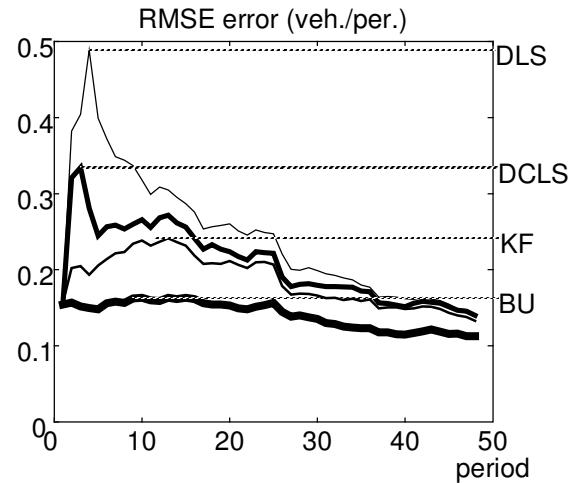


Fig. 2. Error of estimation for respectively the DLS,DCLS, KF and BU methods, when applied to network specification 1 (averaged over 10 simulation runs).

5.1 Experiments with empirical data

A second series of experiments has been done using empirical data collected at the Amsterdam beltway. Nine days are selected which are free from major incidents and disruptions in the data collection system. The selected network covers 11 kilometres of the anti clockwise direction of the Beltway (see Fig. 3). This corridor contains 5 entry ramps and 5 exit ramps (see the emphasised links). Average travel times have been estimated in two independent ways (using speed observations and flow observations) and have been averaged. Using a mean travel speed, all flow observations are converted to an MTCS (see section 3.1), using intervals of 5 and 10 minutes. As an evaluation

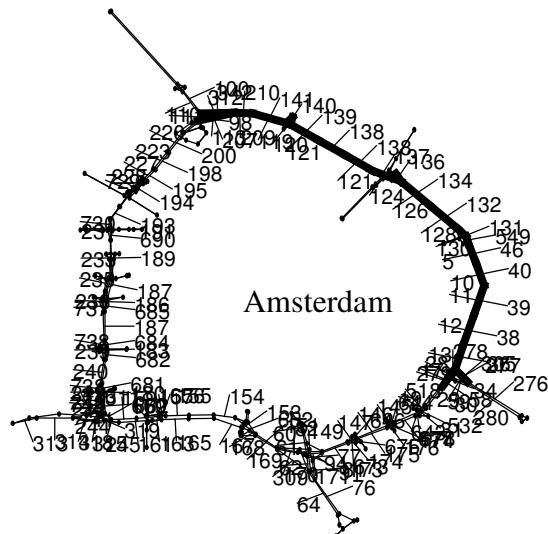


Fig. 3. Representation of the Amsterdam network. The numbers represent the induction loop locations. The emphasized links represent the network that was selected for further tests.

criterion the root mean squared error of the link flow predictions is used, where link flows are predicted based on split proportions that were estimated one period before.

Results. The results are shown in table 3. Surprisingly the accurate estimates of split proportions obtained with the BU method could not be translated into accurate link flow predictions. Detailed analysis has revealed two explanations for this. Firstly the computation of point estimates for the BU method as described in section 4.4 could not always be completed successfully due to low probabilities of sampling values within the interval [0,1]. In such cases the algorithm switches to a suboptimal method in which each split proportion is considered separately and correlations are ignored. Although this does not effect the RMSE defined by (13) to a great extent it might cause a bias in the link flow prediction. A second explanation is that the error measure that is used is quite similar to the objective functions that are minimized by the DCLS and KF methods. A quick fix to these problems would be to rely on the point estimates produced by the KF method in cases where point estimates for the BU method can not be successfully computed. However, further research is needed on this subject.

Table 3: RMSE predicted link-flow volumes, averaged over 9 days (veh./period)

method	interval (min.)	
	5	10
DCLS	13.02	22.19
KF	11.82	19.92
BU	13.48	29.95

6. CONCLUSIONS

The estimation method that is used in dynamic OD estimation has a large effect on the error of estimation. Simulation results indicate that the most accurate estimates of split proportions are obtained by a new method based on Bayesian updating. However, further research is still needed to translate these estimates in accurate link flow predictions.

7. REFERENCES

- Ashok, K. and Ben-Akiva, M.E. (1993) Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems, *Proc. 12th Int. Symp. on Transportation and Traffic Theory*, Berkeley, C.F. Daganzo (Ed)
- Bell, M.G.H. (1991) The Real Time Estimation of Origin-Destination Flows in the Presence of Platoon Dispersion, *Transportation Research-B*,

- Vol. 25-B**, pp. 115-125
- Chang, G. and Wu, J. (1994) Recursive Estimation of Time-Varying Origin-Destination Flows from Traffic Counts in Freeway Corridors, *Transportation Research-B*, **Vol. 28B**, pp. 141-160
- Cremer, M. and Keller, H. (1981) Dynamic Identification of Flows from Traffic Counts at Complex Intersections, In: *Proc. 8th Int. Symposium on Transportation and Traffic Theory*, University of Toronto Press, Toronto Canada
- Cremer, M. and Keller, H. (1987) A New Class of Dynamic Methods for the Identification of Origin-Destination Flows, *Transportation Research-B*, Vol. **21B**, No.2, pp.117-132
- Nihan, N.L. and Davis, G.A. (1987) Recursive Estimation of Origin-Destination Matrices from Input/Output Counts, *Transportation Research B*, Vol. **21B**, No.2, pp. 149-163
- Van der Zijpp, N.J. and Hamerslag, R. (1994) An Improved Kalman Filtering Approach to Estimate Origin-Destination Matrices for Freeway Corridors, *Transportation Research Records*, No. 1443, pp. 54-64
- Van der Zijpp, N.J. (1996) Dynamic OD-Matrix Estimation on Freeway Networks *PhD Thesis*, Delft University of Technology
- Van der Zijpp, N.J. (1997) Dynamic OD-Matrix estimation from Traffic Counts and Automated Vehicle Identification Data, *submitted to the 1997 annual meeting of the Transportation Research Board (TRB)*, Washington, January 1997

Updating trip matrices for Copenhagen using multiple data sources

Otto Anker Nielsen (oan@ctt.dtu.dk) and Christian Overgård Hansen (coh@ctt.dtu.dk), Centre Traffic and Transport, DTU, Building 115, 2800 Kgs. Lyngby

ABSTRACT

Traffic planning in the Greater Copenhagen Region have over the last 10 years mainly been supported by the so-called OTM traffic model. The behavioural models in OTM include advanced state-of-the-art utility-based formulations, which are combined with base-matrices in a pivot-correction procedure. Before and after studies of specific projects have shown, that these matrices might be the Achilles heel of the whole model system. Vuk and Hansen (2006) therefore validated the present version of the OTM traffic model (version 4.0 from summer 2000) and concluded that a major drawback of the model was indeed outdated base 1992 matrices. From January 2005 to March 2007 the OTM model has therefore been in a large-scale process of updating where creation of new trip matrices has been the main focus.

The article describes the undertaken methodology for constructing the travel matrices for Copenhagen, and how the improved matrices influence the performance of the model. A main focus in the work has been to utilise various data sources for estimation of the new matrices. This includes telephone survey data, cordon line surveys and existing transport surveys to construct the base matrices, traffic counts to adjust these, and digital network databases. The article also demonstrates a new approach to adjust public transport matrices to counts.

INTRODUCTION

The Orestad Traffic Model (OTM) is a tactical traffic model for the Greater Copenhagen Region (GCR). OTM consists of demand and assignment models for both passenger and freight transport. The demand models include trip production, trip distribution and mode choice models, all following a utility-based framework, and the road network model is based on Mixed-Probit formulations and equilibrium algorithms. The model system includes feedback cycles to take congestion into account. The behavioural functions have been estimated based on the combination of multiple Revealed and Stated Preference data-sets.

The first version of the model was developed in 1994. The matrices built however on older 1992-matrices which again used adjusted 1989-matrices from a prior model and data sources. Since then the model has been continuously improved, latest in summer 2000 (Jovicic and Hansen, 2003). The matrices describing the 1992 travel patterns built upon older travel analyses have been adjusted to the counted traffic numerous times since the first version of the model was built in 1995. A major re-estimation of the car matrices was carried out in the harbour tunnel model project in 1998 (Paag et.al., 2001).

With respect to planning of the alignment of Copenhagen's Metro's phase 4, the so called Metro City Ring, a group of clients headed by the Ministry of Energy and Transport wished to upgrade OTM 4.0 to a new version, version 5.0, where a number of improvements were proposed. Most importantly the OTM 5.0 includes new base 2004 matrices. The reasons for building new travel matrices for Copenhagen were following:

- Reduction of uncertainty in forecasting, since base matrices are applied in a pivot point procedure where the demand model is adjusted to fit the base-year matrices,
- The metro is now an existing mode (first phase opened October 2002, second phase in fall 2003, and third phase in fall 2007), whereby the behavioural models for metro can be based on Revealed Preference data, and
- The matrices can be applied in numerous types of travel analyses beyond modelling purposes, e.g. for the analyses of travel behaviour.

Improvement of the model was initiated in January 2005 and finalized March 2007. In the period three main types of improvements have been performed. First, base matrices were updated from 1992 to 2004. Second, the demand sub-models were re-estimated. Finally, improvements and updates of the model zone structure, road and public transport networks and zone data were made.

The article introduces the approach used for the estimation of new trip matrices and the data foundation. The main part of the article discusses the method used to construct the base matrices and adjust them to fits counts. The article is concluded with results and experiences gained from the study.

MAIN APPROACH

The main approach in the estimation of the matrices followed the steps;

1. Base matrices were estimated at an aggregated zonal level (90 zones) based on personal interview data. First, interview data was expanded to population stratified by socio economy and home location and adjusted for seasonal and weekly variations. Second, whilst mode was estimated directly on data, splitting by trip purpose was supported by a modelling approach due to the limited amount of data.
2. The matrices were then adjusted according to the postcard survey to take account of children and residents outside Copenhagen (e.g. tourists) and integrated with the postcard data.
3. The zone aggregated matrices were spatial detailed into the finer system of 818 zones by use of zonal generation and attraction rates and formulated as Generation-Attraction (GA) matrices.
4. Time of day factors computed from survey data with respect to travel purpose and distance were applied to split the all day matrices into seven time-of-day matrices.
5. Trips with one or two legs outside GCR defined by 17 port zones were added to the matrices based on counts.
6. The car matrices were re-estimated on the detailed zonal structure based on traffic counts. A modified Multiple Path Matrix Estimation (MPME) procedure where used (Nielsen, 1998), which assumes that car users choose routes according to a stochastic user equilibrium framework (Sheffi, 1985).
7. The public transport matrices were re-estimated based on traffic counts in busses and at stations. Since one zone may connect too many stops, and each stop may have zonal-connectors from several zones, a fairly complex procedure was developed for the estimation. The results where correction factors at zonal level, which were used in a modified Furness (1970) procedure.
8. After the matrix estimations, the matrices were reformulated into GA-based tours.

The resulting OTM 5.0 base travel matrices describe an average weekday in 2004 in the GCR, which is split between 818 internal zones and 17 port zones. An average weekday is defined as Monday-Friday for a year where June, July and August are excluded.

The base 2004 matrices are segmented into travel mode, travel purpose and time periods. There are five main travel modes: Walk, Bicycle, Car driver, Car passenger and Public transport. Compared to the prior model, the split between car drivers and car passengers is a new feature of the model.

Further, there are six travel purposes: Home-Work (HW), Home-Education (HE), Home-Shopping (HS), Home-Leisure (HL), non Home based Leisure (nHL) and Business (BS). The split of “leisure trips” into the 3 purposes HS, HL and nHL is also an improvement of the model.

Finally, the time periods defined in the model are the following: 5 am to 7 am, 7 am to 8 am, 8 am to 9 am, 9 am to 3 pm, 3 pm to 6 pm, 6 pm to 9 pm and 9 pm to 5 am; in total seven time periods, compared to the three in the prior model.

The OTM 5.0 works therefore with $5 \times 6 \times 7 = 210$ base 2004 travel matrices. The all day home-based matrices are formulated as tour matrices (GA matrices), whilst the nHL and BS matrices are trip matrices (OD matrices).

DATA

Traffic surveys

Data sources used to develop the travel matrices are TU data (the Danish annual national transport survey) for the years 1997-2003 and 2005 (newly completed interviews), and 2005 postcard data.

Since OTM 5.0 is applied for planning of the Metro's phase 4 (Metro City Ring) it was decided that the stratification of the 2005 TU interviews should mainly cover the alignment of the new metro line, i.e. most new interviews were completed with respondents living in the Copenhagen and Frederiksberg municipalities. In total, the 2005 TU data includes 16,285 interviews and 60,542 records. The 1997-2003 TU data includes 16,794 interviews and 51,960 records.

The postcard analysis was completed in March 2005 where over 61,000 postcards were handed out to train-, bus- and metro-passengers, car users and bicyclists travelling across "Sø-snippet" (main cordon north of Copenhagen City Centre). 18,376 postcards were successfully coded in a data file. When expanded to an average weekday (based on traffic counts), the survey gave in total 853,663 personal trips across the corridor.

Count data

For the purpose of matrix adjustment of car matrices 2,193 counts were collected from the Danish Road Directorate, counties and municipalities across CGA. A lot of efforts were put into making count data consistent and reliable with respect to vehicle classification, time segments etc.

The bus, train and metro counts were collected from the transit operators. The bus count data were collected for November 2004 based on a sub-set of 5% of counting busses that runs in a stratified schedule during the entire year. The train counts were a train postcard analysis on all passengers completed the first Thursday of November 2004. The metro passengers counted by infrared equipment with a 100% sample referred to November 2004. The side-rail lines had only very imprecise counts (traffic estimates) collected by contacts to the companies, but they only carry a relatively small number of passengers. Bus, train and metro counts were adjusted to account for the average weekday in 2004.

All the automatic counting systems had the problem, that the sum of boarding and alighting passengers along a run did not necessarily add up to 0 during the run due to the uncertainty of the equipment. Therefore a method was used, which corrected the deviations relatively along the run to secure balance and non negative values.

The matrix estimation method adjusts passengers entering and exiting the system. Therefore it was necessary to distinguish between passengers who transferred between lines, and passengers who entered or exited the system. This was supposed to be estimated based on survey data. As they turned

out to be fairly old and quite imprecise, transfer patterns were instead estimated by use of a route choice model (Nielsen & Frederiksen, 2006).

MATRIX ESTIMATION BASED ON SURVEY DATA

Trips represented in the TU data gives an average of 0.15 trips for each zone pair between the 818 internal zones. If decided to build an 818x818 zone matrix directly from this data, this would result in a large number of matrix cells with zero trips. It was therefore decided to aggregate the 818 zones into 90 larger zones. With only 90x90 cells there were in average 13 trips per zone combination. When this was done, it was observed that 7% of zone combinations had zero trips.

Out of the total 18,376 postcards available in the project, 11,358 postcards (62%) describe trips between zone relations comparable to the TU based matrices. When these trips were expanded by applying traffic counts a total of 503,899 trips across the Sø-snippet (cordon line north of the city centre) were achieved. As expected, the postcard total was larger than the TU total because it includes trips made by children under 8, trips made by tourists and temporarily residents, and because the business trips are better described in the postcard data than in TU data.

The statistical error in the two data sources was judged to be about the same. It was therefore decided to combine the two sources when calculating relevant zone-to-zone trips over Sø-snippet based on the following assumptions: a) walk trips were judged to be correct in the TU data, b) trip totals for other travel modes than walk were judged to be correct in the postcard data, c) the ratio between the travel mode totals for postcard data and TU data (factors are over 1.0) were applied for all cells (though not for walk mode).

Table 1 shows GCR travel matrices based on a combination of TU and postcard data. To take account for tourists etc. the zone-to-zone TU trips are expanded by applying the experience from the postcard survey as explained above, and external trips added based on traffic counts.

TABLE 1 Weekday trips including the external traffic, 2004

Travel mode	HW	HE	HS	HL	nHL	BS	Total Trips
Walk	51,285	89,059	310,678	349,350	162,704	11,852	974,927
Bicycle	252,207	184,082	178,620	288,696	151,981	25,360	1,080,946
Car, driver	559,776	37,545	393,361	649,925	355,994	168,199	2,164,799
Car, passenger	135,330	71,324	200,103	465,599	162,936	45,567	1,080,859
Public transport	298,929	137,678	121,552	219,528	111,130	32,434	921,251
Total	1,297,527	519,688	1,204,313	1,973,097	944,745	283,412	6,222,782

The matrix adjustment was then applied on car and public transport matrices based on the existing traffic counts for 2004. There are three reasons for that. First, the TU data applied in the project relates to the period 1997-2005 while the counts are collected for 2004, which is the model base year. Second, there was not enough information in the TU data and postcard data to split the 90x90 zone matrices into 818x818 zone matrices without introducing additional uncertainty. Finally, the TU data does not include information about trips within GCR made by persons living outside GCR. The adjustments for car and public transport respectively are described in the following two sections.

MATRIX ADJUSTMENT, CAR TRANSPORT

Initial preparation

Van and truck matrices originated from OTM 4.0 were simply adjusted to the new zone system and time of day periods. The car passenger matrices as shown in Table 1 were assigned onto the road network together with van and truck matrices in order to compare with the available road traffic counts. Some general deviations were noticed. For instance, the traffic was underestimated across the island of Amager as well as on the corridor along the Motorring 3 (motorway around Copenhagen), whilst it was overestimated on the main access roads towards the city. Therefore, before the matrix adjustment was started the existing matrices were corrected for the above described tendencies.

The Multiple Path Matrix Estimation method

The matrix adjustment applied in the project is a so called Multiple Path Matrix Estimation method (MPME), developed by Nielsen (1998). The MPME is a heuristic method, which simultaneously re-estimates the matrix while the car assignment model iterates.

The method ensures that the estimated link load minimised the weighted square average deviation relative to the available link counts. All paths for each zone-pair are used for the estimation relative to the likelihood of the path being used, and all counts along each path are used.

When zone-to-zone traffic is calculated, the original matrix is adjusted in a heuristic way where the square deviation between the original matrix and the new matrix is minimised. The method therefore adjusts the matrix to fit the available counts as good as possible, and conditional to this changing the original matrix at least as possible.

The more iterations the better the matrices fits to the observed traffic. However, if the number of counts is not optimal (i.e. not enough counts for a sub-area) the matrices can be adjusted wrongly if running with many iterations. In cases with no counts, a Furness-like method is used, where the relative correction of other trips to/from the specific OD-pair is used as an proxy for the estimation (refer to Nielsen, 1998). Naturally, both this and zonal pairs with only one or few counts are much more imprecisely estimated than pairs with many count.

The matrix estimation method was adjusted in order to keep the totals in the port zones, since these usually had very precise counts (e.g. the bridge to Sweden with a toll station).

The route choice model

The route choice model is an integrated part of the matrix estimation procedure, since MPME allow for this as long as it is solved by the Method of Successive Averages (Nielsen, 1998). The route choice model was estimated in a Mixed Logit framework, which beside an error term describes heterogeneity of preferences represented by stochastic coefficients (Nielsen, et.al. 2002). The route choice model was hereafter calibrated compared to the road network data. The following utility function was used based on work in Nielsen (2004);

$$U = k \cdot l + c + \beta_{free} \cdot (t_{free} + \beta_{con} \cdot t_{con}) + \varepsilon$$

The normal driving costs (petrol, etc) are assumed to be proportional k with the length l . k is assumed to be 0.7 DKK per km (0.13 USD). β_{free} is the Value of Time (VoT) for free flow time, t_{free} . β_{free} is estimated from data to be logarithmic normal distributed. It was estimated for four different segments: commute, business, other (leisure, shopping etc.) and van/trucks. β_{con} is the extra VoT due to congestion relatively to β_{free} . Based on AKTA (Nielsen, 2004) this was also estimated to be logarithmic normal distributed $(1+\ln(\mu, \sigma^2))$ i.e. $\log(\beta_{con} - 1) = N(\mu, \sigma^2)$.

Applied adjustment procedure

Because the passenger car matrices estimated from survey data were judged to be quite reliable and the number of counts rather limited, the matrix adjustment was reduced to few iterations.

Whilst the MPME-procedure only increased the number of trips by 0.7%, the average trip length was reduced by 7%. Since the time-of-day factors were estimated on some older travel survey data, a major benefit of MPME was adjustments of the time distribution of trips. For instance, the number of trips in the time segment from 5 am to 7 am was increased by 30% due to a rapid growth in road congestions in the morning peak not existing in older survey data.

The output matrices were made symmetrical over the day and finally they were modified to be GA matrices for the model segments which were home-based. The number of car passengers was computed based on passenger loads in the input matrices.

The number of van trips was increased by 4% in the MPME-procedure, while truck trips were reduced by 21%. The strong correction of trucks is contributed by two factors. First, the vehicle classification scheme has been changed by the road administration. Second, buses were preloaded from bus timetables to the road network and subtracted from truck counts more accurately than in previously matrix estimations. Trip lengths were only marginal changed by the MPME-procedure since it decreases by 4% for vans and increases by 4% for trucks.

The van and trucks matrices would have benefited from a larger number of iterations to improve the quality of the matrices. However, this was not possible without changing the passenger car matrices in the simultaneous assignment and matrix adjustment procedure and priority was given to passenger car matrices.

MATRIX ADJUSTMENT OF THE PUBLIC TRANSPORT

The public transport matrices were also assigned onto the network and adjusted according to the available counts.

Mabit and Nielsen (2006) describe in details the public transport network, which was also used in the matrix estimation project. The model has 3,951 zonal connectors which represent walk/bicycle access to the public transport stops, 270 public transport lines (with 1,170 variants, that has different alignments or stopping patterns) and 5,023 stops (bus-stops, train- and metro stations). The day schedule includes 17,744 runs, where a run represent a given bus or train running from the start to the end of a line-variant.

The timetables were imported from the official timetable database (the organization behind www.rejseplanen.dk), and linked to a digital map (KRAKS Geodatabase, www.krak.dk) in ArcGIS

(www.esri.com) using the Traffic Analyst software package in a modified version (www.rapidis.com). This means that a very high accuracy have been obtained concerning the network data.

There is one important methodological difference between car and public transport matrix adjustments, as the public transport matrix adjustments are based on counts on the stop level (boarding and aligning passengers) whilst the car matrix adjustment are based link counts.

In the applied public transport matrix adjustment method we calculated adjustment factors, which afterwards were corrected manually taking specific local traffic conditions and land use into consideration. A common reason for need for manual corrections is that the number of transfers when travelling from O to D by public transport is not known in counts.

The public transport assignment method applied in the matrix adjustment was based on the actual bus and train timetables. It is a stochastic assignment model, which considers distributed values of travel time (Nielsen & Frederiksen, 2006).

Method

A pure MPME-method for public transport would be too time-consuming and difficult to implement – especially due to the time aspect. A “simplified” approach was therefore adapted, where only boarding and exiting passengers in each zone represented by zonal connectors was considered. The principle in this approach was;

1. The route choice model estimates the traffic flows.
2. The route choice model was used to estimate transfer patterns. This was validated on available surveys on transfer patterns. The share of transfers and new boarding or exiting passengers at each stop could hereby be estimated. The counts were then adjusted so that they only represented the boarding and exiting passengers – NOT transfers.
3. Modelled and counted passenger volumes could then be compared at each stop (station or bus-stop), and the relative deviation be calculated.
4. The zonal connectors to the specific stop could then be assigned the same relative deviation (note that connectors from several zones could be connected to the same stop).
5. The relative deviations for all connectors for each zone could then be compared. If the deviations have different signs, this could mean that there were some problems with the Level of Service (LoS) or the connectors to the zone, or the location of the zonal centroid (e.g. if the activity density centre differs from the calculated one from the population and work place data). The model was then adjusted manually to secure the same sign on the deviations. After this a new route choice model was run, and point 4 and 5 repeated until satisfactory results obtained.
6. A weighted correction factor could then be calculated for each zone. A matrix adjustment procedure was then run by the Furness (1970) method. However, to be conservative (i.e. not changing the matrices too fast and too much) all adjustment factors were evaluated manually before this adjustment was made. After this, a new route choice model was run, and the workflow returned to point 4.

The main challenge in this respect was that a trip starting in a given time-period, may first reach its destination in the next (or even following) periods. And a count at a given bus-stop may consist of trips from the same as well as prior time-periods. In principle the same problem exists for car trips. The

public transport trips are however often slower due to the slower speed of public transport, why the problem is larger than for the car trips.

To make the adjustment feasible, a method had to be developed which took this into accounts. As this has not been published before, it is being described in mathematical and algorithmic terms in the following.

The matrix adjustment is described by using the following notation;

T_{ijkxy} Modelled traffic T from zone i to j with purpose k in beginning in time-interval x and arriving in time-interval y . The absent of an indices indicates that the matrix has been summarised over this dimension.

E_i	Traffic as it ought to have been modelled (Expected) from zone i or j to fit the counts.
T_{is}	Modelled traffic along zonal a zonal-connector between zone i and stop s in the network
E_{is}	Traffic as it should have been modelled along the zonal-connector between zone i and stop s to obtain the correct stop-volume compared to counts
R_{is}	Relative share of traffic from zone i , which uses the zonal-connector to stop s
V_{sap}	Observed traffic volume at stop s . Indices a for exiting passengers and p for boarding passengers
T_{sap}	Modelled traffic at stop s
R_{sap}	Relative deviation between observed and modelled traffic at a stop
R_j	Relative deviation between modelled and estimated traffic from zone i or j (matrix factor)
(n)	Indices for iteration number.

The algorithm for the matrix adjustment is described in pseudo-code in the following:

0. Initialisation All traffic volumes are read:

$$T_{iskxy(0)}, T_{sikxy(0)}, V_{sp}, \text{ and } V_{sa}$$

Expected volumes are set equal to the modelled (to initialise volumes for stops and zonal-connectors for which no counts exists).

$$\begin{aligned} E_{iskx(0)} &:= T_{iskx(0)} \\ E_{siky(0)} &:= T_{siky(0)} \\ R_{sax(0)} &:= 1 \\ R_{spy(0)} &:= 1 \\ R_{ix(0)} &:= 1 \\ R_{jy(0)} &:= 1 \end{aligned}$$

The share of traffic on each zonal-connector is estimated

$$\begin{aligned} R_{iskx} &:= T_{iskx(0)} / T_{ikx} \\ R_{siky} &:= T_{siky(0)} / T_{jky}(j=i) \end{aligned}$$

The iteration number is set to $n=0$

1. Updating, s
The traffic at all stops are calculated as it should have been, if it should equal the weighted sum of counts at the end of the zonal-connectors. The corresponding traffic on the zonal connectors is calculated as it should have been to fit counts:

$$\begin{aligned} \text{If } V_{sp} &\neq \text{Null: } R_{spx} := V_{spx} / T_{spx(n)} \\ \text{If } V_{sa} &\neq \text{Null: } R_{say} := V_{say} / T_{say(n)} \\ E_{iskx(n)} &:= T_{iskx} \bullet R_{spx} \\ E_{siky(n)} &:= T_{siky} \bullet R_{say} \end{aligned}$$

2. Updating, ij
The expected traffic for all zones are calculated as the sum of the traffic along all zonal-connectors. Matrix adjustment factors can hereafter be calculated as follows:

$$\begin{aligned} E_{ikx} &= \sum_{\text{over zonal connector } s} (\sum_{n=\# \text{ time intervals } \epsilon[x;y]} T_{iskyn} R_{spn}) \\ E_{jky} &= \sum_{\text{over zonal connector } s} (\sum_{n=\# \text{ time intervals } \epsilon[x;y]} T_{sikxn} R_{san}) \quad (j=i) \\ R_{ix} &= E_{ix} / T_{ix} \\ R_{jy} &= E_{jy} / T_{jy} \end{aligned}$$

3. 'Assignment'
 $n := n+1$

For all zonal-connectors (the traffic is assigned onto the zonal-connectors after the constant relative ratios adapted from the full assignment of the base-matrix from the prior adjustment round):

$$\begin{aligned} T_{iskx(n)} &= R_{iskx} \bullet E_{ikx} \\ T_{siky(n)} &= R_{siky} \bullet E_{jky} \quad (j=i) \end{aligned}$$

For all zones (stations volumes are calculated as the sum of zonal-connectors to each stop):

$$\begin{aligned} T_{spkx(n)} &= \sum_{\text{over zonal connector } is} T_{iskx(n)} \\ T_{saky(n)} &= \sum_{\text{over zonal connector } si} T_{siky(n)} \end{aligned}$$

4. Stop criterion
If a stop criteria is not reached (e.g. number of iterations as the simplest one), then go to step 1.

5. Matrix-adjustment
 $T_{ijk(\text{new})} := T_{ijk(0)}$

Rows and columns in $T_{ijk(\text{new})}$ are adjusted iteratively to fit with E_{ik} and E_{jk} (equal to a common factor model or Furness method). This is done for each time-period separately.

The pseudo assignment in step 3 assumes that the trip-distribution is independent of the adjustment of the zonal-connectors (the matrix adjustment is in step 5). In principle a full matrix adjustment as in step 3 should be followed by a new full assignment. The assignment is – however – itself an iterative

algorithm, which takes several hours, whilst the pseudo assignment takes few minutes. The overall calculation time with a “true” assignment procedure would therefore increase from minutes to weeks, if a full assignment were to be part of the matrix adjustment algorithm. This is why the “simplification” above was used, and why a full MPME-method could not be used for the large public transport network in the present model.

After the above algorithm has run, it is necessary to run the full assignment to evaluate the results. For each main iteration, two full assignments are therefore needed (before and after the adjustment).

Matrix adjustment

During the estimation procedure, it was realised, that it was inappropriate to have counts on bus-stops that were not connected to zones by zonal connectors, since trips hereby were overlooked. Bus stops which did not offer transfer options nor was connected to zones by zonal connectors was therefore located. Some stops were then additionally connected to zones by zonal connectors. Stops with transfers but no connectors were manually inspected to assess whether connectors should be added.

Following this, counts on bus stops with no transfers or connectors were moved proportionally to the nearby stops, and the non-modelled stops were removed from the network model. A special ArcGIS-based software was developed for this purpose. The problem mainly occurred in the outer parts of the Copenhagen region, where the zones are pretty big, as they only were used to describe the hinterland traffic to Copenhagen, but not local traffic in great details.

The core indicators suggested relative adjustments at the stop-level, at the connector level (which considered the average suggested adjustments of all zones using the stop at the end of the connector), and at the zonal level. Based on thorough manual inspections it was first examined whether justified (e.g. with uneven land use in the zone) relocation of Zonal centroids could improve the fit. This was typical the case where some stops had too much traffic and some too little. Then it was examined whether some connectors were missing, typically to stops with too little modelled traffic, whilst neighbouring stops had too much traffic.

The adjustments of zonal connectors, network and centroids was repeated 3 times before the matrix adjustments was begun. A major challenge in this respect was, that a connector factor had to be added for the long connectors, since the route choice model tend to assign too much traffic on these and too little on local busses to the train network (refer to Mabit & Nielsen, 2006).

The next two steps only adjusted the total daily traffic, while some fine-adjustments on connectors and centroids were still made.

Finally, the matrices were also adjusted with respect to the 7 time-periods during the day. The calculation of these adjustment factors needed also to use a bookkeeping of the start time of trips and when they reach the count. This fine-tuning with respect to time periods was done by using a spreadsheet mode. This fine-tuning of matrices was repeated twice.

Some observations on the public transport matrix adjustment

It is interesting to note, that very limited changes were made on the daily matrices, when the 818x818 matrices where aggregated to the 90x90 matrices that had been estimated based on the travel surveys. This basically indicates a consistency between the surveys and the traffic counts.

The main exception for this was the zones in the new urban development of the “Ørestad”, where the transport surveys cover 2004 as well as prior years, whilst the counts covered 2004. Due to the fast development in this town area, the adjustment procedure increased flows to/from these zones.

The adjustment did, however, make many changes of the “sub-zones” within each of the large base-zones. Some general observations was made of the corrections *after* the pure algorithmic approach had been made, i.e. it is observed changes not used criteria for the adjustments;

- More public transport journeys close to the fast and/or high frequency bus lines (“A” or ”S” busses), train stations or the metro.
- Less public transport journeys in industrial areas with poor public transport service.
- More public transport in newly developed urban areas. This was due to the combination of several years in the transport surveys, while the adjustments were done using the 2004-counts.
- More journeys at stations that opened during the survey period, where the counts where after the opening (two new stations only).
- Much fewer journeys in rural areas. These areas typically only contain a small population share of the larger 90-zones, but are populated with car-owning population segments.
- Many more journeys to areas with summer houses. This could both be due to the use of population and workplaces for the initial split of the 90x90 matrices, or since some of these houses are used illegal for whole-year living, which is most likely not reported in e.g. telephone surveys (since the inhabitants have pro-forma addressed elsewhere)
- More journeys from zones with larger stations in towns in the outer part of the region.
- Various special cases, where sub-zones include special attractions such as museums, aquarium, etc,
- Zones with concentrated education activities (high schools, colleges, universities,...) or student hostels.

VALIDATING THE RESULTING MATRICES

Mode choice and trip length

Table 2 shows the final 2004 day matrix after the matrix adjustment.

TABLE 2 Final base day matrix 2004

Travel mode	HW	HE	HS	HL	nHL	BS	Total Trips
Walk	51,285	89,059	310,678	349,350	162,704	11,852	974,927
Bicycle	252,207	184,082	178,620	288,696	151,981	25,360	1,080,946
Car, driver	549,634	39,827	395,508	623,134	378,187	192,358	2,178,648
Car, passenger	132,697	77,366	201,808	447,111	172,905	52,081	1,083,968
Public transport	276,965	117,921	126,202	212,997	123,222	31,898	889,205
Total	1,262,788	508,255	1,212,816	1,921,288	988,999	313,549	6,207,694

To compare the new trip 2004-matrices with prior trip matrices, a forecast with the previous model - OTM 4.0 - was conducted to update the 1992-matrices to a 2004 reference year. The total number of trips in the new 2004-matrices fits fairly well with the updated 1992-matrices, since it is only 6% less.

While the number of car fits very precise, the divergence of walk and bike trips are large. In the 1992-basis matrices used in pivot-point corrections of OTM 4.0, walk and bike trips have only been roughly estimated and possibly biased. In contrast passenger car matrices were adjusted in 1998 (Paag et. al. 2001) and later using MPME on the older matrices. The number of public transport trips is 6% less than in updated 1992-matrices.

Analyses reveal, however, that the trip length distribution differs between the two set of matrices. The average length of public transport trips is 25% larger in the new 2004-matrices than in the updated 1992-matrices and 20% for passenger car trips. Explaining the differences it is likely;

- That persons travel longer distances today than in 1992 not completely captured in the model forecast.
- That the matrix estimation extensively applied in development of the 1992-matrices has changed the trips length distribution.
- That the respondents in the TU-survey forgot to report short distance trips.

A consequence of the longer trips distances person km. by car and public transport is larger in the new 2004-matrices than prior.

Trip purposes

When looking at the trip purposes, a dramatic change of the matrices are evident. For the model, this is important, since the different trip purposes have different Value of Times, and these result in different changes of behaviour when new policies or services are introduced.

We need to recall the data foundation of the two set of matrices. The new 2004-matrices are based on TU-surveys and minor matrix adjustments of car and public transport. Therefore, TU is the main source for segmentation of the 2004-matrices, whilst segmentation of the 1992-matrices is estimated from a cordon survey downtown in 1994 and commuter statistics.

For instance, there are 40% less home-work trips in the 2004-matrices than in the 1992-matrices. While the 1992-matrices clearly overestimate commuter trips there may be underestimation in the 2004-matrices. Since the prior estimations of trip purpose distribution were based on commuter statistics which did not consider absence and part time employment, this is probably the main reason to overestimation of commuter trips. In OTM 5.0, home-work trips are modelled like simple tours with an out and return leg. Therefore, it does not allow combined trip chain modelling and combined home-work-shopping trips may be split into other trip purposes than commuting although efforts have been put into trip purpose segmentation of TU-data.

The number of business trips are much larger in the updated 1992-matrices than in the 2004-matrices. We conclude again that it must be due to major overestimations in the old matrices and underestimations in the new matrices. In the 1992-matrices, trip purpose split is based on a survey in an area downtown Copenhagen with a proportional high number of business trips compared to GCR as whole. While the 1992-matrices reflect quite well business trips in the City Centre of Copenhagen, larger divergences are likely in the suburban and rural areas of GCR. With respect to the 2004-matrices it is noticed to be difficult to capture business trips in household TU-interviews.

The number of home-education trips differs only slightly. This, however, cover a mix of divergences by mode. While the 2004-matrices contain more education trips by walk, car, and public transport than the 1992-matrices, the number of bike trips is less. The cordon survey in 1994 used in estimation of the 1992-matrices may include a higher proportion of cyclist compared to walk due to the distance to the nearby universities, whereas the use of public transport and car in connection with educational trips may be higher in the rural areas.

Finally, the number of private trips is larger in the new matrices partly caused by some misclassifications of commuter trips and business trips.

Time of day distribution

The number of time periods has been increased from three in OTM 4.0 to seven in OTM 5.0. A comparison of the time segments reveals quite large moves of traffic from the morning rush hours to out of peak periods. This tendency of earlier commuting to Copenhagen compared to 1992 confirms congestion measurements from the AKTA-project (Nielsen, 2004). It should be recalled that the trip length is considerable longer in the 2004-matrices than in the updated 1992-matrices and therefore impose more traffic and road congestion even though the number of trips are less in the morning peak.

CONCLUDING REMARKS

The article describes the applied matrix estimation procedures for OTM 5.0 and demonstrates the appraisals of the new matrices. We believe that quality of modelling depends on the data foundation as documented in Vuk & Hansen (2006) and the hybrid approach has been cost efficient.

The approach using surveys to estimate matrices at a large-zonal level (90 zones), and then to split these to sub-zones using survey data and with an adjustment (fine-tuning) using traffic counts, turned out to be quite successful.

The resulting matrices are very detailed including 6 trip purposes, 7 time-of-day periods and built on a zonal structure with 818 internal zones and 17 port zones. The old matrices built on surveys from the 1980'ies have been adjusted several times since then primarily based on traffic counts. Comparing the new matrices with the old revealed major differences mainly contributed to biases in the 1992-matrices and data definitions. However, there has been a dramatic change of time-of-day and trip distribution that the prior adjustments have not captured.

It is our belief that the new matrices improve the data foundation for OTM considerably. The car matrices will be much better to describe congestion. And the more correct split on purposes will improve forecasts on policy initiatives. The public transport matrices have been changed and hereby improved even more, and this will provide a strong basis for decision making the coming years.

REFERENCES

- Furness K.P. Time Function Interaction. *Traffic Engineering and Control* Vol 7, No 7, pp19-36, 1970.
- Jovicic, Goran and Hansen, Christian Overgaard. A passenger travel demand model for Copenhagen. *Transportation Research Part A*, vol. 37, pp. 333-349, Elsevier Science Ltd., 2003.

Mabit, Stefan & Nielsen, Otto Anker. The effect of correlated Value of Time Savings in public transport assignment. Article presented at European Transport Conference (ETC), September 2006.

Nielsen, Otto Anker. *Two new methods for estimating Trip Matrices from Traffic Counts*. Chapter in *Travel Behaviour Research: Updating the state of play*. Edited by Ortúzar, H. D., Hensher, D & Jara-Díaz, D. Elsevier Science Ltd., pp. 221-250, 1998.

Nielsen, Otto Anker; Simonsen, Nikolaj & Frederiksen, Rasmus Dyhr. Stochastic User Equilibrium Traffic Assignment with TurnDelays in Intersections. *International Transactions in Operational Research*, Vol. 5, No. 6, pp. 555-568. Pergamon, Elsevier Science Ltd, 1998.

Nielsen, Otto Anker; Frederiksen, Rasmus Dyhr & Daly, Andrew. A stochastic multi-class road assignment model with distributed time and cost coefficients. *Networks and spatial economics*. No 2, pp. 327-346. Kluwer, 2002.

Nielsen, Otto Anker. Behavioural responses to pricing schemes: Description of the Danish AKTA experiment. *Journal of Intelligent Transportation Systems*, Vol. 8(4). Pp. 233-251. Taylor & Francis, 2004.

Nielsen, Otto Anker & Frederiksen, Rasmus Dyhr. Optimisation of timetable-based, stochastic transit assignment models based on MSA. *Annals of Operations Research*, Vol. 144, Issue 1 pp 263-285. Kluwer, 2006.

Paag, H., Daly, A. & Rohr, C. *Predicting use of the Copenhagen Harbor Tunnel. Travel behaviour Research: The Leading Edge*. Chapter 36 in Book edited by David Hensher. Pergamon press, Elsevier, 2001.

Sheffi, Y. *Urban Transport Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, New Jersey. 1985.

Vuk, G.V. & Hansen, C.O. Validating the passenger traffic model for Copenhagen. *Transportation*, Volume 33, Issue 4, Page 371-392, Springer, 2006.

1 **A MAXIMUM ENTROPY METHOD FOR SUBNETWORK ORIGIN-DESTINATION
2 TRIP MATRIX ESTIMATION**

3 Chi Xie

4 Research Fellow

5 Department of Civil, Architectural and Environmental Engineering
6 The University of Texas at Austin – 6.506 E. Cockrell Jr. Hall

7 Austin, TX 78712-1076

8 chi.xie@mail.utexas.edu

9 Phone: 512-471-4622 & FAX: 512-475-8744

10 Kara M. Kockelman

11 (Corresponding author)

12 Professor and William J. Murray Jr. Fellow

13 Department of Civil, Architectural and Environmental Engineering

14 The University of Texas at Austin – 6.9 E. Cockrell Jr. Hall

15 Austin, TX 78712-1076

16 kkockelm@mail.utexas.edu

17 Phone: 512-471-0210 & FAX: 512-475-8744

18 S. Travis Waller

19 Associate Professor and Clyde E. Lee Fellow

20 Department of Civil, Architectural and Environmental Engineering

21 The University of Texas at Austin – 6.2 E. Cockrell Jr. Hall

22 Austin, TX 78712-1076

23 stw@mail.utexas.edu

24 Phone: 512-471-4539 & FAX: 512-475-8744

25 The following paper is a pre-print and the final publication can be found in
26 *Transportation Research Record No. 2196: 111-119, 2010.*

27 Presented at the 89th Annual Meeting of the Transportation Research Board, January 2010

28
29
30 **Key Words:** Trip table estimation, subnetwork analysis, maximum entropy, linearization
31 algorithm, column generation

32
33
34
35
36
37

1 ABSTRACT

2 In the context of sketch planning, it is expected that a simplified network (i.e., an abstracted
3 network or subnetwork) model can accurately approximate the travel demand patterns and level-
4 of-service attributes obtained from its full-network counterpart. A data prerequisite in this
5 approximation process is the trip matrix of the simplified network. This paper discusses a
6 maximum entropy method for the subnetwork trip matrix estimation problem, relying only on
7 link flow rates (estimated via full-network traffic assignment or as observed link-level vehicle
8 counts). A linearization algorithm of the Frank-Wolfe type is devised for problem solutions, in
9 which a column generation approach is iteratively used to solve the linearized subproblem
10 without path enumeration. Encouraging results from a numerical example suggest that this
11 method holds much promise for generating trip matrices that can be used to evaluate traffic flow
12 patterns under various network changes.

13 INTRODUCTION

14 Transportation planners almost always rely on simplified representations of roadway networks
15 for traffic analysis. For example, metropolitan areas often have networks with 10,000 or more
16 coded links, yet they ignore most local streets and they simplify intersection signal timing and
17 other details. In general, all models are abstractions of reality, and the level of details used
18 depends on desired accuracy in model outputs as well as available computational resources.

19 When the impacts of changes to a large network (such as that found in an urbanized area or
20 across a state) need to be anticipated, sketch planning is often considered as a cost-effective tool.
21 A sketch network may be a skeleton topology synthesizing only major arterials in the region (i.e.,
22 an abstract network) or may focus on the details of a neighborhood or corridor of larger system
23 (i.e., a subnetwork). Such strategies are appealing when evaluation of the regional network
24 requires specialized expertise and/or is very computationally demanding, thus prohibiting
25 evaluation of one or more scenarios in a limited time frame. These simplified networks provide
26 planners with a less complex platform to facilitate quick-response and relatively informed
27 decision making early on.

28 While a number of studies addressed important theoretical and practical extraction/aggregation
29 issues for network abstraction (see [1-8](#)), subnetwork analysis is still quite limited (see [9, 10](#)).
30 Given a subnetwork extracted from a larger network, the first task we face is to determine the
31 subnetwork's trip table. This is a data prerequisite for any subsequent travel demand analysis.
32 Specifically, this work's objective is to derive a consistent origin-destination (O-D) trip matrix
33 for the subnetwork so that the travel demand analysis in the subnetwork (under changed network
34 conditions) will closely mimic full-network modeling results.

1 For a congested¹ network experiencing user-equilibrium (UE) traffic conditions, one certainly
 2 can estimate any subnetwork's trip table by combining path flows from the full network, using a
 3 procedure described by Haghani and Daskin (5, 6) and Hearn (11). This approach results in a
 4 subnetwork trip matrix that can induce a link flow pattern in the subnetwork exactly as the same
 5 as that in the larger system. However, this approach requires complete information on paths
 6 taken in the full network and hence the resulting subnetwork trip matrix is dependent on the
 7 given full-network path flow pattern. In general, unique path flow patterns do not exist in a UE
 8 context, so this approach does not result in a unique subnetwork trip matrix. Thus, if such a
 9 matrix were used to evaluate traffic shifts under network modifications (e.g., link additions
 10 and/or expansion), the resulting flow estimates will likely deviate from full-network results.

11 Two complementary approaches may be used to eliminate this non-uniqueness issue, and both
 12 make use of the entropy maximization principle (see 12, 13). The first approach is to estimate a
 13 most likely, unique path flow pattern in the full network by means of an entropy-maximizing UE
 14 traffic assignment algorithm (14-18) and then to aggregate corresponding path flows to form a
 15 subnetwork trip matrix. The second approach requires only a link flow pattern from traffic
 16 assignment in the full network; the link flows are used as inputs to a maximum entropy (ME)
 17 method for estimating a most likely, unique O-D flow pattern for the subnetwork.

18 While one can debate which of the two procedures generates a more accurate and robust trip
 19 matrix (in terms of the subnetwork evaluation result), we recommend the second approach, for
 20 two computational and practical reasons. First, the second approach requires only link flow
 21 information and conducts its ME optimization on the subnetwork level, which is much less
 22 computationally demanding than the first approach (which conducts ME over the full network
 23 and must store and manipulate path flows from the full network. The second reason relates to
 24 the availability of input data. In cases where large-scale traffic assignment across the full
 25 network is not feasible, one must rely on other data sources. Almost every traffic management
 26 agency has a long history of collecting and assembling link-based traffic counts (e.g., for the
 27 U.S.'s mandated Highway Performance Monitoring System [HPMS]). Thus, the second
 28 approach is more practical in that it requires just subnetwork flow values, either estimated or
 29 measured. For this reason, the second approach is the focus of this paper's discussion.

30 The following section of this paper provide an overview of existing trip matrix estimation
 31 methods, with a focus on traffic count-based methods. Next, a subnetwork matrix estimation
 32 model based on ME theory is formulated and then analyzed, using a small numerical example for
 33 interpreting the model's optimality conditions. A linearization solution algorithm of the Frank-
 34 Wolfe type is devised for this convex optimization model, in each iteration of which the
 35 linearized subproblem is solved by a column generation approach. (This avoids a reliance on
 36 path enumeration.) The solution method is then applied to a small network with relatively large

¹ The network does not really have to be congested, but the method applied here assumes that travel times are flow dependent (so one cannot simply assume all-or-nothing assignments, for example).

1 network changes to assess their performance by comparing the traffic flow patterns generated by
2 the subnetwork and full-network traffic assignments. Finally, some modeling extensions are
3 suggested and research findings are summarized.

4 RELEVANT RESEARCH

5 Trip matrix estimation methods may be distinguished in terms of their theoretical basis (e.g.,
6 gravity allocation, entropy maximization, and error minimization), traffic routing restrictions
7 (e.g., user equilibrium or proportional assignment), and required inputs (e.g., trip productions
8 and attractions, traffic counts, travel times, or target trip matrix). Count-based estimation
9 problems have been widely investigated and formulated in terms of a few different optimization
10 principles, including ME, least squares (LS), and maximum likelihood (ML), among others.

11 ME theory (or minimum information theory) was first used by Willumsen (19) and Van Zuylen
12 and Willumsen (20) for the most-likely-trip-matrix estimation problem, based on traffic counts.
13 By assuming that the underlying traffic flows follow a known proportional routing pattern, these
14 models resort to a simple iterative balancing method for solutions. In the same modeling
15 framework, Nguyen (21) formulated a ME problem synthesizing both traffic count data and trip
16 production and attraction data. While more information can be helpful, potential inconsistencies
17 across constraints can result in no feasible solutions. Fisk (22) imposed the UE routing principle
18 to a similar matrix estimation problem, resulting in a ME model subject to a variational
19 inequality constraint. His contribution is mostly theoretical, rather than practical, however; its
20 nonconvex feasible region makes the problem hard to solve.

21 Nguyen (23, 24) and others (24-27) pursued an alternative approach, incorporating equilibrium
22 traffic flows. Their approach uses minimum travel costs between all O-D pairs as inputs.
23 Knowing flows and link cost functions, link travel costs and path travel costs can be readily
24 calculated. The model makes no assumption about trip distribution patterns; however,
25 alternative optimal solutions may exist. To ensure convergence to a unique matrix solution,
26 some extra information (for example, a target trip matrix or the ME assumption) is generally
27 needed.

28 The joint use of (full or partial) traffic count and target matrix information has resulted in a
29 number of other trip matrix estimation methods, such as the Bayesian inference approach (28,
30 29), least squares approach (30-33), maximum likelihood approach (34-36, 29), constrained
31 regression approach (37), least absolute norm approach (38-40), and integrated squared error
32 approach (41). Despite various implicit parameter assumptions and optimization principles, all
33 seek a matrix that represents some form of trade-offs between a target trip matrix and observed
34 traffic counts. These trade-offs appear in the model constraints or objective functions. Due to
35 the above methods' requirement of a target trip table, none applies here.

Given the fact that link flows (either estimates or measured counts) are the only data source available in the context under study, we propose an ME model for subnetwork trip matrix estimation. This approach is based on the early work of Willumsen (19) and Van Zuylen and Willumsen (20).

MAXIMUM ENTROPY MODEL

Imagine a subnetwork $G = (N, A)$, where N and A are the node and link sets of the subnetwork, respectively. The origin node set R and the destination node set S are subsets of N , (i.e., $R \subseteq N$ and $S \subseteq N$). The proposed model implies two important assumptions, which greatly reduce the modeling difficulty. (We later discuss how these assumptions can be removed, thereby accommodating more general network conditions.) First, we assume that every node in the network is potentially an origin and destination node (i.e., $R = N$ and $S = N$). If any node r cannot be an origin, one simply sets $x_{rs} = 0, \forall s \in S$; similarly, if any node s cannot be a destination, one sets $x_{rs} = 0, \forall r \in R$. Second, we assume that the to-be-estimated subnetwork trip matrix has fixed values. In other words, every O-D flow rate in the matrix is invariant to any network change. In reality, all flows with external origin and/or external destination (i.e., outside the subnetwork) can well change, as these tripmakers may seek different routes (potentially avoiding the subnetwork entirely, or adding trips to the trip table). Thus, the greater the share of trips involving origins or destinations outside the subnetwork, the more problematic is this second assumption.

Nevertheless, given a complete set of estimated or measured link flow rates, $\hat{v}_a, a \in A$, one can construct the following ME problem (P1):

$$\max_{rs} - \sum_{rs} (x_{rs} \ln x_{rs} - x_{rs}) \quad (1.1)$$

$$\text{or min}_{rs} \sum_{rs} (x_{rs} \ln x_{rs} - x_{rs}) \quad (1.2)$$

$$\text{subject to } \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} = \hat{v}_a \quad \forall a \in A \quad (1.3)$$

$$f_k^{rs} \geq 0 \quad \forall k \in K_{rs}, r \in R, s \in S \quad (1.4)$$

where trip rate x_{rs} is defined as

$$x_{rs} = \sum_k f_k^{rs} \quad \forall r \in R, s \in S \quad (1.5)$$

1 Here f_k^{rs} is the path flow rate of path k between origin r and destination s .

2 This model's functional form is common to the ME specifications used by Willumsen (19) and
 3 Fisk (22). These models' can all rely on link flow rates as the only input. (Production and
 4 attraction data or some target O-D flow pattern is not required.) Note that the proposed
 5 formulation (P1) does not imply any traffic routing assumption (i.e., how the estimated trip
 6 matrix is assigned to generate the observed flow pattern in the network). In general, an
 7 appropriate traffic routing principle needs to be specified and incorporated into the matrix
 8 estimation process. For example, Willumsen's ME model presumes that the proportions of any
 9 O-D flow rate on traversed links are known a priori (so route choice is independent of
 10 congestion), while Fisk's model explicitly contains a UE traffic routing component.

11 In both Willumsen's and Fisk's models, observed link flows may contain noise and be
 12 inconsistent with one another (e.g., flow may not be conserved at nodes); moreover, flow
 13 observations are generally only available on a subset of network links. In the current context, by
 14 contrast, a complete set of link flow rates is available, and the solution-implied flow rates may be
 15 error-free, if these rates are produced by a traffic assignment process in the full network. Our
 16 model does not require an explicit traffic assignment component, since the complete set of link
 17 flows implies the desirable traffic flow pattern. If the given link flow pattern is achieved via a
 18 UE traffic assignment in the full network, then the estimated subnetwork trip matrix can replicate
 19 exactly the same flow pattern through a UE traffic assignment in the subnetwork. In fact, any
 20 feasible solution of the ME model (P1) holds this conclusion, as proven here now.

21 **Property 1.** Assume that both the subnetwork and full-network traffic flow patterns are UE. A
 22 subnetwork traffic assignment based on any feasible trip matrix solution of P1: $\{\mathbf{x} =$
 23 $[x_{rs}]: \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} = \hat{v}_a, \forall a, f_k^{rs} \geq 0, \forall k, r, s, x_{rs} = \sum_k f_k^{rs}, \forall r, s\}$ produces the same
 24 subnetwork flow pattern (in terms of link flows) as the full-network traffic assignment.

25 **Proof.** Since the full-network flow pattern is UE, any part of the traffic flow pattern is UE. Let
 26 us assume that a feasible trip matrix solution \mathbf{x}^* of P1 is obtained by decomposing \hat{v}_a into a
 27 specific set of f_k^{rs*} in terms of $\sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} = \hat{v}_a$ and then summing up this set of f_k^{rs*} for
 28 each O-D pair $r-s$, i.e., $x_{rs}^* = \sum_k f_k^{rs*}$.

29 The UE traffic assignment problem for the subnetwork with this specific trip matrix x_{rs}^* is:
 30 $\min_{\mathbf{v}} \left\{ \sum_a \int_0^{v_a} t_a(\omega) d\omega : \sum_k f_k^{rs} = x_{rs}^*, \forall r, s, v_a = \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs}, \forall a, f_k^{rs} \geq 0, \forall k, r, s \right\}$. It is
 31 obvious that the specific path flow pattern $\mathbf{f}^* = [f_k^{rs*}]$ and thus the link flow pattern $\mathbf{v}^* = [v_a^*]$
 32 exactly satisfy the optimality condition of this traffic assignment problem. Because this traffic
 33 assignment problem has a unique optimal solution (in terms of the link flow pattern), we know
 34 that it is the link flow pattern $[v_a^*]$. ■

1 Fortunately, this result does not mean that the proposed ME model will fail in the face of noisy
 2 and inconsistent input data. Given the assumption that every node in the subnetwork can be an
 3 origin and/or destination node, any set of flow rates can serve as an input data set. In other
 4 words, an arbitrary set of input link flow rates contains some feasible solutions to the maximum
 5 entropy model (P1). A more formal proof of this is as follows:

6 **Property 2.** Assume that every node in the network is potentially an origin and destination node.
 7 Feasible solutions of P1 always exist given an arbitrary set of positive link flow rates.

8 **Proof.** The feasible solution set of P1 is confined by such a system of linear equations: $\{\mathbf{x} =$
 9 $[x_{rs}]: \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} = \hat{v}_a, \forall a, f_k^{rs} \geq 0, \forall k, r, s, x_{rs} = \sum_k f_k^{rs}, \forall r, s\}$. Its feasibility is
 10 equivalent to the feasibility of the reduced linear system of $f_k^{rs}: \{\mathbf{f} = [f_k^{rs}]: \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} =$
 11 $\hat{v}_a, \forall a, f_k^{rs} \geq 0, \forall k, r, s\}$. Given that every $f_k^{rs} \geq 0, \forall r \in R, s \in S, R = S = N$, exists, the
 12 optimal solution of a quadratic program: $\min_y \left\{ \sum_a y_a^2 : \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} + y_a = \hat{v}_a, \forall a, f_k^{rs} \geq$
 13 $0, \forall k, r, s, y_a \geq 0 \right\}$ is $y_a^* = 0, \forall a$, where y_a is a slack variable for link a . The validity of this
 14 least-squares method for checking the feasibility of linear systems can be referenced in Carey and
 15 Revelli (42). Thus, the existence of feasible solutions to P1 is guaranteed. ■

16 Note that, however, if the observed flow rates in the subnetwork are not UE (due to measurement
 17 errors or other factors), the subnetwork trip matrix estimated from this disequilibrium flow
 18 pattern will not result in the same flow pattern emerging from UE traffic assignment. The larger
 19 the measurement errors are, the greater the deviation between the observed and produced traffic
 20 flow patterns is.

21 In cases where traffic flow values on some links are missing, the model still produces a trip
 22 matrix, except that those O-D flows using a path that fully consists of segments with missing
 23 flow values will be underspecified. This is because path flows fully traversing segments with
 24 missing data are unconstrained.

25 The optimality conditions of this subnetwork trip matrix estimation problem with an “ideal”
 26 input data set can be analyzed by using the Lagrangian of the model formulation, incorporating
 27 the link flow conservation constraint:

$$L(\mathbf{f}, \boldsymbol{\lambda}) = \sum_{rs} (x_{rs} \ln x_{rs} + x_{rs}) + \sum_a \lambda_a \left(\hat{v}_a - \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} \right) \quad (2)$$

28 where λ_a is the Lagrangian multiplier on link a ’s flow-conservation constraint. Since the first-
 29 order condition of the Lagrangian with respect to path flow rate f_k^{rs} is:

$$\frac{\partial L(\mathbf{f}, \boldsymbol{\lambda})}{\partial f_k^{rs}} = \ln x_{rs} - \sum_a \lambda_a \delta_{a,k}^{rs} \quad (3)$$

1 the optimality conditions of the problem, $\partial L(\mathbf{f}, \boldsymbol{\lambda})/\partial f_k^{rs} \geq 0$ and $f_k^{rs} \partial L(\mathbf{f}, \boldsymbol{\lambda})/\partial f_k^{rs} = 0$, can be
2 written as:

$$f_k^{rs} = 0 \Rightarrow \ln x_{rs} \geq \sum_a \lambda_a \delta_{a,k}^{rs} \quad (4.1)$$

$$\ln x_{rs} = \sum_a \lambda_a \delta_{a,k}^{rs} \Rightarrow f_k^{rs} \geq 0 \quad (4.2)$$

3 Note that $-\ln x_{rs}$ denotes the minimum path “entropy impedance” among all paths connecting
4 O-D pair (r, s) . One can also define $-\lambda_a$ as the entropy impedance of link $a = (i, j)$ and define
5 $-\sum_a \lambda_a \delta_{a,k}^{rs}$ as the entropy impedance of path k between O-D pair (r, s) , which is the sum of the
6 entropy impedances of all the links along this path. It becomes readily apparent that $-\lambda_a =$
7 $-\ln x_{ij}$, where link $a = (i, j)$, if $x_{ij} > 0$. Note that x_{ij} is the trip rate between O-D pair (i, j) ,
8 the head and tail nodes of link a , which should not be confused with the link flow rate of link a ,
9 v_a . Given these definitions, the optimality conditions of the defined ME problem can be stated
10 as follows.

11 **Property 3.** In an ME O-D flow pattern, all used paths (i.e., paths with a positive flow rate)
12 have their path entropy impedance equal to the minimum entropy impedance, and all unused
13 paths (i.e., paths with zero flow) are associated with a path entropy impedance greater than or
14 equal to the minimum impedance value. ■

15 This property describes the conditions of the path flow distribution in terms of entropy
16 impedance under the estimated ME O-D flows. Such an ME path flow pattern generally differs
17 from a UE path flow pattern derived in terms of travel cost, even if their corresponding link flow
18 patterns and O-D flow patterns are identical. More generally speaking, the path flow space
19 constrained by the ME O-D flow pattern (of the trip matrix problem) differs from the path flow
20 space constrained by the UE link flow pattern (of the traffic assignment problem). After all, the
21 trip matrix problem and the traffic assignment problem follow different optimality principles,
22 which result in different optimality conditions for path flows.

23 The solution uniqueness of the problem in terms of O-D flows is apparent thanks to the fact that
24 the objective function is strictly convex (i.e., its Hessian matrix is positive definite) and the
25 constraints (1.3)-(1.5) forms a convex feasible region. However, in general this ME problem
26 does not have a unique path flow solution.

1 We use a toy network shown in Figure 1 to examine the optimal conditions of the ME problem
 2 ([P1](#)).

3 Given five potential O-D pairs, (1, 2), (1, 3), (1, 4), (2, 3) and (4, 3), this ME problem is written
 4 as follows:

$$\min \sum_{rs} (x_{rs} \ln x_{rs} - x_{rs})$$

5 where $x_{rs} = x_{12}, x_{13}, x_{14}, x_{23}$ and x_{43} ,

$$\text{subject to } f_{1-2} + f_{1-2-3} = 2$$

$$f_{2-3} + f_{1-2-3} = 2$$

$$f_{1-3} = 3$$

$$f_{1-4} + f_{1-4-3} = 1$$

$$f_{4-3} + f_{1-4-3} = 1$$

$$f_{1-2}, f_{1-3}, f_{1-4}, f_{1-2-3}, f_{1-4-3}, f_{2-3}, f_{4-3} \geq 0$$

6 where the O-D flow variables can be decomposed into path flow variables,

$$x_{12} = f_{1-2}$$

$$x_{23} = f_{2-3}$$

$$x_{13} = f_{1-3} + f_{1-2-3}$$

$$x_{14} = f_{1-4}$$

$$x_{43} = f_{4-3}$$

7 This numerical problem can be solved analytically as follows: Given $x_{12} = x_{23}$, $x_{14} = x_{43}$ and
 8 $x_{12} + x_{13} + x_{14} = 6$, one can reduce the objective function to $2(x_{12} \ln x_{12} - x_{12}) +$
 9 $2(x_{14} \ln x_{14} - x_{14}) + (6 - x_{12} - x_{14}) \ln(6 - x_{12} - x_{14}) - (6 - x_{12} - x_{14})$. This single-
 10 objective minimization problem can be readily solved by checking its partial gradient subject to
 11 $0 \leq x_{12} \leq 2$ and $0 \leq x_{14} \leq 1$, which results in $x_{12}^* = 1.791$ and $x_{14}^* = 1$. Moreover, $x_{23}^* =$
 12 1.791 , $x_{43}^* = 1$ and $x_{13}^* = 3.209$ as well.

If one examines, for example, O-D pair (1, 3), the minimum entropy impedance of this O-D pair is $-\ln x_{13}^* = -1.166$. There are three paths between O-D pair (1, 3): 1-3, 1-2-3 and 1-4-3. The entropy impedance of path 1-3 is just the entropy impedance of link 1-3, which is obviously equal to the minimum entropy impedance. The entropy impedance of path 1-2-3 is the sum of the impedance values of links 1-2 and 2-3, $-\ln x_{12}^* - \ln x_{23}^* = -1.166$, which is equal to the minimum entropy impedance. However, the entropy impedance of path 1-4-3 is the sum of those of links 1-4 and 4-3, $-\ln x_{14}^* - \ln x_{43}^* = 0$, which is greater than -1.166 . This result means that between O-D pair (1, 3) there exist positive path flows on paths 1-3 and 1-2-3 while no flow on path 1-4-3. In fact, the path flow pattern for O-D pair (1, 3) is $f_{1-3}^* = 3$, $f_{1-2-3}^* = 0.209$ and $f_{1-4-3}^* = 0$.

SOLUTION ALGORITHM

The Frank-Wolfe algorithm (43) can be adapted for solving the ME problem (P1) defined in this text. The modified algorithmic steps for the ME problem are depicted as follows.

Step 0 (Initialization): Find an initial feasible O-D trip matrix. One possible initial trip matrix can be obtained by setting $x_{rs} = \hat{v}_a$, if nodes r and s are the head and tail nodes of some link a , i.e., $a = (r, s)$, and $x_{rs} = 0$, for all other O-D pairs.

Step 1 (Direction finding): Find an auxiliary trip matrix y_{rs} , $\forall r \in R, s \in S$, by solving the following linearized problem (P2):

$$\min_{rs} \sum_{rs} y_{rs} \ln x_{rs}^n \quad (5.5)$$

$$\text{subject to } \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} = \hat{v}_a \quad \forall a \in A \quad (5.6)$$

$$f_k^{rs} \geq 0 \quad \forall k \in K_{rs}, r \in R, s \in S \quad (5.7)$$

where trip rate y_{rs} is defined as

$$y_{rs} = \sum_k f_k^{rs} \quad \forall k \in K_{rs} \quad (5.8)$$

Step 2 (Line search): Find an optimal α value for $0 \leq \alpha \leq 1$ by solving the following line search problem:

$$\min \sum_{rs} [x_{rs}^n + \alpha(y_{rs} - x_{rs}^n)] \ln[x_{rs}^n + \alpha(y_{rs} - x_{rs}^n)] - [x_{rs}^n + \alpha(y_{rs} - x_{rs}^n)] \quad (6.1)$$

$$\text{subject to } 0 \leq \alpha \leq 1 \quad (6.2)$$

1 Step 3 (Solution update): Set $x_{rs}^{n+1} = x_{rs}^n + \alpha(y_{rs} - x_{rs}^n)$.

2 Step 4 (Convergence test): If a convergence criterion is met (for example, $\sum_{rs} \frac{|x_{rs}^{n+1} - x_{rs}^n|}{x_{rs}^{n+1}} < \varepsilon$),
3 stop; otherwise, go to step 1.

4 It should be noted that the computational bottleneck of the Frank-Wolfe algorithm in solving the
5 ME problem is the linearized ME subproblem formed in step 1. The standard linear
6 programming (LP) solution method — the simplex method — may not be directly applied to this
7 linear problem, because an explicit statement and processing of such an LP problem requires
8 enumeration of all possible path flows between each O-D pair, which is computationally
9 prohibitive for problems of realistic network size. For this reason, an efficient approach that
10 avoids path enumeration is required; otherwise, the application of the Frank-Wolfe algorithm for
11 the ME problem may be limited to subnetworks of small size only.

12 To relax the computational difficulty, this work resorts to the column generation approach,
13 which generates path flows only as and when needed within the solution framework of the
14 revised simplex method (see, e.g., 44 and 45). Given that the linearized problem is in the form
15 of path flows, we label the path set of the network as $P = \bigcup_{r \in R, s \in S} K_{rs}$. Since the optimal
16 solution of this linearized problem is a basic feasible solution, it is readily known that there are at
17 most $|A|$ paths with positive flow rate in the optimal solution.

18 For convenience, one can first rewrite the linearized ME problem (P2) into the following path-
19 based matrix form:

$$\min \mathbf{c}^T \cdot \mathbf{f} \quad (7.1)$$

20 where \mathbf{c} is the negative of the path entropy impedance vector, $\mathbf{c} = [c_{rs}^n]_{|P| \times 1} = [\ln x_{rs}^n]_{|P| \times 1}$, and
21 \mathbf{f} is the path flow vector, $\mathbf{f} = [f_k^{rs}]_{|P| \times 1}$,

$$\text{subject to } \Delta \cdot \mathbf{f} = \hat{\mathbf{v}} \quad (7.2)$$

$$\mathbf{f} \geq \mathbf{0} \quad (7.3)$$

22 where Δ is the link-path incidence matrix, $\Delta = [\delta_{a,k}^{rs}]_{|A| \times |P|}$, and $\hat{\mathbf{v}}$ is the estimated link flow
23 vector, $\hat{\mathbf{v}} = [\hat{v}_a]_{|A| \times 1}$.

1 Suppose that we are at some iteration of the simplex procedure, where the current basic feasible
 2 solution contains $|A|$ basic paths of positive flow. The sets of basic paths and nonbasic paths are
 3 labeled P_B and $P_{\bar{B}}$, respectively. Suppose that the corresponding basis matrix and cost vector are
 4 \mathbf{B} and \mathbf{c}_B , where $\mathbf{B} = [\delta_{a,k}^{rs}]_{|A| \times |A|}$ and $\mathbf{c}_B = [\ln x_{rs}^n]_{|A| \times 1}$. Given the simplex multiplier vector
 5 $\mathbf{w} = \mathbf{c}_B \mathbf{B}^{-1}$, one knows that the reduced cost for a nonbasic path flow variable f_k^{rs} is $c_k^{rs} -$
 6 $z_k^{rs} = \ln x_{rs}^n - \mathbf{c}_B \mathbf{B}^{-1} \Delta_k^{rs}$, where Δ_k^{rs} is the corresponding column of Δ to the nonbasic path k . It
 7 is readily known that if all reduced costs $c_k^{rs} - z_k^{rs} \geq 0, \forall k \in P_{\bar{B}}$, the current basic feasible
 8 solution is optimal; otherwise, one may increase the path flow rate of a nonbasic path with
 9 $c_k^{rs} - z_k^{rs} < 0, k \in P_{\bar{B}}$ from 0 to some positive level so that the objective function value is
 10 decreased while the problem feasibility is maintained. In the latter case, a nonbasic path with the
 11 lowest reduced cost value may be chosen for this purpose, according to Dantzig's rule. Without
 12 enumerating all the nonbasic paths in the set $P_{\bar{B}}$, Dantzig's rule can be implemented by solving
 13 the following minimization problem:

$$\min_{k \in P} \{c_k^{rs} - z_k^{rs}\} \quad (8.1)$$

14 which can be further decomposed into a set of minimization problems by O-D pairs:

$$\min_{r \in R, s \in S} \left\{ \dots, \min_{k \in K_{rs}} \{c_k^{rs} - z_k^{rs}\}, \dots \right\} \quad (8.2)$$

15 Note that the minimization problem for each O-D pair $r-s$ is essentially a shortest path problem
 16 (P4), as follows, given that $c_k^{rs} = \ln x_{rs}^n$ is fixed for all paths between O-D pair $r-s$:

$$\min -z_k^{rs} = -\mathbf{c}_B \mathbf{B}^{-1} \Delta_k^{rs} \quad (9.1)$$

$$\text{subject to } k \in K_{rs} \quad (9.2)$$

17 where Δ_k^{rs} is the link-path incidence vector of path k between O-D pair $r-s$, which exists in Δ as
 18 a column. It is obvious that for this shortest path problem, the negative of the simplex multiplier
 19 vector $-\mathbf{w} = -\mathbf{c}_B \mathbf{B}^{-1}$ specifies the link costs over the network. It should be noted that an
 20 arbitrary element in \mathbf{w} (or $-\mathbf{w}$) (i.e., the cost of an arbitrary link) may be positive or negative, so
 21 a shortest path algorithm that prevents negative cost loops is needed.

22 After executing the shortest path search for each O-D pair, one can then obtain the entering path
 23 flow variable (to the basis matrix) with the lowest $c_k^{rs} - z_k^{rs}$ value over all O-D pairs, which
 24 generates a new column for the basis matrix, Δ_l^{od} . The remaining algorithmic issue is to
 25 determine the value of the entering path flow variable and accordingly identify a leaving path
 26 flow variable (from the basis matrix). Suppose that the entering path is l between O-D pair $o-d$

1 and its flow rate and the link-path incidence vector are f_l^{od} and Δ_l^{od} , respectively. Then the
 2 leaving path flow variable is the one that maximizes the f_l^{od} value while maintaining the
 3 problem feasibility (i.e., all the basic feasible path flow variables must be greater than or equal to
 4 0):

$$\max\{f_l^{od} : \mathbf{f}_B = \mathbf{B}^{-1}\hat{\mathbf{v}} - (\mathbf{B}^{-1}\Delta_l^{od})f_l^{od} \geq \mathbf{0}\} \quad (10)$$

5 where \mathbf{f}_B is the vector of path flow variables corresponding to the current basis matrix and $\hat{\mathbf{v}}$ is
 6 the link flow vector. Since $\mathbf{B} \geq \mathbf{0}$ (where each element $\delta_{a,k}^{rs}$ in \mathbf{B} is equal to 1 or 0), the
 7 inequality in (10) is reduced to $\mathbf{v} - \Delta_l^{od}f_l^{od} \geq \mathbf{0}$, which in turn results in:

$$(f_l^{od})_{\max} = \min\{\hat{v}_a/\delta_{a,l}^{od} : \delta_{a,l}^{od} = 1, \forall a\} \quad (11)$$

8 This result implies that $(f_l^{od})_{\max}$ should be set to equal the minimum link flow along path l .
 9 Accordingly, the path flow variables in the current basis matrix should be updated by $\mathbf{f}_B =$
 10 $\mathbf{B}^{-1}\mathbf{v} - (\mathbf{B}^{-1}\Delta_l^{od})(f_l^{od})_{\max}$, in which the path flow variable whose value is decreased to 0 is the
 11 leaving variable.

12 The algorithmic steps of the column generation approach described above can be summarized as
 13 follows, which synthetically serve as step 1 of the Frank-Wolfe solution framework:

14 *Step 1.1 (Initialization):* Find an initial, feasible O-D trip matrix for the linearized problem.
 15 Such an initial trip matrix can be obtained by setting $f_k^{rs} = \hat{v}_a$ for such a path k between such an
 16 O-D pair $r-s$ that nodes r and s are the head and tail nodes of some link a (i.e., $a = (r,s)$) and
 17 path k contains link a only (i.e., $\delta_{a,k}^{rs} = 1$) and $\delta_{b,k}^{rs} = 0$, $\forall b \neq a$, and by setting $f_l^{rs} = 0$,
 18 $\forall l \neq k$ between O-D pair $r-s$. The values of all other path flow variables are set to be 0.

19 *Step 1.2 (Entering path choosing):* Solve a shortest path problem defined in (9) for each O-D pair
 20 and identify entering path flow variable f_k^{rs} with the minimum $c_k^{rs} - z_k^{rs}$ value over all O-D
 21 pairs. If the minimum $c_k^{rs} - z_k^{rs}$ value is greater than or equal to 0, the current basic feasible
 22 solution is optimal; otherwise, go to step 1.3.

23 *Step 1.3 (Leaving path choosing):* Compute the value of the entering path flow variable by
 24 $(f_l^{od})_{\max} = \min\{v_a/\delta_{a,l}^{od} : \delta_{a,l}^{od} = 1\}$ and identify the leaving path flow variable whose value is
 25 decreased to 0.

26 *Step 1.4 (Basis matrix updating):* Update the basic feasible path flow variables by $\mathbf{f}_B = \mathbf{B}^{-1}\mathbf{v} -$
 27 $(\mathbf{B}^{-1}\Delta_l^{od})f_l^{od}$ and update the basis matrix by inserting the entering path's link-path incidence
 28 vector and removing the leaving path's link-path incidence vector.

1 **NUMERICAL EVALUATION**

2 In this section, the solution method’s performance is evaluated using a numerical example (the
3 Sioux Falls network). The full network has 24 nodes and 76 links, while the subnetwork
4 includes 12 nodes and 34 links (see Figure 2). The subnetwork approximately covers the
5 downtown area of the City of Sioux Falls.

6 UE traffic assignment over the full network was conducted to estimate a link flow pattern
7 (Figure 2), which was then used as input data set for the subnetwork trip matrix estimation model.
8 The Frank-Wolfe algorithm with column generation was then applied to generate an EM trip
9 matrix with 121 O-D pairs for the subnetwork. The estimation’s accuracy was then indirectly
10 assessed by applying the estimated subnetwork trip matrix to generate a traffic flow pattern and
11 comparing it to that generated by the full-network traffic assignment.

12 For evaluation purposes, a list of synthetic network upgrade scenarios was developed, including
13 both capacity expansion and link addition cases, as shown in Table 1. The flow pattern
14 comparison results for these upgrade scenarios are plotted in Figure 3. Two performance
15 measures are used here, to indicate the difference between the full-network and subnetwork flow
16 patterns: R^2 -squared value (R^2) from linear regression (of full-network flows following network
17 change on subnet estimates) and root mean square error (RMSE). Among these various
18 scenarios, R^2 values range from 0.963 to 0.993, indicating a very close match between the full-
19 network and subnetwork link flow patterns. Similarly, the RMSE values always like below 10%,
20 across all scenarios, implying high accuracy in subnetwork assignment results. These results
21 allow us to conclude that such a subnetwork model serves as a good approximation to its
22 corresponding full-network model for sketch planning purposes.

23 It should be noted that the numerical experiment involves rather significant network upgrades,
24 effectively representing what is likely to serve as “worst-case” situations, in terms of model
25 performance. In networks of larger or more realistic size, or where upgrades represent less of
26 change, one can expect that flow estimates will lie even closer to their full-network counterparts,
27 on average, since the impacts of network upgrades will diminish in relation to the larger,
28 relatively stable, network topology.

29 **MODELING EXTENSIONS**

30 There are important modeling possibilities that relax the two demand generation assumptions
31 embedded in the proposed ME model. These are restriction-free demand generation and the
32 invariant-demand trip matrix.

33 Recall that the first assumption allows any link flow errors or imbalances and inconsistencies to
34 be absorbed by relevant O-D pairs without affecting the problem’s solution feasibility and
35 estimation efficiency. However, in a realistic network, some “intermediate” nodes cannot be

either an origin or destination node (e.g., the point of an off-ramp from a freeway link). If one adds this restriction to the model, it may result in intermediate nodes where in-flow total does not match out-flow, thus not satisfying the flow conservation principle. One can use I to denote the set of such intermediate nodes, where $I \subset N$, $I \cap R = \emptyset$ and $I \cap S = \emptyset$. The following, enhanced subnetwork trip matrix model (P3) can accommodate this data inconsistency issue while guaranteeing solution feasibility:

$$\min \sum_{rs} (x_{rs} \ln x_{rs} - x_{rs}) + w \sum_a ((y_a^+)^2 + (y_a^-)^2) \quad (12.1)$$

$$\text{subject to } \sum_{rs} \sum_k f_k^{rs} \delta_{a,k}^{rs} + y_a^+ - y_a^- = \hat{v}_a \quad \forall a \in A \quad (12.2)$$

$$f_k^{rs} \geq 0 \quad \forall k \in K_{rs}, r \in R, s \in S \quad (12.3)$$

$$y_a^+, y_a^- \geq 0 \quad \forall a \in A \quad (12.4)$$

where x_{rs} is defined as

$$x_{rs} = \sum_k f_k^{rs} \quad \forall r \in R, s \in S \quad (12.5)$$

$$\sum_s x_{rs} = 0 \quad \forall r \in I \quad (12.6)$$

$$\sum_r x_{rs} = 0 \quad \forall s \in I \quad (12.7)$$

In this enhanced model, the added non-negative artificial variables y_a^+ and y_a^- represent on link a the difference between link a 's input flows and its estimated link flow, consistent with the estimated trip matrix. The error-minimizing term added into the weighted objective function, $\sum_a ((y_a^+)^2 + (y_a^-)^2)$, seeks to steer the estimated link flow pattern as close as possible to the given link flow pattern, in the form of least squares. Here, w is a weighting coefficient that indicates the relative preference for or strength of the error-minimizing term to the entropy-maximizing term.

Relaxation of the second assumption (that trip tables are fixed) adds an extra modeling dimension—trip generation—into the subnetwork trip matrix estimation problem, which favors a more general demand modeling condition. Without loss of generality, one can categorize subnetwork trips into four groups, in terms of their departure and arrival locations: 1) internal-

1 internal flows; 2) internal-external flows; 3) external-internal flows; and 4) external-internal
 2 flows. Assuming that the O-D trip matrix in the full network is known and fixed, we know that
 3 the internal-internal O-D flow rates are also known and the trip production rates of all internal-
 4 external O-D pairs and the trip attraction rates of all external-internal pairs are known. The
 5 remaining tasks are how to distribute the internal-external O-D flows to their origins, distribute
 6 the external-internal O-D flows to their destinations, and distribute the external-external between
 7 their candidate origins and destinations, where the origins of external-internal O-D flows, the
 8 destinations of internal-external O-D flows and the origins and destinations of external-external
 9 O-D flows are the trip entry and egress points to the subnetwork.

10 In view of such an O-D flow structure embedded within a subnetwork, we propose a combined
 11 trip distribution and traffic assignment model, of which the trip distribution is still based on the
 12 ME theory and the traffic assignment follows the UE principle. Given the input data sets (i.e.,
 13 internal-internal O-D flow rate $\hat{x}_{rs}^{(i-i)}$, $\forall r \in R_i, s \in S_i$, internal-external production rate $\hat{o}_r^{(i-e)}$,
 14 $\forall r \in R_i$, external-internal attraction rate $\hat{d}_s^{(e-i)}$, $\forall s \in S_i$, and maximum external-external O-D
 15 flow rate $\hat{q}_{rs}^{(e-e)}$, $\forall r \in R_e, s \in S_e$, where R_i and S_i are the internal origin and destination node
 16 sets, and R_e and S_e are the external origin and destination node sets, respectively), the combined
 17 model (P4) is given as follows:

$$\begin{aligned} \min \quad & \sum_{rs} x_{rs}^{(i-e)} \ln x_{rs}^{(i-e)} + \sum_{rs} x_{rs}^{(e-i)} \ln x_{rs}^{(e-i)} \\ & + \sum_{rs} (x_{rs}^{(e-e)} \ln x_{rs}^{(e-e)} + x'_{rs}^{(e-e)} \ln x'_{rs}^{(e-e)}) \\ & + w_1 \sum_a \int_0^{v_a} t_a(\omega) d\omega + w_2 \sum_{rs} \int_0^{x'_{rs}^{(e-e)}} u'_{rs}(\omega) d\omega \end{aligned} \quad (13.1)$$

$$\text{subject to } \sum_k f_k^{rs(i-i)} = \hat{x}_{rs}^{(i-i)} \quad \forall r \in R_i, s \in S_i \quad (13.2)$$

$$\sum_s \sum_k f_k^{rs(i-e)} = \hat{o}_r^{(i-e)} \quad \forall r \in R_i \quad (13.3)$$

$$\sum_r \sum_k f_k^{rs(e-i)} = \hat{d}_s^{(e-i)} \quad \forall s \in S_i \quad (13.4)$$

$$\sum_k f_k^{rs(e-e)} + x'_{rs}^{(e-e)} = \hat{q}_{rs}^{(e-e)} \quad \forall r \in R_e, s \in S_e \quad (13.5)$$

$$f_k^{rs(i-i)}, f_k^{rs(i-e)}, f_k^{rs(e-i)}, f_k^{rs(e-e)}, x'_{rs}^{(e-e)} \geq 0 \quad (13.6)$$

1 where O-D flow rates $x_{rs}^{(i-e)}$, $x_{rs}^{(e-i)}$, $x_{rs}^{(e-e)}$ and $x'_{rs}^{(e-e)}$, and link flow rate v_a are defined as,

$$x_{rs}^{(i-e)} = \sum_k f_k^{rs(i-e)} \quad \forall r \in R_i, s \in S_e \quad (13.7)$$

$$x_{rs}^{(e-i)} = \sum_k f_k^{rs(e-i)} \quad \forall r \in R_e, s \in S_i \quad (13.8)$$

$$x_{rs}^{(e-e)} = \sum_k f_k^{rs(e-e)} \quad \forall r \in R_e, s \in S_e \quad (13.9)$$

$$x'_{rs}^{(e-e)} = d'_{rs}^{(e-e)}(u'_{rs}) \quad \forall r \in R_e, s \in S_e \quad (13.10)$$

$$v_a = \sum_{rs} \sum_k f_k^{rs(i-i)} + \sum_{rs} \sum_k f_k^{rs(i-e)} + \sum_{rs} \sum_k f_k^{rs(e-i)} + \sum_{rs} \sum_k f_k^{rs(e-e)} \quad \forall a \in A \quad (13.11)$$

2 Note that $f_k^{rs(i-i)}$, $f_k^{rs(i-e)}$, $f_k^{rs(e-i)}$, $f_k^{rs(e-e)}$, and $x'_{rs}^{(e-e)}$ are decision variables of the model,
3 among which $x'_{rs}^{(e-e)}$ represents the amount of external-external flows that choose not going
4 through the subnetwork. Moreover, it is assumed that link cost t_a is a convex, increasing
5 function of link flow rate x_a , and synthetic external-external O-D cost u'_{rs} is also a convex,
6 increasing function of the relevant congestion level outside the subnetwork affecting the route
7 choice of potential travelers choosing a path between nodes r and s .

8 The validity of these two models can be readily proven by applying standard convex analysis
9 techniques. Due to space limitations, detailed analyses are omitted here. However, it is worth
10 mentioning here that the increasing complexity of the model's structure can still be
11 accommodated by existing solution methods: the enhanced trip matrix estimation problem (P3)
12 can be solved by a similar procedure to the linearization algorithm (43) presented in this text,
13 while the elastic-demand trip matrix estimation problem (P4) can be solved by the partial
14 linearization algorithm (46).

15 CONCLUSIONS AND FURTHER WORK

16 In the domain of sketch planning, a network abstraction (or subtraction) process should satisfy
17 two criteria: 1) the simplified network should be small enough (in terms of the network size) so
18 that it can be managed and processed efficiently; 2) the simplified network should preserve
19 important network characteristics and behaviors so that the impacts of any network change can

1 be properly reflected. Here, the focus is on developing a subnetwork trip matrix that provides
2 link flows consistent with flow observations (from field surveys) or flow estimates (generated by
3 a full-network model).

4 Here, an ME model was suggested for subnetwork trip matrix estimation and a solution method
5 based on the Frank-Wolfe algorithm was devised. The adapted Frank-Wolfe algorithm requires
6 efficient solution of a linearized ME problem in each of its iterations. We proposed a column
7 generation approach that relaxes the minimum reduced cost search during the course of solving
8 the linearized problem to a set of shortest path problems, thus avoiding a computationally
9 prohibitive path enumeration process. The modeling philosophy and solution performance are
10 illustrated via a numerical example, under multiple network modification scenarios, the results of
11 which support the validity and accuracy of our subnetwork modeling methodology for sketch
12 planning.

13 The basic model and enhancements discussed here are likely to prove useful to much larger scale
14 tests, offering an opportunity for thoughtful, real-time network evaluations. Such tools should
15 prove helpful not only for quickly evaluating a variety of network improvement projects, but also
16 work zone constraints, operations policies, evacuation procedures, and other instances of
17 network changes, potentially in a real-time setting.

18 **ACKNOWLEDGEMENTS**

19 The authors are grateful for funding support from the Texas Department of Transportation, under
20 Research Project No. 0-6235 (titled “Sketch Planning Techniques to Assess Regional Air Quality
21 Impacts of Congestion Mitigation Strategies”). They also appreciate the use of Professor Hillel
22 Bar-Gera’s origin-based computer code for all traffic assignment results in the paper.

23

24

25

26

27

28

29

30

31

1 REFERENCES

- 2 [1] Chan, Y. A Method to Simplify Network Representation in Transportation Planning.
3 *Transportation Research*, Vol. 10, No. 3, 1976, pp. 179-191.
- 4 [2] Chan, Y., T.S. Shen, and N.M. Mahaba. Transportation Network Design Problem:
5 Application of a Hierarchical Search Algorithm. In *Transportation Research Record: Journal of*
6 *Transportation Research Board*, No. 1251, 1989, pp. 24-34.
- 7 [3] Eash, R.W., K.S. Chon, Y.J. Lee, and D.E. Boyce. Equilibrium Traffic Assignment on an
8 Aggregated Highway Network for Sketch Planning. In *Transportation Research Record: Journal of*
9 *Transportation Research Board*, No. 944, 1983, pp. 30-37.
- 10 [4] Kaplan, M.P., Y.J. Gur, and A.D. Vyas. Sketch Planning Model for Urban Transportation
11 Policy Analysis. In *Transportation Research Record: Journal of Transportation Research Board*,
12 No. 952, 1984, pp. 32-39.
- 13 [5] Haghani, A.E., and M.S. Daskin. Network Design Application of an Exaction Algorithm for
14 Network Aggregation. In *Transportation Research Record: Journal of Transportation Research*
15 *Board*, No. 944, 1984, pp. 37-46.
- 16 [6] Haghani, A.E., and M.S. Daskin. Aggregation Effects on the Network Design Problem.
17 *Journal of Advanced Transportation*, Vol. 20, No. 3, 1986, pp. 239-258.
- 18 [7] Taylor, M.A.P., W. Young, and P.W. Newton. PC-Based Sketch Planning Methods for
19 Transport and Urban Applications. *Transportation*, Vol. 14, No. 4, 1988, pp. 361-375.
- 20 [8] Rogus, M.J. *Building Sketch Networks: The Development of Sketch Zones Boundaries and*
21 *Aggregation Process of Detailed Highway Networks*. Working Paper No. 96-11, Chicago Area
22 Transportation Study, Chicago, IL, 1996.
- 23 [9] Dowling, R.G., and A.D. May. Comparison of Small-Area O-D Estimation Techniques. In
24 *Transportation Research Record: Journal of Transportation Research Board*, No. 1045, 1985,
25 pp. 9-15.
- 26 [10] Zhou, X., S. Erdogan, and H.S. Mahmassani. Dynamic Origin-Destination Trip Demand
27 Estimation for Subarea Analysis. In *Transportation Research Record: Journal of Transportation*
28 *Research Board*, No. 1964, 2006, pp. 176-184.
- 29 [11] Hearn, D.W. Practical and Theoretical Aspects of Aggregation Problems in Transportation
30 Planning Methods. *Transportation Planning Models*, M. Florian, Eds., North-Holland,
31 Amsterdam, The Netherlands, 1984, pp. 257-287.

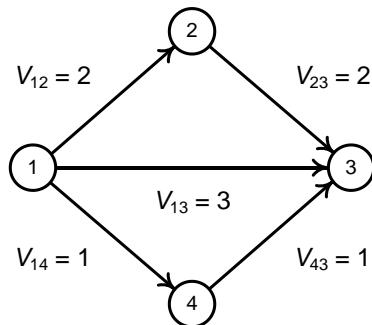
- 1 [12] Wilson, A.G. *Entropy in Urban and Regional Modeling*. Pion, London, England, 1970.
- 2 [13] Oppenheim, N. *Urban Travel Demand Modeling: From Individual Choices to General*
3 *Equilibrium*. Wiley, New York, NY, 1995.
- 4 [14] Rossi, T.F., S. McNeil, and C. Hendrickson. Entropy Model for Consistent Impact-Fee
5 Assessment. *Journal of Urban Planning and Development*, Vol. 115, No. 2, 1989, pp. 51-63.
- 6 [15] Bell, M.G.H., and Y. Iida. *Transportation Network Analysis*. Wiley, New York, NY, 1997.
- 7 [16] Janson, B.N. Most Likely Origin-Destination Link Uses from Equilibrium Assignment.
8 *Transportation Research*, Vol. 27B, No. 5, 1993, pp. 333-350.
- 9 [17] Larsson, T., J.T., Lundgren, C. Rydberg, and M. Patriksson. Most Likely Traffic
10 Equilibrium Route Flows Analysis and Computation. *Equilibrium Problems: Nonsmooth*
11 *Optimization and Variational Inequality Models*, F. Giannessi Eds., Springer, New York, NY,
12 2002, pp. 129-159.
- 13 [18] Bar-Gera, H. Primal Method for Determining the Most Likely Route Flows in Large Road
14 Networks. *Transportation Science*, Vol. 40, No. 3, 2006, pp. 269-286.
- 15 [19] Willumsen, L.G. Simplified Transport Models Based on Traffic Counts. *Transportation*,
16 Vol. 10, No. 3, 1981, pp. 257-278.
- 17 [20] Van Zuylen, H.J., and L.G. Willumsen. The Most Likely Trip Matrix Estimated from
18 Traffic Counts. *Transportation Research*, Vol. 14B, No. 3, 1980, pp. 281-293.
- 19 [21] Nguyen, S. *Modèles de Distribution Spatial Tenant Compte des Itinérarries*. Publication No.
20 225, Centre de Recherché sur les Transports, Univerité de Montréal, Montreal, Canada, 1981.
- 21 [22] Fisk, C.S. On Combining Maximum Entropy Trip Matrix Estimation with User Optimal
22 Assignment. *Transportation Research*, Vol. 22B, No. 1, 1988, pp. 69-79.
- 23 [23] Nguyen, S. *Estimation an O-D Matrix from Network Data: A Network Equilibrium*
24 *Approach*. Publication No. 60, Centre de Recherché sur les Transports, Université de Montréal,
25 Montreal, Canada, 1977.
- 26 [24] Nguyen, S. Estimating Origin-Destination Matrices from Observed Flows. *Transportation*
27 *Planning Models*, F. Florian Eds., Elsevier, Amsterdam, The Netherlands, 1984, pp. 363-380.
- 28 [25] Turnquist, M.A., and Y. Gur. Estimation of Trip Tables from Observed Link Volumes. In
29 *Transportation Research Record: Journal of Transportation Research Board*, No. 730, 1979, pp.
30 1-6.

- 1 [26] LeBlanc, L.J., and K. Farhangian. Selection of a Trip Table Which Reproduces Observed
2 Link Flows. *Transportation Research*, Vol. 16B, No. 2, 1982, pp. 83-88.
- 3 [27] Yang, H., Y. Iida, and T. Sasaki. The Equilibrium-Based Origin-Destination Matrix
4 Estimation Problem. *Transportation Research*, Vol. 28B, No. 1, 1994, pp. 23-33.
- 5 [28] Maher, N.J. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian
6 Statistical Approach. *Transportation Research*, Vol. 17B, No. 6, 1983, pp. 435-447.
- 7 [29] Lo, H.P., N. Zhang, and W.H.K. Lam. Estimation of an Origin-Destination Matrix with
8 Random Link Choice Proportions: A Statistical Approach. *Transportation Research*, Vol. 30B,
9 No. 4, 1996, pp. 309-324.
- 10 [30] Cascetta, E. Estimation of Trip Matrices from Traffic Counts and Survey Data: A
11 Generalized Least Square Estimator. *Transportation Research*, Vol. 18B, No. 4/5, 1984, pp. 289-
12 299.
- 13 [31] Hendrickson, C., and S. McNeil. Matrix Entry Estimation Errors. *Proceedings of the 9th
14 International Symposium on Transportation and Traffic Theory*, Delft, The Netherlands, 1984,
15 pp. 413-430.
- 16 [32] Bell, M.G.H. The Estimation of an Origin-Destination Matrix from Traffic Counts.
17 *Transportation Science*, Vol. 17, No. 1, 1991, pp. 198-217.
- 18 [33] Yang, H., T., Sasaki, Y. Iida, and Y. Asakura. Estimation of Origin-Destination Matrices
19 from Link Traffic Counts on Congested Networks. *Transportation Research*, Vol. 26B, No. 6,
20 1992, pp. 417-434.
- 21 [34] Bell, M.G.H. The Estimation of an Origin-Destination Matrix from Traffic Counts.
22 *Transportation Science*, Vol. 17, No. 2, 1983, pp. 198-217.
- 23 [35] Bell, M.G.H. Variances and Covariances for Origin-Destination Flows When Estimated by
24 Log-Linear Models. *Transportation Research*, Vol. 19B, No. 6, 1985, pp. 497-507.
- 25 [36] Spiess, H. A Maximum Likelihood Model for Estimating Origin-Destination Matrices.
26 *Transportation Research*, Vol. 21B, No. 5, 1987, pp. 395-412.
- 27 [37] McNeil, S., and C. Hendrickson. A Regression Formulation of the Matrix Estimation
28 Problem. *Transportation Science*, Vol. 19, No. 3, 1985, pp. 278-292.
- 29 [38] Yang, H., Y. Iida, and T. Sasaki. An Analysis of the Reliability of an Origin-Destination
30 Trip Matrix Estimated from Traffic Counts. *Transportation Research*, Vol. 25B, No. 5, 1991, pp.
31 351-363.

- 1 [39] Sherali, H.D., R. Sivanandan, and A.G. Hobeika. A Linear Programming Approach for
2 Synthesizing Origin-Destination Trip Tables from Link Traffic Volumes. *Transportation*
3 *Research*, Vol. 28B, No. 3, 1994, pp. 213-233.
- 4 [40] Nie, Y., and D.H. Lee. Uncoupled Method for Equilibrium-Based Linear Path Flow
5 Estimator for Origin-Destination Trip Matrices. In *Transportation Research Record: Journal of*
6 *Transportation Research Board*, No. 1783, 2002, pp. 72-79.
- 7 [41] Gajewski, B.J., L.R., Rilett, M.P. Dixon, and C.H. Spiegelman. Robust Estimation of
8 Origin-Destination Matrices. *Journal of Transportation and Statistics*, Vol. 5, No. 2/3, 2002, pp.
9 37-56.
- 10 [42] Carey, M., and R. Revelli. Constrained Estimation of Direct Demand Functions and Trip
11 Matrices. *Transportation Science*, Vol. 20, No. 3, 1986, pp. 143-152.
- 12 [43] Frank, M., and P. Wolfe. An Algorithm for Quadratic Programming. *Naval Research*
13 *Logistics Quarterly*, Vol. 3, No. 1-2, 1956, pp. 95-110.
- 14 [44] Dantzig, G.B. *Linear Programming and Extensions*. Princeton University Press, Princeton,
15 NJ, 1963.
- 16 [45] Bazaraa, M.S., J.J. Jarvis, and H.D. Sherali. *Linear Programming and Network Flows*.
17 Wiley, New York, NY, 1990.
- 18 [46] Evans, S. Derivation and Analysis of Some Models for Combining Trip Distribution and
19 Assignment. *Transportation Research*, Vol. 9, No. 12, 1976, pp. 241-246.

TABLE 1 List of Network Upgrading Scenarios

Scenario number	Network upgrading location, type and scale
1	Increasing capacity by 50 percent on road segments 4-11-14
2	Increasing capacity by 50 percent on road segments 5-9-10-15
3	Increasing capacity by 50 percent on road segments 6-8-16-17-19
4	Increasing capacity by 100 percent on road segments 4-5-6
5	Increasing capacity by 100 percent on road segments 11-10-16
6	Increasing capacity by 100 percent on road segments 14-15-19
7	Increasing capacity by 50 percent on road segments 10-17
8	Adding two road segments 4-9 and 9-11
9	Adding a road segment 10-14

**FIGURE 1 An Illustrative Example For The Maximum Entropy Problem**

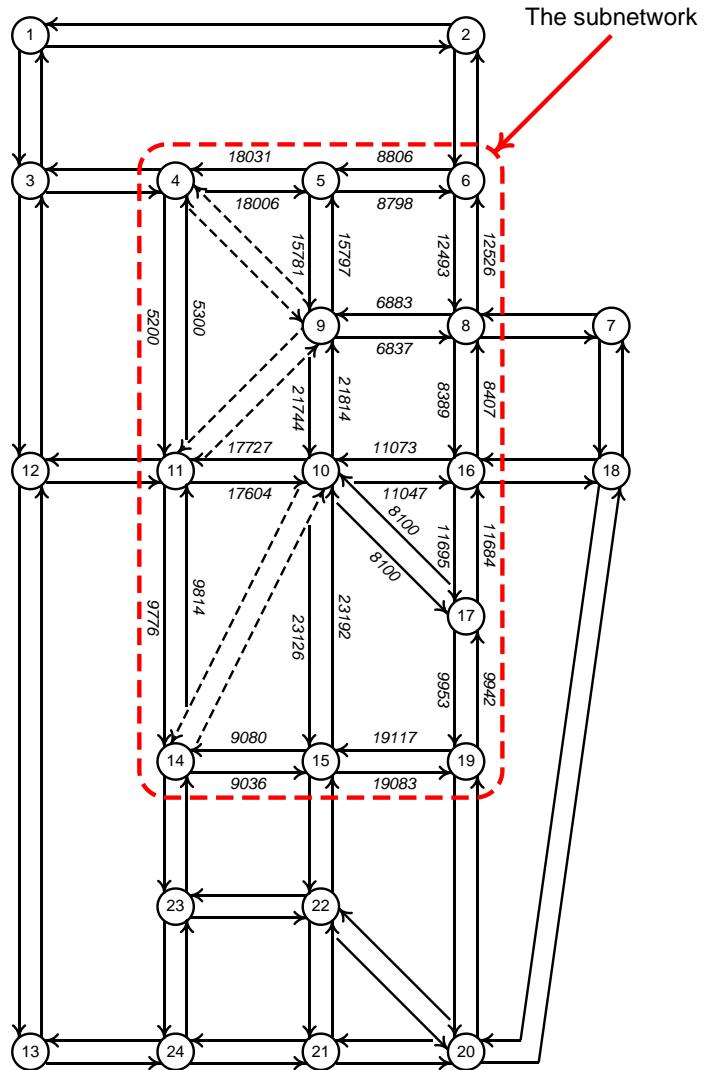


FIGURE 2 The Sioux Falls Network and Its Subnetwork Traffic Flow Pattern

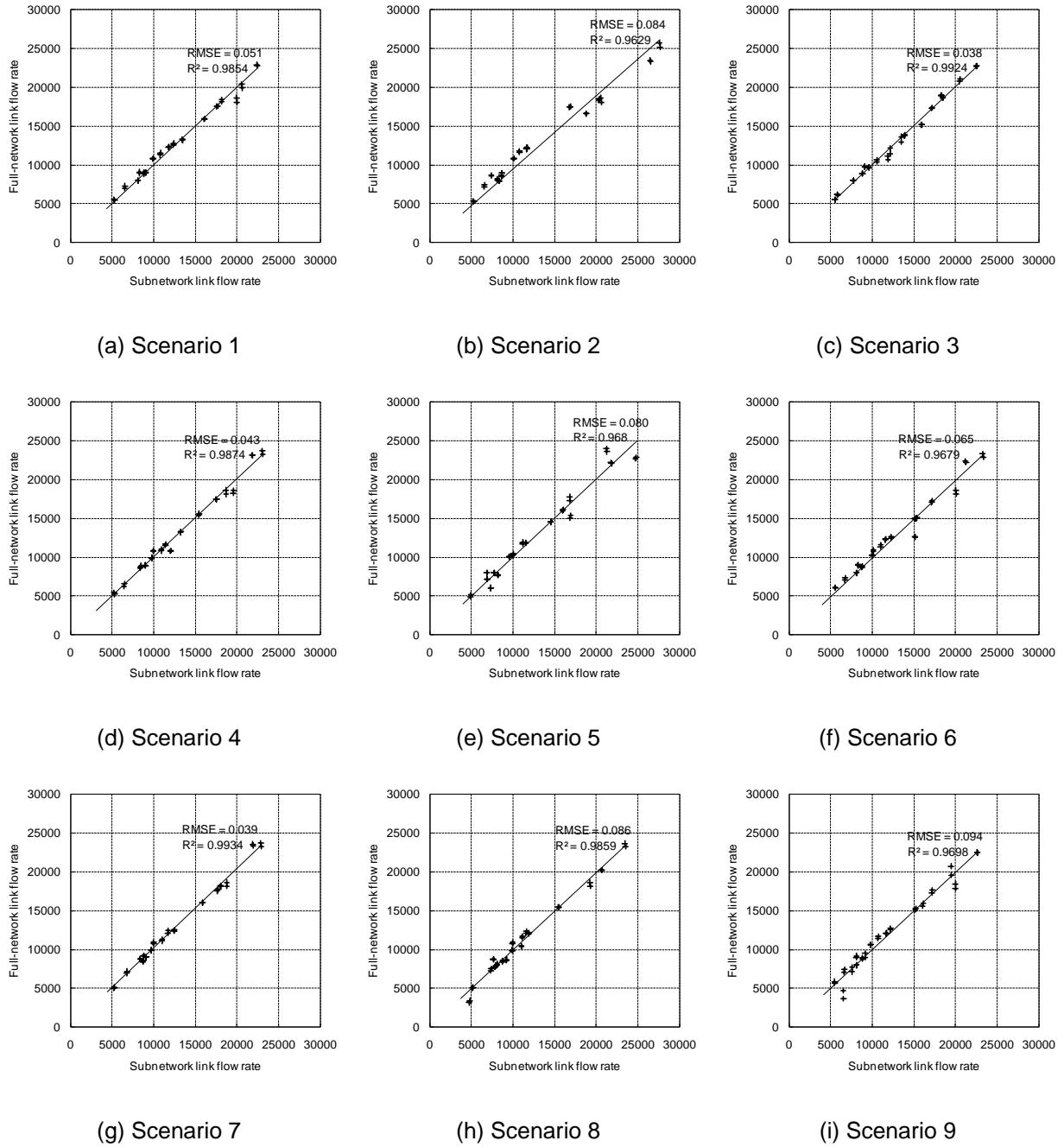


FIGURE 3 Link Flow Rates Estimated Using Full-Network And Subnetwork Traffic Assignments

1 **Daily O-D Matrix Estimation using Cellular Probe Data**

2 Yi Zhang*

3 Department of Civil and Environmental Engineering,

4 University of Wisconsin-Madison,

5 Madison, WI 53706

6 Phone: 1-608-262-2524

7 E-mail: zhang34@wisc.edu

8 Xiao Qin

9 Department of Civil and Environmental Engineering,

10 South Dakota State University

11 Brookings, SD 57007

12 Phone: 1-605-688-6355

13 E-mail: xqin@cae.wisc.edu

14 Shen Dong

15 Department of Civil and Environmental Engineering,

16 University of Wisconsin-Madison,

17 Madison, WI 53706

18 Phone: 1-608-262-2524

19 E-mail: sdong2@wisc.edu

20 Bin Ran

21 Department of Civil and Environmental Engineering,

22 University of Wisconsin-Madison,

23 Madison, WI 53706

24 Phone: 1-608-262-0052

25 E-mail: bran@wisc.edu

36 * Corresponding Author

1 **ABSTRACT**

2 With the fast-growing wireless-communication market, the cellular positioning
3 technologies are becoming one of the important means to monitoring real-time traffic status,
4 providing traveler information, measuring system operations performance, and estimating
5 travel demand. An innovative methodology is presented in this paper to estimate the daily
6 O-D demand using cellular probe trajectory information. Taking advantages of the emerging
7 cell-phone tracking technologies, the cellular probe trajectories are obtained by recording all
8 the signal-transition events and period location update events of cellular probes to determine
9 the trip origins and destinations. To apply the O-D estimation to a broader spectrum, the
10 probability of cell-phone ownership was treated as a conditional probability depending on
11 users' socio-economic factors available in the census data such as age, rage, household
12 income, etc.. A mathematic model was designed to convert the cellular counts into equivalent
13 vehicle counts, using the posterior information obtained from the characteristics of cellular
14 probe trajectories. Next, the traveling population daily O-D demand was estimated via a
15 robust Horvitz-Thompson estimator. Finally, the methodology was tested via a VISSIM
16 simulation and results were compared with a conventional simple random sampling (SRS)
17 method. The comparison outcome shows great potential of using cellular probe trajectory
18 information as a means to estimating daily O-D travel demand.

19 Key words: Daily O-D demand estimation, Cellular probe data, Cell-phone tracking
20 technology, Horvitz-Thompson estimator

1. INTRODUCTION

2 With respect to the increasing needs on the traffic demand forecasting, the estimation and
 3 prediction of O-D matrix has become an important issue in the current transportation planning
 4 and operation scope. The O-D estimation is the essential source for traffic demand information.

5 Generally, there are two types of the O-D estimation methods. One is the survey-based
 6 O-D estimation method, which utilize the trip survey data to generate the O-D matrix (1, 2).
 7 The other one is the traffic-counts-based O-D estimation method, which uses the observed link
 8 traffic counts to reversely derive the O-D matrix (3-5).

9 Traditional survey-based trip diary approach to estimating trip generation and distribution
 10 is time-consuming and cost-prohibitive (1, 6). The estimation may vary from one study as a
 11 result of the limitation of the survey sample size and sampling randomness. The counts-based
 12 methods used the existing traffic devices such as loop detectors and video cameras to obtain
 13 the link traffic counts. The O-D matrix is derived by an opposite way of traffic assignment (3).
 14 But the naturally most of the models are underdetermined (3, 7).

15 In recent years, some positioning technologies such as GPS and cell phone emerged to be
 16 used to monitor real-time traffic status(8, 9), provide traveler information (10, 11), and estimate
 17 travel demand (12-16). With the popularity of cell phone and emerging cell-phone tracking
 18 technologies, using cellular probe data have the great potential to provide a larger sample size
 19 in a timely manner.

20 Pan et al. (12) proposed a method to record the cell-phone positions every 2 hours and
 21 aggregated them to obtain the trip distribution between each O-D pairs. Caceres (13) proposed
 22 another method to record the Location Updates (LU) events to count the O-D trip flows
 23 between each Location Area (LA) and convert the cell-phone counts into vehicle counts. Sohn
 24 and Kim (14) developed an idle Handoff (HO) technology for cell-phone positioning to get the
 25 “virtual” traffic counts on observation links, and use the synthetic method to derive the
 26 time-dependent O-D matrix from the link traffic counts.

27 There are three major limitations existing in current literatures. The first one is the
 28 signal-transition events are not fully used. Since the LA includes tens of cells, and its coverage
 29 is much larger than cell, sometimes, the data fusion of the LU and HO events will increase the
 30 complexity of the O-D estimation problem. Most of literatures use either the HO events or the
 31 Location Update events (Includes periodic location update (PLU)) to determine the trajectories
 32 of cell phones or cell-phone counts. The limitation of using only one type of transition events is
 33 that it only records a part of information of cellular probe trajectories, in which case it may lead
 34 to inaccurate positioning results.

35 The second one is that the socio-economic difference of cell-phone owners is omitted.
 36 Considering the cell-phone owner group as a sample selected from the population, naturally the
 37 sample can be treated as a simple random sample. This is the so-called “Simple Random
 38 Sampling” (SRS) strategy. Most of literatures adopted this method (12, 13, 15, 16). However,
 39 whether a person owns a cell phone depends on several important factors, such as age,
 40 household income, race etc.. Disregarding the difference among those factors may leads to
 41 socio-economic “bias”.

42 The third one is that cell-phone counts are not properly converted into vehicle counts. As

we know, aggregation of the cellular probe trajectories will return the cell-phone counts. However, in transportation area, the interests mainly focus on the vehicle counts. Typically, the cell-phone counts are not equal to vehicle counts, since different vehicles may carry different number of cell-phone owners. It needs to be converted before it can be used. In current literatures, this problem is either omitted (10-12, 14), or treated by predefining an equivalent factor to do the conversions (13).

This paper proposed a method trying to cover the above three limitations. The method uses the full information on the signal-transition events to produce cellular probe trajectories. Also the socio-economic factors are taken into consideration to generate the probability of cell-phone ownerships. Then, the vehicle counts are aggregated by using the characteristics of the cellular probe trajectories.

This paper is organized as follows: The section 2 introduce the cell-phone tracking technology; The section 3 introduce the proposed method to estimate the daily O-D demand; The section 4 gives a simulation based experiment to demonstrate and verify the proposed method; The last section gives out the major conclusion of this paper.

2. CELL-PHONE TRACKING TECHNOLOGY

The cell-phone tracking technology uses the signal transition between two conterminous cells to determine the location of the object (10). Signal transition refers to a phenomenon that some parameters change their values at some “virtual” boundaries of its defined location region. In practice, a cell size and boundary changes with time due to the fluctuation of signal coverage.

Generally, in the GSM network, the parameters which can be used to track the signal transitions are Location Area Code (LAC), serving cell ID (Cell ID) and Timing Advance (TA). The corresponding signal-transition events in GSM network are Location Update (LU) for LAC, HO for Cell ID and the transition of TA values, respectively (17).

When a cell-phone with an on-going phone call crosses the boundary of different cells, a HO operation, in which the cell id and the time stamp are recorded automatically by the system, will be executed. If the cell phone is turned on but not on call, a LU event will be automatically recorded when it crosses the boundary of different Location Areas (LA). The timing advance is used to compensate for the time that takes a wireless signal to travel at the speed of light between a Base Transceiver Station (BTS) and the cell phone (17). Multiplying TA and 550 meters can give the minimum distance to a BTS. The maximum distance will be (TA+1) multiplying with 550 m. Similar to a HO, the timing advance transition can only be collected when a cell phone is in on-call mode.

In addition to the signal-transition events, the cellular system also provides a periodic location update for the cell ID information (12). Generally, the cellular system will update each cell phone's cell IDs periodically and add it into a Database. Here the location information provided by this event is the cell ID and timestamp. This event is called Periodic Location Update (PLU), and the length of the period can be adjusted by the mobile carrier. Usually, the update period is set to 2 hours by default.

Table.1 Typical signal-transition events in Figure.1

Events	Area	Timestamp
--------	------	-----------

PLU	Cell 1	08:00
TA	Cell 1	08:33
HO	Cell 1 → Cell 2	08:38
LU	Cell 8 → Cell 11	08:59
PLU	Cell 12	10:00
PLU	Cell 12	12:00
PLU	Cell 12	14:00
HO	Cell 14 → Cell 16	17:09
TA	Cell 16	17:16
LU	Cell 13 → Cell 10	17:45
PLU	Cell 4	18:00
PLU	Cell 4	20:00

Combining the above cellular location technologies, the cellular probe trajectories can be obtained by recording the signal-transition (HOs, LUs and TAs) and the PLU events. Figure.1 gives an example to illustrate the process of cellular tracking method. Assume a cell phone starts traveling at Cell 1. Its trajectory can be tracked by the signal transitions and the PLUs. Table.1 describes the different events recorded by the system.

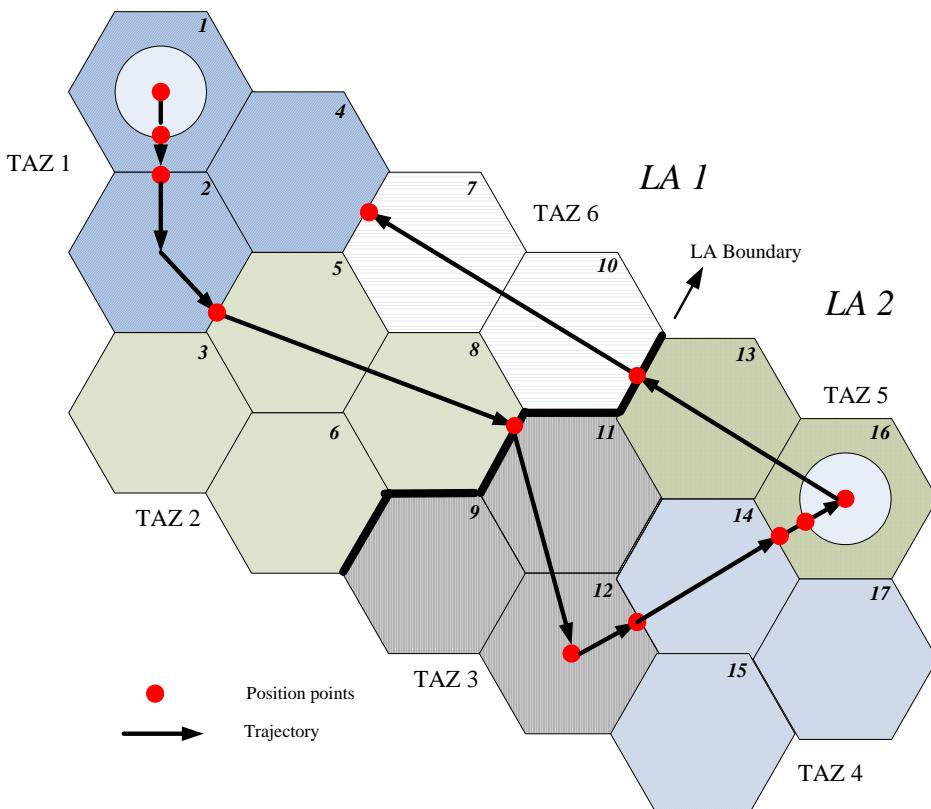


Figure.1 Illustration of cellular tracking technology

The cellular tracking method provides a possibility to record the Origin-Destination information by analyzing the trajectory of cell-phone users. Gur Y.J. et. al. (18) considered the location of the first signal transition event (mostly is the event that the first time turn on the cell phone in the morning and register to GSM network) as the trip origin. In this paper, we adopted this method to identify the trip origins. The problem is how to decide the trip end. One possible

solution is to consider the TAZs (Transportation Analysis Zone) with the longest distance from the trip origin and most PLUs recorded as the destinations. Taking the Figure.1 as an example, the trip starts at TAZ 1 since the first events happens at TAZ 1. According to Table.1, TAZ 3 has the longest distance and it has the longest duration in a day. It is easy to conclude that TAZ 3 is the trip end.

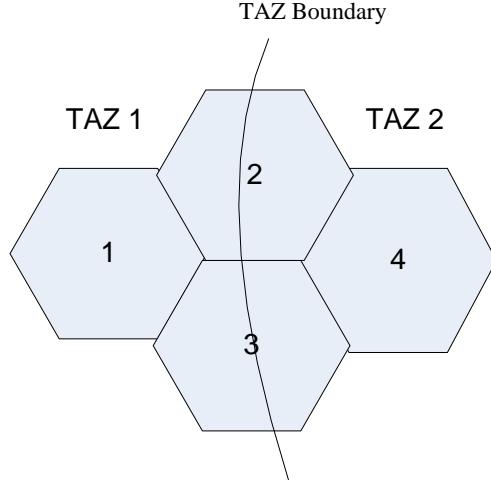


Figure.2 Illustration of imperfectly overlapping between TAZs and cells

Typically, TAZs will not match the boundaries of cells perfectly. A cell may be covered by multiple TAZs, meanwhile a TAZ may cover one or more cells entirely or just a part of a cell. As shown in Figure. 2, TAZ 1 and TAZ 2 cover the entire cell 1 and cell 4 respectively, and share the cell 2 and cell 3. If there is no signal transition event, the system can only tell the cell ID information by the PLU events, which may cause spatial errors because it is hard to determine which TAZ the cell belongs to. Pan et al. (12) used a probability of one cell belonging to a TAZ according to the proportion of area covered in each TAZ to determine it is covered by which TAZ, if there is no additional demographical information provided. We adopted this method in this paper.

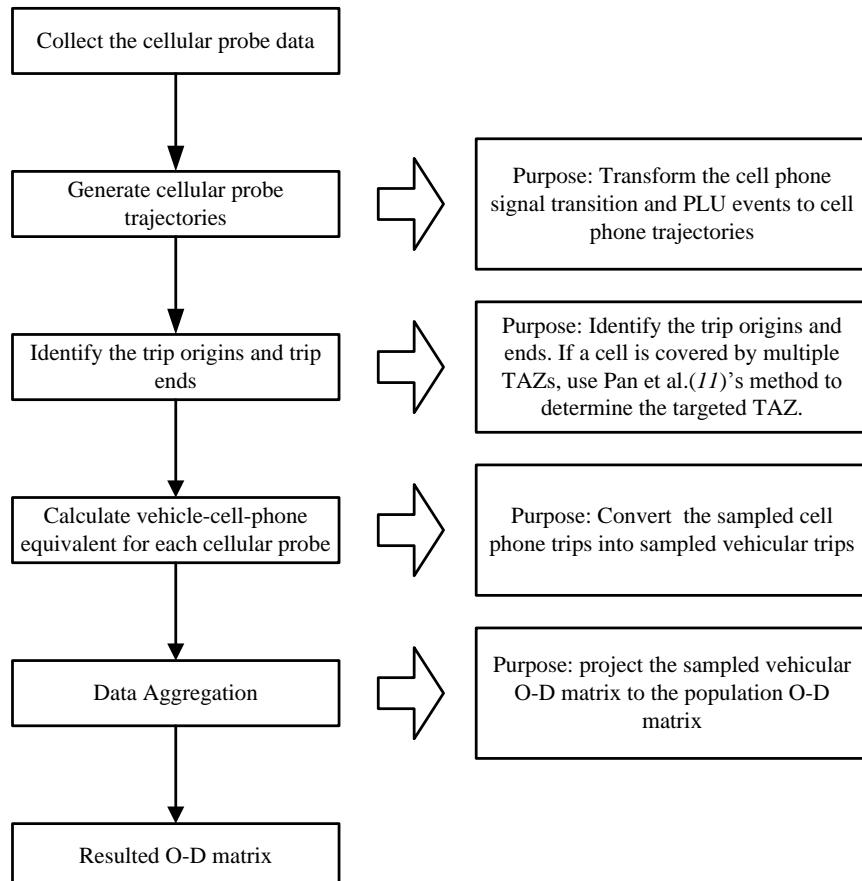
3. METHODOLOGY

3.1 Study Design

The signal-transition events and PLUs associated with the corresponding cell-phone ids and timestamps can be collected and stored in a database at the operating center of the cellular carrier. It is easy to get a specific cell-phone owner's trajectory by just doing a query in the database. The proposed method will first generate the individual cellular probe trajectory, in which the cell-phone signal transition and PLU events are recorded to form the trajectory.

After the collection of cellular probe trajectories, the identification of trip origins and ends will be executed. Then a vehicle-per-cellphone equivalent factor will be calculated to convert the cell-phone trips into vehicle trips, since a vehicle may carry different number of cell-phone owners. Till now, what we got is the cellular trips and the corresponding equivalent vehicle trips. However, those trip makers who don't own cell phones should also be considered. A data aggregation process will be carried out to project the sampled vehicle O-D matrix to the population vehicle O-D matrix. As a result, the actual O-D matrix can be obtained following

1 these above procedures. Figure.3 shows the procedures of the O-D estimation method.



2
3 Figure.3 Flow chart of the estimation process

4 Notations:

- 5 $p(cp | f_1, f_2, \dots, f_n)$ - the conditional probability of cell-phone ownership depending on factors
 6 (f_1, f_2, \dots, f_n)
- 7 $p(cp | f_n)$ - the conditional probability on factor f_n
- 8 p_{c_m} - the market share of a specific carrier m
- 9 $p(cp)$ - the market penetration of cell phones
- 10 f_{vc} - the vehicle-per-cellphone equivalent factor
- 11 \bar{f}_{pvo} - the average passenger-vehicle occupancy
- 12 N_i - the total population in TAZ i
- 13 T_{ij} - the O-D flows from TAZ i to j
- 14 \hat{T}_{ij} - the estimated value of T_{ij} of SRS method
- 15 \tilde{T}_{ij} - the estimated value of T_{ij} from cell-phone owner group of our method
- 16 S^i - the cell-phone owner group in TAZ i
- 17 p_k^i - the posterior probability of cell-phone ownership for one or several carriers of the k -th
 18 people in TAZ i . It may vary in terms of ages, income and sex.

- 1 Y_k^{ij} - the indicating variable, 1 if k -th people in population has a trip between TAZ i and j ,
 2 otherwise 0.
- 3 y_k^{ij} - the indicating variable. 1 if k -th people in cell-phone owner group has a trip between
 4 TAZ i and j , otherwise 0.
- 5 P_{ij} - the proportion of people have trips between TAZ i and j
- 6 \hat{P}_{ij} - the estimated value of P_{ij} of SRS method
- 7 \tilde{P}_{ij} - the estimated value of P_{ij} from cell-phone owner group

8 ***3.2 Important Assumptions***

9 Before introducing the daily O-D demand estimation method, two important assumptions
 10 should be made in order to make the cell-phone tracking method can be used to estimate the
 11 O-D demand between TAZ pairs.

- 12 1. *There might be multiple cellular carriers existing in the research areas. Each of the*
 carriers is operated independently. It means that the owner groups and the signal coverage
 of each cellular carrier are independent. In this case, each cell-phone owner group which
 belongs to a specific cellular carrier can only be treated as individual sample set.
- 13 2. *The cell-phone ownership pattern is identically distributed among different cellular*
 carriers. That means different cellular carriers have the same distribution of their owner's
 age, income, race etc.. This assumption holds in the general conditions, although there are
 some cellular carriers having different distribution in terms of the subscriber's
 demographics, such as MetroPCS has a heavy emphasis on prepaid phone plans, Nextel
 had a strong business focus.

14 The first assumption guarantees the generation of cellular probe trajectories and the
 15 identification of trip origins and destinations can be carried out independently among various
 16 cellular carriers. After the trip origins and destinations are determined, the difference on the
 17 signal coverage of different carriers will no longer influence the accuracy estimation results.
 18 The second assumption guarantees the generation of the cell-phone ownership probabilities can
 19 be applied to multiple carriers.

20 ***3.3 Determine the Probability of Cell-phone Ownership***

21 In traditional trip survey methods, the SRS strategy are generally adopted to design the trip
 22 surveys, in which at most 5% sampling rate are used to get the unbiased estimation of trips in a
 23 TAZ. But sometimes a lower sampling rate makes the sampling error intolerable.

24 The cellular probe data gives another way to aggregate the trip data because of its unique
 25 advantages:

- 26 1. Cellular probe data is easy to be collected.
- 27 2. The size of cell-phone owner group is much larger than the sample size in traditional
 surveys.

28 For example, in United States, the number of cell-phone users reaches approximately 87%
 29 of the total population in 2008 (19), in which the major 3 cellular carriers add up to nearly 80%
 30 market share (Verizon: 32%, AT&T: 29%, Sprint: 18%) (20). Considering the large size of the

existing cell-phone owner group, it would be clear that the sampling error should be substantially less than the traditional O-D surveys. It should be noticed that although the cell phone market penetration rate reaches 80% of the market share, in practice, since the cellular probe data are collected independently among different cellular carriers, it is more possible to collect the data from one or two cellular carriers. Therefore, we need to consider both the cell phone market penetration rate and the market share of individual cellular carriers.

Many researches treated the cellular probe trajectory data as the SRS survey data. Each individual in the sample is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process. In other words, each individual has the same probability to own a cell phone. Here the probability of owing a cell phone is a prior probability and equals to the cell-phone market penetration rate. However, the market penetration rate is a kind of prior probability which is obtained from some market research reports or papers. It may lead to inaccuracy if disregarding the possible social-economical bias. Typically, the probability of whether a person owns cell phones should be related to many factors, such as age, income and race, etc. (21). For example, young people consist of the largest group of the cell-phone owners in terms of the cell-phone owners' age distribution. Therefore, the young people will have larger probability to own cell phones than the old ones. And in some cases, the higher income people will have larger probability to own cell phones.

The conditional probability of cell-phone ownership can be assumed to have the following linear relationship:

$$p(cp | f_1, f_2, \dots, f_n) = \alpha_0 p(cp | f_1) + \alpha_1 p(cp | f_2) + \dots + \alpha_n p(cp | f_n) \quad (1)$$

where $(\alpha_0, \alpha_1, \dots, \alpha_n)$ is the coefficients where $\sum_{i=0}^n \alpha_i = 1$. Generally, equation (1) needs to be calibrated to determine the coefficients. In many situations, the following equation is used to calculate the probability $p(cp | f_n)$:

$$p(cp | f_n) = \frac{p(cp, f_n)}{p(f_n)} = \frac{p(f_n | cp) p(cp)}{P(f_n)}$$

Considering the situation for a specific cellular carrier m , the probability of a person owns a cell phone in TAZ i turns to be:

$$p_k^i = p_{c_m} p(cp | f_1, f_2, \dots, f_n) = p_{c_m} [\alpha_0 p(cp | f_1) + \alpha_1 p(cp | f_2) + \dots + \alpha_n p(cp | f_n)] \quad (2)$$

If multiple carrier data are available, the equation (2) turns to be:

$$p_k^i = [\alpha_0 p(cp | f_1) + \alpha_1 p(cp | f_2) + \dots + \alpha_n p(cp | f_n)] \sum_{m=1}^M p_{c_m} \quad (3)$$

where M is number of cellular carriers from which the data are obtained. Note that

In practice, due to the privacy concerns, most of the personal information required for the equations (1 – 3) cannot be obtained directly from cellular carriers or operators. However, the U.S. census provides a large amount of demographic survey data for us to produce the distributions of the personal information (age, income, race, etc.). We can utilize the information to calculate the probability of cell-phone ownership. The case study part will give a detailed procedure to determine the cell-phone ownership probabilities.

3.4 Vehicle-per-cellphone equivalent factor

Typically, the cell-phone tracking technology will return the cellular probe counts. However, in transportation planning field, the main interest is on vehicle flows rather than the cellular probe flows. Consequently, a vehicle-per-cellphone equivalent factor f_{vc} will be used in our method to convert cellular probe flows into equivalent vehicle flows (13). We designed a method to estimate the f_{vc} based on the posterior information obtained from the characteristics of the cellular probe trajectories.

According to the cellular probe trajectory characteristics, the set of trajectories can be divided into three subsets:

1. The set of trajectories crossing at least two LA boundaries, σ_1 .
2. The set of trajectories crossing just one LA boundaries, σ_2 .
3. The set of trajectories without crossing any LA boundaries, σ_3 .

For the first two subsets of trajectories, here are three assumptions:

1. *Phones in close proximity (i.e. the same car) generate signal transition events at exactly the same time.* In practice, this assumption needs to be relaxed since phone variation is quite high and signal events may have quite large differences in timing even for phones in the same car. The following two assumptions hold based on this assumption.
2. *There cannot be two vehicles crossing two continuous LA boundaries at same timestamps.* Typically, the dimension of a LA is 3-5 miles by 3-5 miles. There is a very small possibility that some parallel travelling cars crossing at least two LA boundaries at two same timestamps. If two cellular probe trajectories crossing two continuous LA boundaries at two timestamps t and $t + \tau$, they should be in the same vehicle.
3. *Within the saturation headway, there is only one vehicle crossing LA boundaries in each lane.* The default saturation headway is 2.0 seconds. Within two timestamps t and $t + 2$, there is only one vehicle crossing LA boundaries in each lane.

The estimation of f_{cpv} for the trips crossing at least two boundaries will be based on the first assumption. Assuming a set σ_s of cellular probe trajectories ($1, 2, \dots, i, \dots, m$) crossing at two LA boundaries at timestamps t and $t + \tau$, so the expected value of the number of passengers in σ_s will be:

$$\psi_{\sigma_s} = \sum_{k \in \sigma_s} \frac{1}{p_k^i} \quad (4)$$

The average passenger-vehicle occupancy \bar{f}_{pvo} (passengers per vehicle) (22) is applied to determine the number of vehicles crossing the two LA boundaries at timestamp t and $t + \tau$:

$$1 \quad Vehs = 1 + \left(\frac{\psi_{\sigma_s} - \|\sigma_s\|}{f_{pvo}} \right)$$

2 So the vehicle-per-cellphone equivalent factor for cell-phone owner i in set σ_s is:

$$3 \quad f_{vc}^i = \frac{1 + \left(\frac{\psi_{\sigma_s} - \|\sigma_s\|}{f_{pvo}} \right)}{\psi_{\sigma_s}}, \quad i \in \sigma_s$$

4 For the second subset of the trips, the third assumption is used. Suppose a cell-phone
 5 owner i in set σ_2 crosses a LA boundary at timestamp t . There are Ω links located at the
 6 boundary. Each of the link j has several lanes. The number of average occupied lanes (the
 7 average number of lanes which are occupied by vehicles) at link during peak hour is π_j . A set
 8 σ_i of cell phones crossing the boundary between time t and $t+2$. Note that σ_i consists of
 9 both the cell phones crossing only one LA boundary and those crossing at least two LA
 10 boundaries at time t and $t+2$. So the vehicle-per-cellphone equivalent factor for set σ_2
 11 will be:

$$12 \quad f_{vc}^i = \frac{\sum_{j=1}^{\Omega} \pi_j - \left(\frac{\psi_{\sigma_i \cap \sigma_1}}{f_{pvo}} \right)}{\psi_{\sigma_i} - \psi_{\sigma_i \cap \sigma_1}}, \quad i \in \sigma_2 \quad (5)$$

13 For the third subset of cellular probe trajectories, the average value of f_{vc} of the first two
 14 subsets is assigned to them:

$$15 \quad f_{vc}^i = \frac{\sum_{j \in \sigma_1} f_{vc}^j + \sum_{j \in \sigma_2} f_{vc}^j}{\|\sigma_1\| + \|\sigma_2\|}, \quad i \in \sigma_3 \quad (6)$$

16 where the operator $\|\bullet\|$ means that the size of the set.

17 3.5 Trip Generation and Distribution

18 The trip generation and distribution are the first two steps in the traditional four-step
 19 transportation planning process. The trip generation is to decide the number of trips which are
 20 produced or attracted in a specific TAZ. The trip distribution process is to distribute the
 21 productions and attractions predicted by trip generation model to the O-D flows from each
 22 production zone i to each attraction zone j .

23 Due to the limitations on the sample sizes of surveys, the traditional trip generation and
 24 distribution model cannot secure an accurate result. The cell-phone tracking technology
 25 provides a larger sample. Here we introduce a new method to obtain the population O-D
 26 demand combining trip generation and distribution together.

1 For the total population, the proportion of people who have trips between TAZ i and j

2 should be:

$$3 P_{ij} = \bar{Y}^{ij} = \frac{1}{N} \sum_{k=1}^N Y_k^{ij}$$

4 To get the value of T_{ij} , it only needs to multiply P_{ij} with N :

$$5 T_{ij} = \sum_{k=1}^N Y_k^{ij} = NP_{ij}$$

6 If treating the cell-phone owner group as a “simple random sample”. The sampling results
7 can directly be estimated by the following equation:

$$8 \hat{P}_{ij} = \frac{1}{\|S^i\|} \sum_{k=1}^n y_k^{ij}$$

9 Note that in SRS method, the size of cell-phone owner group can be estimated by:

$$10 \|S^i\| = Np(cp)p_{c_m}$$

11 So the estimated value of T_{ij} in SRS survey method is:

$$12 \hat{T}_{ij} = N\hat{P}_{ij} = \frac{\sum_{k=1}^n y_k^{ij}}{p(cp)p_{c_m}} \quad (7)$$

13 Since the distribution of cell-phone owners cannot be considered as the “simple random

14 sample”, the \hat{T}_{ij} cannot be inferred directly using equation (7). A Horvitz – Thompson (HT)

15 estimator (23) of the P_{ij} is proposed:

$$16 \tilde{P}_{ij} = \sum_{k \in S^i} \frac{y_k^{ij}}{Np_k^i} \quad (8)$$

17 From equation (8), it can be seen that the higher probability of owning cell phones, the less

18 weight the corresponding y_k^{ij} is given, in this way the HT estimator uses probability to weight

19 the responses in the estimating the total. The HT estimator of T_{ij} can be defined as follows:

$$20 \tilde{T}_{ij} = N\tilde{P}_{ij} = \sum_{k \in S^i} \frac{y_k^{ij}}{p_k^i} \quad (9)$$

1 Note that T_{ij} is the O-D trips between TAZ i and j , but what we need is the vehicle
2 O-D flows. So the vehicle-per-cellphone factor should be added in the estimator:

$$3 \quad \tilde{T}_{ij}^{\text{veh}} = \sum_{k \in S^i} \frac{y_k^{ij} f_{vc}^k}{p_k^i} \quad (10)$$

4 Now to prove the HT estimator of P_{ij} is an unbiased estimator. Let

$$5 \quad \delta_k^i = \begin{cases} 1 & \text{if } k \in S, \text{ that is to say the } k\text{th ppl has cell phone in TAZ } i \\ 0 & \text{Otherwise} \end{cases}$$

6 Then the estimator \tilde{P}_{ij} can be expressed in following form:

$$7 \quad \tilde{P}_{ij} = \sum_{k=1}^N \frac{Y_k^{ij} \delta_k^i}{N p_k^i}$$

8 The expectation of \tilde{P}_{ij} is:

$$9 \quad E(\tilde{P}_{ij}) = \sum_{k=1}^N E\left(\frac{Y_k^{ij} \delta_k^i}{N p_k^i}\right) = \sum_{k=1}^N \left(\frac{Y_k^{ij} E(\delta_k^i)}{N p_k^i}\right) = \sum_{k=1}^N \left(\frac{Y_k^{ij} p_k^i}{N p_k^i}\right) = \bar{Y}^{ij} = P_{ij} \quad (10)$$

10 The estimator \tilde{P}_{ij} is the unbiased estimator of P_{ij} .

11 Furthermore, the variance of the estimator \tilde{P}_{ij} is.

$$12 \quad \text{Var}(\tilde{P}_{ij}) = \text{Var}\left[\sum_{k=1}^N \frac{Y_k^{ij} \delta_k^i}{N p_k^i}\right] = \frac{1}{N^2} \left\{ \sum_{k=1}^N \frac{\left(Y_k^{ij}\right)^2 \text{Var}(\delta_k^i)}{\left(p_k^i\right)^2} + \sum_{k \neq m}^N \frac{Y_k^{ij} Y_m^{ij} \text{Cov}(\delta_k^i, \delta_m^i)}{p_k^i p_m^i} \right\} \quad (11)$$

13 Note that

$$14 \quad \text{Var}(\delta_k^i) = E\left[\left(\delta_k^i\right)^2\right] - \left[E\left(\delta_k^i\right)\right]^2 = p_k^i(1 - p_k^i)$$

15 and

$$16 \quad \text{Cov}(\delta_k^i, \delta_m^i) = E\left(\delta_k^i \delta_m^i\right) - E\left(\delta_k^i\right) E\left(\delta_m^i\right) = p_{km}^i - p_k^i p_m^i$$

17 where p_{km}^i is the joint probability of both the k th people and the m th people own cell
18 phones.

19 Considering two people in sample have an independent probability to own cell phones, the

1 equation (11) turns to be:

$$2 \quad \text{Var}(\tilde{P}_{ij}) = \text{Var}\left[\sum_{k=1}^N \frac{Y_k^{ij} \delta_k^i}{N p_k^i}\right] = \frac{1}{N^2} \left\{ \sum_{k=1}^N \frac{\left(Y_k^{ij}\right)^2 p_k^i (1 - p_k^i)}{\left(p_k^i\right)^2} \right\} = \frac{1}{N^2} \left\{ \sum_{k=1}^N \frac{\left(Y_k^{ij}\right)^2 (1 - p_k^i)}{p_k^i} \right\} \quad (12)$$

3 If assume the people in analyzed TAZ have the same probability p of cell-phone
4 ownership, the expected value of variance turns to be:

$$5 \quad E\left[\text{Var}(\tilde{P}_{ij})\right] = E\left[\frac{(1-p)}{N^2 p} \sum_{k=1}^N \left(Y_k^{ij}\right)^2\right] = \frac{(1-p)}{N p} E\left[P_{ij}\right] = \frac{(1-p)\tilde{P}_{ij}}{N p} \quad (13)$$

6 Then the expected value of standard deviation should be:

$$7 \quad E\left[SD(\tilde{P}_{ij})\right] = E\left[\sqrt{\frac{(1-p)P_{ij}}{N p}}\right] = \sqrt{\frac{(1-p)\tilde{P}_{ij}}{N p}} \quad (14)$$

8 4. CASE STUDY – SIMULATION EXPERIMENTS



(a)

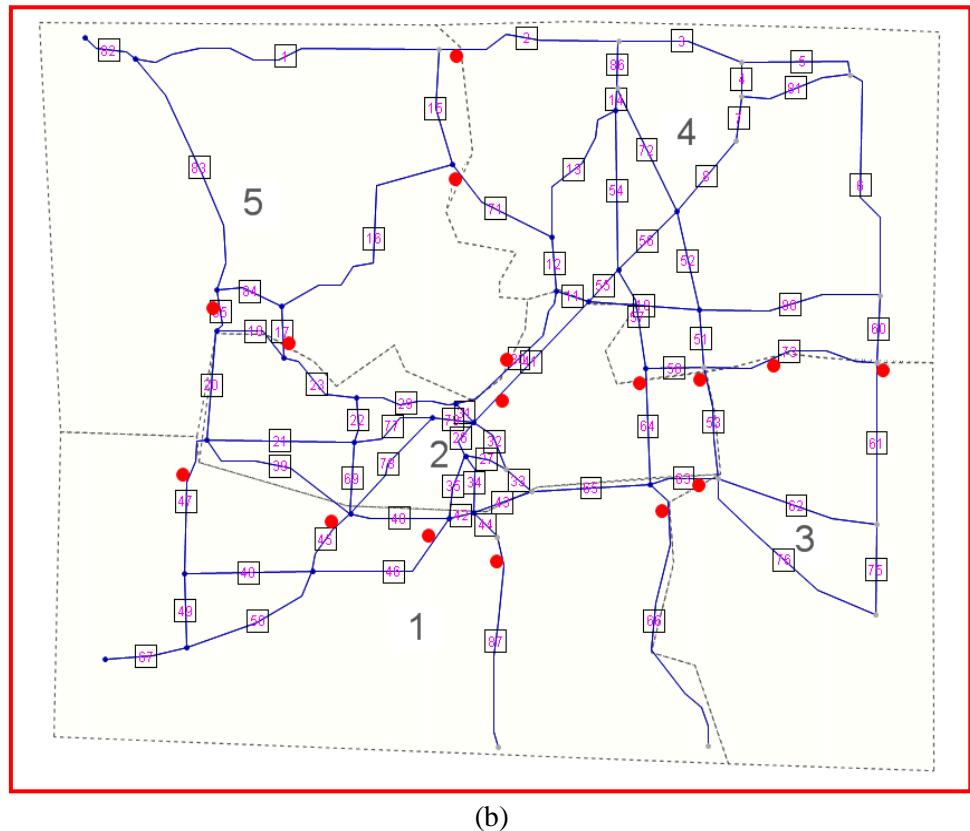
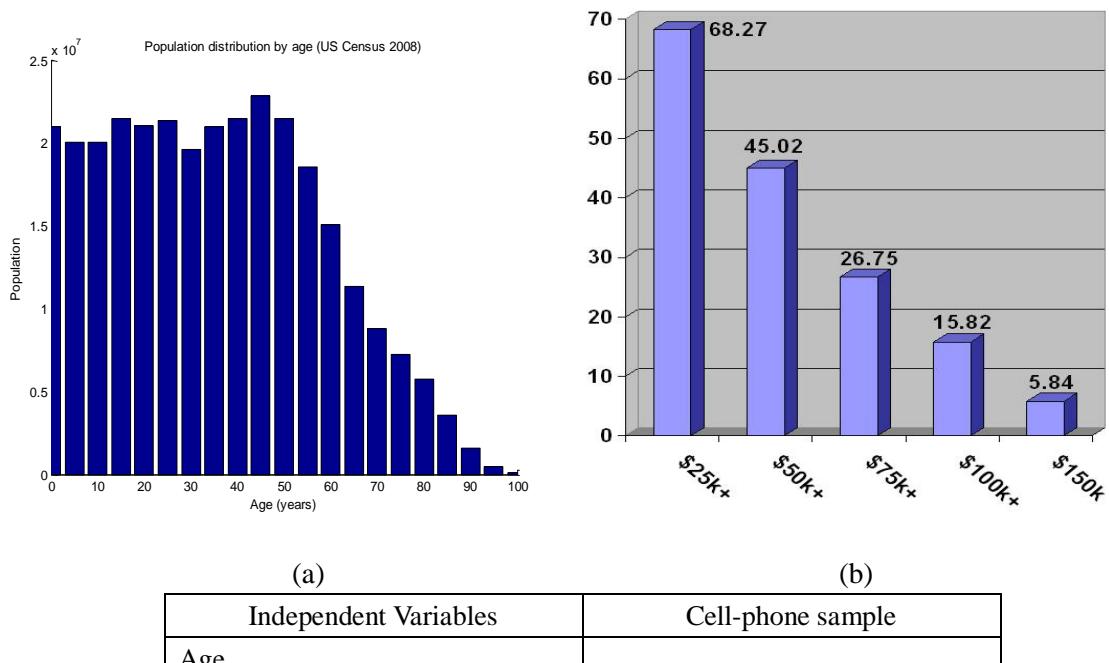


Figure.4 (a) Cell tower location map of the research area
 (b) Corresponding simulation network

This simulation aims to provide a verification of the proposed O-D estimation method. The proposed research area is Dane county in the southwest of the state of Wisconsin. To be simplified, it is divided into 5 TAZs. Figure.4 (a) shows the cell-phone tower locations and Figure.4 (b) shows the corresponding VISSIM simulation layout of the research area. The red circles in Figure.4 (b) are the intersections of links and LA boundaries.



18 - 30	41%
31 - 62	53%
63 – up	6%
Income	
Less than \$50,000	51%
\$50,000 – up	37%
Don't know/Refused	12%

1
2
3
4
5
6
7
8
9
10
11
(c)

Figure.5 (a) U.S. census data of age distribution, 2008 (24)

(b) U.S. census data of income distribution, 2006 (25)

(c) Demographic information on cell-phone ownership pattern (21)

The input data involve with 3 input modules: the trip survey data module, the cell phone ownership distribution module and the vehicle occupancy distribution module. The simulation period is set to be 24 hours to estimate the daily O-D demand data. The 3 input modules prepared the input parameters as well as the input O-D matrix to start the VISSIM simulator. A cell phone signal transition events module will be paired with the VISSIM simulator module to provide the random events such as call-in and call-out events, which can be used to generate the HO events.

The trip survey data module uses the Wisconsin State-wide Trip survey data (26) as the aggregated input daily O-D matrix. Also the trip survey data contains the age and household income information which can be used in the cell phone ownership distribution module.

The cell phone ownership distribution module is designed to assign the cell phone to each trip maker with cell phone ownership probabilities. To simplify the demonstration, only the income and age are taken into consideration for the cell-phone ownership probability. Figure.5 (a) and (b) illustrate the U.S. census demographic data for the population by age and income.

Figure.5 (c) shows the population age and income distribution, from which it is easy to get the conditional probability $p(cp | age)$ and $p(cp | income)$. To get the value of $p(cp | age)$ and $p(cp | income)$, it can be calculated as following:

$$p(cp | age) = \frac{p(cp, age)}{p(age)} = \frac{p(age | cp) p(cp)}{P(age)}$$

$$p(cp | income) = \frac{p(cp, income)}{p(income)} = \frac{p(income | cp) * p(cp)}{p(income)}$$

where $p(cp)$ is the market penetration rate of cell phones. In this simulation, it is set to be 0.8, and the market share of a specific carrier is set to be 0.25.

Use the equation (1) to determine the probability of cell-phone ownership:

$$p(cp | age, income) = \alpha p(cp | age) + (1 - \alpha) p(cp | income)$$

where α is the coefficient between 0 and 1. In this simulation, it is set to be 0.5.

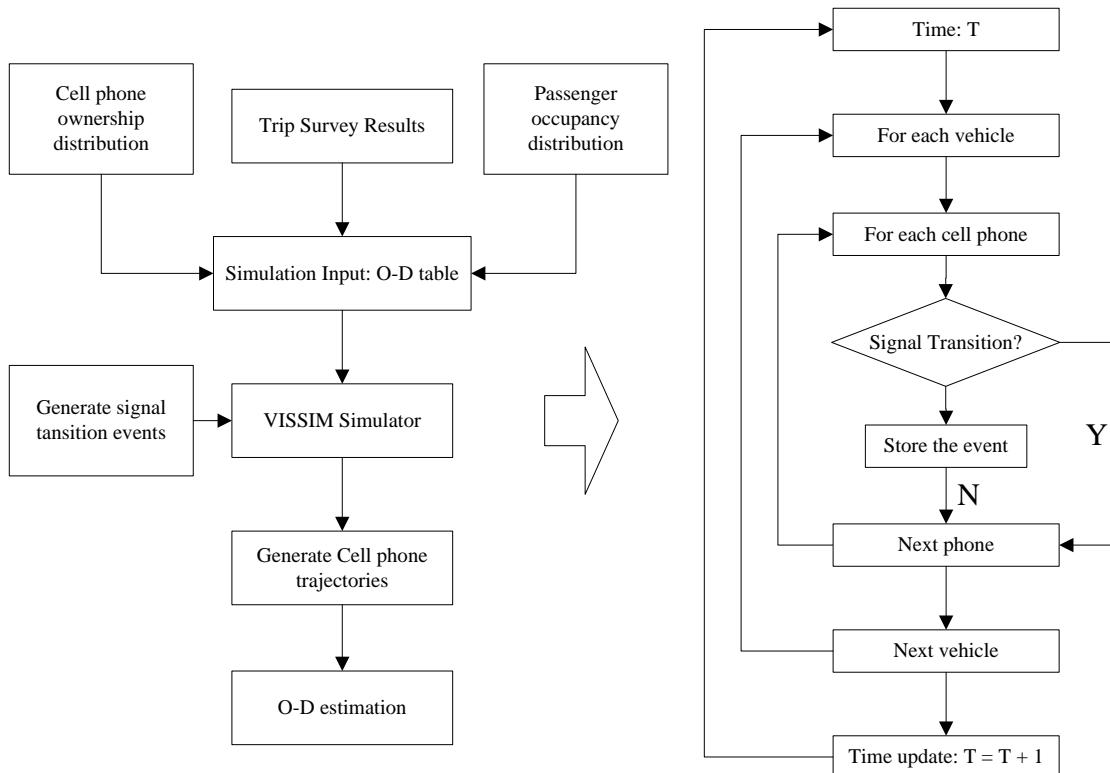
In 2008 U.S. census data (24), Dane county has a 482,705 residents. The cell-phone

1 ownership probability for each people is generated based on the above demographic
2 information. Then the cell-phone owners are assigned to each individual vehicle.

3 The vehicle occupancy distribution module is to generate the random numbers of
4 passengers assigned to each individual vehicle. It is used to convert the trip counts into vehicle
5 counts. The average Passenger-Vehicle Occupancy in United States is 1.89 (27). A [1,3]
6 discrete uniform distribution is used to generate the random numbers.

7 The VISSIM simulation tool is employed to simulate the vehicle movements between each
8 O-D pair. The input O-D table is assigned by VISSIM's built-in Dynamic Traffic Assignment
9 (DTA) algorithm to generate the vehicle flows on links. The centers and boundaries of cells
10 and LAs are predefined without any time-dependent fluctuations in the simulation network.
11 The radius of cell coverage is set to 1000 ft. The boundaries in VISSIM are set as data
12 collection points on links where the cell or LA boundaries intersect with. The data collection
13 points can record each vehicle's ID and timestamp when the vehicle crosses them. In this
14 simulation, the cell phones are assumed to be set in turn-on mode automatically.

15 During each simulation time step, the system will check whether there is any
16 signal-transition event happened. The cell phones are assigned with a small probability to
17 determine the occurrence of call-in and call-out events. The durations are determined by
18 assigned a random number. The HO events will be recorded when the cell phones are in on-call
19 mode and cross the data collection point at cell boundaries. The TA events will be record with
20 its corresponding cell centers when cell phones are in on-call mode. The LU events will be
21 record when cell phones cross the data collection points at LA boundaries. The PLU events will
22 be recorded every two hours with its corresponding cell centers as well. After collecting the
23 cellular probe trajectories, the O-D estimation introduced in Figure.3 is employed to get the
24 estimated O-D matrix. The simulation process is shown in Figure.6.



25

26

Figure.6 Illustration of the simulation process

The SRS method is also implemented in the simulation, and the average passenger-vehicle occupancy is used to convert the population trips into vehicle trips.

The results are shown in Table.2. It can be found that most of the estimated O-D flows have less percentage error than the SRS method. The average percentage error of the proposed method is 7.56%, while the SRS method returns 15.45%. Since the cell phone user group is not naturally a random sample, the HT estimator can give a more accurate estimation results comparing to SRS method. Moreover, our method uses the vehicle-per-cellphone factor to convert the cell phone counts into vehicle counts, while the SRS method employs the prior Passenger-Vehicle Occupancy information. The simulation results fully show the advantage of our method over SRS method.

Table.2 Simulation results of proposed method and SRS method

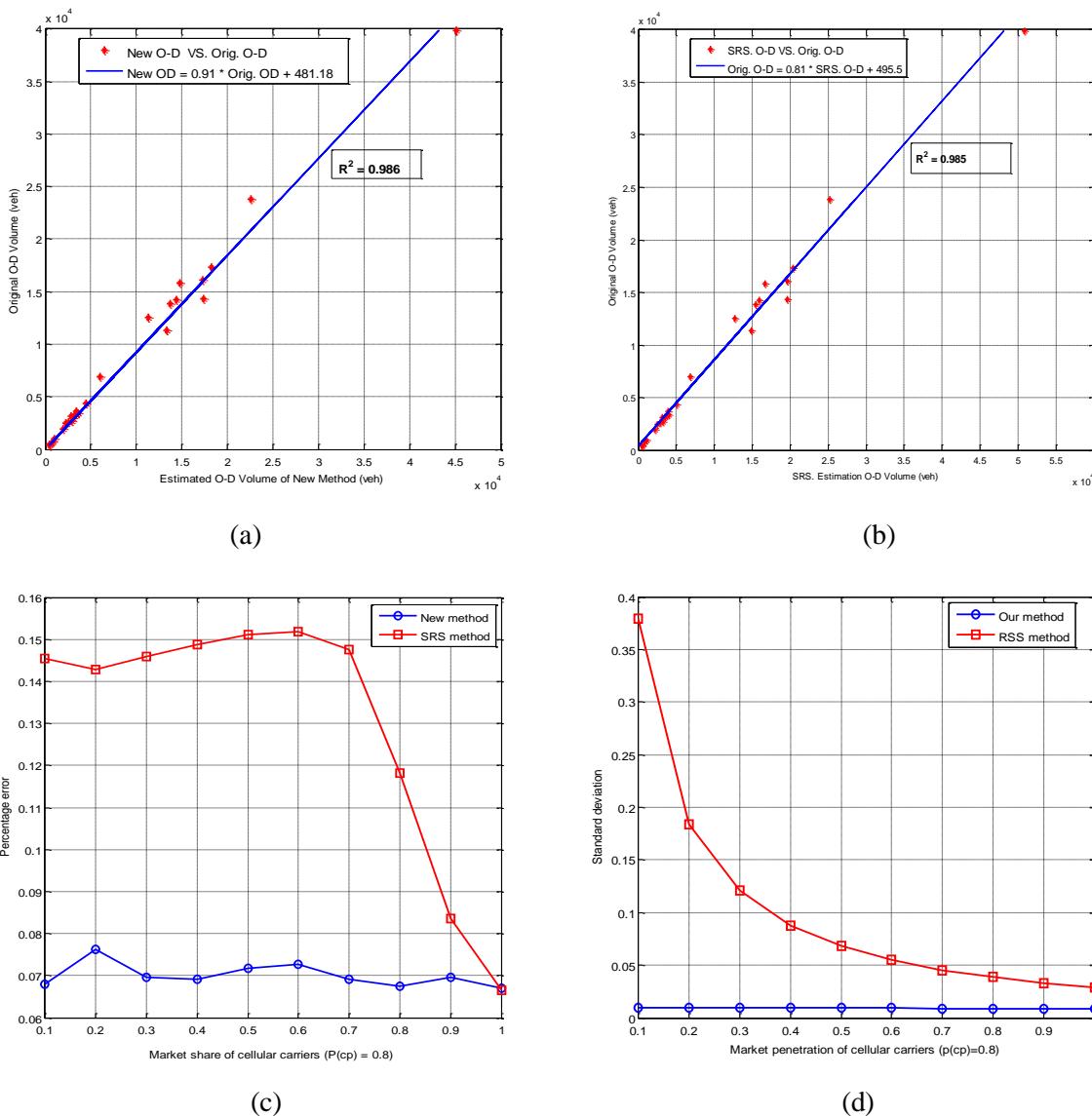
O-D Pair	Orig. O-D	Esti. O-D	Per. Error	SRS O-D	Per. Error
1->1	16074	17325	7.78%	21080	22.23%
1->2	13872	13749	0.89%	15785	11.22%
1->3	750	796	6.13%	1043	21.33%
1->4	2562	2543	0.74%	3238	11.24%
1->5	2442	2274	6.88%	2543	2.70%
2->1	11334	13379	18.04%	16600	31.20%
2->2	23810	22581	5.16%	23413	6.08%
2->3	4332	4441	2.52%	5443	16.41%
2->4	14256	14405	1.05%	16230	12.06%
2->5	15828	14764	6.72%	15048	5.53%
3->1	960	972	1.25%	1000	15.94%
3->2	3384	3645	7.71%	4008	19.77%
3->3	6948	6026	13.27%	6035	1.87%
3->4	3300	3366	2.00%	3900	14.24%
3->5	456	481	5.48%	498	20.61%
4->1	2712	2899	6.90%	3613	21.05%
4->2	17322	18290	5.59%	21863	17.83%
4->3	3138	2823	10.04%	2863	0.06%
4->4	39840	45142	13.31%	55298	27.60%
4->5	3642	3418	6.15%	3730	7.08%
5->1	1947	1943	0.21%	2268	12.12%
5->2	14316	17358	21.25%	22353	37.25%
5->3	390	506	29.74%	585	37.95%
5->4	3174	3202	0.88%	3538	10.84%
5->5	12522	11347	9.38%	11123	2.02%

Furthermore, the estimated O-D volumes versus original O-D volumes are plotted in Figure.7 (a) and (b). It can be seen that both the results from the proposed method and SRS method have strong relationship with the original O-D flow. The regression shows both the lines fits the data very well, in which the coefficients of determination R^2 of our method are 0.986 and 0.985, respectively.

To better illustrate the advantages of the proposed method, a sensitivity analysis is carried

out to see the influence of the cell-phone owner group size (market share of cellular carriers). The cell-phone market penetration rate is fixed at 0.8. The market share of cellular carriers is increased from 0.1 to 1. Note that here the market share of cellular carriers is the total market share of the carriers which are available to provide cellular data. Figure.7 (c) and (d) show the comparison between our method and SRS method in term of the percentage error and standard deviation of P_{ij} with the increasing of market share of cellular carriers.

It can be seen that the percentage error keeps unchanged at about 7% with increasing of cellular carriers' market share. On the other hand, the percentage error of SRS method decreases until the market share increased to 0.7.



(a)

(b)

(c)

(d)

Figure.7 (a) Estimated O-D VS. Original O-D

(b) RSS O-D VS. Original O-D

(c) Market penetration rate VS. Percentage error

(d) Market penetration rate VS. Standard deviation

In Figure.7 (c) and (d), the standard deviation of our method keeps unchanged below 5% with the increasing of market share, while the SRS method will decrease from more than 35% to 5% when the market share increases from 0.1 to 1.

Both the results of percentage error and standard deviation show the proposed method is robust method for the daily O-D matrix estimation. Generally, the smaller market share of the cellular carriers in practice, the less cellular trips can be obtained from the trajectories. Then the accuracy of estimation results will be more difficult to attain. Different with the SRS method, the proposed method can still keep good performance at smaller data set.

5. CONCLUSIONS AND FUTURE RESEARCH EXTENSION

Traditional survey-based trip diary approach to estimating trip generation and distribution is time-consuming and cost-prohibitive. The estimation may vary from one study as a result of the limitation of the survey sample size and sampling randomness. With the popularity of cell phone and emerging cellular tracking technologies, using cellular probe data have the great potential to provide a larger sample size in a timely manner.

In this paper, an exploratory methodology was presented to estimate the daily O-D demand using cellular probe trajectories. They can be obtained by tracking all the signal-transition and periodic location update events of cellular probes to determine the trip origins and destinations. To overcome the potential socio-economic bias, a conditional probability of cell-phone ownership was estimated using traveler's socio-economic factors that are readily available in the census data. Then, a vehicle-per-cellphone equivalent factor was generated based on the posterior information of the characteristics of cellular probe trajectories. In other words, individual cellular trips were converted into equivalent vehicle trips. Next, the trip generation and distribution were obtained simultaneously using a Horvitz-Thompson estimator so that the population O-D demand can be estimated. The Horvitz-Thompson estimator was proved to be an unbiased estimator of the population O-D demand in theory. A VISSIM based simulation was designed to exemplify the proposed method. A "simple random sampling" (SRS) method, the prevailing method in current literature, was also simulated. The comparison between the outcome of cellular probe data and SRS shows that both methods yielded desirable goodness-of-fit in terms of R^2 but the average percentage error of SRS is almost twice of the cellular probe data method, demonstrating the superiority of the proposed methodology. The sensitivity analysis has also shown that the proposed method provides a robust estimation for the daily O-D matrix.

To verify the validity of the assumptions of proposed methodology, a field test is needed in the future study, in which the cellular probe data and cell boundaries will be obtained from cellular carriers. A method should be proposed to eliminate the estimation error caused by variations of cell sizes and boundaries. A more accurate method to determine the trip origins and destinations should be developed. And an additional survey is needed to get accurate demographic information of cell-phone owners. An existing O-D demand matrix will be used as the ground truth to verify the correctness of the estimation results.

1 REFERENCE

- 2 1. Meyer, M.D. and E.J. Miller, *Transportation Planning: A Decision-Oriented Approach*.
3 McGraw-Hill Book Company, INC, New York, NY. 1984
- 4 2. Giaimo, G.T., Modifications To Traditional External Trip Models. 2002. pp. p. 163-171.
- 5 3. Abrahamsson, T., Estimation of Origin-Destination Matrices Using Traffic Counts - A
6 Literature Survey. INTERIM Report. 1998.
- 7 4. Sherali, H.D. and T. Park, Estimation of Dynamic Origin-Destination Trip Tables for A
8 General Network. *Transportation Research Part B: Methodological*, 35(3). 2001. pp. 217-235.
- 9 5. Wong, S.C., et al., Estimation of Multiclass Origin-Destination Matrices from Traffic
10 Counts. *Journal of Urban Planning and Development*, 131(1). 2005. pp. 19-29.
- 11 6. Stopher, P.R. and S.P. Greaves, Household Travel Surveys: Where Are We Going?
12 *Transportation Research Part A: Policy and Practice*, 41(5). 2007. pp. 367-381.
- 13 7. Hazelton, M.L., Some Comments on Origin-Destination Matrix Estimation.
14 *Transportation Research Part A: Policy and Practice*, 37(10). 2003. pp. 811-822.
- 15 8. Qiu, Z., et al. State of the Art and Practice: Cellular Probe Technology Applied in
16 Advanced Traveler Information Systems. In *Transportation Research Board 86th Annual
17 Meeting*. Washington D.C.: Transportation Research Board. 2007.
- 18 9. Astarita, V., et al., Motorway Traffic Parameter Estimation from Mobile Phone Counts.
19 *European Journal of Operational Research*, 175(3). 2006. pp. 1435-1446.
- 20 10. Caceres, N., J.P. Wideberg, and F.G. Benitez, Review of Traffic Data Estimations
21 Extracted from Cellular Networks. *Intelligent Transport Systems, IET*, 2(3). 2008. pp. 179-192.
- 22 11. Asakura, Y. and T. Iryo, Analysis of Tourist Behaviour based on the Tracking Data
23 Collected using a Mobile Communication Instrument. *Transportation Research Part A: Policy
24 and Practice*, 41(7). 2007. pp. 684-690.
- 25 12. Pan, C., et al., Cellular-Based Data-Extracting Method for Trip Distribution.
26 *Transportation Research Record: Journal of the Transportation Research Board*. 2006. pp. pp
27 33-39.
- 28 13. Caceres, N., J.P. Wideberg, and F.G. Benitez, Deriving Origin Destination Data from A
29 Mobile Phone Network. *Intelligent Transport Systems, IET*, 1(1). 2007. pp. 15-26.
- 30 14. Keemin, S. and K. Daehyun, Dynamic Origin-Destination Flow Estimation Using Cellular
31 Communication System. *Vehicular Technology, IEEE Transactions on*, 57(5). 2008. pp.
32 2703-2713.
- 33 15. White, J. and I. Wells. Extracting Origin Destination Information from Mobile Phone Data.
34 In *Road Transport Information and Control*, 2002. *Eleventh International Conference on (Conf.
35 Publ. No. 486)*. 2002.
- 36 16. Lu, J., et al. Applying Cellular-Based Location Data to Urban Transportation Planning. In
37 *Applications of Advanced Technology in Transportation*. Chicago: ASCE. 2006.
- 38 17. Mircea, I., S. Emil, and H. Simona. Cell ID Positioning Method for Virtual Tour Guides
39 Travel Services. In *ECAI 2007 - International Conference*. Pitesti, Romania: Electronics,
40 Computers and Artificial Intelligence. 2007.
- 41 18. Gur, Y.J., S. Bekhor, and C. Solomon. An Aggregate National Transportation Planning
42 Process in Israel: Formulation and Development. In *Transportation Research Board 88th
43 Annual Meeting*. Washington D.C.: Transportation Research Board. 2009.
- 44 19. *Background on CTIA's Semi-Annual Wireless Industry Survey*. 2009; Available from:
45 files.ctia.org/pdf/CTIA_Survey_Year_End_2007_Graphics.pdf.
- 46 20. US Wireless Data Market Update - Q1. Chetan Sharma Consulting Co. Ltd. 2009.
- 47 21. Cell Phone Nation 2009. Marist Institute for Public Opinion. 2009.
- 48 22. Gan, A. and K. Liu, *Vehicle Occupancy Trends in Florida: Evidence from Traffic Accident
49 Records*. Transportation Research Board 87th Annual Meeting. Transportation Research Board.
50 pp. 17p. 2008
- 51 23. Konijn, H.S., *Statistical Theory of Sample Survey Design and Analysis*. American Elsevier
52 Publishing Company, INC., New York, NY. 1973
- 53 24. Age and Sex Distribution in 2005. U.S. Census Bureau. 2005.
- 54 25. Annual Social and Economic Supplement 2006. U.S. Census Bureau. 2006.

-
- 1 26. Statistics, B.o.T. *NHTS/NPTS Database.* 2003; Available from:
2 <http://nhts.ornl.gov/download.shtml>.
3 27. Statistics, B.o.T., NHTS 2001 Highlights Report. U.S Department of Transportation,
4 Washington, DC. 2003.
5
6

RESEARCH ARTICLE

A Two-Stage Algorithm for Origin-Destination Matrices Estimation Considering Dynamic Dispersion Parameter for Route Choice

Yong Wang^{1,2}, Xiaolei Ma^{3,4*}, Yong Liu¹, Ke Gong¹, Kristian C. Henricakson², Maozeng Xu¹, Yinhai Wang^{2*}

1 School of Management, Chongqing Jiaotong University, Chongqing, China, **2** Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington, United States of America, **3** School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure, Systems, and Safety Control, Beihang University, Beijing, China, **4** Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou #2, Nanjing, Jiangsu, China

* xiaolei@buaa.edu.cn (XM); yinhai@uw.edu (YH)



OPEN ACCESS

Citation: Wang Y, Ma X, Liu Y, Gong K, Henricakson KC, Xu M, et al. (2016) A Two-Stage Algorithm for Origin-Destination Matrices Estimation Considering Dynamic Dispersion Parameter for Route Choice. PLoS ONE 11(1): e0146850. doi:10.1371/journal.pone.0146850

Editor: Zhong-Ke Gao, Tianjin University, CHINA

Received: August 11, 2015

Accepted: December 21, 2015

Published: January 13, 2016

Copyright: © 2016 Wang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research is supported by National Natural Science Foundation of China (Project No. 71402011, 71471024, 51408019, 71301180, 51329801), National Social Science Foundation of Chongqing of China (No. 2013YBJJ035), and the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No. KJ1400307), and the Natural Science Foundation of Chongqing of China (No. cstc2015jcyJA30012), National Key Technologies R&D Program of China (2014BAG01B03), Science and Technology Project

Abstract

This paper proposes a two-stage algorithm to simultaneously estimate origin-destination (OD) matrix, link choice proportion, and dispersion parameter using partial traffic counts in a congested network. A non-linear optimization model is developed which incorporates a dynamic dispersion parameter, followed by a two-stage algorithm in which Generalized Least Squares (GLS) estimation and a Stochastic User Equilibrium (SUE) assignment model are iteratively applied until the convergence is reached. To evaluate the performance of the algorithm, the proposed approach is implemented in a hypothetical network using input data with high error, and tested under a range of variation coefficients. The root mean squared error (RMSE) of the estimated OD demand and link flows are used to evaluate the model estimation results. The results indicate that the estimated dispersion parameter theta is insensitive to the choice of variation coefficients. The proposed approach is shown to outperform two established OD estimation methods and produce parameter estimates that are close to the ground truth. In addition, the proposed approach is applied to an empirical network in Seattle, WA to validate the robustness and practicality of this methodology. In summary, this study proposes and evaluates an innovative computational approach to accurately estimate OD matrices using link-level traffic flow data, and provides useful insight for optimal parameter selection in modeling travelers' route choice behavior.

Introduction

Urban sprawl and population growth have resulted in increasingly severe traffic congestion in major cities around the world. City planners and decision makers have recognized the need for comprehensive traffic management strategies to meet the challenges of rapidly evolving built environments and population demographics. Effective transportation policies and control

on Transportation Construction by the Ministry of Transport of China (2015318835200).

Competing Interests: The authors have declared that no competing interests exist.

measures can improve traffic safety and quality of service, as well as promoting economic development and reducing air pollution. Obtaining origin-destination (OD) traffic demand matrix in low-cost and high-accuracy manner not only becomes a problem transportation science, but also draws attentions from many scholars in various scientific fields. For example, researchers in statistical physics and complex systems recently proposed a number of novel methods to estimate OD matrix directly from population data [1, 2, 3, 4, 5, 6]. Reliable OD matrix estimation can provide critical insight for traffic management, operations, and urban planning efforts to mitigate congestion [7, 8]. Thus, a reliable OD matrix estimation method is indispensable for both transportation planners and traffic engineers.

A number of approaches have been developed for estimating OD matrices in the past several decades [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Compared with conventional survey-based method, data-driven OD estimation methods relying on link-level traffic flow measurements require less effort and offer significantly reduced time and cost for data acquisition and processing. For such methods, observed traffic flows at key points throughout the network should be known as prior information for OD matrix initialization.

Past research on this topic has considered range of different optimization methods, including entropy maximizing estimators [21, 22], maximum likelihood estimation [23], Bayesian inference estimation [24], generalized least squares (GLS) [9, 10, 25] to estimate OD demands. Entropy maximizing estimators are used to maximize the spread of trip distributions on all available paths (routes) where the observed traffic flows are used as the only information (i.e. without a target trip matrix). Maximum likelihood estimation aims to maximize the likelihood of the closeness between target OD matrix and estimated OD matrix. In the Bayesian inference approach, the target OD matrix is a prior probability function of the estimated OD matrix on a basis of observed traffic count data. The GLS estimator is a robust and efficient linear unbiased estimator, which can solve the estimation of OD matrix by minimizing the Weighted Euclidean Distances (WED) between the target data and the solution data.

User equilibrium (UE) assignment models are commonly used to obtain path choice behavior based on the estimated OD demand. Deterministic UE assignment models assume that all users have access perfect information about the generalized link travel costs, and select a route with the lowest perceived travel cost [26]. Beckman [27] formulated the UE assignment model by assuming that the OD demands are a function of level of service. A combined distribution and assignment model which relies on link-level traffic flow data was presented by Fisk and Boyce [28], and extended by Lam and Huang [29] to address multiclass-user transportation networks. Fisk [30, 31] proposed a combined entropy maximizing model with UE constraints. Yang et al. [11] integrated the GLS technique with a UE traffic assignment model for OD matrix estimation, presented in the form of a convex bi-level optimization problem. Summaries of the more recent contributions to UE-based traffic assignment are provided in Han [32], Lu et al. [33], Inoue and Maruyama [34], Kumar and Peeta [35].

The stochastic user equilibrium (SUE) principle allows the perceived cost to vary between individuals in a heterogeneous population, which can be seen as a more realistic approach than deterministic UE [15, 36], in which the perceived travel costs cannot vary between travelers. The probit SUE was first formulated as a generalization of user equilibrium by Daganzo and Sheffi [37], and developed by Sheffi and Powell [38] as a mathematical programming problem. Liu and Fricker [39] presented a two-stage SUE approach to estimate OD matrices and the probit dispersion parameter in an iterative manner. Yang et al. [15] improved on the methods described in Liu and Fricker by incorporating link traffic flows and travel cost obtained using logit-based SUE traffic assignment. Meng et al. [40] presented a linearly constrained model and solution algorithm for the probit SUE problem with fixed demand and separable link travel time functions. This modeling approach was extended in Meng et al. [41] using elastic demand and non-

separable link travel time functions. Time-dependent traffic assignment can be also formulated as a multinomial logit model [42, 43, 44], and this has become one of the most common methods for SUE-based traffic assignment [45, 46, 47]. In a fixed-point formulation, fixed target demands or link flows are used to establish model based on UE and SUE principles [13, 14, 19].

In the multinomial logit model formulation, the link choice probability is a function of a dispersion parameter θ [16], which describes road users' perception of travel costs. Though the dispersion parameter θ is predetermined in many previous studies [14, 36, 45, 46, 47, 48], here we assume that this value should be allowed to change with traffic conditions. In addition, Lo and Chan [16] proposed a maximum likelihood procedure for simultaneously estimating the OD matrix and the dispersion parameter θ , while the link choice proportions and link flows can be further calculated based on the maximum likelihood estimators of OD matrix and θ . Compared with the previous studies, the main contributions of this paper lie in: (1) A fixed-point model is formulated with a dynamic dispersion parameter θ , where the estimation of link choice proportions is integrated into the optimization procedure; (2) A GLS estimator is utilized to train this model, and the link choice proportions can be simultaneously calculated based on the OD matrix and dispersion parameter through a multinomial logit model; (3) A two-stage iterative algorithm is presented to refine the OD matrix and dispersion parameter estimates, and Sequential Quadratic Programming (SQP) from the extended quasi-Newton method is applied in the two-stage algorithm process[49].

The remainder of this paper is organized as follows: In Section 2, relevant notation, definitions, and model formulations are presented, followed by a link choice proportion approach to calculate the observed link flow using a true OD matrix. A two-stage algorithm is described in Section 3, along with model implementation details. The performance of the proposed approach is tested in a hypothetical network, and a sensitivity analysis is conducted using a range of variation coefficients. Results are presented and compared with those obtained through other established OD estimation methods. In section 4, results are presented for a real-world network using loop detector data in the city of Seattle, WA to demonstrate the practicality of the proposed approach. Finally, conclusions are summarized in Section 5.

Model Formulation

Related Notations and Definitions

The notation and parameter definitions used throughout the paper are as follows:

K the set of network links $k \in K$, where T denotes the total number of links

L the set of observed links $l \in L$, where Γ denotes the number of observed links

J the set of OD pairs $j \in J$, where τ indicates the total number of OD pairs

A the set of paths connecting the OD pair j , $a \in A$

c_k travel cost of link k

c_{rj} travel cost of path r connecting the OD pair j

t_k the free flow travel time of link k

C_k the capacity of link k

α_k the performance function parameter of link k

β_k the exponential value of link k 's performance function

M the set of nodes in the network

- f the vector for estimated link flows, where f_k is the estimated flow for link k
- \hat{f} the vector for observed link flow, where \hat{f}_l is the observed link flow for link l
- d the estimated OD vector matrix, where d_j is the j th element of d for OD pair j
- \bar{d} the target OD vector matrix, where \bar{d}_j is the j th element of \bar{d} for OD pair j
- \tilde{d} the initial OD vector matrix for SQP algorithm optimization
- W' the initial weight matrix in the group of all paths connecting each OD pair
- w'_{mn} the initial weight element of all paths connecting nodes m and n , $m,n \in M$
- W the weight matrix in the group of all paths connecting each OD pair
- w_{mn} the weight element of all paths connecting nodes m and n , $m,n \in M$
- E the identity matrix with the same dimension as the initial weight matrix
- G the vector matrix of observed link flows, where G_i is the i th element of G
- U the covariance matrix for the target OD vector and estimated OD vector
- V the covariance matrix for the observed link flows and estimated link flows
- P the matrix of link choice proportions, where p_{kj} is the kj th element of P . This is equivalent to the proportion of OD pair j traveling on the observed link k
- P_{rj} the probability of path r that connects OD pair j being chosen for a trip
- x_k the observed traffic flow of link k
- $x_k^{(s)}$ the estimated traffic flow of link k at the s th iteration in the traffic assignment stage
- $y_k^{(s)}$ the auxiliary mean traffic flow of link k at the s th iteration in the traffic assignment stage
- θ the estimated dispersion parameter for OD estimation
- $\bar{\theta}$ the target dispersion parameter for OD estimation
- $\tilde{\theta}$ the initial dispersion parameter for SQP algorithm optimization
- Q the covariance matrix for θ and $\bar{\theta}$
- S_d the feasible solution set for OD matrix
- S_θ the feasible solution set for θ parameter
- $\sigma_{d_j}^2$ the variance for OD demands
- $\sigma_{x_k}^2$ the variance for link flows
- σ_θ^2 the variance for dispersion parameter
- λ_d random term for the target OD matrix
- λ_f the random term for observed link flows
- F_1 the “distance” between the estimated OD vector matrix d and target demand OD matrix \bar{d}
- F_2 the “distance” between the estimated link flow vector f and observed link flow vector \hat{f}

F_3 the “distance” between the estimated dispersion parameter θ and the target dispersion parameter $\bar{\theta}$

a_{krj} decision variable, if the link k lies on path r connecting the OD pair j , and set $a_{krj} = 1$, and 0 otherwise

η the percentage of traffic flow traveling from each node to the most adjacent node

$RMSE(OD)$ the root mean squared error between estimated and true OD matrices

$RMSE(LF)$ the root mean squared error based on estimated and observed (true) link flows

Based on the above notations, the OD estimation model development and validation procedure can be described as follows:

1. A ground truth OD matrix is used as prior information to calculate the link flows based on the link choice proportion model. These link flows represent the measured traffic flows obtained through fixed mechanical sensors or other means;
2. The observed link flows are chosen from those calculated link flows at fixed points throughout the network;
3. The estimated OD matrix, link flows, and dispersion parameter are obtained via the fixed-point model and two-stage iterative algorithm using the partial observed link flows from step (2);
4. Results are evaluated and compared with the ground truth as established in step (1).

The Fixed Point Model with Dynamic Dispersion Parameter. As presented in the previous subsection, the estimated OD vector matrix is expressed as $d = [d_1, d_2, \dots, d_r, \dots, d_\tau]'$, where d_j denotes the mean traffic flow of the j th element of d for OD pair j . Consider an OD pair j connected by a link k which is associated with a link performance cost function $c_k(f_k)$ equal to the cost of using link k . The link performance cost function [50] is expressed during the traffic assignment procedure in Eq 1:

$$c_k(f_k) = t_k[1.0 + \alpha_k(\frac{f_k}{C_k})^{\beta_k}], \forall k \in K \quad (1)$$

The link flow vector is defined as $f = [f_1, f_2, \dots, f_\Gamma]'$, and the matrix of link choice proportions is denoted as $P = [p_{kj}]$, where $0 \leq p_{kj} \leq 1$. This represents the proportion of OD pair j connected by the link k . Thus, the mathematical expectation of link flow vector f can be calculated as $E[f] = [Pd]_{\Gamma \times 1}$, where Pd is the product of the observed mean link flow vector and the matrix of link choice proportions P . P can be adjusted by the link flows and the dispersion parameter θ .

The OD matrix can be estimated via a fixed point formulation by considering the target OD matrix and observed link flows as follows [9, 10, 13, 14, 15, 16, 18, 19, 51]:

$$\begin{aligned} d &= \arg \min_{d \in S_d} [F_1(d, \bar{d}) + F_2(f, \hat{f})] \\ &= \arg \min_{d \in S_d} [(d - \bar{d})^T U^{-1}(d - \bar{d}) + (P(d)d - \hat{f})^T V^{-1}(P(d)d - \hat{f})] \end{aligned} \quad (2)$$

Where:

$P(d) = \{p_{kj}(d_j)\}$ is the assignment matrix, which represents the proportion of OD pair j using the observed link k ;

$f = P(d)d$ is the estimated link flow vector. $f = \{f_k\}$, where $f_k = \sum_j p_{kj}(d_j)d_j$.

In this study, the dispersion parameter is integrated into the objective function ([Eq 2](#)) [[16](#), [19](#)] as follows:

$$(d, \theta) = \arg \min_{\substack{d \in S_d \\ \theta \in S_t}} [F_1(d, \bar{d}) + F_2(f, \hat{f}) + F_3(\theta, \bar{\theta})] \quad (3)$$

This model can be seen as a Stochastic User Equilibrium (SUE) problem [[13](#)]. The Generalized Least Square (GLS) estimator can be used to solve [Eq 3](#) by minimizing the Weighted Euclidean Distances (WED) between the target data and the solution vector, and [Eq 3](#) can be then reorganized as shown in [Eq 4](#) [[9](#), [10](#), [15](#), [48](#)]:

$$(d, \theta)^{GLS} = \arg \min_{\substack{d \in S_d \\ \theta \in S_t}} [(d - \bar{d})^T U^{-1} (d - \bar{d}) + (P(d, \theta)d - \hat{f})^T V^{-1} (P(d, \theta)d - \hat{f}) + (\theta - \bar{\theta})^2 Q^{-1}] \quad (4)$$

Where:

$P(d, \theta) = \{p_{kj}(d_j, \theta)\}$ is the assignment matrix, and is a function of both OD matrix and dispersion parameter θ .

$f = P(d, \theta)d$ is the estimated link flow vector. $f = \{f_k\}$, where $f_k = \sum_j p_{kj}(d_j, \theta)d_j$.

The matrix for link choice proportions P can be generally assumed fixed during the optimization procedure [[10](#), [13](#), [14](#), [19](#)]. This procedure performs well for uncongested traffic conditions or an idealized traffic network with fixed link costs. However, when the network becomes congested, users' choices are increasingly influenced by adverse traffic condition. In this case, link flow and cost are not independent, and the assignment matrix P should be assumed to vary within each optimization step for link flow and OD estimation. Similarly, the GLS estimators of d and θ can be also obtained by solving [Eq 4](#).

The Link Choice Proportion Calculation Using the Dispersion Parameter. As mentioned in notation and definitions subsection, the link flow and cost will be updated when a new set of values of d and θ is received. Drivers' link choice decisions are influenced by the network-wide traffic condition, and thus the link choice proportion matrix P should be allowed to vary as well. The method of successive average (MSA) is adopted to calculate equilibrium link flows in the traffic assignment procedure [[7](#), [16](#), [45](#), [52](#)].

The cost of path r connecting the OD pair j can be expressed as:

$$c_{rj} = \sum_k a_{krj} c_k, \forall k \in K \quad (5)$$

The probability P_{rj} can be then computed according to the path choice logit model [[45](#)]:

$$\begin{aligned} P_{rj} &= \frac{\exp(-c_{rj}\theta)}{\sum_a \exp(-c_{aj}\theta)} \\ &= \frac{1}{\sum_a \exp(-c_{aj}\theta)} \exp(-\theta \sum_k a_{krj} c_k) \\ &= \frac{1}{\sum_a \exp(-c_{aj}\theta)} [\exp(-\theta a_{1rj} c_1) \cdot \exp(-\theta a_{2rj} c_2) \cdot \dots \cdot \exp(-\theta a_{Trj} c_T)] \end{aligned} \quad (6)$$

For a driver traveling along the path r , the weight assigned to link k is equal to $\exp(-c_k\theta)$. It is worth noting that the sum of probabilities over all feasible paths for each OD pair is equal to one.

As previously noted, $W' = [w'_{mn}]$ is the initial weight matrix of all possible paths connecting each OD pair. With the initial weight set to $w'_{mn} = \exp(-c_k\theta)$, then W' , W'_2 , and W'_3 represent the weight matrix in the group of paths with one link, two links and three links respectively. Therefore, the weight matrix for all possible paths can be formulated as:

$$W' + W'_2 + W'_3 + \dots = (E - W')^{-1} - E \quad (7)$$

Wong [53] and Lo and Chan [16] have proven that the right side of Eq 7 is convergent for any acyclic networks, and is equal to $W = (E - W')^{-1} - E$. Therefore, the probability of a trip from node m to node n (OD pair j) choosing link k can be calculated as follows:

$$p_{kj} = \frac{w_{mg} \exp(-\theta c_k) w_{vn}}{w_{mn}} \quad (8)$$

Where link k connects node g and node v , and w_{mn} expresses weight matrix of all possible paths connecting nodes m and n , $m, n \in M$. w_{mn} is set to 1 for all nodes in the network.

Following the previous definition, the auxiliary mean traffic flow $y_k^{(s)}$ of link k is defined for each incoming d and θ via the following equation:

$$\begin{aligned} y_k^{(s)} &= [Pd]_k \\ &= \sum_j p_{kj} d_j, \forall k \in K \end{aligned} \quad (9)$$

The equilibrium traffic link flows can be then obtained using the MSA method. Specifically, the flow of link k can be calculated at the $(s+1)$ th iteration with the following equation:

$$\begin{aligned} x_k^{(s+1)} &= x_k^{(s)} + \frac{1}{s} (y_k^{(s)} - x_k^{(s)}) \\ &= \frac{1}{s} \sum_{j=1}^s y_k^{(j)}, \forall k \in K \end{aligned} \quad (10)$$

As shown in Eq 10, the flow of link k at the $(s+1)$ th iteration is equal to the mean of the auxiliary traffic flow of link k in the previous s iterations.

When a new set of values of d and θ is received, the matrix P of link choice proportions is updated following the procedure described above, and is then integrated into the Eq 4 to update the values of d and θ . This optimization procedure continues until convergence of the OD matrix and dispersion parameter estimation is reached.

Model Solution Algorithm

To solve the Stochastic User Equilibrium (SUE) problem described above, a two-stage algorithm for GLS estimation and SUE traffic assignment is proposed: First, the OD matrix d and the dispersion parameter θ are simultaneously estimated under the condition of the fixed link flows, link costs, and weight matrix. Second, the link flows, link costs, and link choice proportions are updated according to the new values of d and θ in the SUE assignment process. The two-stage algorithm is executed iteratively until the convergence of values of d and θ is reached. Sequential quadratic programming (SQP) from the extended quasi-Newton method is chosen as the solution method [49].

Two-Stage Algorithm

The initialization procedure of the two-stage algorithm can be described as follows:

1. Initialize the counter $t = 0$, set the initial OD vector matrix $d^{(0)} = \bar{d}$, the initial dispersion parameter $\theta^{(0)} = \bar{\theta}$, and the initial link flow $x_k^{(0)} = 0, k \in K$.
2. Calculate the initial link costs for all links in the network using [Eq 1](#), and calculate the weight matrix W for all paths based on the initial link costs and $\theta^{(0)}$.
3. Calculate the link choice proportion matrix P using the weight matrix W and $\theta^{(0)}$.
4. Calculate the initial mean auxiliary traffic flow for all the observed links with [Eq 9](#), and update $t = t + 1$.

The first stage of the algorithm is described as follows:

Step 1. The objective function ([Eq 4](#)) can be updated with the new mean auxiliary observed link flows as follows:

$$(d^{(t)}, \theta^{(t)})^{GLS} = \arg \min_{\substack{d \geq 0 \\ \theta > 0}} [(d^{(t)} - \bar{d})^T U^{-1} (d^{(t)} - \bar{d}) + (P^{(t)} d^{(t)} - \hat{f})^T V^{-1} (P^{(t)} d^{(t)} - \hat{f}) + (\theta^{(t)} - \bar{\theta})^2 Q^{-1}] \quad (11)$$

Where:

$U^{-1}, P^{(t)}, V^{-1}$, and Q^{-1} can be updated using the new mean auxiliary observed link flows, estimated OD vector matrix, dispersion parameter, and link flow vector respectively;

The feasible set for d and θ should meet the requirements $d \geq 0, \theta > 0$. When the value of θ approaches zero, the path choice probabilities for all paths tend to be equal. As the value of θ increases, the path choice probabilities tend to be deterministic.

Step 2. Use the SQP algorithm to obtain a new set of values of $d^{(t)}$ and $\theta^{(t)}$ that minimizes the objective function. The starting point for optimizing the OD vector $\tilde{d}^{(t)}$ and dispersion parameter $\tilde{\theta}^{(t)}$ should be fixed in advance. During the iterative process of the SQP algorithm, whenever a new value θ is received, the link choice proportion matrix P will be updated by changing the value of $\exp(-\theta c_k)$ in [Eq 8](#), while the link cost and weight matrix should remain unchanged.

The second stage of the algorithm can be described as follows:

Step 3. Initialize the counter $s = 1$.

Step 4. Calculate the weight matrix W with the new dispersion parameter $\theta^{(t)}$.

Step 5. Calculate the link choice proportion matrix $P^{(t)}$ using the weight matrix W and dispersion parameter $\theta^{(t)}$.

Step 6. Calculate the mean auxiliary traffic flow for all observed links as follows:

$$y_l^{(s)} = [P^{(t)} d^{(t)}]_l = \sum_j p_{lj} d_j, l \in L$$

Step 7. Calculate the equilibrium traffic link flow of link k via the MSA method:

$$x_l^{(s+1)} = x_l^{(s)} + \frac{1}{s} (y_l^{(s)} - x_l^{(s)}), l \in L$$

Step 8. The maximum relative difference between current and previous mean link flows should satisfy the following requirement:

$$\max_{\forall l \in L} \left\{ \frac{|x_l^{(s+1)} - x_l^{(s)}|}{x_l^{(s+1)}} \right\} \leq \epsilon_1 \quad (12)$$

If the above requirement is met, the algorithm proceeds directly to step 11, otherwise proceed to step 9.

Step 9. Calculate the new link costs according to $x_l^{(s+1)}$, $l \in L$.

Step 10. Calculate the weight matrix using the updated link costs, set $s = s + 1$, and return to step 5.

Step 11. The maximum relative difference between the current and previous OD matrix estimates should satisfy the following requirement:

$$\max_{\forall j \in J} \left\{ \frac{|d_j^{(t)} - d_j^{(t-1)}|}{d_j^{(t)}} \right\} \leq \epsilon_2 \quad (13)$$

If the above requirement is met, terminate the procedure and output the current estimates of OD vector matrix d and dispersion parameter θ as $d^{(t)}$ and $\theta^{(t)}$. Otherwise, set $t = t + 1$, and proceed to step 12.

Step 12. Calculate the new starting points as follows: $\tilde{d}^{(t+1)} = \frac{1}{t} d^{(t)} + \frac{t-1}{t} d^{(t-1)}$, $\tilde{\theta}^{(t+1)} = \frac{1}{t} \tilde{\theta}^{(t)} + \frac{t-1}{t} \tilde{\theta}^{(t-1)}$, and return to step 1.

Model Evaluation

To evaluate the performance of the proposed method, the root mean squared errors (RMSE) for OD matrix and link flows after convergence are defined as follows:

(1) The root mean squared error (RMSE) of the estimated link flows $x_l^{(s+1)}$ relative to the true link flow x_l is computed as follows:

$$RMSE(LF) = \sqrt{\frac{1}{\Gamma} \sum_{l=1}^{\Gamma} (x_l^{(s+1)} - x_l)^2} \quad (14)$$

Similarly, the RMSE of the observed (target) link flows \hat{f}_l relative to the true link flows x_l can be defined as RMSE ($L\bar{F}$), where $x_l^{(s+1)}$ is replaced by \hat{f}_l in Eq 14.

(2) The RMSE of the estimated OD matrix $d^{(t)}$ relative to the true OD matrix d can be defined as RMSE (OD):

$$RMSE(OD) = \sqrt{\frac{1}{\tau} \sum_{j=1}^{\tau} (d_j^{(t)} - d_j)^2} \quad (15)$$

Likewise, the RMSE of target OD matrix \bar{d} relative to the true OD matrix d is defined as RMSE ($O\bar{D}$), where $d^{(t)}$ is replaced by \bar{d}_j in Eq 15.

Numerical Experiment and Result Analysis

A Hypothetical Network Test

In this section, the performance of the proposed approach is tested in a hypothetical network. The network and data proposed by Yang et al. [15] and Caggiani et al. [19] are adopted as the test bed with some slight modifications. The network (presented in Fig 1), is composed of 9 nodes (3 origin centroids and 3 destination centroids), and 14 links. The true and initial OD vector matrices d and \tilde{d} for the SQP algorithm are shown in Table 1. The initial dispersion parameter $\tilde{\theta}$ is assumed to be 4, and the true dispersion parameter $\hat{\theta}$ is fixed to 1.5. Note that the initial OD matrix \tilde{d} and dispersion parameter $\tilde{\theta}$ are quite dissimilar from those of the ground truth data.

The following parameters in the Bureau of Public Roads (BPR) [50] link performance function are used: $\alpha_k = 0.15$ and $\beta_k = 4, \forall k \in K$. In addition, the free flow travel time (t_k) and capacity (C_k) for each link are predetermined as shown in Table 2.

The ground truth link flows can be generated by allocating the true OD matrix to the traffic network using SUE-Logit assignment method presented in Section 2.3. The true dispersion parameter is $\theta = 1.5$, resulting in the link flows shown in Table 3. The set of links {5, 6, 7, 11, 13} is selected as the observed links.

In this example, we assume that the OD vector and link flow vector follow the Poisson distribution. The covariance matrices U (for OD demands) and V (for link flows) in Eq 4 can be assumed to be diagonal matrices [9, 14, 54]. The diagonal element for U , V and Q can be computed respectively through the following equations:

$$\sigma_{d_j}^2 = (cv_d \cdot d_j)^2, \sigma_{x_k}^2 = (cv_x \cdot x_k^{(s+1)})^2, \sigma_\theta^2 = (cv_\theta \cdot \theta)^2$$

Where cv_d , cv_x and cv_θ represent the variation coefficients for OD demands, link flows, and dispersion parameter respectively. Specifically, these parameters are set as $cv_d = 0.3$, $cv_x = 0.05$, and $cv_\theta = 0.1$.

The target OD matrix \bar{d} , observed link flow vectors \hat{f} , and target dispersion parameter $\bar{\theta}$ can be generated separately by adding random terms into the corresponding true values. The random terms are sampled from independent normal variables with zero means. For instance, the target OD matrix can be calculated by adding a random term with $\lambda_d = 0.3$ to the values of the true OD matrix divided by two, the observed link flow vectors can be generated by adding a random term with $\lambda_f = 0.1$, and the target parameter can be set as $\bar{\theta} = 4$. In addition, the error tolerance threshold used in the optimization is set to $\varepsilon_1 = \varepsilon_2 = 10^{-3}$. The convergence for theta is plotted in Fig 2, which shows the estimate slowly falling in the first 120 iterations before rapidly converging to the true value at 1.5099. This is a very slight deviation with the true value of 1.5. In addition, the convergence of the objective function is presented in Fig 3, where the value of the objective function sharply falls at the first iteration and then gradually decreases and levels off at a lower value. Poor initial choices of OD input vector and dispersion parameter may lead to the slower convergence.

In order to further evaluate the effectiveness of the proposed approach, a sensitivity analysis is conducted with parameter cv_θ (CVT) varying from 0.1 to 0.5 and cv_d (CVD) changing from

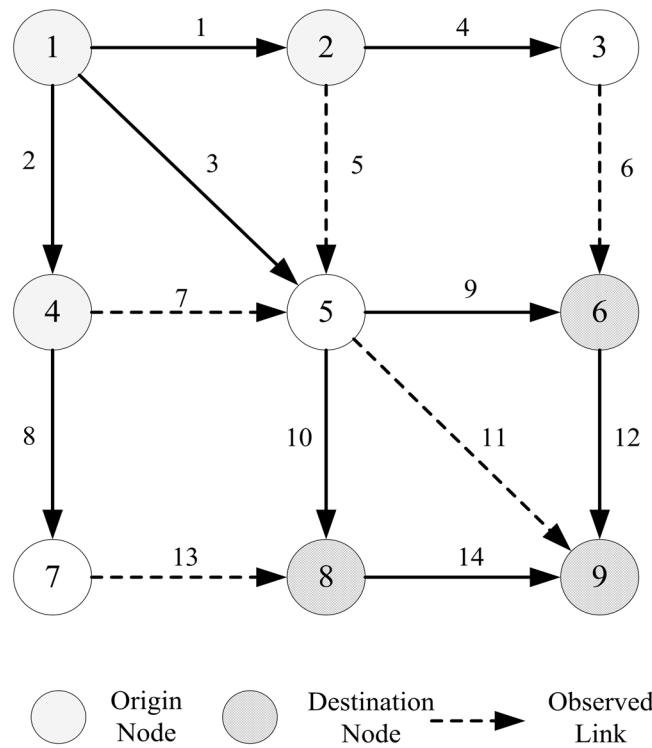


Fig 1. The test network used in the numerical example.

doi:10.1371/journal.pone.0146850.g001

Table 1. The true and initial OD vector matrices.

OD pair	1–6	1–8	1–9	2–6	2–8	2–9	4–6	4–8	4–9
j	1	2	3	4	5	6	7	8	9
d	120	150	100	130	200	90	80	180	110
\tilde{d}	30	20	10	30	30	30	30	40	20

doi:10.1371/journal.pone.0146850.t001

Table 2. Free flow travel time and capacity for each link.

link	1	2	3	4	5	6	7	8	9	10	11	12	13	14
t_k	2.0	1.5	3.0	1.0	1.0	2.0	2.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0
C_k	280	290	280	280	600	300	500	400	500	700	250	300	350	520

doi:10.1371/journal.pone.0146850.t002

Table 3. True link flows in the hypothetical network.

link	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_k	125	143	103	172	474	172	201	313	307	393	279	148	313	475

doi:10.1371/journal.pone.0146850.t003

0.1 to 1. This generates 50 different estimates for RMSE (OD), RMSE (LF) and Theta as presented in Figs 4–6.

As shown in Fig 4, RMSE (OD) increases with the variation coefficient cv_θ when cv_d falls between 0.2 and 0.8. With cv_θ fixed between 0.3 and 0.5, RMSE(OD) can be seen as a convex

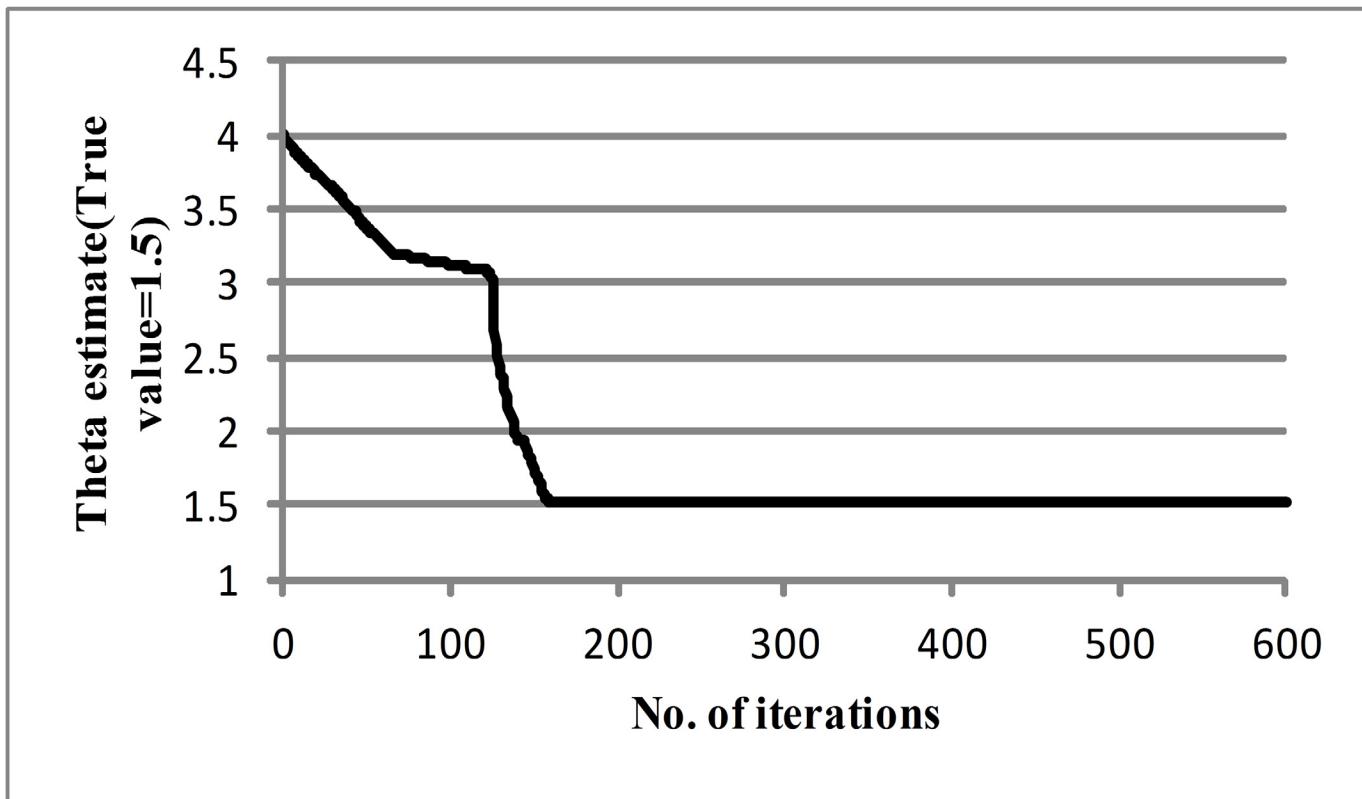


Fig 2. Convergence of the theta estimate.

doi:10.1371/journal.pone.0146850.g002

function of cv_d . Alternatively, when cv_θ is between 0.1 and 0.2, cv_d has a negligible impact on RMSE (OD). Thus, the maximum value (19.9022) of RMSE (OD) can be found at $cv_d = 0.5$ and $cv_\theta = 0.5$, and the minimum value (4.6859) is obtained at $cv_d = 0.3$ and $cv_\theta = 0.1$. Compared with the initial RMSE (OD) of 93.8971 calculated from Table 1, a 78.8% reduction is achieved at the maximum RMSE (OD), and a 95% reduction is obtained at the minimum RMSE (OD).

Fig 5 shows the impact of cv_d and cv_θ on RMSE (LF). With the value of cv_d fixed, RMSE (LF) increases with the variation coefficient cv_θ . For a fixed value of cv_θ , the RMSE (LF) decreases with an increase in cv_d . Thus, we can conclude that maximum RMSE (LF) value of 19.6597 is located at $cv_d = 0.1$ and $cv_\theta = 0.5$, and the minimum value of 8.6696 can be found at $cv_d = 0.6$ and $cv_\theta = 0.1$. Compared with the initial RMSE (LF) value of 30.8347, 36.2% and 71.9% reductions can be achieved for the maximum value of RMSE (LF) and minimum value of RMSE (LF) respectively.

As shown in Fig 6, the value of theta varies negligibly with the choice of cv_d and cv_θ . In other words, the estimated value of theta always converges to approximately the true value. As shown in Fig 6, for a fixed value of cv_d , the estimated θ is close to the true value for any given cv_θ . For example, the value of θ fluctuates between 1.37 and 1.51 when $cv_d = 0.3$. Likewise, for any fixed cv_θ , the estimated θ varies minimally about the true value of θ using the proposed method. For example, the estimated θ is between 1.35 and 1.52 for $cv_\theta = 0.1$.

The above discussion reveals a fact that the initial value of \bar{d} , $\bar{\theta}$, and observed link flow vectors \hat{f} do not affect the theta estimation performance. This is equivalent to a convex optimization problem, where the optimal results tend to converge near the true dispersion parameter value. This implies that the estimate of θ is insensitive to the variation coefficients, and can be used as a stable and accurate parameter to determine travelers' route decisions.

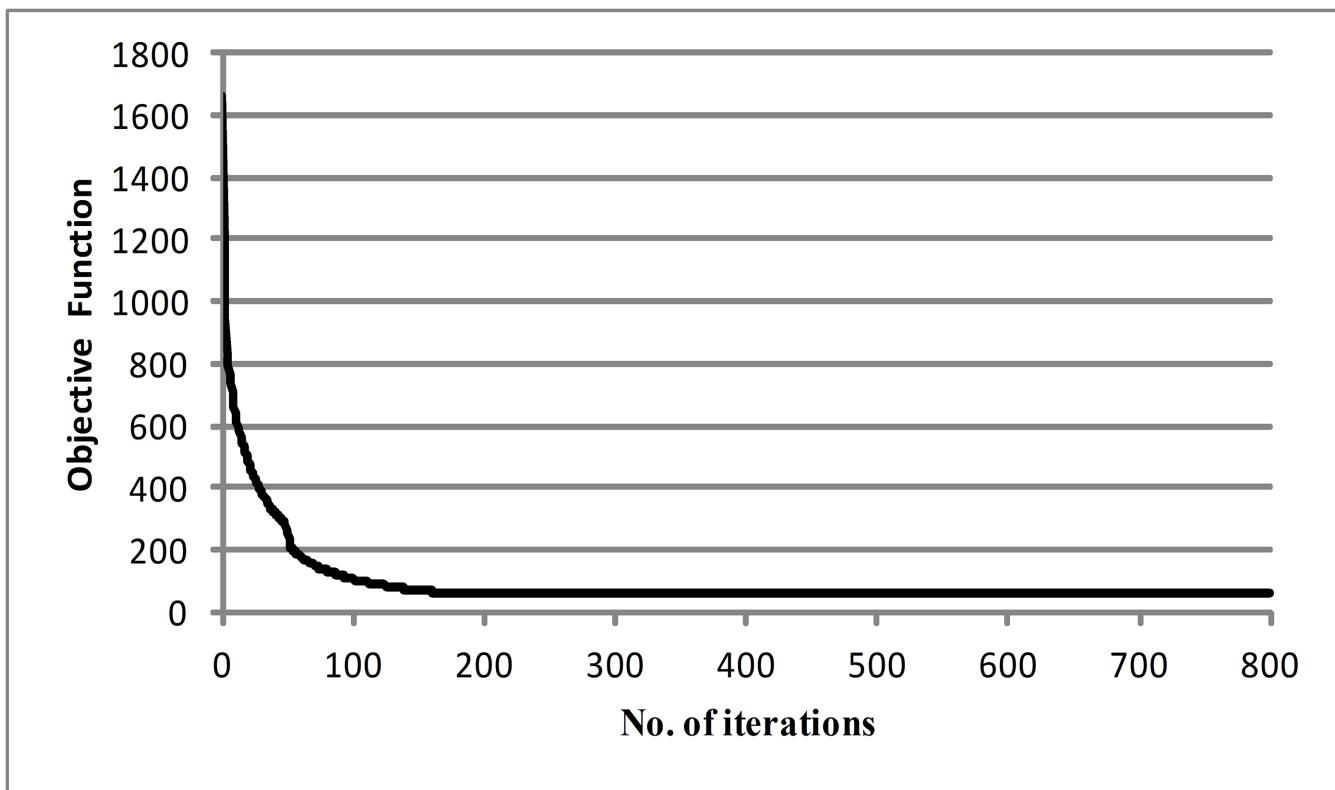


Fig 3. Convergence of the objective function.

doi:10.1371/journal.pone.0146850.g003

Comparison and Analysis

To further demonstrate the advantages of the proposed methodology, two OD matrix estimation methods are implemented and compared with the proposed approach. To make this comparison, we first implement the algorithm described in Yang et al. [15], which presents an optimization model for OD matrix estimation in congested networks using the logit-based

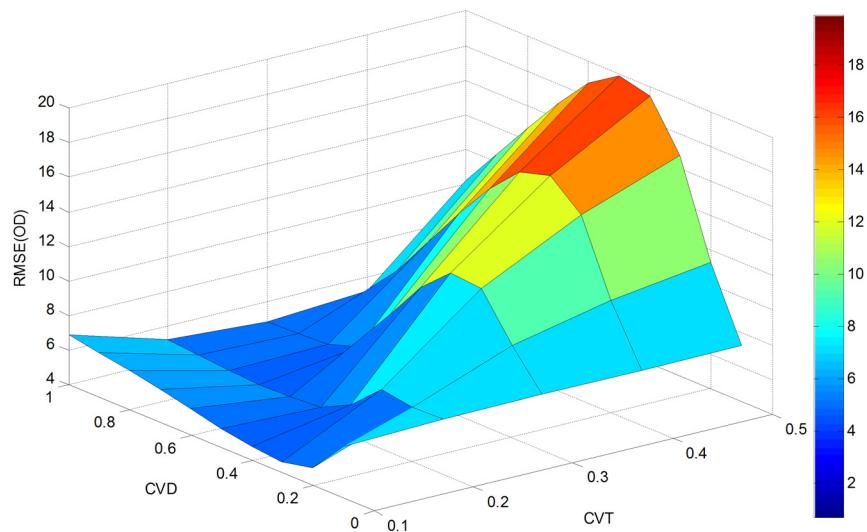


Fig 4. RMSE(OD) versus cv_d and cv_θ .

doi:10.1371/journal.pone.0146850.g004

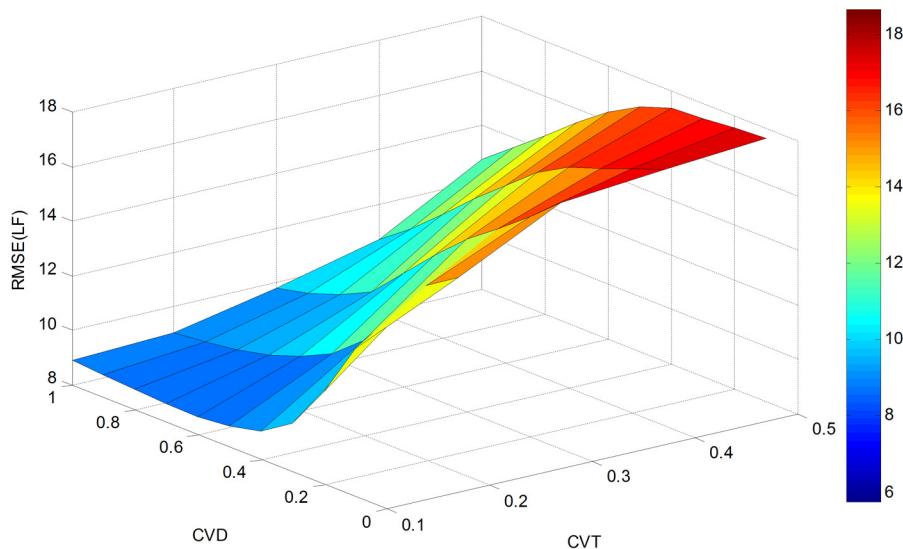


Fig 5. RMSE(LF) versus cv_d and cv_θ .

doi:10.1371/journal.pone.0146850.g005

SUE. The method described in Lo and Chan [16] is implemented for the second comparison. This method applies both statistical estimation and traffic assignment to simultaneously calculate the OD matrix and link choice proportions based on OD survey data and traffic counts. To maintain a fair comparison, the same test network and data set are applied in all cases.

The OD matrix estimation method proposed by Yang et al. [15] is given in section 2.2. The objective function is shown in Eq 16.

$$\begin{aligned}
 (d, \theta) &= \arg \min_{d \in S_d} [F_1(d, \bar{d}) + F_2(f, \hat{f})] \\
 &= \arg \min_{d \geq 0} \left[\frac{1}{2}(d - \bar{d})^2 + \frac{1}{2}(Pd - \hat{f})^2 \right] \\
 \theta &> 0
 \end{aligned} \tag{16}$$

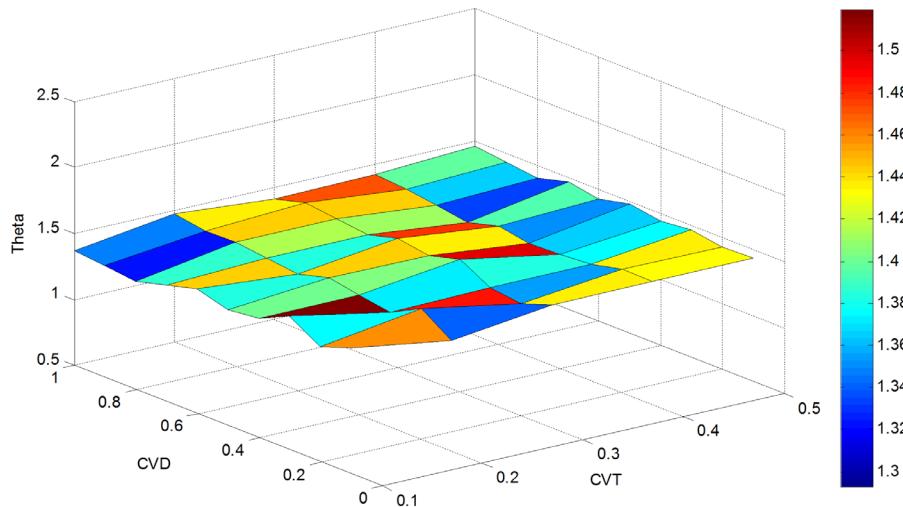


Fig 6. Estimated Theta versus cv_d and cv_θ .

doi:10.1371/journal.pone.0146850.g006

Table 4. Comparison between Yang et al.'s approach and the proposed approach with OD matrix and link flow estimation.

Approach	Estimated OD matrix		Estimated Link flows	
	RMSE (\bar{OD})	RMSE (OD)	RMSE (\bar{LF})	RMSE (LF)
Yang et al.'s approach	24.27	20.85	26.65	19.02
Proposed approach	24.27	18.79	26.65	17.39

doi:10.1371/journal.pone.0146850.t004

In Yang et al.'s work, the weighted Euclidean distance function is used to develop a unit weighting matrix and the value of theta is set to 1.5. The RMSE(OD), RMSE(LF), RMSE (\bar{OD}) and RMSE(\bar{LF}) for Yang et al.'s approach are calculated and compared with the proposed approach in [Table 4](#):

As shown in [Table 4](#), the proposed method yields significantly lower RMSE (OD) and RMSE (LF) relative to Yang et al.'s approach. Compared with the initial RMSE values, a 22.6% reduction in RMSE (OD) is achieved using the proposed approach, while only a 14.1% reduction is achieved using the method described in Yang et al. Similarly, the proposed approach resulted in a 34.7% reduction in RMSE(LF), while only a 28.6% reduction was achieved using Yang et al.'s approach. One reason that the dispersion parameter is estimated and integrated into the [Eq 3](#) by $F_3(\theta, \bar{\theta})$ in the proposed method, and it is able to yield a better estimate of the dispersion parameter than previous approaches. The other reason is that the covariance matrices U (for OD demands), V (for link flows) and Q (for dispersion parameter) are not a fixed variable during the calculation. These improvements can help the method enhance the estimation performance for the OD matrix and link flow vectors.

Lo and Chan [16] present the following maximum likelihood objective function:

$$(d, \theta) = \arg \max_{d \geq 0, \theta > 0} \ln L(\theta, d | \hat{f}, \bar{d}) \quad (17)$$

In Lo and Chan [16], it is assumed that the observed flows are equal to the true flows in the test network. For Lo and Chan's algorithm, we set the target dispersion parameter to $\bar{\theta} = 4$ (This is also equal to the initial dispersion parameter value used in Lo and Chan [16]'s work), and the variation coefficients as follows: $cv_\theta = 0.1$, $cv_x = 0.05$, and $cv_d = 0.3$. In order to evaluate the performance of the proposed approach relative to that of Lo and Chan [16]'s method, RMSE (OD), RMSE (LF), and the estimated Theta are selected for comparison and shown in [Table 5](#).

Unlike Lo and Chan's method, random terms are added to the observed link flows in the proposed approach, thus introducing additional challenges for estimation. However, the results presented in [Table 5](#) demonstrate that the method proposed in this paper outperforms Lo and Chan's approach in terms of OD matrix, link flow, and Theta estimation accuracy.

Table 5. Comparison between Lo and Chan's approach and the proposed approach with OD matrix, link flow and Theta estimation.

Approach	Estimated OD matrix		Estimated Theta	
	RMSE (OD)	RMSE (LF)	Theta target	Theta estimated
Lo and Chan's approach	5.34	12.08	4	1.572
Proposed approach	4.69	9.77	4	1.509

doi:10.1371/journal.pone.0146850.t005

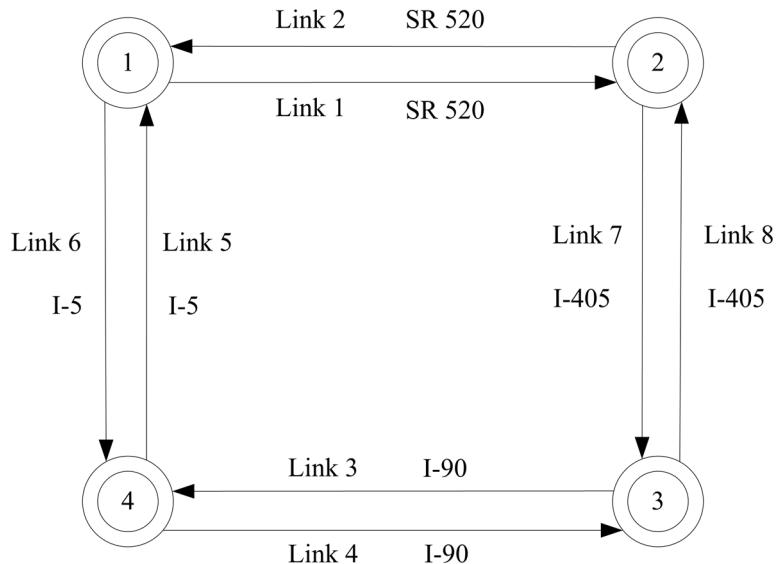


Fig 7. Square network in Seattle. Double circle nodes represent zone centroids (origins and destinations).

doi:10.1371/journal.pone.0146850.g007

Application to A Square Network in Seattle

A square network in Seattle is used as a congested network case study to demonstrate the applicability and transferability of the proposed approach in a real-world traffic network (Shown in Fig 7). Empirical data was collected from loop detectors located along one freeway section in Seattle area, and obtained for this research through the Strategic Highway Research Program 2 (SHRP 2 program) supported by Washington State Department of Transportation (WSDOT) [55].

The square test network used in this case study consists of 4 nodes and 8 links, where all nodes are centroids (origins and destinations). The topology of the test network is outlined in Fig 7. We assume that the study network is acyclic, such that the traffic flow starting from one node will leave the network before returning to the original node. Specifically, Links 1 and 2 represent the SR 520 Bridge connecting I-5 in Seattle and SR 202 in Redmond. Interstate 90 (I-90) is represented by Links 3 and 4, and Interstate 5 (I-5) is represented by Links 5 and 6. Links 7 and 8 represent Interstate 405 (I-405), which intersects I-90 in the south and SR 520 in the north.

Traffic flows were obtained from loop detectors installed at nodes 1, 2, 3 and 4, illustrated in Fig 8. The parameters for the BPR link performance cost function (Eq 18) were estimated based on the empirical data and are presented in Table 6.

$$c_k(f_k) = t_k \left[1.0 + \alpha_k \left(\frac{f_k}{C_k} \right)^{\beta_k} \right], \forall k \in K \quad (18)$$

Table 7 indicates the external traffic flow recorded for each node during peak hour, where 1-Link 1 represents the external traffic flow on Link 1 from node 1, and 2-Link 7 represents the external traffic flow on Link 7 from node 2, and so forth. To convert true link flows into a ground truth OD matrix, the flow proportion for each node $\eta = 0.6$ is assumed based on extensive video records and field surveys. This implies that, for the traffic leaving each node, 60% exits the network from an adjacent node while 40% exits from the other nodes. In order to avoid circular flow in the OD calculation process, it is assumed that the final remaining traffic

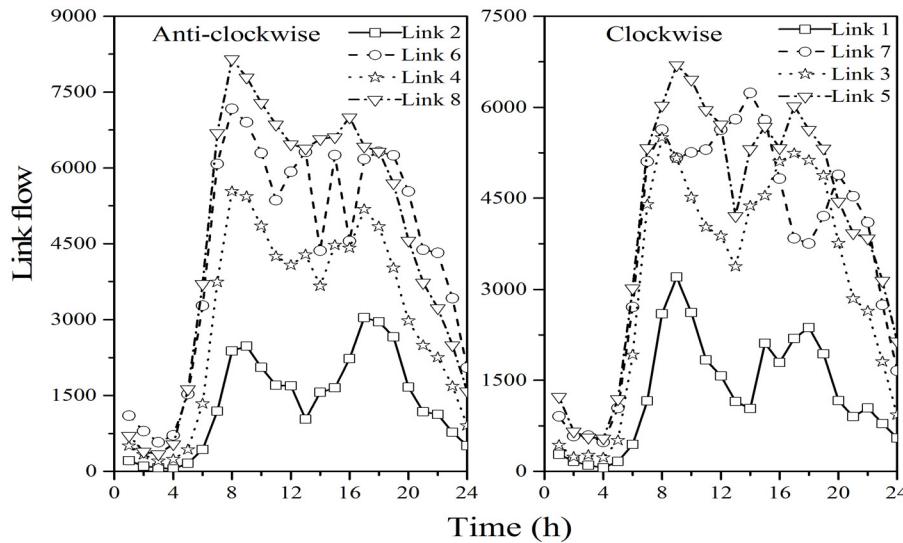


Fig 8. Traffic flow fluctuation by time of day.

doi:10.1371/journal.pone.0146850.g008

flow will leave from the last node before returning to the original node. Based on these assumptions, the ground truth OD matrix is calculated and shown in [Table 8](#). In addition, the initial OD matrix \tilde{d} can be computed by rounding the last digit of the true OD matrix as shown in [Table 8](#).

The true OD matrix in [Table 8](#) is then used to assign the corresponding traffic flow into each link according to [Eq 5](#) through [Eq 10](#). The calculated traffic flows can be assumed to represent the true link flows, where link 1, 3, 5, 6, and 8 are selected as the observed links to estimate OD matrix shown in [Table 9](#).

Similar to the hypothetical network, we assume that the OD demands and observed link flows follow the Poisson distribution, and the covariance matrices U and V can be assumed to be diagonal. The initial value of the dispersion parameter $\tilde{\theta}$ is set to 40.5. The remaining input parameters are set identically to the hypothetical network. In addition, a sensitivity analysis with 50 different combinations of variation coefficients cv_d and cv_θ was conducted to investigate the optimal parameter initialization for the proposed approach. The results of this sensitivity analysis are shown in Figs [9–11](#).

Table 6. BPR link performance cost function parameter calibration.

Links	Road Name	Length(km)	$t_k(h)$	$C_k(\text{pcu/h})$	α_k	β_k
1,2	SR 520	7.5	0.1162	4149	0.1450	3.5
3,4	I-90	3.6	0.0667	8685	0.1035	2.7
5,6	I-5	3.5	0.1016	9683	0.0988	2.7
7,8	I405	7.8	0.1332	7961	0.1242	3.5

doi:10.1371/journal.pone.0146850.t006

Table 7. The external true traffic flow for each node at peak hour.

Link direction	1-Link 1	2-Link 2	3-Link 3	4-Link 4	4-Link 5	1-Link 6	2-Link 7	3-Link 8
Traffic flow	3199	2480	5499	5535	6018	7169	5628	8153

doi:10.1371/journal.pone.0146850.t007

Table 8. True OD matrix and initial OD matrix at peak hour for each OD pair.

OD pair	1–2	1–3	1–4	2–1	2–3	2–4	3–1	3–2	3–4	4–1	4–2	4–3
j	1	2	3	4	5	6	7	8	9	10	11	12
d	3067	2489	4814	2389	3774	1946	3277	5772	4604	4497	2773	4284
\tilde{d}	307	249	482	239	378	195	328	578	461	450	278	429

doi:10.1371/journal.pone.0146850.t008

Table 9. Observed link flows at peak hour.

Link No.	1	2	3	4	5	6	7	8
x_k	3711	-	8282	-	8439	7857	-	7591

doi:10.1371/journal.pone.0146850.t009

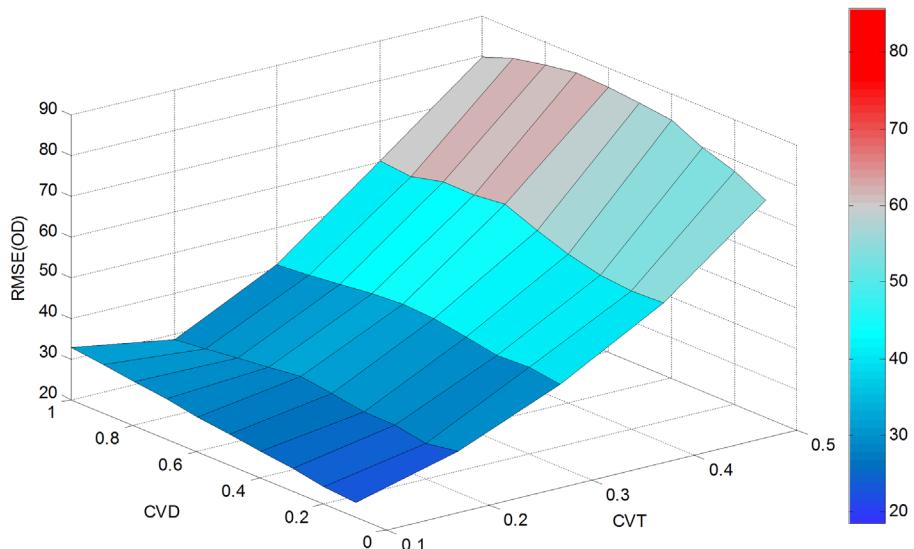


Fig 9. RMSE(OD) versus cv_d and cv_θ in the actual network.

doi:10.1371/journal.pone.0146850.g009

[Fig 9](#) shows the value of RSME (OD) versus cv_d and cv_θ . For a fixed value of cv_d between 0.1 and 0.7, RMSE(OD) increases with cv_θ . RMSE(OD) is a concave function of cv_θ when cv_d is fixed between 0.8 and 1.0, and a convex function of cv_d for a fixed value of cv_θ between 0.1 and 0.5. Thus, the maximum RMSE (OD) of 85.6123 can be obtained at $cv_d = 0.7$ and $cv_\theta = 0.5$, and the minimum value of 23.5917 can be obtained at $cv_d = 0.1$ and $cv_\theta = 0.1$.

As shown in [Fig 10](#), the value of RMSE (LF) increases with cv_θ for a fixed value cv_d . For a fixed value of cv_θ , the value of RMSE (LF) decreases with an increase of cv_d . The maximum RMSE (LF) of 82.5113 is found at $cv_d = 0.1$ and $cv_\theta = 0.5$, and the minimum value of 7.3277 at $cv_d = 1.0$ and $cv_\theta = 0.1$.

As noted in the hypothetical case, the choice of cv_d and cv_θ has very little impact on the estimation of Theta. As shown in [Fig 11](#), the estimated dispersion parameter θ is between 20.8327 ($cv_d = 0.5$ and $cv_\theta = 0.5$) and 22.7165 ($cv_d = 0.7$ and $cv_\theta = 0.2$) in all cases. The best estimate of dispersion parameter θ can be found between 20.8327 and 22.7165.

Finally, using the BPR link performance cost function parameters described in [Table 6](#), different combinations of variation coefficient $cv_d = 0.3$ and $cv_\theta = 0.1$; $cv_d = 0.5$ and $cv_\theta = 0.5$; $cv_d = 0.7$ and $cv_\theta = 0.2$ are used to estimate theta for the actual network.

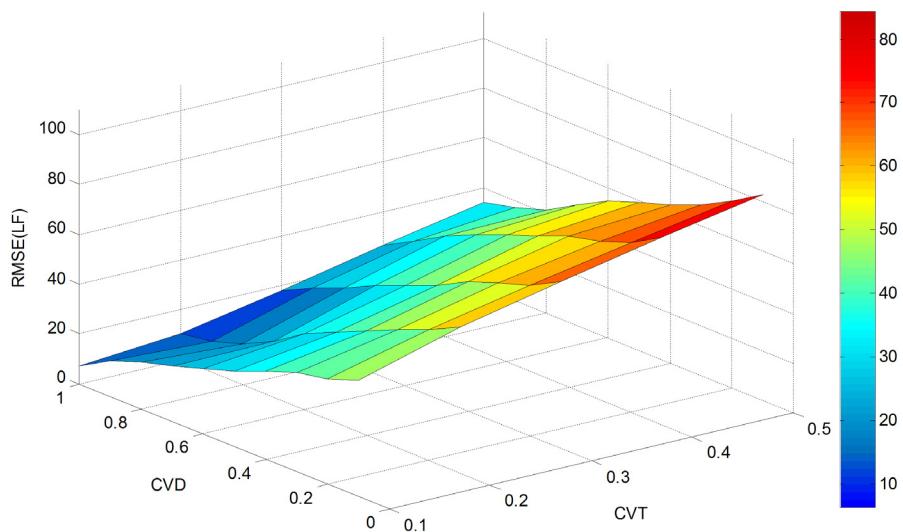


Fig 10. RMSE(LF) versus cv_d and cv_θ in the actual network.

doi:10.1371/journal.pone.0146850.g010

It is interesting to observe that the estimated RMSE(OD), RMSE(LF), and Theta for both hypothetical and actual networks exhibit a similar trend yet have obvious differences. Two primary reasons may explain these differences: First, the network topology is quite different for the two scenarios. The hypothetical network is unidirectional, where each node can be either origin or destination. In contrast, the actual network is bidirectional, where each node is both origin and destination, and thus multiple paths may exist between each OD pair. For example, the traffic flows on both 1-Link 1 and 1-Link 6 contribute to the OD demands from node 1 to node 2. Second, compared with the hypothetical network with equal cost parameters for all links, a more realistic BPR link performance cost function is adopted for the actual network. In the real-world network, the parameters (e.g. free-flow travel time and link capacity) are

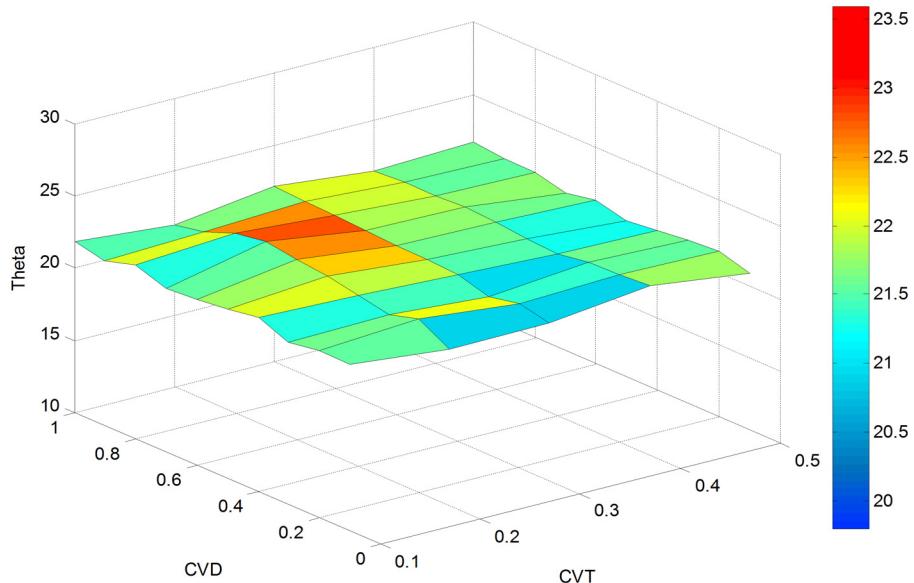


Fig 11. Theta estimated versus cv_d and cv_θ in the actual network.

doi:10.1371/journal.pone.0146850.g011

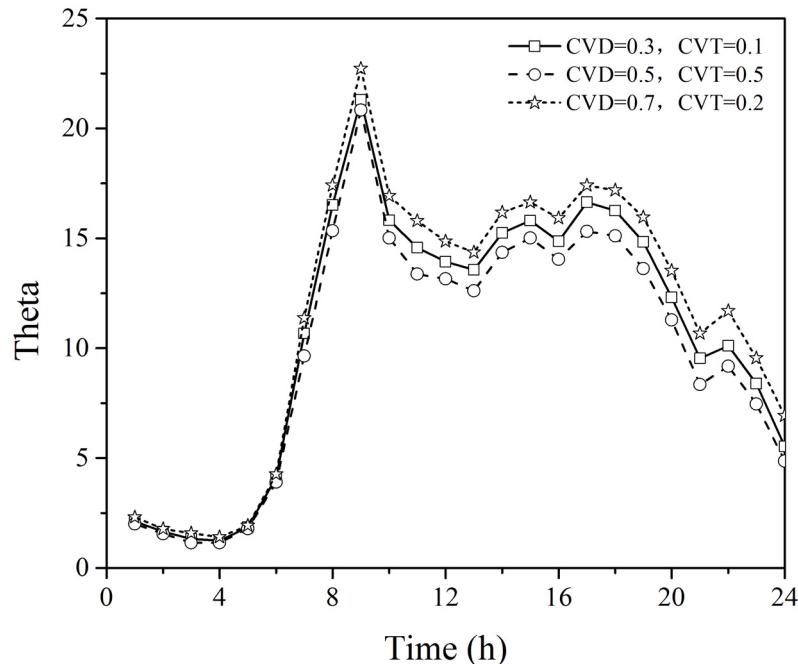


Fig 12. Estimated Theta values in the actual network by time of day.

doi:10.1371/journal.pone.0146850.g012

calibrated for each link based on empirical data. That said, the sensitivity analysis for Theta produced similar results for both the hypothetical and actual networks, indicating that this parameter is not sensitive to the choice of variation coefficients. In addition, the theta estimates obtained using a range of different parameter settings exhibits a similar and regular trend over time of day as shown in Fig 12. These findings provide guidance for initial parameter selection, and offer useful insight for interpreting modeling results.

Conclusions

This paper proposes a two-stage algorithm to simultaneously estimate origin-destination matrices and link choice proportions by incorporating a dynamic dispersion parameter into the route choice model. The dispersion parameter θ is of practical significance in describing travelers' route choice decisions, but has typically been assumed constant in previous studies. Finding the optimal dispersion parameter is not a straightforward task. To address this issue, this paper presents a model calibration procedure to simultaneously estimate the dispersion parameter θ , link choice proportions, and OD matrix. In order to obtain the Generalized Least Square (GLS) estimators of the above listed parameters, a two-stage algorithm is proposed which integrates GLS estimation into the SUE traffic assignment procedure. The first and second stages of the algorithm are applied iteratively until the maximum relative difference presented in Step 11 is achieved, after which the estimated OD matrix, link choice proportion, and dispersion parameter θ can be obtained. The SQP approach based on the extended quasi-Newton method is used to search for the optimal solution in the first stage of the algorithm. The SUE traffic assignment procedure is applied to incorporate both OD matrix and link choice proportion estimation into the second stage of the algorithm, and MSA is used to obtain the equilibrium link flows.

A hypothetical network was constructed to test the performance of the proposed approach, followed by a comprehensive sensitivity analysis with 50 combinations of variation coefficient

combinations cv_d (CVD) and cv_θ (CVT) to investigate the stability of the estimated OD matrix, link flows, and Theta. A comparison with two different methods described in Yang et al. [15] and Lo and Chan [16] suggests that the proposed approach can achieve superior performance in terms of RMSE (OD), RMSE (LF), and accuracy of the estimated Theta parameter. Moreover, a case study is presented using a real-world congested square network in Seattle, WA to demonstrate the practicality of the proposed approach, in which the true OD matrix and observed link flows are calculated via ground-truth traffic count data collected by loop detectors. The proposed method is shown to be robust under a range of initial parameter values. The RMSE (OD) can be reduced from 3426.9 to 23.6 at $cv_d = 0.1$ and $cv_\theta = 0.1$ when traffic flows are observed on five out of eight links. In addition, the estimated dispersion parameter exhibits a consistent and regular trend by time of day for all combinations of initial parameters. For future research, the proposed approach should be tested on a network of greater complexity and size, and the impact of input data inaccuracy should be considered. Additionally, further work is needed to determine the number and location of observed links required for accurate OD estimation using the proposed approach.

Supporting Information

S1 Dataset. The dataset includes the Link Speed Data and Link Volume data, and the data were collected from loop detectors located along the freeway section (I-5, I-90, I-405 and SR 520) in Seattle area, and are retrieved via the Strategic Highway Research Program 2 (SHRP 2 program). The file named as “S1 Link Speed Data” records the average speed for all links every 20-second time interval, and the other file named as “S1 Link Volume data” records volume for all links every 20-second time interval.

(RAR)

Acknowledgments

The helpful comments from the two anonymous reviewers are gratefully acknowledged. This research is supported by National Natural Science Foundation of China (Project No. 71402011, 71471024, 51408019, 71301180, 51329801), National Social Science Foundation of Chongqing of China (No. 2013YBJJ035), and the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No. KJ1400307), and the Natural Science Foundation of Chongqing of China (No. cstc2015jcyjA30012), National Key Technologies R&D Program of China (2014BAG01B03), Science and Technology Project on Transportation Construction by the Ministry of Transport of China (2015318835200).

Author Contributions

Conceived and designed the experiments: Yong W. XM Yinhai W. Performed the experiments: Yong W. XM YL KG. Analyzed the data: Yong W. XM YL KG MX. Contributed reagents/materials/analysis tools: KCH Yinhai W. Wrote the paper: Yong W.

References

1. Simini F, González M C, Maritan A, Barabási A-L. A universal model for mobility and migration patterns. *Nature*, 2012; 484: 96–100. doi: [10.1038/nature10856](https://doi.org/10.1038/nature10856) PMID: [22367540](https://pubmed.ncbi.nlm.nih.gov/22367540/)
2. Simini F, Maritan A, Néda Z. Human mobility in a continuum approach. *PloS one*, 2013; 8(3): e60069. doi: [10.1371/journal.pone.0060069](https://doi.org/10.1371/journal.pone.0060069) PMID: [23555885](https://pubmed.ncbi.nlm.nih.gov/23555885/)
3. Yan X-Y, Zhao C, Fan Y, Di Z, Wang W-X. Universal predictability of mobility patterns in cities. *Journal of the Royal Society Interface*, 2014; 11: 0834.

4. Gao Z K, Jin N D. A directed weighted complex network for characterizing chaotic dynamics from time series. *Nonlinear Analysis: Real World Applications*, 2012; 13(2): 947–952.
5. Tang J J, Wang Y H, Wang H, Zhang S, Liu F. Dynamic analysis of traffic time series at different temporal scales: A complex networks approach. *Physica A: Statistical Mechanics and its Applications*, 2014; 405, 303–315.
6. Gao Z K, Yang Y X, Fang P C, Zou Y, Xia C Y, Du M. Multiscale complex network for analyzing experimental multivariate time series. *Europhysics Letters*, 2015; 109(3): 30005.
7. Cheung WM, Wong SC, Tong CO. Estimation of a time-dependent origin-destination matrix for congested highway networks. *Journal of Advanced Transportation*, 2010; 40(1): 95–117.
8. Ma XL, Yu HY, Wang YP, Wang YH. Large-scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. *PloS one*, 2015; 10(3): e0119044. doi: [10.1371/journal.pone.0119044](https://doi.org/10.1371/journal.pone.0119044) PMID: [25780910](#)
9. Cascetta E. Estimation of trip matrices from traffic counts and survey data: a generalized least squares approach. *Transportation Research Part B: Methodological*, 1984; 18(4–5): 289–299.
10. Cascetta E, Nguyen S. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 1988; 22(6): 437–455.
11. Yang H, Sasaki T, Iida Y, Asakura Y. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transportation Research Part B: Methodological*, 1992; 26(6), 417–434.
12. Hazelton ML. Some comments on origin-destination matrix estimation. *Transportation Research Part A: Policy and Practice*, 2003; 37(10), 811–822.
13. Cantarella GE. A general fixed-point approach to multimode multi-user equilibrium assignment with elastic demand. *Transportation Science*, 1997; 31(2), 107–128.
14. Cascetta E, Postorino MN. Fixed point approaches to the estimation of O/D matrices using traffic counts on congested networks. *Transportation Science*, 2001; 35(2): 134–147.
15. Yang H, Meng Q, Bell MGH. Simultaneous estimation of the origin-destination matrices and travel-cost coefficient for congested networks in a stochastic user equilibrium. *Transportation Science*, 2001; 35 (2): 107–123.
16. Lo HP, Chan CP. Simultaneous estimation of an origin-destination matrix and link choice proportions using traffic counts. *Transportation Research Part A: Policy and Practice*, 2003; 37(9): 771–788.
17. Manley E. Estimating urban traffic patterns through probabilistic interconnectivity of road network junctions. *PLOS ONE*, 2015; 10(5): e0127095. doi: [10.1371/journal.pone.0127095](https://doi.org/10.1371/journal.pone.0127095) PMID: [26009884](#)
18. Ródenas RG, Marín Á. Simultaneous estimation of the origin–destination matrices and the parameters of a nested logit model in a combined network equilibrium model. *European Journal of Operational Research*, 2009; 197(1): 320–331.
19. Caggiani L, Ottomanelli M, Sassanelli D. A fixed point approach to origin-destination matrices estimation using uncertain data and fuzzy programming on congested networks. *Transportation Research Part C: Emerging Technologies*, 2013; 28: 130–141.
20. Shao H, Lam WHK, Sumalee A, Chen A, Hazelton ML. Estimation of mean and covariance of peak hour origin-destination demands from day-to-day traffic counts. *Transportation Research Part B: Methodological*, 2014; 68: 52–75.
21. Van Zuylen HJ, Willumsen LG. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 1980; 14(3): 281–293.
22. Bell MGH. The estimation of an origin destination matrix from traffic counts. *Transportation Science*, 1983; 17(2): 198–217.
23. Spiess H. A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 1987; 21(5): 395–412.
24. Maher MJ. Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach. *Transportation Research Part B: Methodological*, 1983; 17(6): 435–447.
25. Bell MGH. The estimation of origin-destination matrices by constrained generalized least squares. *Transportation Research Part B: Methodological*, 1991; 25(1): 13–22.
26. Sheu JB. A composite traffic flow modeling approach for incident-responsive network traffic assignment. *Physica A: Statistical Mechanics and its Applications*, 2006; 367: 461–478.
27. Beckmann M, McGuire CB, Winsten CB. *Studies in the Economics of Transportation*. New Haven: Yale University Press; 1956.
28. Fisk CS, Boyce DE. A note on trip matrix estimation from link traffic count data. *Transportation Research Part B: Methodological*, 1983; 17(3): 245–250.

29. Lam WHK, Huang HJ. A combined trip distribution and assignment model for multiple user classes. *Transportation Research Part B: Methodological*, 1992; 26(4): 275–287.
30. Fisk CS. On combining maximum entropy trip matrix with user optimal assignment. *Transportation Research Part B: Methodological*, 1988; 22(1): 69–73.
31. Fisk CS. Trip matrix estimation from link traffic counts: The congested network case. *Transportation Research Part B: Methodological*, 1989; 23(5): 331–336.
32. Han SJ. A route-based solution algorithm for dynamic user equilibrium assignments. *Transportation Research Part B: Methodological*, 2007; 41(10): 1094–1113.
33. Lu CC, Mahmassani HS, Zhou XS. A bi-criterion dynamic user equilibrium traffic assignment model and solution algorithm for evaluating dynamic road pricing strategies. *Transportation Research Part C: Emerging Technologies*, 2008; 16(4): 371–389.
34. Inoue SI, Maruyama T. Computational Experience on Advanced Algorithms for User Equilibrium Traffic Assignment Problem and Its Convergence Error. *Procedia-Social and Behavioral Sciences*, 2012; 43: 445–456.
35. Kumar A, Peeta S. Entropy weighted average method for the determination of a single representative path flow solution for the static user equilibrium traffic assignment problem. *Transportation Research Part B: Methodological*, 2015; 71: 213–229.
36. Zhang HL, Mahmassani HS, Lu CC. Dynamic pricing, heterogeneous users and perception error: Probit-based bi-criterion dynamic stochastic user equilibrium assignment. *Transportation Research Part C: Emerging Technologies*, 2013; 27: 189–204.
37. Daganzo CF, Sheffi Y. On stochastic models of traffic assignment. *Transportation Science*, 1977; 11(3): 253–274.
38. Sheffi Y, Powell WB. An algorithm for the equilibrium assignment problem with random link times. *Networks*, 1982; 12(2): 191–207.
39. Liu S, Fricker JD. Estimation of a trip table and the θ parameter in a stochastic network. *Transportation Research Part A: Policy and Practice*, 1996; 30(4): 287–305.
40. Meng Q, Lam WH, Yang L. General stochastic user equilibrium traffic assignment problem with link capacity constraints. *Journal of Advanced Transportation*, 2008; 42(4): 429–465.
41. Meng Q, Liu Z. Mathematical models and computational algorithms for probit-based asymmetric stochastic user equilibrium problem with elastic demand. *Transportmetrica*, 2012; 8(4): 261–290.
42. Lam WHK, Yin YF. An activity-based time-dependent traffic assignment model. *Transportation Research Part B: Methodological*, 2001; 35(6): 549–574.
43. Lam WHK, Li ZC, Huang HJ, Wong SC. Modeling time-dependent travel choice problems in road networks with multiple user classes and multiple parking facilities. *Transportation Research Part B: Methodological*, 2006; 40(5): 368–395.
44. Londono G, Lozano A. Dissuasive queues in the time dependent traffic assignment problem. *Procedia-Social and Behavioral Sciences*, 2014; 162: 378–387.
45. Bell MGH. Alternatives to dial's logit assignment algorithm. *Transportation Research Part B: Methodological*, 1995; 29(4): 287–295.
46. Conti PL, Giovanni LD, Naldi M. Blind maximum likelihood estimation of traffic matrices under long-range dependent traffic. *Computer Networks*, 2010; 54(15): 2626–2639.
47. Guo XL, Yang H, Liu TL. Bounding the inefficiency of logit-based stochastic user equilibrium. *European Journal of Operational Research*, 2010; 201(2), 463–469.
48. Akamatsu T. A dynamic traffic equilibrium assignment paradox. *Transportation Research Part B: Methodological*, 2000; 34(6): 515–531.
49. Boggs PT, Tolle JW. Sequential quadratic programming for large-scale nonlinear optimization. *Journal of Computational and Applied Mathematics*, 2000; 124(1–2): 123–137.
50. Bureau of Public Roads. Traffic assignment manual. U.S. Department of Commerce, Urban Planning Division, Washington, D. C., 1964.
51. Lu ZB, Rao WM, Wu YJ, Guo L, Xia JX. A Kalman filter approach to dynamic OD flow estimation for urban road networks using multi-sensor data. *Journal of Advanced Transportation*, 2015; 49(2): 210–227.
52. Liu HX, He XZ, He BS. Method of successive weighted averages (MSWA) and self-regulated averaging schemes for solving stochastic user equilibrium problem. *Networks and Spatial Economics*, 2009; 9(4): 485–503.
53. Wong SC. On the convergence of Bell's logit assignment formulation. *Transportation Research Part B: Methodological*, 1999; 33(8): 609–616.

54. Cascetta E, Russo F. Calibrating aggregate travel demand model with traffic counts: estimators and statistical performance. *Transportation*, 1997; 24(3): 271–293.
55. Ma X, Wu Y, and Wang Y. DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis. *Transportation Research Record: Journal of the Transportation Research Board*. 2011; 2215: 37–49.

Assessing the Quality of Origin–Destination Matrices Derived from Activity Travel Surveys

Results from a Monte Carlo Experiment

Mario Cools, Elke Moons, and Geert Wets

To support policy makers combating travel-related externalities, quality data are required for the design and management of transportation systems and policies. To this end, much money has been spent on collecting household- and person-based data. The main objective of this paper is to assess the quality of origin–destination (O-D) matrices derived from household activity travel surveys. To this purpose, a Monte Carlo experiment is set up to estimate the precision of O-D matrices given different sampling rates. The Belgian 2001 census data, containing work- and school-related travel information for all 10,296,350 residents, are used for the experiment. For different sampling rates, 2,000 random stratified samples are drawn. For each sample, three O-D matrices are composed: one at the municipality level, one at the district level, and one at the provincial level. The correspondence between the samples and the population is assessed by using the mean absolute percentage error (MAPE) and a censored version of the MAPE (MCAPE). The results show that no accurate O-D matrices can be derived directly from these surveys. Only when half of the population is queried is an acceptable O-D matrix obtained at the provincial level. Therefore, use of additional information to grasp better the behavioral realism underlying destination choices and collection of information about particular O-D pairs by means of vehicle intercept surveys are recommended. In addition, results suggest using the MCAPE next to traditional criteria to examine dissimilarities between different O-D matrices. An important avenue for further research is the investigation of the effect of sampling proportions on travel demand model outcomes.

In modern cosmopolitan society, travel is a cornerstone for human development, for both personal and commercial reasons: travel is not only regarded as one of the boosting forces behind economic growth, but is also seen as a social need providing people the opportunity for self-fulfillment and relaxation. As a result of the continuous evolution of modern society (e.g., urban sprawl, increasing female participation in labor, decline in traditional household structures), transportation challenges have accrued and have become more complex (1). Consequently, combating environmental (e.g., greenhouse

gas emissions, noise), economic (e.g., use of nonrenewable energy sources, time lost due to congestion), and societal (e.g., health problems such as cardiovascular and respiratory diseases, traffic casualties, community severance and loss of community space) repercussions is a tremendous task (2).

To support policy makers in addressing these externalities, quality data are required for the design and management of transportation systems and policies (3). To this end, during the last four decades, a lot of money has been spent on collecting household- and person-based data. For most metropolitan areas, the largest part of planning budgets (an estimated \$7.4 million per year) was devoted to the conduct of household and person travel surveys (4). The data collected by these surveys are used for a wide variety of applications, including traffic forecasting, transportation planning and policy, and system monitoring (3).

The main objective of this paper is to assess the quality of origin–destination (O-D) matrices derived from travel surveys. O-D matrices are core components in both traditional four-step and modern activity-based travel demand models. A sample size experiment is set up to estimate the precision of the O-D matrices given different sampling rates. Thus, an assessment of the appropriateness of travel surveys for deriving O-D relations can be made. Note that different types of travel surveys exist: Cambridge Systematics (5) distinguished seven different commonly used types of surveys (household activity travel; vehicle intercept and external; transit on-board; commercial vehicle; workplace and establishment; hotel–visitor; and parking). Each of these survey types provides a unique perspective for input into travel demand models. In this paper, the term “travel survey” is confined to the household activity travel survey.

In a household activity travel survey, respondents are queried about their household characteristics, the personal characteristics of household members, and about recent activity travel experiences of some or all household members. For most regions, household activity travel surveys remain the best source of trip generation and distribution data, and therefore are an important building block for travel demand models. In addition to model building purposes, these surveys are also used to poll specific target populations (such as transit users and nonusers), to assess the potential demand and level of public support for major infrastructural projects, and to create a deeper understanding of travel behavior in the region (5). For a more elaborate discussion concerning travel surveys the reader is referred to *The Online Travel Survey Manual: A Dynamic Document for Transportation Professionals* (3), Cambridge Systematics’ *Travel Survey Manual* (5), and Tourangeau et al. (6). Recent trends in household travel surveys are discussed by Stopher and Greaves (7).

Transportation Research Institute, Hasselt University, Wetenschapspark 5, Bus 6, BE-3590 Diepenbeek, Belgium. Corresponding author: G. Wets, geert.wets@uhasselt.be.

Transportation Research Record: Journal of the Transportation Research Board, No. 2183, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 49–59.
DOI: 10.3141/2183-06

The remainder of this paper is organized as follows. The next section provides an extended discussion on the setup of the sample size experiment. The relationship between sampling rates and the precision of a general statistic (i.e., the proportion of the commuting population) follow, along with the results and corresponding discussion of the statistical analysis of the main sample size experiment. Finally, some general conclusions are formulated and avenues for further research indicated.

SETUP OF SAMPLE SIZE EXPERIMENT

As mentioned in the introduction, the main goal of this paper is the assessment of the quality of O-D matrices derived from household activity travel surveys and, consequently, providing an answer to the question of how large a sample size should be to provide accurate O-D information in a region. To this end a Monte Carlo experiment is set up to estimate the precision of the O-D matrices given different sampling rates. A Monte Carlo experiment involves the use of random sampling techniques and computer simulation to obtain approximate solutions to mathematical problems. It involves repeating a simulation process, using in each simulation a particular set of values of random variables generated in accordance with their corresponding probability distribution functions (8). A Monte Carlo experiment is a viable approach for obtaining information about the sampling distribution of a statistic (in this study the precision of an O-D matrix) of which a theoretical sampling distribution may not be available due to the complexity. Monte Carlo simulation is generally suitable for addressing questions related to sampling distribution, especially when (a) the theoretical assumptions of the statistical theory are violated; (b) the theory about the statistic of interest is weak; or (c) no theory exists about the statistic of interest (9). The latter is the case in this study (i.e., the precision of O-D matrices given different sampling rates).

The Monte Carlo experiment reported in this paper focuses on commuting (i.e., work- and school-related) trips made in Belgium. The 2001 census data will be used for the experiment. In particular, the census queried information about the departure and arrival times and locations of work and school trips (when applicable) for all 10,296,350 residents. For different sampling rates, ranging from one (the full population) to a million, 2,000 random stratified samples were drawn (2,000 for each sampling rate). Note that this number is common in transportation-oriented simulation experiments [see, for example, Patel and Thompson (10) and Awasthi et al. (11)]. To ensure that the persons in the samples were geographically distributed, the sample was stratified by geographical area: three nested stratification levels—province, district, and municipality—were taken into account. The sample was proportionately allocated to the strata. In other words, the sample in each stratum was selected with the same probabilities of selection (12).

For each sample, the proportion of persons making commuting trips was calculated, and three corresponding (morning commute) O-D matrices were composed: one O-D matrix on the municipality level (589×589), one O-D matrix on the district level (43×43), and one on the provincial level (11×11). A side note has to be made for the latter O-D matrix: actually there are only 10 provinces in Belgium, but the Brussels metropolitan capital area (accounting for about one-tenth of the entire population) was treated as a separate province. The correspondence of the sample proportion and sample O-D matrices with the population (census) proportion and O-D matrices was then tabulated.

The correspondence between the sample and the population is assessed by using the mean absolute percentage error (MAPE) and

an accommodated version of the MAPE. The MAPE is the mean of the absolute percentage error (APE) and is calculated by

$$\text{MAPE}_{ij} = \frac{\sum_i \sum_j \text{APE}_{ij}}{N} \quad \text{APE}_{ij} = \left| \frac{A_{ij} - E_{ij}}{A_{ij}} \right| \times 100$$

where

A_{ij} = population count for the morning commute from origin i to destination j ,

E_{ij} = sample count (scaled up to population level) for this morning commute, and

N = total number of O-D cells.

Despite its widespread use, the MAPE has several disadvantages. Armstrong and Collopy (13) for instance, argued that the MAPE is bounded on the low side by an error of 100% (O-D counts are all positive integers), but there is no bound on the high side. In response to this comment, Makridakis (14) proposed a modified MAPE (MDAPE), which is often referred to as SAPE (smoothed absolute percentage error) or SMAPE (symmetric mean absolute percentage error). This modified MAPE (MDAPE) is given by

$$\text{MDAPE}_{ij} = \frac{\sum_i \sum_j \text{MDAPE}_{ij}}{N}$$

where

$$\text{MDAPE}_{ij} = \left| \frac{A_{ij} - E_{ij}}{\frac{(A_{ij} + E_{ij})}{2}} \right| \times 100$$

Although this modification accommodates the above described problem, it treats large positive and negative errors very differently (15). Therefore, in this paper, a new modification of the MAPE is proposed, named the mean censored absolute percentage error (MCAPE). This new statistic takes into account the above described comments by limiting the positive values to a maximum of 100. Mathematically, the MCAPE is given by the following formula:

$$\text{MCAPE}_{ij} = \frac{\sum_i \sum_j \text{CAPE}_{ij}}{N}$$

where

$$\text{CAPE}_{ij} = \min \left\{ 100, \left| \frac{A_{ij} - E_{ij}}{A_{ij}} \right| \times 100 \right\}$$

When A_{ij} in the above formulas would be equal to zero, the different criteria would be undefined. This has been remedied by equalizing the APE_{ij} , MDAPE_{ij} , and CAPE_{ij} to zero in these occasions. After all, when the true population count equals zero (no person in the full population corresponds to the considered O-D pair) the up-scaled sample count also equals zero, and thus the true zero is correctly estimated.

The correspondence between the sample proportion (p) of persons making commuting trips and population proportion (π) is calculated by simply calculating the APE:

$$\text{APE} = \left| \frac{\pi - p}{\pi} \right| \times 100$$

No accommodation of this APE was required, as the population proportion (π) was equal to 62.59%, and consequentially the APE could not exceed 100.

To recapitulate, for each sampling rate, 2,000 MAPE and MCAPE values are calculated for the O-D matrix on the municipality level, for the O-D matrix on the district level, and for the O-D matrix on the provincial level. In addition 2,000 APE values are computed for the commuting proportion. For each of these sets of 2,000 values, the 2.5th, 5th, 95th, and 97.5th percentiles were calculated. The k th percentile is that value x , such that the probability that an observation drawn at random from the population is smaller than x , equals k percent (16). The 2.5th and 97.5th percentiles are used to construct the 95th percentile interval, which will be illustrated graphically as lower and upper bounds for the median. The 5th and 95th percentiles will be displayed in the corresponding tables because one is most often only interested in the one-sided alternative. In addition, the median (the 50th percentile) and the arithmetic mean are also computed.

To guarantee that the Monte Carlo experiment is estimating the precision of the O-D matrices in function of different sample rates, rather than in function of other (unobserved) effects, one could take a look at the different sources of errors and biases in surveys. Groves (17) distinguished different sources of inaccuracy in surveys, of which an overview is given in Figure 1. Because in this experiment the true population values are known, and samples are drawn under ideal circumstances (no response bias, no selection bias, no observation errors, no nonresponse, and perfect coverage), the resulting variations in the experiment are only a consequence of the sampling variance (indicated with a gray box, framed with a thick black line in Figure 1). Thus, as intended, the relationship between different sample sizes and the precision and accuracy (sample variance) of the quantities under study are investigated.

RESULTS

Proportion of Commuting Population

Before elaboration on the quality of O-D matrices, an assessment of the appropriateness of travel surveys for deriving traditional indices—such as the mean number of trips made or the mean number of activities performed by individuals or households, or the proportion of the population making work- and school-related trips—is made. For traditional indices such as the mean number of trips made—activities performed by individuals or households, classical sample size calculations can be used to determine optimal sample sizes. Cools et al. (18), for instance, calculated the required number of households for a household activity survey using the following formula:

$$n \geq \frac{z^2 p(1-p)}{md^2}$$

where

n = sample size,

p = sample (survey) proportion,

md = maximal deviation, and

z = z -value of desired confidence interval.

For the “safest” case (i.e., $p = 0.5$), a maximal deviation of 2% and a confidence level of 95% would require a minimum of at least 2,401 households. This example illustrates that for aggregate indices, such as the proportion of the commuting population, a clear theory exists and Monte Carlo simulation is not per se required. Notwithstanding,

an investigation of the relationship between sampling rates and precision (sample variance) is still valuable, and especially contributes to the literature when the focus is turned to the different percentiles that are examined.

Results from the Monte Carlo experiment for the proportion of commuters in the population are graphically displayed in Figure 2 and numerically represented in Table 1. Figure 2 shows a clear relationship between the APE and the sampling proportion. As expected, the additional improvement in precision decreases as the sampling rate increases: for instance the increase in precision (decrease in APE) from a sampling rate of one-millionth (Base-10 logarithm of the sampling proportion = -6) to one-hundred-thousandth (Base-10 logarithm = -5) is considerably larger than the increase in precision from a sampling rate of 1,000 to 100. This is especially so for the upper bound of the 95% percentile interval (97.5th percentile).

The results also show that when the full population is sampled, an absolute precision is obtained (absence of all variation). By definition this result should be obtained. When an average deviation of 5% is considered acceptable, a sample rate between one-hundred- and two-hundred-thousandths is required (5% lies between the mean values 4.293 and 6.083). On the other hand, from the median value one could conclude that in 50% of the cases the maximal deviation (APE) is smaller than 5.192%. A more cautious approach entails the use of the 95th percentiles. If the APE were allowed to exceed 2 in only 5% of the cases, then a sampling rate of about 5 ten-thousandths would be required, which roughly corresponds to sampling 5,000 persons.

Precision of O-D Matrices

In this part of the result section, an assessment of the appropriateness of household activity travel surveys for deriving O-D matrices is made. Recall that a Monte Carlo simulation is particularly suitable for addressing the questions concerning the distribution of the precision of these O-D matrices, as no real theoretical background of this distribution exists. First, attention will be paid to O-D matrices at the municipality level. Afterward, the focus is on O-D matrices at district and provincial levels.

O-D Matrices at Municipality Level

Before expanding on the results of the Monte Carlo experiment, it is important to mention that the true O-D matrix (O-D matrix composed from the full population) is a very large and sparse matrix: of the 346,921 O-D pairs (589×589), 77.8% are zero-cells. As zero-cells in the full population are by definition correctly predicted by taking a sample from this population, the actual overall precision is significantly boosted by the sparseness of the true O-D matrix. Therefore, the decision was made to present the results based on the 76,882 non-zero-cells. To derive the values that include the zero-cells, one only needs to divide the MAPE and MCAPE values by 4.512 [= all cells/(all cells—zero-cells)].

Inspection of Table 2 immediately reveals that no accurate O-D matrices are obtained at the municipality level, even if zero-cells are taken into account: a survey that would query half of the population still would have an average APE of 11.99% when zero-cells are included and correspondingly of 54.11% when only the actual predictions (non-zero-cells) are taken into account. This clearly indicates that the direct derivation of O-D matrices from household activity travel surveys should be avoided. Notwithstanding, O-D matrices derived from household activity travel surveys are very

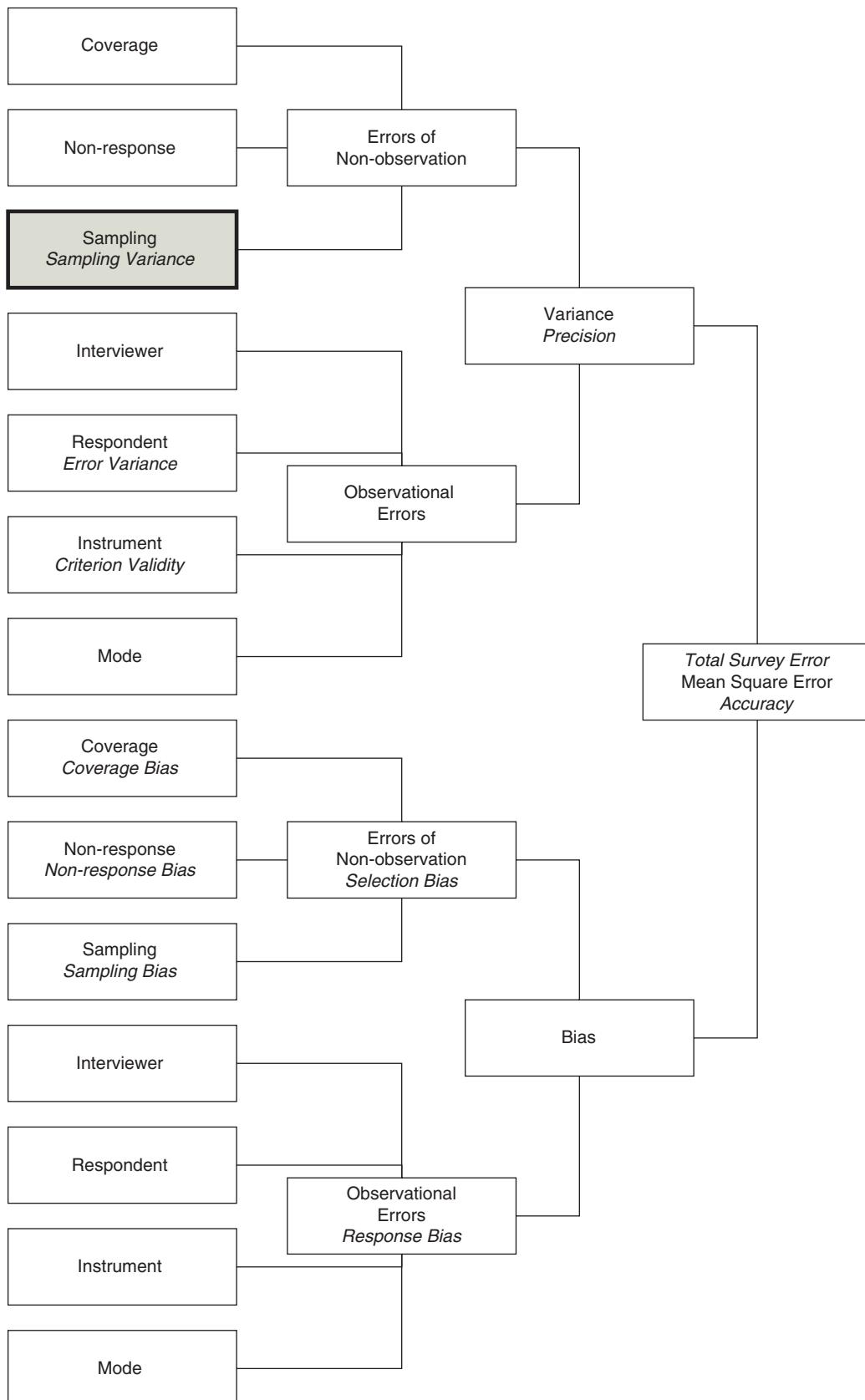


FIGURE 1 Potential sources of total survey error.

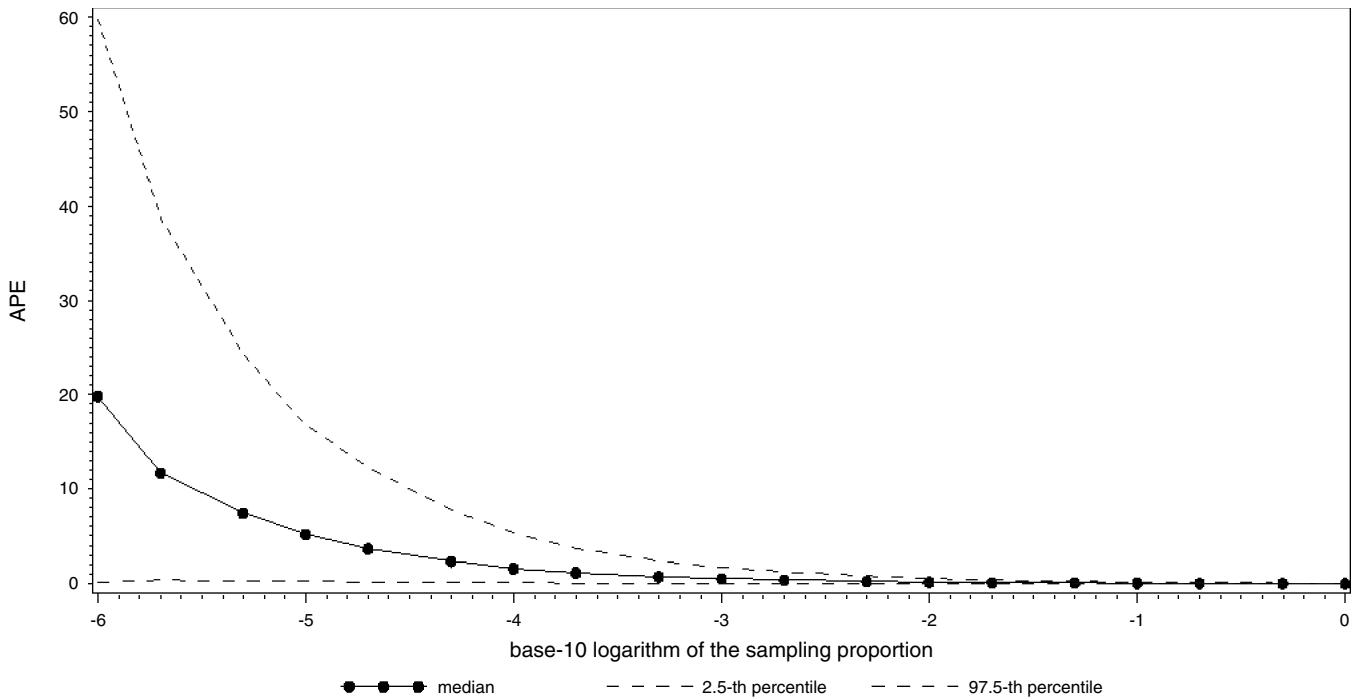


FIGURE 2 Relationship between APE and sampling rate for commuting proportion.

valuable: even a simple gravity model with the inverse squared distance as deterrence function, taking into account the productions and attractions derived from the surveys, already results in a clear improvement of the O-D matrices. This is certainly a plea for travel demand models that incorporate the behavioral underpinnings of destination choices (activity location choices) given a certain origin,

TABLE 1 APE Statistics for Commuting Proportion Given Different Sampling Rates

Sampling Rate (SR)	\log_{10} SR	Mean	P5	Median	P95
0.000001	-6.00	20.270	1.670	19.826	46.744
0.000002	-5.70	13.915	1.096	11.684	34.377
0.000005	-5.30	8.642	0.659	7.412	21.509
0.000010	-5.00	6.083	0.474	5.192	14.849
0.000020	-4.70	4.293	0.343	3.671	10.618
0.000050	-4.30	2.741	0.213	2.317	6.793
0.000100	-4.00	1.879	0.137	1.574	4.586
0.000200	-3.70	1.330	0.097	1.140	3.219
0.000500	-3.30	0.853	0.072	0.723	2.097
0.001000	-3.00	0.602	0.052	0.508	1.485
0.002000	-2.70	0.430	0.040	0.362	1.054
0.005000	-2.30	0.269	0.022	0.227	0.659
0.010000	-2.00	0.190	0.015	0.160	0.462
0.020000	-1.70	0.129	0.010	0.109	0.313
0.050000	-1.30	0.080	0.007	0.067	0.197
0.100000	-1.00	0.053	0.004	0.045	0.132
0.200000	-0.70	0.034	0.003	0.029	0.082
0.500000	-0.30	0.016	0.001	0.013	0.038
1.000000	0.00	0.000	0.000	0.000	0.000

NOTE: P = percentile; e.g., P5 stands for 5th percentile.

like, for instance, models that make use of space–time prisms [e.g., Pendyala et al. (19)], and models that combine data from different sources, such as data integration tools [e.g., Nakamya et al. (20)]. In addition, O-D matrices derived from travel surveys form a good basis for O-D matrices derived from traffic counts: as multiple O-D matrices can be derived from the same set of traffic counts, O-D matrices derived from travel surveys provide a good basis for constraining the matrices derived from traffic counts (27). A thorough look at Table 2 also reveals that when half the population is sampled, the values for the MAPE and MCAPE are the same. This can be explained by the fact that when half of the population is used, none of the 2,000 samples has a MAPE higher than 1.

When the general tendency of the precision of the O-D matrices derived from travel surveys is discussed, Figures 3 and 4 provide a clear insight into the relationship between the precision and the sampling rate. From Figure 3 one can clearly see that the median MAPE first increases when samples are becoming larger and then starts to decrease. The increase in median MAPE for the smallest sampling rates can be accounted for by the fact that on average more cells are seriously overestimated, whereas the maximum underestimations are bounded by 100%. This effect is filtered out by using the MCAPE, as can be seen from Figure 4; a clear decreasing relationship is visible here. Next to the difference in relationships between the MAPE and MCAPE, one could also observe a clear difference between the percentile interval for the MAPE and the percentile interval for the MCAPE. By condensing the APE to a maximum of 1 (i.e., the CAPE), almost all variability around the median value is filtered out: the 2.5th and 97.5th percentiles almost coincide with the median values in case of the MCAPE.

When this decreasing pattern of the MCAPE (Figure 4) is compared with the one of the proportions (Figure 2), a clear contrast in the tendency can be seen: while the pattern for proportion is a convex decreasing function, for the O-D matrices this is a concave

TABLE 2 MAPE and MCAPE for O-D Matrices Derived at Municipality Level

Sampling Rate (SR)	\log_{10} SR	MAPE				MCAPE			
		Mean	P5	Median	P95	Mean	P5	Median	P95
0.000001	-6.00	205.996	100.902	115.559	753.493	100.000	100.000	100.000	100.000
0.000002	-5.70	200.992	102.528	129.575	754.266	100.000	100.000	100.000	100.000
0.000005	-5.30	199.861	109.279	155.540	431.119	99.999	99.998	99.999	100.000
0.000010	-5.00	199.089	116.768	175.588	352.896	99.997	99.995	99.997	99.999
0.000020	-4.70	198.647	127.682	187.707	304.125	99.994	99.992	99.994	99.996
0.000050	-4.30	198.338	148.261	194.723	259.440	99.977	99.972	99.977	99.981
0.000100	-4.00	199.452	160.661	198.325	243.313	99.927	99.919	99.927	99.936
0.000200	-3.70	197.459	170.778	196.746	226.279	99.784	99.769	99.784	99.798
0.000500	-3.30	195.930	178.611	195.344	215.156	99.414	99.391	99.414	99.437
0.001000	-3.00	193.624	181.284	193.511	206.508	98.999	98.973	98.999	99.026
0.002000	-2.70	190.182	181.664	189.982	199.411	98.393	98.360	98.392	98.427
0.005000	-2.30	183.425	177.916	183.465	188.907	97.033	96.990	97.033	97.077
0.010000	-2.00	175.654	171.907	175.659	179.551	95.352	95.298	95.353	95.407
0.020000	-1.70	164.993	162.373	165.044	167.739	92.797	92.730	92.798	92.865
0.050000	-1.30	145.263	143.745	145.271	146.836	87.514	87.424	87.515	87.598
0.100000	-1.00	124.970	124.078	124.960	125.866	81.369	81.273	81.369	81.469
0.200000	-0.70	99.172	98.724	99.169	99.631	72.293	72.193	72.294	72.392
0.500000	-0.30	54.108	54.089	54.108	54.128	54.108	54.089	54.108	54.128
1.000000	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

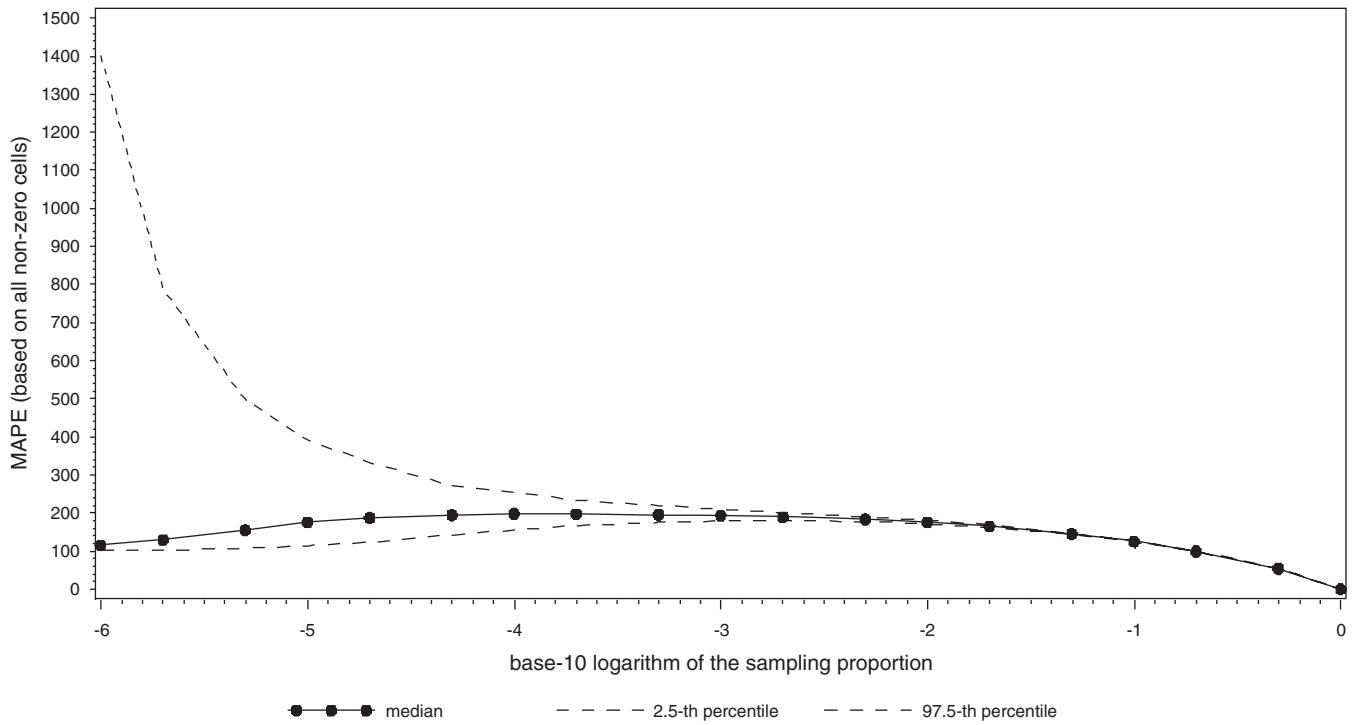


FIGURE 3 Relationship at municipality level between MAPE and Base-10 logarithm of sampling rate.

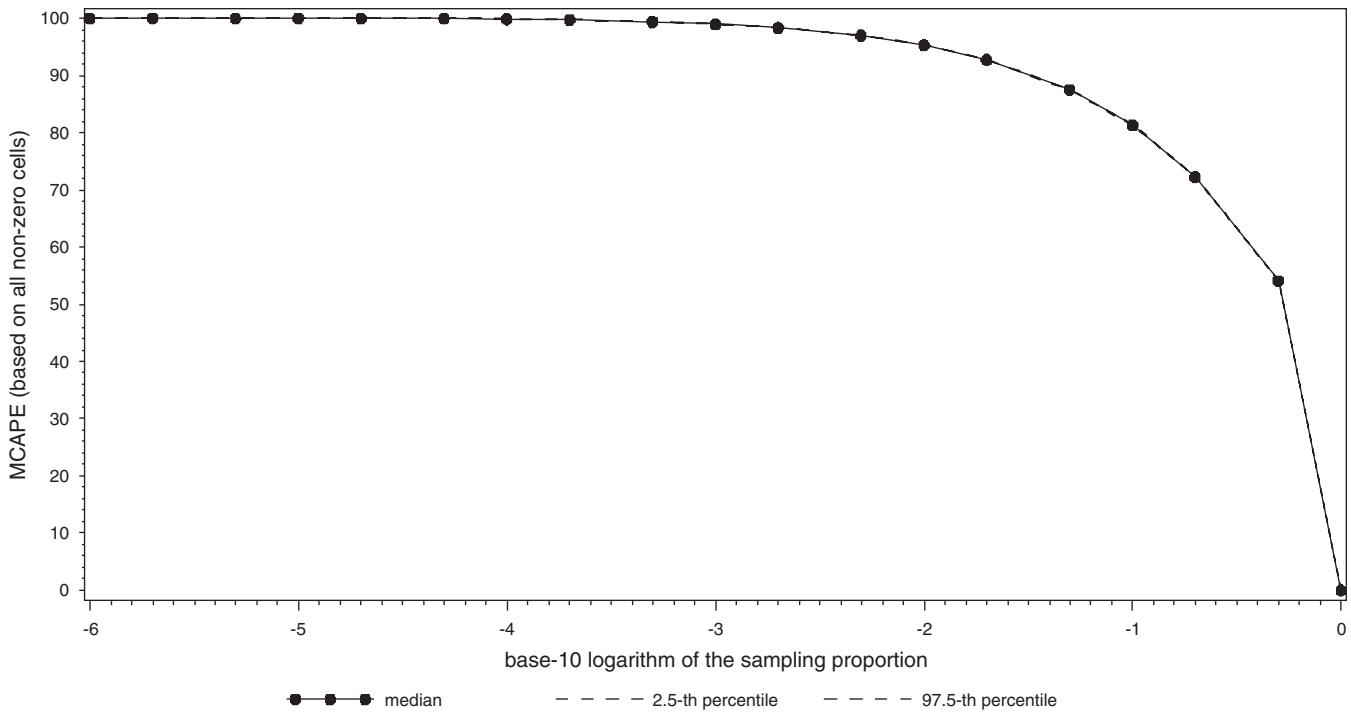


FIGURE 4 Relationship at municipality level between MCAPE and Base-10 logarithm of sampling rate.

decreasing function. This difference in pattern, as well as the difference in precision, can be explained by the fact that proportions are aggregate indices, and that surveys are extremely suitable for capturing these aggregate figures, while in O-D matrices all individual information is used.

O-D Matrices at District Level

Similar to the true O-D matrix at the municipality level, the true O-D matrix at the district level (43×43) comprises a nonnegligible amount of zero-cells. Nonetheless, the number of non-zero-cells is considerably smaller: 10.4% of the 1,849 O-D pairs are zero-cells. Recall that zero-cells in the full population are by definition correctly predicted by taking a sample from this population. Therefore, similar to the previous paragraph, the results are based on the 1,657 non-zero-cells. The values that include the zero-cells can be calculated by dividing the MAPE and MCAPE values by 1.116.

A thorough look at Table 3 shows that also at the district level no accurate O-D matrices can be derived. Even if zero-cells are included in the calculations, surveying half of the population would result in an average APE of 22.25% (24.832 divided by 1.116). When compared with the results of O-D matrices derived at the municipality level, the results including the zero-cells are worse at the district level (an average MAPE of 22.25% versus one of 11.99%). This is because at the municipality level a much larger share (77.8% versus 10.4%) of zero-cells is automatically correctly predicted. In contrast, when the results of only the non-zero-cells are compared, the precision of the O-D matrices derived at the district level is higher than the precision of the O-D matrices derived at the municipality level. This confirms that predictions on a more aggregate level are more precise.

A visual representation of the relationship between the precision of the O-D matrices derived at the district level and the sampling

rate is provided in Figures 5 and 6. Inspection of Figure 5 reveals a pattern very similar to the one observed in Figure 3: the MAPE first increases when samples are becoming larger, and then starts to decrease. Recall that the increase in median MAPE for the smallest sampling rates can be accounted for by the fact that more cells are seriously overestimated on average, while the maximum underestimations are bounded by 100%. By analogy with the results at the municipality level, this effect is filtered out by using the MCAPE, as could be noticed from Figure 6. Moreover, the relationship between the MCAPE and sampling proportion is a concave decreasing function, similar to the relationship between the MCAPE and sampling rate at the municipality level.

O-D Matrices at Provincial Level

In contrast to the true O-D matrices at the municipality and district levels, the true O-D matrix at the provincial level (11×11) only comprises non-zero-cells. Examination of Table 4 reveals that at the provincial level, hardly any accurate O-D matrices can be derived. Nonetheless, in contrast to the results at the municipality and district levels, for the largest sample sizes acceptable results are obtained: sampling half of the population would result in an average APE of 3.4%, and surveying one-fifth of the population results in an average APE of 7.4%. Results from Table 4 also confirm that predictions related with a more aggregate level are more precise. Notwithstanding, results at the provincial level confirm the finding unraveled at the lower levels (municipality and district)—that the direct derivation of O-D matrices from household activity travel surveys should be avoided.

The visualization of the relationship between the precision of the O-D matrices derived at the provincial level and the sampling proportion is shown in Figures 7 and 8. Analogous to the relationships between the sample rate and the precision of the O-D matrices at the municipality and district levels, the MAPE first increases when

TABLE 3 MAPE and MCAPE for O-D Matrices Derived at District Level

Sampling Rate (SR)	\log_{10} SR	MAPE				MCAPE			
		Mean	P5	Median	P95	Mean	P5	Median	P95
0.000001	-6.00	171.196	100.978	110.178	299.050	99.996	99.987	99.993	100.000
0.000002	-5.70	206.097	101.875	113.822	385.419	99.937	99.869	99.937	100.000
0.000005	-5.30	197.675	104.054	121.673	442.676	99.716	99.589	99.717	99.841
0.000010	-5.00	200.483	105.752	127.559	481.848	99.352	99.174	99.351	99.535
0.000020	-4.70	191.212	108.363	136.036	417.340	98.750	98.524	98.749	98.970
0.000050	-4.30	186.436	111.940	145.249	383.240	97.578	97.331	97.572	97.860
0.000100	-4.00	187.614	114.861	152.048	359.850	96.444	96.145	96.449	96.738
0.000200	-3.70	177.715	118.461	156.391	307.348	94.788	94.416	94.791	95.131
0.000500	-3.30	172.565	122.683	160.022	274.422	92.023	91.596	92.027	92.443
0.001000	-3.00	164.271	124.944	157.032	230.024	89.557	89.113	89.557	90.004
0.002000	-2.70	154.307	123.783	150.441	196.883	86.232	85.711	86.228	86.753
0.005000	-2.30	138.952	118.396	137.611	164.985	81.028	80.476	81.014	81.605
0.010000	-2.00	124.538	110.153	123.828	141.608	75.874	75.224	75.874	76.529
0.020000	-1.70	109.048	99.198	108.610	120.450	69.540	68.823	69.548	70.242
0.050000	-1.30	85.890	80.198	85.743	92.146	59.623	58.825	59.617	60.407
0.100000	-1.00	67.276	63.761	67.238	70.863	50.649	49.818	50.645	51.482
0.200000	-0.70	48.653	46.789	48.640	50.602	40.040	39.253	40.042	40.890
0.500000	-0.30	24.832	24.110	24.826	25.529	24.832	24.110	24.826	25.529
1.000000	0.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

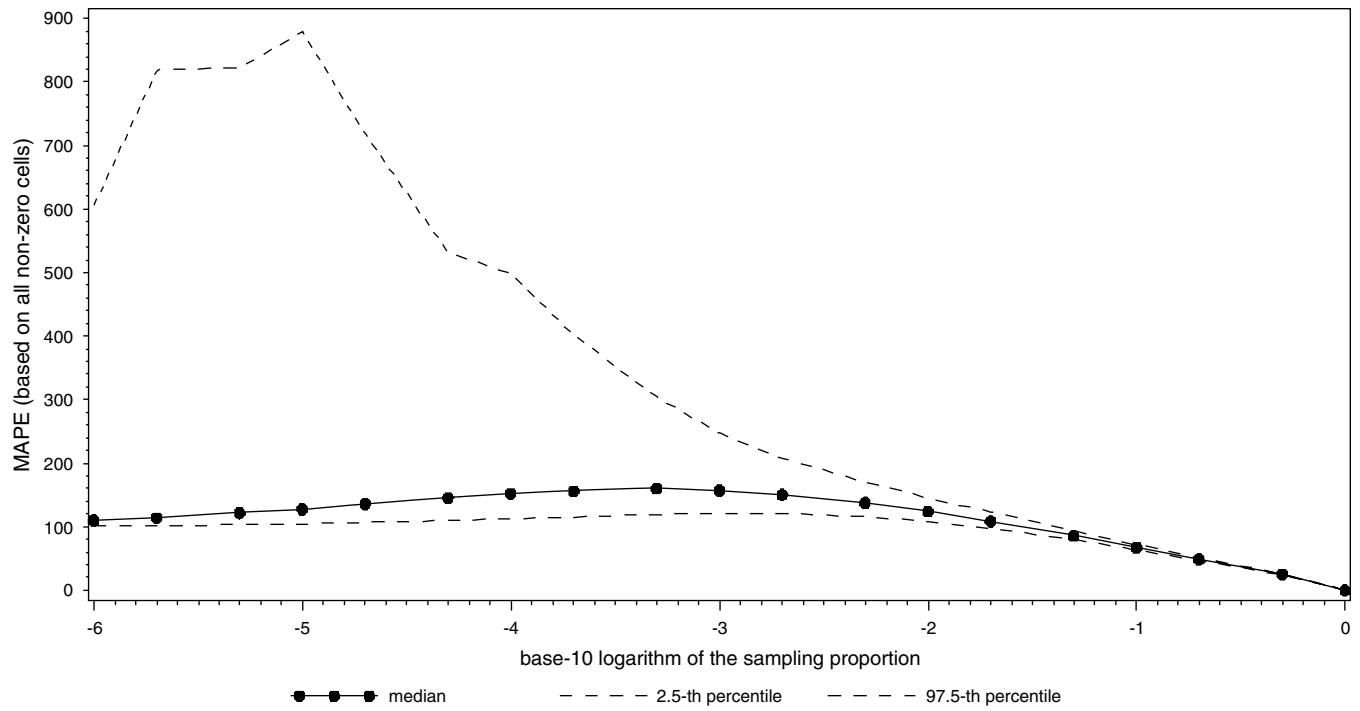


FIGURE 5 Relationship at district level between MAPE and Base-10 logarithm of sampling rate.

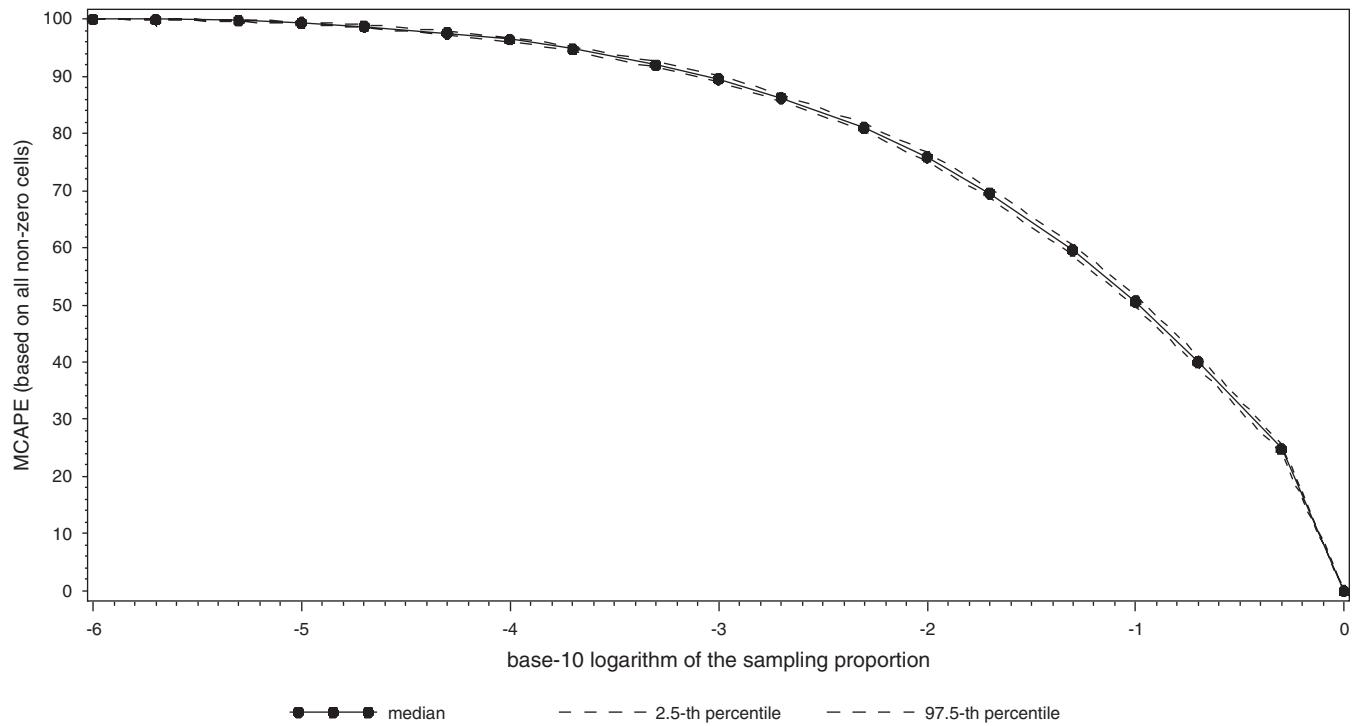


FIGURE 6 Relationship at district level between MCAPE and Base-10 logarithm of sampling rate.

TABLE 4 MAPE and MCAPE for O-D Matrices Derived at Provincial Level

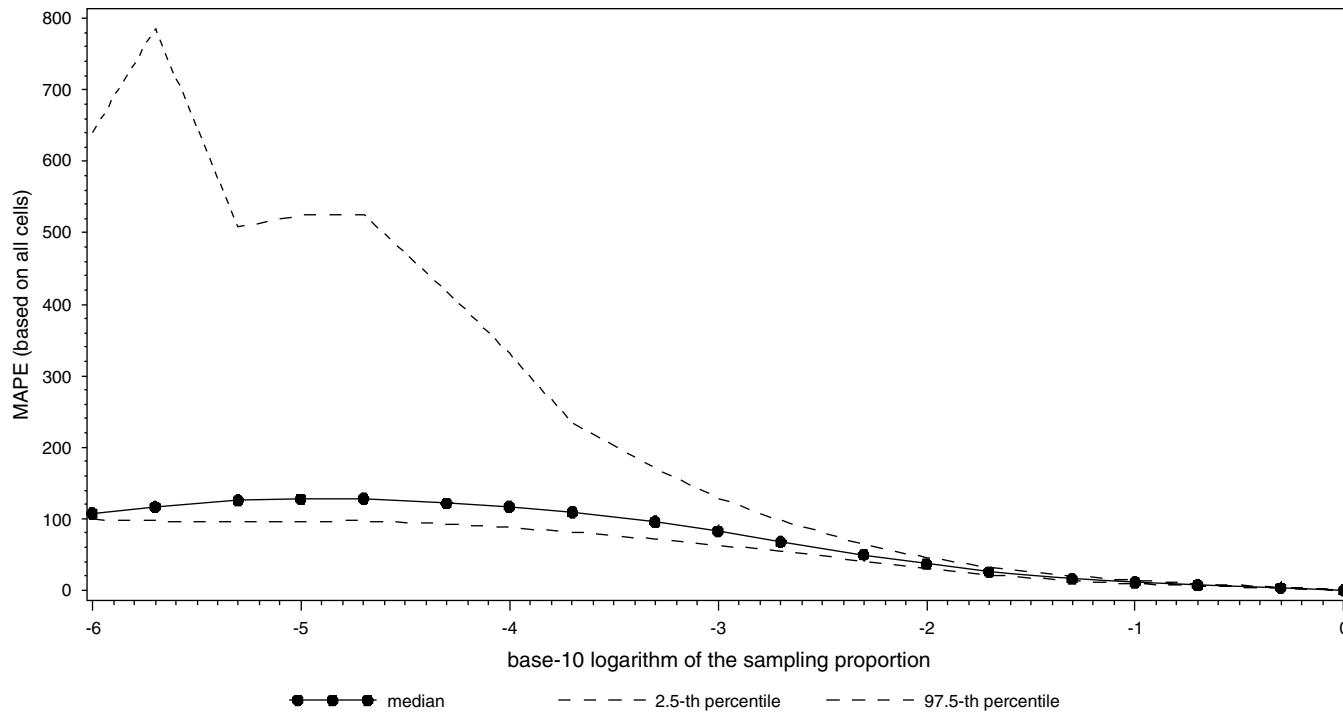


FIGURE 7 Relationship at provincial level between MAPE and Base-10 logarithm of sampling rate.

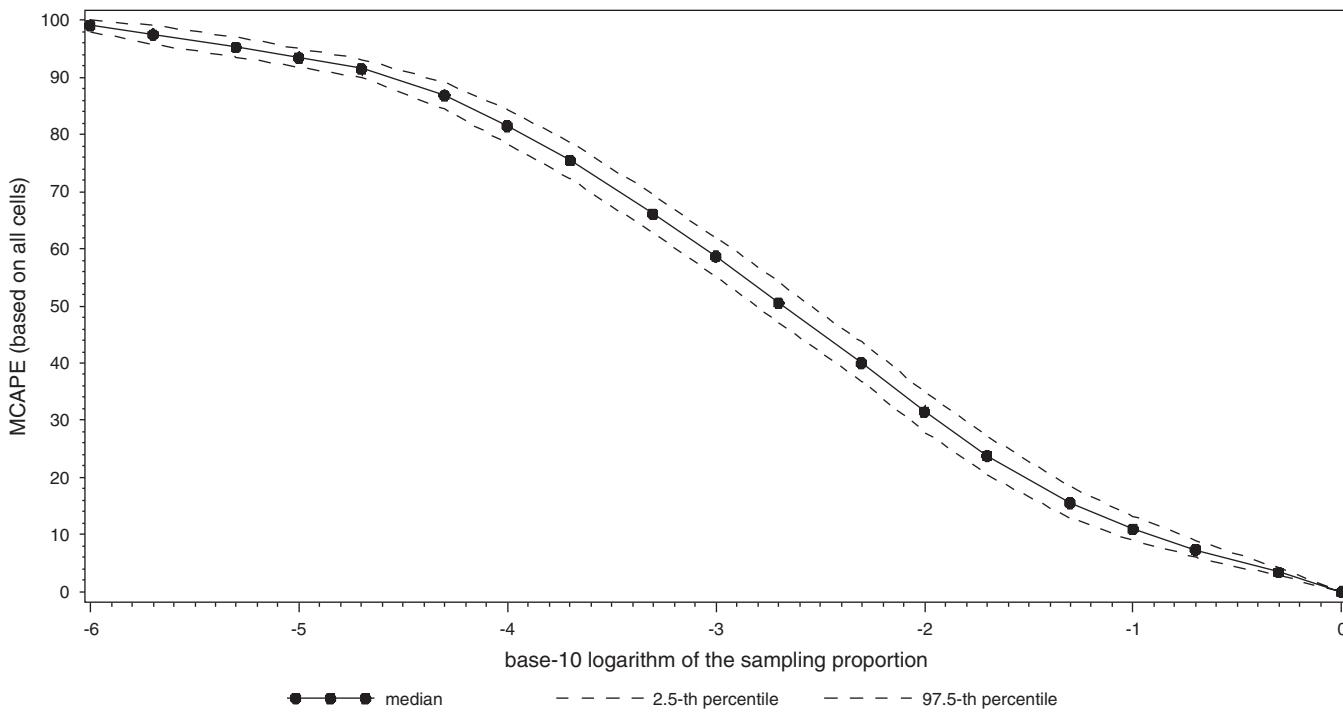


FIGURE 8 Relationship at provincial level between MCAPE and Base-10 logarithm of sampling rate.

samples are becoming larger and then starts to decrease (Figure 7). Again, the use of the MCAPE filters out this effect. In contrast to the results at the municipality and district levels, the relationship between the precision and the MCAPE reveals an s-shaped decreasing function: for the smallest sampling rates the relationship is concavely decreasing, similar to the O-D matrices at the municipality and district levels; but for the larger sampling rates the relationship is a convex decreasing function. Moreover, the 95th percentile interval is much wider than for the O-D matrices at less aggregated levels. The most important reasons for this are the degree of sparseness and size of the matrix: for the less aggregate levels (municipality and district levels), a lot of the variability of the precision is taken away by the large amounts of (zero-)cells.

CONCLUSIONS

In this paper, an assessment of the quality of O-D matrices derived from household activity travel surveys was made. The results showed that no accurate O-D matrices can be directly derived from these surveys. Only when half of the population is queried is an acceptable O-D matrix obtained at the provincial level. Therefore, use of additional information to better grasp the behavioral realism underlying destination choices is recommended. This is certainly a plea for travel demand models that incorporate the behavioral underpinnings of destination choices (activity location choices) given a certain origin. Moreover, matrix calibration techniques could seriously improve the quality of the matrices derived from these household activity travel surveys. In addition, collection of information about particular O-D pairs by means of vehicle intercept surveys rather than household activity travel surveys is recommended, as vehicle intercept surveys are tailored for collecting specific O-D data. Note that the results presented in this paper do not negate the value of travel surveys as was shown in the example of deriving the commuting population, but indicate that sophistication is needed in the manner in which the data are employed.

A second important finding in this paper is that traditional methods to assess the comparability of two O-D matrices could be enhanced: the MCAPE index that was proposed has clear advantages over the traditional indices, the most important being that the MCAPE filters out the noise created by the asymmetry of the traditional criteria. Therefore, when dissimilarities between different O-D matrices are investigated, use of the MCAPE index next to traditional criteria is highly recommended.

An important avenue for further research is the investigation of the relationship between the variability in the outcomes of travel demand models and underlying survey data. Triangulation of both travel demand modeling and small area estimation models could prove to be a pathway for success. An empirical investigation of the effect of sampling proportions in household activity travel surveys on final model outcomes would further illuminate the quest for optimal sample sizes. A thorough examination of the minimum required sampling rate of a household travel survey such that trip distribution models (e.g., a gravity model) could help fill in the full trip table certainly is an important step in further analyses. Model complexity and computability will certainly be key challenges in this pursuit.

ACKNOWLEDGMENT

The authors thank Katrien Declerq for her advice on the implementation of the experiment.

REFERENCES

- Haustein, S., and M. Hunecke. Reduced Use of Environmentally Friendly Modes of Transportation Caused by Perceived Mobility Necessities: An Extension of the Theory of Planned Behavior. *Journal of Applied Social Psychology*, Vol. 37, No. 8, 2007, pp. 1856–1883.
- Steg, L. Can Public Transport Compete with the Private Car? *IATSS Research*, Vol. 27, No. 2, 2003, pp. 27–35.
- TRB Committee on Travel Survey Methods. *The Online Travel Survey Manual: A Dynamic Document for Transportation Professionals*. <http://trbtsm.wiki.zoho.com>. Accessed July 22, 2009.
- Stopher, P., R. Alsnih, C. Wilmot, C. Stecher, J. Pratt, J. Zmud, W. Mix, M. Freedman, K. Axhausen, M. Lee-Gosselin, A. Pisarski, and W. Brög. *NCHRP Report 571: Standardized Procedures for Personal Travel Surveys*. Transportation Research Board of the National Academies, Washington, D.C., 2008.
- Travel Survey Manual*. Cambridge Systematics, Washington, D.C., 1996.
- Tourangeau, R., M. Zimowski, and R. Ghadially. *An Introduction to Panel Surveys in Transportation Studies*. Report DOT-T-84. FHWA, U.S. Department of Transportation, 1997.
- Stopher, P. R., and S. P. Greaves. Household Travel Surveys: Where Are We Going? *Transportation Research Part A: Policy and Practice*, Vol. 41, No. 5, 2007, pp. 33–40.
- Rubinstein, R. Y. *Simulation and the Monte Carlo Method*. John Wiley and Sons, Inc., New York, 1981.
- Fan, X., A. Felsővályi, S. A. Sivo, and S. C. Keenan. *SAS® for Monte Carlo Studies: A Guide for Quantitative Researchers*. SAS Institute, Cary, N.C., 2000.
- Patel, A., and M. Thompson. Consideration and Characterization of Pavement Construction Variability. In *Transportation Research Record 1632*, TRB, National Research Council, Washington, D.C., 1998, pp. 40–50.
- Awasthi, A., S. S. Chauhan, S. K. Goyal, and J.-M. Proth. Supplier Selection Problem for a Single Manufacturing Unit Under Stochastic Demand. *International Journal of Production Economics*, Vol. 117, No. 1, 2009, pp. 229–233.
- Groves, R. M., F. J. Fowler, M. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*. John Wiley and Sons, Inc., Hoboken, N.J., 2004.
- Armstrong, J. S., and F. Collopy. Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting*, Vol. 8, No. 1, 1992, pp. 69–80.
- Makridakis, S. Accuracy Measures: Theoretical and Practical Concerns. *International Journal of Forecasting*, Vol. 9, No. 4, 1993, pp. 527–529.
- Goodwin, P., and R. Lawton. On the Asymmetry of the Symmetric MAPE. *International Journal of Forecasting*, Vol. 15, No. 4., 1999, pp. 405–408.
- Good, P. I. *Resampling Methods: A Practical Guide to Data Analysis, 3rd ed.* Birkhäuser, Boston, Mass., 2006.
- Groves, R. M. *Survey Errors and Survey Costs*. John Wiley and Sons, Inc., Hoboken, N.J., 1989.
- Cools, M., E. Moons, T. Bellemans, D. Janssens, and G. Wets. Surveying Activity–Travel Behavior in Flanders: Assessing the Impact of the Survey Design. In *Proceedings of the BIVEC–GIBET Transport Research Day 2009, Part II* (C. Macharis and L. Turckin, eds.). VUB-PRESS, Brussels, 2009, pp. 727–741.
- Pendyala, R. M., T. Yamamoto, and R. Kitamura. On the Formulation of Time–Space Prism to Model Constraints on Personal Activity–Travel Engagement. *Transportation*, Vol. 29, No. 1, 2002, pp. 73–94.
- Nakanya, J., E. A. Moons, S. Koelet, and G. Wets. Impact of Data Integration on Some Important Travel Behavior Indicators. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1993*, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 89–94.
- Abrahamsson, T. *Estimation of Origin–Destination Matrices Using Traffic Counts: A Literature Survey*. IIASA Interim Report IR-98-021. International Institute for Applied Systems Analysis, Laxenburg, Austria, May 1998.

The Travel Survey Methods Committee peer-reviewed this paper.

CHALMERS



Obtaining Origin/Destination-matrices from cellular network data

Master's Thesis in Engineering Mathematics

Erik Mellegård

Department of Mathematical Sciences
Division of Mathematics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2011
Master's Thesis 2011:

Abstract

Mobile network operators are collecting a lot of data on people's calls, like who they call, for how long they call etc., that are stored for billing and network purposes. Among that data, there is also information on which base station they are connected to, something that could be used to obtain valuable information about people's movements. The main reason that this is not being done today is that the operators are afraid of what would happen if someone would mistreat the data and use it to track people. This thesis presents a method for obtaining Origin/Destination-matrices from the mobile network data in a way that keeps the individuals' privacy. Since the operators are reluctant to let us use any real data, the method has been applied to synthetic data and some call data records. The method is applied to each individual separately, which makes it possible to run it in parallel on a cluster of computers, something that is very important when having months of data on million of users. Running the algorithm on call data records, this thesis shows that it is feasible to obtain Origin/Destination-matrices from mobile network data.

Acknowledgements

Without the help and support of a number of people, the making of this thesis would not have been possible. Most thanks I owe to my supervisor at Ericsson, **Simon Moritz**, who has given me a lot of help and feedback. Other people at Ericsson supporting me include **Apostolos Georgakis**, **Muhamed Zahoor** and **Marika Stålnacke**. I would also like to thank the members of the Consider8 project working at SICS; **Olof Görnerup**, **Pedro Sanches** and **Martin Nordström**.

Working at Ericsson would not have been as fun as it was without you, **Jonathan Magnusson**, **Viktor Kärnstrand** and **Abhiroop Gupta**.

Erik Mellegård
Göteborg
October 4, 2011

Contents

1	Introduction	1
1.1	Traffic link counts	2
1.2	Surveys	2
2	Background	3
2.1	Data	3
2.1.1	Available data	3
2.2	Mobile networks	4
2.3	Reverse geocoding APIs	4
3	Related work	6
3.1	Simulations	6
3.2	Using real data	7
3.3	Other related work	7
4	Use cases	8
4.1	Building new metro and mono rail in India	8
4.2	Traffic planning	8
4.3	Bridges and overpasses	9
4.4	Unprotected railroad crossings	9
5	Methodology	10
5.1	Initial study	10
5.2	Method development	10
5.3	Implementation and analysis	10
6	Framework & Results	12
6.1	Framework	12
6.1.1	Convert from cell id	12
6.1.2	Finding trips	13
6.1.3	Compare and adjust stations	15

6.1.4	Group stations using open street map	15
6.1.5	Aggregate	16
6.2	Implementation	16
6.2.1	Architecture	16
6.3	Visualization	17
6.4	Results	17
6.4.1	Finding trips	17
6.4.2	Simulated data	18
6.4.3	Real data	18
7	Discussion	24
7.1	Method	24
7.1.1	Advantages	24
7.1.2	Disadvantages	25
7.1.3	Improvements	25
7.2	Result	25
8	Conclusion and future work	27
	Bibliography	29
A	Mean shift clustering	30
B	Hadoop MapReduce	31
C	Data collection	33

List of Figures

6.1	An overview of the method for obtaining O/D-matrices.	12
6.2	All the cells in Sweden that have been used in this thesis.	14
6.3	Flowchart describing how to find trips made by an individual	15
6.4	An illustration of the algorithm for finding the trips made by an individual	19
6.5	All stations for one user of the Android application and how they were clustered.	20
6.6	All the "places" in Sweden, extracted from Open Street Map [6].	21
6.7	The O/D-matrices displayed in a web browser using Google Earth.	22
6.8	The O/D-matrices displayed in a web browser using Google Maps.	23
6.9	The distribution of trips over the day from the real data set.	23
B.1	An illustration of how hadoop operates on key-value pairs.	32
C.1	An illustration of where the data is collected in the mobile network. . . .	34

1

Introduction

This thesis is part of a project called Consider8. Consider8 is a collaboration between Ericsson AB and Swedish Institute of Computer Science (SICS) that aims at finding human mobility patterns from mobile network data.

When doing all sorts of traffic planning, it is important to have information on people's movements. One of the more fundamental kind of information you can have on people's movements is what is called Origin/Destination-matrices (O/D-matrices). An O/D-matrix is a list of places and how many people travelled from place A to place B (within a given time frame). A simple example of an O/D-matrix can be seen in table 1.1. Even though O/D-matrices seem very simple, they are actually quite useful in many applications. It is also remarkably hard to obtain accurate O/D-matrices that are up-to-date. Today, O/D-matrices are produced in one of two ways, either by surveys or by using measurements of traffic link counts, from cameras and magnetic sensors positioned along the road.

Table 1.1: An example of what an O/D-matrix looks like. From this matrix one can see that three people have travelled between Kista and Solna.

	Kista	Solna	Bromma	Gamla stan	Norrmalm
Kista	-	3	2	9	7
Solna	14	-	12	6	4
Bromma	5	2	-	5	2
Gamla stan	1	2	9	-	7
Norrmalm	4	3	7	9	-

1.1 Traffic link counts

Since the hardware required for measuring traffic counts is relatively inexpensive today, that technique is mostly used today. Usually, a lot of different O/D matrices can reproduce the same traffic counts; therefore the problem of deducing the O/D-matrix from traffic link counts is underdetermined. There are a lot of different techniques for finding the most likely O/D matrix from the link traffic counts, for example; maximum likelihood, generalized least square and Bayesian inference [13]. This way of estimating the O/D-matrix will only give you an estimation on the roads on which you have measuring equipment. Furthermore, the equipment is expensive and requires a lot of maintenance.

1.2 Surveys

Household surveys are conducted from time to time on a country level. People are asked to fill in forms, answering questions about where and how they travel. Household surveys can produce good O/D-estimation matrices if a lot of people participate. The problem is that people might not remember correctly when they fill in their logs and they might have reasons for not being honest when doing so. Going through all these logs are expensive, which means that these surveys are not conducted on a regular basis, resulting in O/D-matrices that are often to old to reflect recent changes in peoples movements.

2

Background

THE MOBILE CARRIERS around the world collect a lot of information on peoples calls as part of their normal operation, like who they call, for how long they call etc. This is collected for billing and network purposes. Among the information the carriers collect, there is information on which base station the phone is connected to during the call.

2.1 Data

When a phone is in active mode, either in a call or sending data, the base station the phone is connected to is logged twice a second in the network carrier systems. When the phone is in idle mode, the information on which base station the phone is connected to is only logged about once an hour (this vary between carriers, but is usually between 20 minutes and 2 hours). Among this data, there is information on the cell id of the base station the phone is connected to and a time stamp. This kind of data is what the method described in this thesis has been developed for.

2.1.1 Available data

The mobile operators are very reluctant to release any data of the kind described above for privacy reasons. However, Ericsson has access to some call data records. In addition to that data, some data has been collected from an Android application and some synthetic data has been generated. The three sets of data are described below.

Call data records

The first set of data is call data records from a mobile network operator. Call data records contains data on all outgoing calls that has been made from a network operator, including time and cell id at the start of every call. The call data records used in this

thesis contains data from two million subscribers during one month's time. Unfortunately the data does⁴ not contain any passive mode information or any information on which base station the phone has been connected to during the call, which in reality would also be consider as input information.

Application data

The second set of data has been collected using an Android application, installed on Ericsson and Swedish Institute of Computer Science (SICS) employees' phones. This application collects information on the base station the phone is connected to, but also the GPS-position. This data can therefore be used to compare the results from the method with what we can manually deduce from the GPS data. The application collects information on which cell the phone is connected to at all times, meaning that this data differs from the data collected in the mobile operators' networks. Therefore, the data has to be modified to look like network data before using it to test and evaluate the method.

Synthetic data

The third set of data is synthetic data created by Viktor Kärnstrand, who worked at Ericsson during this summer. The method generating this data by creating an O/D-matrix from demographic data and then finds the closest route between all points in the O/D-matrix using an A^* -algorithm. From these routes, synthetic network data and GPS data are created. Since there is information on the actual O/D-matrix used to create the network data, this data be used to evaluate methods developed for obtaining O/D-matrices from mobile network data.

2.2 Mobile networks

A mobile network is organized in an hierarchical structure, with what is called base transceiver station (BTS) as the lowest component in the hierarchy (see appendix C.1). The area a BTS covers is usually called a cell. Each cell has a cell id that is unique within a certain location are code (LAC) for a specified operator and country. In this thesis it will be important to know the location of cell from its cell id (and its LAC, country and operator). There are a few APIs online from where you can get the longitude and latitude of a large part of the cells around the world, for example location-api [5] , open cell id [7] and Ericsson Labs [1].

2.3 Reverse geocoding APIs

Reverse geocoding is any process that puts a name or a place to a point (i.e. a longitude and latitude). This includes everything from just identifying the country a certain point belongs to, to identifying what road is closest to that point. There are a few freely

available reverse geocoding APIs available online, for example the Google Geocoding API [3] and Openstreetmap [6]. In this thesis, openstreetmap has been used to find all places with one of the tags: "city", "town", "village" and "suburb" in the areas of interest.

3

Related work

Several projects have used cellular network data to analyze traffic in real time. This is usually achieved using cell handover information from active phones to calculate the current speed on a road network. The purpose of this is to be able to give the drivers information on the current speed and congestions on the route they are traveling. Mobile network data as input to traffic analysis has been used by TomTom [12], in the STRESS project [15] and by AirSage [10] among other. Since their work is not that related to the aim of this thesis, their approach will not be explained in any more detail.

3.1 Simulations

There have not been made that many attempts at analyzing traffic in a more long-term perspective from mobile network data, for example obtaining O/D-matrices. Some simulations and studies have been made on the subject, for example Caceres N., Wideberg J.P. and Benitez F.G [17] concluded from their simulation that it is possible to extract O/D-matrices from their simulated cellular network data. Sohn Keemin and Kim Daehyun [22] concluded the same thing in their paper. Their simulation also showed how the result depends on some network characteristics, such as network penetration (proportion of people using the network in question) and the cell size. They both used handover information to estimate the link traffic volume and then used that to estimate O/D-matrices. One of the drawback with this approach is that it can only use data from phones in active mode (when it is in a call or sending/receiving data). It can also only give an approximate solution, since the problem of finding O/D-matrices from traffic link counts is underdetermined. In their simulation they also assumed that a phone is always connected to the closest base station, which is a simplification which is not true. From the data we have access to we can see that a phone can switch between several base stations even while being stationary. This kind of behavior of the network would make it very hard to obtain O/D-matrices with the approach they used.

3.2 Using real data

Some studies on real data have been made in smaller scale, to determining the feasibility of acquiring O/D-matrices from a cellular network. Ahas Rein, Aasa Anto, Silm Siiri and Tiri Margus [14] published a case-study on 277 persons living outside the city of Tallinn. The study was conducted over a time period of 8 days in 2006. The current cell each of the phones where connected to where recorded once every 15 minutes. With this information they were able to determine where people live and where they work and how much time they spend at each place.

Solomon Charles, Yehuda Gur J, Shlomo Bekhor and Leonid Kheifits published an article describing how cell phone data can be used to help transport planning in Israel [23]. From the network carrier Orange Salomon et al. obtained data from 160 000 persons. One week from each person, spread out over 16 weeks. The 2200 antennas Orange had in the country was were divided into 600 traffic analysis zones. A person was defined as stationary when more than 20 minutes were spent in the same zone. Their were unable to obtain an accurate O/D-matrix down to the zone-level, but they claim to have been able to obtain an accurate O/D-matrix on a district level (there are 34 districts in Israel).

3.3 Other related work

The project "Geographic Privacy-aware Knowledge Discover and Delivery" (GeoPKDD) [2] may be one of the more ambitious projects of mapping human mobility. It was a collaboration between over 40 persons from around Europe that ended in 2008. They publiched a lot of papers regarding how to find out what routes people travel in privacy-aware manner, using mobile network data as well as GPS data. They did not, however, do much work on O/D-matrices.

Kang Jong Hee, Welbourne William, Stewart Benjamin and Borriello Gaetano [20] and Asakura Yasuo and Eiji Hato [16] have proposed algorithms for deciding if a user is stationary or moving from positioning data. They used GPS data collected on the mobile device, as opposed to cell id data collected in the cellular network. Using cell id data is much harder since the accuracy of the measurements is lower, the measurements are not evenly spaced in time and they are usually less dense in time.

4

Use cases

O/D-matrices by themselves has a lot of applications, but could also be used together with other kinds of information on people's movements, or together with demographic data. Following are some examples of what O/D-matrices can be used for.

4.1 Building new metro and mono rail in India

India is a country with many fast growing cities. One of them, Chennai, is planning to build metros and mono rails to support the inhabitants need of transportation. The aim is to build metros connecting the most important hubs of Chennai, with mono rails connecting the hubs to less travelled parts of Chennai. In order to make informed decisions on where to put the metros and the mono rails, accurate O/D-matrices are needed. The information on how people travel is already inside the operators' networks, but the methods for extracting information is not yet in place.

4.2 Traffic planning

Accurate O/D-estimations is essential when planning new infrastructure. Measuring traffic link flows is a good way of finding out what roads are inadequate for the amount of traffic traveling on them. However, the best solution might not always be to build a bigger road, but instead build a new road along another stretch. In order to know if a new road might lead to less traffic on an existing road, you need to know between what places people travel and when they travel. As an example, there has been talk about building a new road connecting Nacka with the northern parts of Stockholm, to divert traffic from the center of Stockholm and the heavily trafficked Söderleden. In order to know if this road will divert traffic from the center of Stockholm, you need to know how many of the people leaving from Nacka will travel to the northern parts of Stockholm.

4.3 Bridges and overpasses

Today, the decision of which crossings should be replaced by overpasses in India is based on information collected by people manually counting cars in crossroads. This is costly and takes a long time. Accurate O/D-matrices could be used to determine what crossings are the most important ones.

4.4 Unprotected railroad crossings

A problem many countries are facing is the large number of unprotected railroad crossings [19]. India has the largest railway network in the world with over 30000 level crossings, nearly 15000 of them being unmanned [9]. Since it is too expensive to make all of the crossings safer, it would be of interest to have information on how often each crossing is crossed, to know which crossings to focus on. Since it is too costly to place a measuring device or a person at every crossing, O/D-matrices may be of use, even though it might not give a definite answer to this question.

5

Methodology

5.1 Initial study

As a first step in the work on this thesis, literature and commercial applications were studied in order to deduce what approaches to O/D-estimation had already been conducted and how to best build on these and make them better.

When this thesis work started, the only data available was the data collected with the android application. This data necessarily shares some characteristics with real data, even though the sample frequency is not the same as in real data. This data was modified to look like data obtained from a mobile network and then studied in Matlab to determine the best way of obtaining O/D-matrices from it.

5.2 Method development

When a rough picture of what the final framework would look like were becoming clearer, the methods were implemented in Java instead. A paper was written and sent to the "International Workshop on Spatial and Spatiotemporal Data Mining" [4] and a patent application protecting the method developed was filed.

5.3 Implementation and analysis

When it became clear that the method could run in parallel, it was altered so that it could run on a Hadoop cluster. This allows for faster calculations which is of great importance if this should ever be implemented in a commercial application, since the application would have to be able to handle data on millions of mobile users, amounting to several terabytes.

During the summer of 2011 Viktor Kärnstrand worked at Ericsson. His main goal was to create synthetic data with the same characteristics as real network data that

could be used by me and other people working with the Consider8 project to evaluate our methods. When everything else was in place, this data, together with the android application data, was used to evaluate the method developed in this thesis.

6

Framework & Results

This section describes the method developed for obtaining O/D-matrices, how it has been implemented and how to evaluate it. It also describes the program developed for visualizing the O/D-matrices.

6.1 Framework

A schematic overview of the method used for calculating O/D-matrices can be seen in figure 6.1. After converting from cell id to location using a cell id database, each user is analyzed separately. The trips each user has made is calculated using only positioning data, no knowledge of the location of any cities is used at this step. After that, a reverse geocoding API is used to identify the places the user has been traveling to/from. The steps of the method (depicted in 6.1) are explained below.

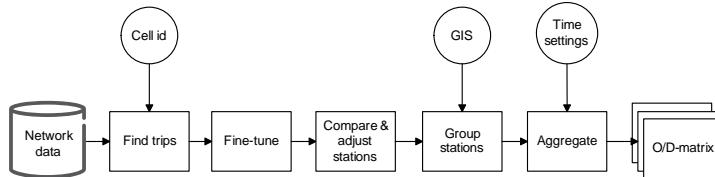


Figure 6.1: An overview of the method for obtaining O/D-matrices.

6.1.1 Convert from cell id

The input data only contains information on which base station the phone was connected to, not the location of that base station. Fortunately, there are free APIs online that you can query for that location of base stations (or, equivalently, cells). They collect the location of cells using applications installed on mobile devices, registering the gps-position and the cell it is currently connected to. Therefore, they do not have the actual

place of the base station, but rather the average location the mobile devices have been at when connected to that base station (which actually suits our need better).

There are a few public cell id databases online, for example Ericsson labs [1], opencellid [7] and location-api [5]. Location-api is the api that has the location of the most cells, at the moment of writing they have the location of more than 17 million cells around the world. Figure 6.2 shows the cells in Sweden that has been used in this thesis (there are more, but they were not needed, so the location of them has not been downloaded).

6.1.2 Finding trips

The first step towards finding movement patterns is to determine if an individual is moving or not. This might seem like an easy task, but the data is very sparse and there are a lot of fluctuations occurring even if an individual is perfectly still. In the data collected by Ericsson and SICS, a stationary phone has been noticed to jump between up to 5-6 base stations with only a few seconds on each base station. In [20] they define an easy algorithm for determining if a person is stationary or moving. The method used in this thesis is similar, with a few modifications. A basic description of the algorithm can be seen in figure 6.3. As input, the method takes all the location data points on a subscriber. The output from the method is a list of all the points the subscriber has been stationary at. These stations are then used to find all the trips made by a subscriber. The algorithm has three parameters, two spatial and one temporal, whose optimum values have to be decided empirically. Figure 6.4 shows an example of how the algorithm works.

Parameters

The algorithm has three parameters; two of them are spatial parameters and one is a temporal parameter. The first spatial parameter defines the longest distance a point can be from the previous point to be classified as stationary (in case the previous point was classified as moving). The second spatial parameter defines the maximal distance a point can be from a mean in order for the new point to be classified as stationary (in case the previous point was stationary). The time parameter defines the minimum time a person has to be stationary in order to count it as a station.

These parameters could either be fixed or dynamic. The main argument for making them dynamic is that the optimal values for the parameters might differ between urban and rural areas.

Fine-tune

The method for finding trips described above is good at finding the origin and the destination, but unfortunately has a harder time of finding the exact time of departure and arrival. This is a result from the construction of the algorithm, and from the fact that the data is very sparse in time. One way of improving the accuracy of the arrival

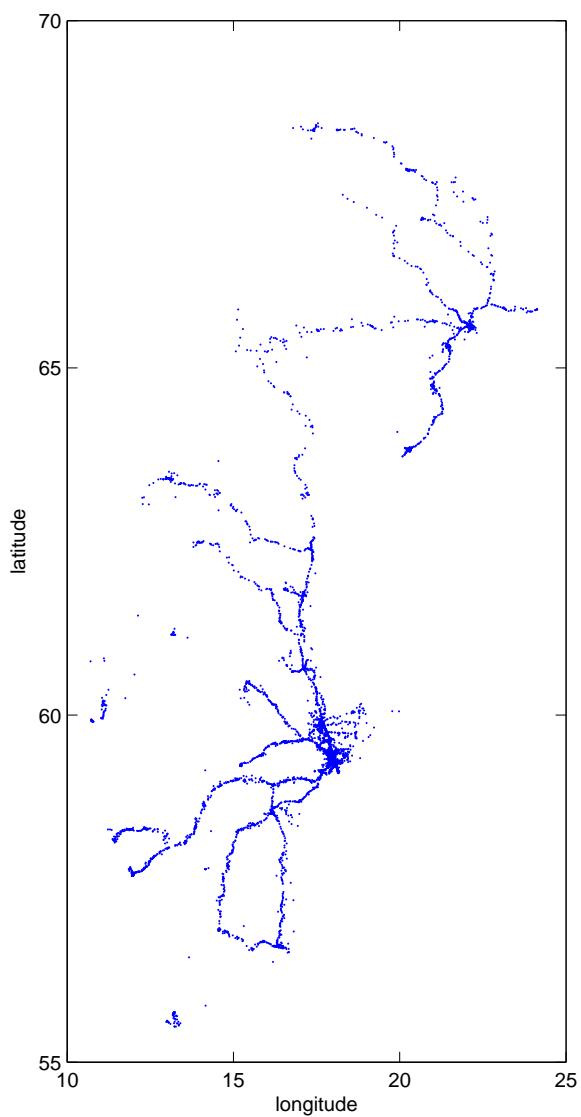


Figure 6.2: All the cells in Sweden that have been used in this thesis.

time would be to do a linear extrapolation from the points between the stations, thereby getting a better estimation of the arrival time.

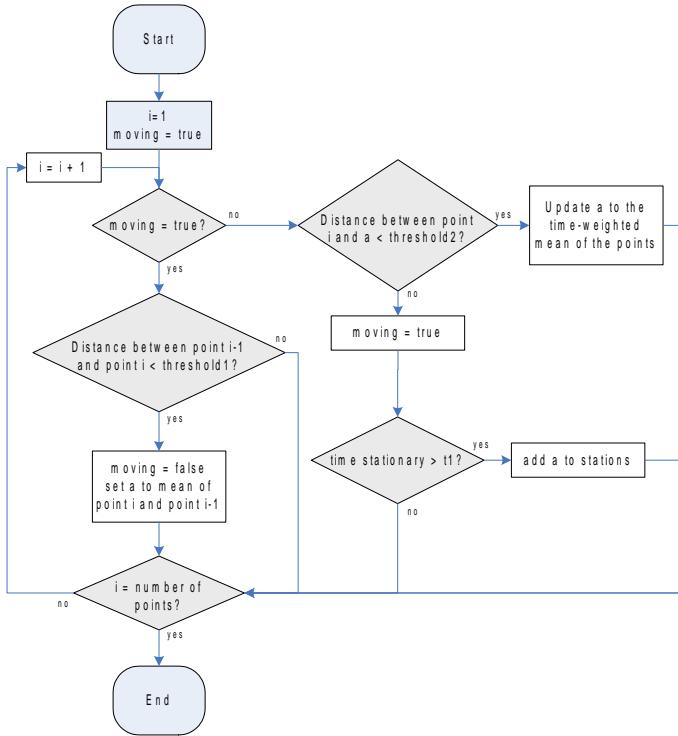


Figure 6.3: Flowchart describing how to find trips made by an individual

6.1.3 Compare and adjust stations

Since the data is very sparse, a station identified by the algorithm may contain as few as two points. This may result in a bad accuracy of the station. However, a person is likely to visit a certain place more than once. This can be used to get a better approximation of the exact location a user was stationary at. In this thesis a clustering algorithm called mean-shift clustering [21] has been used to cluster the stations. An explanation of how this algorithm is used in this thesis is described in appendix A. The reason for choosing the mean shift clustering algorithm is that it is easily implemented and it does not require any prior knowledge of the number of clusters. An example of how the algorithm clusters a set of points can be seen in figure 6.5. The figure also displays the name of the place each cluster belongs to, which is not known at this stage of the algorithm.

6.1.4 Group stations using open street map

For each trip, the location of the origin and the location of the destination are mapped to places defined by Open Street Map [6]. All the places in Sweden can be seen in figure 6.6. After this step, all trips have a name on both the origin and the destination, as can be seen in figure 6.5. Now, a matrix can be created from the trips, with the element at index i,j representing the number of trips from place with index i to the place with index j .

6.1.5 Aggregate

The O/D-estimation can either be aggregated into a big O/D-matrix or many O/D-matrices representing different times. The trips can be grouped by month, weekday/holiday, or by what time of the day they were made. A combination of those is probably the most interesting. In this thesis, a separate O/D-matrix has been created for each hour of the day and by what month the trip was made. This means that for three months data, $3 \cdot 24 = 72$ O/D-matrices would be created.

6.2 Implementation

The framework described above has been implemented in Java with location-api [5] as cell-id database and Open Street Map [3] as reverse geocoding API. The parameters in the algorithm for finding the trips has been implemented as stationary. In order to find the optimal values for the parameter, a swarm optimization algorithm has been used minimize the error between the trips calculated by the program and the trips manually observed in the data. Since the data from the android application have information on the GPS-position, this data has been used for this optimization. The step called fine-tune in the framework has not been implemented.

6.2.1 Architecture

Since every user is evaluated separately, the calculations can be made in parallel on a cluster of computers. In this implementation, Hadoop MapReduce has been used. It is a framework for doing calculations on a cluster of computers developed by Google in 2004. In short, MapReduce requires two user defined functions, a "mapper" and a "reducer", both of them using key-value pairs as both input and output. A more thorough description of how MapReduce work can be found in appendix B. The procedure of finding O/D-matrices is split up into four parts; find unique cell-ids; look up location of cells; sort data; find O/D-matrices (the main part). These steps will be explained below.

Find unique cell-ids

The input data this method uses only contains the cell id the user is connected to, not the location of the cell. Fortunately, there are a lot of free APIs online that you can query for the location of a cell from a cell id. Querying such an API more than once for the same cell id would be unnecessary and the APIs usually have restrictions on how many queries you are allowed to make. Therefore, the first step of finding O/D-matrices uses MapReduce to find all unique cell ids in the data, so that the location of each unique cell only has to be queried for once.

Look up location of cells

In this step, the location of all unique cells are downloaded from a cell id database. In this project, location-api [5] has been used.

Sort data

The input data is typically sorted by time and not by user. A rather simple MapReduce program sorts all the data by user, and also adds location, collected in the previous step, to all data points.

Find O/D-matrices

This is obviously the main part of the program. It uses MapReduce to run the algorithm described in 6.1. An explanation of how it is implemented is provided in appendix B.

6.3 Visualization

Even though the O/D-matrices is the main result of this thesis, they are rather boring to look at and hard to get any real knowledge from. Therefore, a program for visualizing the O/D-matrices has been developed. The program is just an example of how you can use the O/D-matrices to get more knowledge from them then by just looking at a list of numbers.

A Tomcat server was used to run the method of obtaining O/D-matrices described above and to return the O/D-matrices in xml-format. A client running javascript and using the Google Earth web plugin was created, displaying the O/D-matrices as shown in 6.7. A client was also created displaying the O/D-matrices in Google Maps 6.8. Since both of these were rather slow when a lot of points were added to them, a third alternative were created, displaying the O/D-matrices in a java application instead. The java application is also capable of showing how each individual has moved and can be used to evaluate how well the method called "finding trips" works.

6.4 Results

6.4.1 Finding trips

The method "finding trips" correctly classifies all the trips made in the simulated gps data, where the time stationary in both the origin and destination is greater then around 10 minutes (this time depends on value of the time parameter used in the method, which in this case is set to 10 minutes).

Values

The optimal values for the parameters used in the algorithm called "Finding trips" were calculated using a swarm optimization algorithm running on the data collected by the Android application. The optimal values obtained was about 1 km for the spatial parameters and about 10 minutes for the time parameter.

6.4.2 Simulated data

The simulated data used to evaluate the method contains data on 513 users for about 3 hours time. The first run showed that the method could only find 775 out of 1985 trips. When investigating why this was, it turned out that the simulated data only started saving data points for an agent from when he first started moving until he stopped moving for the last time. The method for obtaining O/D-matrices will then not detect that the agent has been stationary at those places. Taking this into account, it turns out that the method is capable of finding 775 out of 959 trips. Investigating this further, one finds that in the trips the method is not capable of finding, the agent has been stationary for less than around 10 minutes at one of the stations.

6.4.3 Real data

It is hard to evaluate the real data, since there are no O/D-matrices to compare the results to. However, you can study how many trips each person makes and how they are distributed over the day. This does not give a definite answer to whether or not the method actually does what it is supposed to do, but it gives a hint at the answer. In figure 6.9 the distribution of trips over the day obtained from the real data set is plotted. In total there are 867209 trips, which is very low considering that it is data from roughly two million users over one months time. Of course, this data is only call data records, which is much less dense in time compared to the data the method was written for. It should also be mentioned that location-api only had the location of less than 10% of the cells in the country the call data records originated from, making this analysis even harder. Still, from figure 6.9, one can draw the conclusion that the distribution of the trips over the day seems reasonable, which is an indication that the method is working.

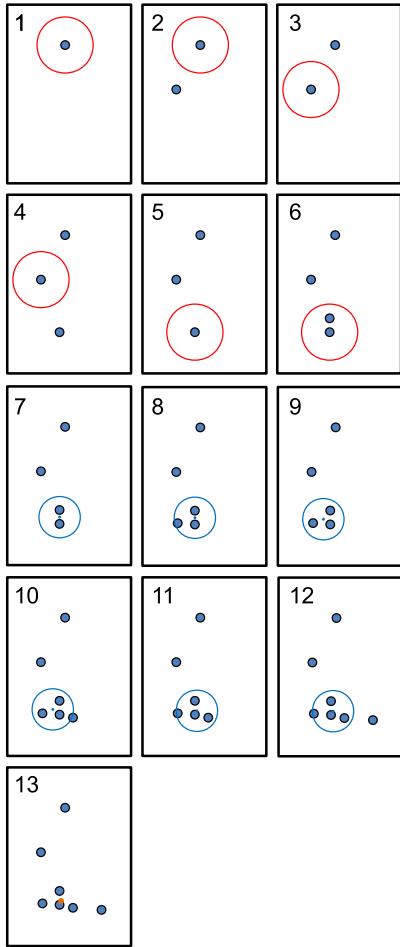


Figure 6.4: An illustration the algorithm for finding the trips made by an individual. The dots (or rather: smaller, filled circles) represent a location of a cell the subscriber has been connected to before the time the box represents. In the first box, the subscriber has only been at one cell. If the next data point is within the circle, the point will be classified as temporarily stationary. In this case, the next point is outside the circle, as can be seen in box 2, and the point is therefore classified as moving. The same procedure is then repeated in box 3 and 4. In box 6 however, the next point is within the circle. The point is therefore classified as stationary, and a mean point is created (box 7). As long as the subsequent points are within (the now smaller) radius, the points keep being classified as stationary. As soon as one point falls outside of the circle (box 12), the new point is classified as moving again. If the time between the first and the last stationary point (the point added in box 4 and the point added in box 10) is larger than a certain threshold, the mean point (weighted by the time spent in each point) is added as a station. When two stations have been added without any longer time between data points (e.g. the phone has been off for two days) in between the two stations, a trip will be created. This figure also explains the three parameters of the algorithm, the two radii and the time threshold.

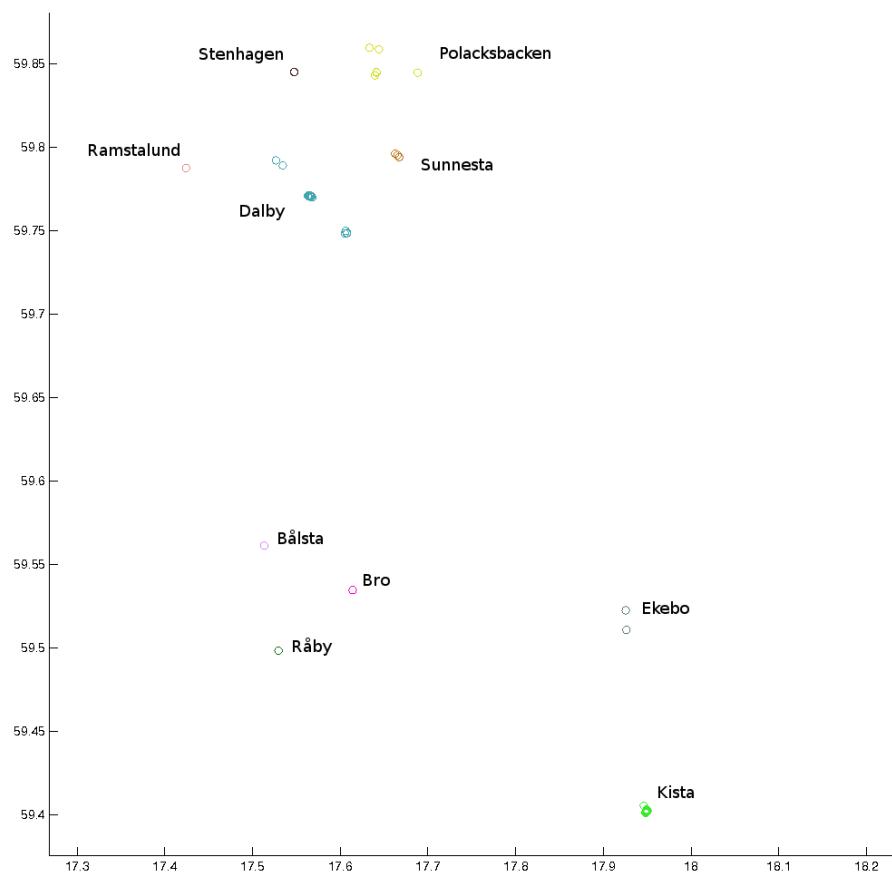


Figure 6.5: All stations for one user of the Android application and how they were clustered.

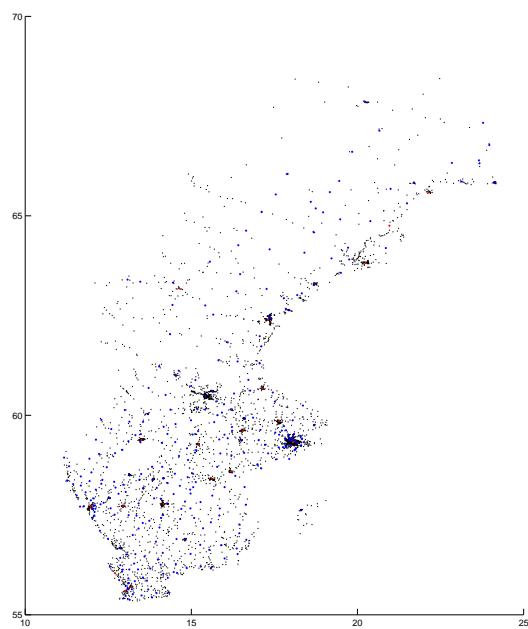


Figure 6.6: All the "places" in Sweden, extracted from Open Street Map [6].

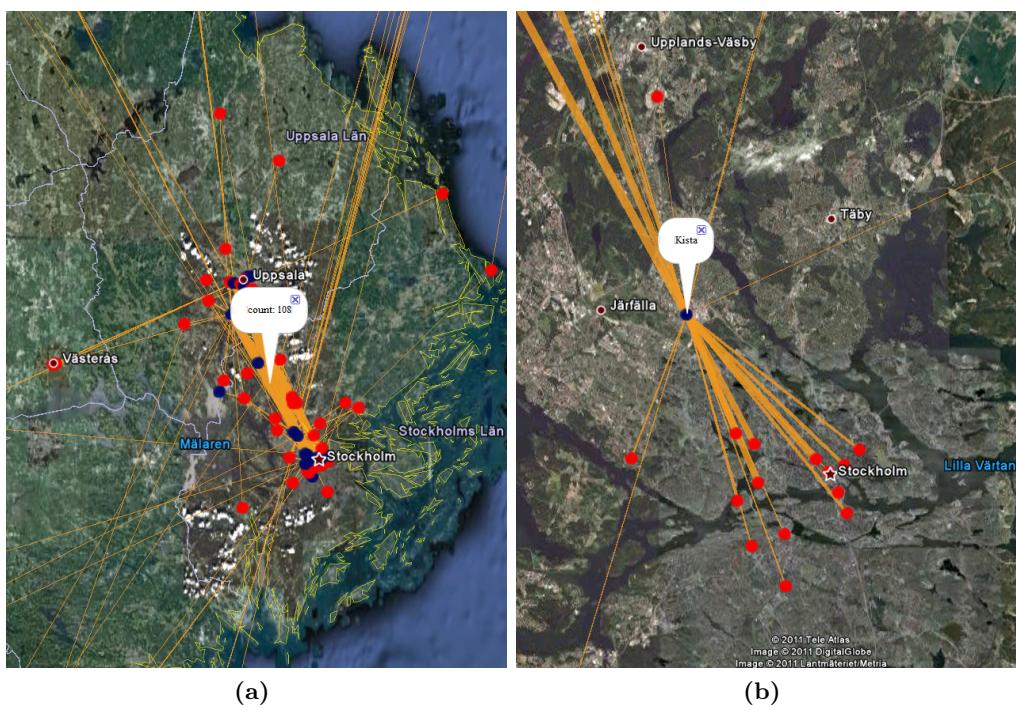


Figure 6.7: The O/D-matrices displayed in a web browser using Google Earth. The red and blue circles represents origins and destinations and the lines represents the trips from one place to another. The more trips that were made, the thicker the line. Clicking on a place shows the name of the place and clicking on a line shows the total number of trips between the two places.

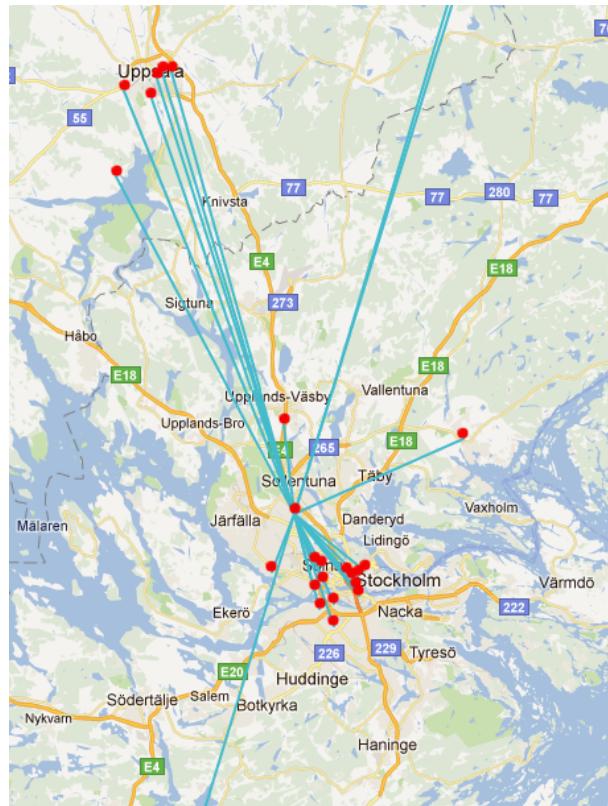


Figure 6.8: The O/D-matrices displayed in a web browser using Google Maps.

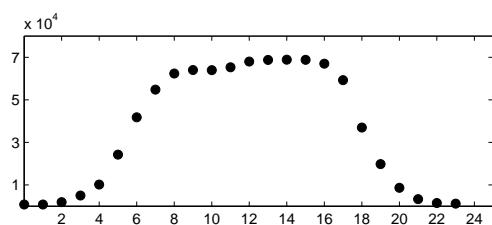


Figure 6.9: The distribution of trips over the day from the real data set.

7

Discussion

This chapter will discuss the advantages and disadvantages of the method developed and the results from running this method on the data available to Ericsson now.

7.1 Method

7.1.1 Advantages

The methodology proposed has a lot of advantages. Comparing all stations for a single person makes it possible to get a better approximation of the location the user is actually at. The reason this is possible is that a person is likely to visit the same place more than once. Comparing the stations a person has visited and clustering them therefore makes it possible to get a better prediction of where the person actually was. Another advantage of treating each person individually is that it makes it possible to parallelize the problem using for example Hadoop and thereby making it a lot faster.

Privacy

When creating the O/D-matrices, all information that could be used to identify an individual is lost. This is however not enough to guarantee that an individual cannot be identified [18]. Another way of assuring anonymity is k-anonymity, which means that every individual should be indistinguishable from $k-1$ other, where k is a predefined number [18]. From the O/D-matrices obtained by the method described in this thesis, it is impossible to tell what day someone traveled from one place to another, you can only tell which month the trip took place. If you know at what time a person left a place and there are less than $k-1$ other people leaving that place during that month within the same time interval, then the method would not be k-anonymized. K-anonymity could be enforced either by removing information from the O/D-matrices that does not satisfy k-anonymity, or the matrices could be modified to enforce k-anonymity. Less than k

trips from a place within a certain time frame would typically mean that the place is small village and the time interval is in the middle of the night. K-anonymity could then be enforced by either clustering villages together or by making the time intervals longer when fewer people travel.

7.1.2 Disadvantages

The biggest disadvantage of the method has to do with the definition of a trip. In the method developed in this thesis, a trip is defined as a travel between two places where the person spend at least 10 minutes. But taking the subway to the train station, just to leave the city by train 15 minutes later would in that case be split up into two trips. This would not be considered two separate trips in reality. But taking the car to the supermarket and then going back 15 minutes later would be consider two trips (both in reality and by the method). This is a difficult problem, and it cannot be solved by simple tweaking the parameters of the method. What you could do is to make this time limit vary depending on how much time you spent traveling before and after as well as the place where the person were stationary.

Another, somewhat related problem, is that some trips will not be identified because of the sparsity of the data. If a subscriber does not use his or her phone while shopping at the supermarket, chances are that there will not be any data collected in the mobile network of his/her presence there. This is not really a disadvantage of the method, but rather the data itself. It could however be dealt with, for example by analyzing data from the same user over a long time. That way, you might be able to draw the conclusions like: if he/she left his/her home and passed that certain place just to return home half an hour later, he/she probably visited this supermarket, since that has happened three times before (when there were more data available).

7.1.3 Improvements

The values of the parameters in the result, the spatial parameter at 1 km and the time parameter at 10 minutes, should optimally not be fixed, but should vary depending on where the user is. They should be smaller in densely populated areas, where the distance between base stations are shorter. These values are an early result and more work is needed to find the optimal values of these parameters.

One could also consider adding more parameters to the algorithm for finding trips. You could for example add a parameter specifying the minimum travel time or a minimum travel distance. This way, two stations would be considered the same if they are not sufficiently separated in time or space. Tries were made using these parameters, but they were deemed unnecessary for the classification error of the trips.

7.2 Result

Running the algorithm on the call data records from two million users has shown that it possible to get O/D-matrices from real cellular network data. This result was from

using call data records, containing only location information for a user at the beginning of each phone call. The algorithm is constructed for using on much more fine-grained data, containing data throughout the phone call as well as passive data collected when the phone is in idle mode. With this data, the method would hopefully be able to perform better.

Unfortunately, no real O/D-matrices obtained in some other way has been used to verify and compare the results. However, when studying the data and the movement distribution over the days, the results seems fair.

The more often a person uses his/her phone, the more accurate O/D-estimation it is possible to make. Smart phones generate much more data, since they are more often connected to the network. This makes it easier to make O/D-estimations. The usage of smart phones is only increasing, which will make this method even better in the future.

8

Conclusion and future work

There are a few companies and organizations working on obtaining patterns of human mobility from mobile network data, and it will be interesting to see what company comes out ahead. There are also a lot of location-based services that are growing fast, for example four-square, facebook places, Google places etc. These are not really competitors to the method proposed in this thesis, since they do not collect information detailed enough for this kind of analysis. However, the data collected by Apple and Google latitude might have the resolution needed to do this kind of analysis. Whether or not they are working on projects similar to Consider8 is not known.

In the case of Google latitude, people are giving up the information voluntarily, Apple however saved information on a persons location in an iPhone without the users consent. Even though they never sent this information to be collected somewhere, this started some commotion [11], and they were blamed of tracking people. It is therefore apparent that it will be of great importance that it is impossible to track a single individual from the O/D-matrices obtained by the method developed in this thesis work. It is also likely that the personal privacy of the first attempt of releasing patterns of human mobility will determine the future of this kind of location based information.

This thesis has shown that it is feasible to obtain O/D-matrices from mobile network data while keeping the privacy of the subscribers. The most important step that remains before this method could be used in a commercial application is to convince the mobile operators that they can run this method without risking to lose the subscribers trust, so that the method can be tested on real data. The ultimate goal here would be to sell a program containing this method to the operators, so that they themselves can run the method and release the O/D-matrices. That way, the sensitive data does not have to leave the operators networks.

Bibliography

- [1] Ericsson labs. <https://labs.ericsson.com>.
- [2] Geographic privacy-aware knowledge discovery and delivery. <http://www.geopkdd.eu>.
- [3] google geocoding. <http://code.google.com/apis/maps/documentation/geocoding/>.
- [4] International workshop on spatial and spatiotemporal data mining.
- [5] location-api. <http://www.location-api.com>.
- [6] Open street map.
- [7] Opencellid. <http://opencellid.org>.
- [8] Your movements speak for themselves: Space-time travel data is analytic super-food! http://jeffjonas.typepad.com/jeff_jonas/2009/08/your-movements-speak-for-themselves-spacetime-travel-data-is-analytic-superfood.html, August 2009.
- [9] 50 percent railway crossings are unmanned. <http://facenfacts.com/NewsDetails/12307/50-percent-railway-crossings-are-unmanned.htm>, July 2011.
- [10] Airsage. <http://www.airsage.com>, 2011.
- [11] Apple under pressure over iphone location tracking. <http://www.telegraph.co.uk/technology/apple/8466357/Apple-under-pressure-over-iPhone-location-tracking.html>, April 2011.
- [12] Traffic transformation part deux gps probes vs. handoff data. <http://blogs.strategyanalytics.com/auto/?p=109>, May 2011.

BIBLIOGRAPHY

- [13] Torgil Abrahamsson. Estimation of origin-destination matrices using traffic counts - a literature survey, 1998.
- [14] Rein Ahas, Anto Aasa, Siiri Silm, and Margus Tiru. Daily rhythms of suburban commuters movements in the tallinn, 2009.
- [15] Andreas Allström and David GrundlegÅrd. Stockholm mobile millennium - modeller fär realtidsestimering av restider. s.l., January 2011.
- [16] Yasuo Asakura and Hato Eiji. Tracking survey for individual travel behaviour using, 2003.
- [17] N. Caceres, J.P. Wideberg, and F.G. Benitez. Deriving origin-destination data from mobile phone network, 2007.
- [18] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-anonymity, 2007.
- [19] Emily Ford. Salisbury city council debates unprotected railroad crossings. <http://www.salisburypost.com/News/031711-Salisbury-City-Council-debates-unprotected-railroad-crossings-qcd>, March 2011.
- [20] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. Extracting places from traces of locations, 2005.
- [21] G. Derpanis Konstantinos. Mean shift clustering, August 2005.
- [22] Keemin Sohn and Daehyun Kim. Dynamic origin-destination flow estimation using cellular communication system.
- [23] Charles Solomon, Gur J Yehuda, Bekhor Shlomo, and Kheifits Leonid. Intercity person trip table for nationwide transportation planning in israel, 2009.

A

Mean shift clustering

The mean shift clustering algorithm does not need any prior knowledge of the number of clusters the points belong to, as apposed to many other clustering algorithms. This appendix will only explain how the algorithm works for the specific purpose it is used for in this thesis. For a general explanation of the algorithm, see [21].

The aim of the algorithm is cluster n points in a two dimensional space into an unknown number of cluster. We define a density:

$$f(x) = c \sum_{i=1}^n k(\mathbf{x} - \mathbf{x}_i),$$

where c is a constant and

$$k(x) = e^{-ax^2},$$

where a is another constant. The function $f(x)$ has a number of local maxima, where the density is (locally) higher. In such a point,

$$0 = \nabla f(x) = \sum_{i=1}^n -a(\mathbf{x} - \mathbf{x}_i) e^{a(\mathbf{x}-\mathbf{x}_i)^2} = a \left[\sum_{i=1}^n e^{-a(\mathbf{x}-\mathbf{x}_i)^2} \right] \left[\frac{\sum_{i=1}^n \mathbf{x}_i e^{-a(\mathbf{x}-\mathbf{x}_i)^2}}{\sum_{i=1}^n e^{-a(\mathbf{x}-\mathbf{x}_i)^2}} - \mathbf{x} \right].$$

The first term is a scalar, therefore we only have to consider the second term when deciding the direction of maximum increase. Define

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{x}_i e^{-a(\mathbf{x}-\mathbf{x}_i)^2}}{\sum_{i=1}^n e^{-a(\mathbf{x}-\mathbf{x}_i)^2}} - \mathbf{x}$$

and

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \mathbf{m}(\mathbf{x}^t). \quad (\text{A.1})$$

The sequence in A.1 will converge. All the points that converge to the same point are considered to belong to a cluster.

B

Hadoop MapReduce

Hadoop MapReduce is a framework developed by Google in 2004. It was developed for being able to do calculations on large amounts of data. It does the calculation in two steps, called "map" and "reduce", both of them using key-value pairs as input and output.

Figure B.1 describes how MapReduces was used to find O/D-matrices. As input, the user id is used as key, and a list of all data points (longitude, latitude and timestamp) are used as value. The MapReduce framework then takes care of sending all data with the same key to one computer, and then start the "mapper" function, where the main calculations happen. Output from the mapper is the O/D-matrices for that specific user in the form of key-value pairs, where the key is the position in the O/D-matrix (the id of the origin, the id of the destination and the time of departure) and the value is the number of trips that user made with that key. Again, MapReduce makes sure all key-value pairs having the same key is copied to the same computer and then starts the "reducer" function. In this case, the reducer function is very simple, the output key being the same as the input key, and the output value the sum of all input values. The output is therefore the actual O/D-matrices we were trying to obtain.

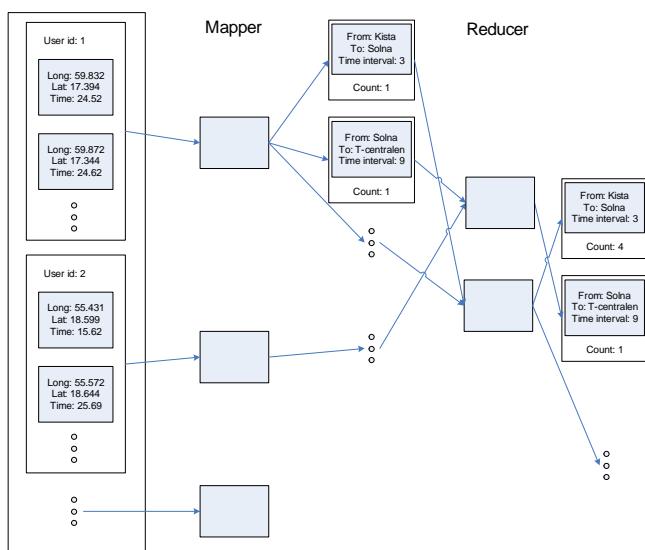


Figure B.1: An illustration of how hadoop operates on key-value pairs.

C

Data collection

The accuracy with which one can determine where a person is depends a lot on what kind of data is available, specifically the frequency and accuracy of the positioning for the mobile phones. The accuracy with which one can pinpoint a certain phone depends on where in the mobile network the information is gathered. Figure C.1 displays a simplification of the mobile network as it is in Europe.

A number of Base Transceiver Stations (BTS) are connected to a Base Station Controller (BSC). A number of BSC is in turn connected to a Mobile Switching Center (MSC). To be able to reach a phone when someone calls it, the network needs to know where the phone is situated. Therefore, all mobile devices are required to give their position to the mobile network about once an hour (or somewhere between 20 minutes and two hours depending on the mobile operator).

During the call, or when using some other kind of service, for example surfing the web, the MSC gets an update on which BTS the phone is connected to twice a second. This information is then stored for billing and network purposes.

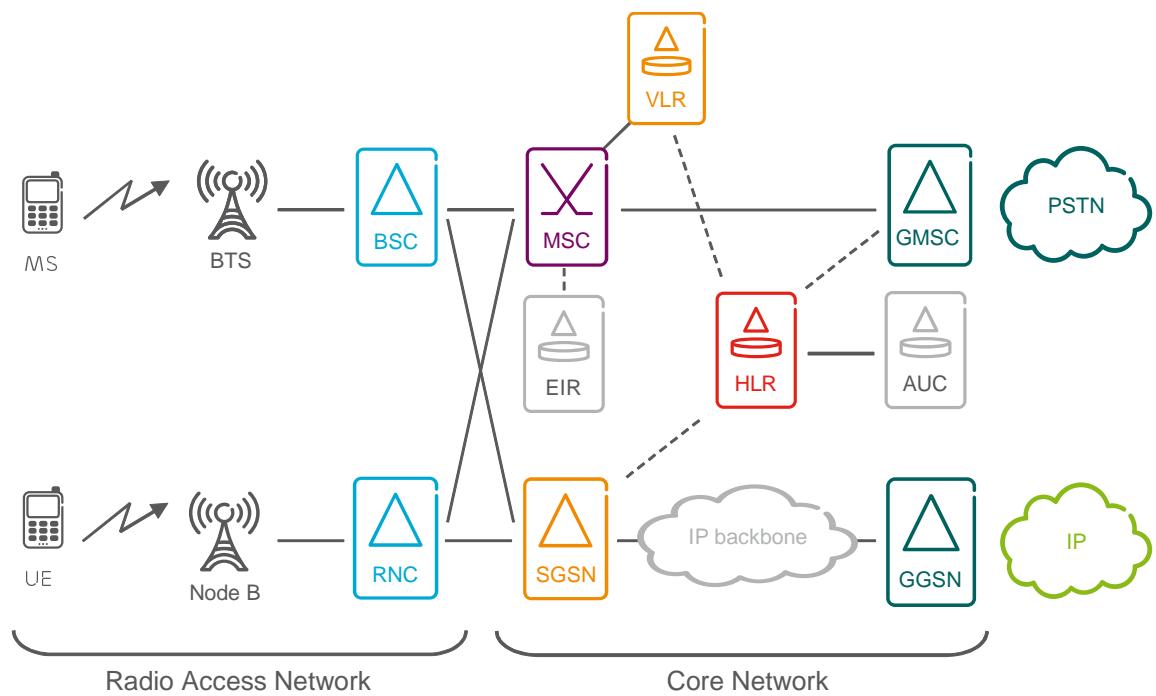


Figure C.1: An illustration of where the data is collected in the mobile network. The upper part of the figure shows the GSM (or 2G, second generation) network, while the lower part shows the corresponding functions for the 3G network.

COMPARING THE QUALITY OF OD MATRICES: IN TIME AND BETWEEN DATA SOURCES

Tim Pollard, Nick Taylor, Tom van Vuren
Mott MacDonald

Abstract

Essential but expensive to carry out, roadside interviews are important building blocks for travel demand models. As a rule of thumb, they cost approx. 10 euros per trip record, and capture around 10% of the total passing traffic – expansion factors of 10 are not uncommon. Many (highway) assignment models rely heavily on these types of survey.

The UK Department for Transport requires data in models to be not older than 6 years. In a period of austerity, it is not unreasonable to query whether older roadside interviews will suffice for a model application, or at least to try and use older RSIs, appropriately weighted, in OD matrix estimation. Alternatively, other sources for observed travel patterns are considerably cheaper, but somehow the profession demands that their value is proven by comparison with traditional roadside interviews. An interesting paper by Potter et al (2011) raises concerns about the dependence on roadside interviews in modern transport models.

Recent literature (Djukic et al, 2013) suggests a comparator between OD matrices, based on a similarity index developed for comparing images at pixel level.

In our paper we discuss the following:

- The complexities in comparing OD matrices using standard statistical tests,
- The application of the MSSIM test as described by Djukic et al

We conclude that the MSSIM test is a valuable addition to the transport modeller's toolkit for comparing matrices, in time, between sources, and pre and post matrix estimation from counts. However, further work is required in the operational application of the MSSIM test, which we hope to present in a future paper.

1 INTRODUCTION

Origin-destination, OD or trip matrices are crucial elements of travel demand models, reflecting the travel patterns in the area of study. Often they are the most expensive, and certainly the most complex element in the implementation of a transport model, but it is difficult to determine the quality of either the matrix components or the overall matrix. There are many reasons why there is value in understanding the quality of an OD matrix:

- Avoiding unnecessary surveys where existing data may suffice
- Comparing matrices derived from cheaper (BIG) data sources with those derived by traditional means, to confirm their acceptability
- Comparing the quality of synthesized matrices with observed ones, to strengthen our belief in the underlying model

Comparing the difference between prior and estimated matrix, when refining OD matrices by using counts

Essential but expensive to carry out, roadside interviews are important building blocks for trip matrices. As a rule of thumb, they cost approximately 10 euros per trip record, and capture around 10% of the total passing traffic – expansion factors of 10 are not uncommon. Many (highway) assignment models rely heavily on these types of survey.

The UK Department for Transport guidance in DfT (2012) requires data in models to be no more than 6 years old and we expect similar rules apply elsewhere in Europe. In a period of austerity, it is not unreasonable to query whether older roadside interviews or any other data containing travel pattern information will suffice for a model application, or at least to try and use older data sources, appropriately weighted, in OD matrix estimation. An interesting paper by Potter et al (2011) raises concerns about the dependence on roadside interviews in modern transport models.

Alternatively, other (passively collected) sources for observed travel patterns may be considered. We have extensive experience in the development of OD matrices from GPS data (Van Vuren and Carey, 2011). A recent paper by Calabrese et al (2013) explored mobile phone data. These data sources are substantially cheaper, but until better established the profession demands that their value is proven by comparison with traditional sources, including roadside interviews or other observation samples.

Finally, when using matrix estimation techniques from counts, it is a requirement to prove that the estimated matrix does not differ too much from the original, prior matrix. The UK DfT's WebTAG guidance states maximum allowable differences in the form of slopes, intercepts and R^2 values of a regression line between before and after OD cells and trip ends. But are those appropriate comparators for spatial patterns? Are correlations ignored? And is an R^2 of 0.99 a real necessity or just a convenient number?

In our paper we explore alternative ways of assessing the quality of OD matrices, and we intend to identify not only more appropriate metrics for comparison, but also appropriate values for these metrics that provide more evidence that the quality of matrices is actually sufficient. In this exploration we make use of a new comparison measure, the MSSIM, as presented by Djukic et al (2013).

The rest of our paper is structured as follows:

- In section 2 we describe other published work in the area;
- In section 3 we discuss in more detail the MSSIM and how we have implemented this in practice;
- In section 4 we describe the results of our comparisons, and investigate at what values the MSSIM might indicate matrix comparisons of sufficient quality;
- In section 5 we draw conclusions, explore deficiencies in our work to date and how future work may improve on this.

2. PREVIOUS WORK

There is remarkably little published material on the comparison of trip matrices. Perhaps this is not surprising as they are in general not observed, but constructed from different datasets. If they are based on observations, these datasets may have known statistical properties (e.g. determined by sample rates) and these are taken into consideration when constructing matrices. Other datasets, however, such as synthesized matrices from models, have unknown statistical properties. Anyway, even if such a comparison were possible, the results will not be able to show if they are statistically similar enough.

As a result, comparisons of matrices and determination of the acceptability or not of their similarity is done on the basis of their practical application, after assignment. If assigned flows are similar, the matrices are considered to be similar as well. The main problem of this approach is that it smooths over differences between matrices

and as a result can give a false impression of similarity. Figure 1 below illustrates how two structurally different matrices lead to exactly the same link flows.

In Figure 1 the black numbers indicate observed flows in-between the 4 junctions/zones of 10, 15 and 12 vehicles. The red numbers show two possible matrix solutions, one with many short trips, and the second with fewer trips, but over longer distances.

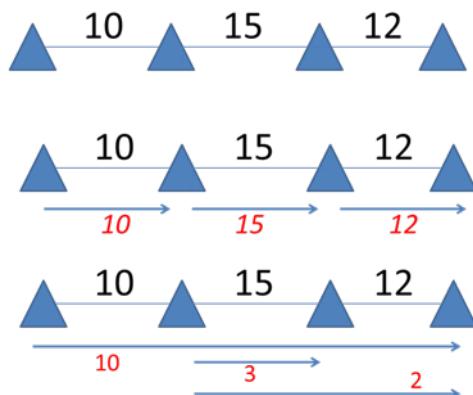


Figure 1: structurally different matrices lead to exactly the same link flows

Sometimes, a further comparison of the trip length distribution in a matrix will be made. This would certainly have detected the structural difference between the matrices in Figure 1, but still there is no clear indicator that will identify an acceptable level of similarity.

Djukic et al (2013) describe the Mean Structural SIMilarity (MSSIM) index in their paper and make a convincing case why this is potentially a more appropriate comparator of matrices. They refer to an earlier, original paper by Zhou et al (2004), who describe the index against a background of comparing images. From a structural point of view, if we liken an image to an OD matrix, pixels equate to individual OD cells; this is attractive as neighbouring pixels / cells are expected to have a degree of correlation just as in images, and the MSSIM index is designed to capture this.

Whereas Djukic et al describe tests that illustrate the value of the MSSIM index in comparing estimated matrices, they do not go as far as exploring which values of the index might indicate a certain level of acceptability. In this paper we discuss tests that we have carried out during the re-estimation of the PRISM model (see van Vuren et al (2004)) for the Greater Birmingham region in the UK, based on new surveys carried out between 2009 and 2011, and also incorporating Big Data from GPS satnav sources.

The work reported in this paper does not yet lead to firm conclusions. To a large extent, only when trying to pursue this line of enquiry did we start recognising practical issues to consider when using the MSSIM index as a matrix comparator. However, our initial findings are encouraging in that they appear to add information over and above the traditional statistical comparisons on the matrix itself before and after assignment.

3. THE MSSIM IN A TRAFFIC CONTEXT

The MSSIM as described by Zhou et al (2004) is a method for comparing two greyscale images. Briefly, the MSSIM computes statistics on groups of pixels and then takes the average (mean), rather than computing statistics based on all the pixels in the image together (such as an R^2 test).

Before we consider how or if this may be applicable to transport applications we shall first look at the details of how it is calculated.

Imagine two pictures A and B (of the same size in pixels). By converting each pixel to a number we can consider a collection of numbers arranged in a matrix instead of a rectangular picture. Consider these two picture-matrices side by side, and then look at an NxN square window in the top left of each (in the example Fig 2 below we are using a 4x4 square window). For each of the windows (a and b) calculate the mean and the variance of each, and then the covariance of the two windows together. In the notation below let μ_a and μ_b refer to the mean of numbers in window a of picture-matrix A and the mean of numbers in window b of picture-matrix B. Similarly, σ_a^2 , σ_b^2 refer to the variance in windows a and b, and σ_{ab} to the covariance.

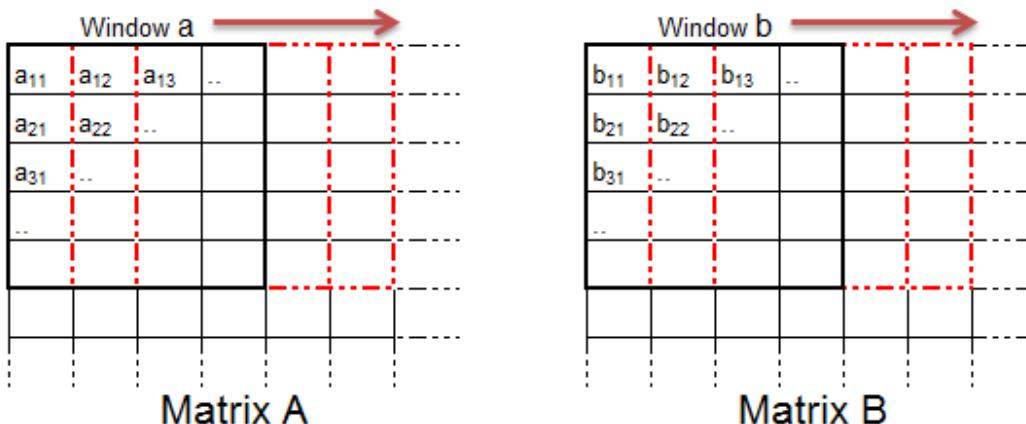


Figure 2: Comparison of two pictures using MSSIM

These numbers are then combined in equation [Fig 3] to form a single Structural SIMilarity Index (SSIM). The equation also has constants C_1 and C_2 ; these will be explored later in this paper.

$$SSIM(a, b) = \frac{(2\mu_a\mu_b + C_1)(2\sigma_{ab} + C_2)}{(\mu_a^2 + \mu_b^2 + C_1)(\sigma_a^2 + \sigma_b^2 + C_2)}$$

Figure 3: SSIM between two windows a and b.

After calculating the SSIM, the windows a, b, are then shifted one element to the right, and another SSIM is calculated, followed by another shift to produce another SSIM. Once it has moved all the way to the right, it starts again from the left but moves one element down.

Finally, once all possible SSIMs have been calculated from all possible windows within the picture-matrices, the mean of these SSIMs is taken to produce the MSSIM.

$$MSSIM = \frac{1}{N} \sum_{i=1}^N SSIM(a_i, b_i)$$

Figure 4: MSSIM between two matrices.

N is the number of windows the SSIM is calculated on

The MSSIM is bounded (between -1 and +1) and is designed to give a value of 1 for identical pictures, and 0 (or close to 0) for two random pictures.

By grouping pixels, this technique assumes that if two pictures are similar then on average the windows will be similar in magnitude of brightness (comparing means), have a similar range of brightness (comparing the variances), and that the arrangement of light-to-dark pixels is similar in both (compared with the covariance). The paper by Zhou et al (2004) goes into detail showing how this technique is combining three comparisons (on mean, variance and covariance) and that the MSSIM provides additional information when comparing pictures making it more useful than standard image comparison tests.

When applying this technique (or any other matrix comparison technique from a different field) to transport OD matrices, one must consider how this relates to the physical meaning of the matrix, and hence if it is still a useful comparison tool.

For example, consider two OD matrices A and B, and associated 6x6 windows a and b. Each window represents movements from 6 origin zones to 6 destination zones (Fig 4).

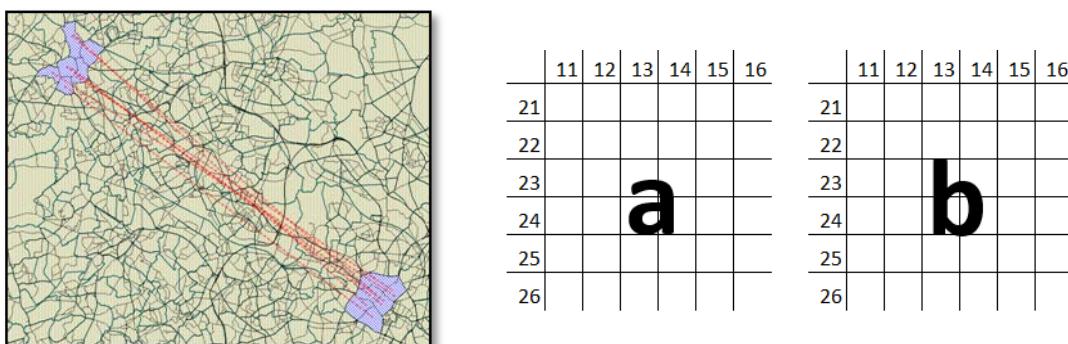


Figure 5: OD matrices A and B with associated 6x6 windows a and b.

If the two matrices describe similar transport patterns then, as with pixels, it is reasonable to expect similarities on windows of OD patterns. Within each pair of windows a, b on two similar matrices we would expect:

- the total volume and hence mean volume of traffic to be similar in each;
- the variance of those travel patterns to be similar (e.g. very small variance if both are city centres with high demand, or perhaps a high variance if one of

- the zones has a very low population);
- the pattern of travel demand to be similar.

Since the three components of the SSIM measure (mean, variance and covariance comparisons) compare aspects of transport patterns that we would expect to be similar in similar OD matrices, it becomes clear that the MSSIM index should be as valuable in comparing OD matrices as with images.

3.1 Testing MSSIM on OD Matrices

In this paper we calculate the MSSIM between matrices, and compare these results with other statistical tests. Before applying the MSSIM to a matrix we need to consider both the values of constants C_1 , C_2 and the size of the windows being used. The default values suggested by Zhou et al (2004) of constants C_1 and C_2 taking values 6.5... and 58.5...¹ and using a window of 10x10 pixels did not seem applicable to our context.

The constants in the SSIM equation (fig 3) stabilise the result in the case where either the mean or variance is close to zero. We had chosen to use a matrix from the PRISM model (see van Vuren et al (2004)) as the basis for comparison, however the mean cell value in the matrix is close to 10^{-6} . Clearly for such a matrix the constants will dominate the SSIM equation for most windows, resulting in MSSIM values of near 1 regardless of the matrices you compare with. This was indeed confirmed by testing.

In trying to calculate suitable constants for MSSIM in this context, we decided to consider the relationship between the constants and typical values in their original application – namely greyscale images where each pixel ranges between 256 possible values of grey Zhou et al (2004). If we assume pixel values are randomly² chosen between 0 and 255 then the mean squared is $\sim 10^4$ and variance is $\sim 10^3$. In the SSIM equation (fig 3) the constant C_1 stabilises the effects of the mean squared, and the default value for C_1 is 4 orders of magnitude less than this expected mean from a greyscale image. Likewise the variance is stabilised by C_2 which is 2 orders of magnitude less than the expected variance from the greyscale. We used these crude relationships to guide us in choosing more appropriate constants for the equation for an OD matrix context.

Taking median of values in the PRISM validated matrix we found a mean squared of $\sim 10^{-6}$, and variance $\sim 10^0$. Hence we adjusted our constants to values of $C_1 = 10^{-10}$, $C_2 = 10^{-2}$.

When considering the size of the window, a detailed look at the size of the prism zones was undertaken, and a typical consecutive collection of 6 zones numbers produced a pattern of 6 adjacent zones approximately 3km across in total. Further work is needed in the future to compare the impact to MSSIM values of different

¹ Zhou et al (2004) calculates $c = k * L^2$ with k equalling 0.01 and 0.03 for C_1 and C_2 .

² A test was carried out on 100 such samples, from which median values of mean and variance were used.

window sizes.

4. RESULTS OF COMPARISONS

The method and results of the investigation can been split into two parts. Firstly we looked at the possibility of using the MSSIM index to see how altering the structure of a matrix affects the flows on an assignment. Secondly we looked at comparing the MSSIM index with another measure for assessing the difference between matrices; the R^2 measure. Our overall aim was to see how the MSSIM index behaves with structurally different matrices and how it can be used in a practical situation alongside other measures.

In this paper we focus on one particular matrix, the 2011 Employer-Business matrix from a large and detailed transport model named PRISM. PRISM, is a multi-model, multi-purpose transport model based on the West Midlands area in the UK described in more detail in van Vuren et al (2004).

We have chosen the Employer-Business matrix in particular as it has been calibrated and validated to a high standard using nearly 1500 traffic counts and it is reasonable large containing nearly 50,000 trips in total. The matrix was the result of a merge of a Road Side Interview (RSI) matrix constructed from surveys undertaken from 2009 to 2011; a GPS matrix built from various GPS devices including satellite navigation devices; and a Synthetic matrix created from a demand model applied to the network using 2011 land use data. We shall use this matrix as the 'ground truth' matrix for our comparisons.

To enable testing, we artificially created a set of matrices that are structurally different from the ground truth matrix. We added different types of 'noise' to the Employer Business matrix by adding varying amounts of the source matrices. We then scaled the resulting matrices so that their totals were equal to the total of the original matrix.

Four types of matrix have been created:

$$\text{Type 1. } B_1 = A + \alpha \text{Random}$$

$$\text{Type 2. } B_2 = A + \alpha \text{RSI}$$

$$\text{Type 3. } B_3 = A + \alpha \text{RSI} + \beta \text{GPS}$$

$$\text{Type 4. } B_4 = A + \alpha \text{RSI} + \beta \text{GPS} + \gamma \text{Random}$$

In the list above, A denotes the Employer-Business matrix; α , β and γ are random numbers from -1 to 1; Random denotes a matrix of random values from 0 to 1 (inclusive); and B is the resulting altered matrix³. After this it was important to scale each matrix so that its total size was equal to that of the ground truth. This is to remove bias from the test, which would otherwise be created as some of the matrices created would be larger than others due to the method in which they were made. It was felt such a removal of bias was necessary as in any real-world matrix build

³ It should be noted that we capped our matrix values below at 0, and so the resultant B matrices will contain some 0 values. Also the random matrix used is different for each new test matrix produced.

application (before comparing two matrices for a merge) they would be scaled in some way to result in similar volumes of traffic.

The resulting matrices were then compared to Matrix A (the original Employer-Business Matrix) using the MSSIM index.

The first series of tests compares the MSSIM index between matrices with the assigned flows of these same matrices. To compare assigned flows of different matrices we have made use of the UK's WebTAG validation criteria and the GEH measure defined in DfT (2012). This states that given a modelled link flow M and an observed traffic count C, the GEH is as follows:

$$GEH = \sqrt{\frac{2(M - C)^2}{M + C}}$$

Figure 6: Definition of GEH from WebTAG guidance

DfT (2012) recommends aiming to achieve a GEH of less than 5 for 85% of links. In place of the C values in the equation above we have used the assigned flows from the original Employer-Business Matrix and for the M values we have used the flows from assigning a test matrix. This is because we are only interested in drawing comparisons with flows from the ground truth matrix. During the network assignment, as well as the Employer-Business matrix being assigned there are other matrices (i.e. other Car, LGV, and HGV) that are also assigned. These fixed matrices are assigned so that the resulting route choice and flows for Employer-Business would be as realistic as possible. These other matrices are fixed, in that they do not change between assignments of the test matrices. Assigning these extra matrices has some implications to the results. Although this was taken account of in our test for GEH (only comparing the resulting flow from the Employer-Business matrix, rather than total flow), this may still have stabilised the GEH results, as the capacity on links, and route choices available will have been affected by the consistent assignment of these fixed matrices.

The results of the first comparisons can be seen below in figure 7.

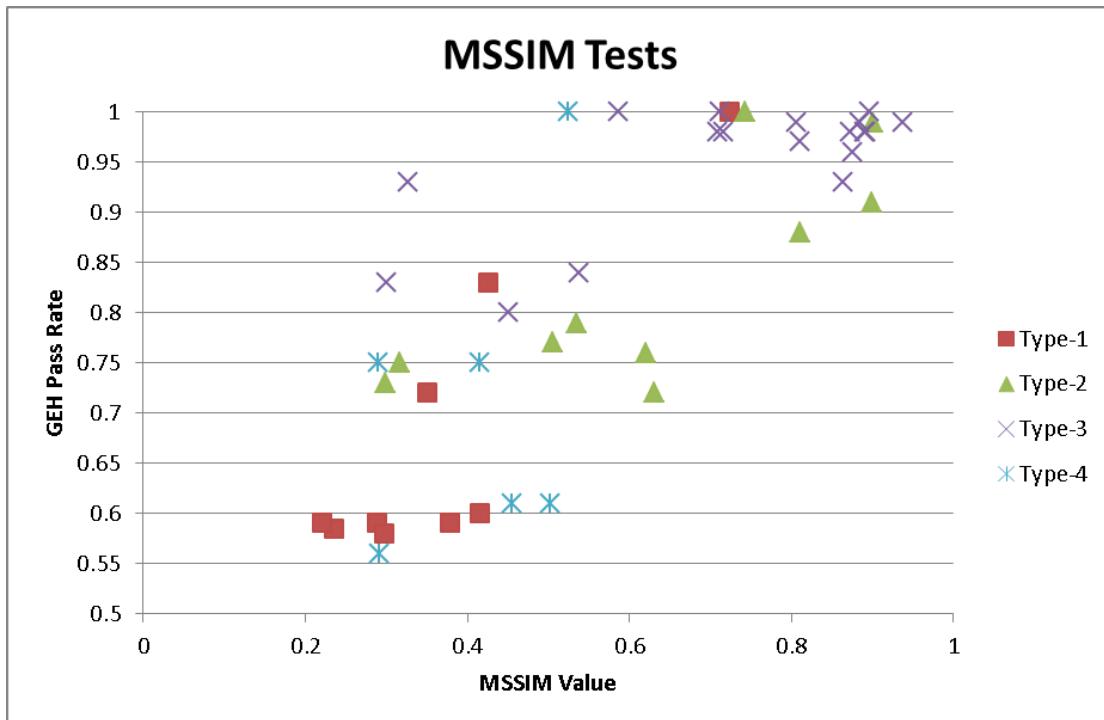


Figure 7: Comparison of GEH pass rate and MSSIM index

Interestingly in our test we did not manage to create a matrix with MSSIM less than 0.2 or with GEH pass rate less than 50% compared to the original (even with 100% random entries) – the latter is due to the stability introduced by the fixed matrices in assignment remaining constant and thus affecting the GEH pass rate (as mentioned earlier). There is a weak correlation between the MSSIM values and the pass rates, but not as strong in the way we had hoped.

It is still possible to infer something from these results.

The worst results for GEH pass rates and MSSIM values came from type 1 matrices, which contain a large random component and tend cluster to the bottom left of the graph.

The type 2 Matrices seem to have a lower GEH per MSSIM compared with the type 3 matrices. Remembering that the type 2 matrices resulted from multiples of the RSI matrix being taken away or added while type 3 matrices have multiples of RSI and GPS matrices taken away or added we can begin to understand this result. Perhaps as the RSI matrix have low cell coverage (i.e. small percentage of non-zero values) then they can make a large change to flows through specific links at the interview sites used to make the RSI matrix, affecting GEH, without making many changes to the overall matrix, and hence MSSIM. On the other hand with type 3 matrices, the GPS matrix has a very large coverage, and so can affect many more OD pairs, affecting the MSSIM, without necessarily affecting the flow as much.

Results for the type 4 matrices, which like type 1 also include a large random component, cluster towards the left of the graph.

Generally lower MSSIM values match to low GEH pass rates and high MSSIM values match to high GEH pass rates. In particular something seems to happen around an MSSIM value of 0.65.

We can split the results roughly into two groups (as shown in figure 8). To the right of 0.65, matrices generally have a GEH pass rate of 85% or above (85% being the normal criteria for passing) and to the left of 0.65 most of the results are below a pass rate of 85%. As shown in the graph we had obtained some results with a low MSSIM value and a high GEH pass rate, which underlines the earlier point in the introduction that assignment of matrices alone is not enough to determine similarity.

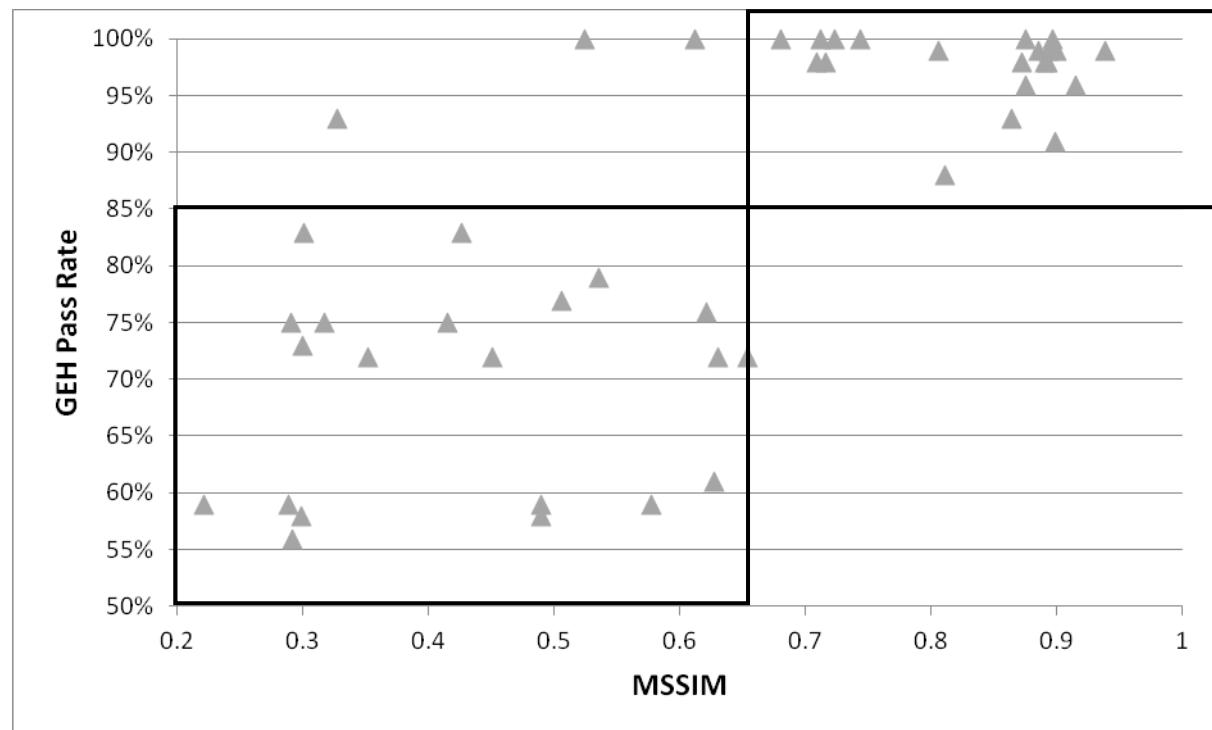


Figure 8: Comparison of GEH pass rate and MSSIM index with splits

To try and get a better understanding of what is happening we decided to carry out a second series of tests on the matrices using the R^2 measure. DfT (2012) section 8.3.13 describes this measure as being recommended before and after matrix estimation with traffic counts. A selection of 8 matrices has been randomly chosen to complete this test. 4 matrices have been chosen with a MSSIM value of less than 0.65 and 4 matrices have been chosen with an MSSIM value of greater than 0.65. These matrices are depicted in Figure 9.

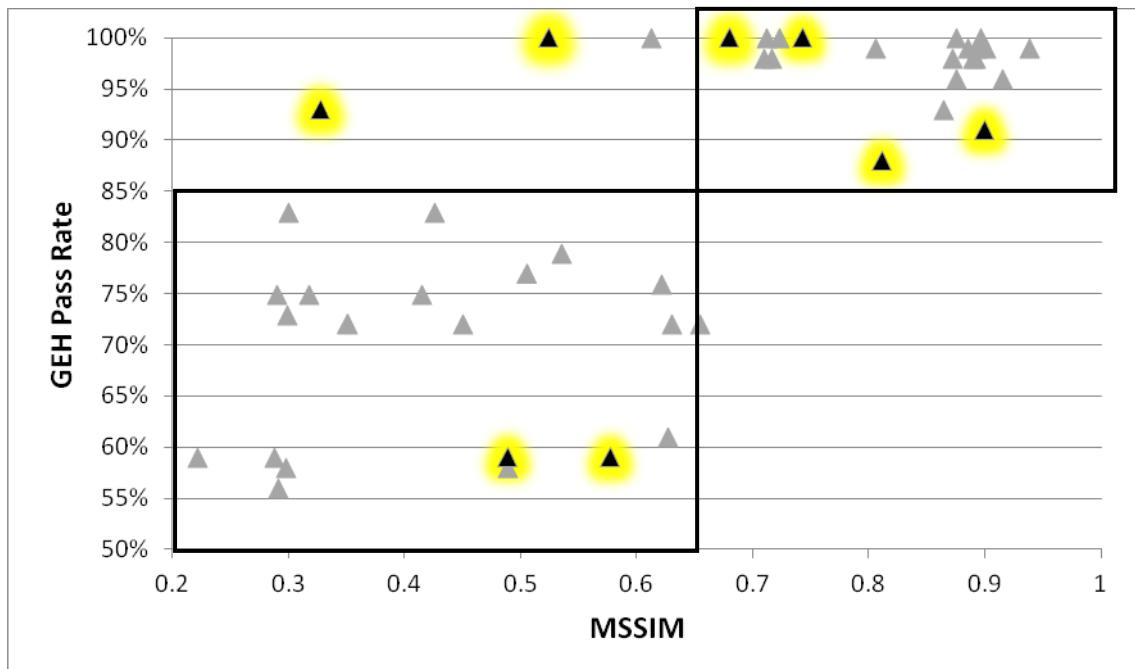


Figure 9: Matrices chosen for R^2 test

The following graph (Fig 10) shows a comparison between the calculated MSSIM values and the related R^2 values. The graph has been ordered from left to right with decreasing MSSIM values.

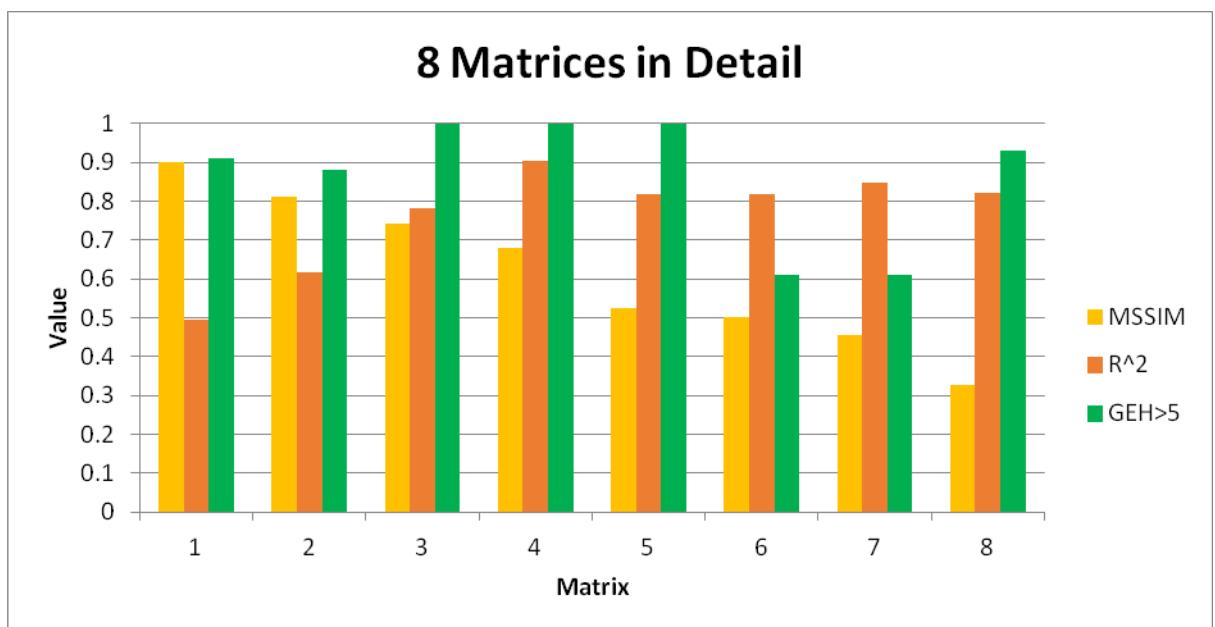


Figure 10: Correlation between calculated MSSIM and R^2 .

Again there seems to be no obvious correlation between the values, however as the MSSIM decreases we seem to get an increase in R^2 which stabilises for the lower MSSIM values. This indicates that the R^2 measure does not pick up the same level of change in matrices as the MSSIM index. The R^2 measure itself relates to individual cell values and not the structural aspects of how cells relate to one another. The results also illustrate that the R^2 measure is not a reliable approximation of GEH pass rate.

To try and understand differences further, a final series of tests was done on 2 types of random matrices. In the figure below 4 matrices (left) are constructed by looping through every cell of the original Employer-Business matrix and multiplying by a random number between 0 and 1 exclusive (a different random number for each cell), a further 4 (right) were created purely from random numbers between 0 and 1 exclusive. As before these test matrices were then scaled to have the same total as the original Employer-Business matrix. The tests GEH, R^2 and MSSIM were then carried out, with results displayed below in figure 11.

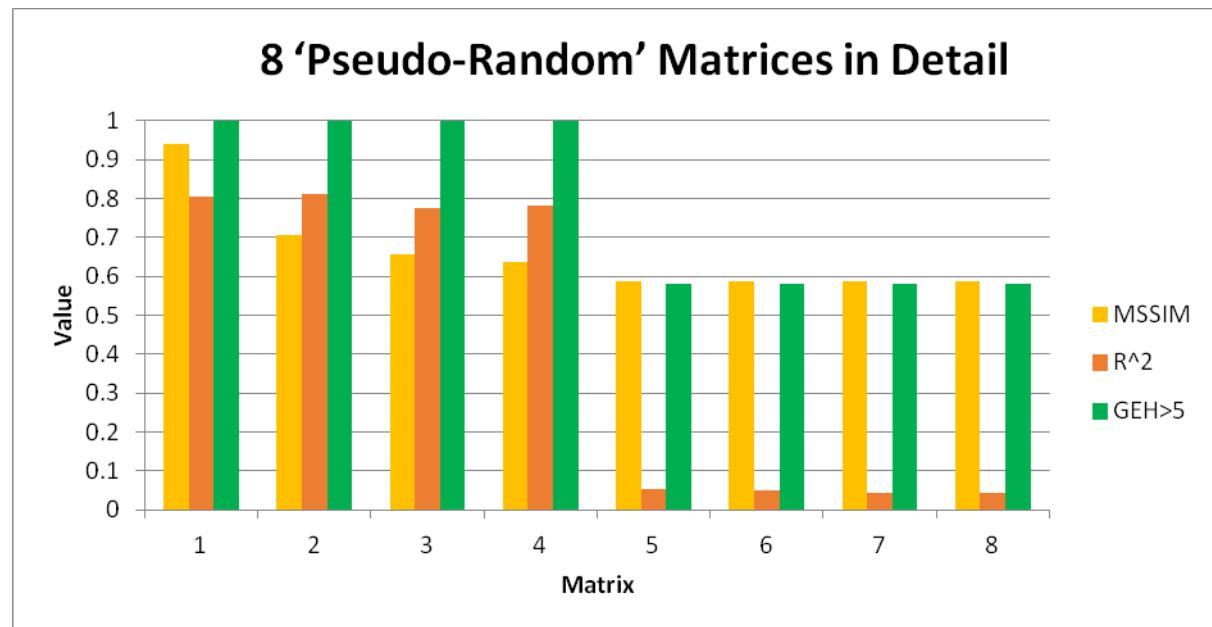


Figure 11: MSSIM and R^2 comparison for B_1 and B_4 matrices.

The 4 purely random matrices consistently have an MSSIM around 0.59, which is higher than some of the values we had before. This is likely because these random matrices were constructed to have no 0 values unlike the previous test matrices. R^2 and GEH seem to give consistent values within both sets of matrices, however the MSSIM has more variation.

There is a clear barrier between the two sets of matrices in the results, with those of a purely random nature scoring much lower on comparisons. These results seem to back up the indication we were getting that GEH was being stabilised around 50%, and that something key happened with MSSIM along the 0.6 value. Looking back to results in figure 7, type 1 and type 4 matrices (those that had a strong purely random element but including some zeros) were consistently to the left of 0.65 and often with

low GEH. This would indicate that our method for calculating stabilising constants needs to be refined even further so that purely random matrices compared with Employer-Business have an MSSIM closer to 0. Nevertheless looking at the matrices in the right of figure 7, combined with the data from figures 10 and 11 it seems clear that the MSSIM is giving us new information about the structure of these matrices that the standard measures do not reveal.

It is clear that further investigation needs to be carried out, however we believe we have shown that there is more information that can be learnt about matrix similarity from the MSSIM index.

5. CONCLUSIONS AND FURTHER WORK

There are many reasons why in practical transport modelling we want to compare OD matrices and determine if they are similar enough:

- Avoiding unnecessary surveys where existing data may suffice
- Comparing matrices derived from cheaper (BIG) data sources with those derived by traditional means, to confirm their acceptability
- Comparing the quality of synthesized matrices with observed ones, to strengthen our belief in the underlying model
- Comparing the difference between prior and estimated matrix, when refining OD matrices by using counts

And because there are so many reasons why we need to compare matrices we feel more work is required to ensure we understand how the MSSIM indicator can be applied best in practice.

We have used the MSSIM indicator to compare artificially constructed matrices with a ground truth matrix, and compared those results with results from other comparison techniques. Based on our tests, we believe that the MSSIM indicator should be refined further when used on transport matrices, as they are different from the pixel matrices for which the indicator was originally developed.

We believe that care needs to be taken in calculating constants used within each step of the MSSIM calculation paying attention to known relationships within the field of image comparisons (as we have done) but also with the result of comparisons with a random matrix.

While Djukic et al (2013) pointed out that zoning systems needed to be consistent in both matrices being compared; we feel also the zone numbering needs to be refined so that consecutive numbers always result in adjacent zones. For various modelling reasons this isn't always the case but before using an MSSIM measure in practice this should be reviewed.

Another possible avenue for further work is to consider weighting some windows (i.e. some area to area movements) as being more or less important, to produce a weighted-Mean-SSIM. This situation may occur when the model the matrices are being built for is more interested in movements into and out of eg a city-centre, and less concerned with surrounding traffic.

The MSSIM index may not be the only comparator that captures the spatial basis of OD matrices; we have come across others, for example the method discussed by

Turner et al (1989). There appears to be a richer literature out there than has been recognised by the transport modelling profession.

Concluding, we believe these results are encouraging enough to continue our investigations, perhaps using different models. We have identified a number of further strands of investigation and testing which should help to provide solid guidance on the application of MSSIM in an OD context that would add value over and above currently used comparators.

6. Acknowledgements

This paper is based on a number of tests with the PRISM strategic transport model for the West Midlands region in the UK. None of the test results reflects an actual model application for either policy or infrastructure investment. We are grateful for the PRISM Management Group's permission to use PRISM for these investigations. The results and their interpretation are solely the responsibility of the authors and cannot be attributed to their employers or the model owners.

7. References

- Calabrese, F; Diao, M; Di Lorenzo, G., Ferreira Jr, J; and Ratti, C (2013) **Understanding individual mobility patterns from urban sensing data: A mobile phone trace example**, Transportation Research Part C: Emerging Technologies, Volume 26, Pages 301-313, <http://www.sciencedirect.com/science/article/pii/S0968090X12001192>, January 2013.
- DfT (2012) WebTAG 3.19 Highway Assignment Modelling, www.dft.gov.uk/webtag, August 2012.
- Djukic, T; Hoogendoorn, SP; and Lint, JWC van (2013) **Reliability assessment of dynamic OD estimation methods based on structural similarity index**, presented at TRB, January 2013
- Potter, A; and Birks, S (2011) **Are Road Side Interviews (RSIs) still what we think they are?**, presented at 9th Transport Practitioners Meeting, Liverpool, July 2011.
- Turner, MG; O'Neill, RV; Gardner, RH; and Milne, BT (1989), **Effects of changing spatial scale on the analysis of landscape pattern**, Landscape Ecology vol. 3 nos. 3/4 pp 153-162
- van Vuren, T; and Carey, C (2011) **Building Practical Origin-Destination (OD/Trip) Matrices from Automatically Collected GPS data**, European Transport Conference <http://abstracts.aetransport.org/paper/index/id/3737/confid/17> Glasgow, October 2011.
- Van Vuren, T; Gordon, A; Daly, A; Fox, J; and Rohr, C (2004) **PRISM: Modelling 21st Century Transport Policies in the West Midlands Region**, Proceedings of European Transport Conference, Strasbourg, Seminar on Applied Transport Methods, October 2004.

Zhou, W; Bovik, AC; Sheikh, HR; and Simoncelli, EP (2004) **Image quality assessment: from error visibility to structural similarity**, *Image Processing, IEEE Transactions on*, 13(4): p. 600-612, April 2004.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/267098616>

Reliability assessment of dynamic OD estimation methods based on structural similarity index

Conference Paper · January 2013

DOI: 10.13140/RG.2.1.4174.1929

CITATIONS

4

READS

148

1 author:



Tamara Djukic
TSS - Transport Simulation Systems, Barcelona, Spain

21 PUBLICATIONS 110 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



SETA Mobility [View project](#)

1 **TRB 13-XXXX**

2 **Reliability assessment of dynamic OD estimation methods based on structural similarity**
3 **index**

4 Tamara Djukic, Ms.C. (corresponding author)
5 T.Djukic@tudelft.nl

6 Transportation and Planning Department
7 Faculty of Civil Engineering and Geosciences
8 Delft University of Technology
9 Stevinweg 1, P.O. Box 5048,
10 2600 GA Delft, the Netherlands

11

12 Professor S.P. Hoogendoorn, Ph.D., Ms.C.
13 S.P.Hoogendoorn@tudelft.nl

14 Transportation and Planning Department
15 Faculty of Civil Engineering and Geosciences
16 Delft University of Technology
17 Stevinweg 1, P.O. Box 5048,
18 2600 GA Delft, the Netherlands

19

20 J.W.C. van Lint, Ph.D., Ms.C.
21 J.W.C.vanLint@tudelft.nl

22 Transportation and Planning Department
23 Faculty of Civil Engineering and Geosciences
24 Delft University of Technology
25 Stevinweg 1, P.O. Box 5048,
26 2600 GA Delft, the Netherlands

27

28

29 Word count:
30 Words (including abstract and references): 6132
31 Tables and figures: 8 * 250 = 2000
32 Total: = 8132

33

34

35

36

37

38 *Submitted for publication and presentation for the 92nd meeting of the Transportation Research Board, 13-17*
39 *January 2013, Washington D.C.*

40

1 Abstract

2 The statistical measures used for quality assessment of estimated OD matrices typically quantify the difference
3 between the estimated OD matrix and available true/reference OD matrix. Although the underlying rational
4 makes sense intuitively, the actual statistical measures in literature, such as MSE, do not capture the most
5 important aspect: the structural similarity of the estimated and reference OD matrix. In this paper we propose a
6 new quality measure that does incorporate such a term, so-called Structural SIMilarity (SSIM) index.

7 In this paper we explore the application perspectives of SSIM index for this purpose. First, we investigate the
8 properties of SSIM index compared to some statistical measures. Then, we show how SSIM index can be used as
9 additional performance measure for benchmarking the dynamic OD estimation methods. More over, we provide
10 insight into how SSIM index can be used further as a new performance function to estimate dynamic OD
11 matrices.

1 INTRODUCTION

2 The ex-post and ex-ante evaluation of traffic and demand management measures, and transport policy measures
 3 requires a very high quality of the traffic variables. In particular, important input to the models used for assessing
 4 such measures are OD matrices. The important role of OD demand has resulted in a variety of mathematical
 5 approaches to estimate and predict dynamic OD matrices, such as the Generalized least square models (1-3), the
 6 Maximum Entropy models (4, 5), the Maximum likelihood (6), Bayesian inference model (7, 8) and the Kalman
 7 filter models (9, 10). Thus, there is increased need for quantitative objective measures to assess the quality of
 8 existing OD estimation methods to pinpoint their strengths and weakness and applicability and validity under
 9 different circumstances.

10 In this paper we will focus on an important and often-overlooked aspect in the benchmark of different OD
 11 estimation methods, namely the selection of an appropriate measure or a set of criteria that can be used to
 12 compute the quality of estimated OD matrices. Few studies have focused on evaluation of the reliability and
 13 accuracy of the estimated OD matrices in absence of ground truth OD matrix (1, 11), and on available ground
 14 truth OD matrix such as (12, 13). A number of statistical measures have been proposed and used in literature to
 15 evaluate the quality of an OD estimator, such as root mean square error and mean percentage error. However, the
 16 basic principal of these performance indices is that they are expressed as deviations in terms of OD demand or
 17 traffic counts in respect to ground truth data. These statistical measures are widely used because they are simple
 18 to calculate and have clear meanings. It is worth noting that these statistical measures are not very well matched
 19 to capture the structural patterns between estimated and ground truth OD pairs. In the ideal case, the estimated
 20 OD matrix reflects the OD matrix which is very close to the actual OD matrix, in the sense that it has a similar
 21 structure, for example in terms of distribution of trips over destinations, and the trip length distribution.

22 The OD matrices may be determined from different sources of information (e.g. land-use models, travel surveys)
 23 using different methods, which represent a common spatial and temporal behavior of travelers (e.g. choice of
 24 destination, departure time, mode choice). For example, the gravity model illustrates the macroscopic
 25 relationships between places (say homes and workplaces). It has long been postulated that the interaction
 26 between two locations declines with increasing (distance, time, and cost) between them, but at the same time, the
 27 interaction is positively associated with the amount of activity at each location. These rules yield the structural
 28 correlation between OD pairs. Such OD matrices are thus highly structured and stem from the combination of
 29 various kinds of information, such as OD matrix structure and the correlation between OD pairs.

30 Statistical measures that ignore this spatial correlation between OD pairs in OD matrices may fail to provide
 31 effective and accurate quality measures. To show this, we will use the mean square error (MSE) as an example.
 32 In FIGURE 1 (b) and (c), we compare the two estimated OD matrices using two different OD estimation
 33 methods with the available “ground truth” OD matrix (a). For better visual examination of the structure in the
 34 OD matrices, where the origin zones are given in rows and destinations in columns, they are represented as
 35 images where the number of trips per OD pair is used as index into the colormap that determine the color for
 36 each OD pair. Then, the number of trips in $X_{l,l}$ (FIGURE 1 (a)) represents the color of the OD pair at position as
 37 an index into the colormap. For example, if $X_{l,l} = k$; then the color of OD pair $X_{l,l}$ is the color represented by
 38 row k of the color map. In Figure 1, the OD pair with demand of 20 trips has a light yellow color, and OD pair
 39 with demand of 400 trips has a dark green color. The MSE between ground truth OD matrix and both of the
 40 estimated OD matrices are exactly the same. However, the visual examination of the two estimated OD matrices
 41 clearly indicate that they capture different structural patterns. Therefore, such measures are only designed to find
 42 the distances between a pair of attributes in a data set or overall distance amongst all data. They are unable to
 43 find and distinguish different correlation structures in data, and cannot provide an in depth explanation of the
 44 patterns in OD demand.

45 One possible solution approach to tackle such a problem is to propose a new quality assessment measure that
 46 will incorporate the structural correlation between OD pairs. In this paper we propose the application of the so-
 47 called Structural Similarity (SSIM) Index (14), which is a measure to quantify the similarity of two data sets,
 48 taking into account their structure that is expressed through correlation between OD pairs. Origination from
 49 image processing and analysis, the SSIM approach was motivated by the observation that image signals are
 50 highly structured, meaning that samples of natural image signals have strong dependencies. These dependencies
 51 carry important information about the structures of the objects in the visual scene.

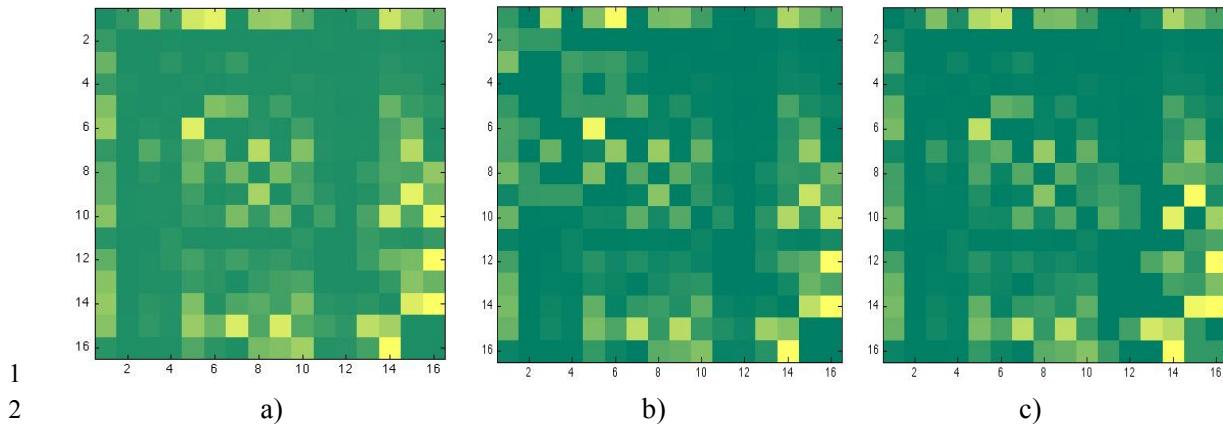


FIGURE 1 Comparison of patterns in reference and estimated OD matrices: a) “ground truth” OD matrix; b and c): estimated OD matrices that have the same MSE with respect to the reference OD matrix, but different structural patterns

In this paper, we propose using the SSIM index as a new quality metric for sensitivity assessment of dynamic OD estimation methods, and study its applicability. The new metric would enable researchers and practitioners better insight into how to assess the quality of the estimated OD matrix, and to make a strict conclusion about the quality and efficiency of OD estimation methods. More specifically, we argue that traffic engineers have rethink whether the statistical measures such as MSE or RMSE are the most useful criteria of choice in their comparative studies and applications. Also, we explore application potentials of SSIM index as a performance function to estimate the OD demand.

The paper is organized as follows. In the first part of the paper we will outline the theoretical background of SSIM and explain its main properties. Next we will present the insensitivity of the statistical measure MSE to identify and evaluate the structural pattern in OD matrices. In the second part of the paper we will first define the experimental data set and benchmark framework to assess the reliability of the several OD estimation methods. Next, we assess the quality of estimated OD matrices with different statistical measures and SSIM index. The paper closes with a discussion on further application perspectives of SSIM index in estimation and prediction of OD demand and further research.

QUALITY ASSESSMENT BASED ON STRUCTURAL SIMILARITY INDEX

The idea of structural similarity

The statistical measures used for quality assessment of estimated OD matrices typically measure the difference between the estimated OD matrix and available true/reference OD matrix. Although the underlying rational makes sense intuitively, the actual statistical measures in literature, such as MSE, do not encapsulate the most important aspect: the structural similarity of the estimated and reference OD matrix. By structural similarity, we mean that the *spatial and temporal behavior of travelers reflected in OD trip patterns have a strong spatial and temporal correlation reflected by the OD pairs*. In the ideal case, the estimated OD matrix reflects the OD matrix which is very close to the actual OD matrix, in the sense that it has a similar structure, for example in terms of distribution of trips over destinations, and the trip length distribution. The statistical measures that ignore correlation in OD matrix data may fail to provide effective and accurate quality measures. Therefore, the key idea in our approach is to define a quality metric, in such way that the structural pattern in estimated OD matrix is quantified adequately.

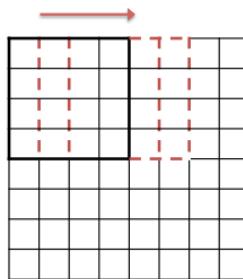
We propose the application of the Structural SIMilarity (SSIM) index (14) that is typically used as a method for measuring the similarity between two images, based on the degradation of the structural information in one image compared to the reference image. In general, images are highly structured: their pixels exhibit strong dependencies, especially when they are close to each other, and these dependencies carry important information about the structure of the objects in the visual scene. The key idea is now, that if we represent the OD demand in the matrix form, the OD pairs can be seen as pixels in image that exhibit strong dependencies as well. Hence, the SSIM seems to be a good measure to compare OD matrices as well.

In next subsection we define and explain in detail the Structural Similarity (SSIM) index and its application to OD matrices. The measure ensures that the amount of structural information in reference OD matrix is preserved in estimated OD matrix. We will demonstrate and compare the suitability of the MSE and SSIM index under different scenarios, to illustrate the key benefits of the SSIM.

1 **The Structural Similarity Index**

2 The Structural SIMilarity (SSIM) index that is typically used as a method for measuring the similarity between
 3 two images, based on the degradation of the structural information in one image compared to the reference one.
 4 For instance, if we assume that any prior OD matrix or available true OD matrix contains the best pattern
 5 information to our knowledge, then the SSIM index can be viewed as an indication of the quality of the
 6 estimated OD matrix compared to the prior OD matrix or true OD matrix, respectively. Note that Wang et al.
 7 (14) introduced the SSIM index in the context of similarity measure to explore and compare the structural
 8 information between images, a problem that is in a many respects similar to exploring the structures in OD
 9 matrices.

10 To explain the metric, we follow a similar rationale as in (14). Assume that the OD demand for a particular time
 11 interval t is defined in form of the matrix where the rows of matrix represents the origins i , $i=1,2,\dots,I$, and
 12 columns represent the destinations j , with $j=1,2,\dots,J$, of trips. To evaluate the structural similarity between two
 13 OD matrices, let $d = \{d_n | n=1,2,\dots,N\}$ and $\hat{d} = \{\hat{d}_n | n=1,2,\dots,N\}$ be two vectors that have been extracted from
 14 the same spatial location from reference OD matrix $D = \{d_{ij}\}$ and estimated OD matrix $\hat{D} = \{\hat{d}_{ij}\}$, as is shown in
 15 FIGURE 2. The SSIM index is computed within a local $N \times N$ square box, which moves cell-by-cell over entire
 16 OD matrix.



17

18 **FIGURE 2 Computation of local SSIM index per sliding $N \times N$ square box**

19 The most general form of the metric that is used to measure the structural similarity between two vectors d and
 20 \hat{d} consist of *three main components* and is given as

$$21 \quad SSIM(d, \hat{d}) = [l(d, \hat{d})^\alpha][c(d, \hat{d})^\beta][s(d, \hat{d})^\gamma] \quad (1)$$

22 In this equation l is used as a distance metric to compare the mean values of the two matrices, c compares the
 23 standard deviation of the matrices, and finally s compares the matrix structure. Now let us look at each of the
 24 components in detail.

25 As said, the term $l(d, \hat{d})$ compares the mean values of the vectors d and \hat{d} , $\mu_d = \frac{1}{N} \sum_{n=1}^N d_n$, and is defined by
 26 the following expression

$$27 \quad l(d, \hat{d}) = \frac{2\mu_d\mu_{\hat{d}} + C_1}{\mu_d^2 + \mu_{\hat{d}}^2 + C_1} \quad (2)$$

28 The term $c(d, \hat{d})$ compares the standard deviation (the square root of variances) of the vectors,

$$29 \quad \sigma_d = \sqrt{\frac{1}{N} \sum_{n=1}^N (d_n - \mu_d)^2}, \text{ and takes the similar form given by}$$

$$30 \quad c(d, \hat{d}) = \frac{2\sigma_d\sigma_{\hat{d}} + C_2}{\sigma_d^2 + \sigma_{\hat{d}}^2 + C_2} \quad (3)$$

31 Finally, the structure term $s(d, \hat{d})$ is defined as the correlation (inner product) between the normalized OD
 32 demand vectors d and \hat{d} , $d - \mu_d / \sigma_d$ and $\hat{d} - \mu_{\hat{d}} / \sigma_{\hat{d}}$, and is effective measure to quantify the structural
 33 similarity. This is equivalent to the correlation coefficient which measures the degree of linear correlation

1 between vectors d and \hat{d} . Geometrically, $s(d, \hat{d})$ correspond to the cosine of the angle between two vectors
 2 $d - \mu_d$ and $\hat{d} - \mu_{\hat{d}}$, independent of the lengths of these vectors.

3 Thus, the structure term $s(d, \hat{d})$ is define as follows:

$$4 \quad s(d, \hat{d}) = \frac{\sigma_{d\hat{d}} + C_3}{\sigma_d \sigma_{\hat{d}} + C_3} \quad (4)$$

$$5 \quad \text{where } \sigma_{d\hat{d}} = \frac{1}{N-1} \sum_{n=1}^N (d_n - \mu_d)(\hat{d}_n - \mu_{\hat{d}}).$$

6 The structure term $s(d, \hat{d})$ reflects the similarity between two OD demand vectors – it equals one if and only if
 7 the structures of the two demand vectors being compared are exactly the same.

8 The constants C_1, C_2, C_3 in Eqn. (2), (3) and (4) are used to stabilize the metric for the case where the means and
 9 variances become close to zero. The parameters in Eqn. (1), $\alpha > 0, \beta > 0$ and $\gamma > 0$, are used to adjust the relative
 10 importance of the three components. In order to simplify the expression, as is recommended in (14) we set
 11 $\alpha = \beta = \gamma = 1$, and $C_3 = C_2 / 2$. This results in a final form of the SSIM index between two OD matrices

$$12 \quad SSIM(d, \hat{d}) = \frac{(2\mu_d \mu_{\hat{d}} + C_1)(2\sigma_{d\hat{d}} + C_2)}{(\mu_d^2 + \mu_{\hat{d}}^2 + C_1)(\sigma_d^2 + \sigma_{\hat{d}}^2 + C_2)} \quad (5)$$

13 Finally, at each step we calculate the local statistics (μ_d, σ_d and $\sigma_{d\hat{d}}$) and SSIM index within the square box. The
 14 overall quality measure of the entire estimated OD matrix is given as a mean of the local SSIM indexes as

$$15 \quad MSSIM(D, \hat{D}) = \frac{1}{M} \sum_{m=1}^M SSIM(d_m, \hat{d}_m) \quad (6)$$

16 where D and \hat{D} are the reference and the estimated OD matrices, respectively, d_m and \hat{d}_m are the OD matrix
 17 contents at the m^{th} local square box; and M is the number of local square boxes of the entire OD matrix.
 18 The SSIM index is symmetric: $SSIM(x, y) = SSIM(y, x)$, so that two OD matrices being compared give the
 19 same index value regardless of their ordering. It is also bounded: $-1 \leq SSIM(d, \hat{d}) \leq 1$, achieving the maximum
 20 value $SSIM(d, \hat{d}) = 1$ if and only if $d = \hat{d}$ and value $SSIM(d, \hat{d}) = 0$ represent that estimated OD matrix does not
 21 capture the spatial correlation between OD pairs as is given in reference OD matrix. Also, the order of origins in
 22 rows and destinations in columns in reference and estimated OD matrix must be the same. Otherwise, the SSIM
 23 index will give a bias result.

24 Although the computation of SSIM index seems a more complex then other statistical measures, it is efficient
 25 method to assess the quality of estimated OD matrices.

26 The reliability of the MSE error and SSIM index

27 To illustrate the properties and the advantages of the SSIM index over statistical measures that are often used in
 28 sensitivity analysis of OD estimation methods or in the optimization process, we examine the following several
 29 scenarios that reflect the importance of assumptions that an engineer is making when she/he decides to use the
 30 MSE. For better visual examination of the structure in the OD matrices, they are represented as images where the
 31 values of the OD flows are used as indices into the colormap that determine the color for each OD pair.

32 In this example, we will show how use of the MSE error is not sufficient for researchers and practitioners to
 33 pinpoint the strengths and weakness of different OD estimation methods. For example, we want to assess the
 34 quality of estimated OD matrices from two different OD estimation methods (FIGURE 3 (b) and (c)) in respect
 35 to the reference OD matrix (e.g. ground truth OD), FIGURE 3(a). In the FIGURE 3, the first perturbed OD
 36 matrix (b) was obtained by adding a constant value to all cells in the “ground truth” OD matrix (a) and reflects
 37 the estimated OD matrix from one model. The second perturbed OD matrix (c) was generated by the same
 38 method, except that the signs of the constant were randomly chosen to be positive or negative, to reflect the
 39 estimated OD matrix from model two.

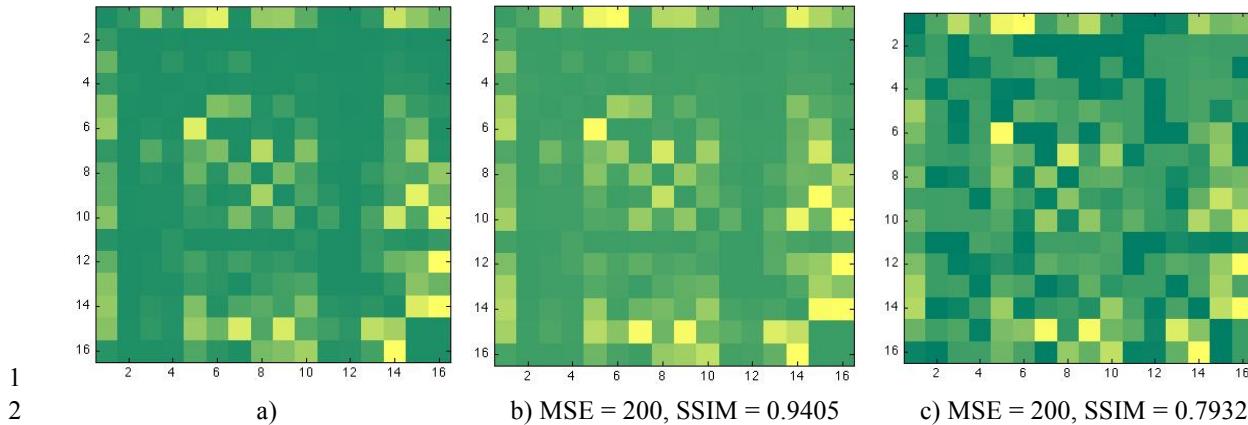


FIGURE 3 Comparison of patterns in reference and estimated OD matrices: a) “ground truth” OD matrix; b and c): estimated OD matrices that have the same MSE with respect to the reference OD matrix, but different structural patterns

The visual difference of the two estimated OD matrices is clearly different. Via visual inspection is it quite clear that the first perturbed matrix resembles the “ground truth” matrix much better than the second one. Yet, the MSE ignores the effect of signs and reports the same value for both perturbed OD matrices while the SSIM index captures the structural difference in the matrices. This result indicates that if we want to compare the performance of the two OD estimation methods and if we look only at the MSE error, we could conclude that both methods perform similarly. Contrary, if we look at the SSIM index values, we can conclude that the method one performs better than other.

Let us consider another example, where the estimated OD matrices have different MSE values but very similar patterns. In the FIGURE 4, the both perturbed OD matrices (b) and (c) were obtained by adding an independent Gaussian noise to the “ground truth” OD matrix (a).

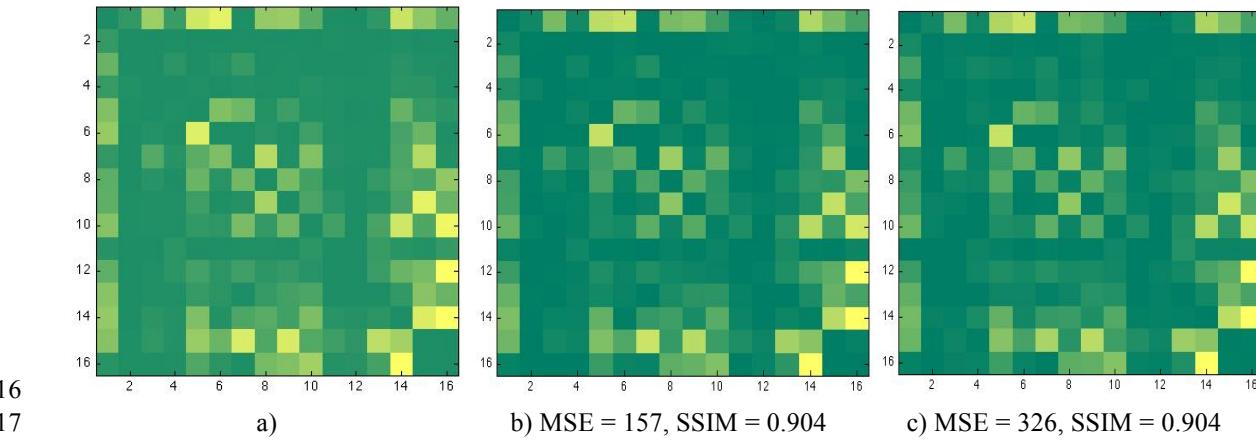


FIGURE 4 Comparison of patterns in reference and estimated OD matrices: a) “ground truth” OD matrix; b and c): estimated OD matrices that have the same SSIM with respect to the reference OD matrix, but different MSE values.

Apparently, OD matrices that undergo small geometrical modifications have a very large MSE values relative to the “ground truth” OD matrix, yet show a negligible loss of perceived quality. In this case, results indicate that the method with estimated OD matrix (b) performs better than method estimated OD matrix (c), which is consistent with MSE value. Therefore, the SSIM index can be used as additional information (goodness of fit measure) in performance assessment of OD estimation methods.

The examples presented in FIGURE 3 and FIGURE 4, indicate that simplified OD estimation models with the associated assumptions about violated or ignored structural correlation may fail to provide efficient and accurate estimates of OD demand. Cascetta et al. (15) have shown that most OD demand estimators can be obtained by solving a constrained optimization problem, where distance functions are defined by considering the distance measures between unknown OD demand and the prior OD demand. These distance measures depend on particular estimation framework, such as maximum likelihood, generalized least square, Bayesian inference, et

cetera. For example, the generalized least square function (2, 15, 16) is equal to the Mahanalobis distance measure which is often turned into Euclidean distance measure under commonly used simplifying assumptions, such as the variance covariance matrix is equal to the identity matrix or it is diagonal. The models that aim to minimize the Euclidean distance between the prior OD and estimated OD matrix can provide unrealistic estimation results due to fact that too little information is taken from prior OD matrix.

In the next section we will show the application perspectives of the SSIM index as an additional performance measure to statistical measures in benchmark study. In the rest of the paper, we provide insight into how SSIM index can be used further as a new performance function to estimate dynamic OD matrices.

BENCHMARKING FRAMEWORK: PERFORMANCE MEASURES AND CONSIDERED SCENARIOS

In this section, we assess the performance of two OD estimation methods with a least square modeling approach for solving the OD estimation problem. In this study, we consider offline estimation without prediction of the future OD flows. The use of least square approaches to solve dynamic OD estimation problem has been proposed by Cascetta (15). In this study we consider two solution algorithms for the least square problem, Kalman filter algorithm proposed by Ashok and Ben Akiva (9) and LSQR algorithm proposed by Bierlaire (3) based on proposal of using deviations between historical and actual OD flows as state variables.

Our benchmark framework is as a simulation-based approach where the OD estimator is considered as black box, providing a certain outcome (i.e. the OD matrix estimate) given certain input. We applied the stratified sampling method (Latin hypercube (LHC) method from (17) that provides a computationally much more efficient alternative to random (Monte Carlo) sampling for estimating the conditional distribution. A requirement for applying this method is that prior distributions for the inputs need to be known. This feature allows us to perform a comprehensive sensitivity analysis with different assumptions on input data that exhibit certain properties of the OD estimation method. We refer to reader to the original paper on LHC method (17) for more details on the approach.

Given such an efficient sampling method exists, we can define a set of input scenarios varying in terms of network topology, traffic conditions, and data availability and quality. First, we will define the set of performance measures to assess the sensitivity of OD estimation methods. Then, in the next subsection we will provide some more detail on the method used to sample the scenarios. For more detail explanation of benchmark framework we refer to the paper (13).

Performance measures

The choice of performance criteria plays an important role in benchmarking OD estimation methods under input uncertainty. From a macroscopic viewpoint, there are two kinds of performance measures that relate to the goodness of fit of the OD estimation algorithm, and to the computational efficiency. In this paper, we are interested in defining a set of statistical measures that allow us to compare the respective performance of the tested OD estimators with our new proposed performance measure, SSIM index. The output will provide us insights into the relative merit of the various performance measures.

The comparison of estimated and true OD demand can be assessed by traditional statistical measures (e.g. error measures such as RMSE, MSE, etc.). To examine the accuracy and robustness of OD estimation methods, we use the following set of well-known performance measures:

- The root mean square error (RMSE) is chosen as an estimate of the variance present in the estimated results (as the average magnitude that the estimate will deviate from the true value). Let \hat{d}_n denote the n^{th} simulation, for $n = 1, 2, \dots, N$, estimated OD matrix for time interval t

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N (\hat{d}_n - d_n)^2} \quad (7)$$

Since the scale of OD flows considerably vary, where OD flows with larger volume might dominate comparison, we applied in addition the set of following relative error measures:

- Normalized root mean square error (NRMSE):

$$NRMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N (\hat{D}_n - D_n)^2} / \sum_{n=1}^N D_n \quad (8)$$

- Mean percentage error (MPE):

$$MPE = \frac{1}{N} \sum_{n=1}^N \left[\frac{\hat{D}_n - D_n}{D_n} \right] \quad (9)$$

The relative measures, which are unit free, eliminate the influence of the input data scale by calculating the error score relative to either the true measurement, or alternatively, the score of another estimation algorithm. In addition, the mean percentage error (MPE) indicates the existence of systematic under or over estimation in the estimated data.

We have seen in the previous section that we can use SSIM index to quantify how well the estimated OD matrices capture the pattern structure of the true OD matrix. Therefore, we include the SSIM index as a quality metric to assess the reliability of OD estimation methods.

Experiment: Synthetic data

In this section numerical experiments are presented to evaluate the performance of the OD estimation methods and solution algorithms in the terms of uncertainty in input data on academic network example and simulated data. First, the OD estimation methods used in the assessment will be described briefly. Then, we will describe the input data and assessment scenarios used in this paper. Finally, we will analyze the performance of OD estimation methods based on defined performance criteria.

Network topology

Prior to methods performance evaluation, we define a simplified synthetic network that consists of 5 nodes, 25 OD pairs with a single route between them and 16 corresponding links (FIGURE 5). This network was chosen because we could model and assume the availability of the “true” OD demand and assignment matrix to the analysts. The “true” assignment matrix is arbitrary derived assuming network is congested.

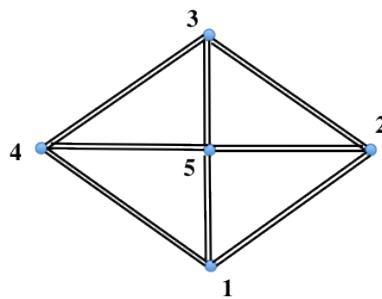


FIGURE 5 The synthetic network

Considered scenarios and results

A major problem with assessing the performance of OD estimators is to obtain meaningful evaluations of the algorithms results and performance, because the ground truth OD data are generally not available for comparison when working with real data. One solution is to use simulated OD demand data, where underlying sources and phenomena are known. To generate a simulated a priory OD matrix dataset for this purpose requires us to define an arbitrary model for OD demand generation, which represents a common spatial and temporal behavior of travelers.

Scenario 1:

By implementing the LHC method described in previous section, we can simulate the fact that the dynamic prior OD matrix denoted as \tilde{d} with elements $\tilde{d}_{i,j}$ may contain errors. This scenario is based on the assumption that the

prior OD matrix is the best estimate of the mean of the dynamic OD matrices. In this case $\tilde{d}_{i,t}$ is varied by adding uniformly random components to the ground truth OD matrix, with standard deviation of 20% representing the difference between the smoothed historical estimate and the particular daily realization:

$$\tilde{d}_{i,t} = d_{i,t}[0.8 + 0.4u_{i,t}] \quad (9)$$

where $u_{i,t} \sim U[0,1]$, and the $d_{i,t}$ the assumed ground truth OD demand. Next to the prior OD, the available traffic data presents an important source of information. We assume that the true traffic counts $c_{i,t}$ resulting from the assignment of the true OD demand are available on all detectors and has been randomly perturbed to obtain the traffic count data. In this case $c_{i,t}$ is varied by adding uniformly random components to the true traffic counts, with standard deviation of 5%, 10% and 20%:

$$\begin{aligned} \tilde{c}_{i,t} &= c_{i,t}[(1-\delta) + 2u_{i,t}\delta] \\ u_{i,t} &\sim U[0,1]; \delta : [0.05, 0.10, 0.20] \end{aligned} \quad (10)$$

The results for Scenario 1 are presented in FIGURE 6. We can observe that the prior OD matrices with randomly distributed perturbations are improved by a fairly consistent percentage. This implies that the estimation accuracy that can be obtained is directly proportional to the random error in the prior OD matrix. Also, we can see from MPE value that both methods slightly systematically underestimate the OD demand. However, the high SSIM index values ($SSIM_{KF} = 0.9816$ and $SSIM_{LSQR} = 0.9803$) indicate that structural patterns in estimated OD matrices are preserved and very close to true OD matrix.

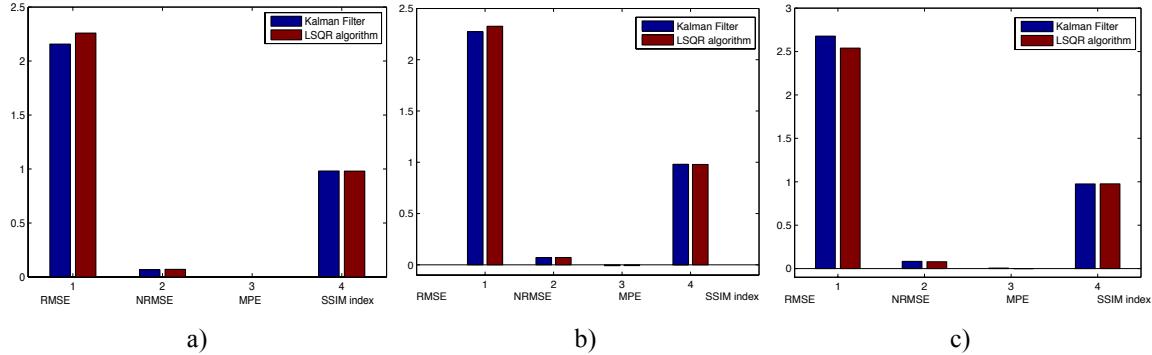


FIGURE 6 Estimation results for random prior OD matrix with: a) traffic counts error $\sigma 2\%$; b) traffic counts error $\sigma 5\%$; and c) traffic counts error $\sigma 10\%$

Scenario 2:

This scenario addresses situations where the prior OD demand might contain other, *structural* errors besides the random daily fluctuations. The demand per each origin over destinations is generated from positively and negatively skewed mean values of distribution from random demand scenario defined in two prior OD data sets:

$$\tilde{d}_{i,t} = d_{i,t}[0.9 + 0.4u_{i,t}] \quad (11)$$

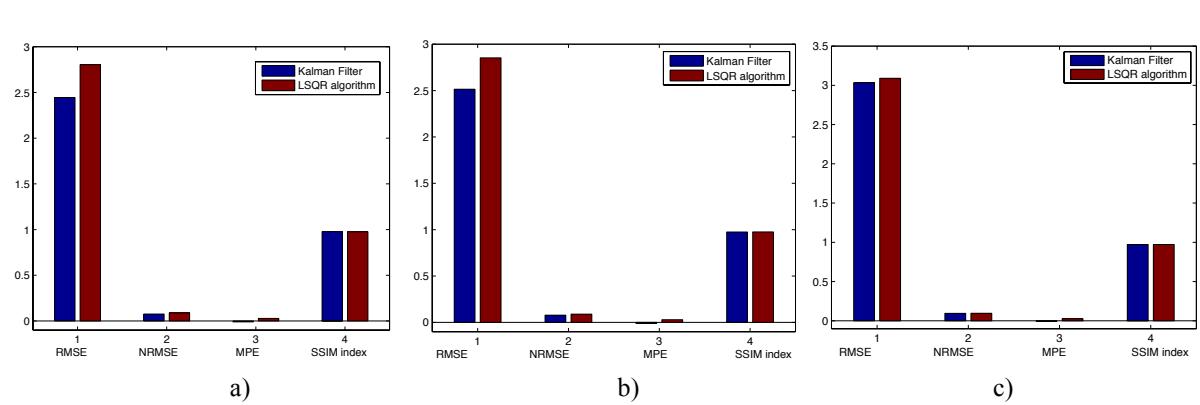
and

$$\tilde{d}_{i,t} = d_{i,t}[0.7 + 0.4u_{i,t}] \quad (12)$$

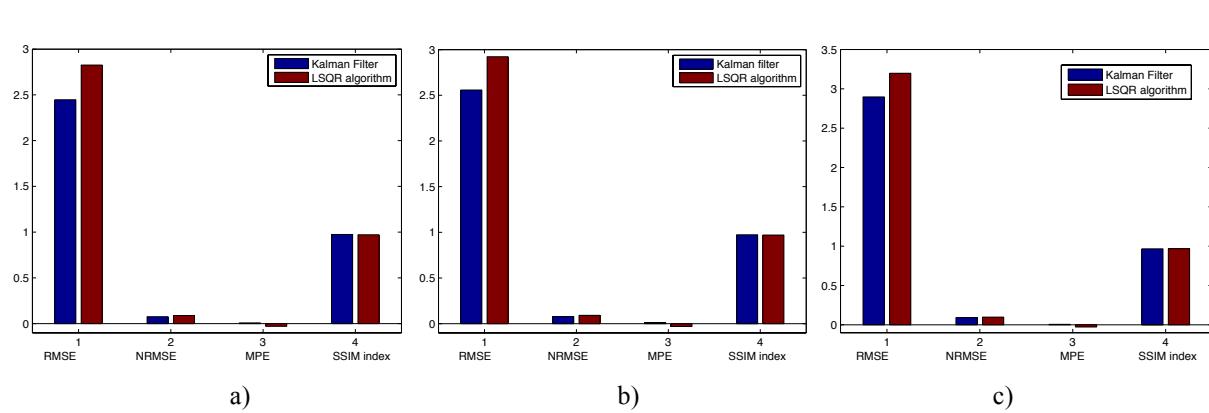
where $u_{i,t} \sim U[0,1]$, and the $d_{i,t}$ the assumed ground truth OD demand. The Eqn. (12) reflects the overestimated prior OD matrix in total demand compare to the true OD matrix, while Eqn. (13) represents the underestimated prior OD matrix. In this scenario we used the same traffic counts scenarios as defined in Scenario 1.

The results for Scenario 2 are presented in FIGURE 7 and FIGURE 8. The estimated OD matrices for the overestimated and underestimated prior OD matrix result in a slightly higher RMSE values, where the Kalman Filer solution algorithm for least square model formulation shows a better performance then LSQR algorithm. The improvement of the Kalman Filter algorithm can be explained by the structure of the used minimization algorithm. We can also observe that both methods show more sensitivity with the increase of deviations in traffic counts. Also, we can see from MPE value that both methods slightly systematically underestimate the OD demand. The SSIM index values ($SSIM_{KF} = 0.9671$ and $SSIM_{LSQR} = 0.9605$) are slightly lower then in Scenario

1 yet they still indicate that structural patterns in estimated OD matrices are preserved and very close to true OD
 2 matrix. Note that next to the RMSE-type measures, this provides us additional insight into this importation
 3 characteristic of the estimated OD matrix.



7 **FIGURE 7 Estimation results for overestimated prior OD matrix with:** a) traffic counts error $\sigma 2\%$; b)
 8 traffic counts error $\sigma 5\%$; and c) traffic counts error $\sigma 10\%$



12 **FIGURE 8 Estimation results for underestimated prior OD matrix with:** a) traffic counts error $\sigma 2\%$; b)
 13 traffic counts error $\sigma 5\%$; and c) traffic counts error $\sigma 10\%$

14 Although, we benchmark the performance of two OD estimation methods in a simple network example, we can
 15 observe that SSIM index provide more insight into the quality of estimation results. For more involved cases,
 16 with more complex networks and error structures, we expect the differences between the different measures to
 17 become more considerable.

19 TOWARDS A NEW OD ESTIMATION FRAMEWORK

20 In FIGURE 3 and FIGURE 4, we have seen that Euclidean distance measure as a performance function is only
 21 designed to find the distances between a pair of attributes in a data set or overall distance among all data. They
 22 are unable to find and distinguish different correlation structures in data, and cannot provide an in depth
 23 explanation of the patterns in OD demand.

24 One possible solution approach to tackle this problem is to use a performance function that will incorporate the
 25 structural correlation between OD pairs and that allows the modeler to control the trade off between simplicity of
 26 the model and the level of realism. In this section we propose such a framework that incorporate the structural
 27 similarity index in performance function as a penalty factor. Since the penalty factor has to be applicable to a
 28 whole range of OD estimation methods, the workings of which are not explicitly considered, the new
 29 performance function can be interpreted as consisting of two elements:

$$30 \quad \text{PerformanceFunction} = \text{Best_fit} \times \text{SSIM_index} \quad (13)$$

31 A higher best fit likelihood leads to optimal OD matrix solution that can explain and fit the available data well,
 32 i.e. that have a low estimation error $\sum(d_{est} - d_{prior})^2$. However, if only this measure would be investigated the

1 overfitting problem would occur as when the estimation error is used for prediction of OD matrices. Therefore,
 2 the models performance is penalized by the SSIM index, which always takes a value between -1 and 1. The
 3 estimated OD matrix that has a significantly different structure than the prior OD matrix has a lower SSIM index
 4 and therefore receives lower performance value.

5 The formulation of the new framework for the estimation (and assessment of the prediction) of OD demand
 6 given in Eqn. (13) has several advantages over existing OD estimation and prediction methods: (1) the most
 7 important feature is that it introduces additional information in the estimation process on the basis of structural
 8 patterns in the OD matrix. This allows in estimation process to select most probably best estimated OD matrix,
 9 taking into account both the estimation error as well as the structural similarity between OD matrices; (2) the
 10 structural OD pattern information is included when estimating the OD demand to rule out unrealistic estimation
 11 results due to the fact that too little information is taken from prior OD demand; (3) the approach can be used to
 12 combine with a weighted performance function, where the weighting value is determined by the corresponding
 13 covariance matrices representing the dispersion of collected data or by the analyst's relative confidence in either
 14 input information.

15 Note that the purpose of this section is to demonstrate the potential application of SSIM index and main features
 16 of the novel performance function in estimation and prediction of OD matrices. The presented framework is still
 17 academic in nature and must be interpreted as concept idea. In the future work we will rigorously derive the new
 18 performance function mathematically to ascertain that the method performs well in practice.

19

20 CONCLUSIONS AND FUTURE WORK

21 This paper discusses the potential of using the structural similarity (SSIM) index as a quality measure that
 22 quantifies the similarity between two OD matrices (e.g. between an OD matrix estimate and a reference OD,
 23 such as the prior OD matrix, or ground truth OD matrix). The most important feature of the new metric is that it
 24 includes additional information in evaluation process on the basis of the structural patterns of OD matrices. We
 25 showed that this quality metric has several advantages over existing statistical measures, such as the Mean
 26 Squared Error MSE. As an example, it turns out to be more sensitive to capture the structural correlation
 27 between OD pairs; it ignores the effect of the signs of the error in estimated OD matrix. Further, we recommend
 28 the application of SSIM index as a performance measure in addition to existing statistical measures in
 29 benchmarking studies.

30 Since our final objective was to show potential application of SSIM index as a new performance function, we
 31 presented a new framework for the estimation of OD demand. This allows in estimation process to select the
 32 optimal OD matrix, taking into account both the estimation error as well as the structural similarity between OD
 33 matrix to be estimated and prior OD matrix. The presented framework is still theoretical in nature and must be
 34 interpreted as such. In particular the mathematical characteristics of the measure needs to be investigated further
 35 (e.g. non-convexity) and its implications for estimation need to be formulated and if need be adapted. More
 36 results in more realistic settings will be obtained in future research to ascertain that the method performs well in
 37 practice.

38 ACKNOWLEDGEMENTS

39 This research is partly funded by the ITS Edulab, a collaboration between TU Delft and Rijkswaterstaat.

40 REFERENCES

- 41 1. Bierlaire, M., *The total demand scale: a new measure of quality for static and dynamic origin-*
42 destination trip tables. Transportation Research Part B: Methodological, 2002. **36**(9): p. 837-850.
- 43 2. Bell, M.G.H., *The estimation of origin-destination matrices by constrained generalised least squares*.
44 Transportation Research Part B: Methodological, 1991. **25**(1): p. 13-22.
- 45 3. Bierlaire, M. and F. Crittin, *An Efficient Algorithm for Real-Time Estimation and Prediction of*
46 Dynamic OD Tables. OPERATIONS RESEARCH, 2004. **52**(1): p. 116-127.
- 47 4. Van Zuylen, H.J. and L.G. Willumsen, *The most likely trip matrix estimated from traffic counts*.
48 Transportation Research Part B: Methodological, 1980. **14**(3): p. 281-293.
- 49 5. Wu, J., *A real-time origin-destination matrix updating algorithm for on-line applications*.
50 Transportation Research Part B: Methodological, 1997. **31**(5): p. 381-396.
- 51 6. Watling, D.P., *Maximum likelihood estimation of an origin-destination matrix from a partial*
52 registration plate survey. Transportation Research Part B: Methodological, 1994. **28**(4): p. 289-314.
- 53 7. Hazelton, M.L., *Inference for origin-destination matrices: estimation, prediction and reconstruction*.
54 Transportation Research Part B: Methodological, 2001. **35**(7): p. 667-676.

- 1 8. Van Der Zijpp, N., *Dynamic Origin-Destination Matrix Estimation from Traffic Counts and Automated*
2 *Vehicle Identification Data*. Transportation Research Record: Journal of the Transportation Research
3 Board, 1997. **1607**(-1): p. 87-94.
- 4 9. Ashok, K., M.E. Ben-Akiva, and T. Massachusetts Institute of, *Dynamic origin-destination matrix*
5 *estimation and prediction for real-time traffic management systems*. Transportation and traffic theory.,
6 1993.
- 7 10. Ashok, K. and M.E. Ben-Akiva, *Alternative Approaches for Real-Time Estimation and Prediction of*
8 *Time-Dependent Origin-Destination Flows*. TRANSPORTATION SCIENCE, 2000. **34**(1): p. 21-36.
- 9 11. Yang, H., Y. Iida, and T. Sasaki, *An analysis of the reliability of an origin-destination trip matrix*
10 *estimated from traffic counts*. Transportation Research Part B: Methodological, 1991. **25**(5): p. 351-363.
- 11 12. Marzano, V., A. Papola, and F. Simonelli, *Limits and perspectives of effective O-D matrix correction*
12 *using traffic counts*. Transportation Research Part C: Emerging Technologies, 2009. **17**(2): p. 120-132.
- 13 13. Djukic, T., J. van Lint, and S. Hoogendoorn, *Efficient Methodology for Benchmarking Dynamic Origin-*
14 *Destination Demand Estimation Methods*. Transportation Research Record: Journal of the
15 Transportation Research Board, 2011. **2263**(-1): p. 35-44.
- 16 14. Zhou, W., A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, *Image quality assessment: from error*
17 *visibility to structural similarity*. Image Processing, IEEE Transactions on, 2004. **13**(4): p. 600-612.
- 18 15. Casetta, E., D. Inaudi, and G. Marquis, *Dynamic Estimators of Origin-Destination Matrices Using*
19 *Traffic Counts*. TRANSPORTATION SCIENCE, 1993. **27**(4): p. 363-373.
- 20 16. Antoniou, C., M. Ben-Akiva, and H.N. Koutsopoulos, *Dynamic traffic demand prediction using*
21 *conventional and emerging data sources*. Intelligent Transport Systems, IEE Proceedings, 2006. **153**(1):
22 p. 97-104.
- 23 17. McKay, M.D., R.J. Beckman, and W.J. Conover, *A Comparison of Three Methods for Selecting Values*
24 *of Input Variables in the Analysis of Output from a Computer Code*. Technometrics, 1979. **21**(2): p.
25 239-245.
- 26
- 27

256 shades of grey – comparing OD matrices using image quality assessment techniques

Tom van Vuren
Tim Day-Pollard
tom.vanvuren@mottmac.com

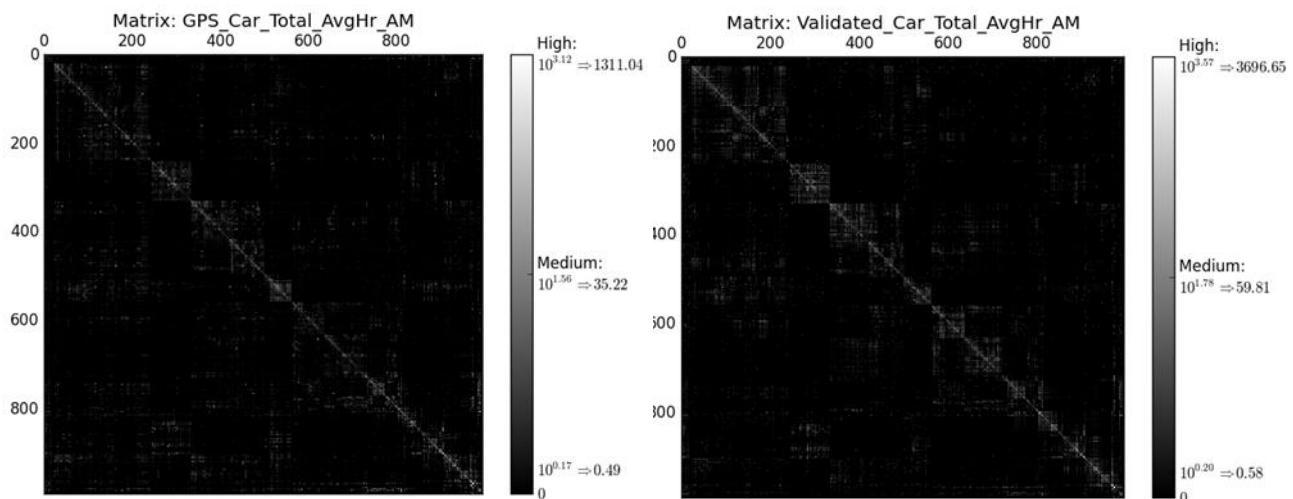
1. OUR PROPOSITION

Imagine a trip matrix, an OD matrix. Usually square, consisting of anything between a few tens of cells for a simple junction to millions of cells for a strategic, regional transport model.

If the matrix is sparse, for example the result of a roadside interview (RSI), the matrix will contain mainly zeroes and a few quite large numbers (interviews expanded to observed flows). If the matrix is based on Big Data, such as mobile phone data or TrafficMaster GPS data, most of the cells in the matrix will be filled, probably with quite small numbers.

Now colour the cells with 256 shades of grey – the zeroes are black, the maximum value is white and use the 254 shades in-between in equal size blocks. Look at the matrix now, and it has become a black and white picture. Matrices that are similar will look very similar. Image processing techniques can now be used to compare them.

Figure 1: Two similar OD-matrices (a GPS component matrix and the final validated matrix) compared as grey-scale images



2. WHY IS SUCH A TECHNIQUE HELPFUL?

OD matrices reflect dynamic spatial movements of travellers but forced into a discrete cell-based structure determined by subjective zoning systems and time period definitions. Both spatially and temporally, these movements are continuous. They differ between days and surveys can only observe a small overall sample whilst estimation techniques tend to be inevitably crude.

Standard statistical techniques do not really reflects such complexities. For example, the r-squared statistic (as recommended in DfT, 2014) can neither recognise the spatial or temporal proximity of movements in adjacent cells. Think again of the image processing analogy. Squinting through your eye-lashes you will recognise that two photographs of which one has moved a fraction left, right, up or down are essentially the same (or at least structurally similar); the r-squared statistic cannot do that.

What we are looking for is a comparator that does not just look at the cell values in isolation, but also considers adjacent cells and cells adjacent to these (in space and potentially in time). The comparator

must also be able to deal with some of the idiosyncrasies of OD matrices, for example comparing sparse matrices with dense matrices.

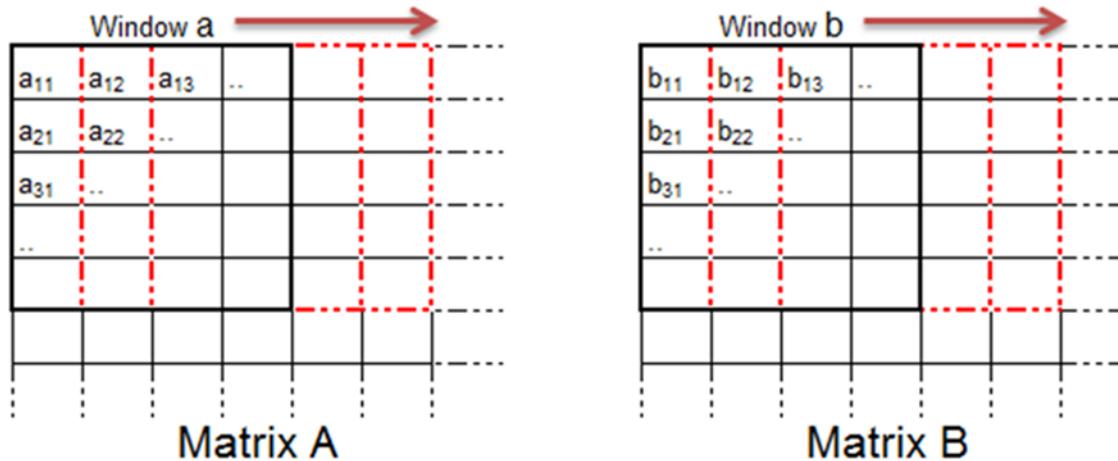
3. WHAT HAS BEEN DONE BEFORE?

In previous publications (Pollard et al, 2013; Day-Pollard and Van Vuren, 2014) we explained how we discovered in existing literature an image processing measure of comparison, the MSSIM (mean structural similarity index) which looks promising for use in OD-matrix comparisons too. The index was originally discovered by Tamara Djukic (Djukic et al, 2013) who refers back to Zhou et al (2004) as the source. Others have also used the MSSIM in OD-matrix comparisons, such as Bringardner et al (2014) and Ruiz de Villa et al (2014).

4. HOW DO I INTERPRET THE MSSIM?

The MSSIM is calculated by summing and averaging SSIM values, as illustrated below, across a whole matrix. The SSIM is calculated over a part or square block of the matrix, generally a few cells wide by a few cells high. Pollard et al (2013) describe the mechanics, illustrated in Figure 2 below.

Figure 2: Calculating the MSSIM by summing and averaging the SSIM value for consecutive blocks in matrices a and b, moving across the whole matrix



What is important to recognise is that the index calculates three characteristics of OD-matrices a and b as in Equation (1) and sums and averages as in Equation (2):

Equation 1

$$SSIM(a, b) = \frac{(2\mu_a\mu_b + C_1)(2\sigma_{ab} + C_2)}{(\mu_a^2 + \mu_b^2 + C_1)(\sigma_a^2 + \sigma_b^2 + C_2)}$$

- The mean value within this part of the matrix, represented by μ_a and μ_b
- The variance (in pixels or OD values), represented by σ_a and σ_b
- The covariance between pixels or OD values, represented by σ_{ab}

Equation 2

$$MSSIM = \frac{1}{N} \sum_{i=1}^N SSIM(a_i, b_i)$$

In the language of pictures, μ , σ_{ab} and σ are interpreted as luminosity, contrast and structure, and the SSIM can be simplified to the form in Equation (3):

Equation 3

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma$$

The calculation involves two constants C1 and C2, for which limited guidance is provided in Zhou et al (2004). We understand that their main job is to avoid divisions by zero when values for μ and σ are zero (or sufficiently close to zero to cause computational errors), and we have derived an alternative formulation without these constants.

5. PRACTICAL ISSUES WHEN USING MSSIM

There are a number of immediate issues that arise when comparing OD-matrices using the MSSIM indicator:

1. How big should the block of matrix cells be that are included in the SSIM comparison? This in turn raises the question of how to determine whether zones are close in matrices with different and user-specific zone numbering systems?
2. How to compare sparse and dense matrices?
3. What is an acceptable MSSIM value, or in other words, when are OD-matrices similar enough?

In the next sections we will show that the first question can be answered by using a proximity measure between OD-cells that we developed in Day-Pollard and Van Vuren (2015). The third question remains an open one, and can only be answered by more testing with more matrices and more networks. We are always interested in hearing from you if you are willing to test MSSIM in your own transport modelling work.

6. WHAT IS A NEARBY OD-PAIR?

Whereas in standard images physically nearby pixels are also of greatest MSSIM interest, this is not necessarily the case when comparing OD-matrices. For example, in a hierarchical zoning system based on administrative areas, zones i and $i+1$ may well be quite distant. There is no guarantee that adjacency in the zone numbering also implies adjacency spatially. Also, and importantly, the comparison we want is between OD-pairs, not origin or destination zones.

Fortunately, there is a way in which proximity between OD-pairs can be calculated. Every OD-pair has an origin and destination centroid, each of which has x- and y-coordinates. Hence, every OD-pair can be defined by (x_O, y_O, x_D, y_D) . The proximity between two OD-pairs can now be established by the Euclidean distance between these coordinates (a_1, a_2, a_3, a_4) for OD pair a and (b_1, b_2, b_3, b_4) for OD pair b, as follows:

Equation 4

$$distance = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2}$$

As an example, compare in the map below OD-pair 1511-7043 (bottom right to top left) with four other OD pairs, two running parallel and of similar length, one almost perpendicular and one very short but with a shared origin. The Euclidean distances are sensible – providing confidence that this measure rather than the zone numbering is appropriate for determining which OD pairs to consider in the SSIM calculations.

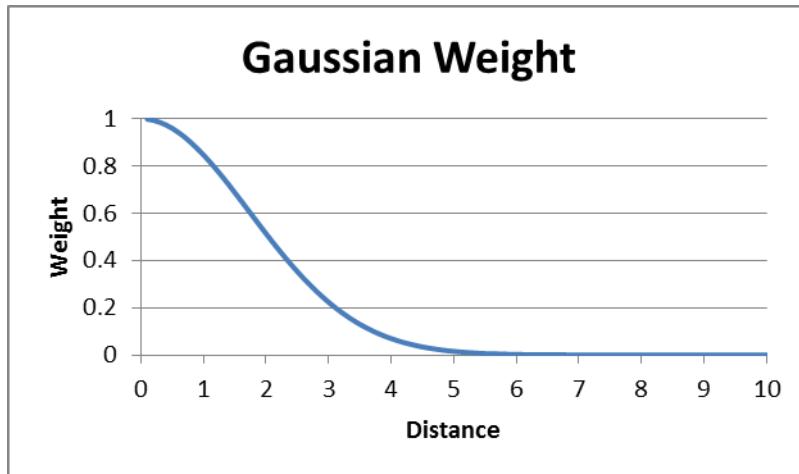
Figure 3: Example comparison of several OD-pairs on the basis of their Euclidean distance



As the Euclidean distance as calculated above does not just determine which OD-pairs are near but also *how* near, the actual Euclidean distance can be used as a weight for the contribution of *any* OD-pair to the SSIM value. In our implementation we use a Gaussian weight as advised by Zhou et al (2004), with further OD-pairs contributing less. In Zhou et al (2004) the Gaussian function is chosen because of the comparatively high weight it gives to the closest pixels, and then after a pre-defined boundary the weight drops significantly. Applying such a function directly to a matrix picture such as Figure 1 would not do justice to the spatial peculiarities of OD-movements for the reasons already discussed. In our implementation we have adjusted the formula to take the Euclidean distance as an input rather than position a proxy grid-picture.

We call this enhancement 4D-MSSIM, the 4-dimensional mean structural similarity index. We aim for a shape similar to that in Figure 4, with a falling contribution from OD-pairs further away. The functional form in Equation (5) works well. The value in the denominator deserves some attention; it would make sense to link this to attributes of the network eg the average trip length or average zone diameter and would thus be mode and possibly purpose-specific. We have not yet analysed the effect of different assumptions in this respect. There is a calculation overhead associated with the 4D extension, as the nearest OD-pairs need to be calculated for every OD-pair in turn. Also, it may be reasonable to apply a cut-off in the contributions of further OD-pairs to the 4D-MSSIM value. So far we have limited our calculations to the nearest 625 OD pairs to aid run-time, and used a parameter σ of 7.5 in the Gauss-equation. This parameter was estimated based on the diameter of a sample of zones in the core model area.

Figure 4: Example shape of Gaussian weighting of contributions by OD-pairs to the 4D-MSSIM value as a function of their Euclidean distance



Equation 5

$$w = \exp\left(-x^2/2\sigma^2\right)$$

7. COMPARING SPARSE AND DENSE OD-MATRICES

As explained earlier, the MSSIM compares structural similarity calculations of mean, variance and covariance in parts of the matrix, summed and averaged. Whereas the mean is a valid comparator between sparse and dense matrices, the variances and covariances can be expected to be structurally rather different so that the standard MSSIM would not be able to capture the similarities that we seek to find.

As discussed in the context of Equation (3) the standard MSSIM comparator assesses similarities in mean or luminosity (I), structure (s) and contrast (c) in equal weight with α , β and $\gamma = 1$. When comparing sparse matrices, a greater weight might need to be placed on the mean, for example $\alpha = 1$, $\beta = 0.1$ and $\gamma = 0.1$. We have carried out some comparative tests but have not yet concluded whether this aids in the comparisons and what appropriate values are for α , β and γ .

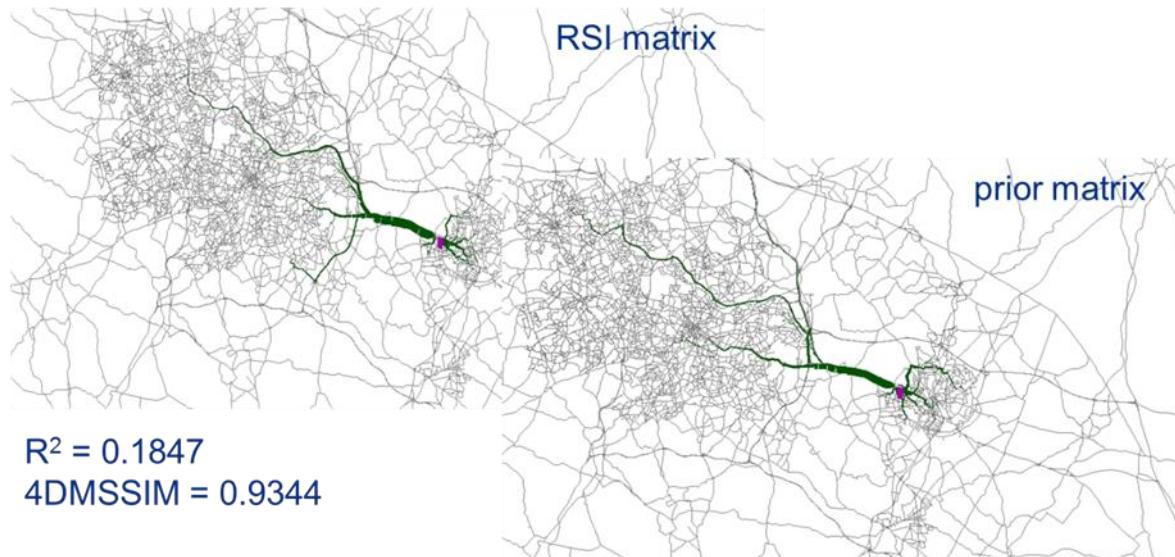
8. DOES IT WORK?

Our previous papers (Pollard et al, 2013; Day-Pollard and Van Vuren, 2015) contain some numerical comparisons between MSSIM and standard OD-matrix comparators such as the r-squared and Chi-squared statistics and the GEH of link flows after assignment. Our conclusions in these papers were, based on artificially constructed matrices:

- The MSSIM index can identify structural differences better than traditional measures such as r-squared or GEH comparisons;
- Conversely, the r-squared value can fail to pick up similarities that MSSIM can detect.

Since developing the 4D-MSSIM implementation we have carried out further tests, using real-life matrices from the PRISM strategic transport model for Greater Birmingham (Van Vuren, 2008); in particular the prior matrix before matrix estimation from counts and one of the (partial) roadside interview (RSI) matrices. We created an RSI equivalent matrix from the validated matrix through a select link analysis at the RSI site.

Figure 5: Assignment of observed roadside interview matrix and equivalent prior matrix in PRISM strategic model plus associated 4D-MSSIM and r-squared values between the two matrices



The results show:

- That the assignments of the two matrices are very similar, as expected;
- That the 4D-MSSIM detects great similarity between the two matrices (value close to 1)
- That the r-squared comparison detects very little similarity (value close to zero), possibly because of the sparseness of the RSI matrix

We also carried out a comparison between the prior matrix and the validated matrix, i.e. that post matrix estimation from counts. WebTAG advises an r-squared comparison, with a target value in excess of 0.95. In the PRISM case an r-squared value of 0.958 is achieved, i.e. a pass. However, it is interesting to note that the 4D-MSSIM comparator only achieves a mediocre 0.611 value.

When considering which is the better comparator for matrix-estimation we must consider the question, what level/type of similarity are we hoping to achieve after matrix estimation? This is particularly relevant given that we are hoping to achieve a different pattern of flows in the assignment. It is acknowledged that some significant level of change is expected and required on some of the OD pairs. The r-squared comparator tests the overall correlation i.e. the average change, whereas the 4D-MSSIM tests the similarity in groups of ODs. This raises the question: is the r-squared a good and sufficient comparator? Are the matrices really similar enough?

9. NEXT STEPS

Our MSSIM work has been carried out after work and in weekends, and as a result our analyses are limited and as yet inconclusive. However, we hope you agree that the results are promising; and in particular that existing matrix comparisons using r-squared or Chi-squared statistics may well be inappropriate; whilst GEH analyses of link flows after assignment are insufficient.

Further testing of the MSSIM or 4D-MSSIM implementations on a wider range of networks and under a wider set of circumstances should provide us insight a) whether the MSSIM index should have a place in our transport modellers' toolkit; b) which of the implementations is most suitable and c) which values we should aim for to be confident that matrices are similar enough. We are always keen to hear from those interested in testing MSSIM in their own work; please get in touch with us using the contact details in the paper.

REFERENCES

- Bringardner, JW, Gemar, MD, Boyles SD and Machemehl RB (2014): Establishing the Variation of Dynamic Traffic Assignment Results Using Subnetwork Origin-Destination Matrices, presented at TRB Annual Meeting, Washington DC, January 2014
- Day-Pollard T and Van Vuren T (2015): When are origin-destination matrices similar enough?, presented at 94th Annual TRB Meeting, Washington DC, 2015
- Djukic, T, Hoogendoorn, SP and Lint, JWC van (2013) Reliability assessment of dynamic OD estimation methods based on structural similarity index, presented at TRB Annual Meeting, Washington DC, January 2013
- DfT (2014) WebTAG Unit M3.1 – Highway Assignment Modelling, Department for Transport, January 2014, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/427124/webtag-tag-unit-m3-1-highway-assignment-modelling.pdf, accessed June 2015
- Pollard T, Taylor N and Van Vuren T (2013): Comparing the quality of OD matrices in time and between data sources, European Transport Conference (<http://etcproceedings.org>), Frankfurt, 2013
- Ruiz de Villa, A, Casas J and Breen M (2014): OD matrix structural similarity: Wasserstein metric, presented at TRB Annual Meeting, Washington DC, January 2014
- Van Vuren, T (2008): A transport policy toolkit for the West Midlands, Proceedings of the Institution of Civil Engineers, Transport 161, pp 85-90, London, 2008.
- Zhou, W, Bovik, AC, Sheikh, HR and Simoncelli, EP (2004): Image quality assessment: from error visibility to structural similarity, Image Processing, IEEE Transactions on, 13(4): p. 600-612, April 2004.



Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations

Patrick Bonnel, Etienne Hombourger, Ana-Maria Olteanu-Raimond, Zbigniew Smoreda

► To cite this version:

Patrick Bonnel, Etienne Hombourger, Ana-Maria Olteanu-Raimond, Zbigniew Smoreda. Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations. *Transportation Research Procedia*, Elsevier, 2015, 11, pp.381 - 398. <10.1016/j.trpro.2015.12.032>. <halshs-01664219>

HAL Id: halshs-01664219

<https://halshs.archives-ouvertes.fr/halshs-01664219>

Submitted on 15 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 11 (2015) 381 – 398

Transportation
Research
Procedia
www.elsevier.com/locate/procedia

10th International Conference on Transport Survey Methods

Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations

Patrick Bonnel ^{a*}, Etienne Hombourger ^b, Ana-Maria Olteanu-Raimond ^c, Zbigniew Smoreda ^d

^a Laboratoire d'Economie des Transports, ENTPE, Lyon, France, patrick.bonnel@entpe.fr

^b DTecITM, CEREMA, Paris, France, Etienne.hombourger@cerema.fr

^c COGIT, IGN, Paris, France, Ana-Maria.Raimond@ign.fr

^d SENSE, Orange Labs, Paris, France, zbigniew.smoreda@orange.com

Abstract

Mobile phone operators produce enormous amounts of data. In this paper we present applications performed with a dataset (communication events + handover and Location Area Up-date) collected by the operator Orange from 31 March to 11 April 2009 for the whole Paris Region. Trips are deduced from the spatio-temporal trajectory of devices through a hypothesis of stationarity within a Location Area in order to define activities. Trips are then aggregated in an origin-destination matrix which is compared with traditional data (census data and household travel survey).

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of International Steering Committee for Transport Survey Conferences ISCTSC

Keywords: origin-destination matrix; mobile phone data; travel survey; passive data

1. Introduction

Data on spatial mobility are essential in order to build and use travel demand forecasting models, for transport planning purposes and for the appraisal of transport policies... (Arentze et al., 2000; Ortuzar, Bates, 2000). They must also be of good quality and, in particular, accuracy, to ensure that investment or transport policy decisions are

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .

E-mail address: author@institute.xxx

based on reliable analyses. While travel surveys provide extremely useful data in order to formalize and estimate behavioural choice models (for example the choice of a destination or mode of transportation), they are much less useful for constructing origin-destination (O-D) matrices due to an inadequate number of trips in many of the matrix elements. In addition, surveys are increasingly confronted by issues during the sample construction phase (Stopher, Greaves, 2007), by falling response rates (Atrostic, Burt, 1999; Ampt, 1997; Bonnel, 2003; Zmud, 2003) and by unreported trips (Wolf et al., 2003), which reduce even further the quality of the resulting matrices. Consequently, trip matrices are also generated from other sources, in some cases in combination with survey data, examples being roadside traffic counts, cordon or screen-line surveys and public transport surveys. The resulting data are useful for improving the quality of matrices, but do not always contain the necessary information. This applies, for example, to road traffic counts which provide information on the traffic volumes at a given point on a road, but not on trip origins and destinations. A variety of techniques have been developed for processing and combining the data from different sources. However, the reliability of the resulting matrices is uncertain and cannot always be measured statistically.

The advent of large volumes of data that are produced automatically and passively such as ticketing data (Arana et al., 2014, Morency et al., 2007; Munizaga et al., 2010; Pelletier et al., 2011), bank cards... and mobile phone data makes it possible to identify the presence of individuals in both space and time in a way which, while admittedly irregular is becoming less so. A number of techniques have been developed for converting these data into trips, but little research has attempted to "validate" them by comparing them with data from other sources in order to identify possible biases and gain a clearer idea of their potential. The aim of this paper is therefore to test the potential of these data for producing origin-destination matrices compared with other sources of available data. The analysis has been conducted within the Greater Paris Region (Ile-de-France) for which we were able to analyse the mobile phone data from the operator Orange and compare them on the one hand with the census data on commuting trips from home to place of work or study, and on the other hand with the data obtained from the "*Enquête Globale Transport*" (EGT), which is the name given to the household travel surveys conducted in the Ile-de-France Region.

We shall begin this paper with a literature survey (Section 1) before presenting the data we have used (Section 2) and (Section 3) the data processing methodology employed to produce the origin-destination matrices, which allow us to make the comparison with external data (Section 4). Finally, we shall present the principal lessons from this research, and some suggested directions for the future (Section 5).

2. Literature survey

Cell phone networks have existed for two decades, and mobile phones have achieved a high rate of penetration: there were 76.8 million active SIM (Subscriber Identity Module) cards in France at the end of 2013, for a total population of 65 million (ARCEP, 2014). Mobile devices (mobile phones, smartphones and tablets) have become indispensable tools, bearing witness to our activities and trips. Mobile phone operators, who are obliged for legal or billing purposes to record information about the use of these devices, therefore find themselves with increasingly informative databases. The reason for this is that each time a mobile terminal is used to make a call, send an SMS (Short Message Service), the operator generates a call detail record (CDR) that contains the timestamp, the terminal's identifier of the base station to which the user is connected and quantitative data about the call (call duration, volume of data exchanged).

As a result of the size of the samples, which in the case of some operators can involve as many as 40-50% of a country's population, and the non-intrusive way the data is collected, the exploitation of mobile phone data logs has enormous potential. Recent cases include using the data to analyse behavioural differences between men and women (Frias-Martinez et al., 2010), studying the propagation of an epidemic (Tizzoni et al., 2013), mapping activities within a city (Noulas et al., 2013), or improving the paging efficiency of the cellular network (Zhang, Bolot, 2007).

But the usefulness of mobile phone data has above all been proven for the study of human mobility, in spite of the fact that the localisation data associated with each log is limited to the position of the base station used, which results in a positioning uncertainty ranging from approximately a hundred metres in a dense urban zone (Calabrese et al., 2013) to several kilometres in rural zones. Gonzalez et al. (2008) were amongst the first scholars to carry out a large-scale study of the mobility of users, with a sample of over 100,000 individuals. This study demonstrated that human mobility may be modelled using a random technique and that trips follow a truncated power-law distribution.

The authors of this paper also found that individuals have a strong tendency to visit a limited number of places many times periodically and many other places just once. Cho et al. (2001) also factored in the impact of social ties, obtained from an online social network. They concluded that short journeys (less than 100 km) are in most cases periodic in nature, while long journeys are much more influenced by the individual's social network (i.e. the presence of friends).

However, even if human mobility seems to comply with these laws in a generic manner, the environment has a strong influence on the parameters of the various distributions. In a series of studies, Isaacman et al. (2010, 2011) have shown that there are important differences between cities (New York and Los Angeles) and seasons (fewer trips in the winter than the summer). Temporary tourist attractions play a major role and may modify a city's normal mobility patterns (Calabrese et al., 2010).

The use of mobile phone data to construct origin-destination matrices in an urban region was first proposed in Italy by Bolla and Davoli (2000) and tested on a small sample in (White and Wells, 2002) with the aim of studying traffic on specific roads. In 2002, Akin and Sisiopiku (2002) selected just 500 individuals in the city of Birmingham in the United States. One of the first studies to use the whole population rather than a sample was carried out in Israel in 2007 (Bar-Gera, 2007). The research in question set out to estimate the traffic and obtain mean speed data on a 14 km road in Israel with 10 interchanges. Calabrese et al. (2011) were the first to produce O-D matrices from a detailed dataset, for the Boston region in Massachusetts.

Data from the mobile phone network can also be used to estimate individual trajectories. In 2009, Schlaich et al. (2010) developed an algorithm that was able to precisely identify a GSM network user's trajectory between the cities of Karlsruhe and Stuttgart in Germany. Two years later Jiang and a group of researchers (Jiang et al., 2011) went further in this area, assigning each user to the transport network in the city of Lisbon.

Mobile phone data can also be used to study mean speeds and journey times. One of the first studies to do this was led by Ygnace (2001) and carried out in the South of France on a rural motorway which became an urban motorway near Lyon. The findings showed that in rural areas the data from the mobile phone network matched those obtained from road traffic surveys but there was a great difference between the data from the two sources in urban areas. More recently, Calabrese et al. (2011, 2013), working in the Boston conurbation, used all the data collected by a telecom operator to study mean speed, mean trip length and the distribution according to the time of day.

The research conducted by Bekhor et al. (2013) is without doubt the most extensive, as it concerns the analysis of the long-distance trips carried out over the entire area of Israel. It illustrates the considerable potential of mobile phone data for the analysis of long-distance trips.

However, matrices obtained in the course of these studies are only representative of the individuals using the network at a given time. Representativeness is of prime importance for these data which describe the mobility of the population of a region or mobility within a region if it is envisaged to use them for planning purposes or for regulating or optimising the use of transport networks. To our knowledge, few studies have tackled this issue. Moreover, the small number of published studies frequently employs different methodologies, pursue different goals and do not always use the same types of mobile phone data.

In 2002, two simultaneous research projects attempted to extract origin-destination matrices from mobile phone network data. One of these (Akin, Sisiopiku, 2002), working in the city of Birmingham (USA), developed an algorithm which calculated origins and destinations and divided the day into three periods:

- from midnight to 8 am, when the person was theoretically at home;
- from 8 am to 4 pm when the person was theoretically at work;
- from 4 pm to midnight when the person was theoretically engaged in activities.

To compute the subject's position during these three time periods, they took the largest number of connections in a zone. Next, during each time period, they considered the largest number of connections as an origin-destination pair, and thus generated an origin-destination matrix. This study has certain limitations, as the matrix which is generated only takes account of trips which are identified as home-work, work-leisure and leisure-home. However, no verification was conducted in this paper on the basis of a comparison with data obtained from other processing methods.

In England, at the same time, (White and Wells, 2002) tested the feasibility, in the county of Kent, of creating an

origin-destination matrix from billing data (Call Detail Record, CDR). They then compared the results with a survey-based origin-destination matrix. They concluded that the billing data were not accurate enough to provide a reliable origin-destination matrix.

In 2007, Caceres et al. (2007) calculated an origin-destination matrix for a road between the cities of Huelva and Seville in Spain. They considered four possible origin-destination pairs based on the positioning of the motorway interchanges. To construct the origin-destination matrix, it was deemed that as soon as a user left the road he/she was no longer visible on the studied network. Moreover, road users had to change zones at a speed which was compatible with below the speed limit in the area. The team then compared the results with those obtained from a road traffic count. The results were very satisfactory: the error did not exceed 4% on any of the possible origin-destination pairs.

More recently, Mellegard (2011) conducted a study that covered a large part of Sweden. To generate the origin-destination matrix the algorithmic method described by Kang et al. (2004) was applied. The method extracts from position data, and with a high degree of accuracy, the places where an individual has stayed for some time or the places an individual has passed through. (Kang et al. (2004) then applied the algorithm to GPS (Global Positioning System) data. Mellegard (2011) adapted the algorithm to the constraints imposed by the database he used in order to obtain an origin-destination matrix. However this study made no sophisticated comparison for the entire O-D matrix, but merely compared a very small number of origin-destination pairs with the data obtained from other surveys.

In 2012, a major study was conducted in two American cities, San Francisco and Boston, by Wang et al. (2012). This team of researchers constructed hour-by-hour origin-destination matrices in order to observe the level of saturation of the network during morning peak periods. The method only took account of journeys taking less than one hour. The results were then analysed by segmenting the population into three groups based on the amount of data collected to verify that frequency of mobile phone use did not introduce a bias. The study was based on a train/road modal split which was subsequently compared with the road traffic count data. The results were deemed to be very satisfactory.

Calabrese et al. (2013) conducted a dual analysis using data from Boston. First, they compared the number of trips per person to the data from the National Travel Survey. The results are fairly close, although the number of trips is slightly greater in the mobile phone data. The authors consider that this disparity can be explained on the one hand by the fact that the scope of the data differs in Boston from the rest of the USA and the fact that underestimates are frequent in travel surveys (Wolf et al., 2003). They then compared the estimated distances with those given by the odometer readings from the annual safety inspections of all private vehicles. The results reveal considerable differences in levels, but fairly similar structures.

Chen et al. (2014) emphasised the shortcomings of the work conducted to validate the mobility data obtained from mobile phone data, and mention that Calabrese's research is the most sophisticated in this respect. Chen's team made a contribution to data validation, but working from a sample of mobile phone data that was simulated on the basis of a household travel survey and mobile phone data. The goal was to have an "accurate" database about which everything is known (the household travel survey) and work on the simulated mobile phone database in order to identify its ability to reproduce the "accurate" data. In this way they have shown that they can reproduce the location of individuals' home and work with a fairly high degree of accuracy, and, with less accuracy, the location of the places they visit.

Our aim in this research is therefore to make an additional contribution to the existing work on the representativeness of mobile phone data. Our analysis relates to the validation of origin-destination matrices obtained from mobile phones by comparing them to external data sources. We shall present our data in the following section.

3. Data used: mobile phone data, commuting data and household travel survey data

In this section the mobile phone data and the external sources used to compare the origin-destination matrices are presented.

3.1. Orange positioning data

The mobile phone network is made up of a set of base stations each of which has a coverage zone (Figure 1). In practice, the zone area is variable with small zones in dense area and much larger zones in low density area as indicated in Figure 2. Furthermore coverage zone is not fixed as it depends on the activity of each base station, as an overloaded base station can pass on some traffic to its neighbours. It also depends on the topography and meteorological conditions. In theory the base station coverage is often represented by Voronoi polygons. The base stations are grouped together to form Location Areas (LA) for reasons to do with management of the mobile phone network as this makes it possible to identify mobile phones more rapidly in the case of a call or an SMS. The LA in which a mobile phone is located is known all the time, while at base station level its position is only known in the event of a call.

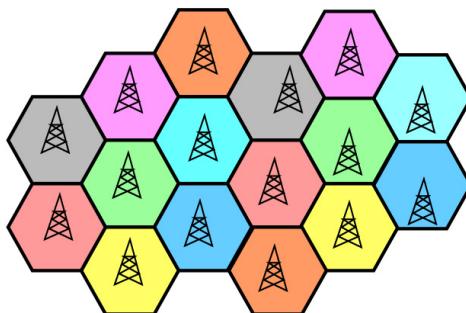


Fig. 1. The cell phone system. Source: (BRISSON, 2008)

For the purposes of this research, Orange has given us access to the data from the network of base stations in Île-de-France, which covers an area that approximates to the administrative region. Data were collected directly from the base stations between 31 March 2009 and 11 April 2009. The Ile-de-France region, which includes the conurbation of Paris, has 12 million inhabitants and covers an area of 12,000 km². The mobile phone data take two forms:

- Billing data (CDR-Call Detail Records), these list the base station through which the information was transmitted every time an individual receives or sends a call or an SMS;
- Signalling data, which are all the data that pass through the base stations. In addition to the billing data, these contain details of handovers (i.e. changes of base station during a call), LA updates, and logs of when the mobile phone is switched on or off. This data is collected via network quality probing systems.

We have worked with the signalling data which have the advantage of informing us in which LA the mobile phone is located on a permanent basis. However, its spatial resolution is much lower (Figures 2 and 3). The Ile-de-France region has almost 10,000 base stations, and 32 LA, each of which has between 150 and 500 base stations.

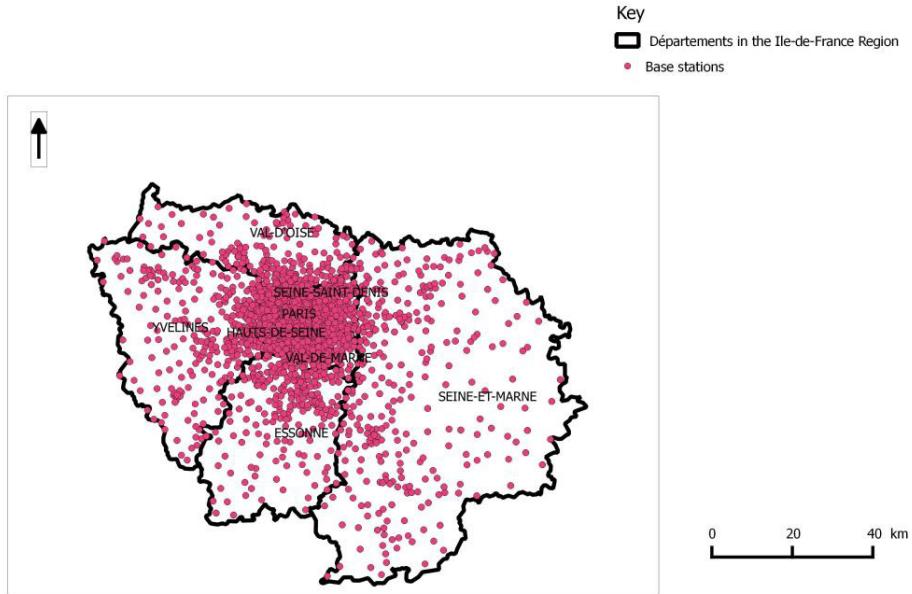


Fig. 2. Position of the base stations in the Île-de-France Region. Source: Orange

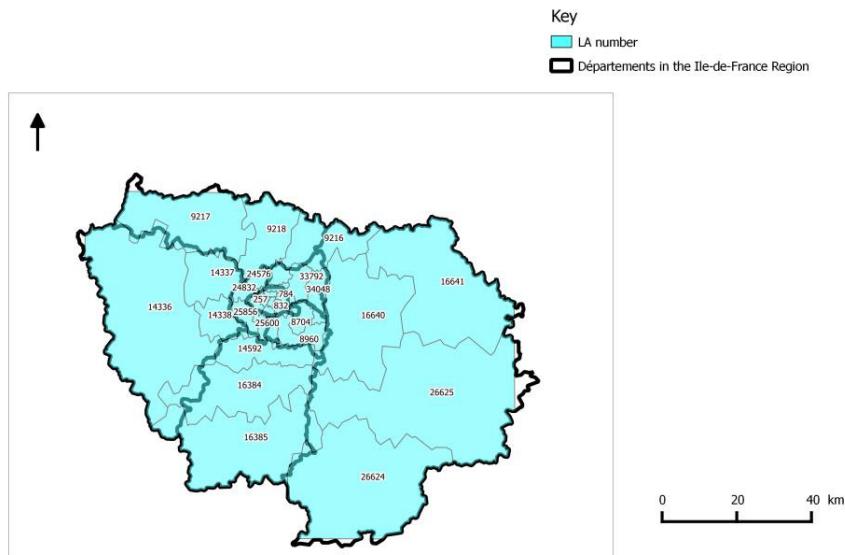


Fig. 3. Area covered by the LA in the Île-de-France region. Source: Orange

The database contains 1.5 billion logs. The availability of a unique anonymised code for each mobile phone means it is possible to find out the number of users who use the Orange network on a given day (Figure 4). According to IDATE (2009), Orange had 43.5% of the SIM card market at the end of 2008 and according to the Sofres TNS, almost 80% of the French population aged over 12 years owned a mobile phone in 2008. In view of the

total population of the Ile-de-France Region and the fact that the figures may be slightly different for the Ile-de-France Region than for the rest of France, the mean number of phones identified per day seems to be consistent with the available statistics.

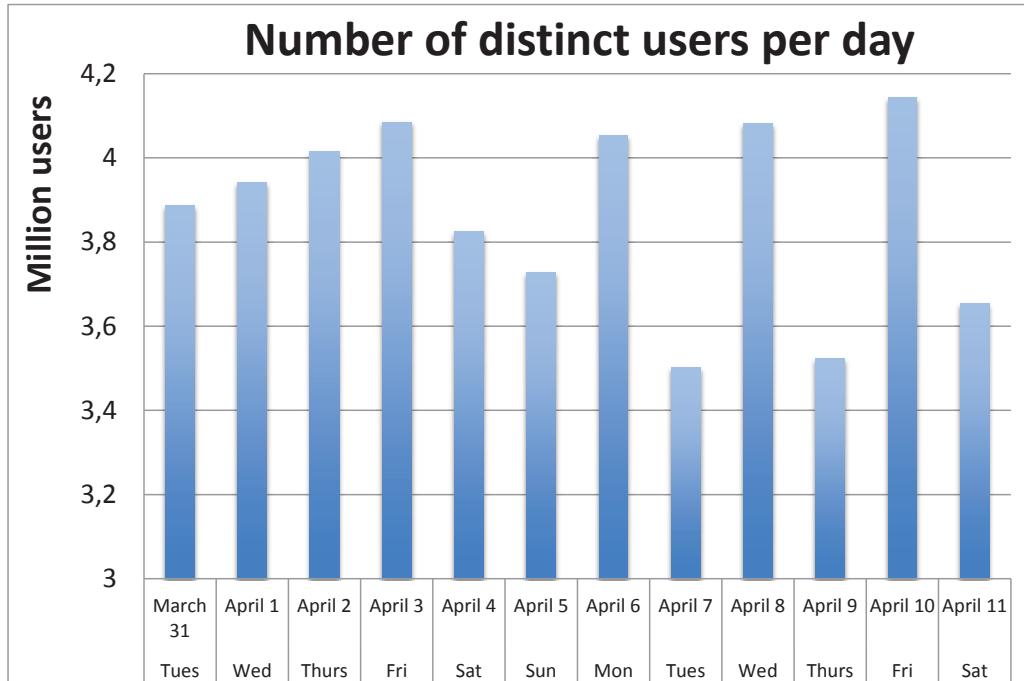


Fig. 4. Daily number of individuals using the network. Source of data: Orange

3.2. Data validation: commuting data and the *Enquête Globale Transport*

Commuting and travel data are the two main sources of mobility data available in France for urban areas.

The commuting data was provided by the population census conducted by the French National Institute for Statistics and Economic Studies (INSEE - *Institut National de la Statistique et des Etudes Economiques*). Since 2007, the French census is only conducted continuously each year on a part of the population. The data for each year are obtained by processing that collected during five years. The collected data include the location of the individual's home and place of work or study. These are used to produce the commuting matrix which lists the municipality in which the individual lives and that in which he/she works or studies. Based on the assumption that individuals travel every weekday, it is possible to produce a matrix that contains the flows from home to work and place of study at the municipal level. INSEE[†] provides an order of magnitude for the statistical precision of this flow data.

The *Enquête Globale Transport* (EGT) is “the main source of information about the trips made by the population of Ile-de-France since 1976” (STIF, 2010). It is a household travel survey which was last performed in 2010 when it included 18,000 households and a total of 43,000 individuals and 150,000 trips. The survey covered the population of the Île-de-France Region aged over five years. The region was divided into 109 sectors each of which contained approximately 100,000 inhabitants. Between about 400 and 500 households were surveyed in each of the 109

[†] “However, in view of, in particular, the sampling, low flows (less than 200 individuals) should be considered merely as orders of magnitude” (INSEE, 2012)

sectors, so as to construct a geographically stratified random sample. The survey collected sociodemographic data on the household and each of its members. All individuals aged five years and over were then interviewed personally in order to collect all the trips that were made the day before the survey day. The principal characteristics of each trip were collected, in particular the time it started and ended and the origin and destination zone using a grid with 100 metre squares.

The commuting data are produced at the national level and therefore make it possible to identify all the home-based trips to work or study made by individuals residing in France with at least one end in Ile-de-France. However, they have the shortcoming of only covering home-based trips that are made for the purposes of work or study to the exclusion of all other purposes. Conversely, the data from the EGT relate to all trip purposes, but only cover residents of the Ile-de-France Region. Neither of the databases therefore completely matches the mobile phone data which cover all individuals using the Orange network who are present in the Ile-de-France region, irrespective of where they live or the purpose of their trip. These differences need to be taken into account when the origin-destination matrices obtained from each database are compared.

4. Construction of the origin-destination matrices

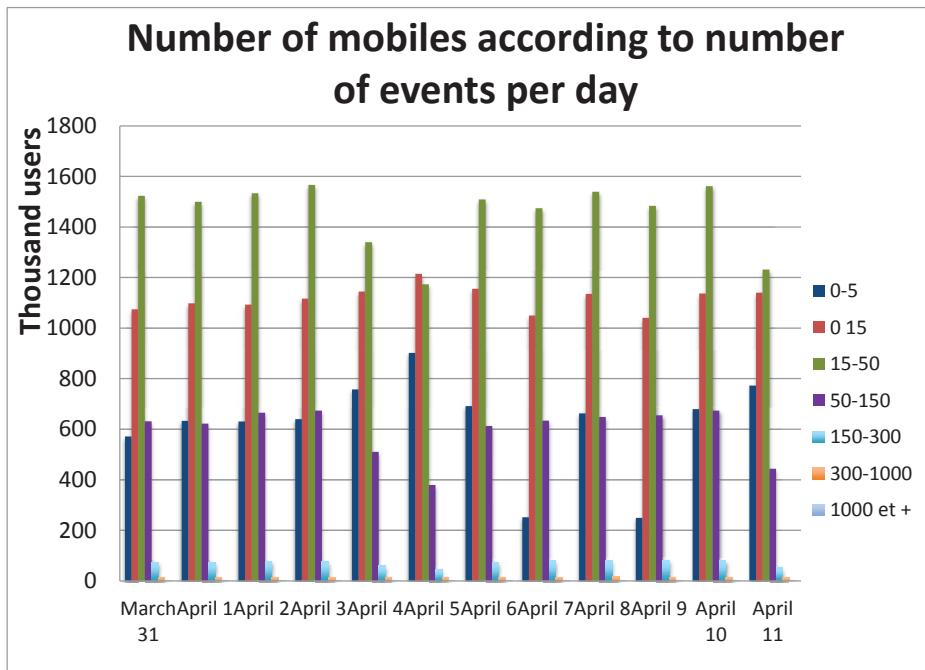


Fig. 5. Distribution of the numbers of users based on the number of events per day. Source of data: Orange

A trip has been defined for the purposes of the EGT as follows by CERTU (2008): a “*trip is the movement of one person conducted for a certain purpose on infrastructure open to the public, between an origin and a destination with a departure time and an arrival time using one or more means of transport*”. It is therefore necessary to specify an origin and a destination which will correspond to a purpose therefore a stationary activity in order to apply the CERTU’s definition. We shall deal first of all with our decision to use the signalling database rather than the billing database. The second provides data for each base station, but only when the mobile phone is communicating (sending or receiving an SMS or call in the case of the mobile phone data for 2009). Consequently the amount of information available depends to a very high degree on the amount of mobile phone activity. As the amount of mobile phone activity an individual engages in is strongly correlated to sociodemographic characteristics such as age, there is a risk of bias if this data is used to construct origin-destination matrices. Moreover, the location is only

known when communication takes place. It is therefore not possible to determine the precise location of the mobile phone throughout the day. It is consequently much more difficult to develop hypotheses in order to identify the individual's stationary activities, particularly for phones which are not frequently used, in view of the fact that there are less than 15 events per day for almost half the mobile phones (Figure 5).

However, the signalling data gives us the position of the phone on a permanent basis, but only at LA level. When a mobile phone changes LA an event is generated (a location area update or LAU) in the signalling file, but not in the billing file. The signalling file also contains an LAU every six hours if the mobile phone has remained inactive (new probing systems do this every 3 hours). This database thus allows us to track the mobile phone in a spatially continuous manner and with a maximum time step of six hours on condition that it does not move outside Ile-de-France and remains connected.

The way the trip is defined means that we have to identify an origin and a destination, and hence a stationary activity at the origin and another at the destination. The size of the LA means that most trips between two LA are made by motorised transport, except for adjacent LAs, but in this case trip duration is in general relatively short. In view of the mean speed of motorised trips in each LA as reported in the data from the EGT, we have made the assumption that if an individual is present for at least one hour in an LA he/she performed a stationary activity there and therefore that the origin or the destination of a trip is located in it[‡]. In order to determine that an activity has taken place, we therefore need at least two events. To determine a trip has been made (therefore an activity has been performed at the origin and at the destination) we therefore need at least 4 events. To reduce the size of our events database mobile phones with three events or less were excluded (Figure 6).

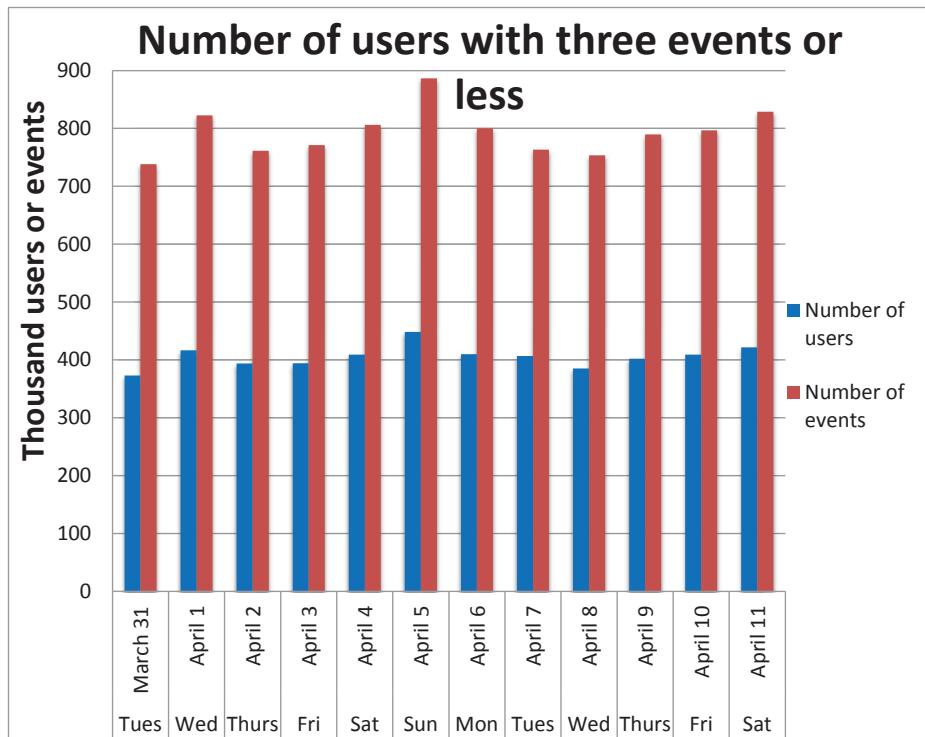


Fig. 6. Number of users with 3 events or less and their respective number of events. Source of data: Orange

[‡] We shall return to this assumption in the conclusion

We have also made allowances for the ping-pong effect (Pollini, 1996, Pierdomenico et al., 2011) which occurs when a stationary mobile phone changes base station (this may be for a number of reasons, in particular base station traffic load). The method we have used is similar to that described by (Iovan et al., 2013).

After this data processing and applying the assumption of stationary time, we can construct an origin-destination matrix. It matches the zoning of Ile-de-France into LA. Obviously, this zoning is the outcome of the telecom operator's needs and has nothing in common with the zonings used in travel surveys whether these are intended to consider commuting (municipality level zoning) or the EGT (a grid with 100 metre squares). We therefore constructed a conversion matrix to make the transition between the different zoning systems. The spatial mobility within a zone is to some extent proportional to the population of the zone and the activities conducted there. We can obtain the population of a zone, but it is not straightforward to obtain the volume of possible activities in it. We therefore used building polygons from the BDTopo database produced by IGN France (the French Mapping Agency). If we make the assumption that origin-destination pairs are uniformly distributed within the built-up zone, we can construct a conversion matrix to move between the different zoning systems. Let us take an example of a zone C_i in the EGT that generates N_{Ci} trips. In order to simplify the calculation, let us assume that this zone straddles two location areas, LA_1 and LA_2 . Using the BDTopo database, for each C_i , it is possible to compute the proportion of the built-up surface area that corresponds to LA_1 and LA_2 , which are denoted respectively by $p(LA_1)$ and $p(LA_2)$. The N_{Ci} trips can be then distributed in the zone C_i using the following formula:

$$\text{Number of trips generated by } LA_1 = p(LA_1) * N_{Ci}$$

$$\text{Number of trips generated by } LA_2 = p(LA_2) * N_{Ci}$$

Generalisation is straightforward for all the zoning systems for both generation and attraction, which means we can construct conversion matrices in order to move from one zoning system to another and thus estimate the trip matrices obtained from the commuting data and the EGT with LA zoning.

5. Comparison between the trip matrices obtained from mobile phones with those obtained from travel surveys

The daily origin-destination matrix obtained from mobile phone data only contains the trips made by individuals who use Orange's network. However, the other two matrices contain data for the entire French population or the entire population of the Ile-de-France Region. We therefore need to adjust the mobile phone data. The penetration rate of mobiles using Orange's network is not precisely known in the Ile-de-France Region and we have no information about the sociodemographic characteristics of these mobile users for reasons of confidentiality and privacy. We are therefore forced to make a new assumption. As we know the (anonymised) identifier of each mobile phone, we are able to estimate the number of mobile phones which use the Orange network every day. If we assume that mobile phone users are representative of the population of Ile-de-France, we can determine a daily expansion factor f_i thus:

$$f_i = \frac{\text{population of Ile - de - France}}{\text{number of persons using network}}$$

This is obviously a strong assumption:

- We have no data that allow us to check that the travel practices of Orange network users are representative of the entire population of mobile phone users. We do however know that Orange is the principal mobile phone operator in France, with about a third of the market. We can therefore hope that the population of Orange users in Ile-de-France is not too atypical;
- Some members of the population, children and older people in particular, do not own a mobile phone. These people have a much lower level of mobility than the rest of the population;
- Some of the people who use the telecom network in the Ile-de-France Region do not live in the region. It is possible to identify individuals who live abroad and exclude them. Identifying where other users live is more complex and we have preferred to avoid this issue to begin with, as the first straightforward analysis which we attempted gave poor results. Once again, it is likely that the mobility within Ile-de-France of individuals who do not live in the region differs from that of individuals who do;

- Some of the individuals who live in the Ile-de-France Region are not in the region on the days when data was collected and so have zero mobility within the region;
- Last, we do not have a sufficient number of events to be able to identify a trip in the case of a non-marginal proportion of mobile phones (of the order of 10%, Figures 6 and 4). We therefore excluded these mobile phones from the working database. Some of these mobile phones necessarily belong to individuals who do not live in Ile-de-France and who are just entering, leaving or passing through the region, but other mobile phones might also be switched off throughout the day (because a mobile phone that is in Ile-de-France and switched on throughout the day should generate, even if it is stationary and not making calls, at least a location area update every 6 hours and therefore at least 4 events a day). This would lead to an underestimation of mobility.

These general comments aside, it would be risky to attempt to precisely estimate biases. As the effects of biases are to some extent contradictory, we have made the (strong) assumption that they compensate for each other. We therefore have matrices that can be compared in order to analyse the number of trips they contain, but also their structure in terms of origin-destination pairs.

5.1. Comparing the mobile phone data with the commuting data

Table 1 shows that the mobile phone data lead to a marked overestimation of the number of trips. This overestimation is however understandable insofar as commuting trips consist only of trips between an individual's home and their place of work or study. On average, these trips involve longer distances than trips for other purposes and therefore have a greater likelihood of resulting in a change of LA. However, analysis of the EGT data show that a high proportion of long trips are made for purposes other than commuting.

Table 1: Number of trips in the Ile-de-France Region based on the commuting data and the mobile phone data (working days only). Source of date of the data: Orange, INSEE

	Commuting matrix	Mobile phone matrix
Number of trips	8,926,000	13,494,000
Mean for each origin-destination pair	9,000	13,600

Looking beyond this major disparity, we attempted to analyse the structure of the two matrices in order to identify any similarities, by means of a variety of analyses (Bonnel et al., 2013). We shall present here the analysis of the correlation between the two matrices after linearisation. The aim was to identify a coefficient of proportionality between the number of trips in each cell of the two matrices (Figure 7). We obtained the following result, where y_{ij} is the number of trips given by the mobile phone matrix for the O-D pair ij and x_{ij} is the number of commuting trips for the same O-D pair:

$$y_{ij} = 1.36 * x_{ij} + 1\ 332, \text{ with } R^2=0.82; \text{ student t for constant } = 4.5 \text{ and slope } = 66.9$$

The constant is relatively small compared to the mean number of trips on the O-D pairs, but it is not null. The R^2 value of 0.82 is acceptable, but when the regression plot (Figure 7) is analysed, we can see there are a large number of origin-destination pairs which are at some distance from the regression line. Even if in very general terms the structure of the O-D matrix is similar, there are clearly quite major deviations from the regression line.

However, the commuting data do not cover all trip purposes, and we should therefore expect deviations. Moreover, analysis of the comparison with the data from the EGT which contains all trip purposes strikes us as being more promising.

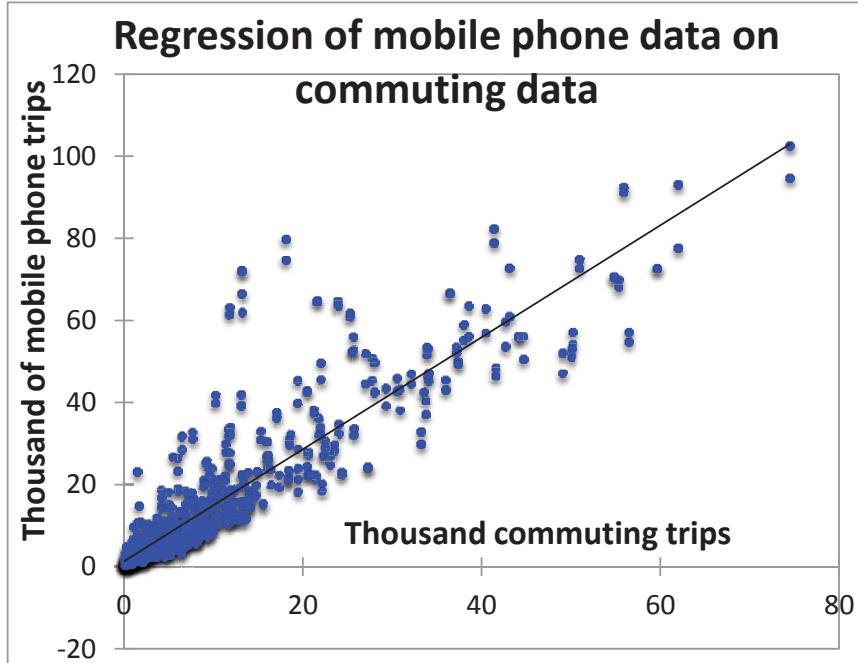


Fig. 7. Linear regression of mobile phone data (working days only) on commuting data. Source of data: Orange, INSEE

5.2. Comparison between the mobile phone data and the data from the travel survey (EGT)

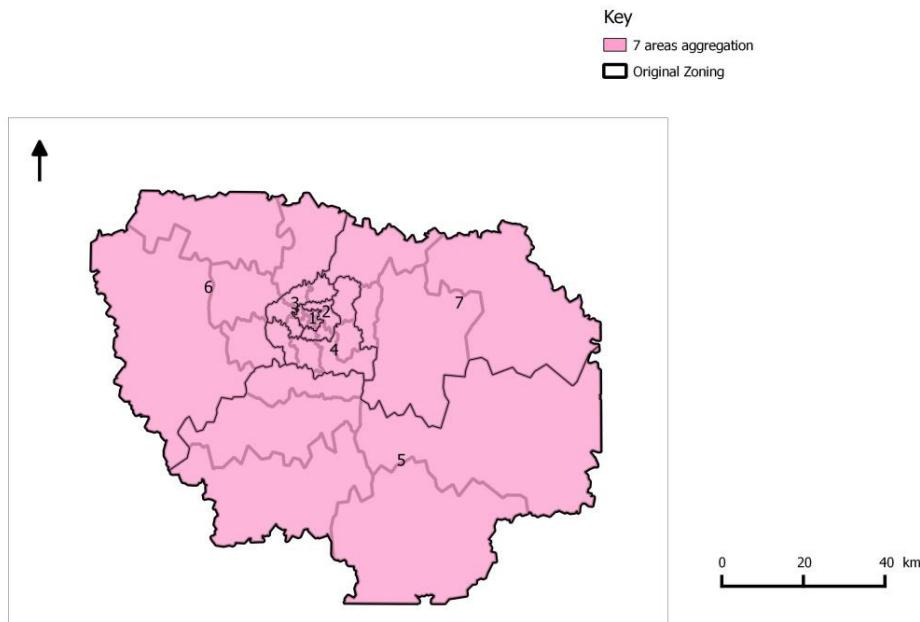
The data from the EGT contain all the trips made by residents of the Ile-de-France Region irrespective of the purpose and the duration of the activity on an average working day. However, we have made an assumption of a minimum of one hour stationary time in an LA in order to identify an origin or a destination in the case of mobile phone data. We have therefore applied the same assumption to the data from the EGT in order to exclude very short activities which cannot be identified as a result of our stationary time assumption. Last, analysis of the total survey sample of the 32-zone EGT zoning system shows that the confidence intervals are very wide for many O-D pairs. We therefore aggregated the location areas in order to produce seven-zone origin-destination matrices (Figure 8) which give a sufficient number of trips for almost all the origin-destination pairs in the EGT. This makes it possible to make a comparison with the mobile phone data matrix which has also been aggregated to correspond to give the seven-zone zoning system.

We can state that the number of trips estimated on the basis of mobile phone data is similar to the estimation based on the EGT, as the difference between the two is less than 10% (Table 2). Furthermore, the regression provides excellent results, and according to the R^2 value almost 100% of the variance is explained by the regression. The slope is close to 1, even if the constant is not null. It is nevertheless relatively small compared to the mean value for the O-D pairs.

$$y_{ij} = 0.963 * x_{ij} + 28\,230, \text{ with } R^2=0.96; \text{ student t for constant } = 3.46 \text{ and slope } = 30.8$$

Table 2: Number of trips from mobile phone data (working days only) and the EGT (division into 7 zones). Source of data: Orange, STIF

7 zones	
EGT	8,739,000
Mobile phones	9,601,000



Analysis of the regression plot (Figure 9) shows that all the points are fairly close to the regression line. However, when we calculated the percentage disparity between the mobile phone data and those from the EGT for each O-D pair (Table 3), we observed that some of these percentage differences were very high. In all cases these corresponded to low flows, which explains why their values in absolute terms are weak. Most of them relate to flows in the outer suburbs or between the second suburban ring and the centre of the Paris conurbation.

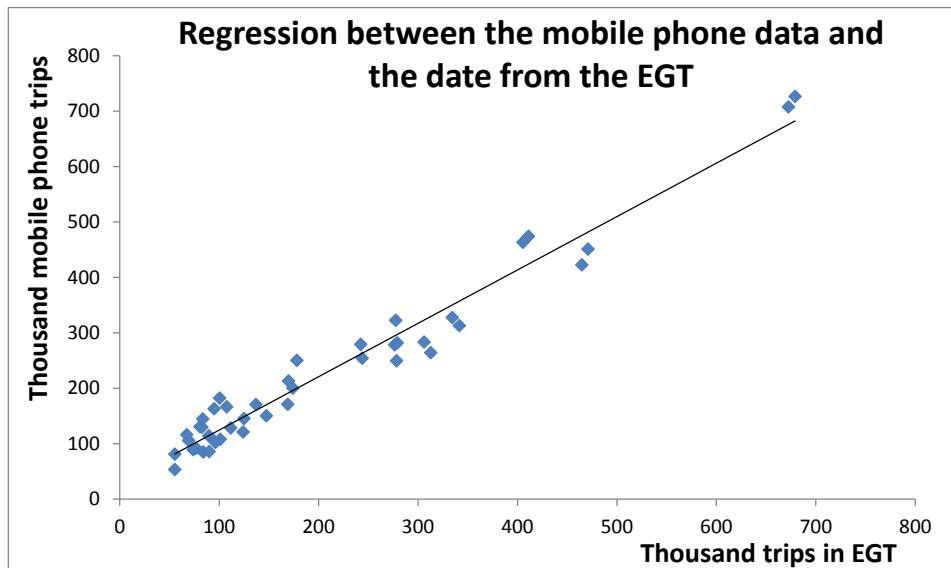


Fig. 9. Regression plot between the mobile phone data (working days only) and the data from the EGT. Source of data: Orange, STIF

We are not really able to explain them in view of the fact that there are many possible sources of error, beginning with the hypotheses which we have been forced to make throughout this study. In addition, the boundaries of the LA do not coincide with the administrative boundaries of Ile-de-France. It is therefore possible that some of the trips that have been assigned to Ile-de-France also involved neighbouring regions. This may go some way to explaining our over-estimates, but it certainly cannot be the sole cause of the disparities.

Table 3: Origin-destination matrix for the percentage difference between the mobile phone data (working days only) and the data from the EGT, Source of data: Orange, STIF

	1	2	3	4	5	6	7
1		5%	15%	4%	1%	7%	1%
2	7%		-2%	-9%	7%	15%	-2%
3	14%	-8%		-16%	-4%	-5%	1%
4	15%	-4%	-7%		15%	-10%	2%
5	62%	81%	46%	41%		73%	72%
6	72%	55%	27%	1%	57%		21%
7	26%	16%	16%	25%	53%	20%	

6. Discussion and conclusion

As many studies have already shown, mobile phone data allow us to construct origin-destination matrices. These matrices were generated from signalling data collected in the spring of 2009 on the mobile phone network operated by Orange in the Ile-de-France Region. However, to our knowledge, and as has been very recently highlighted by Chen et al. (2014), the origin-destination matrices generated with this type of data have never been validated at the scale of a region like Ile-de-France (12 million inhabitants with an area of 12,000 km²). More generally, world-wide, very few studies have been undertaken to validate the travel or traffic data obtained from mobile phone data (White, Wells, 2002; Caceres et al., 2007 Mellegard, 2011; Wang et al., 2012; Bekhor et al., 2013; Calabrese et al., 2013; Chen et al., 2014).

This work is therefore among the first studies in this area. Chen et al. (2014) have attempted to validate the mobility data produced from mobile phone data by constructing ad hoc data that provided an accurate portrayal of “reality”. Most other studies have attempted to compare mobile phone data with external sources of mobility data. These take quite a large variety of forms, for example road traffic counts in the case of Caceres et al. (2007), odometer data obtained from annual vehicle inspections in the USA in the case of Calabrese et al. (2013), but most commonly, the comparison has been with travel survey data. This is what we have done with the commuting data and the data obtained from the *Enquête Globale Transport* (EGT) which is a household travel survey carried out in 2010 of a representative sample of the population of Ile-de-France.

The comparison between the matrices obtained from commuting and mobile phone data provide quite limited results. It is reasonable for the estimated number of trips to differ, as commuting data only relates to trips from the individual’s home to their place of work or study, while mobile phone data cover all trip purposes. But the analysis we have performed also show that the structures of the matrices have little in common. There is thus a high degree of dispersion in the rates of variation between the two sources of origin-destination data. This preliminary work was conducted on the basis of fairly strong hypotheses. It is therefore quite possible that more detailed analysis would make it possible to moderate some of the strongest hypotheses and improve comparability. Nevertheless, the only trip purposes covered by commuting data are work and study. It would therefore be necessary to be able to estimate the location of the home and place of work or study of the individuals from mobile phone data in order to significantly improve the matrices produced from this source. Even if a number of algorithms have been described in the literature (Chen et al., 2014; Calabrese et al., 2013; Phithakkitnukoon et al., 2012), our identification of these locations is bound to be uncertain unless we have a large number of events for each mobile phone, which is not the case with the data that we have used. It seems certain that the analysis of the mobile phone data from smartphones which are frequently connected to web-based applications would make it possible to attempt this type of analysis.

Comparison with the data from the *Enquête Globale Transport* (EGT) performed by STIF in 2010 is much more promising. This database has the advantage of covering all trip purposes, not just those related to work and study. However, it only contains individuals who live in the Ile-de-France Region. Using this database we were able construct a matrix that set out to reproduce the assumption of minimum stationary time within a location area which is necessary in order to produce origin-destination matrices from mobile phone data. In view of the size of the survey (43,000 individuals and 150,000 trips), zones were aggregated into seven zones in order to reduce the confidence interval for each element in the EGT matrix. This meant that we were able to obtain a total number of trips in Ile-de-France from the mobile phone data that was similar to that given by the EGT (a difference of 9%). Above all, the linear regression we performed on the number of trips in each element in the two matrices showed that the structure of the two matrices is very similar with an R^2 value of 0.96 and a slope that is very close to 1. But these very encouraging results should not distract us from the fact that the results are less satisfactory in the case of some origin-destination pairs for which the disparities can attain 70 to 80%, even if in terms of numbers, the disparities are smaller as the largest percentage disparities are for those origin-destination pairs with a fairly small number of trips.

A large number of hypotheses need to be made to construct trip matrices from mobile phone data. In order to identify possible approaches for further investigation, we shall restate these below:

- The mobile phone data related to all the trips made by individuals who were present in the Ile-de-France Region. However, the data from the EGT only covered Ile-de-France residents. It would therefore be interesting to attempt to identify where the mobile phone owners in the database live. As we have already mentioned, this has not been done because of the short period covered by our data. However, it would be worthwhile to carry out a similar analysis on smartphone data for which there are many more events for each mobile phone, and possibly on data that covers a longer period. This would make it possible to extract solely the residents of the Ile-de-France Region in order to improve the validity of the comparison with household travel survey data. This of course assumes that the travel of smartphone owners is representative of that of the population as a whole, which remains to be verified;
- We have applied a uniform assumption of minimum stationary time of one hour for all the location areas. It would certainly be possible to refine this and vary it according to the characteristics of each LA in terms of surface area and travel speeds. Moreover, the duration of one hour is necessarily somewhat arbitrary, even if it was based on an analysis of the data from the EGT. We have therefore tested the impact of this threshold on the number of trips generated. This analysis was conducted on a single day (Figure 10). It shows that the results are highly sensitive to this assumption. It would therefore be of interest to be able to refine this threshold for each LA and also test the sensitivity of the number of trips to the selected threshold as it is by no means certain that changing the threshold would lead to a proportional change in all the matrix elements.
- The data obtained contain spatial information whose resolution corresponds to the location areas. However, the events in the Orange database pass through base stations. This information is potentially useful. Initially, the zoning would not be changed, as only the database which contains changes in LA gives the mobile traces the spatial continuity which our trip generation method requires. It would however enable us to refine our analyses, in particular as regards ping-pong effects, or alternatively refine our assumption as regards the minimum stationary time within an LA;
- After this, it would be interesting to analyse the data from 3G probes which monitor the exchange of data in addition to phone calls, SMS messages, LA updates and handovers. These databases contain more events, allowing us to make novel hypotheses for trip generation (Chen et al., 2014);
- The boundaries of the location areas are identified by analysing Voronoi polygons. This means there is a high degree of uncertainty about the boundaries. Moreover the actual limits to the coverage of the base stations varies according to mobile phone traffic, the weather and the local topography. It would be interesting both to refine the base station boundaries and hence those of the LA and also study the impact of uncertainty about LA boundaries on the construction of origin-destination matrices;
- Each database uses its own zoning. This means we have to construct conversion matrices to convert one zoning system into another. The analysis we have conducted is based on the surface area of the built up part of each zone. New databases are now available that contain not only the built-up surface area, but also the built volume. INSEE has recently started to make census data available which is much more fine-grained than the IRIS database (Aggregated Units for Statistical Information - *Ilots Regroupés pour l'Information Statistique*). This

makes it possible to assign individuals and jobs to each block of buildings, considerably increasing the accuracy of the conversion matrices that are used for the transformation from one zoning system to another (Manout, 2014);

- The expansion of matrices obtained from mobile phone data is based on the very simple assumption that the population of mobile phone owners for whom we have been able to constitute at least one trip is representative. It is unlikely that we will be able to access demographic data on the users of the Orange network for obvious commercial reasons, but it is not impossible to try to collect information from other sources. Calabrese et al. (2011) and Bekhor et al. (2013) have analysed the spatial distribution of mobile phone users by comparing it to census data. Bekhor et al. (2013) have also used travel survey data which contained questions about mobile phone use. These data could be used to identify any bias affecting the samples of mobile phone data in order to adjust the data using travel data from household travel surveys;
- Finally, we undertook no analysis of the data for mobile phone owners for whom we had fewer than four events. It would nevertheless be useful to identify those who switch their mobile phone on or off during the study day in order to distinguish between them and individuals who entered or left the study zone during the day.

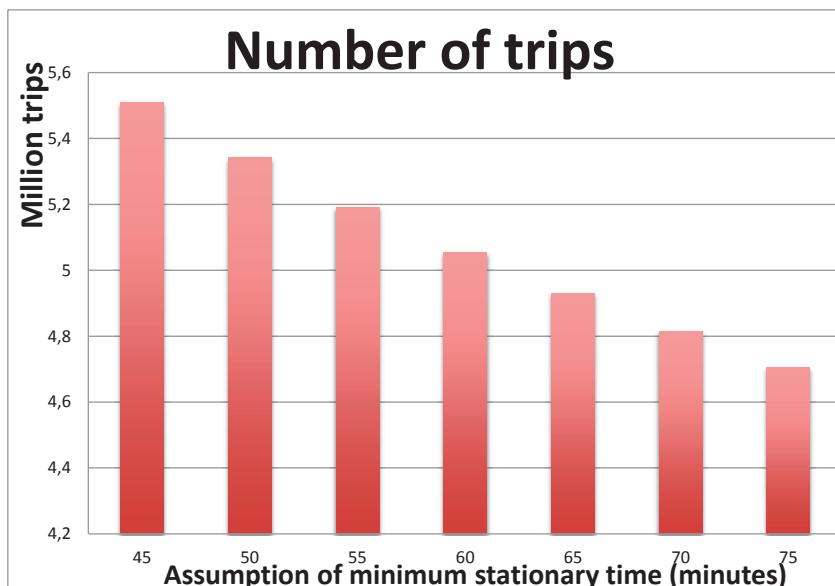


Fig. 10. Number of trips made according to the minimum stationary time assumption. Source of data: Orange

Mobile phone data therefore seem very promising for the analysis of spatial mobility, but a considerable amount of further research is required in order to be able to fully validate their use in order to construct origin-destination matrices for transport modelling or transport planning purposes.

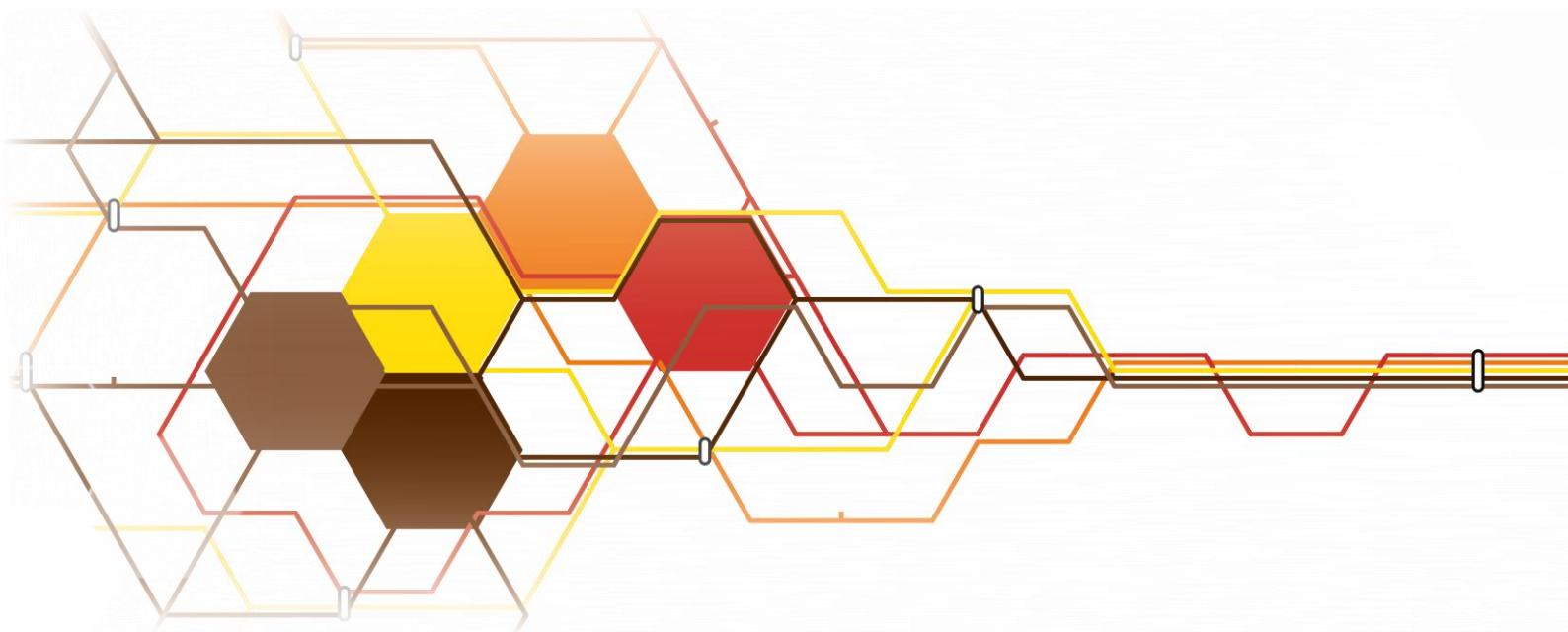
Acknowledgements: The authors would like to thank Orange for giving them access to the mobile phone data collected in the Ile-de-France Region and STIF for making the EGT data available. However, the authors alone are responsible for the contents of this paper.

References

- Akin D, Sisiopiku V (2002), *Estimating Origin-Destination Matrices Using Location Information from Cellular Phones*, Proc. NARSC RSAI, Puerto Rico, USA.
- Ampt ES (1997) Response Rates - Do they matter? In: Bonnel P, Chapleau R, Lee-Gosselin M, Raux C (eds.) *Les enquêtes de déplacements urbains: mesurer le présent, simuler le futur*, Programme Rhône-Alpes Recherches en Sciences Humaines, Lyon, pp 115-125
- Arana P, Cabezudo S, Peñalba M (2014) Influence of weather conditions on transit ridership: A statistical study using data from Smartcards, *Transportation Research part A*, 59 pp. 1-12.

- ARCEP (2014) Autorité de Régulation des Communications Electroniques et des Postes, *Observatoires / Services Mobiles*.
- Arentze T, Timmermans H, Hofman F, Kalfs N (2000), Data needs, data collection, and data quality requirements of activity-based transport demand models, In: *Transport surveys, raising the standard*, TRB transport circular E-C008, pp. II-J-1/30.
- Atrostic BK, Burt G (1999) *Household non-response: what we have learned and a framework for the future*, Statistical Policy working paper 28, Federal Committee on Statistical methodology, Office of Management and Budget, Washington, pp 153-180.
- Bar-Gera H (2007), Evaluation of a cellular Phone-Based System for Measurement of Traffic Speeds and Travel Times: A Case Study from Israel, *Transportation Research Part C*, 15 (6) pp. 380-391.
- Bekhor S, Cohen Y, Solomon C (2013), Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions, *Journal of Advanced Transportation*, 47 (4) pp. 435-446.
- Bolla R, Davoli F (2000), *Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network*, Proc. IEEE WCNC, Chicago, USA.
- Bonnel P (2003) Postal, telephone and face-to-face surveys: how comparable are they? In: Stopher PR, Jones PM (eds.) *Transport Survey Quality and Innovation*, Elsevier, London, pp 215-237.
- Bonnel P, Hombourger E, Smoreda Z (2013), *Quel potentiel des données de la téléphonie mobile pour la construction de matrices origines-destinations de déplacement – application à l'Ile-de-France*, Rapport de Recherche, Laboratoire d'Economie des Transports, Orange Labs, 133p.
- Brisson P (2008), *Global system for mobile communication*. Université de Montreal.
- Caceres N, Wiedeberg JP, Benitez FG (2007), Deriving Origin-Destination Data from a Mobile Phone Network, *IET Intelligent Transport System*, 1 (1) pp. 15-26.
- Calabrese F, Diao M, Di Lorenzo G, Ferreira Jr J, Ratti C (2013), Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C*, 26, pp. 301-313.
- Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating Origin-Destination Flows using Mobile Phone Location Data, *IEEE Pervasive Computing*, 10 (4) pp. 36-44.
- Calabrese F, Pereira F, Di Lorenzo G, Liu L, Ratti C (2010), *The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events*, Proc. Pervasive Computing, Helsinki, Finland.
- CERTU (2008), *L'enquête ménages déplacements standard CERTU*, éditions du CERTU, Lyon, 203p.
- Chen C, Bian L, Mac J (2014), From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C*, 46, pp. 326–337
- Cho E, Myers SA, Leskovec J (2011) *Friendship and Mobility: User Movement in Location-based Social Networks*, Proc. ACM SIGKDD, San Diego, USA.
- Enquête Globale Transport de l'Ile-de-France : http://www.stif.org/IMG/pdf%20Enquete_globale_transport_BD-2.pdf
- Frias-Martinez V, Frias-Martinez E, Oliver N (2010) *A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records*, Proc. AAAI AI-D, Palo Alto, USA.
- Gonzalez MC, Hidalgo CA, Barabasi A-L (2008), Understanding Individual Human Mobility Patterns, *Nature*, 453 (7196) pp. 779-782.
- IDATE (2008), Observatoire économique de la téléphonie mobile – faits et chiffres 2008, *Mobile et société*, 9, pp. 6-15. http://www.fftelecoms.org/sites/default/files/contenus_lies/mobile_et_societe_9.pdf
- INSEE (2012), *Bases sur les flux de mobilité : documentation*. INSEE, Paris.
- Iovan C, Olteanu-Raimond A-M, Couronné T, Smoreda Z (2013), Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies, In: *Geographic Information Science at the Heart of Europe*, Springer, pp. 247-265.
- Isaacman S, Becker R, Caceres R, Kobourov S, Rowland J, Varshavsky A (2010) *A Tale of Two Cities*, Proc. ACM HotMobile, Annapolis, USA.
- Isaacman S, Becker R, Caceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011), *Ranges of Human Mobility in Los Angeles and New York*, Proc. IEEE PerCom Workshops, Seattle, USA.
- Jiang S, Vina-Arias L, Ferreira J, Zegras C, Gonzalez MC (2011), *Calling for Validation: Demonstrating the use of Mobile Phone data to Validate integrated land use Transportation models*, Proc. 7VCT, Lisbon, Portugal.
- Kang JH, Welbourne W, Stewart B, Borriello G (2004), *Extracting Places from Traces of Locations*, Proc. ACM WMASH, Philadelphia, USA.
- Manout O (2014), *Codification des connecteurs de zones pour les transports en commun*, mémoire de master TER, Université Lumière Lyon2 – ENTPE, Lyon
- Mellegard E (2011), *Obtaining Origin/Destination-Matrices from Cellular Network Data*. Master's Thesis, Chalmers University of Technology.
- Morency C, Trépanier M, Agard B (2007), Measuring transit use variability with smart-card data. *Transport Policy*, 14 (3), pp. 193–203.
- Munizaga M, Palma C, Mora P (2010), Public transport OD matrix estimation from smart card payment system data. In: *12th World Conference on Transport Research*, Lisbon, Paper No. 2988.
- Noulas A, Mascolo C, Frias-Martinez E (2013) *Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments*, Proc. IEEE MDM, Milan, Italie.
- Ortuzar J de D, Bates J (2000), Workshop summary, in: *Transport surveys, raising the standard*, TRB transport circular E-C008, pp. II-J/31-35.
- Pelletier MP, Trépanier M, Morency C (2011), Smart card data use in public transit: A literature review, *Transportation Research Part C*, 19, pp. 557–568.
- Phithakkitnukoon S, Smoreda Z, Olivier P (2012), Social-geography of human mobility: a study using longitudinal mobile phone data. *PLoS ONE* 7 (6), e39253.
- Pierdomenico F, Valerio D, Ricciato F, Hummel K (2011), Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data, In: *Traffic Monitoring and Analysis*, Springer Berlin Heidelberg, pp. 66-80.
- Pollini GP (1996), Trends in handover design, *IEEE Commun Mag*, 34(3), pp. 82-90.

- Schlaich J, Otterstatter T, Friedrich M (2010), *Generating Trajectories from Mobile Phone Data*, Proc. TRB Annual Meeting, Washington D.C, USA.
- Smoreda Z, Olteanu-Raimond A-M, Couronné T (2013), Spatiotemporal data from mobile phones for personal mobility assessment, In: Zmud , Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), *Transport Survey Methods: Best Practice for Decision Making*, Emerald, pp. 745-767.
- STIF (2010), *Notice méthodologique de l'Enquête Globale Transport du Syndicat des Transports d'Île de France*, rapport du Syndicat des Transports de l'Île-de-France, Paris.
- Stopher PR, Greaves SP (2007), Household travel surveys: where are we going? *Transportation Research Part A*, 41, pp. 367–381.
- Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider C, Blondel V, Smoreda Z, Gonzalez MC, Colizza V (2014) On the Use of Human Mobility Proxies for Modeling Epidemics, *PLOS Computational Biology*, 10(7), e1003716.
- Wang P, Hunter T, Bayen A, Schechtner K, Gonzalez MC (2012) Understanding Road Patterns in Urban Areas, *Scientific Reports*, 2 (1001) pp. 1-6.
- White J, Wells I (2002) *Extracting Origin Destination Information from Mobile Phone Data*, Proc. IEEE RTIC, London, UK.
- Wolf J, Oliveira M, Thompson M (2003), Impact of underreporting on mileage and travel time estimate – results from Global Postionning System enhanced household travel survey, *Transportation research record*, 1854, pp. 189-198.
- Ygnace J-L (2001) *Travel Time/Speed Estimates on the French Rhone Corridor Network using Cellular Phones as Probes*, INRETS STRIP Project Technical Report.
- Zhang H, Bolot J (2007) *Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks*, Proc. ACM MobiCom, Montreal, Canada.
- Zmud J (2003) Designing instruments to improve response: keeping the horse before the cart, In: Stopher PR, Jones PM (Eds) *Transport Survey Quality and Innovation*, Elsevier, Pergamon, Oxford, pp 89-1



The construction of purpose specific OD matrices using public transport smart card data

W. Kuhlman

Colophon

This document contains the master thesis report of Wouter Kuhlman, in fulfilment of the requirements for the degree of Master of Science in Civil Engineering, at Delft University of Technology. The research was performed at Panteia, located in Zoetermeer.

The thesis report has been submitted on Monday October 19th 2015, and will be publically defended on Monday October 26th 2015, at the faculty of Civil Engineering and Geosciences of Delft University of Technology.

An electronic version of this document is available at <http://repository.tudelft.nl>.

For any questions regarding the content, please contact:

- Wouter Kuhlman (author) wouterkuhlman@gmail.com or
- Jan Kiel (Panteia) j.kiel@panteia.nl or
- Bert Schepers (Panteia) b.schepers@panteia.nl

Assessment committee

- Prof.dr.ir. E. de Romph – chair (Delft University of Technology)
- Dr.ir. R. van Nes – daily supervisor (Delft University of Technology)
- Drs. J. Kiel – company supervisor (Panteia)
- Dr.ir. M. Kroesen – external supervisor (Delft University of Technology)
- Ir. P.B.L. Wiggenraad – thesis coordinator (Delft University of Technology)

- Drs. B. Schepers – external advisor (Panteia)
- Drs. S. Kieft – external advisor (SRA)
- Drs. N. in 't Veld – external advisor (GVB)

Key words

Smart card data; travel survey data; data enrichment; public transport; OD matrix; travel purpose inference; travel demand modelling

Document statistics

Pages: 149 (127 main text)

Words: 52.049

Chapters: 9

Tables: 27

Figures: 40

Appendices: 4 (external document)

Preface

During my master I got fascinated by the possibilities of smart card data in travel demand studies. This led me to do an internship investigating the potential of a combination of different data sources. Unfortunately, we did not obtain any smart card data during my internship. So, when Panteia and GVB, the public transport operator in Amsterdam, provided the possibility to do my thesis using actual smart card data, I was eager to take it.

One year later, this has resulted in the report that now lies before you. It has been an exciting and challenging journey through the many phases of research. I have learned that, after diving into a subject, I need to come up for air every once in a while. Fortunately, many people supported me during the process.

I am grateful for the opportunity to have performed this research at Panteia. I want to thank my colleagues at Panteia: Bert, Gerben, Ferry, Dick, Yuko, thanks for sharing your experience with me. Thanks Jan, for your guidance and the trips we made to several conferences, at which I was allowed to present my research. I also want to thank the other interns at Panteia for the required distractions at the office.

The creation of this report would not have been possible without the expertise of my graduation committee: Erik de Romph, Rob van Nes, Maarten Kroesen, your feedback and expertise greatly improved the quality of my work. Special thanks to Rob for the sessions in which we structured the many possibilities within this research topic.

I would like to thank the GVB and EBS for making their data available for this research. Special thanks go out to Natalie in 't Veld for her help and enthusiasm during my time at the GVB office in Amsterdam. Also the support from Suzanne Kieft from the City region of Amsterdam is greatly appreciated. I am proud that all these parties committed to this research.

I want to thank my parents for their faith and encouragement and, most of all, my girlfriend Petra, thank you for your love and composure.

Wouter Kuhlman
Utrecht, October 2015



Summary

With the introduction of the smart card as fare collection system in public transport, a new data source emerged. The smart card is able to collect a previously unattainable large sample of travel data by recording all check-in and check-out transactions. When a smart card is the only valid ticketing system, smart card data approximately cover the complete public transport demand. Therefore, smart card data are considered as a rich data source for an abundant amount of topics, predominantly for the description of travel demand.

Problem description

Currently, travel demand models are not able to accurately represent the public transport travel demand. Travel surveys and counts are required to increase the accuracy by means of enrichment and calibration processes. However, collecting this empirical data is expensive. The collection of smart card data is already incorporated in the system and allows for the collection of large data samples at low costs.

Smart card data are very accurate in space and time, as transactions are stored together with their corresponding stop and the time of the transaction. In case the smart card is the only valid ticketing system, like the Dutch OV-chipkaart, the described volumes also accurately represent the volume of the public transport travel demand. Moreover, the continuous data collection allows for longitudinal description of the travel demand.

Yet, the smart card data are also limited regarding their usability in travel demand studies. Smart card data lack information about exact origin and destination locations, where activities are performed. Transactions are recorded at the used stops, hence smart card data do not consider access and egress trip legs to and from the stops. Stop locations are not considered to be stable indicators of the travel demand, as they are not to the actual locations where activities are performed. The used stops relate to the route choice and thus depend on the public transport supply.

Moreover, the interpretability of the travel data is limited due to the passive data collection. Smart card data do not contain information about travellers' motivation to travel, nor personal characteristics and preferences. Because of this lacking information, the interpretation of the travel demand is limited and, consequently, the forecast capabilities of travel demand models using these data are too.

Research motivation

The main objective of this study is the improvement of public transport travel demand forecasts. In order to obtain this objective, we have aimed at increasing the interpretability of smart card data so they can be used as input data for travel demand models.

By enriching smart card data with information about travel purposes and activity locations, smart card data can be used to describe the current travel demand by means of OD matrices. With the high volumes and veracity of smart card data, this results in a more accurate description of the current travel demand. Consequently, travel demand models can provide more accurate forecasts, as the description of current travel demand constitutes the foundation of travel demand forecasts. In



In addition, increased interpretability is required for other applications of smart card data in travel demand modelling.

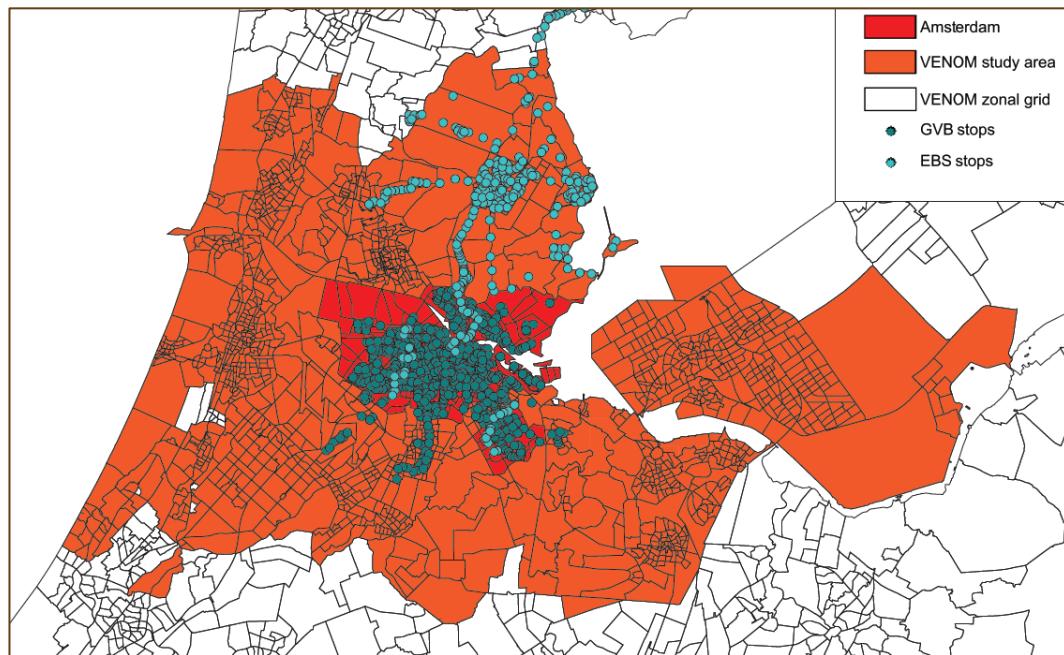
Methodology and data

In order to enrich smart card data, we have developed enrichment models that estimate the lacking information. In order to comply with the highly disaggregated structure of smart card data, travel is represented by individual trips. Three enrichment procedures are applied to every trip, which (1) allocate trips to origin zones, (2) allocate trips to destination zones and (3) infer the travel purpose.

The enrichment models are estimated on data collected by travel surveys. The required information in the survey data consists of the information to be estimated, the dependent variables, and attributes that explain the information to be estimated, the independent variables. Since the models have to be applicable to smart card data, the independent variables are limited to key variables, which are available in both smart card data and survey data.

Data

The smart card data available for this study consisted of the OV-chipkaart data of the year 2014 for two public transport concessions: Amsterdam and Waterland. These concessions are dissimilar, as Amsterdam is highly urbanized and includes bus, tram and metro services, while Waterland is a more rural area to the north of Amsterdam, with only bus services. Data of the national railways and adjacent concessions were not available.



The survey data used for estimation of the enrichment models consist of the WROOV studies, performed yearly between 2003 and 2009. The data for these seven consecutive years are stacked, which results in a dataset of 1.7 million trips with bus and light rail in the Amsterdam and Waterland concessions. The WROOV data include trips by bus and light rail, made with any public transport ticket before implementation of the OV-chipkaart, except student cards and local tickets.



The land-use data available for this study originate from the strategic transport model of the Amsterdam City Region: VENOM. The VENOM model uses a relatively high resolution zonal structure, with corresponding land-use data from the base year 2010.

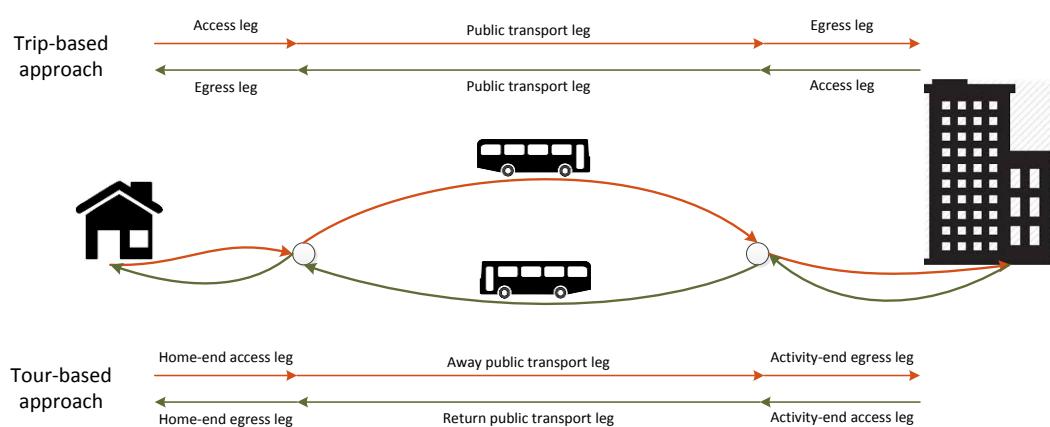
Inevitably, the available data have several limitations. Transfers to and from the train network are not observed, since data from the national railways is not available. The WROOV surveys did not include the student public transport card and local tickets. As a result, the survey data is not suitable to estimate enrichment models for students and tourists. For this reason, we have filtered students (20%) and short term contracts (9%) from the OV-chipkaart data.

Methodology

Literature suggests rule-based methods for the allocation of origins and destinations and the inference of the travel purpose. However, deterministic rules imply crude simplifications. While this may be appropriate for the interpretation of system averages, it is not for the construction of purpose-specific OD matrices. The simplifications result in inaccurate allocation of trip to OD pairs and to wrongly inferred purposes. Therefore, we propose probabilistic methods for the allocation of trips to origin zones, destination zones and the inference of the travel purpose. The enrichment procedures are based on logit allocation models, a technique derived from discrete choice modelling.

We have applied two approaches for the probabilistic enrichment of smart card data: a trip-based approach and a tour-based approach. Trips describe travel between successive activity locations: the origin and destination. The available attributes consist of trip characteristics and land-use characteristics at both trip-ends.

Tours describe travel starting and ending at the home location, via one or several activity locations. The tour-based approach defines trip-ends by their geographical location: the home-end and activity-end, instead of the chronological order in the trip-based approach. This distinction allowed for the application of land-use data related to homes or activities. Regarding the purpose inference, the tour-based approach also included the additional attribute activity duration.



We have compared the probabilistic approaches with a rule-based method at different levels of zonal resolution in order to assess the accuracy increase at each level. In order to do so, we applied the enrichment models to the WROOV data to construct purpose-specific OD matrices for each method. Subsequently, we compared the



resulting matrices with the observed WROOV OD matrix by means of linear regression on the cell values.

In addition, we applied the enrichment models to the OV-chipkaart data and compared the resulting matrices with the WROOV matrices to assess the difference of the described travel patterns between the sources.

Results

Analysis of travel patterns in the WROOV data

Regarding the zonal allocation models, the distribution of access and egress distances proved to differ substantially between the used modes. Especially for metro, travellers cover larger distances to stops on average. Moreover, a distinction is visible between the home-end, where access and egress distances mainly depend on the level of service, and the activity-end, where the access and egress distances depend more on the travel purpose.

Regarding the purpose inference, the distribution of travel purposes in the survey data show large deviations over many key variables, including the departure time, activity duration, travel frequency and ticket types. Age and gender are also correlated to the travel purpose, but not available in OV-chipkaart data and thus not applicable in the enrichment models.

A quantitative comparison of the travel described by the WROOV data and the OV-chipkaart data regarding key variables shows several dissimilarities. Most imbalances are related to the same problem: the overrepresentation of contracts in the WROOV data, resulting in an overestimation of work trips. These dissimilarities could also be caused by the time gap between the sources and different data collection methods. Moreover, the WROOV data mainly consist of tours with two trips due to the survey format, where the OV-chipkaart data also contains tours with more than two trips.

Enrichment models

In the trip-based zonal allocation models, similar results were found for the origin inference and the destination inference, due to the high share of tours in the dataset. Three attributes proved to have a significant explanatory value to the origin and destination zones:

- The share of the catchment area in the zone [%];
- The stop density [stops/ha];
- The level of urbanization [addresses/ha].

The share of the catchment area in the zone relates to the nearness of a zone to the used stop in terms of area. It is positively related to the probability of alternatives and has the largest effect on the model fit. The stop density and the level of urbanization relate to the production and attraction of zones. Where the level of urbanization directly relates to potential activity locations, the stop density is an indirect indicator. Although direct indicators are preferable, the stop density was a better and more stable indicator than other direct attributes related to activity locations, such as the number of jobs and the number of student places in a zone.

In the tour-based approach, the same attributes are included in the home zone allocation model. However, the effect of the level of urbanization is substantially larger. On the other end, in the activity zone allocation, the effect of the level of



urbanization is not a stable indicator and thus excluded from the model. As a result, the home zone allocation model has a better fit. The attributes related to the activities, like jobs and student places, did not have a stable influence on the probability of zones.

The trip-based purpose inference models include five attributes, four categorical and one numeric:

- The concessions travelled in;
- The contract duration;
- The travel frequency [tours/week];
- The used modes;
- The departure time.

A comparison of the estimated probabilities and the observed travel purposes in WROOV show that the model is very capable of distinguishing work trips and other trips. Conversely, the identification of education trips and shopping trips is less accurate. This is caused by the similarities between these less occurring purposes and their more frequently observed counterparts, respectively work and other.

The tour-based purpose inference model has a substantially better fit, mainly due to the inclusion of the activity duration. Nonetheless, the distinction between work and education trips is still limited. The inclusion of the activity duration as categorical variable might improve this distinction.

Matrix evaluation

Assessing the matrices constructed with the different approaches, we conclude that the trip-based and tour-based zonal allocation models result in a significantly better representation of the travel demand than a rule-based approach. Especially at a high resolution, like the VENOM zonal grid, the improvement is substantial. The effect reduces at the level of PC3 zones, which indicates that the effect of access and egress legs is relevant for travel demand models with a higher resolution than PC3 zones. This includes nearly all Dutch travel demand models.

Individual matrix cell values, however, still show substantial deviations. The largest deviations are all related to transfers with the train network. Since the train network is highly intertwined with the urban and regional public transport in the Amsterdam region, the unavailability of train data constitutes a limitation to this study.

The comparisons of purpose-specific matrices verify the high accuracy of the inference of the purposes *work* and *other* and the lower accuracy of the inference of *education* and *shopping* purposes. Between the lower scoring purposes, the shopping matrix indicates a better fit than the education matrix. This leads to the conclusion that education trips have more specific spatial patterns, which deviate from commuting patterns, while shopping trips have a similar pattern as trips made for *other* purposes. Consequently, an accurate distinction is possible between compulsory purposes, work and education, and discretionary purposes *shopping* and *other*.

Assessing the differences between the matrices constructed with the different sources, WROOV and OV-chipkaart, we conclude that the described travel demand differs substantially. This can be attributed to differences between the samples, an inevitable



issue when coupling two different data sources, and shifted demand between the data collection periods.

Conclusions

The applied methodology of constructing purpose-specific OD matrices based on smart card data shows great potential. During this research, an operational method has been developed, which results in a more accurate description of public transport travel demand than previously available.

Based on the accuracy of the enrichment models and the constructed OD matrices, we conclude that the tour-based approach is the most accurate enrichment method. Furthermore, the tour-based approach contains a higher level of behavioural richness and thus it is preferable over the rule-based approach and the trip-based approach.

However, restrictions of the available data have resulted in limited applicability and durability of the method. The survey data do not cover students in higher education and international travellers. With additional fine-tuning of and increased availability of smart card data, including all public transport operators in the study area, the method presented in this report can be enhanced to a fully applicable approach and lead to valuable improvements of the quality of public transport demand forecasts.

Recommendations

This study has shown that the probabilistic method of enrichment of smart card data improves the description of public transport travel demand and increases the interpretability of the data. However, due to limitations in the available data the constructed matrices do not represent the complete travel demand. Especially the interaction with train travellers is essential to the description of travel demand by OD matrices. Furthermore, the lacking students and international travellers in the survey data resulted in a substantial gap in the total travel demand.

Therefore, we recommend starting a new, online travel survey in connection with OV-chipkaart data. In order to increase the interpretability of smart card data, periodic travel surveys are still required, as the durability of the method requires periodical updating of the model parameters. With a new survey, a representative sample of travel in the complete public transport system can be selected for estimation of the enrichment models. It is recommended to include the smart card data from all public transport operators in the data in order to identify concession traversing transfers.

Regarding the enrichment methodology, we recommend further research on combining the activity zone allocation and the purpose inference into a single allocation model. The data analysis has shown a correlation between the travel purpose and the access and egress distances at the activity end. Moreover, by combining these models, the effects of purpose-specific land-use characteristics are specifically estimated for their corresponding purpose.



Table of contents

Preface	iii
Summary	v
Table of contents	xi
Tables	xiii
Figures	xv
Glossary of terms	xvii
1 Introduction	1
1.1 Problem description	1
1.2 Research motivation	4
1.3 Research questions and approach	5
1.4 Definitions used in this report	7
1.5 Report structure	8
2 Literature review	11
2.1 Introduction to smart card data	11
2.2 Smart card data for OD matrix construction	14
2.3 Travel purpose inference	19
2.4 Conclusions from the literature study	23
3 Methodology and data	27
3.1 The Dutch smart card: the OV-chipkaart	27
3.2 The WROOV surveys	28
3.3 Land use data	29
3.4 Qualitative comparison of the OV-chipkaart and WROOV	30
3.5 Methodology of enriching smart card data	31
3.6 Research outline	33
3.7 Analysis framework	36
3.8 Modelling estimation framework	36
3.9 Matrix evaluation framework	40
3.10 Conclusions regarding the methodology and data	42
4 The Amsterdam region case study	45
4.1 Eventual application of OD matrices in the VENOM model	45
4.2 Availability of OV-chipkaart data	46
4.3 Matching the WROOV dataset to the OV-chipkaart dataset	47
4.4 Generalizability of the case study	48



5	Public transport travel analysis	51
5.1	Access and egress trip legs	51
5.2	Travel purpose	54
5.3	Concession traversing transfers	58
5.4	Quantitative comparison of key variables	60
5.5	Conclusions regarding the travel analysis	65
6	Estimation of enrichment models	67
6.1	Rule based reference models	67
6.2	Probabilistic zonal allocation models	68
6.3	Probabilistic purpose inference models	82
6.4	Identification of concession traversing transfers	93
6.5	Conclusions regarding the model estimations	94
7	Evaluation of OD matrices	97
7.1	Procedure of model applications	97
7.2	Model validation on WROOV data	99
7.3	Evaluation of differences between modelling approaches	104
7.4	Source comparison on travel patterns	105
7.5	Conclusions regarding the matrix evaluation	106
8	Conclusions	109
8.1	Relevant attributes	109
8.2	Transferability of information	110
8.3	Quality of the enrichment models	112
8.4	Matrix evaluation	114
8.5	Answer to the main research question	116
9	Recommendations	119
9.1	Follow-up research	119
9.2	Utilization of the results	122
Bibliography		123



Tables

Table 1: Locations of sub-questions in the report	9
Table 2: Overview of applicable methodologies found in literature	24
Table 3: Methodology assessment criteria.....	25
Table 4: Available attributes in OV-chipkaart data	27
Table 5: Available attributes in WROOV data	28
Table 6: Available attributes in land use data	29
Table 7: Key variables for transferring information	30
Table 8: Qualitative comparison of the data sources OV-chipkaart and WROOV	31
Table 9: Generic model specification for zonal allocation in Biogeme	39
Table 10: Matrix divisions	41
Table 11: Public transport concessions in the Amsterdam region	46
Table 12: Shares of BTM trips transferring to or from the train network per train station	59
Table 13: Characteristics of different choice set generation amplifications	70
Table 14: Shares of trip-ends within catchment areas	70
Table 15: Available attributes for zonal allocation and their expected effect on utility	71
Table 16: Final estimation results of trip-based zonal allocation models	75
Table 17: Final estimation results of tour-based zonal allocation models	76
Table 18: Final estimation results of non-home-based trip zonal allocation models...	77
Table 19: Descriptive statistics of the attributes in the zonal allocation models	81
Table 20: Zonal allocation example 1.....	81
Table 21: Zonal allocation example 2.....	82
Table 22: Available attributes for purpose inference	84
Table 23: Model statistics for the three specific purpose inference models	86
Table 24: Final parameter estimates for specific purpose inference models	87
Table 25: Examples of travel purpose inference	91
Table 26: Multiplication factors of OD cells for equal trip totals	98
Table 27: Key variables and their representation in the purpose inference models .	111



Figures

Figure 1: Amsterdam in three different zonal resolutions	2
Figure 2: Research objectives, aims and goals.....	7
Figure 3: The primary dimensions of travel behaviour (Bagchi & White, 2005)	13
Figure 4: Research set-up	33
Figure 5: Research outline.....	35
Figure 6: Classification of zonal allocation models	38
Figure 7: Data handling process of generic model structure for zonal allocation	39
Figure 8: Data handling process of purpose inference models	40
Figure 9: VENOM study area	45
Figure 10: Share of tours and non-home-based trips in the WROOV data	48
Figure 11: Trip-based and Tour-based definitions of trip legs.....	52
Figure 12: Access and egress distance distributions by mode at the home-end	53
Figure 13: Access and egress distance distributions by mode at the activity-end	53
Figure 14: Longitudinal analysis of access and egress distances	54
Figure 15: Overall purpose shares in the WROOV data	55
Figure 16: Activity duration distributions per purpose	56
Figure 17: Departure time distributions per purpose	56
Figure 18: Distribution of contract types per purpose	57
Figure 19: Longitudinal analysis of the purpose shares in the WROOV data.....	58
Figure 20: Flow diagram of distinction between transfers and activities	60
Figure 21: Trips per week in the OV-chipkaart data	61
Figure 22: Trip shares per time of day over the year.....	62
Figure 23: Comparison of trip shares per mode.....	63
Figure 24: Comparison of trip shares over the activity duration.....	63
Figure 25: Comparison of trip shares over departure time.....	64
Figure 26: Comparison of trip shares per contract duration	64
Figure 27: Stop locations in the VENOM zonal grid of Amsterdam	68
Figure 28: Catchment areas of stops	70
Figure 29: Model enhancement strategy for zonal allocation	74
Figure 30: Origin allocation model parameter values for yearly WROOV datasets	78
Figure 31: Home allocation model parameter values for yearly WROOV datasets	79
Figure 32: Enhancement strategy of purpose inference models	85
Figure 33: Probability distributions per purpose for the trip-based model	92
Figure 34: Probability distributions per purpose for the tour-based model	93
Figure 35: r^2 statistics of the rule-based model validation	100
Figure 36: r^2 statistics of the trip-based model validation	101
Figure 37: r^2 statistics of the tour-based model validation.....	102
Figure 38: Model validation regression lines	102
Figure 39: r^2 statistics of approach comparison of OV-chipkaart total matrices	104
Figure 40: r^2 statistics of source comparison per purpose	105



Glossary of terms

Access leg	The trip leg between the origin and the boarding stop.
Activity	An act performed outside home, which requires travelling.
Activity-end	The trip-end at the activity side of the trip. In case of an away trip, equal to the destination, in case of a return trip, equal to the origin.
Alighting	Disembarking a public transport vehicle at a stop or station.
Anonymised data	Data that cannot be traced back to an individual.
Automated fare collection (AFC)	A digitalized ticketing system that releases the driver of a public transport vehicle from the task of collecting the fares of travellers.
Boarding	Entering a public transport vehicle at a stop or station
Catchment area	The area around a public transport stop or station from which travellers are drawn.
Check-in	The smart card transaction made when boarding a vehicle (in an open system) or entering a station (in a closed system)
Check-out	The smart card transaction made when alighting a vehicle (in an open system) or leaving a station (in a closed system)
Closed system	A closed public transport smart card system involves smart card readers at the entrances and exits of stations, generally integrated in gates (generally applicable to metro systems).
Concession	A set of public transport services granted to an operator by the regional transport authority.
Destination	The final location of a trip, where the next activity is performed.
Egress leg	The trip leg between the alighting stop and the destination.
Euclidean distance	The distance between two locations measured by a straight line between them.
Evening peak	The peak period of travel demand during the evening. In this research the evening peak is defined from 4 pm to 6 pm.
Fare	The pricing system used for travel by public transport.



Home-end	The trip-end at the home side of the trip. In case of an away trip, equal to the origin. In case of a return trip, equal to the destination.
Key variable	A variable available in both smart card data and survey data, which can be used to compare the sources and transfer information between them.
Land-use characteristics	Characteristics of the functions of land.
Mode	The type of transportations, relating to the infrastructure and vehicles used.
Morning peak	The peak period of travel demand during the morning. In this research the morning peak is defined from 7 am to 9 am.
Non-home-based trip	A trip that does not start at home. (In this study a trip that is not part of a tour is assumed to be non-home-based.)
OD matrix	Describes the number of trips from each Origin zone to each Destination zone, used to describe the travel demand.
Open system	An open public transport smart card system involves smart card readers in the vehicles and does not require gates (generally applicable to bus and tram).
Origin	The start location of a trip.
Smart card	A plastic card with a chip that stores data. In case of public transport smart cards, the chip contains information on the loaded contracts or stored value on the card.
Smart card reader	A device that registers transactions made with the smart card. The reader can read and write data from and to the smart card. In case of public transport smart cards, the reader registers check-in and check-out transactions, as well as loading stored value or contracts.
Time of day (TOD)	A specific period during the day.
Tour	The travel of an individual from the home location, conceivably via activity locations, back to the home location. (In this study, we assume all tours to be home-based.)
Transfer	Change of mode or vehicle during a trip, linking two consecutive trip legs. A transfer can be made between different services at a single stop or consist of a walking leg between different stops.
Travel	The movement of people between different locations.



Travel purpose	The reason for travelling, related to the activity that is performed.
Trip	The travel of an individual from one activity location to another, which may consist of one or more trip legs
Trip leg	Part of a trip covered with a single mode or vehicle, without discontinuations or transfers.
Trip-ends	The two locations between which a trip is made. From a trip perspective the trip-ends are defined chronologically, as the origin and the destination. From a tour perspective trip-ends are defined geographically, as the home-end and the activity-end.
Working days	Monday to Friday
Zone (traffic analysis zone)	A specific area defined by the transport planner, often derived from postal codes. Zones can be defined in different levels of spatial resolution, depending on the

Abbreviations used in this report

AFC	Automated Fare Collection
BTM	Bus, Tram and Metro
LMS	Dutch National Transport Model (Landelijke Model Systeem)
MON/OViN	Dutch Mobility Study (Mobiliteitsonderzoek Nederland/ Onderzoek Verplaatsingen in Nederland)
NRM	Dutch Regional Transport Model (Nederlands Regionaal Model)
NVB	National ticketing system (Nationale Vervoerbewijzen)
OD	Origin-Destination
PC3	3-digit postal code
PC4	4-digit postal code
SRA	City region of Amsterdam (Stadsregio Amsterdam)
TOD	Time-Of-Day
VENOM	Amsterdam Region Transport Model (Verkeerskundig Noordvleugel Model)
WROOV	The Netherlands Public Transport Allocation System (Werkgroep Reizigers Omvang en Omvang Verkopen)



1 Introduction

Since 2009, a new ticketing system was introduced in the Dutch public transport sector: the OV-chipkaart. This smart card system is currently employed as an Automatic Fare Collection (AFC) system by all public transport operators in the Netherlands. The fare collection works as follows. When boarding, travellers present their smart card to a reader, which generates a check-in transaction. Subsequently, the smart card is again presented to the reader at alighting, generating the check-out transaction. Fares are calculated based on the times and locations of these two transactions. Over the year 2014, the smart card system operator Translink registered 2.2 billion¹ transactions (Translink, 2015), which were all stored in the central back office.

The OV-chipkaart replaced the National Ticketing System (NVB), together with the complementing survey studies of The Netherlands Public Transport Allocation System (WROOV). The NVB contained the main ticket types in public transport, which were available throughout the Netherlands, and allowed travellers to travel with all public transport operators. The WROOV studies consisted of surveys that were used to allocate the revenues of the NVB to the operators and public transport authorities. The survey involved questions about the use of public transport based on ticket sales. Besides the information required for the fare box allocation, the survey also generated data on personal characteristics, the origins and destinations of travellers and the travel purpose. Consequently, the WROOV surveys were the primary data source of travel patterns regarding bus and light rail until 2009, the year the OV-chipkaart was introduced countywide.

This transition from conventional data collection towards new digital sources provides both opportunities, as well as challenges, for the use of public transport travel data. Digital data collection offers the possibility to handle large amounts of data at lower costs, resulting in high veracity on the number of travellers. However, since the data collection is passive, no additional information is collected on the characteristics and preferences of travellers. Combining the high veracity of smart card data with information derived from survey data might provide new opportunities for the application of public transport travel data.

1.1 Problem description

Public transport smart card data contain valuable information for operators, planners and authorities, and therefore they are applied in numerous fields. Some examples are transport performance monitoring and analysis, travel demand modelling, route choice modelling, advertising and even detection of contagious outbreaks. This study focusses on the utilization of smart card data for OD matrix construction, as input for travel demand models. Travel demand forecasts are essential information for authorities that take strategic decisions on investments in transport supply. Travel demand models are the primary tool to render these forecasts, therefore accuracy and credibility of these models are key in the decision making process (Ortúzar & Willumsen, 2011). These performance statistics depend on the quality of the input data. However, collecting this data by means of surveys and counts is expensive and

¹ Besides check-in and check-out transactions, the 2.2 billion transactions also include transactions recharging the stored value on smart cards.



sample sizes, and thus the quality of the results, is under pressure in these times of budget cuts.

The formulation of Origin-Destination (OD) matrices is a primary method for describing the travel demand and therefore it plays a prominent role in several modelling techniques (Ortúzar & Willumsen, 2011). These matrices contain the number of trips from every Origin zone to every Destination zone. In practice, the synthetic OD matrices generated by *trip generation* and *trip distribution* models often do not represent the current travel demand as it is observed in reality. This deviation can be associated with two issues:

- 1) The model does not take into account the daily variation of travel, but generates the travel of an average working day;
- 2) The model does not capture all factors that influence travel behaviour.

The first issue affects both aggregate and disaggregate modelling approaches and traditional data collection techniques do not provide the tools to compensate for it due to the limited sample sizes (Ortúzar & Willumsen, 2011). Regarding the second issue, correction is sought by means of matrix calibration. This process alters the matrix values in order to improve the fit of the matrix to calibration data, traditionally collected by means of counts or OD surveys. However, the influence of the calibration process should not be too large. The input data may not represent the average working day and large alterations can compromise the model consistency. Ortúzar & Willumsen (2011) state that the objective of the calibration process should not be to replicate the observations, but to estimate a matrix that captures their main features.

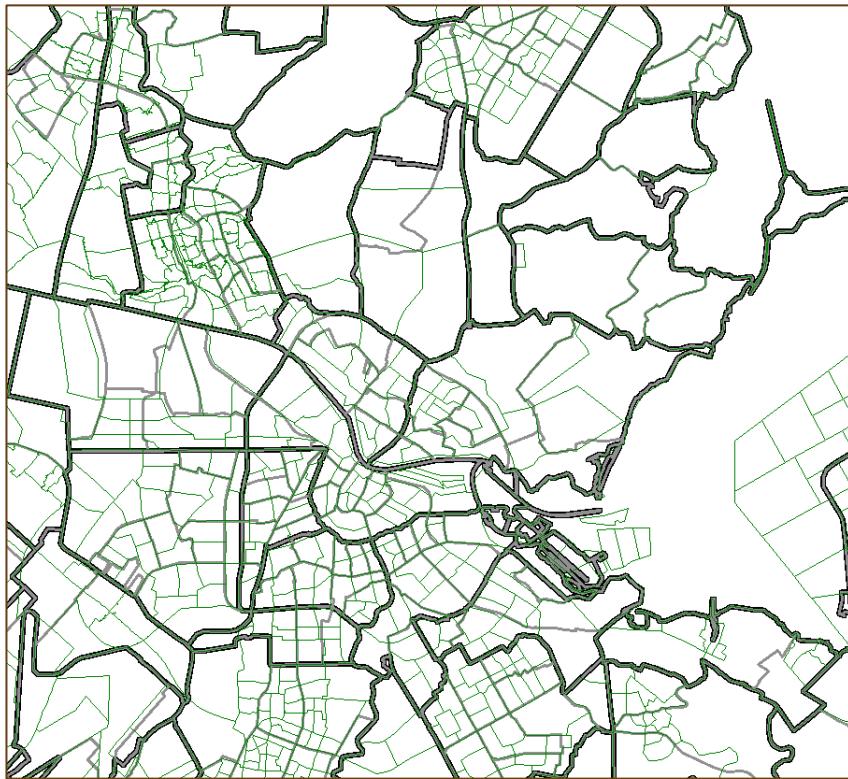


Figure 1: Amsterdam in three different zonal resolutions

Public transport smart cards provide continuous, disaggregate data that is very detailed in both time and space. Boarding and alighting times and locations are



recorded in detail. As they are often used to determine the price of the travel, they are the primary goal of collecting the data. However, the information that these features provide is still limited with regard to OD matrix construction. In order to specifically implement the effects of policy measures and socio-economic forecasts in the model, OD matrices are split into different categories that relate to the travel behaviour of travellers. These characteristics relate to the choices a traveller makes, and are affected by policy measures and changes in socio-economic data (Rijkswaterstaat, 2012). Therefore, OD matrices are categorized by:

- Travel purpose (why travel?);
- Mode (how to travel?);
- Time of Day (TOD) (At what time to travel?).

Considering public transport smart card data, the travel purpose is not observed. The modes are limited to public transport modes and can be derived from the data. The total number of trips made and the times of travelling are also directly available from the data, as well as the locations of the used stops. Hence, the specification of stop-based matrices by mode and by time of day is uncomplicated, but the distinction by travel purpose

The locations of the used stops are, however, not the equal to the origin and destination. The origin and destination of a trip are the locations where activities have been performed, for example working, but also staying at home. In order to arrive at the boarding stop, a traveller has to make an access trip leg. On the other end of the trip an egress trip leg leads from the alighting stop to the destination. Access and egress trip legs are usually short and mostly covered on foot or by bike (Goeverden). Therefore, the stop location can be a close approximation of the origin or destination. Yet, the use of stop locations can lead to wrongly allocated zones since access and egress trip legs can cross zonal borders, especially with a high resolution zonal grid.

Furthermore, the travel purpose is not available in smart card data. The data collection is passive, and therefore traveller characteristics are unavailable. For privacy reasons, is it prohibited by law to couple travel data to personal data². Consequently, for the construction of OD matrices, trips observed by smart card data lack three essential pieces of information, the information we aim to add:

1. The origin of the trip;
2. The destination of the trips;
3. The travel purpose.

Where smart card data provide large numbers of observations, survey data can provide more insight in the background of the traveller. Surveys can include questions on the travels made, providing information on the origins and destinations, used modes and TOD, but also questions regarding the traveller, which provide information on the gender, age and the travel purpose. Hence, survey data contains all the required information for the construction of OD matrices.

However, to construct OD matrices based on surveys, the sample size should be sufficiently large to capture the main features of the travel patterns in the system. For high resolution zonal grids, this requires large sample sizes, which increase the expenses substantially.

² ² Wet Bescherming Persoonsgegevens (WBP, BWBR0011468)



Where smart card data can provide high veracity on the number of trips, used modes and TOD, survey data can provide information on the access and egress patterns and travel purposes. Together these sources contain the information to construct OD matrices with high veracity, at low costs. Therefore, this study focusses on coupling of these two data sources.

1.2 Research motivation

We have distinguished one main objective of this study and two supporting objectives. This paragraph describes the motivation for these objectives, with the matching aims and specific goals to achieve them.

1.2.1 Objectives, aims and goals

The main objective of this study is to improve the public transport travel demand forecasts, and thereby enhance the decision making process of authorities concerning investments in public transport. We aimed to achieve this by starting at the front side of transport models and improve the quality of the input data that describe the current travel demand by means of OD matrices.

Other options to enhance travel demand forecasts with smart card data are to calibrate transport models with the large amount of observations or to create new modelling techniques based on this new source. We have focussed on the description of the current travel demand as this can be applied in current methodologies and it provides the essential insight of the competences of smart card data. Thereby, this study provides a foundation for research in the application for model calibration and development of new modelling techniques, which can be encouraging follow-up studies.

The implementation of new data sources is not only a way to improve travel demand forecasts, but it is also an objective in itself. Globally, more and more developments are information driven. This trend has also penetrated into the transport market. Travellers want to know their route options and travel times before departure, so they can choose their route and departure time deliberately, based on accurate information. Transport modelling needs to keep up with these developments, and the information producing travellers provide several options to do so.

While modelling techniques are usually very much settled over time in order to preserve consistency, new data sources need to be adopted to comply with current developments. In addition to improving the modelling accuracy, this is also important for the replacement of traditional data sources with declining efficacy. New data sources are, however, not direct replacements of traditional data sources. This study aimed to increase insight in the differences between smart card data and survey data and the requirements for application in OD matrix construction.

The implementation of new data sources is in turn supported by the objective to increase insight in public transport travel patterns. By relating the travel purpose and access and egress trip legs to other travel characteristics we have gained insight in the behaviour related to these lacking pieces of information in smart card data.

1.2.2 Relevant actors

Different types of actors can benefit from this research. Primarily, users of strategic transport models benefit from an increased quality of the OD matrices. These users



consist of authorities responsible for procurement of PT concessions and municipalities. Improved modelling of travel with public transport will help them to make better choices regarding investments in transport supply and also increase the acceptance of model results by involved parties. In addition, PT operators will benefit from this study with increased perception of the travel behaviour in their system as well as an improved long-term forecast of the travel demand.

Furthermore, transport consultancies and research institutes can benefit from the newly developed modelling techniques and acquired knowledge on the combination of travel patterns in public transport. The conclusions and recommendations on combining sources might be relevant for future research, not only on the use of smart card data, but also on the implementation of other Big Data sources in transport modelling.

1.3 Research questions and approach

Directing the research motivation to the problem at hand, one main research question has been formulated. In order to structure the research process, the main question is divided into four sub questions, targeting the distinct aspects of this study.

1.3.1 Research questions

To achieve the main objective, this study has answered the following research questions:

"To what extent can the travel purpose, origins and destinations of public transport trips derived from smart card data be inferred based on information from survey data, in order to construct purpose-specific OD matrices suitable as base matrices in transport models?"

Survey data contain all the required information for the construction of OD matrices per purpose, mode and TOD. Hence, we are able to estimate the relations between the information that is lacking in smart card data and characteristics that are available, based on survey data. These relations are represented by enrichment models, which can project the relations found in survey data onto smart card data. Consequently, we combine the high veracity of smart card data on the number of trips, with the lacking information from survey data.

Only attributes that are available in both sources, the key variables, can be used to transfer information from survey data to smart card data. Hence, the enrichment models can only apply key variables as independent variables to estimate the lacking information.

First, we aim to identify attributes that are correlated to the information we want to add to the OV-chipkaart data by means of an analysis onto WROOV data. Second, we aim to assess the appropriateness of these attributes as medium of transferring information by comparing the data sources both qualitatively and quantitatively. Based on this assessment, the datasets are adapted in order to match their coverage of the system. In addition, attributes are selected for implementation in the enrichment models. Third, we aim to optimize enrichment models that can be applied for the construction of purpose-specific OD matrices, based on the appropriate attributes with potential explanatory value of the lacking information. Fourth, and finally, we aim to evaluate the constructed OD matrices by comparing matrices



constructed by different modelling approaches in order to assess their quality and value.

In order to represent these different subjects within the research, the main research question is divided into the following sub-questions:

1. Which travel characteristics are correlated to the information to be added to OV-chipkaart data and to what extent?
 - 1.1 Which travel characteristics, and to what extent, are correlated to the access and egress trip legs to and from public transport stops?
 - 1.2 Which travel characteristics, and to what extent, are correlated to the travel purpose?
2. How do the data sources OV-chipkaart and WROOV compare to each other?
 - 2.1 How do the data sources OV-chipkaart and WROOV compare to each other qualitatively, concerning data collection method, available information, target populations and coverage of the transport system?
 - 2.2 Which attributes are key variables and how can these be represented in order to construct comparative data sets from OV-chipkaart and WROOV?
 - 2.3 How do the data sources compare quantitatively, regarding the travel they describe?
3. To what extent can information, lacking in the OV-chipkaart data, be inferred based on WROOV data, in order to construct purpose specific OD matrices?
 - 3.1 To what extent can origins and destinations of trips derived from smart card data be inferred in order to convert boarding-alighting matrices, based on stops, to OD matrices, based on zones?
 - 3.2 To what extent can travel purposes of trips derived from smart card data be inferred in order to distinguish base matrices by purpose?
4. How do base matrices created by different methods compare to each other?
 - 4.1 To what extent do the trip-based and tour-based approaches result in different matrices for the total average working day, matrices per time of day and matrices per purpose?
 - 4.2 At what level of resolution do these differences appear?
 - 4.3 How can continuity of this method of OD matrix construction be attained regarding the required data sources?

1.3.2 Approach

The foundation of this study was laid during the literature study on the applications of smart card data for travel demand forecasts. The literature study provided possible methodologies to couple the data sources as well as indications of which attributes could have explanatory value.

In order to describe the travel purpose and the access and egress distances based on other travel characteristics, an analysis of the survey data was performed. By means of this analysis, and with input from the literature study, we answered sub-question 1 and learned which characteristics could possibly have an explanatory value in the coupling of information to smart card data.

Subsequently, the data sources were compared in order to derive the transferability of information. In order to couple the data sources, key variables were identified. These variables are available in both data sources and information can be added to smart card data via these attributes. In order to determine the appropriateness of these variables as medium for transferring information, their patterns in both sources were



compared. Combined with a qualitative comparison of the data sources, we answered sub-question 2. This leads to the selection of key variables to start the model estimations.

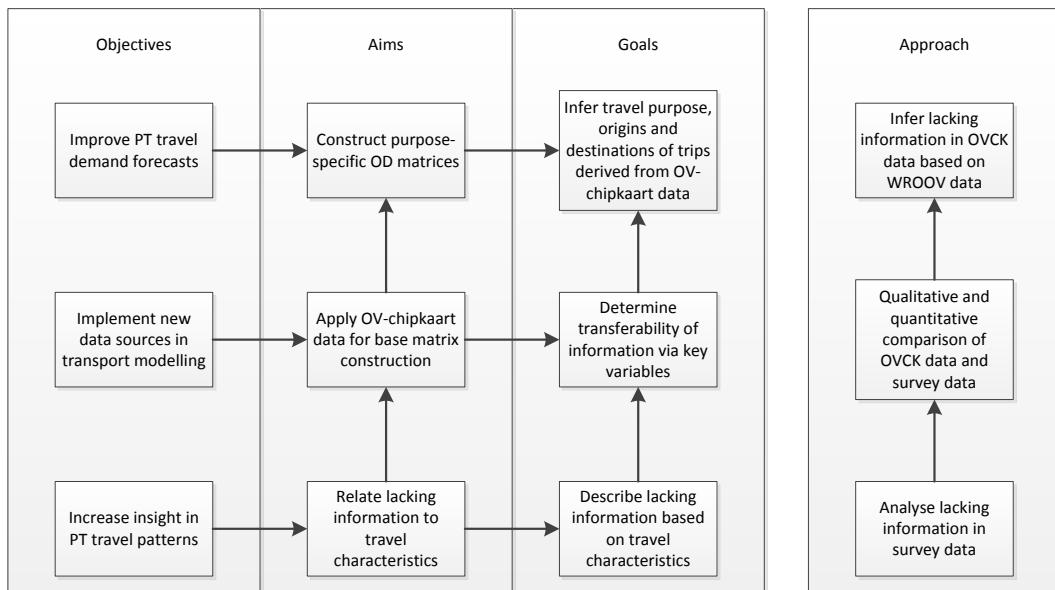


Figure 2: Research objectives, aims and goals

Using these key variables as attributes, we estimated zonal allocation models and purpose inference models. These have been applied to enrich smart card data with the information required to construct purpose specific OD matrices. During the model estimation phase, the explanatory variables were further selected based on the model performance. Three different approaches have been applied, with different levels of enhancement. These models were validated on WROOV data, to investigate their accuracy and answer sub-question 3.

Finally, we have evaluated the differences between the model approaches by comparing their resulting OD matrices. The evaluation consists of comparisons at different levels of spatial resolution, as well as time resolution and purpose specification. These comparisons have provided insight in the appropriate modelling approaches for specific OD matrix qualifications, and thereby answered sub-question 4.

1.4 Definitions used in this report

The perception of several definitions used in this report is essential to grasp the subtle distinctions that have played a significant role in this study. Therefore, we present the most essential terms here. A complete glossary of terms can be found at the front of the report.

Firstly, there are several ways to describe travels, which are closely linked and should not be confound. The principal unit of travel used in this study is a *trip*. A trip is made from one activity location, the origin, to the next activity location, the destination. OD matrices are based on trips, as these are directly related to the performance of the transport system. Zooming in, a trip can consist of several *trip legs*, which relate to a certain mode or vehicle that is used. Consecutive trip legs within a trip are connected by transfers.



Zooming out, research shows that consecutive trips made by a traveller, are often related. Therefore, an alternative approach in describing travel was introduced, based on *tours*. Tours describe the travel from one location up until return at that location. Tours can be classified as home-based, thus starting and ending at home, or non-home-based, starting and ending at, for example, the office. Several activities can be performed within one tour, increasing the number of trips within the tour. Hence, a tour with one trip relates to a tour without activities. A tour with two trips relates to a tour with one activity, et cetera. The distinction between transfers, that connect trip legs, and activities, that connect trips, is that a transfer only serves the purpose of getting onto the next vehicle (Devillaine, Munizaga, & Trépanier, 2012).

The trip-based and tour-based approach result in different taxonomies of the trip ends. The trip-based definitions of *origin* and *destination* are applied in the OD matrix. Conversely, applying the tour-based approach, these are classified as the *home-end* and *activity end*. By using this definition, trip ends are categorized by location instead of departure or arrival. Compared to the trip-based approach, this allows for the anticipation of specific characteristics at certain locations and ensures consistency between trips. This distinction has proven to be rather important in the allocation of trip ends to zones.

1.5 Report structure

The remainder of this thesis report is structured as follows.

- Chapter 2 covers the results from the literature review. Here, relevant references from literature are discussed regarding the use of PT smart card data in travel demand modelling. In addition, this chapter lists methodologies applied in other studies, used to handle similar questions as the ones covered in this study;
- Chapter 3 presents the methodology used for answering the research questions and the data sources employed;
- Chapter 4 contains a description of the case study used for answering the research questions. This chapter covers the availability of data for this study and the complementary boundaries, as well as the relation between these boundaries and the generalizability of this study;
- Chapter 5 comprises the results from the WROOV data analysis, regarding travel characteristics related to access and egress distances and travel characteristics related to the travel purpose. Furthermore, the data sources WROOV and OV-chipkaart are compared, both qualitatively and quantitatively;
- Chapter 6 covers the process and the final results of the estimation of zonal allocation models and purpose inference models. Here, the interpretation, stability and generalizability of the final models are discussed;
- Chapter 7 encompasses the evaluation: the comparison of the OD matrices constructed by different modelling approaches;
- Chapter 8 lists the conclusions from this study and discusses the implications of the findings. The conclusions are divided in the answers of the research questions and additional conclusions;
- Chapter 9, finally, contains the recommendations to relevant actors based on the conclusions. The recommendations are categorized by the utilization of the results and recommendations regarding further research.



Table 1: Locations of sub-questions in the report

<i>Sub-question</i>	<i>Paragraph</i>	<i>page</i>
1. Which travel characteristics are correlated to the information to be added to OV-chipkaart data and to what extent?	5.1 / 5.2	51 / 54
2. How do the data sources OV-chipkaart and WROOV compare to each other?	3.4 / 5.4	30 / 60
3. To what extent can information, lacking in the OV-chipkaart data, be inferred based on WROOV data, in order to construct purpose specific OD matrices?	6.2 / 6.3	68 / 82
4. How do base matrices created by different methods compare to each other?	7.2	99



2 Literature review

This chapter contains the results of the literature study that has been performed throughout the research. The literature study functions as the foundation of the research and has four objectives:

- *Providing background information*: ensuring comprehension of the possibilities and complexities of smart card data and their utilization;
- *Listing reference studies*: placing this research in context of other studies, extracting relevant information and methodologies and identifying gaps in the existing literature;
- *Discovering methodology options*: determining feasible options to answer the research questions;
- *Supplying input for the assessment framework*: formulating a comprehensive and consistent framework of assessment criteria for the methodology choice.

The chapter is structured as follows. First, the smart card data are introduced (section 2.1). This paragraph describes the implementation of the smart card in PT, the data structure and the potential of smart card data for travel demand modelling. Second, methodologies to construct OD matrices from smart card data are discussed (section 2.2). This leads to an indication of possible solutions to answer research question 3.1. Third, techniques for travel purpose inference are discussed (section 2.3). This leads to an indication of possible solutions to answer research question 3.2. The chapter culminates with conclusions on the discovered literature and implications for this study (section 2.4), supplying an overview of possible methodologies and assessment criteria.

2.1 Introduction to smart card data

Smart cards have become a popular means of fare collection in PT systems. Renowned examples are the Octopus Card in Hong Kong and the Oyster card in London, but there are many more smart card systems emerging around the world. Without going into detail on the smart card technology, this paragraph describes the data that is collected through smart card systems used in PT.

2.1.1 *Introduction of the smartcard in public transport*

Early contributors to the research on the use of public transport smart card data did not directly focus on the possibilities for transport demand modelling. Blythe was primarily interested in the interoperability of smart cards between operators across the UK and additional advantages for the operability of public transport (Blythe & Holland, Integrated ticketing - Smart cards in transport, 1998). These advantages consist of less labour-intensive revenue collection, reduced boarding times and increased security (Blythe, Improving public transport ticketing through smart cards, 2004). Dinant & Keuleers (2004) investigated the use of smart card data from a data protection perspective. They identify the privacy concerns that emerge with cross-profiling of databases and present several cryptographic solutions to prevent the possibility of cross-profiling. Hence, enriching smart card data by cross-profiling is not considered as a desirable solution to this research into travel demand modelling. Moreover, the Dutch law on protection of personal information³ prohibits the use of personal data for purposes other than the ones specified in advance.

³ Wet Bescherming Persoonsgegevens (WBP, BWBR0011468)



2.1.2 Structure of smart card travel data

The characteristics of smart card systems in PT can differ in several ways. The universal characteristic is the use for automated fare collection (AFC). However, the recording of relevant data for travel demand studies with these AFC systems depends on several specifications. There are six main differences in the description of smart card data structure observed in literature:

1. The smart card system can cover one or several modes of PT, where differences in the structure of collected data might occur (Seaborn, Attanucci, & Wilson, 2009). A higher share of smart card deployment in the total PT system, results in a more integral representation of the travel behaviour (Cui, Wilson, & Attanucci, 2006).
1. The availability of other ticket types besides the smart card will result in an incomplete, and possibly biased, representation of travel behaviour. Moreover, smart cards are not necessarily equal to unique travellers, since travellers can have more than one card or share their card (Morency, Trepanier, & Agard, 2007) (Robinson, Narayanan, Toh, & Pereira, 2014);
2. Different fare policies result in different transaction requirements. Flat fares only require one transaction (check-in) that reduces the stored value on the card with the fare for one ride. Alternatively, distance based fares require two transactions: when boarding and when alighting (check-in and check-out), to determine the distance travelled. Distance based fares therefore collect more relevant information for travel demand studies, including the alighting location, distance travelled and travel time;
3. The availability of stop or station information where the transaction took place depends on the placement of the smart card reader. Smart card readers can be placed in vehicles or at stations, depending on the transport system. Readers placed at stations are usual for train and metro systems and are generally placed at the station entrances and exits. Since these readers are stationary, they can be automatically coupled to location information. Readers in bus and tram systems are usually placed in the vehicle, thus not stationary. Hence, the availability of stop information depends on the integration with an Automated Vehicle Location (AVL) system, which records the GPS coordinates of the vehicle at the time of the transaction. For distance based fares, the availability of an AVL system is required, but it is also applicable for flat rate fares. In most systems, the AFC and the AVL are one integrated system, assigning every check-in and check-out transaction to a stop or station. If AFC and AVL are separated, stops are not directly recorded in the data, but can still be derived by additional data processing (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013) (Liao & Liu, 2010) (Zhao, Rahbee, & Wilson, 2007);
4. The identification of transfers can be automated in the data collection, indirectly traceable or completely absent in the data. This depends on the fare policy for trips with several rides and the smart card software. The automated identification of transfers might not always match the definition of a transfer used in this study (see paragraph 2.2.2);
5. The quality of the data is influenced by both system errors and user errors. System errors can include software bugs, hacks, erroneous input data of the network or fare structure, broken hardware and communication malfunctions. User errors can include failure to check-in or check-out and untimely check-outs. Both cases can be either deliberate or not (Robinson, Narayanan, Toh, & Pereira, 2014) (Chu & Chapleau, Imputation techniques for missing fields and implausible values in public transit smart card data, 2007). Despite these possible errors, the overall quality of smart card is high, certainly compared to survey data. In addition, correction for



some of these errors is possible by means of rule-based inference. (Chu & Chapleau, Imputation techniques for missing fields and implausible values in public transit smart card data, 2007)

Smart cards are often personal cards, so technically the smart card data could also contain personal data. Although collected by purchase transactions, personal data are usually not combined with smart card data due to privacy concerns. One study, however, did have access to billing addresses. Utsunomiya et al. (2006) used the data from the Chicago Card to study the access and egress distances to stops.

2.1.3 Potential for travel demand modelling

Bagchi & White (2005) pioneered in the potential of smart card data for travel demand modelling. The authors provide a insightful overview of possibilities and constraints of smart card data and compare the smart card to traditional data sources. They define a conceptual framework for travel behaviour analysis, which they apply to the problem of transport turnover, also referred to as churn. This framework consists of three generic dimensions in which travel behaviour takes place: time, space and structure (see Figure 3). In order to analyse travel behaviour, boundaries for these dimensions need to be determined. Spatial boundaries can be defined by administrative areas, like postal codes, or by service areas of public transport operators. Regarding time boundaries, the authors suggest a period of one year, to balance effects of seasonal variation in behaviour. The structure can be constrained by transport operators, modes or routes.

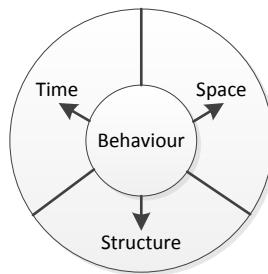


Figure 3: The primary dimensions of travel behaviour (Bagchi & White, 2005)

Analysing the potential of smart card data, the authors encounter essential issues that complicate the use for travel demand modelling. Smart card data do not indicate the exact origins and destinations on the level of street addresses, nor the travel purpose. Furthermore, the authors also acknowledge the problems concerning definitions. Smart cards record passengers boarding and in some systems also alighting, which are not equal to passenger-trips. The UK National Travel Survey defines trips as one way journeys from one activity to another, which may consist of trip legs. Trips can be constructed by data processing, for example by assuming a transfer in case of an alighting followed by a boarding within a certain time interval. This accentuates the importance of concept definitions and rule-based processing, as the interval is often chosen arbitrarily. Another issue is the generalizability of the group of smart card users, which might not be typical for the whole population when several ticket types are in operation.

On the other hand, smart card data also have substantial advantages over survey data. Most surveys are momentary representations of behaviour, while counts, of both



ticket sales and passengers, cannot be traced back to individuals. The personal and continuous characteristics of smart card data are considered to be a valuable enhancement in the examination of travel behaviour, even though smart cards are not necessarily equivalent to individuals. Bagchi and White conclude that smart card data have a large potential due to the large sample size and long periods of coverage, but smart card data alone are not sufficient for travel demand modelling. They suggest a routine survey to verify and complement the smart card data with information on travel purpose, origins and destinations.

Building on the fundamental analysis by Bagchi & White (2005), Pelletier et al. (2011) propose deliberate research topics to employ the potential of smart card data for travel demand modelling. The authors constructed a comprehensive literature review of the use of smart card data for transport planners up to 2011. They differentiated a wide range of studies into operational, tactical and strategic planning. They conclude that smart card data can be useful for both researchers and transport planners, and list challenges to overcome in order to use the full potential of smart card data in transport planning. The authors state that, from a researchers perspective, this new data source requires new modelling approaches that are fit for such a detailed level of resolution. A primary issue is linking socio-demographic data to anonymous smart card data, while complying with privacy regulations. Furthermore, the continuous data will encourage new methods of data analysis for longitudinal studies. The authors end with the notion that the smart cards will be commonly used in public transport, ensuing research on travel behaviour. And indeed, much research has been done since 2011.

Concluding the research from Bagchi & White and Pelletier et al., there are two main issues to solve in order to employ the full potential of smart card data in travel demand modelling:

1. Distinguishing the actual origins and destinations of the trips observed in smart card data;
2. Estimating the travel purpose;
3. Adapting modelling techniques in order to utilize the continuous character of smart card data.

The first two issues are the main topics of this study. Literature on these issues is discussed in the following paragraphs. The third issue is another subject, which could be dealt with in a follow-up study.

2.2 Smart card data for OD matrix construction

In order to take up the first issue of this study, this paragraph discusses the literature on the construction of OD matrices from smart card data. The formulation of OD matrices is a primary method for describing the travel demand (Ortúzar & Willumsen, 2011), and so, this has been a recurring topic in the research on smart card data. As described in the previous paragraph, smart card systems and their collected data can differ substantially.

Depending on the specifications of the smart card system, several procedures have to be performed in order to convert the smart card data into OD matrices. First, if the alighting stop is not recorded, it has to be inferred. Successively, if the used stops are known, the trip legs have to be converted into trips by identification of transfers. Transfers within the described system can be derived through data processing. On the



other hand, transfers to other systems that are not included in the available data cause another problem. In the final step, the stop-based matrices that are derived from smart card data can be converted into OD matrices by inferring the actual origins and destinations of trips.

2.2.1 Alighting stop estimation

In smart card systems with flat rate fares, the alighting stop is not recorded in the data. Consequently, the first step in constructing a stop-based OD matrix is the inference of the alighting stop. Several studies describe the procedure of estimating the alighting stop based on rule-based processing. The fundamental processing rules are (Cui, Wilson, & Attanucci, 2006) (Barry, Freimer, & Slavin, 2009) (Wang, Attanucci, & Wilson, 2011):

1. The alighting stop of a trip leg is the nearest stop to the boarding stop of the next trip leg that day;
2. The last alighting stop of the day equals the first boarding stop of that day.

These rules are based on the assumptions that 1) travellers will return to their alighting stop after their activity, since the walking distances are small, and 2) travellers start and end their travels at home.

In addition to these elementary rules, several refining rules have been applied. The search procedure for the first rule is delimited to stops on the used line, in the right direction. Constraints can be set on the maximum distance and the travel time between alighting stop and boarding stop (Zhao, Rahbee, & Wilson, 2007) (Munizaga & Palma, 2012). If the second rule does not result in a valid alighting stop, the search procedure can be extended to the next day, as tours may cross the day border (Munizaga, Devillaine, Navarrete, & Silva, 2014). Another option to extend the search for an alighting stop is the examination of regularity in trips and estimate the alighting stop based on travel patterns (Trépanier, Tranchant, & Chapleau, 2007) (Jun & Dongyuan, 2013).

Where Barry et al. (2009) find 90% valid alighting stops for the New York metro system, Trépanier et al. (2007) report a success rate of 66%. However, the alighting stops might be valid, but still wrongly estimated since there is no direct validation possible. The need for validation of smartcard data is therefore generally acknowledged and has been studied using survey data. The results of various studies differ considerably: Munizaga et al. (2014) find a correct estimation of the alighting stop in 84% of the cases, where Wang et al. (2011) report only 60% of correctly estimated alighting stops.

2.2.2 Identification of transfers within the described structure

The travel demand we want to describe consists of the number of trips between activities, since activities form the reasons to travel to their corresponding location. Therefore, OD matrices are trip based, where one trip can consist of several trip legs (see paragraph 1.4 for the definitions). Transfers are not activities in this regard, as they only serve the purpose of reaching the destination (Devillaine, Munizaga, & Trépanier, 2012). The smart card does not record trips, but trip legs.⁴ In order to

⁴ In closed systems, with stationary smart card readers at the entrances and exits of stations, transfers do not require additional check-in and check-out transactions. Therefore, consecutive trip legs within a closed system are considered as one trip leg in the data. Reddy et al. (2009) derive the number of actual trip legs in a closed system based on a survey.



derive trips, transfers need to be identified between consecutive trip legs that are part of the same trip.

While some smart card systems automatically record transfers in the data, transfers are not actually observed. What a traveller does between alighting and a consecutive boarding is unknown. It could be the traveller is walking to another stop or waiting for a connecting ride, but it is also possible that a short activity takes place. The identification involves rule-based processing, usually by means of a time constraint. Therefore, automated identification of transfers is not necessarily the right identification (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Robinson, Narayanan, Toh, & Pereira, 2014). The specified time interval that an operator uses for fare calculations, in case of reduced fares for transfer trips, might even be a cost-based incentive for short activities (Jang, 2010).

For systems without automated recording of transfers, similar rule-based processing methods are applied to derive transfers and combine trip legs into trips. From Bagchi and White (2005), we know constraints can be set on time, space and structure. Constraints on these dimensions are frequently identified as rules to infer transfers within the transport structure described by the smart card data. The proposed constraints, however, differ substantially over the studies found in literature. A range of time constraints is used:

- 30 minutes (Bagchi & White, 2005) (Munizaga & Palma, 2012) (Devillaine, Munizaga, & Trépanier, 2012);
- 35 minutes (Nijenstein & Bussink, 2014);
- 60 minutes (Chakirov & Erath, 2012);
- 90 minutes (Hofmann & O'Mahony, 2005) (Hofmann, Wilson, & White, 2009);
- 120 minutes (Utsunomiya, Attanucci, & Wilson, 2006).

If an operator uses a flat fare structure, the alighting time is not available and the on-board time should be included in the time constraint. This results in different time constraints for transfers between different modes (Seaborn, Attanucci, & Wilson, 2009). Transfer times that lie in between the aforementioned time constraints can be ambiguous. Variance between average transfer times in different networks is observed: in Seoul 80% of the transfers take less than 10 minutes (Jang, 2010), while in Gatineau this share is reached at 18 minutes (Chu & Chapleau, Enriching Archived Smart Card Transaction Data for Transit Demand Modelling, 2008). The network density and level of service influence the transfer times (Jang, 2010) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013). In addition, transfer behaviour might also differ between user groups (Bagchi & White, 2005). Therefore, refinement of the transfer identification process has been pursued in the other two dimensions of travel behaviour.

Two types of spatial constraints are proposed:

1. *Distance covered during transfer*: the distance between the alighting stop and the next boarding stop should be low as it is assumed to be covered on foot. The distance can also be converted into walking time using a mean walking speed. The use of a buffer is recommended as walking distances depend on the infrastructure layout (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013);
2. *Distance between trip ends*: in order to account for return trips after short activities, constraints can be set on the directness of the trip. This can be



done by comparing the Euclidian distance between the first boarding stop and the last alighting stop to the total distance travelled during the trip (Munizaga, Devillaine, Navarrete, & Silva, 2014) (Robinson, Narayanan, Toh, & Pereira, 2014). A similar method is to use the circuitry ratio of the path travelled within a trip. If a transfer location lies outside a specified circuitry area around the trip ends, it is assumed that the trip is not direct and an activity has been performed (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013). In case three or more consecutive trip legs are observed that describe a tour, it is not trivial at which transfer locations an activity took place, in other words: where to cut the tour in outbound and inbound trips. The activity location can be selected using the highest likelihood based on directness of the trips (Munizaga, Devillaine, Navarrete, & Silva, 2014).

In addition, two kinds of structural constraints are proposed:

1. *No transfers on the same route*: transfers on the same route identify an activity at the transfer location. Alighting and then boarding a vehicle on the same route results in deliberate time loss, which only provides utility to the traveller in case of an activity. This holds for vehicles in both directions of the route (Munizaga & Palma, 2012) (Hofmann & O'Mahony, 2005) (Hofmann, Wilson, & White, 2009) (Chu & Chapleau, Enriching Archived Smart Card Transaction Data for Transit Demand Modelling, 2008);
2. *First opportunity*: using bus route schedules it can be checked if the transfer was actually the first opportunity for the traveller. Again using the utility maximization principle, letting a opportunity for a faster transfer pass only makes sense in case of an activity (Chu & Chapleau, Enriching Archived Smart Card Transaction Data for Transit Demand Modelling, 2008) (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013).

Different combinations of processing rules can be used in order to identify transfers. The transfer identification procedures proposed by Chu & Chapleau (2008) Nassir et al. (2011) and Gordon et al. (2013) combine constraints in all three dimensions. They define a walking time to estimate the arrival time at the boarding stop. Using the scheduled bus times, they estimate the number of boarding opportunities between arriving at the stop and boarding. If the first opportunity is taken, a transfer is inferred. Using these constraints, Chu & Chapleau (2008) found that the automated identification of transfers in Gatineau overestimates the number of transfers with nearly 40%.

2.2.3 Identification of transfers to other systems

Transfers within the described system only constitute part of the possible transfers that can be made by travellers. The described system is limited by the available data, usually the data from one operator or region. However, the entire transport system available to travellers consists of several modes, where the PT system is a section of the total system. In addition, the PT system consists of several operators that provide their services in different areas and possibly with different modes. Service areas may be adjacent or overlapping. These constraints on the transport structure, caused by the availability of data, are not desirable in the context of trip analysis. We aim for information on the total trip, including transfers to other modes and operators.

In case the available data encompasses the data from one operator, there are three categories of transfers to other systems, which have different consequences for the construction of OD matrices:



1. *Transfers to private modes*: private modes comprise car, which can be subdivided into driver and passenger, and bicycle. Walking is considered a general component of travel and therefore is not included as mode (Nes, 2002). Transfers to private modes cannot be obtained from smart card data. However, these transfers are rare. The options to transfer between public and private modes are limited. PT users are usually bound to walking (Munizaga, Devillaine, Navarrete, & Silva, 2014) (Nassir, Khani, Lee, Noh, & Hickman, 2011) or cycling, which is popular mode of transportation in the Netherlands (Goeverden);
2. *Transfers to adjacent PT operators*: as transactions with other PT operators are not available in the data, transfers between operators cannot be obtained and result in incomplete visibility of the trip in de data. Additional trips can be made with other operators on either side of the observed trip legs. Depending on which side of the observed trip leg the transfer occurs, the origin or the destination lies outside the study area. This causes the trip to be entered in an incorrect OD cell, where the origin or destination should be an external cell. The optimal solution to the problem of transfers to other operators is combining the data from different operators, which has been done in the Netherlands by Nijenstein & Bussink (2014);
3. *Transfers to PT operators within overlapping service area*: a distinction of transfers to other operators can be made if the service areas of the operators overlap. If the traveller uses another operator to travel within the same service area, the trip is entered in an incorrect OD cell, where the origin or destination should, in this case, be an internal cell. When a different operator is used in between trip legs that are described in the available data, they can be identified by looking at the spatial-temporal distribution of consecutive trip legs (Chakirov & Erath, 2012). Depending on the used rules for transfer identification, the observed data would otherwise be interpreted as two separate trips or as one trip with a large transfer distance.

2.2.4 Conversion from stops to origins and destinations

Smart card data provide travel information on the level of used stops and stations during travels. The previous paragraphs describe the construction of matrices from these data on the level of stops. Several studies (Bagchi & White, 2005) (Pelletier, Trépanier, & Morency, 2011) acknowledged that the used stops are not equal to origins and destinations. Therefore, these matrices are not actual OD matrices, but stop-based matrices. Stop-based matrices may accurately describe the current travel demand, since the used stops are assumed to be in the vicinity of the activity locations. However, travel demand is derived demand from travellers' origins and destinations, as these are the locations where activities are performed. The actual origin and destination addresses are considered more stable indicators for future travel demand, since the used stops are, to a larger extent than origins and destinations, subject to the travel supply. The stability of indicators for travel demand is especially important in strategic level studies (Ortúzar & Willumsen, 2011).

Not all studies into the use of smart card data take this refinement into account, the stop-based matrix is then presented as OD matrix (Munizaga & Palma, 2012) (Nassir, Khani, Lee, Noh, & Hickman, 2011) (Nijenstein & Bussink, 2014). Studies that do consider the conversion from stop-based matrices to OD matrices can be divided into two categories:



1. *Clustering of stops:* several studies combine stops into stop clusters in order to reduce the number of cells in the matrix. At the lower scale this includes the aggregation of stops in opposite directions of the same line. On a higher scale the aggregation can also include the clustering of stops that serve the same catchment area (Lee, Hickman, & Tong, 2012) (Lee & Hickman, Are Transit Trips Symmetrical in Time and Space?, 2013). A second method used to cluster stops is based on movement patterns. This method includes the assessment of similarities of the used stops at both ends of the trip instead of only one trip end. Movement patterns can be clustered based on travel directions and adjacency of stops at either end of the trip (Kim, Oh, Lee, Kim, & Jung, 2014). A third method applied to cluster stops is the clustering by travel analysis zones, which equates to a direct conversion from the stops to the zone in which they are situated. This method is often used because of its applicability in transport modelling (Lianfu, Shuzhi, Yonggang, & Ziyin, 2007) (Farzin, 2008) (Zhou, Murphy, & Long, 2014) (Oort, Drost, & Brand, 2014);
2. *Allocation of origins and destinations:* trips can be allocated to specific origins and destinations. Several techniques have been applied for this allocation procedure, with different levels of resolution of the origins and destinations. Trips can be allocated to zones using a logit allocation, based on the walking distances to nearby zones and zonal characteristics (Barry, Freimer, & Slavin, 2009). In their study, Barry et al. use variable zonal characteristics, like population and employment, for different times of day, indicating the relation with the travel purpose. At a higher level of resolution, trips can also be allocated to addresses. Utsunomiya et al. (2006) use the billing address of cardholders in Chicago to assess the access distances to stops, with the assumption that the billing address corresponds to the home address. Chu & Chapleau (2010) allocate trips with student-cards to school buildings if the used stop is within 500 meter of the school address. Other trips are allocated to specific locations within the area around the stops, based on a probabilistic approach. The applied density function depends on the walking distance to the stop and the population distribution in the area. Ordóñez & Erath (2013) allocate commuting trips to work locations by minimizing the total walking distance from stops to work locations in Singapore. To do so, they use high resolution GIS data, to estimate work space capacities on parcel level and determine walking distances between stops and office buildings, and correct for the use of other transport modes used for commuting.

The clustering techniques can be perceived as OD matrices, since stops are clustered based on spatial characteristics. However, these techniques do not specifically take into account the access and egress trip legs.

Utsunomiya, et al. (2006) find that the access distance differs between rail and bus services, due to differences in the stop density and the service quality. Furthermore, access distances differ over individual stops.

2.3 Travel purpose inference

In order to take up the second issue of this study, this paragraph discusses the literature on inference of the travel purpose. As indicated by the literature in the previous paragraph, the travel purpose is closely related to the allocation of origins and destinations to trips observed in smart card data. In fact, both issues are determined by the activities performed at either end of the trip: the activity location depends on the activity type. However, the travel purpose inference is also a research



topic of its own. The purpose of travelling is a key subject of policy measures, making it an important element for grasping the influence of these measures by strategic transport models (Ortúzar & Willumsen, 2011).

The lack of information about travel purposes is a common issue for passive data collection of travel trajectories. Movements of individuals can be traced using GPS data, mobile phone data, smart card data or data from social media. All these sources provide spatial-temporal information of individuals, with different penetration rates and levels of resolution (Yue, Lan, Yeh, & Li, 2014), but lack information on travel purposes. The discovered literature on travel purpose inference focusses on two sources: smart card data and GPS data, which are discussed sequentially below.

2.3.1 Purpose inference from smart card data

Since the smart card data do not provide information on the travel purpose, but do offer detailed information on the use of PT, researchers have shifted their focus to the identification of user groups. User groups can be categorized by means of a K-means clustering method. The number of clusters can be either pre-defined (Morency, Trepanier, & Agard, 2007) or determined using a more elaborated version of clustering. Agard et al. (2006) (2009) use a Hierarchical Ascending Clustering (HAC) method to determine the number of clusters, while Ma et al. (2013) use a K-means++ clustering to find the optimal number of clusters.

The clusters are defined based on temporal variables or a combination of temporal and spatial variables. Three scales of temporal variables are defined: *times* during the day (TOD), *frequency* of travel during working days of one week and *regularity* of travel during a period of several weeks (Agard, Morency, & Trépanier, 2009) (Agard, Morency, & Trépanier, 2006). The used lines and stops can be applied as spatial clustering variables (Ma, Wu, Wang, Chen, & Liu, 2013). The user groups seem to have a high correlation with the age groups deduced from card types, which indicates the relevance for market analysis and segmentation (Agard, Morency, & Trépanier, 2009). Furthermore, these clustering methods can support short term predictions, for operators to enhance their service quality (Morency, Trepanier, & Agard, 2007) (Agard, Morency, & Trépanier, 2009) (Ma, Wu, Wang, Chen, & Liu, 2013).

Although it is argued that trip attributes from smart card data could possibly better characterize trips than the travel purpose can (Chu & Chapleau, 2010), it is also reasoned that clustering techniques cannot capture the complexity of travel patterns (Kim K. , 2014). In addition, user groups based on spatial and temporal clustering variables depend on the level of service and do not reflect a traveller's motivation for travelling, where a classification based on travel purpose does. Hence, these clusters are not the stable indicators of travel demand that are pursued in long-term planning.

Expanding the protocols of smart card data mining, the rule-based processing approach is also applied on the inference of travel purposes. Four attributes are generally considered to have explanatory value of the travel purpose:

- Activity duration
- Departure time
- Frequency
- Card type

The used rules in literature consist of constraints to these attributes:



- Activities longer than six hours are *work* activities (Chakirov & Erath, 2012);
- Activities longer than five hours with adult cards are *work* activities (Devillaine, Munizaga, & Trépanier, 2012) (Gatineau, Canada);
- Activities longer than two hours with adult cards are *work* activities (Devillaine, Munizaga, & Trépanier, 2012) (Santiago, Chile);
- First trips of the day that take place during the morning peak, and have a corresponding return trip in the evening peak, are allocated to the purpose *work* (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014);
- Trips made with student cards with alighting near schools or universities are *educational* trips (Chu & Chapleau, 2010) (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014);
- Trips that are the last of the day, and not the only one, are allocated to the purpose *home* (Devillaine, Munizaga, & Trépanier, 2012).

The differentiated travel purposes differ between several studies. Some studies only take into account the most prevalent travel purpose: *work* (Jun & Dongyuan, 2013) (Zhou, Murphy, & Long, 2014). If all trips are to be incorporated, the basic purposes *work*, *home* and *other* are distinguished (Chakirov & Erath, 2012), and in some papers also the purpose *education* is considered (Devillaine, Munizaga, & Trépanier, 2012) (Chu & Chapleau, 2010). Using survey data, it is also possible to consider a wider range of purposes, like shopping and business (Kusakabe & Asakura, 2014) (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014). The availability of highly detailed GIS data initiates the focus on one specific travel purpose. The purpose *education* can be allocated to trips made with student cards when alighting takes place in the vicinity of a school address. Return trips can then be allocated to the *home* purpose (Chu & Chapleau, 2010). The purpose *work* can be allocated to trips based on parcel data of office buildings (Ordóñez & Erath, 2013).

Besides the use of rule-based processing, several other approaches have been applied, all with the aim of coupling information from survey data to smart card data. A rather direct coupling is the use of a Naïve Bayes classifier. This method assumes the same distribution of purposes relative to key variables, such as arrival time and activity duration, which are available in both survey data and smart card data (Kusakabe & Asakura, 2014). Another method is the use of a logit model for the allocation of purposes to trips. Such a model that determines the relative possibilities of a trip having a certain purpose can be estimated with survey data and subsequently applied to smart card data. Parameters with a significant influence on the distribution of chances are the activity duration, the start time and purpose-specific land-use information (Chakirov & Erath, 2012). Another method is the use of a decision tree algorithm with a learning module to classify trips into travel purpose bins (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014).

Comparing a rule-based approach with a logit allocation procedure, Chakirov & Erath (2012) find that the share of correctly estimated purposes differs only slightly in favour of the logit allocation. This is mainly because the simpler rule-based model has a surprisingly high fit, with almost 87% correctly estimated purposes. Furthermore, the addition of land-use information in the logit model only marginally increases the share of correctly inferred purposes compared to a logit model without land-use information. Train (2009) issues that the share of correctly estimated choices is not a decent indicator of the model fit, since it does not capture the underlying theory of probabilities.



2.3.2 Purpose inference from GPS data

GPS sensors are commonly employed in navigation tools, smartphones and special GPS tracking devices. The sensors continuously records location, speed and direction. For a more detailed specification of GPS devices and data structures, see Wolf et al. (2001) and Stopher et al. (2008). Many studies have been performed to investigate the use of GPS data as support or replacement of travel diary studies. The employment of GPS data generates similar issues as the use of smart card data. As GPS devices continuously record data, trip ends have to be derived in a comparable way as the identification of transfers with smart card data. In case of GPS devices in cars, the relation between exact activity location and the parking location is similar to the relation between activity locations and used PT stops (Axhausen, Schönfelder, Wolf, Oliveira, & Samaga, 2003).

A comprehensive literature review on the use of GPS data to identify person-trips is provided by Gong et al. (2014), indicating the inference of the travel purpose as one of the main research topics. The authors specify three categories of inference methods, which are also found in the literature on purpose inference of smart card data:

- *Rule-based processing*: high resolution land-use information can be combined with trip attributes such as activity duration and arrival time to infer the travel purpose (Wolf, Guensler, & Bachman, 2001) (Shen & Stopher, 2013). Purposes can also be allocated to purpose-specific points of interest (POI) in the vicinity of the trip end (Stopher, FitzGerald, & Zhang, 2008). In case of multiple POI within reach of the trip end, the closest POI can be allocated, or survey participants can be asked to provide the right purpose (Bohte & Maat, 2009);
- *Probabilistic approach*: Different POI within a specified distance of the trip end can be ranked on probabilities based on distance, the socio-demographic attributes sex, car availability and occupation status, and the trip attributes starting time, activity duration and day of the week (Axhausen, Schönfelder, Wolf, Oliveira, & Samaga, 2003). Another approach is the estimation of chances of specific purposes with a logit allocation model. Two categories of attributed are found to have a significant influence of the probabilities of purposes: the temporal attributes time of day and activity duration and the spatial attributes of land-use information. Attributes from previous trips proved insignificant (Chen, Gong, Lawson, & Bialostozky, 2010);
- *Machine learning*: based on survey data, learning models can find the relation between attributes and specific purposes and assign purposes to trips based on tree building classification. Used attributes for classification are socio-demographic attributes, trip attributes, such as activity duration and activity start time, and land-use attributes (McGowen & McNally, 2007) (Deng & Ji, 2010) (Montini, Rieser-Schüssler, Horni, & Axhausen, 2014).

Validation of the purpose inference results is possible with external surveys or by means of integrating the GPS tracks in the travel diary survey. Results of correctly classified purposes range from 43% (Bohte & Maat, 2009) to approximately 90% (Wolf, Guensler, & Bachman, 2001) (Deng & Ji, 2010). These large differences can be largely explained by the quality of land-use data, since that is the most prominent category of attributes relating to the travel purpose.



2.4 Conclusions from the literature study

This paragraph summarizes the findings from the literature study. The conclusions are categorized by findings on the discovered literature and the implications for this study.

2.4.1 Conclusions on the discovered literature

Smart card technology has entered all sorts of markets around the world with its many different applications. The data generated by smartcards are one of the main drivers of their success, as they hold valuable information on user behaviour. Likewise, the penetration in the public transport market, by means of AFC, has increased over the years and looks to be increasing still. PT Smart card data is used for many appliances: performance monitoring, market analysis and demand modelling.

Focussing on the smart card data use for travel demand modelling, the literature indicates a high potential thanks to the longitudinal character of data collection and the large sample size, which can approximate complete coverage of the described system. These factors can contribute to a significant quality improvement of current modelling techniques, since no equally rich source has previously been available for travel demand modelling.

However, smart card data cannot be seen as a direct replacement of currently used travel data, collected with surveys and counts. Compared to survey data, essential information is lacking as a result of passive data collection: travel purpose, origins and destinations are unobserved. This information is required for the interpretation of the data. The travel purpose, or activity, provides the reason for travelling, which is essential to long term demand forecasts that deal with policy changes. Origin and destination provide the locations of activities, which are similarly essential to long term demand forecasts as they are subject to changes in land use. Furthermore, the interpretation of check-in and check-out transactions in terms of trips, tours and transfers, depends on definitions used in rule-based processing.

The method of OD matrix construction with smart card data depends on the data structure and the utilisation of the OD matrix. Regarding the data structure, various datasets exist from different AFC systems. Essential elements are the availability of the alighting transactions and the integration with AVL systems. Quality differences exist in the rule-based processing procedures to derive transfers within the system. Additional processing rules can be added to basic rules in order to allow for more exceptions. In contrast, transfers to other systems are not a subject of interest in the discovered literature. Regarding the utilization of the matrices, most studies focus on short term planning and therefore do not incorporate the conversion from matrices from stop-level to zone-level. For long term planning, however, stops do not have the desired stability of indicators for trip production and attraction.

The travel purpose inference is acknowledged as a main issue for strategic planning as it incorporates the motivation for travelling, which is subject to policy measures and societal changes. Therefore, it is a recurring topic in literature on the implementation of new data sources in travel demand modelling. All passive data collection methods encounter this problem and, despite differences in data structure, similar methods of purpose inference have been studied.

Inspired by the large amount of data, several studies deal with the absence of travel purpose information by application of user group clustering. For short term planning,



this method provides insight in travellers affected by disruptions and rerouting. Yet, this method does not suit long term planning as it does not determine the motivation to travel.

The most simple method builds on the data mining procedures and applies rule-based processing to infer travel purposes. These rules involve crude simplifications and cannot differentiate between many travel purposes. More sophisticated methods are based on probabilities and machine learning principles. These methods identify the same variables to be explaining information on the travel purpose. These variables can be categorized in three groups: (1) trip characteristics like activity duration, start time, frequency and card type, (2) socio-demographic characteristics like age and occupancy status and (3) land-use characteristics. The influence of land-use characteristics highly depends on the resolution of the available data. The resolution may vary from aggregated zonal data to high resolution land-use data, which contain individual buildings.

2.4.2 Implications for this study

Assessing the discovered literature, we can conclude that the research questions of this study have been studied frequently in similar contexts. Against the background of this literature, however, we have also found that dissimilarities in available data and eventual employment have great impact on the used methods to answer them.

Regarding the data processing rules to interpret the crude data, the literature provides several options, with different levels of detail, for the identification of transfers and activities. Several issues arise with the interrelation between them. First, inconsistent use of PT results in single trips, for which it is not possible to derive a tour. Another possible result from inconsistent use is the occurrence of "gaps" between subsequent alighting and boarding. These gaps can be interpreted as an activity or as an unobserved transfer: a trip leg made in another system, which could be a private mode or another operator. Second, transfers on the same line distinguish activities, but not all of these activities might be relevant for this study. Different processing rules result in contradictory "observations". The implications of these issues are especially relevant for modelling context based on tours.

Rule-based processing can also provide feasible solutions for the purpose inference, although this reduces the possibilities to incorporate a larger scale of travel purposes. Other feasible methodologies for the purpose inference have a wide range of detail. The Naïve Bayes Classifier used by Kusakabe & Asakura (2014) seems suitable for the enrichment with survey data, but depends on the stability of this data. Since the survey available for this study has been terminated, this method might not be very stable. More deliberate approaches by (Chakirov & Erath, 2012) and (Lee & Hickman, Trip purpose inference using automated fare collection data, 2014) also seem appropriate for this study. Although only applied to high resolution land registration data, the literature has also indicated that the combined estimation of the exact location and the activity holds potential. Therefore the option of combining the models for zone allocation and purpose inference might be an fertile approach.

2.4.3 Overview of applicable methodologies to answer the research questions

The literature has provided several methodologies to tackle the two main problems at hand. Table 2 presents the methodologies that are deemed applicable for this study, with the corresponding attributes and data sources used.

Table 2: Overview of applicable methodologies found in literature



<i>Problem</i>	<i>Approach</i>	<i>Applied attributes</i>	<i>Applied data sources</i>
Zonal allocation	Direct conversion to zones	- Stop locations	- Stop coordinates - zonal boundaries
	Clustering of stops	- catchment areas - movement patterns	- OD matrix - zonal centroids
	Logit allocation	- distances between stops and zones	- stop coordinates - land-use data
Purpose inference	Rule-based processing	- activity duration - departure times - frequencies - card types - purpose specific land use attributes	- survey data - land-use data
	Naïve classifier	Bayes - arrival time - activity duration	- survey data
	Logit allocation	- activity duration - departure time - purpose-specific land use attributes	- survey data - land-use data
	Machine learning three building classification	- Gender - age - job status - activity duration - departure times - purpose-specific land use attributes	- survey data - land-use data - traveller data

The literature does not provide options for the identification of concession traversing transfers. For this specific issue, a straightforward approach has been pursued, based on the available data.

2.4.4 Overview of assessment criteria for the methodology choice

The methodologies found in literature have indicated differences in the eventual application of the results. The eventual application determines the desired quality of the method. On the other hand, the feasibility of a methodology depends on the available data and budget constraints. Table 3 contains the seven assessment criteria that were used in the methodology choice.

Table 3: Methodology assessment criteria

<i>Quality of the method</i>	<i>• Feasibility of the method</i>
Behavioural richness	Fit for available data
Level of detail	Budget constraints
Durability of the method	Flexibility in development
Interpretability	



The methodology choice was based on these criteria. With knowledge of the available data sources, and their qualities and limitations, a selection of feasible methods was derived for both the zonal allocation and the purpose inference. Subsequently, these methods were assessed on their qualities, in relation to the eventual application of the OD matrices in transport models. In addition, feedback on these criteria provided input for the evaluation, which in turn led to recommendations for future research.

The next chapter focusses on this process by first describing the available data sources and, subsequently, the applied methodology.



3 Methodology and data

This chapter considers the methodology to answer the research questions and the employed data sources. In the introduction, we introduced the limitations of smart card data and the aspired coupling with survey data, in order to construct purpose specific OD matrices. Subsequently, the literature review provided an overview of possible methods to achieve this objective and criteria to assess them. In this chapter, we describe the applied methodology and the motivation for choosing it over the other alternatives.

In order to do so, the chapter starts with the specification of the features of the data sources, starting with the OV-chipkaart (paragraph 3.1), followed by the available travel surveys (paragraph 3.2) and the land use data (paragraph 3.3). Based on the available sources and the methodology assessment criteria derived from literature, we motivate the choice of methodology for the enrichment models (paragraph 3.4). Subsequently, the research outline is presented (paragraph 3.5), followed by a more detailed description of the three research phases (paragraphs 3.6 to 3.8). Finally, conclusions about the available data sources and the applied methodology are listed (paragraph 3.9).

3.1 The Dutch smart card: the OV-chipkaart

The Dutch smart card, the OV-chipkaart, is employed by all public transport operators in The Netherlands. Almost all travel products have been converted to the smart card. Therefore, the OV-chipkaart approximates complete coverage of the public transport system, recording all trips made by train and by bus, tram and metro (BTM). The exceptions consist of several operators that still sell un-chipped tickets at the driver and trips with missing check-in or check-out transactions.

The literature on the use of smart card data shows substantial differences between smart card systems applied for revenue collection in public transport. Similar to the Singapore EZ link card, the OV-chipkaart applies revenue collection with distance-based fares. Hence, both at boarding and alighting a transaction is required to determine the distance travelled. The smart card system is coupled with a GPS tracking system, which directly translates the location of transactions to stops. The accurate registration of transaction times allows operators to differentiate their fares between peak and off-peak periods. Furthermore, the personal OV-chipkaart allows for fare differentiation based on age groups, children and seniors, by registering the birth year. The anonymous OV-chipkaart does not register a birth year.

The OV-chipkaart registers the data per transaction. Table 4 contains the attributes relevant for this study.

Table 4: Available attributes in OV-chipkaart data

Attributes	Specifics
Card number	hashed, but uniquely identifiable
Transaction sequence number	Counts transactions by card
Transaction type	check-in or check-out
Date and time	accurate to seconds



Stop/station	coded
Entry stop/station	only for check-out transactions
Concession	also indicates operator
Mode	Train, bus, tram or metro
Line	only for bus and tram
Distance travelled	based on route
Card type	Personal or anonymous
Travel product	Contract type or e-purse
Fare	Full, reduction or unlimited travel

From these attributes, additional information can be derived. Transactions can be aggregated into trip legs via the transaction sequence number, transaction type and the entry station. Subsequently, trips can be aggregated into trips based on a combination of transaction times and locations, depending on the distinction between transfers and activities. With a similar procedure, but different processing rules, trips can be aggregated into tours. The metadata of the available OV-chipkaart data and the applied processing-rules are described in paragraph 4.2.

3.2 The WROOV surveys

The survey data employed for this study comprise the WROOV-light studies. These studies consisted of a yearly travel survey, running from 2003 up to 2009, with the purpose of revenue allocation of the NVB ticketing system. The NVB contained all national tickets for bus and light rail, but did not include the student card and regional tickets. Moreover, the NVB tickets were not valid on the majority of national railways, hence the WROOV survey does not include train travels.

During one month a year, purchased travel products were accompanied by a survey form, requesting the travels made with the product. For *stripenkaart* tickets, one trip was to be declared. For contracts, the most frequent trip was asked, with the corresponding frequency, and one occasional trip. For all trips, the additional question was asked if that trip was also made in the opposite direction. This resulted in 110.000 to 150.000 completed surveys a year, adding up to a total of 1.7 million trips over seven years. This sample size is extraordinary large for a travel survey, especially one focussed on bus and light rail.

The WROOV data contains information on the traveller and on the travels made. However, the dataset is not structured based on trips, but on revenue allocation elements. These can subsequently be aggregated into trip legs, trips and tours, based on the available attributes. After aggregation into trips, the dataset contains the attributes presented in Table 5.

Table 5: Available attributes in WROOV data

Attributes	Specifics
Survey form number	hashed, but uniquely identifiable
Gender	
Age	at the time of the survey
Tour number	always 1 for <i>stripenkaart</i> tickets,



	max 2 for contracts
Trip number	away trip or return trip
number of trip legs in trip	based on indicated transfers
Origin	at the level of PC6
Destination	at the level of PC6
Boarding stop	coded
Alighting stop	coded
Weekday	Monday to Sunday
Departure time	Accurate to minutes
Mode	Bus, tram or metro
Route	for bus, tram and metro
Distance travelled	mostly based on route
Travel product	Contract type or <i>strippenkaart</i> -tickets
Fare	Full, reduction or unlimited travel
Concession	also indicates operator
Frequency	number of tours per week
Travel purpose	7 distinct purposes, as well as <i>multiple</i> and <i>other</i>

The WROOV data does not require data processing to determine transfers or activities, as these are indicated by the respondent. The activity duration can be derived based on the departure times of consecutive trips. Since the alighting time is lacking, the activity duration includes the travel time of the away trip. Therefore, the activity duration derived from WROOV data is slightly overestimated.

One alternative option for survey data in the Amsterdam region was the MON/OViN survey. This survey consists of a yearly survey that is still running, although a change in the data collection method resulted in a trend reversal between 2009 and 2010. However, the complete sample size is much smaller than WROOV, with the essential distinction that MON/OViN includes all transport modes. Hence, the sample size per year for public transport is less than 50 times smaller compared to WROOV (Kuhlman, 2014). Moreover, the MON/OViN data does not include the used public transport stops, which are essential for the allocation of origin and destination zones. Consequently, the MON/OViN data is considered not suitable for this study and has not been employed.

3.3 Land use data

Besides the previously introduced travel data sources, this study has also employed land use data. Other studies have indicated that land use data are valuable for both the zonal allocation as well as the purpose inference. The land use data available for this study originates from the Amsterdam Region transport model VENOM, which is introduced in paragraph 4.1. The land use data contains averaged data per zone in the VENOM zonal grid. Table 6 presents the attributes in the land use data relevant for this study. We classified the attributes, based on their expected influence on the zonal allocation.

Table 6: Available attributes in land use data



<i>Attributes</i>	<i>Classification</i>	<i>Specifics</i>
centroid coordinates	Geographical	based on gravitational centre of the zone
area	Geographical	rounded to hectares
residents	Home-end	
students	Home-end	
working population	Home-end	
cars	Home-end	registered by residents
households	Home-end / Activity-end	/
jobs	Activity-end	
student places	Activity-end	for 5 school categories

The zonal data contains attributes related to the activity end of the purposes work and education, but does not include any data specifically related to the purpose shopping.

3.4 Qualitative comparison of the OV-chipkaart and WROOV

A qualitative comparison between the OV-chipkaart and WROOV has been performed in order to determine the possibilities and appropriateness of transferring information between these sources. First, key variables have been identified by comparing the available information in the sources. Second, the coverage and target population of both sources have been compared, which has provided input for the construction of comparable data sets.

3.4.1 Key variables for coupling of information

By comparing the available attributes in both travel data sources, key variables can be determined. These variables are the instruments of transferring information between sources. In order to do so, the information to be added is to be expressed in terms of these key variables.

Table 7: Key variables for transferring information

<i>Key variables between OV-chipkaart and WROOV data</i>
Activity duration
Frequency
Departure time
Travel distance
Contract duration
Fare
Mode
Operator
number of legs within trip
number of trips within tour
Zonal data at the stop locations



3.4.2 Sample sizes, target populations and data collection periods

In order to determine their validity as medium of transfer, the description of key variables in both data sets has been compared. The more equally distributed the values of key variables are, the higher their validity. The results of this quantitative comparison between data sources are presented in paragraph 5.4. When perceived as valid, attributes have been included in the model estimation process, where their explanatory value has been assessed.

Table 8: Qualitative comparison of the data sources OV-chipkaart and WROOV

	<i>WROOV</i>	<i>OV-chipkaart</i>
Information	Surveyed trips Used modes Used stops Departure time Origin and destination Travel purpose	Observed trips Used modes Used stops Boarding and alighting time
Period	yearly from 2003-2009	1 week in 2014
Coverage	National tickets No local cards No Student cards	All tickets
Location	Amsterdam + Waterland	Amsterdam + Waterland

3.5 Methodology of enriching smart card data

In order to construct purpose-specific OD matrices with OV-chipkaart data, three pieces of information have to be coupled: the origin, the destination and the travel purpose. The WROOV data contains all these three elements, and therefore these can be applied as source of the required information. The literature study has provided several options to allocate trips to zones and infer the travel purpose based on survey data (see paragraph 2.4.3). In addition, a set of assessment criteria was constructed based on the evaluation of this study in context of reference studies (see paragraph 2.4.4).

The number of feasible methods is limited by the fundamental objective and the specifications of the available data. The objective is to create matrices that are applicable as base matrices in transport models. Therefore, the allocation of origins and destinations needs to comply with the zonal structure of the model. Furthermore, the level of resolution of the zonal allocation based on land-use data is limited to an aggregated level due to data constraints. Since origins and destinations recorded in the WROOV data are not recorded as addresses and the available land use data are aggregated at the zonal level, the allocation of origins and destinations to specific addresses is not feasible.



3.5.1 Zonal allocation

Initially, three applicable methods for the allocation of origin and destinations were derived from the literature:

1. Direct conversion of stop locations to zones;
2. Clustering of stops;
3. Logit allocation models.

The direct conversion of stop locations to zones inevitably complies with the zonal level of resolution. However, the accuracy of direct conversion depends on the size of the zones in relation to the access and egress distances. Within a high resolution zonal grid, the chance increases that origins and destinations are not situated in the same zone as the stop. The clustering of stops results in origins and destinations which do not comply with the zonal structure of a transport model. Therefore this method would not result in OD matrices with the desired level of resolution.

Hence, the logit allocation models are perceived as the most appropriate method for this specific combination of research objectives and available sources. Consequently, this study has applied logit allocation models for the allocation of origins and destinations to trips. These origins and destinations consist of the distinct zones classified in the transport model, with the corresponding land use data. Logit models estimate the chance of a specific option being "chosen", relative to the other options. The model parameters to be estimated determine the influence of specific land use variables on the chance a trip originates from or terminates in a specific zone.

In order to investigate the level of resolution at which logit allocation models actually perform better than a direct conversion of stop location to zones, this method has also been applied as a reference. Paragraph 3.8 continues in more detail about the framework of the enrichment models.

3.5.2 Purpose inference

Regarding the purpose inference methods, the literature has provided four applicable methods:

1. Rule-based processing;
2. Naïve Bayes classifier;
3. Logit allocation;
4. Machine learning classification tree.

For this study, we aimed for a disaggregate approach, to match the disaggregate nature of the OV-chipkaart data, eliminating the Naïve Bayes Classifier as desired approach. The rule-based processing approach applies crude simplifications, with resulting errors that might cancel out at an aggregated level. The probabilistic approach of logit allocation fits better to the disaggregate OV-chipkaart data and the uncertainties in the distribution of travel purposes.

The machine learning approach was also considered a feasible method, but for practical reasons we applied the logit allocation. Since logit allocation is also applied for the zonal allocation, both model types are estimated in the same software package Biogeme, which is freely available. Using the same modelling framework and software for both problems limited the required time to get familiar with the software.



Moreover, a combination of both problems in a single model was considered. A combined model, estimating both the destination zone and the travel purpose, proved to be possible. However, the large number of attributes and alternatives⁵ made this model hard and time expensive to interpret. Therefore, this approach has not been sustained.

Similar to the zonal allocation models, the purpose inference models have been estimated in both the trip-based and the tour-based approach. In addition, a simple rule-based processing approach has been applied. Consequently, purpose-specific OD matrices have been constructed for all three approaches. These allowed for a comparison between the approaches, and thereby drawing conclusions on the differences.

3.6 Research outline

Following the literature study and the methodology choice, the remainder of the research has been set-up in three different phases. Primarily, the data analysis was performed in order to provide insight in the correlation in the WROOV data between key variables and the lacking information. In addition, a quantitative comparison on key variables between WROOV data and the OV-chipkaart data provided insight in the transferability of information and the appropriateness of attributes as explanatory variables in the enrichment models.

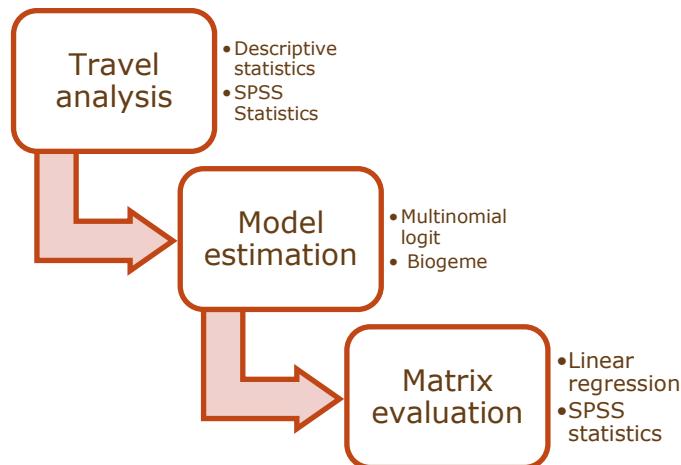


Figure 4: Research set-up

Successively, in the second phase, the estimation of the enrichment models has been performed. This process consisted of the selection of explanatory variables and calibration of the model parameters. Where the initial selection of explanatory variables was based on their explanatory value described in literature and their correlation in the data, this selection process was based on the model performance and the interpretation of the model parameters. Two approaches have been applied, starting with the less demanding trip-based approach, which was successively enhanced to a tour-based approach. For both approaches, this resulted in models that allocate origins and destinations to trips observed in OV-chipkaart data and models that infer the travel purpose of these trips.

⁵ For the combined zonal allocation and purpose inference model, the number of alternatives is equal to the number of available zones, multiplied with the number of distinct purposes.



Finally, in the third phase, the estimated models were applied to both WROOV data, and to OV-chipkaart, in order to construct purpose specific OD-matrices for both sources. The resulting matrices were compared and resulted in the evaluation of five different aspects:

1. Model validation: the application of models onto WROOV data served the model validation by comparing the resulting matrices with the observed matrix;
2. Source comparison: the OD matrices resulting from the same modelling approach on both sources allowed for a comparison between the movement patterns described by WROOV and by the OV-chipkaart;
3. Model approach comparison: the comparison on different levels of resolution allows for the assessment of the added value of more complex model approaches in relation to the level of resolution;
4. Face validation OV-chipkaart matrices: the OD matrices based on OV-chipkaart data allow for a face-validity check based on several high-profile movement patterns.

The research outline is visualized in Figure 5. In the following paragraphs, the methods applied in the successive research phases are described in more detail.



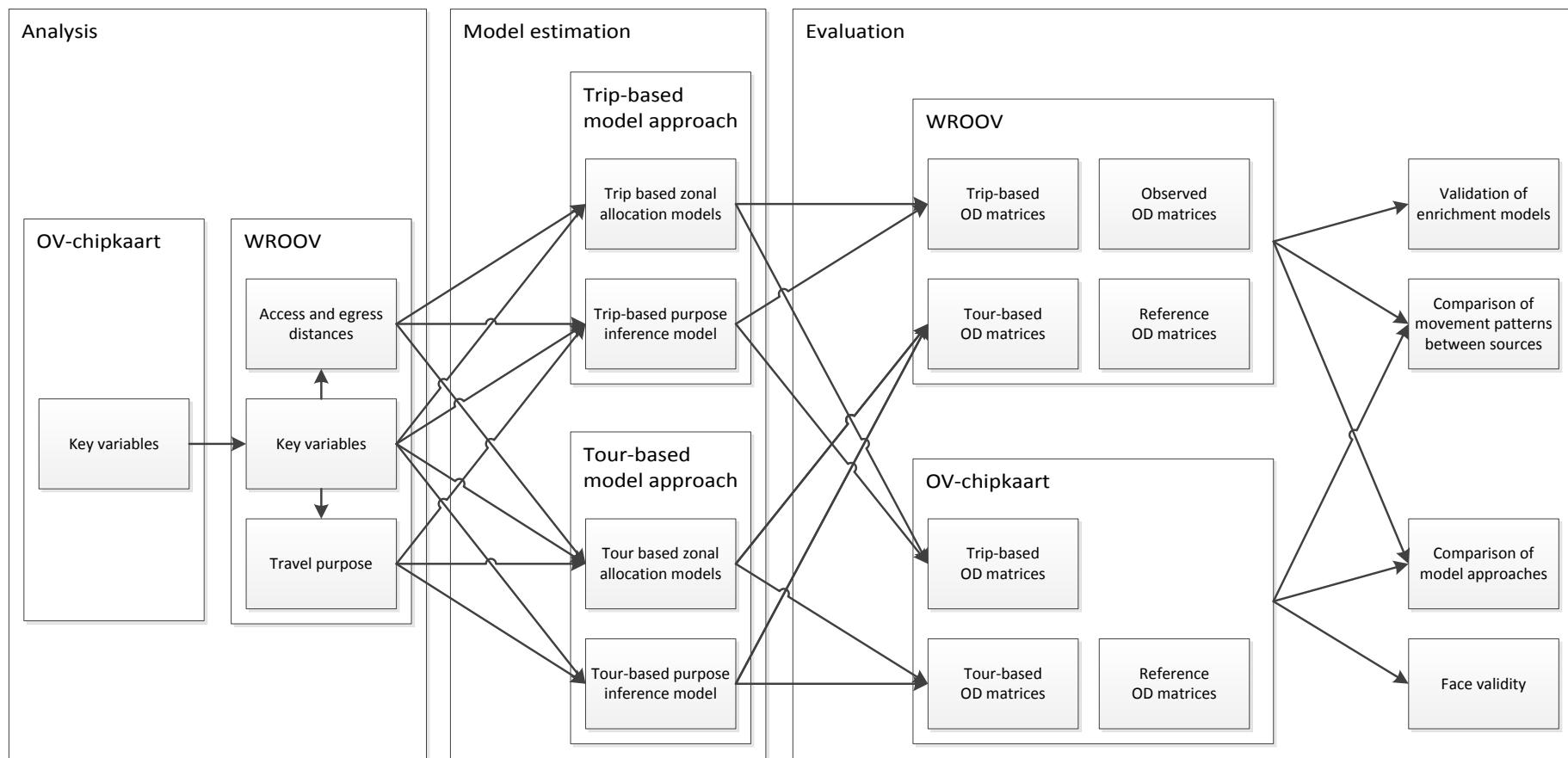


Figure 5: Research outline



3.7 Analysis framework

The literature review has indicated attributes which might have explanatory value in the estimation of access and egress distances as well as attributes correlated with the travel purpose. Several of these attributes are available in both OV-chipkaart data as well as WROOV data, and therefore belong to the key variables which can be used to transfer information between these sources.

The analysis of WROOV data was focussed at relating the access and egress distances and the travel purpose to key variables. This has been done by analysing the descriptive statistics of these variables in the software package of IBM SPSS Statistics and visualization of the distributions over key variables. Regarding the access and egress distances, these analyses have been performed for both the trip-based definitions of origin and destination, as well as the tour-based definitions of home-end and activity-end, in order to evaluate the possible explanatory value of both approaches.

Successively, the description of key variables within both datasets has been compared in order to determine the appropriateness of key variables as medium of the information transfer. This analysis is based on relative frequencies of key variables in both sources.

Key variables that were perceived as appropriate or valuable to the estimation of lacking information were selected into the initial set of attributes for the estimation of the enrichment models. The results from the travel data analysis are presented in chapter 5.

3.8 Modelling estimation framework

The framework of the enrichment models for both the zonal allocation and the purpose inference consisted of logit models, based on their employment in discrete choice modelling. Discrete choice models have been applied in many fields, for example the travel demand related mode choice and route choice problems. Here, we describe the basic workings of this modelling framework. For more deliberate explanations and background, we advise reading the work by Ben-Akiva and Lerman (1985) and Train (2009).

3.8.1 Theoretical background

Discrete choice modelling is based on random utility theory⁶. This theory states that individuals optimize their utility in the choices they make. The utility function of alternatives can be divided into a measurable part, which can be explained by attributes, and an error term.

$$U_{ni} = V_{ni} + \varepsilon_{ni}, \quad \forall i \in I \quad 3.1$$

With:
 U_{ni} = utility of alternative i for individual n
 V_{ni} = systematic utility of alternative i for individual n
 ε_{ni} = error term of alternative i for individual n
 I = choice set

⁶ An alternative to the optimization of utility is the minimization of regret. This theory has been developed based on the idea that not all choice processes are based on utility, but different choice strategies exist. (Chorus, Arentze, & Timmermans, 2008)



The measurable part of the utility function consists of a vector of attributes multiplied with their corresponding utility coefficients. These coefficients are assumed to be constant over individuals.

$$V_{ni} = \sum_{k=1}^K \beta_k * X_{ik}, \quad \forall i \in I \quad 3.2$$

With:
 V_{ni} = systematic utility of alternative i for individual n
 β_k = Utility coefficient for attribute k
 X_{ik} = value of attribute k

The probability of an alternative being chosen depends on the utilities of all available alternatives. Depending on the assumptions regarding the error term, different model specifications result in different expressions of the probability formula. Multinomial logit (MNL), which is the most simplified and commonly used form of a logit model, assumes the error term to be Gumbel distributed with the Independent of Irrelevant Alternatives (IIA) property. This implies that the choice set contains all relevant alternatives and error terms are uncorrelated between individuals and alternatives. While this may not be realistic in many situations (Train, 2009), this causes the error terms to cancel out in the probability formula, leading to the MNL formula in equation 3.3 (Ben-Akiva & Lerman, 1985).

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_{j \in I} e^{V_{nj}}} \quad 3.3$$

With: P_{ni} = the probability of individual n choosing alternative i from choice set I
 V_{ni} = measurable utility of alternative i for individual n

The model estimation process consists of the optimization of utility coefficients of the specific attributes to fit a dataset with choice observations. Several methods are available for this process. We have used the software package Biogeme, which uses optimization of the log-likelihood of the model describing the data. In addition to the assessment of the model fit, the significance of individual parameters can be tested by means of the t-test. Based on this test, relevant attributes to be included in the model can be selected.

3.8.2 Application in allocation problems

The problems at hand do not comply with the theoretic definition of a choice problem. Although the home zone of a traveller is related to the residential location choice, this is not a choice we consider here. On the activity end of the trip, the traveller generally has made the choice of destination long before alighting at the PT stop. The travel purpose is also a choice that is made before the trip is initiated. Hence, the model framework is in this case only used for the optimization of the probability that a trip originates or ends in a specific zone, or is made for a specific purpose. Hence, the discrete choice modelling framework is used for three separate allocation problems.

3.8.3 Distinct zonal allocation models

Two approaches have been pursued in the estimation of logit allocation models: a trip-based approach and a tour-based approach. In the-trip-based approach, trips are handled as uncorrelated units of travel. This resulted in two distinct models for both trip ends: one for the origin zone allocation and one for the destination zone



allocation. Conversely, in the tour based approach, trips are treated as correlated within tours. This ensures consistency of origins and destinations between trips within the same tour. The tour-based approach resulted in four distinct models. Besides the home zone and activity zone allocation models, which are estimated for tours, these also include origin zone and destination zone models for non-home-based trips (Figure 6).

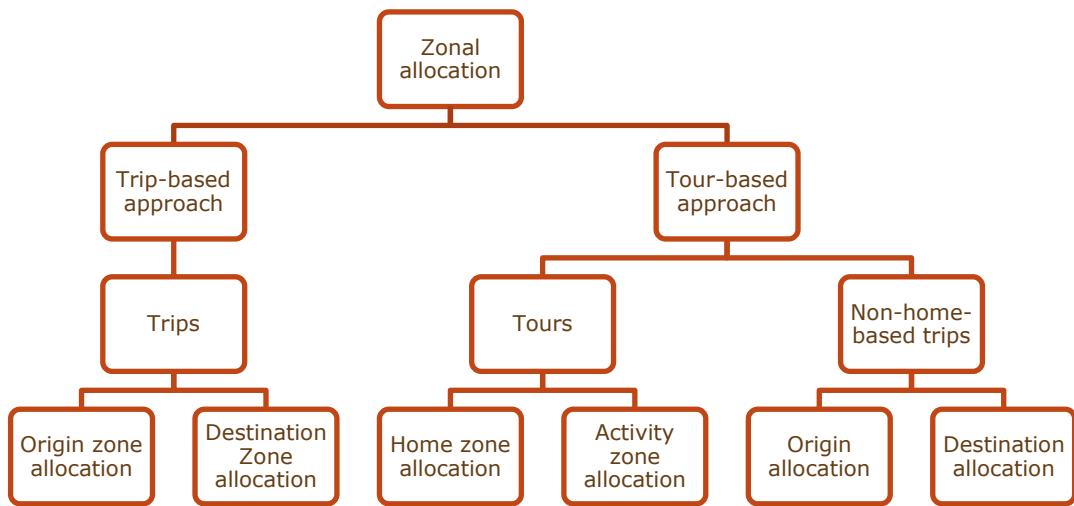


Figure 6: Classification of zonal allocation models

3.8.4 Generic model structure for zonal allocation models

The choice set generation yields the available alternatives per stop. Yet, the zonal allocation needs to be applicable for every stop, since it was not preferable to estimate a distinct allocation model for every stop. Stops at different locations have different zones as alternative origins or destinations. Consequently, the available alternatives have to be generic.

Since the utility of alternatives does not depend on trip characteristics, as these are equal for all alternatives, the utility only depends on zonal characteristics. In order to create generic alternatives, the alternative zones are numbered based on their share of the catchment area and successively matched to the zonal data obtained from the VENOM model (see Figure 7). The data file then consists of records that each consist of the observed trip-end zone and the zonal data for every alternative. If the number of available alternatives is lower than the maximum number of alternatives, the remaining alternative numbers are marked as unavailable with an availability identifier. In order to prevent the ranking of alternatives to influence the choice probabilities, the alternatives are randomly distributed over the maximum number of alternatives. By including the zone number and availability identifier in the randomization, it is still possible to identify to which zone each alternative corresponds.



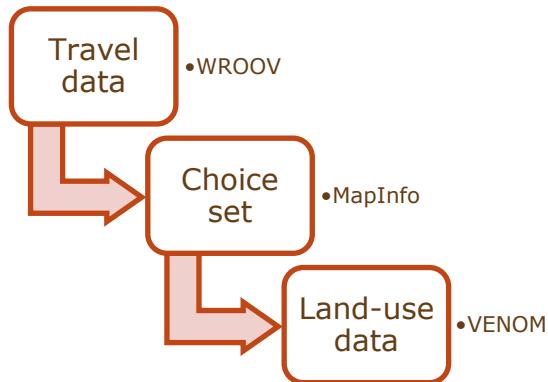


Figure 7: Data handling process of generic model structure for zonal allocation

Table 9 provides a simplified version of the specification of choice alternatives in the Biogeme model files, indicating the generic character of the alternatives. The number of specified alternatives (n) depends on the choice set generation. To which zone the alternatives correspond depends on the used stop and can be determined from the data file. The zonal characteristics, attributes X_1 to X_m , are included for all n alternatives, but set to the Biogeme missing value if the alternative is unavailable.

Table 9: Generic model specification for zonal allocation in Biogeme

<i>Alternative</i>	<i>Availability</i>	<i>Utility specification</i>
Zone ₁	Availability ₁	$B_1 * X_1(\text{Zone}_1) + B_2 * X_2(\text{Zone}_1) + \dots + B_m * X_m(\text{Zone}_1)$
Zone ₂	Availability ₂	$B_1 * X_1(\text{Zone}_2) + B_2 * X_2(\text{Zone}_2) + \dots + B_m * X_m(\text{Zone}_2)$
:		
Zone _n	Availability _n	$B_1 * X_1(\text{Zone}_n) + B_2 * X_2(\text{Zone}_n) + \dots + B_m * X_m(\text{Zone}_n)$

The generic specification of alternatives provided the opportunity to apply identical model specification files for both the trip-based as well as the tour-based approach. Moreover, adapting the number of alternatives ensured the same structure was also applicable for the alternative approaches of choice set generation.

3.8.5 Distinct purpose inference models

Similar to the zonal allocation models, the purpose inference models have been estimated with the trip-based approach and the tour-based approach. This ensures that the complete construction process of purpose-specific OD matrices is executed within these two approaches.

Consequently, three distinct purpose inference models have been estimated: purpose inference model for all trips, a purpose inference model for tours, and a purpose inference model for non-home-based trips.

3.8.6 Model structure of purpose inference models

In contrast to the zonal allocation models, the purpose inference models do not require a generic selection of alternatives. The available alternatives consist of a predetermined number of distinct purposes. The WROOV surveys differentiated nine travel purposes, including *multiple* and *other*. Previous research indicated that many of these purposes cannot be distinguished based on travel patterns (Kuhlman, 2014). Based on their frequency, four distinct purposes are identified as the most relevant purposes to be included in the model:



- Work;
- Education;
- Shopping;
- Other.

The attributes of the purpose inference model consist of trip characteristics. Since information can only be transferred via attributes that are available in OV-chipkaart data as well, these characteristics are limited to key variables.

In addition to these trip characteristics, land-use data at both ends of the trip has been implemented as attributes. Since the origins and destinations are not exactly at the used stops, the land-use data has been averaged over an area within a radius of 400 metres around the stop. This distance is chosen conservatively, since the averaging effect of the land-use data relates to the area, which increases quadratic with the radius. As a result of the unknown origin or destination, the land-use data values applied might not be consistent with the real land-use data at the origin or destination.

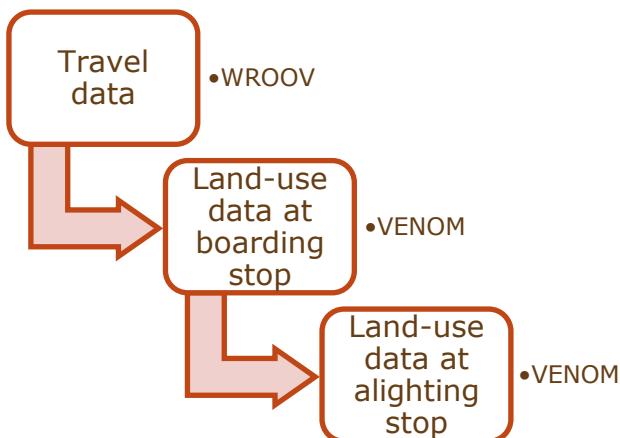


Figure 8: Data handling process of purpose inference models

The results from the estimation of the enrichment models are presented in chapter 6.

3.9 Matrix evaluation framework

The model estimation phase produced three different sets of enrichment models: the rule-based reference models, the trip-based models and the tour-based models. All three sets have been applied to both the WROOV data and the OV-chipkaart data, resulting in three different sets of OD matrices per data source. In addition, we also have observed OD matrices from the WROOV data, leading to a total of seven sets of OD matrices. All of these sets consist of OD matrices for three different divisions in space, four divisions in time, and five divisions by purpose. Hence, a total of $7 * 3 * 4 * 5 = 420$ specific matrices are available for the evaluation. Not all of these matrices are worthwhile to compare with each other.

First, in order to draw conclusions on the model validity, the OD matrices based on observations in WROOV have been compared with the OD matrices constructed with the three modelling approaches. Since observed OD matrices from OV-chipkaart data are not available, we assess distinct OD pairs on face validity. Second, in order to draw conclusions on the comparability of the sources, the OD matrices of WROOV and



OV-chipkaart based on the tour-based models have been compared. Finally, in order to draw conclusions on the differences between sources, the OD matrices of OV-chipkaart based on the tour-based models have been compared to the OD matrices based on the trip-based models and the reference models. An overview of the exact matrix comparisons is presented in chapter 7.

Table 10: Matrix divisions

<i>Matrix division dimension</i>	<i>Divisions</i>
Space	PC3
	PC4
	VENOM zones
Time	Average working day
	Morning peak/Off-peak/Evening peak
Purpose	All purposes
	Work/education/shopping/other

The method of evaluation of OD matrices is not straightforward. The number of cells within OD matrices increases quadratic with the number of distinct zones in the study area of the transport model. With a high level of resolution, this implies a large number of specific movement patterns between zones. Moreover, the numbering of zones within a study area, usually based on postal codes, might not be related to their geographical situation. Therefore, comparing OD matrices is limited to techniques which have their limitations (Allos, Merrall, Smithies, & Fishburn, 2014) (Pollard, Taylor, & van Vuren, 2013).

The British Department for Transport applies a quite rigid method by comparing OD matrices based on a linear regression of the cell values. This method is applied to determine the difference between base matrices before and after the calibration process (Department for Transport, 2014). This objective is similar to the comparison pursued in this study.

Another assessment option is the Mean Structural Similarity (MSSIM) index. This method compares cells in the OD matrix per block in order to assess structural differences (Djukic, Hoogendoorn, & Lint, 2013). However, techniques based on structural similarity are still under discussion (Pollard, Taylor, & van Vuren, 2013). Since the structural similarity in the matrices based on OV-chipkaart data is guaranteed based on stop locations, this method has limited additional value in the evaluation of the applied estimation models.

In addition to the evaluation with linear regression, the comparison of different modelling approaches based on the network assignment would provide additional assessment options which directly relate to the influence of the differences between modelling approaches on the network performance. However, this comparison would require many time-consuming model runs and insight in the assignment models, which is considered a very interesting follow-up study.

Hence, the applied method of evaluation consists of linear regression. The equation of the linear regression line is presented in 3.4. This technique estimates the cell values of one OD matrix (the dependent variable) based on the cell values of another OD matrix (the one independent variable). This results in three different parameters that can be assessed:



1. The r^2 statistic: this statistic is a measure of the model fit, which represents the explained variance in the dependent variable by the independent variables. Its value ranges from 0, indicating no explained variance at all, to 1, indicating fully explained variance;
2. The a parameter: this parameter represents the intersection of the regression line with the y axis in the regression formula;
3. The b parameter: this parameter represents the slope of the regression line, which is the derivative of y in respect to x .

$$y = a + b * x$$

3.5

In case of complete equal matrices, the results would indicate a r^2 statistic of 1, a parameter a with value 0 and a parameter b with value 1. The closer the statistics approach these values, the more matrices are alike. In order to assure enough comparability between matrices before and after calibration, the Department For Transport requires r^2 values of 0.99. This value is not considered feasible for comparing OD matrices based on different modelling approaches.

The results from the matrix evaluation are presented in chapter 7.

3.10 Conclusions regarding the methodology and data

In order to provide an overview of the implications of the available data sources and the applied methodology on the results of this study, this paragraph lists the essential conclusions from this chapter.

3.10.1 Conclusions regarding the data sources

The literature already indicated the importance of the data structure of smart card data in relation to the quality of its employment in travel demand studies. In addition, we found that this also holds for the use of survey data and land-use data used to enrich smart card data.

The Dutch OV-chipkaart has a rich data structure, since it contains both boarding and alighting transactions, as well as the used travel product. Moreover, the transactions are automatically coupled to stops by an integrated GPS system. Besides the rich data structure, the coverage of the system is also high compared to similar systems around the world, as it is the only valid ticketing system in most regions.

The WROOV data contain all the required information for the enrichment of smart card data. This survey includes the used stops, which allows for an analysis of the access and egress distances, as well as the travel purpose. The sample size is relatively large for a travel survey and fully focussed on transport with bus and light rail. The only drawback of this source is that the survey has been terminated in 2009, which limits the appropriateness of the results over time.

The land-use data from the VENOM model match the zonal grid, which allows for a straightforward linkage with zones as alternatives in the logit zonal allocation models. The data contain specific attributes related to the home-end and to the activity-end for purposes work and education. However, no directly related attributes are available for the activity-end of the purpose shopping.



3.10.2 *Conclusions regarding the methodology*

The methodology of enrichment of OV-chipkaart data is based on logit allocation. These models estimate the probability a specific alternative is “chosen” relative to the other alternatives, based on their explained utility. During the model estimation, the influence of attributes on the utility of an alternative is determined. Concerning the zonal allocation models, these alternatives consist of zones nearby the used stop. Attributes that influence the utility of these attributes are zonal characteristics from the land-use data. Concerning the purpose inference models, the alternatives consist of specific travel purposes. Attributes that influence the utility of purposes consist of travel characteristics and land-use data near the used stops. The land-use data values are averaged around the used stops, and therefore might not be consistent with the real land-use characteristics at the origin or destination.

Nonetheless, the models estimating the lacking information are handled separately. Combined models might increase the added value, as they consider the complete movement patterns. This ensures consistency of the land-use data with the allocated zones. Due to high complexity and computation times, this approach has not been persevered.

The logit models allow for a disaggregate approach, estimating the origin zone, the destination zone and the travel purpose for individual trips. The specific attributes influencing the allocation are interpretable by the estimated model parameters.

Three different approaches have been applied to construct purpose specific OD matrices based on OV-chipkaart data. The trip-based approach and the tour-based approach use logit allocation models, where the trip-based approach does not take into account the correlation between successive trips, and the tour-based approach does. In addition to these logit models, a straightforward rule-based processing approach has been applied in order to assess the added value of the more complex logit models.

The evaluation of the resulting matrices consists of a series of comparisons between differently constructed OD matrices. Differences consist of the data source, the spatial resolution, the time resolution and the purposes specified. The comparisons are based on linear regression, which indicates overall comparability of OD matrices but does not include structural similarity. Evaluation by means of a MSSIM index has not been pursued since the added value is unknown. A network assignment of the OD matrices can provide additional assessment options, but is time expensive. This topic of route choice calibration, based on OV-chipkaart data, is considered as an interesting follow-up study, but not feasible within this research.



4 The Amsterdam region case study

This chapter presents the case study that has been performed in order to verify the potential of the proposed methodology of constructing purpose-specific OD matrices. In the previous chapter, this methodology has been introduced, based on the available data sources in The Netherlands. The OV-chipkaart is a national smart card system, containing many different regions, with unique public transport systems. This study was concentrated on the Amsterdam region. Here, we go into further detail on the eventual application of the resulting OD matrices and the available data sets.

The literature review already indicated that the preferred methodology depends on the eventual application. Therefore, the chapter starts with the specifications of the intended application in the strategic transport model of the Amsterdam region: the VENOM model (paragraph 4.1). Subsequently, the availability of OV-chipkaart data for this study is presented (paragraph 4.2), followed by a qualitative comparison between the data sets (paragraph 4.3). Finally, the implications of using this specific case study for the generalizability of the method are discussed (paragraph 4.4).

4.1 Eventual application of OD matrices in the VENOM model

The VENOM model is the strategic transport model of the City Region of Amsterdam (SRA), which has been officially in use since 2012. The model forecasts the travel demand for an average working day in the Amsterdam region. The study area covers the larger metropole area of Amsterdam, which includes adjacent areas that have a large influence on the travels in and around Amsterdam. For example, the city of Almere is not part of the Amsterdam region, but is known to inhabit many commuters travelling to Amsterdam. The models zonal grid is adapted from the Dutch postal code system. The spatial resolution is slightly higher than the PC4 level, which means that PC4 zones are generally divided into several VENOM zones, with smaller zones in more urbanized areas.

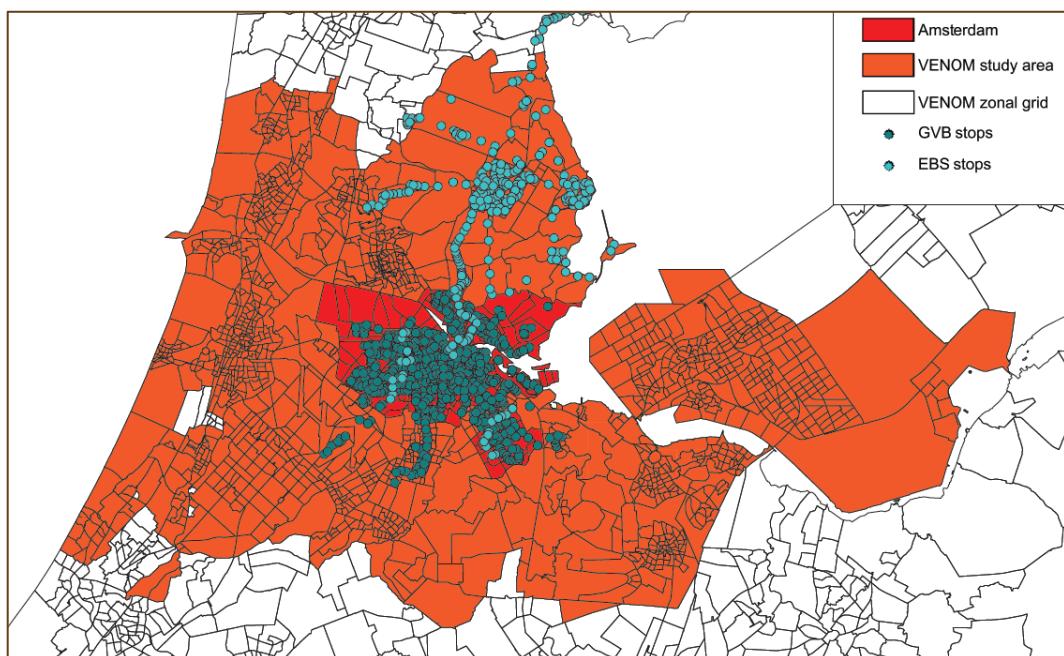


Figure 9: VENOM study area



The VENOM model applies the same pivot point procedure as the main Dutch transport model systems LMS and NRM (Rijkswaterstaat, 2012), but applies a different approach regarding the travel demand of public transport. The VENOM model generates synthetic OD matrices for both train and BTM. These matrices are enriched with survey data because the model does not provide consistent matrices. The matrices for train and BTM are then summed and this public transport matrix is successively assigned to the network. Based on the network loads, the public transport OD matrix is calibrated to improve the fit with cross-section counts (Kieft, Herder, & Pieters, 2013).

The influences of the enrichment and calibration processes on the OD matrix are substantial. During the enrichment the internal trips are increased by approximately 50% for the average working day. During the calibration process, an additional 16% is added to the number of trips. These large alterations invalidate the pursued consistency within the pivot point method. Therefore, the application of OV-chipkaart data for the construction of OD matrices is assumed to provide higher quality OD matrices, although these do not ensure consistency with the synthetic forecast matrices either.

Compared to the OD matrices for road transport, the public transport matrices are not distinguished by purpose. This decreases the model capabilities to specifically implement the influences of policy measures. Furthermore, the currently applied empirical sources for calibration require updating.

This case study initiates several improvements to the currently used OD matrices, which can expand the capabilities of the VENOM model:

- Employment of recent data;
- Employment of high volumes of observed trips;
- Differentiation of matrices by travel purposes.

4.2 Availability of OV-chipkaart data

The responsibility of governance of the Dutch public transport systems lies with regional authorities. The only exception is the concession of the national railways, which is covered by the national government. The SRA is responsible for the procurement of four public transport concessions in its region. This results in the presence of five different public transport concessions in the Amsterdam region⁷.

Table 11: Public transport concessions in the Amsterdam region

<i>Concession</i>	<i>Operator</i>	<i>Modes</i>	<i>Available</i>
Amsterdam	GVB	Bus, tram and metro	yes
Waterland	EBS	Bus	yes
Amstelland-Meerlanden	Connexxion	Bus	no
Zaanstreek	Connexxion	Bus	no
National railways	NS	Train	no

⁷ It has to be noted that more public transport concessions exist within the study area, since that also includes adjacent regions under control of other authorities than the SRA.



For this study, the data from the concessions Amsterdam, operated by GVB, and Waterland, operated by EBS, have been used. Their area of operation is indicated by their stop locations in Figure 9. This means that the OD matrix of the study area can only be constructed partially. Moreover, concession traversing transfers to the other operators cannot be determined. Since these issues result from the unavailability of data from other operators, rather than limitations of the data itself, we have not pursued enhanced identification of these transfers. When more operators are prepared to contribute to studies that exceed concessionary boundaries, it will be possible to study complete public transport trips. This case study does combine OV-chipkaart data from two different operators, and thereby demonstrates that the proposed methodology of enrichment works for concession-traversing studies. Consequently, this study can stimulate the availability of OV-chipkaart data for future research.

The period for which the travel data is available includes the entire year 2014, for both GVB and EBS data. This long period allows for a longitudinal analysis of the year. However, the vast amount of data also results in long computation times. Therefore, a longitudinal analysis has been performed on several travel characteristics in order to determine a single week, which best represents an average working week. The data of the selected week has been applied in the construction of the OD matrices. The longitudinal analysis is presented in paragraph 5.4.2.

4.3 Matching the WROOV dataset to the OV-chipkaart dataset

The stacked WROOV data of the period 2003 – 2009 contains 279.374 trips within the structural boundaries of the public transport concessions Amsterdam and Waterland. In order to comply the dataset with the target population and to remove missing values, the dataset is filtered with the following selections:

- Only weekdays;
- No student cards⁸;
- No missing origins and destinations;
- No missing stops;
- No missing travel purpose.

After this selection procedure, 204.041 trips (73%) remain in the dataset, which comes down to an average of nearly 30.000 observations per year. Trips with missing origin, destination or stops are removed from the dataset. Since Biogeme cannot deal with missing values, these cannot be used in the estimation. The removal of trips with missing travel purpose is applied with the aim of equivalent datasets for the zonal allocation and the purpose inference.

Shifting to the tour-based perspective, the remaining total of 204.041 trips can be classified by tours and non-home-based trips. The used survey for WROOV data collection ensured tours can only consist of two trips: an away trip and a return trip. Only 10% of the trips are classified as non-home-based trip, the majority of 90% is part of a tour (see Figure 10)⁹.

⁸ In general, student cards are not included in the WROOV studies. However, students were allowed to travel with NVB tickets on a reduced fare, in times their student card was not valid.

⁹ The trips within tours and single trips add up to 204.033 trips. The remaining 8 trips that complete the dataset to the 204.041 trips are tours with missing values in one of the trips.



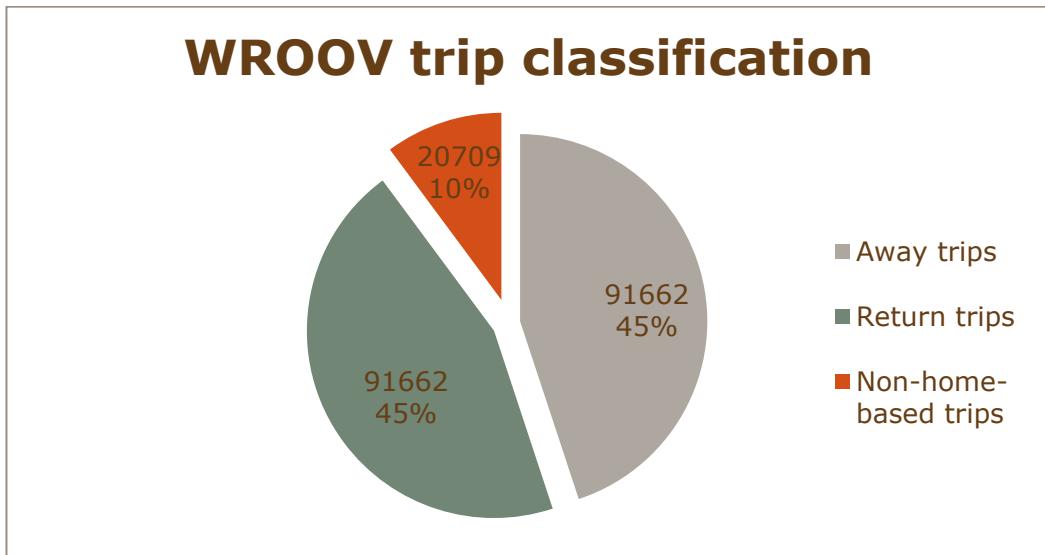


Figure 10: Share of tours and non-home-based trips in the WROOV data

In order to reduce the computation times, for both the trip based as the tour-based modelling approach, a random sample of 40.000 records (20%) is used as training set. This implies that the allocation to zones is independent from time. That is, final model parameter values were expected to be equal for every year within the WROOV data. To validate this assumption, the total dataset is divided per year as well (see paragraph 6.2.4).

Finally, one last selection is applied to the datasets used for the model estimation: the “chosen” zone has to be available in the choice set. From discrete choice theory, we know the choice set should always contain all potential alternatives. Regarding this specific allocation problem, however, it does not make sense to take every possible exception into account. In theory, every zone in the study area is an alternative since it is possible to walk, cycle or drive to any zone. For the construction of base matrices we do not want to consider exceptions where travellers cover large distances to and from the PT stops they use. Therefore, chosen zones that are not within the catchment area of a stop are filtered from the data. Paragraph 6.2.1 provides the details of the choice set generation by means of catchment areas and presents the number of observations used of the model estimation per modelling approach and choice set generation method.

While the Biogeme software does provide the option to apply weights on observations, the weights available in the WROOV data are not applied in the estimation. These weights are based on the division of the revenues of ticket sales and are not constructed with the goal to represent the origins and destinations of travellers. In addition, the selection procedures that have been applied on the original WROOV dataset invalidate the weight factors (see Appendix A). It is possible to compensate for the applied selections, but because of the complex and unclear determination of these factors, we decided to use the observations without weights.

4.4 Generalizability of the case study

With this case study we aim to demonstrate the value of the method of enriching smart card data. In order to determine the implications of using this specific case study, this paragraph discusses the generalizability of the case study, concerning the



structural boundaries due to limited availability of data and the eventual application of OD matrices.

This study focussed on the travel with bus and light rail, where a complete description of the travel with all public transport modes would have been more valuable. However, for both the WROOV survey and the OV-chipkaart, no data concerning travel by train was available. Regarding OV-chipkaart data, this is a result of unavailability of data, in contrast to WROOV, which does not cover train travel.

Amsterdam is a unique situation for public transport in the Netherlands, which may comprehend significantly different travel patterns compared to other regions. Only the larger cities in The Netherlands provide public transport by light rail. Furthermore, Amsterdam attracts high volumes of tourists, both domestic and foreign. These factors might result in different model parameters. The estimated enrichment models for the case study, therefore, may not be applicable for other regions. Nonetheless, the WROOV data provide the opportunity to estimate the models for specific regions, since it covers the entire county. Moreover, the OV-chipkaart data structure is equal over the entire country. Hence, a similar procedure is applicable for other regions.

Besides different travel patterns and utilization of the public transport system, also differences in the smart card system require consideration when relating this case study to application of smart card data in travel demand studies abroad. The OV-chipkaart registers both boarding and alighting stop, as well as transaction locations. For ticketing system based on flat fares or systems without an integrated GPS system, techniques are available to infer the required information to employ the methodology of this study (see paragraph 2.2). However, the data quality will be less for inferred attributes than observed attributes.

The aimed application of OD matrices in the VENOM model is limited due to the data constraints. The fact that VENOM applies combined public transport matrices increases the limitations due to the unavailability of train data. On the other hand, the resulting matrices, based on data from the concessions Amsterdam and Waterland might also be valuable for the Amsterdam City Model (VMA).



5 Public transport travel analysis

This chapter presents the results from the travel analysis that has been performed on the WROOV and OV-chipkaart datasets. The literature review has provided an indication of attributes relevant to the information we want to add to the OV-chipkaart data. Chapter 3 described the data sources and which attributes are applicable as key variables in the model estimation. These attributes have been examined in the WROOV data in order to provide an indication of their predictive value. In addition, characteristics which are not available in OV-chipkaart data, but could have a possible predictive value, have been examined to comprehend the overall predictive value of the key variables. Furthermore, the stability of the information to be added has been investigated in order to derive the durability of their predictive value. Subsequently, a comparison of the datasets on key variables provides insight in the suitability of key variables as predictors. These classifications of predictors form the foundation of the model estimation, described in the next chapter.

First, the lacking information in OV-chipkaart data is analysed in the WROOV data:

- the access and egress trip legs are expounded as a foundation for the conversion of stop-based matrices to zonal matrices (paragraph 5.1);
- a depiction of the distribution of travel purposes provides input for the estimation of the purpose inference models (paragraph 5.2);
- the description of concession traversing transfers provides information for the filtering of trips with origins or destinations outside the study area (paragraph 5.3).

Then, after these analyses of WROOV data, a comparison with OV-chipkaart data on key variables is presented (paragraph 5.4). The chapter is concluded by a summary of the findings from the data analysis (paragraph 5.5).

5.1 Access and egress trip legs

In paragraph 1.4 we introduced different definitions to describe the trip ends: the trip-based *origin and destination* or the tour-based *home-end and activity-end*. In the analysis of access and egress trip legs, the tour-based definition provides more insight through the additional information it contains. Since the WROOV data set contains mostly tours with two trips (see Figure 10), the access and egress trip legs are registered once at both ends, resulting in almost equal statistics for both access and egress trip legs in the trip-based definition. Considering the tour-based definition, access and egress trip leg statistics are specific for the home-end and activity-end. Hence, the trip-based definition allows for analysis on access and egress distances combined as one phenomenon, where the tour-based approach allows for analysis of the specific trip ends.



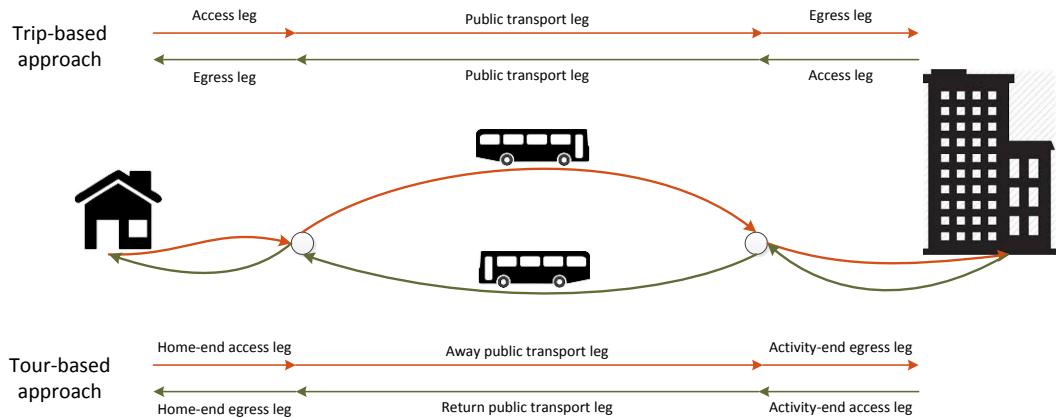


Figure 11: Trip-based and Tour-based definitions of trip legs

Origins and destinations, as well as the boarding and alighting stops are registered in the WROOV dataset. Origins and destinations are available at the level of PC6 zones, which are not equivalent to addresses, but comparable to housing blocks. These small zones are represented by the coordinates of their centre. Stops are represented by the coordinates of the stop cluster, which means that the clustered stops in opposite directions of the same line are represented by the same coordinates. Access and egress distances are calculated by means of the Euclidean distance between these coordinates, which is not equal to actual walking or cycling distance, as these depend on the local infrastructure. Hence, the access and egress distances considered in this study are a slight underestimation¹⁰ of the actual distances travelled.

Data cleaning has been applied to reduce the effects of wrongly registered origins and destinations, resulting in approximately 13% of the data to be excluded for the analysis of access and egress trip legs. The data cleaning process is described in Appendix A. However, some erroneous data remains, with very large access and egress distances. Most likely, these are caused by switched origins and destinations. Since approximately 93% of the distances are below 1500 metres, we focus on this interval in the analysis in the remainder of this paragraph.

5.1.1 Key variables

First, we have aimed to relate the access and egress distances to attributes available in OV-chipkaart data. Previous studies (Utsunomiya, Attanucci, & Wilson, 2006) (Alshalalfah & Shalaby, 2007) have indicated that the access and egress distances depend on the level of service provided at the considered stop. The level of service is described by the frequency, the speed and the directness of the transport service. Travellers are prepared to walk or cycle further for transport with higher speeds and higher frequencies. Since these characteristics are not readily available, we investigated the distances by mode, as these pertain different levels of service.

The analysis shows different distributions of access and egress distances between modes. On average, travellers cover longer distances to and from metro stops. The difference on the home side is larger than on the activity side. Moreover, on the home side travellers cover larger distances to and from tram stops compared to bus stops, while this deviation is not observed at the activity end.

¹⁰ Assuming a rectangular infrastructure pattern, the maximum underestimation is a factor $\sqrt{2} \approx 1.4$



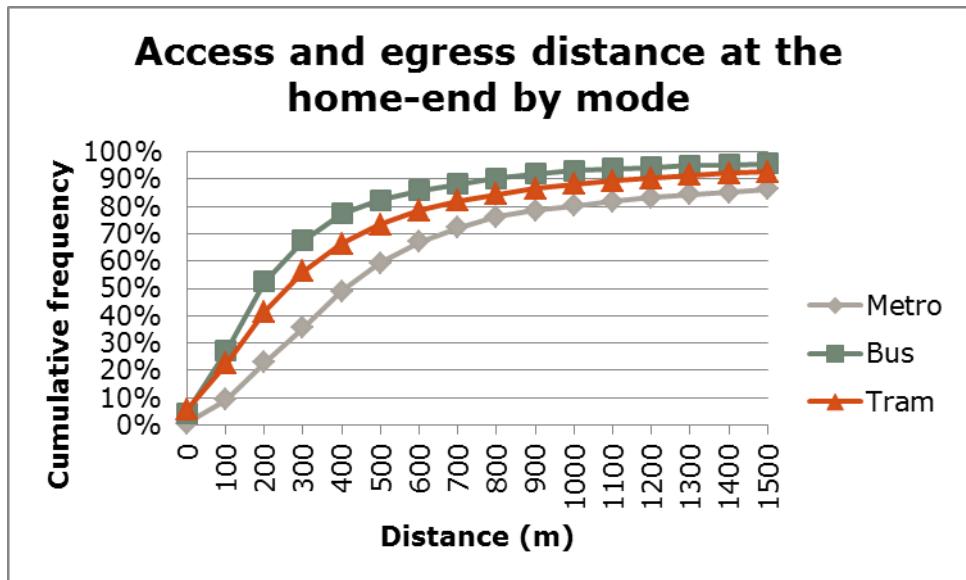


Figure 12: Access and egress distance distributions by mode at the home-end

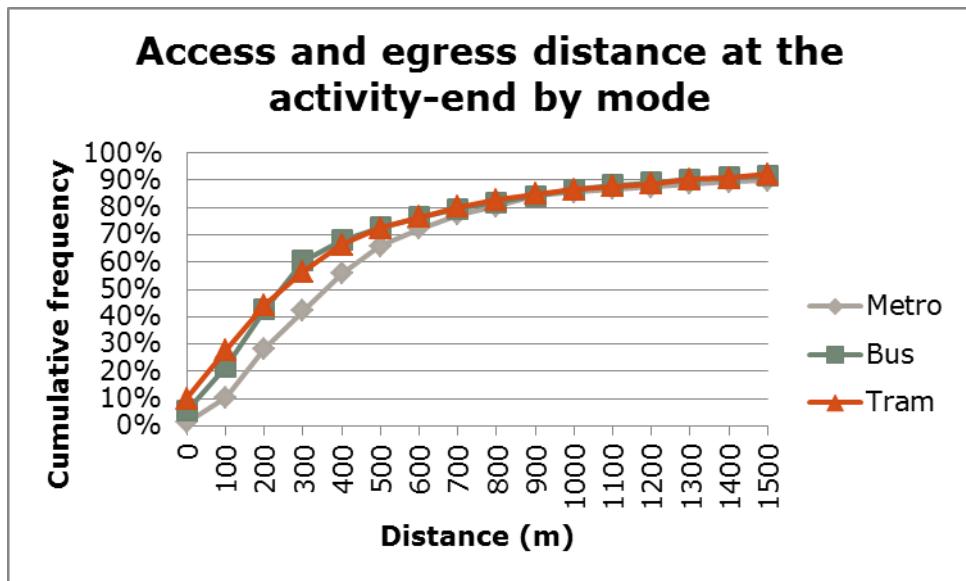


Figure 13: Access and egress distance distributions by mode at the activity-end

Besides the mode, we also expected a relation between the degree of urbanization and the level of service. In more urban areas, travellers have more options in their stop and route choice, while in more rural areas, travellers are more likely to have lesser options and can be classified as "captives" of the network, and consequently need to cover longer access and distances. However, the analysis of the degree of urbanization at the used stop does not present a clear correlation. While on average longer distances are covered in very rural areas (class 5), in very urban areas (class 1), distances are similar to other degrees of urbanization. This might be caused by the fact that the metro system only serves highly urban areas. Again, differences are only observed at the home-end and are less apparent at the activity-end.

For travel distances below 15 km, no differences are observed between the home-end and the activity-end. However, for longer travel distances, the average access and egress distances decrease, where they remain equal at the activity-end.



Distributions of access and egress variables with additional key variables are presented in Appendix B.

5.1.2 Characteristics unavailable in OV-chipkaart data

The influence on the access and egress distances of three personal traveller attributes, which are not available in OV-chipkaart data, has been investigated: the travel purpose, the gender and the age. The results of these analyses provide additional insight in the relative explanatory value of the key variables. Graphs related to these analyses can be found in Appendix B.

From these analyses we conclude that the gender is not related to the access and egress distances on either end of the trip. Regarding the travel purpose and the age of the traveller, differences in access and egress distances are observed at the activity end. Children and especially senior travellers cover shorter distances on average between the used stop and their activity location. Furthermore, trips made for the purpose of *shopping*, and to lesser extent the purpose *other*, contain shorter access and egress legs. This can be explained by two possible phenomena. Either travellers are prepared to cover larger distances for the purposes *work* and *education*, or shopping locations are better served by public transport than offices and schools.

At the home end, access and egress distances do not differ for different travel purposes or ages, which indicates that the access and egress distances are mainly influenced by the level of service.

5.1.3 Longitudinal analysis

The access and egress distances are stable over the WROOV years. Both the trip-based definitions and the tour-based definitions only have slight variation over the years. Figure 14 shows the 75 percentile of the access and egress distances, since the mean value is to a larger extent influenced by erroneous data.

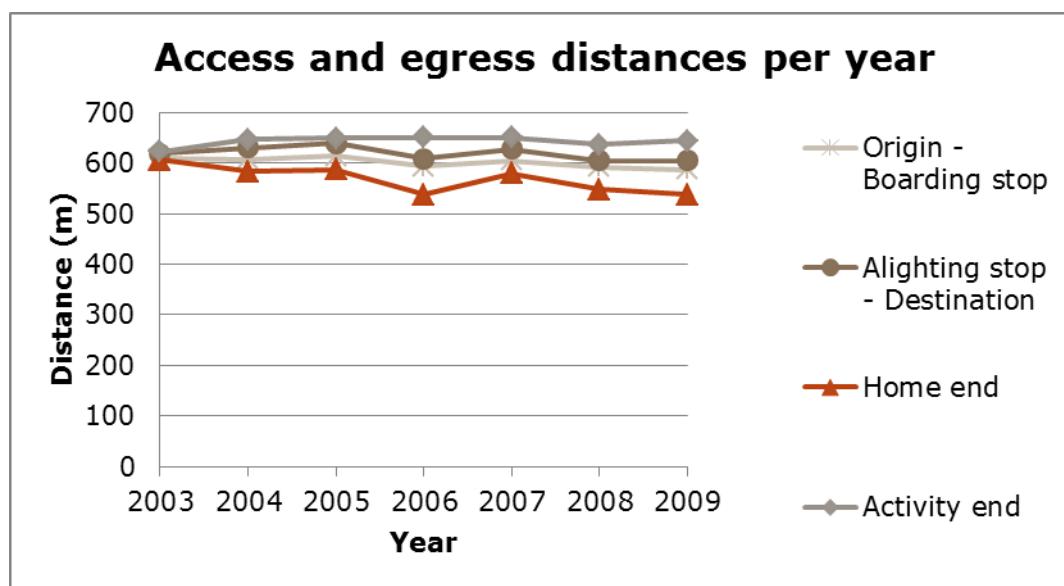


Figure 14: Longitudinal analysis of access and egress distances

5.2 Travel purpose

For the identification of relevant attributes for the purpose inference, a similar method has been pursued as with the access and egress distances. First, the relation with key-



variables has been investigated. Second, to place the explanatory value into context, unavailable attributes in OV-chipkaart data have been analysed. It has to be noted that the relative frequencies per purpose are based on shares per purpose, hence adding to 100% for each purpose. This does not relate to the absolute frequency of each purpose. Figure 15 presents the overall shares of the four identified purposes.

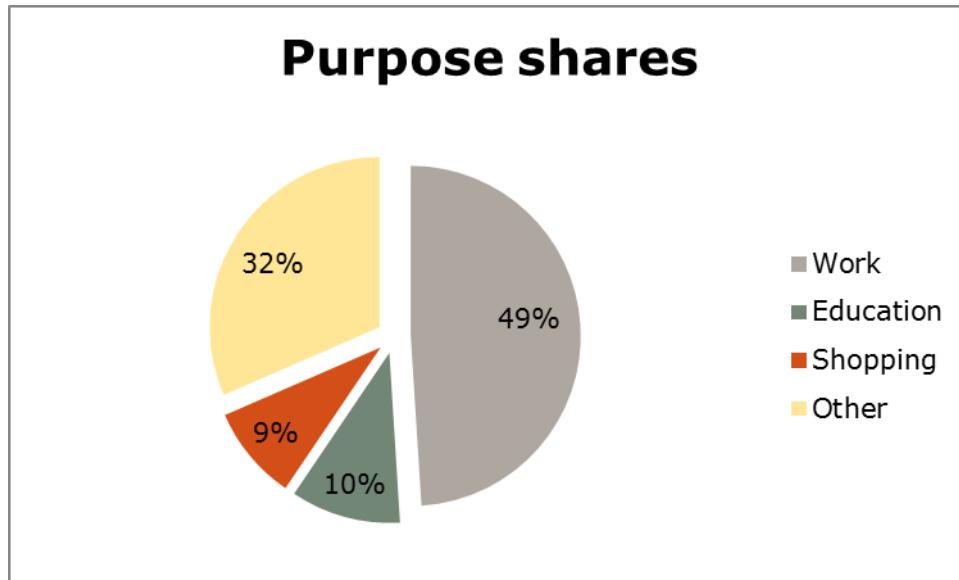


Figure 15: Overall purpose shares in the WROOV data

5.2.1 Key variables

The literature indicates a clear relation between the activity duration and the travel purpose. Therefore, it was applied as the main attribute in the rule-based processing approach for purpose inference by (Chakirov & Erath, 2012).

The distribution of activity durations per purpose in WROOV data confirms the indicated relevance of the travel purpose. Clear peaks are visible for the purposes work, between nine and ten hours, and education, between seven and eight hours¹¹. The purposes shopping and other show less sharp peaks in activity duration, but consist of mostly activities shorter than six hours.

¹¹ Note that the activity duration includes the travel time of the away trip due to the lacking alighting time in WROOV data.



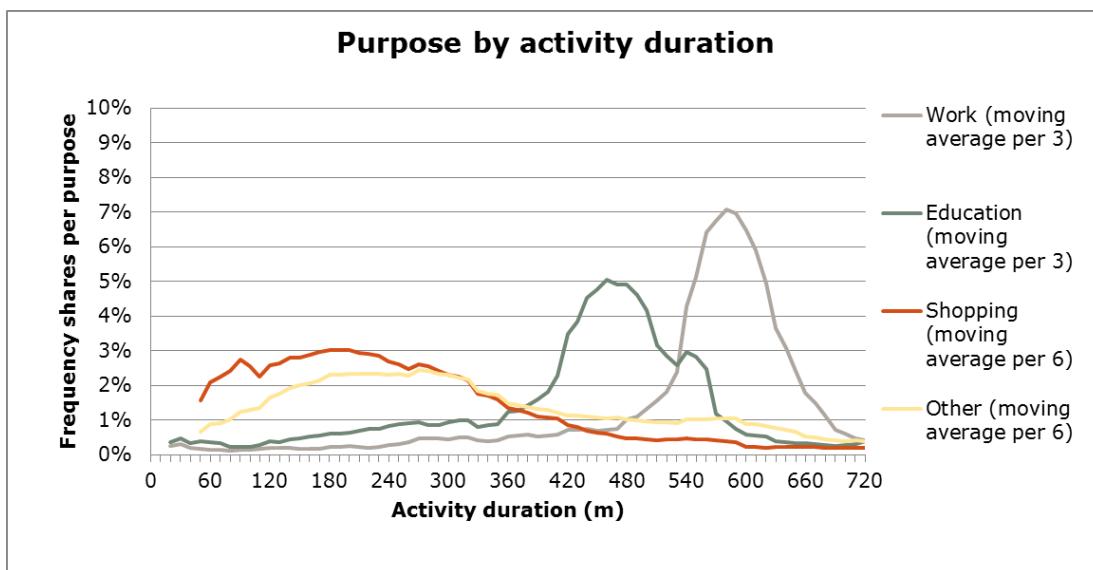


Figure 16: Activity duration distributions per purpose¹²

The second-most attribute related to the travel purpose in the literature is the departure time. The shares of departure time per purpose show similar patterns for the purposes work and education. Both purposes have strong peaks in the morning and afternoon, with the distinction that the afternoon peak of education trips is earlier than the afternoon peak of work trips. Furthermore, the departure time patterns of the purposes shopping and other are similar. These purposes mostly occur during the day, between the morning and afternoon peaks. A distinction is that the purpose other is relatively frequent in the night.

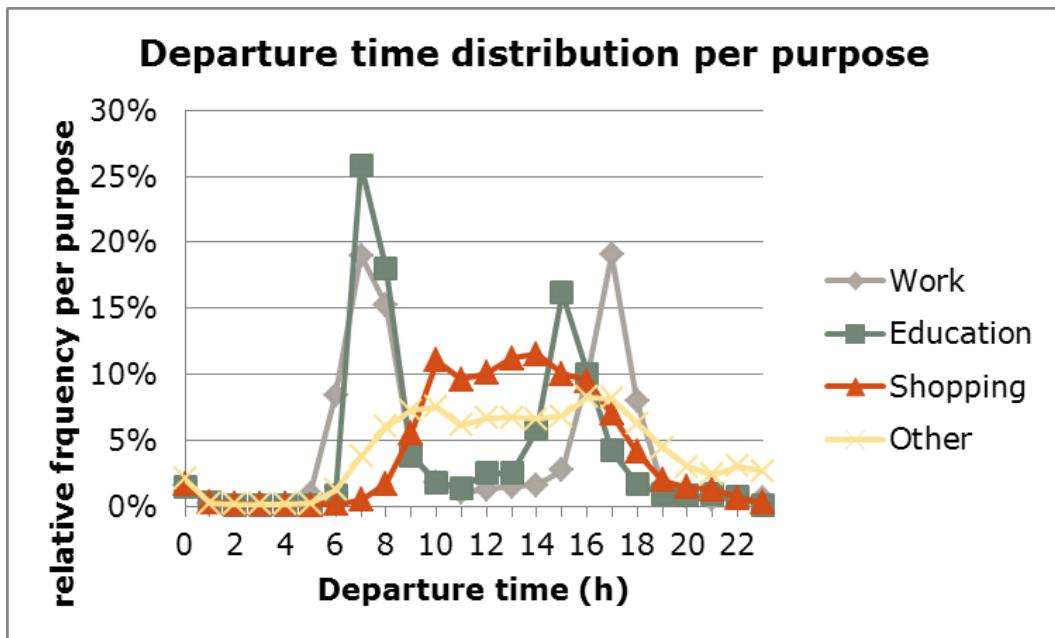


Figure 17: Departure time distributions per purpose

A third attribute indicated as relevant for the travel purpose in literature is the travel frequency. The travel frequency distribution shows a similar distinction between the

¹² Moving averages are applied in order to correct for rounding of times by respondents. The observed frequencies are obtainable in Appendix B.



compulsory travel purposes *work* and *education* on the one hand, and the discretionary purposes *shopping* and *other* on the other hand.

In addition to these three attributes frequently identified as possible explanatory variables for the travel purpose, the distributions of further key variables per purpose have been investigated. Many of these attributes show different distributions per purpose, and therefore are potential attributes for the purpose inference. Here, only the distribution over contracts is presented, since it shows the largest distinction between purposes. In Appendix B distributions are presented for the travel frequency, travel distance, number of legs per trip, operators used, and the product fares.

The distribution of contract types per purpose does show a distinction between the purposes *work* and *education*. Travellers with the purpose *work* mostly used year contracts, while travellers with the purpose *education* mostly used monthly contracts. On the other hand, the purposes *shopping* and *other* mostly travel without a contract. The distribution of travel purposes per fare show a very similar distribution as the contract durations, indicating a high correlation between the attributes.

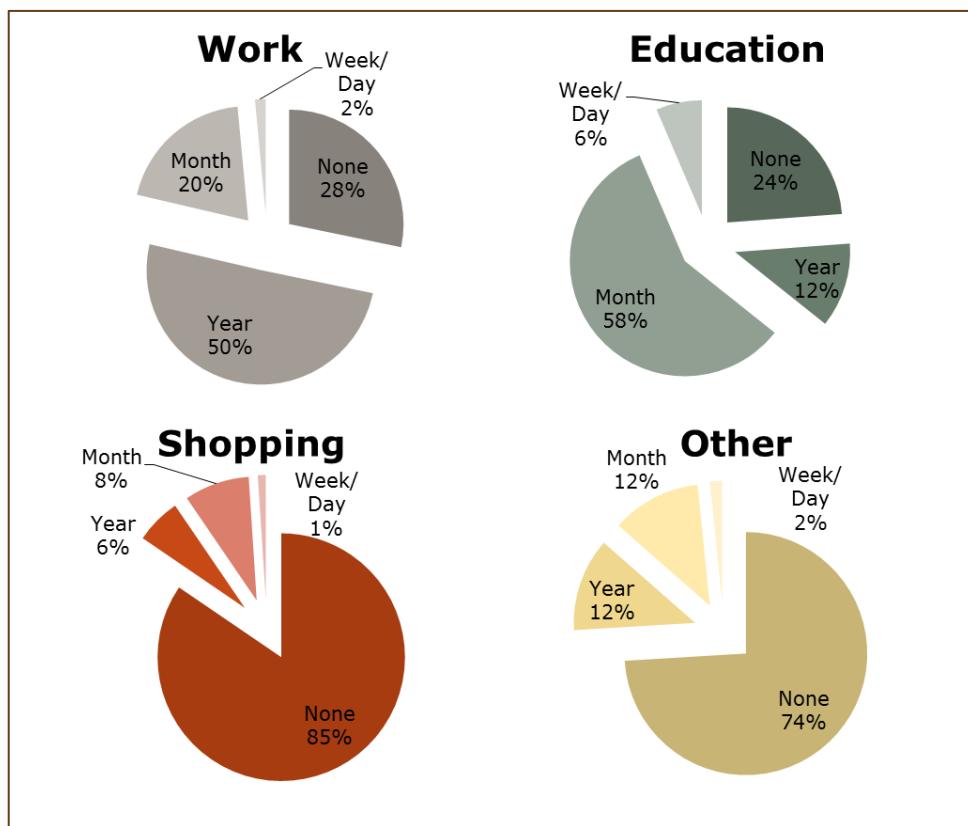


Figure 18: Distribution of contract types per purpose

The distributions of travel distances per purpose only show slight differences between the purposes. However, looking at the relative purpose shares over the travel distance, the share of the purpose *work* steadily increases with longer distances. The purpose *shopping* has a strong peak for short distances, but also longer trips are made for shopping. This indicates the difference of shopping for daily groceries and shopping as a recreational activity, which are both included in this category.



5.2.2 Characteristics unavailable in OV-chipkaart data

The OV-chipkaart data does not include information on the gender of the travellers, but in case of personal cards, it does contain the age of the traveller. This information has not been used in this study to reduce the possibility of privacy violations. Moreover, the age sample of personal cards is not considered as representative for the entire system. Nonetheless, the correlation between age and purpose has been investigated, since it is considered as a valuable predictor.

The distribution of age groups per purpose confirms the expectations. The purpose *work* is mostly applicable to adults, while *education* trips consist of mostly children. Seniors have low shares of trips in these compulsory purposes and mostly travel for *shopping* or *other* purposes. Hence, it can be concluded that the age would be a valuable estimator for the purpose inference.

5.2.3 Longitudinal analysis

The longitudinal analysis of the purpose distribution shows a slight increase for the share of the purpose work over time. It has to be noted that the WROOV data of 2009 do not cover the entire system due to the partial implementation of the OV-chipkaart. The e-purse travel was implemented in that year, replacing the *stripenkaart* tickets, while contracts were still paper tickets, included in the WROOV study. Furthermore, the year 2003 shows a slight distinction with the trend, which can be explained by initialization of the study.

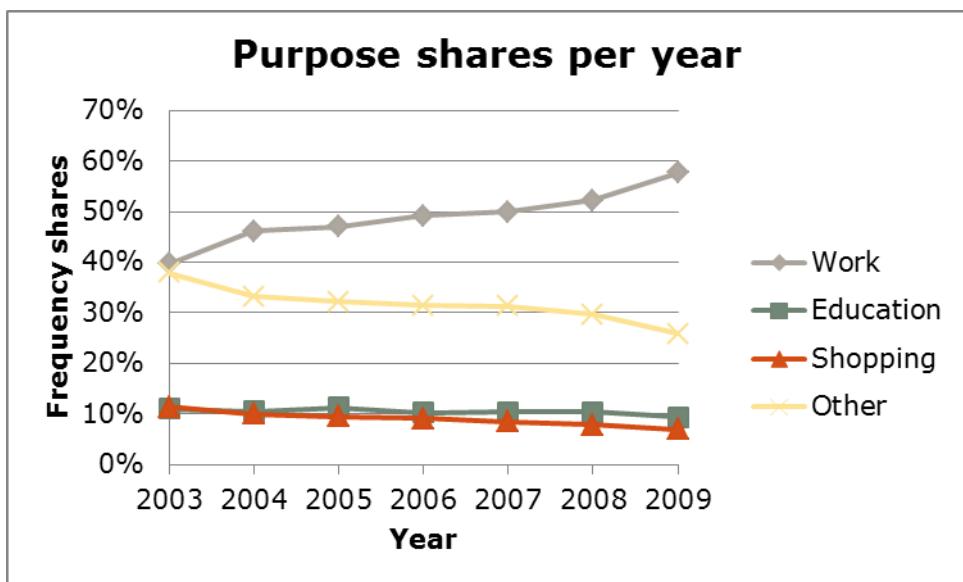


Figure 19: Longitudinal analysis of the purpose shares in the WROOV data

5.3 Concession traversing transfers

Since we do not have the data from all public transport operators in the study area available for this study, it is not possible to derive transfers to all operators from the data. These transfers do have an impact on the OD matrix construction, as the origins and destinations are outside the catchment area of the stop where the transfer is made. Since this problem is not part of the research focus, but a consequence of the unavailability of data for this study, we have aimed at a straightforward solution. When data from all operators would be available, this problem would cease to exist.



The WROOV data contain data for all BTM operators, but not for the national railways. However, respondents did indicate whether a trip by train was made before or after their travels with bus or light rail. Although this indication is not irrefutable, as it was not a specific question, it provides an overview of the share of travellers transferring to and from the train. These shares have been analysed per train station and are presented in Table 12.

Table 12: Shares of BTM trips transferring to or from the train network per train station

Train station	Trips		Tours		Non-home-based trips	
	Boarding	Alighting	Home-end	Activity-end	Boarding	Alighting
Amstel	31%	31%	50%	12%	32%	16%
Bijlmer	13%	14%	33%	8%	14%	7%
Centraal	40%	39%	61%	15%	41%	17%
Diemen	23%	34%	10%	60%	0%	20%
Diemen Zuid	14%	15%	17%	11%	14%	9%
Duivendrecht	75%	74%	85%	39%	69%	30%
Holendrecht	9%	10%	8%	9%	11%	12%
Lelylaan	23%	23%	26%	19%	26%	11%
Muiderpoot	16%	17%	20%	10%	14%	11%
Purmerend	34%	33%	26%	36%	25%	20%
Purmerend Overwhere	11%	11%	5%	33%	4%	6%
RAI	19%	19%	31%	10%	29%	12%
Schiphol	33%	32%	65%	17%	24%	13%
Sloterdijk	53%	52%	76%	17%	51%	21%
Weesp	45%	46%	51%	38%	41%	15%
Wormerveer	40%	37%	35%	42%	20%	21%
Zuid	32%	31%	59%	12%	40%	17%

The shares of transfers at train stations relate to the shares of trips that use BTM as access transport, in case of a transfer at the alighting stop, or egress transport, in case of a transfer at the boarding stop. The fact that the metropole area of Amsterdam has fourteen train stations indicates that the train system also has a regional function. The transfer shares also prove that the train system and the BTM systems are closely related. The shares of trips used for access or egress transport are very high for the stations Duivendrecht, Sloterdijk and Centraal. These stations are very well connected to the GVB network, which explains these large shares.

The transfers shares deviate substantially between stations and, moreover, between the home-end and the activity-end. In general, train stations in Amsterdam indicate higher transfer shares at the home-end compared to the activity-end. Train stations



outside Amsterdam show a different picture, with in some cases even an opposite relation.

5.4 Quantitative comparison of key variables

After determining the relation of specific attributes to the travel purpose, their appropriateness as medium of the information transfer has been assessed by means of a quantitative comparison of the key variables of both sources. In order to create a comparable dataset from OV-chipkaart data, however, rule-based processing had to be applied for the identification of trips. Therefore, we discuss these processing rules first and the applied OV-chipkaart dataset first, before we present the results of the quantitative comparison.

5.4.1 Construction of trips and tours from OV-chipkaart data

In order to compare the WROOV data with OV-chipkaart data, the raw OV-chipkaart data had to be interpreted. This included the application of processing rules to distinct transfers and activities between consecutive trip legs, which influenced the key variables. Figure 20 contains the flow diagram with the applied processing rules. The complete overview of the data handling procedures is included in Appendix A.

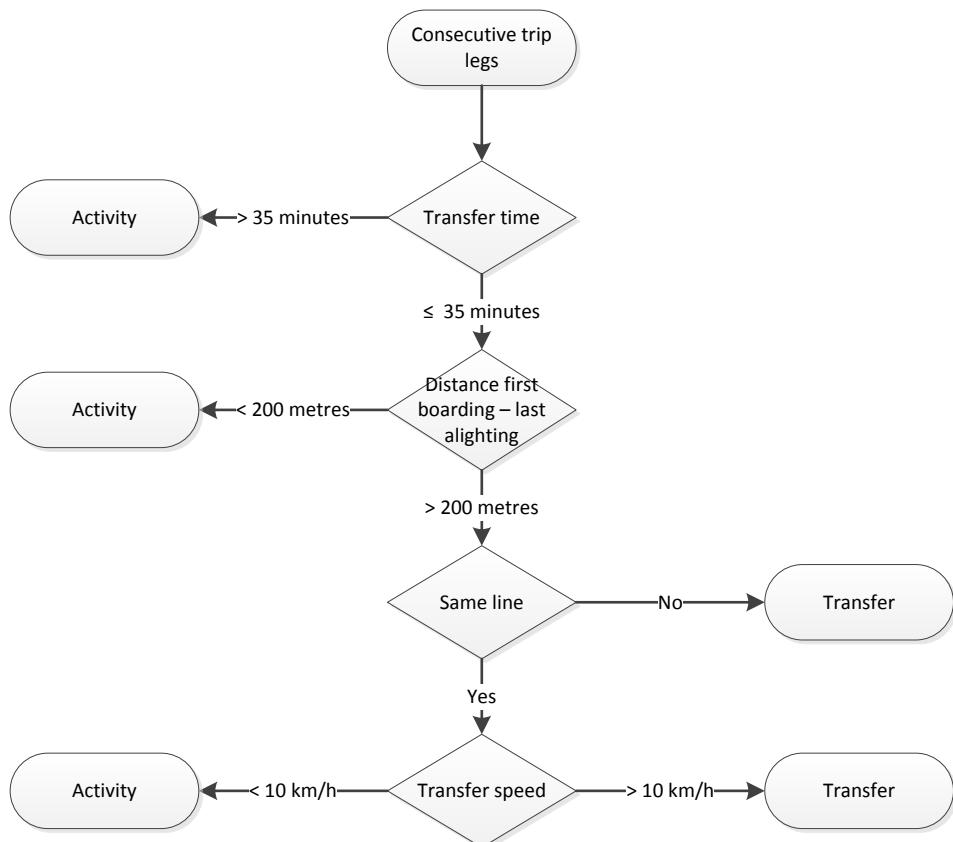


Figure 20: Flow diagram of distinction between transfers and activities

The distinction between transfers and activities is based on constraints in the three dimensions indicated in the literature (see paragraph 2.2.2): time, space and structure. The transfer time is the primary indicator and includes the constraint of 35 minutes commonly applied in Dutch public transport (Nijenstein & Bussink, 2014). This time constraint is brought about by the assumption that transfers options are always available within this time window, hence it is also related to the public



transport structure. The second indicator is related to the identification of tours. When two consecutive trips consist of an away trip and a return trip, thus when a tour is made, an activity is inferred, regardless to the transfer time. Therefore, the spatial constraint of 200 metres between the first boarding stop and the last alighting stop has been applied as indicator of a return trip. The third indicator is a structural constraint: an activity is inferred if a transfer is made on the same line. Devillaine and colleagues (2012) argue that a transfer on the same line, in both directions, implies an activity. However, the analysis of transfers on the same line in OV-chipkaart indicated that traveller's check-in and check-out in the same vehicle, while driving. In order to correct for this fare-dodging behaviour, we implemented an additional constraint of the transfer speed for transfers on the same line. The basic application of only the time-constraint of 35 minutes results in an overestimation of transfers by 22% compared to the enhanced distinction between transfers and activities.

5.4.2 Longitudinal analysis of OV-chipkaart data

The second step toward a quantitative comparison between WROOV and OV-chipkaart data, is the selection of a data collection period. In strategic transport modelling, common practice is to use an average working day as capacity of the travel demand. This average working day does not actually exist, as it is an average, but it does provide a more stable measurement, which contemplates with the requirements of long-term forecasts. The availability of a full year disaggregated travel data has provided the unique opportunity to place an average working day in context. Moreover, this allows for a more deliberate "construction" of the average working day.

The total number of trips per week clearly depicts the holiday periods and, less obviously, the seasonal variance. Excluding the weeks with holidays, the average travel demand of work weeks looks constant, with slightly decreased demand in the spring and slightly increased demand in winter. From this, we conclude that the construction of an average working day does not need to be based on a long period but can be based on a single week, which drastically decreases the computation times.

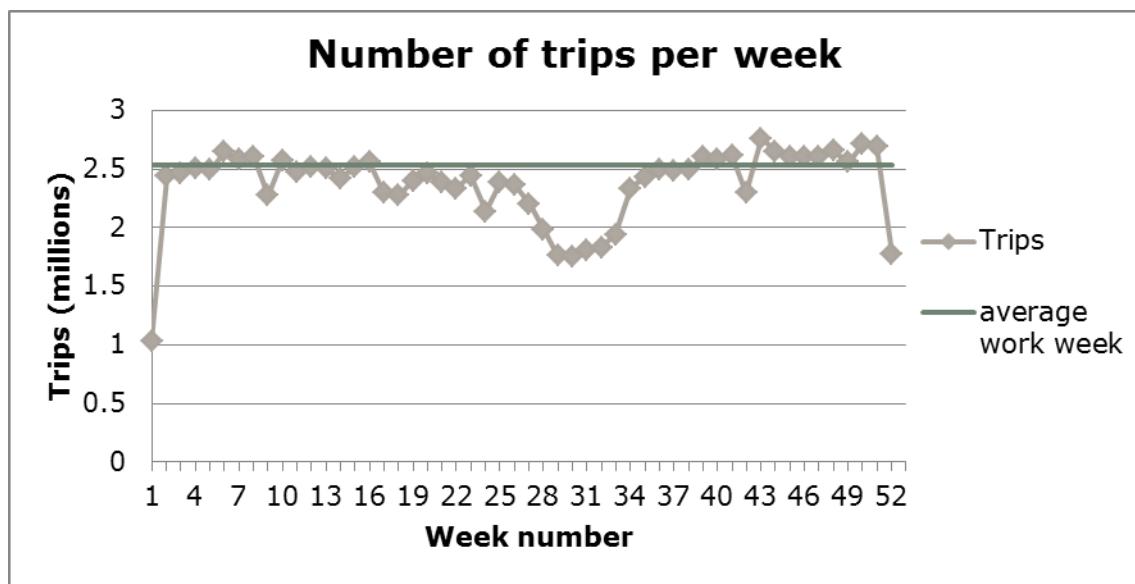


Figure 21: Trips per week in the OV-chipkaart data

Analysing the stability of key variables in OV-chipkaart data, we have found that these are rather stable as well, again with the exception of holidays. This indicates that



these variables are appropriate for enriching the OV-chipkaart data in context of an average working day.

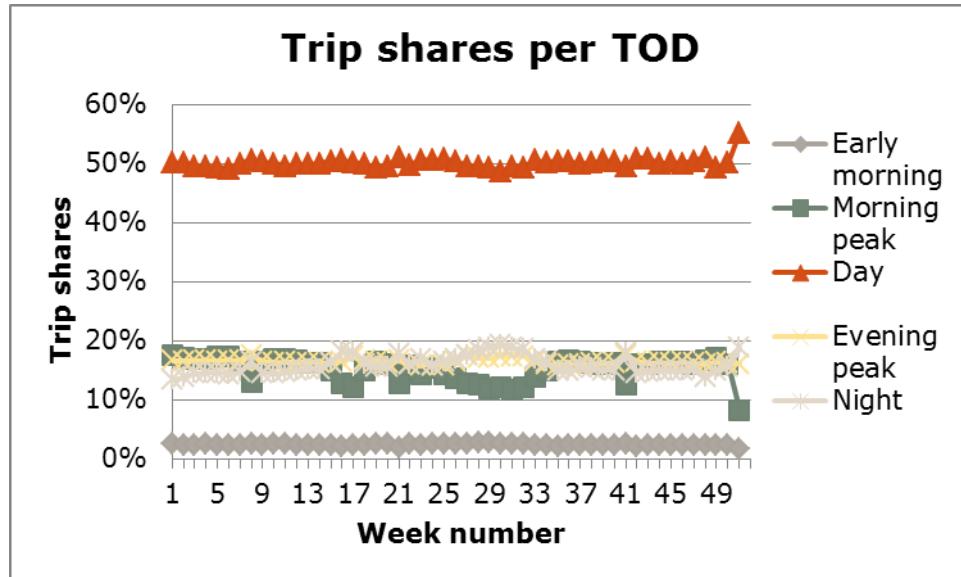


Figure 22: Trip shares per time of day over the year

The overall statistics are fairly constant over the year, with exceptions of holiday periods and national days. For computational reasons, a selection of 1 week has been made to construct the average work day.

Week 49 is closest to the average work week; hence this week has been used in further analyses and for the construction of the OD matrix for an average working day. This one week of crude OV-chipkaart data has been processed, resulting in a trip dataset, representing the average working week of 2014. An overview of the data processing procedures can be found in Appendix A.

5.4.3 Characteristics related to access and egress trip legs

Based on the selected week, we have compared the description of the travel demand by the OV-chipkaart with the description of travel demand by WROOV. Ideally, the distribution of key variables agrees between the sources. In that case, the key variables are appropriate for transferring the information between the sources. In order to match the target populations of the OV-chipkaart dataset and the WROOV dataset, students (20%) and short term contracts (9%) are filtered from the data, since these are not included in the WROOV survey.

The key variable most related to the access and egress distance is the mode. The comparison of both sources on the trip distribution over modes shows an overestimation of the trips made with multiple modes by the WROOV data. The modal shares of bus are very similar, so the overestimation of multiple modes is at the expense of tram and metro trips.



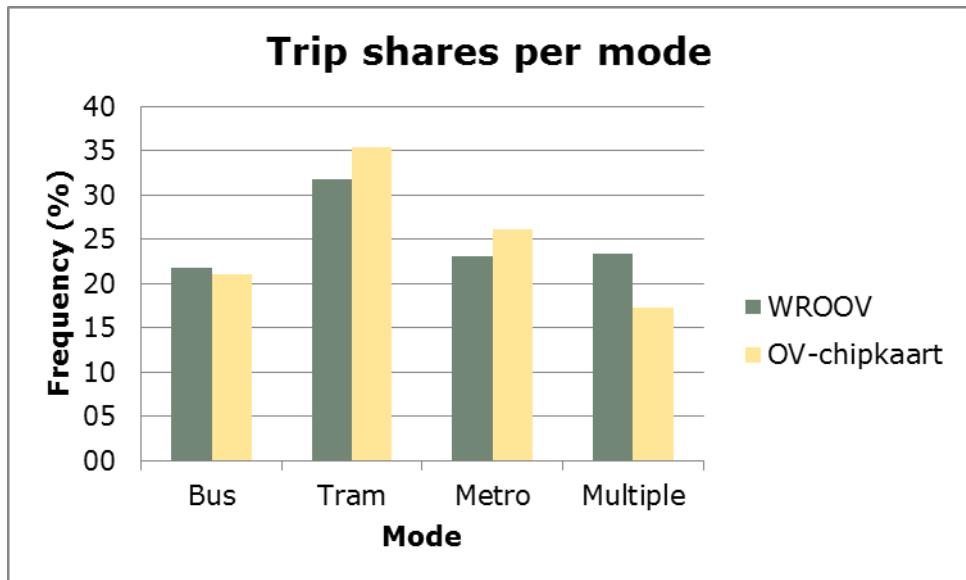


Figure 23: Comparison of trip shares per mode

5.4.4 Characteristics related to the travel purpose

The mode is also correlated to the travel purpose, showing a higher share of work trips for the modes metro and multiple modes. Consequently, WROOV overrepresents the purpose work.

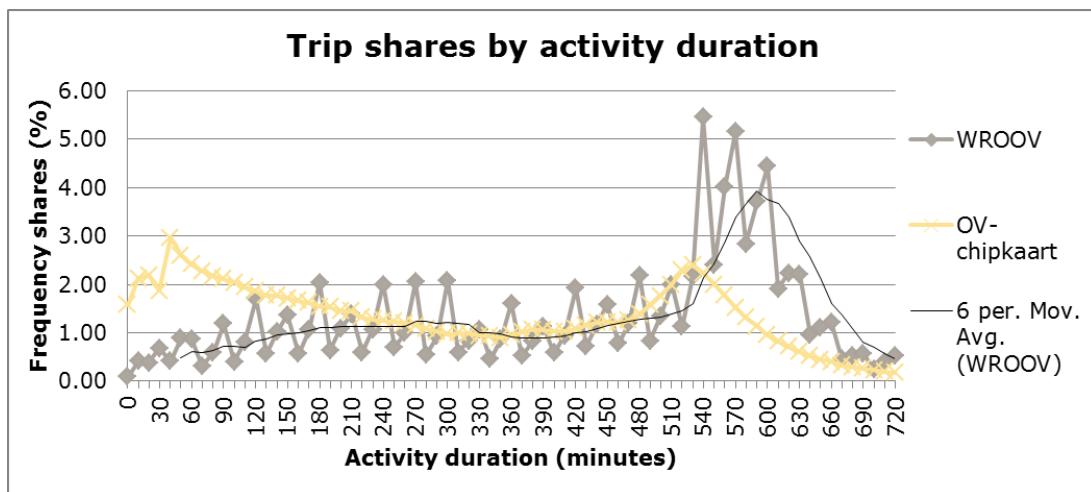


Figure 24: Comparison of trip shares over the activity duration

The comparison of activity durations in the sources indicates an overrepresentation of long activities in the WROOV data. Consequently, work trips are overrepresented. In addition, short activities are underrepresented in the WROOV data. This is a well-known problem of travel surveys due to respondents that forget to report short activities or think these are not important. The distribution of activity durations in OV-chipkaart data also shows a clear peak at the 35 minute interval. This indicates that the applied processing rules for the distinction between transfers and activities do not catch all activities. As a result, short activities are underrepresented by OV-chipkaart data as well. On the other hand, the applied processing rules also indicate activities close to zero minutes. This is not realistic and probably caused by errors in the processing rules and errors in the time recording.



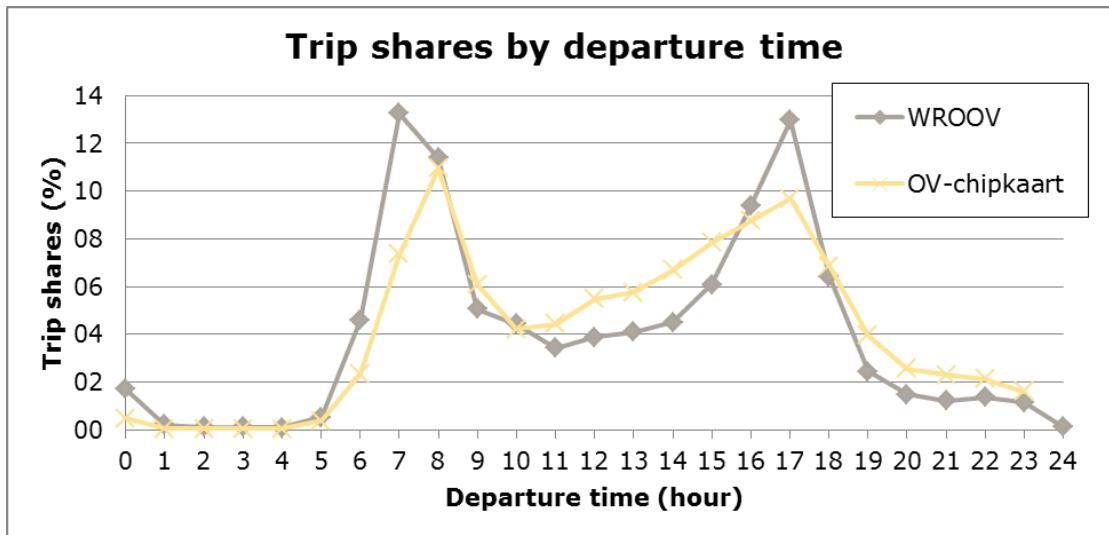


Figure 25: Comparison of trip shares over departure time

The distribution of trips over the day shows clear peaks in the morning and evening in both sources. However, the peaks are sharper in the WROOV data compared to the OV-chipkaart data. The OV-chipkaart shows a larger share of trips during the day, between the peaks, and also during the night. This indicates a larger share of discretionary purposes.

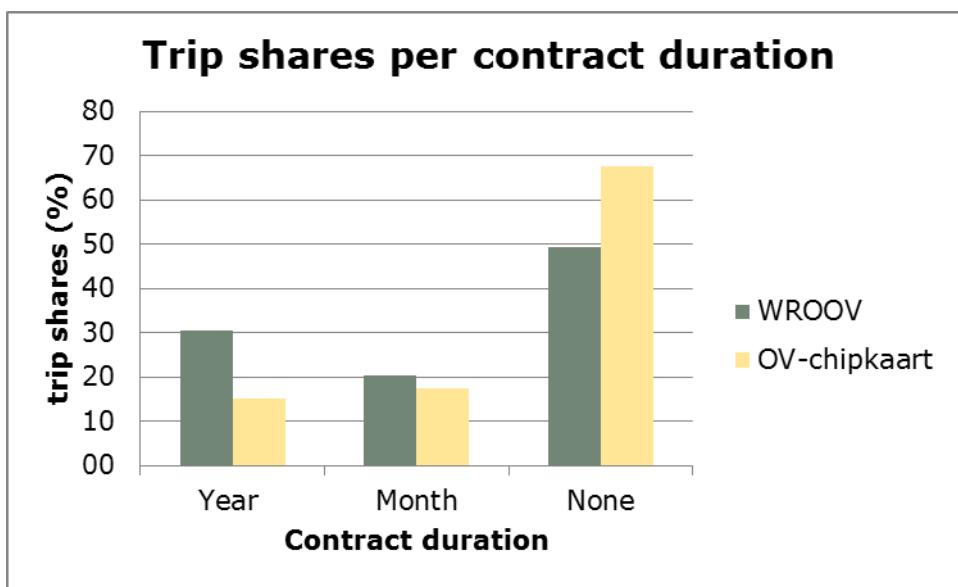


Figure 26: Comparison of trip shares per contract duration

The distribution of contract durations also dissimilarities between the sources, as the WROOV data over represent trips made with contracts, mostly year contracts. The OV-chipkaart has a substantially larger share of trips made without a contract. As a result, WROOV over represents work trips and, to lesser extent, educational trips.

The WROOV surveys also over represent trips with higher frequencies, showing many tours that are made four or five times a week, where the OV-chipkaart shows mostly tours made only once or twice a week.



The share of trips per concession in WROOV also shows a deviation with the OV-chipkaart data. WROOV slightly overrepresents the share of trips made in Waterland.

Overall, we can conclude that most of these imbalances are related to the same problem: the larger share of contracts in the WROOV data. Consequently, WROOV represents a larger share of work trips. The higher share of trips outside the peak periods, without contract and with shorter activities can also be caused by the insufficient filtering of international travellers from the OV-chipkaart dataset, as these can also travel with a regular OV-chipkaart with stored value.

5.5 Conclusions regarding the travel analysis

The travel analysis had two main goals: the identification of relevant attributes for the purpose inference model and the identification of attributes that can be applied in the differentiation of catchment areas for the zonal allocation models.

5.5.1 Predictive value of attributes

The literature relates the level of service to the access and egress distances to public transport stops. However, line frequencies and travel speeds are not available in the WROOV data. Therefore, the mode of transport has been indicated as the main key variable in order to distinguish between catchment areas of stops, since it attains different levels of service for each mode.

In addition, differences were found between the access and egress distances at the home-end and the activity-end. At the home-end, access and egress distances seem to be mainly related to the level of service. Conversely, on the activity end, the access and egress distances show higher correlation with the travel purpose. These differences are related to the geographical location of the used stops and therefore are not considered as a valid distinction between access and egress distances at specific public transport stops.

The variance in the access and egress distances cannot be fully explained by the available attributes in the data. Even between stops with the same modes, large differences are observed in the distribution of access and egress distances. In order to obtain more insight in the catchment areas of stops, it is advisable to include the number of lines serving a stop, their frequencies and their operational speeds.

Considering the attributes related to the travel purpose, several key variables are indicated as potential explanatory variables for the purpose inference. Besides the frequently mentioned attributes of activity duration, departure time and frequency, also the contract duration, the fare and the distances travelled are correlated to the travel purpose. The distribution over key variables shows a strong correlation between the purposes work and education, with similar patterns over the departure time and frequency. In addition, the purposes shopping and other show a correlation in the distribution over these attributes as well. Attributes that do indicate differences between compulsory and discretionary purposes are the activity duration and the contract duration. Another relevant, but unavailable attribute is the age of the traveller.

5.5.2 Appropriateness of key variables as medium of information transfer

All differences that have come to light in the quantitative comparison of both sources can be associated with the higher share of contracts in the WROOV data compared to OV-chipkaart data. This dissimilarity can probably be explained by the coverage of



WROOV survey, which did not include regional tickets or international respondents. Since the popularity of Amsterdam as tourist attraction has only increased in recent years, together with the indirect filtering of tourists based on short term contracts, this leads to difference between the samples of the two sources.

Travellers with contracts tend to travel more for compulsory purposes *work* and *education*, whereas travellers without contract mainly travel for the discretionary purposes *shopping* and *other*. Hence, estimation of the influence of key variables based on WROOV data and application of these estimates on OV-chipkaart data might result in biased estimates due to differences between the data sets. This problem can only be solved by estimating the influence of model attributes based on a representative sample of the current travellers. This is not a feasible solution within this research due to budget constraints, but might be required in future years for the durability of this method.



6 Estimation of enrichment models

This chapter describes the process and presents the results of the model estimation for the three allocation subjects. First, the construction of reference matrices is described (paragraph 6.1), constructed by rule-based processing. Then, the zonal allocation models are described (paragraph 6.2), including the model structure, choice set generation, available attributes, model enhancement and the final parameter estimates. The allocation models for both trip-ends are closely related and therefore presented collectively. Subsequently, the same set-up is used for the purpose inference models (paragraph 6.3), where the model estimation for the purpose inference is principally different from the zonal allocation. Next, the identification of concession-traversing transfers is discussed (paragraph 6.4). Finally, the conclusions from the model estimations are listed (paragraph 6.5).

6.1 Rule based reference models

The goal of constructing reference matrices is to investigate the effect of more complex models and evaluate if the extra effort is worthwhile. Since there is no ground truth to compare the OV-chipkaart data with, the reference matrices provide an alternative assessment option. The reference models are based on rule-based processing and therefore easy to apply. The literature provides several reference studies that apply similar procedures which attain reasonable results (see Chapter 2).

6.1.1 Zonal allocation

Origins and destinations can be directly allocated to trips based on the zones the used stops are situated in. That is, the origin of a trip is the zone where the first boarding stop is situated in, and the destination of the trip is the zone where the last alighting stop is situated in. This is done by matching the used stops in the trips dataset to their coordinates and successively allocating origins and destinations to, respectively, the first and last used stops in the GIS software of MapInfo Professional.

The travel analysis on WROOV data shows approximately half of the origins and destinations are situated in the same zone as the used stop, hence substantial discrepancies occur when using individual allocations to zones based on the zone where the used stop is situated. The cause of this large discrepancy becomes clear when the stop locations are depicted in the VENOM zonal grid (Figure 27). Many stops are located on arterial roads of the city. These roads frequently form the borders of the zonal grid.

Currently, rule-based processing is commonly applied in applications of smart card data. Therefore, we applied this simple method as reference for the improvement of an probabilistic approach.



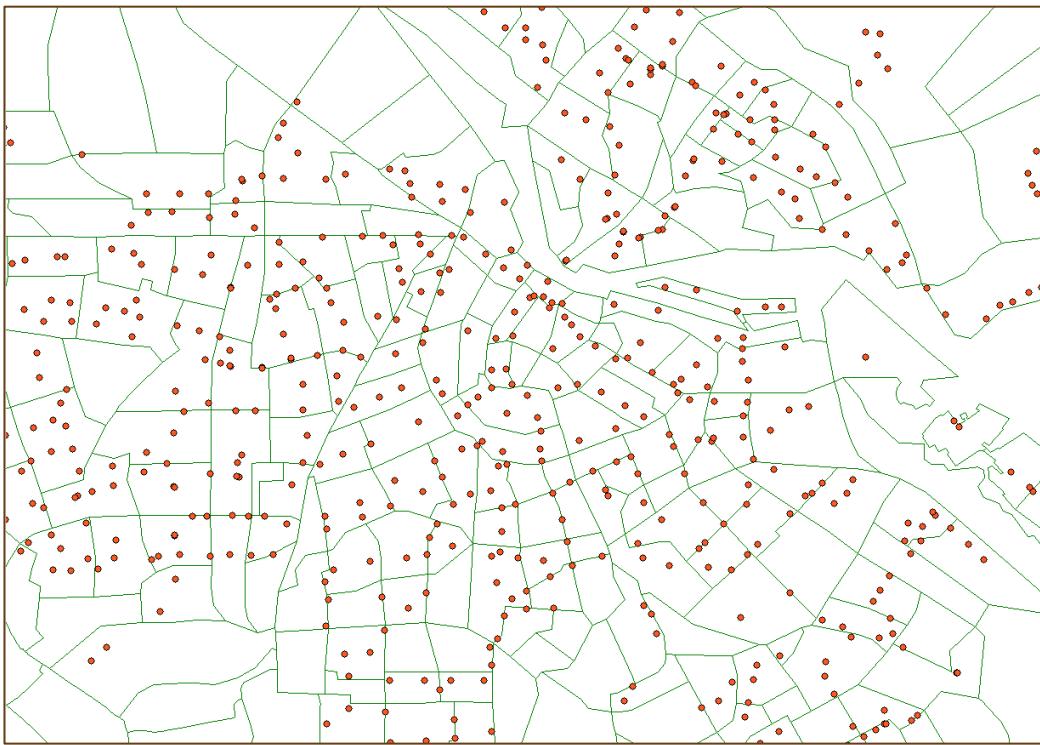


Figure 27: Stop locations in the VENOM zonal grid of Amsterdam

6.1.2 Purpose inference

The reference model for purpose inference is based on Rule-based processing, since the aim is to construct a simple model for comparison. Inspired by the study of Chakirov and Erath, we analysed the travel purpose distribution over the activity duration. This resulted in the classification of purposes based on three time intervals:

- The purpose *work* is inferred for tours with activities longer than 8 hours;
- The purpose *education* is inferred for tours with activities between 6 and 8 hours;
- The purpose *other* is inferred for tours with activities shorter than 6 hours.

The purpose *shopping* is not distinguishable from *other* by the activity duration. Accordingly, Chakirov and Erath did not incorporate this purpose. In order to construct comparable matrices, we allocated one fourth of the trips with the purpose *other* to the purpose *shopping*. This share is based on the overall distribution of purposes and is distributed randomly.

Non-home-based trips, which lack an activity duration, are allocated to the purpose *other*, since this purpose makes up the largest part of non-home-based trips. Travel behaviour analysis shows these rules are crude simplifications. However, the overall shares of purposes seem to match reasonably well.

6.2 Probabilistic zonal allocation models

The zonal allocation is performed by means of a multinomial logit model, estimated based on WROOV data with the Biogeme software (Bierlaire, 2003). The theoretical framework of multinomial logit modes can be found in paragraph 3.8.1. The procedure allocates trip ends, which consist of the used stops at either end of the trip, to the traffic analysis zones, of which the OD matrices are composed. The model estimates



the chance of each available zone to be "chosen", based on characteristics of that zone. In this case, the zones situated near the used stop comprise the available alternatives, also referred to as the choice set. As discussed in paragraph 3.8.2, this problem does not actually consist of a choice problem, but an allocation problem. However, throughout this chapter the terminology of discrete choice modelling is used.

Two different approaches concerning the classification of trip ends have been applied during the estimation process: trip-based and tour-based. In the trip-based approach, trip ends are considered separately for every trip. That is, for every trip the origin zone is allocated to the first boarding stop and the destination zone is allocated to the last alighting stop. The end-result consists of a trip from the origin zone to the destination zone. In the tour-based approach, on the other hand, trips are considered as linked within tours (see paragraph 1.4 for the definitions). In this context, the trip ends are classified as home-end and activity-end. The first trip of the day is assumed to start at home, hence the home zone is allocated to the first boarding stop, and the activity zone is allocated to the last alighting stop. Then, the return trip is allocated to the same activity zone and home zone, which creates consistency between trips in tours.

6.2.1 Alternative generation

The choice set generation embodies the identification of relevant alternatives, in this case zones at trip-ends. We assume that origins and destinations are in the vicinity of the used stops. Hence, the identification of alternative zones is based on the catchment areas of stops. There are, however, no definite boundaries of catchment areas. From the travel analysis on characteristics related to access and egress trip legs (see paragraph 5.1), we concluded that substantial differences exist between the sizes of catchment areas of specific stops. Although we have shown that the distances vary significantly between modes, and not between trip end classifications, we were not able to allocate the full variation of access and egress distances to specific attributes that are available in the WROOV data. Therefore, different approaches in the estimation of catchment areas have been investigated. In order to keep the procedure of choice set generation by means of catchment areas manageable, we limited the shape of catchment areas to spherical. The approaches vary in the radius of the catchment areas:

1. Uniform catchment areas: an equal radius of 400 metres for all stops, with the constraint that zones contain at least 1% of catchment area. This distance is based on concession requirement of a stop within 400 metres for all residents in Amsterdam;
 2. Mode-specific catchment areas: a radius of 750 metres for metro stations, 600 metres for tram stops and 500 metres for bus stops. These distances are based on the 75 percentile of the distance covered during access and egress trip legs per mode. For simplicity reasons, specific home-end and activity-end catchment areas have not been applied, as these would require substantial added handling times;
 3. Stop-specific catchment areas: Individual radius per stop, based on 90 percentile distances covered during access and egress trip legs. The 90 percentile is only applied to stops with more than ten observations, with a minimum of 500 metres and a maximum of 2500 metres. A radius of 500 metres is applied to stops with less than ten observations.
- 2.



The zones that overlap with the catchment areas are identified as alternative. The number of alternatives, then, varies, depending on the size of the catchment areas and to the location of the stop relative to zonal borders. For example, a stop at the centre of a large zone might just yield one alternative and a stop with a large catchment area in a zonal grid with small zones might yield many alternatives.

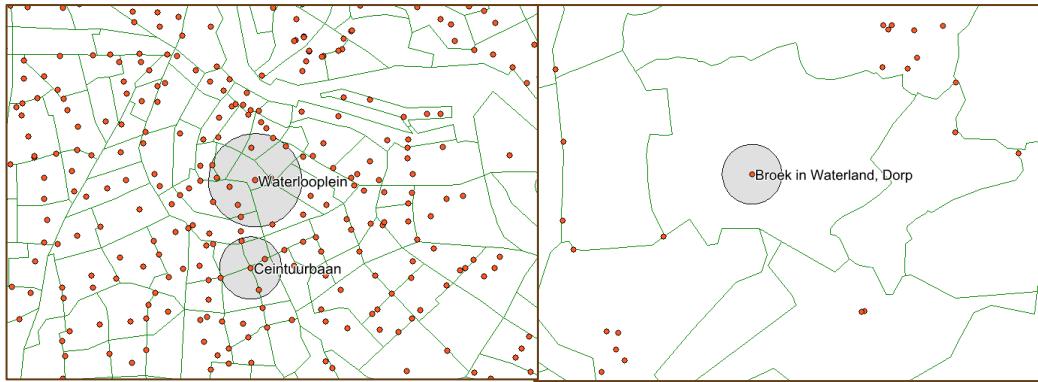


Figure 28: Catchment areas of stops

Maximum number of available alternatives determines the number of zones to be defined in the model specification. This number depends on the radius of catchment areas. Table 3 provides the maximum number of alternatives for the three different approaches of choice set generation.

Table 13: Characteristics of different choice set generation amplifications

<i>Catchment areas</i>	<i>maximum radius (m)</i>	<i>maximum number of zones in catchment area</i>
Uniform	400 (all)	12
Mode-specific	750 (metro)	17
Stop-specific	2500 (upper bound)	54

As indicated in the previous paragraph, trips with origins and destination outside the catchment area of the used stop are filtered from the dataset. This means that the different classifications of trip ends result in different sizes of the final datasets used for the model estimation. Table 14 provides the shares of trip-ends within the catchment areas of the used stop for the three approaches of catchment areas and both the trip-based and the tour-based classification of trip-ends.

Table 14: Shares of trip-ends within catchment areas

<i>Catchment areas</i>	<i>Trip-based approach</i>			<i>Tour-based approach</i>			
	All rips		Tours		Non-home-based trips		
	Origin	Destination	Home	Activity	Origin	Destination	
Complete dataset	204 041 (100%)	204 041 (100%)	91 662 (100%)	91 662 (100%)	20 709 (100%)	20 709 (100%)	
Uniform	148 835 (72.9%)	148 655 (72.9%)	63 294 (69.1%)	70 091 (76.5%)	15 451 (74.6%)	15 274 (73.8%)	
Mode-specific	159 634	159 516	67 715	75 468	16 450	16 338	



	(78.2%)	(78.2%)	(73.9%)	(82.3%)	(79.4%)	(78.9%)
Stop-specific	169 535 (83.1%)	169 955 (83.3%)	71 059 (77.5%)	77 364 (84.3%)	17 369 (83.9%)	17 591 (84.9%)

The WROOV dataset contains only trip legs within the structural boundaries of the concessions Amsterdam and Waterland, analogous to the OV-chipkaart dataset available for this study. Hence, the fact that origins and destinations are not within the catchment areas of the used stops could indicate a transfer to another system. At the tour-based level, the difference between home-end and activity-end, which does not exist at the trip-based level, implies that there is a smaller share of transfers to other systems at the activity-end compared to the home-end. This finding suggests that Amsterdam generates more inbound activities than outbound activities. Considering the high attraction of the capital, this seems a consistent finding. However, the share of trip-ends within the catchment area also increases with larger radii of catchment areas. This implies that a small share of travellers covers a larger distance during access or egress transport.

6.2.2 Available attributes and expectations

The available attributes for the estimation of the trip-end zone consist of zonal characteristics. Most of these characteristics are obtained from the VENOM data, which have the same format as the zonal data from the LMS/NRM models. Several attributes have been created from VENOM data, WROOV data or both, or were constructed with the MapInfo GIS software (see Table 15):

- The attribute *share of catchment area* [%] consists of the areal share of the catchment area that overlaps with the respective zone;
- The attribute *distance to centroid* [m] encompasses the Euclidean distance from the used stop to the respective zonal centroid, where the zonal centroid represents the gravitational centre of the zone;
- The attribute *schools* [student places] consist of the sum of student places at secondary schools and vocational education. The student places of primary school are not considered relevant, as these students generally do not travel with PT. Moreover, the student places at higher education are not considered because these student have student PT cards (SOV) which are not included in the target population of the WROOV surveys;
- The *household size* [residents/household] is the average number of residents per household;
- The *level of urbanization* [addresses/ha] is the number of households situated in a zone divided by the zonal area;
- The attribute *cars* [cars/household] consists of the average number of cars available per household;
- The attribute *stop density* [stops/ha] is the number of public transport stops situated in a zone divided by the zonal area;

The available attributes for the estimation of the origin zone are categorized into (1) attributes regarding the geographic information of the zones, (2) attributes regarding the build environment in zones and (3) attributes regarding the population within zones. These categories are applied in the model optimization strategy, which is presented in the next paragraph.

Table 15: Available attributes for zonal allocation and their expected effect on utility



Category	Attribute	Source	Origin	Destination	Home	Activity
Geographical	Share of catchment area	Created with VENOM	++	++	++	++
	Distance to centroid	Created with VENOM	--	--	--	--
	Area	VENOM	+	+	+	+
Activity end	PC6	postal codes	+	+	+	+
	Households	VENOM	+	+	++	+
	Urbanization	Created with VENOM	++	++	++	++
	Stops	WROOV	+	+	0	0
	Stop density	Created with stops and VENOM	+	+	+	++
	Jobs	VENOM	+	+	0	++
Home end	Schools	Created with VENOM	+	+	0	+
	Residents	VENOM	+	+	++	+
	Household size	Created with VENOM	0	0	+	-
	Working residents	VENOM	+	+	++	0
	Students	VENOM	+	+	+	0
	Cars	Created with VENOM	-	-	-	0
	Average income	VENOM	-	-	-	0

The expectations of the attributes are based on factors that influence the production and attraction in trip generation modelling. In the trip-based approach, production and attraction cannot be considered separately because the dataset consists of both away trips and return trips. The tour-based approach does provide the opportunity to distinguish differences between factors related to trip production and factors related to trip attraction, which is displayed in Table 15.

Evidently, some of the attributes are highly correlated, which should be avoided in the estimation of logit models. Highly correlated attributes in the model impair the interpretation of the influence that specific attributes have on the utility. Where other attributes were initially removed, these attributes have been investigated to compare their influence. This holds for the number of postal codes, the number of households and the number of residents. These three factors indicate the amount of potential travellers at different levels of resolution: at the level of housing blocks, houses and individuals. In addition, the geographical attributes *share of catchment area* and *distance to centroid* are also correlated: the share of the catchment area will increase as the distance to the zonal centroid decreases. It was expected that at high resolution, the share of catchment area would have more explanatory value compared to the distance to the zonal centroid since the actual trip-end might be close to the zonal border. With a wider scope, however, the explanatory value of the distance to the centroid was expected to increase, since the intra-zonal distances become



relatively smaller. Larger catchment areas might contain entire zones, so the share of catchment area is determined by the zonal area.

The tables of correlations in zonal data and trip data can be found in Appendix C.

6.2.3 Model enhancement

The model enhancement strategy involved the selection of attributes based on the interpretation of the following values:

- Significance
- Parameter value
- Rho squared statistic
- Correlation of attributes

The evaluation of individual parameter values is based on t-tests, which assess if the parameter value is significantly different from zero at a certain confidence level and a number of degrees of freedom. The degrees of freedom are determined by the sample size, which is large in this case. Therefore, the t-test values are assessed at infinite degrees of freedom. At a 95% confidence level, a t-value larger than 1.96 indicates a significant difference between the parameter value and zero. At the 99% confidence interval, the t-value should be larger than 2.575 to indicate a significant difference and at a 99.9% confidence interval this is 3.291. Because the available dataset is relatively large, t-test values are relatively high, which may cause the significance of parameter values to be overestimated. Therefore, additional assessment criteria are required to evaluate the implementation of specific attributes in the model.

The sign of the parameter value indicates if the attribute has a positive or negative effect on the utility, and thus on the probability an alternative is chosen. Parameter values with an influence that cannot be theoretically substantiated are considered undesirable, as these parameters might result in overestimation of the model on the dataset. With the size of the available data set, this is an apparent issue.

High correlations between attributes are undesired since including both attributes results in multicollinearity. The individual effects of both parameters cannot be interpreted when they partly describe the same effect. Logit models do allow little multicollinearity, but in case of two highly correlated attributes it is preferable to include only one in the model.

The goodness of fit of the model can be evaluated by means of the Rho-squared statistic, also referred to as the likelihood ratio index. This statistic specifies the increase in log-likelihood of the estimated model compared to the null-model, where the log-likelihood indicates the probability that the dataset is described by the estimated model. The value of the Rho-squared statistic lies on the interval from 0 to 1. A value of zero indicates no improvement compared to the null model, and a value of one indicates a perfect fit, which is practically impossible in logit modelling. The likelihood-ratio test provides the possibility to compare different model configurations and asses if one configuration is significantly better than the other configuration.

The formula of the Rho-squared statistic (Train, 2009):

$$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(0)} \quad 6.1$$

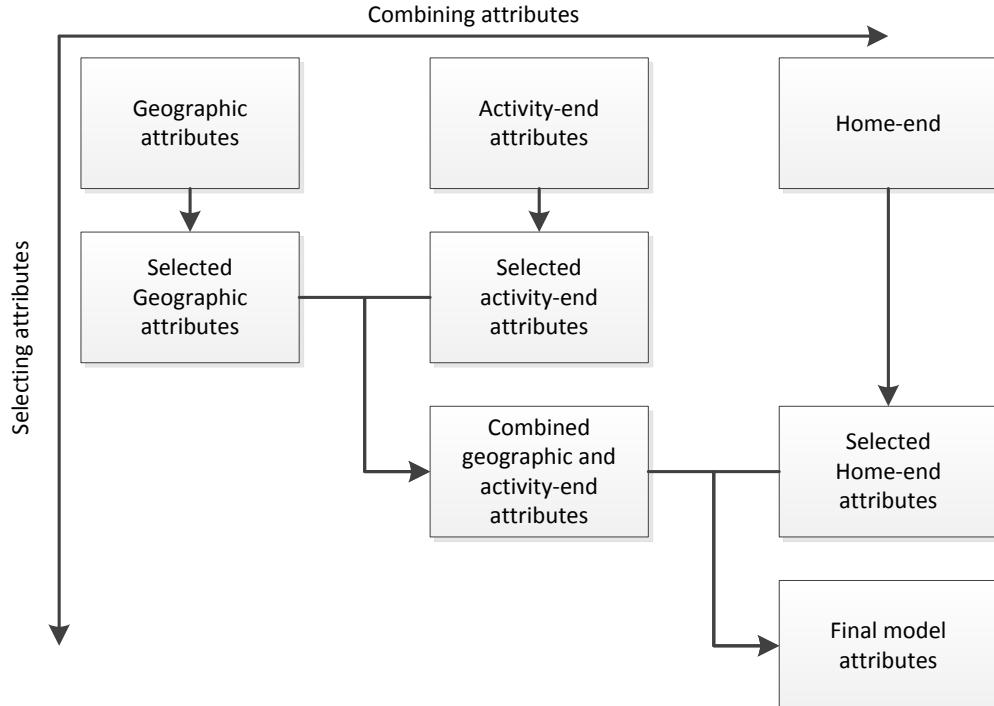


Where ρ^2 is the Rho-squared statistic, $LL(\hat{\beta})$ is the log-likelihood of the estimated model with parameters $\hat{\beta}$ and $LL(0)$ is the log-likelihood of the model with all parameters set to zero.

With this many available attributes, the number of possible model configurations is too large to assess all of them. In order to identify viable model configurations, a sequential strategy has been applied, with the following steps:

1. A model with only alternative specific constant parameters was estimated to test if the randomization resulted in generic alternatives, independent from the alternative number. All alternative specific constants tested insignificant, verifying that the utilities of alternatives are independent of the alternative numbering;
2. The different categories of attributes were assessed separately. This provided insight in the explanatory value of each attribute block and filtered irrelevant and highly correlated parameters. The optimization of each category block resulted in a selection of attributes with explanatory value, that are significant and have minimal correlation;
3. The remaining attributes of different category blocks are combined. Because the estimated parameters in logit models are all relative, combining attribute blocks results in additional correlations that influence the model stability. Therefore, after adding an addition block, another selection sequence has been applied.

3.



4.

Figure 29: Model enhancement strategy for zonal allocation

6.2.4 Convergence complications

During the model optimization process, we encountered complications regarding the model convergence. Depending on the applied algorithm, Biogeme then produces output with one of the following the diagnostics:

- Radius of the trust region is too small (BIO algorithm)
- No significant improvement possible (DONLP2 algorithm)



- Normal termination (SOLVOPT algorithm)

In case the model has not converged, the estimated parameter values cannot be interpreted as valid estimates. Therefore, the cause of these complications has been investigated.

"Radius of the trust region too small" can indicate that the model is close to singular which can be caused by inclusion of irrelevant parameters or totally correlated parameters. However, in the majority of model configurations the smallest singular value was found to be close to 200. This leads to believe that the converging issues were not caused by singularity of the matrix.

When confronted with convergence complications, Bierlaire recommends assessment of the final gradient norm, which should be close to zero. The final gradient norm differs between optimization algorithms, but the estimated parameters are equal. In some cases the DONLP2 algorithm diagnoses convergence, where the BIO algorithm does not. The SOLVOPT algorithm generally estimates a larger final gradient norm and converges less easily.

Convergence complications are mainly caused by correlations between attributes. The issues caused by correlated attributes can be solved by excluding one of the two correlated parameters. This exclusion is at the expense of the model fit. However, in several instances, a model with only the attribute distance to the centroid still resulted in an unstable model. Hence, correlations between attributes were not the only cause of model instability. We have not been able to discover other causes for instability, but these might be related to the high values of land-use characteristics in relation to the low variance between adjacent zones. However, this theory has not been investigated.

6.2.5 Final parameter estimates

The following tables present the final model statistics and parameter estimates for the trip-based model (Table 16), the tour-based model (Table 17) and the model for non-home-based trips (Table 18). Only the final model configurations are presented here, intermediate results from the estimation process are presented in Appendix C.

The final model configurations only include two or three attributes, since most available attributes were insignificant or caused instability of the model. Paragraph 6.2.7 deliberates on the interpretation of the individual model parameters.

Table 16: Final estimation results of trip-based zonal allocation models

Model statistics	Origin Inference		Destination Inference	
Model	Multinomial Logit		Multinomial Logit	
Number of parameters	3		3	
Rho-square	0.322		0.319	
Diagnostic	Convergence reached...		Convergence reached...	
Parameters	Value	t-test	Value	t-test
$\beta_{i,1}$ (share of catchment area)	0.0851	320.46	0.0851	320.82
$\beta_{i,2}$ (stop density)	2.97	170.15	2.94	170.22



$\beta_{i,3}$ (urbanization level)	0.00183	16.78	0.00143	13.14
------------------------------------	---------	-------	---------	-------

The estimated trip-based zonal allocation models are represented in formula by the following equations. Equation 6.3 presents the *origin zone allocation model* and equation 6.4 presents the *destination zone allocation model*. The model parameters and the model fit are approximately similar. This can be explained by the trip-based definition of trip-ends in combination with the high share of symmetry in the trip-based data set used for estimated. As a result, access legs in the away trip are made in opposite direction during the return trip as egress legs. The slight differences between origin and destination allocation can be attributed to the differences in non-home-based trips.

$$V_{O,i} = 8.51 * 10^{-2} * X_{i1} [\%] + 2.97 * X_{i2} [\text{stops/ha}] + 1.83 * 10^{-3} * X_{i3} [\text{addresses/ha}] \quad 6.2$$

$$V_{D,i} = 8.51 * 10^{-2} * X_{i1} [\%] + 2.94 * X_{i2} [\text{stops/ha}] + 1.43 * 10^{-3} * X_{i3} [\text{addresses/ha}] \quad 6.3$$

With:

$V_{O,i}$ = systematic utility of alternative i for origin zone O

$V_{D,i}$ = systematic utility of alternative i for destination zone D

X_{i1} = value of share of catchment area of alternative i

X_{i2} = value of stop density of alternative i

X_{i3} = value of urbanization level of alternative i

Table 17: Final estimation results of tour-based zonal allocation models

Model statistics	Home zone inference		Activity zone inference	
Model	Multinomial Logit			Multinomial Logit
Number of parameters	3			2
Rho-square	0.347			0.311
Diagnostic	Convergence reached...			Convergence reached...
Parameters	Value	t-test	Value	t-test
$\beta_{i,1}$ (share of catchment area)	0.0824	81.69	0.0901	92.56
$\beta_{i,2}$ (stop density)	2.81	36.78	2.86	51
$\beta_{i,3}$ (urbanization level)	0.00613	14.23	0	(fixed)

The estimated tour-based zonal allocation models are represented in formula by the following equations. Equation 6.5 presents the *home zone allocation model* and equation 6.6 presents the *activity zone allocation model*. In contrast to the trip-based models, the tour-based zonal allocation models differ substantially between trip-ends. At the activity end, the influence of the urbanization level did not prove to be significant and stable and therefore was omitted from the model. At the home end, the same attributes are included as in the trip-based models. The fit statistics indicate a better fit of the home zone allocation model compared to the activity zone allocation model, which can be explained by the additional attribute.

$$V_{H,i} = 8.24 * 10^{-2} * X_{i1} [\%] + 2.81 * X_{i2} [\text{stops/ha}] + 6.13 * 10^{-3} * X_{i3} [\text{addresses/ha}] \quad 6.4$$



$$V_{A,i} = 9.01 * 10^{-2} * X_{i1} [\%] + 2.86 * X_{i2} [stops/ha] \quad 6.5$$

With:

- $V_{H,i}$ = systematic utility of alternative i for home zone H
- $V_{A,i}$ = systematic utility of alternative i for activity zone A
- X_{i1} = value of share of catchment area of alternative i
- X_{i2} = value of stop density of alternative i
- X_{i3} = value of urbanization level of alternative i

In order to complete the tour-based modelling approach, also non-home-based trips have to be considered. These trips are not part of a tour, so the tour-based zonal allocation models are not applicable. Moreover, non-home-based trips have different travel patterns than home-based trips (see appendix B). Therefore, specific non-home-based zonal allocation models are required.

However, the quantitative comparison of the data sources has shown that the share of non-home-based trips is substantially larger in the OV-chipkaart dataset compared to the WROOV dataset. Consequently, the WROOV data are not optimal for estimation of the non-home-based models. Since no better alternatives are available, we continued with the WROOV data.

Table 18: Final estimation results of non-home-based trip zonal allocation models

<i>Model statistics</i>		<i>Origin inference</i>		<i>Destination inference</i>	
Model		Multinomial Logit		Multinomial Logit	
Number of estimated parameters		3		3	
Number of observations		2631		2628	
Rho-square		0.377		0.345	
Diagnostic		Convergence reached...		Convergence reached...	
Parameters		Value	t-test	Value	t-test
$\beta_{i,1}$ (share of catchment area)		0.0519	17.9	0.0525	18.53
$\beta_{i,2}$ (stop density)		2.55	18.77	2.33	18.71
$\beta_{i,3}$ (distance to centroid)		-0.00222	-14.75	-0.00219	-15.03

The estimated non home-based zonal allocation models are represented in formula by the following equations. Equation 6.7 presents the *origin zone allocation model* and equation 6.8 presents the *destination zone allocation model*.

$$V_{O,i} = 5.19 * 10^{-2} * X_{i1} [\%] + 2.55 * X_{i2} [stops/ha] - 2.22 * 10^{-3} * X_{i3} [m] \quad 6.6$$

$$V_{D,i} = 5.25 * 10^{-2} * X_{i1} [\%] + 2.33 * X_{i2} [stops/ha] - 2.19 * 10^{-3} * X_{i3} [m] \quad 6.7$$

With:

- $V_{O,i}$ = systematic utility of alternative i for origin zone O
- $V_{D,i}$ = systematic utility of alternative i for destination zone D
- X_{i1} = value of share of catchment area of alternative i
- X_{i2} = value of stop density of alternative i
- X_{i3} = value of distance to centroid of alternative i



6.2.6 Model stability analysis

In order to assess the transferability of the models, a stability analysis has been performed on the final model configurations. This analysis comprises the estimation of these models on the yearly WROOV data sets, for the years 2003 to 2009.

From this analysis it can be concluded that the parameter estimates for the attributes *share of catchment area* and *stop density* are stable over the years. The parameter estimate for the *urbanization level* is less stable and shows a larger variation over the years. This was also indicated by the relatively low t-value compared to the other parameters. The distinction between home-end and activity-end shows that this variation is mainly caused by the urbanization at the activity-end. For the activity allocation model, this attribute is excluded because it unsettles the model convergence. In the home allocation models, on the other end, the parameter estimates are more stable over time compared to the trip based approach.

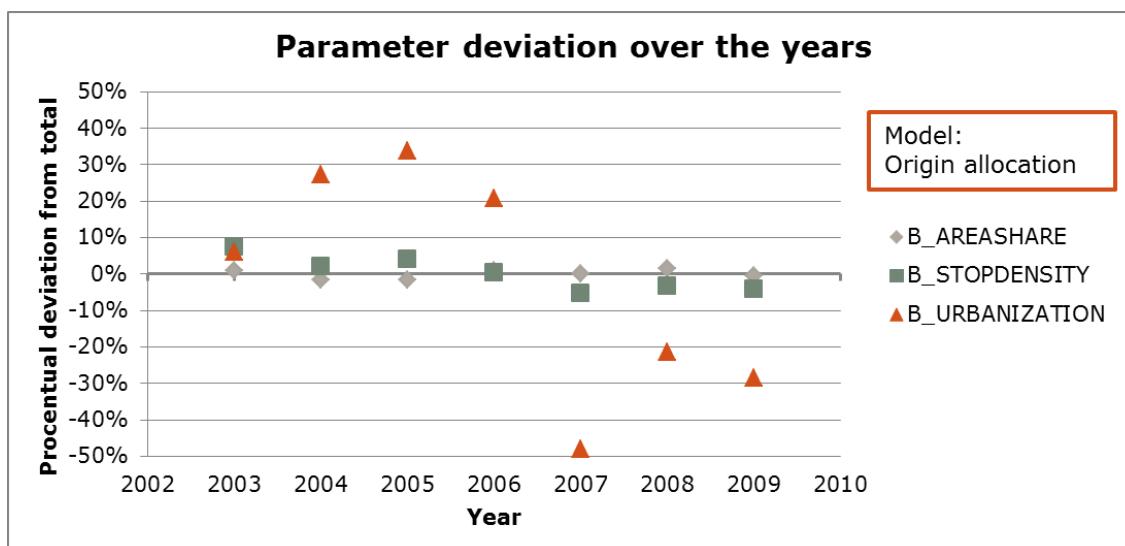


Figure 30: Origin allocation model parameter values for yearly WROOV datasets

The relative deviation of the coefficient for the urbanization level in the trip-based models shows a distinction between two periods: up to 2006 and after 2006 (see Figure 30). The effect of the urbanization level is estimated higher in the years up to 2006, compared to the effect estimated on the entire dataset, and lower in the years after. This is counterintuitive since the urbanization level is determined from data of the year 2010. The distinction between these periods is not observed for the home zone allocation model (see Figure 31). The coefficient for urbanization level still deviates more than the coefficients for share of the catchment area and stop density, but substantially less compared to the deviation in the origin allocation model.



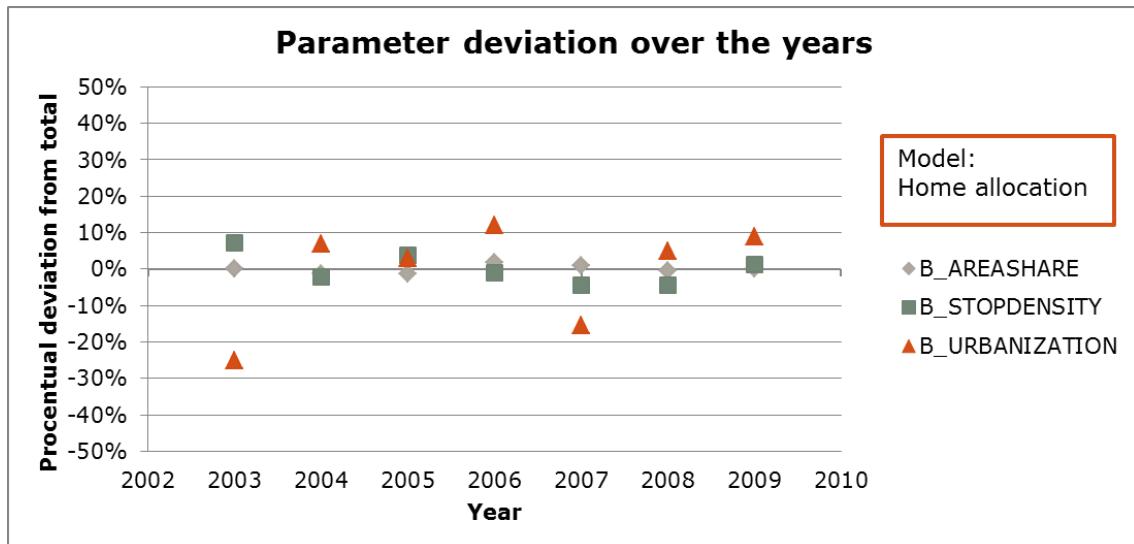


Figure 31: Home allocation model parameter values for yearly WROOV datasets

6.2.7 Interpretation of the model parameters

The trip-based models for origin allocation and destination allocation consist of almost equal parameters, which was expected due to the mix of away trips and return trips in the dataset. First, we interpret the three attributes in the trip-based zonal allocation models, before comparing the coefficients with the distinct tour-based zonal allocation models.

First, the geographical attribute share of the catchment area has the largest effect on the model fit, based on the Rho-squared increase) and proved very stable over time. A higher share of the catchment area in a zone results in a higher utility for that zone, and thus a higher probability. This large effect was expected, since travellers tend to optimize their travel time. Especially in a highly urban environment like Amsterdam, the public transport supply allows travellers to choose stops situated near their origins and destinations, resulting in short access and egress legs. Consequently, origins and destinations are more likely to be near the used stop.

Second, a higher stop density also results in a higher probability of that zone being the origin or destination zone. Since stop locations are deliberately positioned near locations with high public transport demand, these can be interpreted as indicators of trip production and attraction. The stability analysis also indicates a stable effect of the stop density on the probabilities of zones over time.

Third, and last, the level of urbanization indicates a positive relation with the utility of zones. Similar to the stop density, the density of addresses in a zone relates to the trip production and attraction. Addresses are potential origins and destinations, explaining the higher probability of zones with a higher urbanization level.

Compared to the trip-based approach, the tour-based zonal allocation models show differences between the trip-ends. Most notably, the level of urbanization is not included in the activity zone allocation model, since it has a small negative effect on the utility. Moreover, the effect was not significant for all years. Therefore, the attribute was excluded from the activity allocation model. At the home end, the coefficient of the level of urbanization more than tripled compared to its value in the



origin allocation model. Since the coefficients for the share of the catchment area and the stop density are very similar to the coefficients in the origin allocation model, the influence of the level of urbanization on the utility of alternative zones is also greater. This indicates that the effect of the urbanization level in the trip-based zonal allocation models is in fact an averaged effect of the large influence at the home-end and the absent effect at the activity end.

The fact that, of all available land-use attributes, only the stop density and the urbanization are included in the final model results from insignificant or unstable effects of other land-use characteristics. Especially for the activity zone allocation, we expected to find significant effects for the number of jobs and schools in a zone. While these attributes showed significant effects before combining them with geographic attributes, the combination destabilized the model, as it did not converge. A model configuration with only the attributes share of the catchment area and number of jobs did converge. However, replacing the number of jobs by the stop density resulted in a significantly better model fit.

The stop density does not directly relate to origins and destinations of trips, as stops are not the actual locations where the distinguished activities are performed. Hence, the stop density serves as a proxy for activity locations. Since stops are generally located near activity sites, the stop density can be interpreted as an indirect indicator of activity locations. It is preferable to include only direct indicators in the model, since the interpretation of indirect attributes cannot be completely disconnected from other factors. Moreover, direct attributes are more stable over time. Nonetheless, the stop density has been included in the absence of direct indicators of activity locations.

The attributes *share of the catchment area* and *distance to the centroid* are highly correlated. By including both attributes, the model stability was compromised and the interpretation of individual attributes distorted by multicollinearity. Therefore, we have only included one of them in the trip-based and tour-based zonal allocation models. We have chosen the attribute share of the catchment area over the distance to the centroid for three reasons. The primary reason is the better model fit of the share of the catchment area. Second, the attribute distance to the centroid resulted in instability of the models. In several instances, the model did not converge with only this attribute. We have not found the exact cause of this instability, but it might be that the relation of the distance to the centroid is not linear with the log odds of the alternatives. Third, the share of the catchment area is considered a better indicator of the nearness of a zone to the used stop than the distance to the centroid in case of small catchment areas. The distance of the centroid depends on the size of the zone and its relative location to the used stop. In case larger catchment areas are applied, the distance to the centroid showed a better fit than the share of the catchment area.

In the non-home-based zonal allocation models, we aimed for the highest model fit. Since the available dataset of non-home-based trips from WROOV is not comparable to the non-home-based trips in OV-chipkaart data we focussed on the tour-based models. In order to limit the effects of non-home-based trips, the interpretation of the models was considered less important. The models include the correlated attributes share of the catchment area and the distance to the centroid, since including both did not compromise the model stability. As a result, the interpretation of the individual attributes is distorted. However, it can be concluded that the effects of share of the catchment area and stop density are similar to the effects in the trip-based and tour-based models. The distance to the centroid has a negative effect on the utility. This



was expected, as travellers tend to optimize their travel time, and thus minimize the access and egress distances.

In order to get a feeling of the effect of the attributes on the probabilities in the different models, we present two examples of trip-ends with corresponding alternatives and their probabilities: example 1 in Table 20 and example 2 in Table 21. The values in the examples are fictional, but based on the descriptive statistics from the dataset presented in Table 19.

Table 19: Descriptive statistics of the attributes in the zonal allocation models

Attributes	Minimum	Maximum	Mean	Std. Deviation
Share of catchment area [%]	0	100	14.1	15.7
Stop density [stops/ha]	0	1.25	0.18	0.15
Level of urbanization [addresses/ha]	0	207	48.9	41.8

The first example shows that, with few alternatives, the influence of the attribute share of the catchment area is overestimated due to the relative scale, which results in high absolute differences of the percentage between zones. With a higher stop density and higher urbanization level, zone 2 was expected to obtain a higher probability than 4%. The relative scale of the share of the catchment area results in heteroscedasticity of the residuals. The probabilities differ per approach, caused by the different effects of the level of urbanization. Even though the differences are small, it does clearly indicate the absence of the urbanization level in the activity allocation and the averaging effect in the trip-based models.

Table 20: Zonal allocation example 1

Attributes	Zone 1	Zone 2
Share of catchment area [%]	80	20
Stop density [stops/ha]	0.1	0.5
Urbanization level [addresses/ha]	100	200
Probabilities	$P_{n,1}$	$P_{n,2}$
Origin zone allocation	98%	2%
Destination zone allocation	98%	2%
Home zone allocation	96%	4%
Activity zone allocation	99%	1%

Example 2 is more representative for the dataset, as it has more alternatives. Consequently, the differences between the values of share of the catchment area are smaller. Consequently, the probabilities are more dispersed over the alternatives. However, the effect is still very large, which complies with the effect on the model fit. The effect of the stop density is also substantial, looking at the probabilities of alternatives 4 and 5. The effect of the urbanization level is not prominent in the trip-based models. On the other hand, the tour-based models do show a larger effect. In the home zone allocation model, differences are noticeable between the probabilities



of zones 2 and 3, while the probabilities of these zones are equal in the activity zone allocation.

Table 21: Zonal allocation example 2

Attributes	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
Share of catchment area [%]	5	10	10	20	20	35
Stop density [stops/ha]	1	0.2	0.2	0.2	0.5	0.2
Level of urbanization [addresses/ha]	150	100	200	100	100	100
Probabilities	$P_{n,1}$	$P_{n,2}$	$P_{n,3}$	$P_{n,4}$	$P_{n,5}$	$P_{n,6}$
Origin zone allocation	29%	4%	5%	9%	22%	32%
Destination zone allocation	28%	4%	4%	9%	22%	32%
Home zone allocation	32%	4%	7%	9%	20%	29%
Activity zone allocation	24%	4%	4%	9%	22%	36%

6.3 Probabilistic purpose inference models

The purpose inference models have been approached separately from the zonal allocation models. The model structure is different due to the pre-determined number of alternatives, which are not generic. Because the influence of trip characteristics is not equal for the distinct purposes, these can be implemented as model attributes (see also paragraph 3.8.6).

6.3.1 Alternative selection

Although the WROOV study specified more alternatives, only four alternatives are distinguished in this study. The purposes *work*, *education* and *shopping* are the most frequent travel purposes in bus and light rail. Remaining purposes are aggregated in the purpose *other*.

The purposes *work*, *education* and *shopping* are the most relevant purposes for long term forecasts. Besides the fact that they are the most frequent purposes, they are also the most susceptible for policy measures. Commuting traffic is influenced by policy on work hours, flex-workers, home-workers. Educational traffic is influenced by policy measures regarding the student PT cards and study financing. Shopping traffic is influenced by increased opening hours of shops at specific locations. On the other hand, visiting family and friends or hospitals are less prone to policy measures due to their optional character. The identification of business trips would be interesting regarding the influence of policies, but it generates few trips with bus and light rail and therefore cannot be distinguished from other discretionary purposes.

The travel data analysis has indicated that the compulsory purposes *work* and *education* are correlated in terms of key variables. The same holds for the discretionary purposes *shopping* and *other* (see paragraph 5.2.1). Nonetheless, we have aimed at the inference of these four purposes due to the inclusion of the activity duration and contract duration.



6.3.2 Available attributes and model structure

The logit models estimate relative utilities of the alternatives. Therefore, one alternative has to be normalized. Since only differences in the utilities matter to the final probabilities of alternatives being chosen, the most logical normalization is to set one alternative utility to zero. Because the purpose other is the least specific of the distinguished travel purposes, we have normalized this purpose to zero.

The categorical attributes are implemented by means of dummy variables. This means for every category of an attribute a specific coefficient is estimated, with exception of one reference category. Similarly to the alternatives, the categories have to be normalized by setting one category to zero. The choices of references categories are based on our perception of the least specific category of every attribute, related to the specified purposes work, education and shopping. Hence, we aimed at the highest possible coefficients.

The available attributes, their categorization and their values are presented in

Table 22. The following attributes have been investigated during the model estimation process:

- *Concession*: the concession where the trip took place. We expected different travel patterns for (1, the reference) Amsterdam, (2) Waterland and (3) travellers traversing both concessions in their trips;
- *Mode*: the mode used during the trip, in this study limited to (1, the reference) bus, (2) tram and (3) metro or (4) multiple modes;
- *Departure time*: the departure time of the trip, categorized in five times of day: (1) the early morning, between 4 am and 7 am, (2) the morning peak, between 7 am and 9 am, (3, the reference) the day, between 9 am and 4 pm, (4) the evening peak, between 4 pm and 6 pm and (5) the night, between 6 pm and 4 am;
- *Distance travelled*: the distance covered by public transportation, based on the route travelled, measured in kilometres;
- *Number of trip legs*: the number of trip legs within the trip, also perceptible as the number of transfers made within the trip plus one;
- *Activity duration*: the duration between consecutive trips, calculated by the time between consecutive boardings, measured in minutes;
- *The contract duration*: the validity period of the travel product. Contracts durations are limited to (1, the reference) no contract, (2) year contracts and (3) month contracts;
- The frequency: the travel frequency, measured by the number of tours per week.

In addition to these trip characteristics and travel pattern characteristics, land use characteristics at both ends of the trips have been investigated. Depending on the definitions used, the trip-based origin and destination, or the tour-based home-end and activity-end, their expected influence differs. Because the exact origin and destination of trips are unknown, we aggregated the land-use characteristics over a 400 metre radius around the used stop. The analysis of access and egress distances shows that a substantial amount of access and egress legs cover longer distances than 400 metres (see paragraph 5.1). Nonetheless, we applied a conservative radius because the averaging effect over larger areas increases. This means that the land-use attributes might not reflect the exact values at the origins and destinations.



Table 22: Available attributes for purpose inference

<i>Attribute category</i>	<i>Attribute</i>	<i>Measurement level</i>	<i>Values/ [Units]</i>
Trip characteristics	concession	categorical	Amsterdam Waterland Both
	mode	categorical	Bus Tram Metro Multiple
	Departure time	categorical	Early morning Morning peak Midday Afternoon peak Night
	distance travelled	continuous	[km]
	number of trip legs	continuous	[trip legs within trip]
	activity duration	continuous	[minutes]
	contract duration	categorical	Year Month None
	frequency	continuous (-)	[trips per week]
	households	continuous (-)	[addresses]
	average income	continuous (-)	[€/year]
Land use characteristics	jobs	continuous (-)	[# of jobs]
	residents	continuous (-)	[# of residents]
	schools	continuous (-)	[# of student places]
	students	continuous (-)	[# of students]
	working population	continuous (-)	[# of working residents]

Since we expected different effects of attributes on specific purposes, we implemented purpose-specific coefficients for each attribute. In combination with the implementation of dummy variables for each categorical attribute, this resulted in a large number of parameters to be estimated.

6.3.3 Model enhancement

The model enhancement strategy applied for the purpose inference models is similar to the approach of the enhancement of zonal allocation models (see paragraph 6.2.3), but contains different categories of attributes. Figure 32 presents the framework of the selection of attributes to include in the purpose inference models.



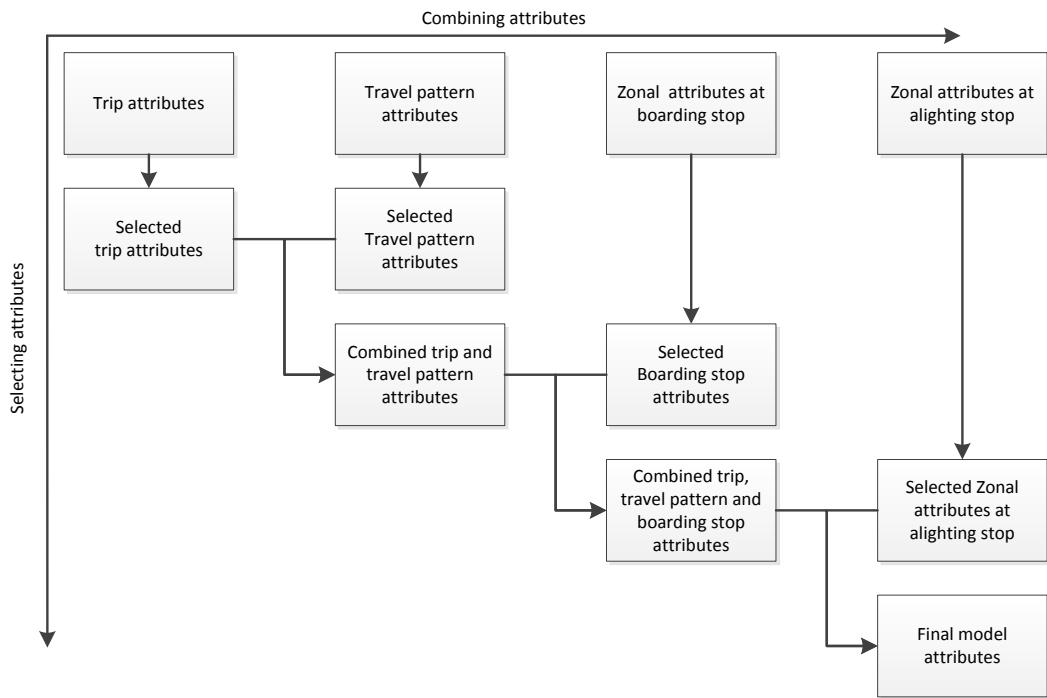


Figure 32: Enhancement strategy of purpose inference models

Since several land-use attributes are highly correlated, especially the ones related to the home end, we applied distinct attributes related to a specific purpose. At the home-end this included the number of households for the purpose shopping, the working population for the purpose work and the number of students for the purpose education. At the activity-end, the attribute *jobs* was included in the utility function for the purpose *work* and the attribute *schools* in the utility function for the purpose *education*. An attribute related to shops was not available to implement in the utility of the purpose *shopping*.

In the trip-based approach, the distinction between home-end and activity-end is not made. Therefore, all attributes were initially implemented at both trip-ends. Without alternative specific constants, the land-use attributes showed a substantial improvement of the log-likelihood. After adding alternative specific constants, however, these proved to explain the same increase in log-likelihood.

In the tour-based approach, the fit of the model configurations with only land-use characteristics was higher compared to the trip-based approach. The attributes are all significant in the configurations without trip and travel pattern attributes, but the models do not converge. The same problem with land-use attributes arose in the zonal allocation models. The exact origin of the model instability resulting from the land-use attributes has not been found.

Although the travel analysis indicated an increased share of work trips for longer distances, the travel distance was excluded from the final model configurations. Different implementation methods have been investigated: linear, quadratic and exponential, but in every case the effect was insignificant. Therefore, it is concluded that the effect of the travel distance on the utility of specific purposes is already explained by other attributes.



During the model estimation process, the same data sample was used as for the zonal allocation models. The final model attributes were then tested for stability over time on the yearly WROOV datasets. Since the longitudinal data analysis showed a slight trend in the purpose shares, the final model coefficients from 2008 are used. This was the most recent year without interference of the OV-chipkaart on the WROOV survey in the Amsterdam region.

6.3.4 Final model parameter estimates

In the final model configurations, only trip characteristics and travel pattern characteristics are included in all model approaches. The trip-based purpose inference model incorporates five attributes, plus alternative specific constants. The tour-based approach allows for the incorporation of the additional attribute *activity duration*. Because the activity duration is correlated to several other attributes, its inclusion results in several alternations compared to the trip-based model. In the final configuration, seven attributes are included. The number of estimated parameters, however, is lower compared to the trip-based model due to the exclusion of correlated parameters. The non-home-based model only includes three attributes, as the other attributes proved insignificant.

Table 23: Model statistics for the three specific purpose inference models

Model statistics	All trips	Tours	Non-home-based trips
Model	Multinomial Logit	Multinomial Logit	Multinomial Logit
Number parameters	of 25	24	16
Rho-square	0.376	0.460	0.318
Diagnostic	Convergence reached...	Convergence reached...	Convergence reached...

The model fit statistics indicate an increased fit for the tour-based model, which can be attributed to the inclusion of the attribute activity duration. Table 24 presents the estimated coefficients for each model parameter in the three model approaches. The purpose-specific coefficients make this table hard to read. It does, however, provide an easy comparison of the coefficient values in the different approaches. In order to increase the readability of the final model configurations, Table 24 is translated into formulas that describe the utility functions for each purpose in all model approaches. The utility functions are presented in equation 6.8 to equation 6.9.



Table 24: Final parameter estimates for specific purpose inference models

<i>Parameters</i>	<i>Purpose</i>	<i>All trips</i>	<i>Tours</i>	<i>Non-home-based</i>		
Attribute		β	t-test	β	t-test	β
ASC Education	E	-1.78	-43.64	-2.38	-35.83	-3.07
ASC Shopping	S	-0.862	-28.99	0	(fixed)	-1.55
ASC Work	W	-1.33	-48.26	-2.53	-27.09	-1.24
Frequency	E	0.342	26.91	0.467	26.63	0.179
Frequency	S	-0.246	-12.81	0	(fixed)	0
Frequency	W	0.267	31.03	0.346	22.41	0.131
Concession Waterland	S	0.585	7.73	0.694	6.03	0
Concession Waterland	W	0.532	9.47	0	(fixed)	0
Concession Both	W	0.525	10.11	0.285	3.48	0
Contract duration Year	E	-0.779	-9.8	-1.61	-16.59	-1.02
Contract duration Year	S	0	(fixed)	-0.264	-2.06	0
Contract duration Year	W	0.954	24.41	0	(fixed)	0.0801
Contract duration Month	E	0.849	13.75	0	(fixed)	0.534
Contract duration Month	S	0	(fixed)	0	(fixed)	0
Contract duration Month	W	0	(fixed)	-0.812	-11.56	-0.836
Mode tram	E	-0.609	-12.48	-0.49	-6.62	0
Mode metro	S	-0.631	-9.05	-0.614	-5.98	0
Mode metro	W	0.662	18.53	0.527	9.06	0
Mode multiple	W	0.31	8.41	0.431	5.7	0
TOD early morning	S	0	(fixed)	-2.17	-3.68	0
TOD early morning	W	2.28	28.46	1.88	17.98	2.13
TOD morning peak	E	1.14	21.3	2.08	26.22	1.98
TOD morning peak	S	-1.71	-12.57	-1.52	-9.08	-1.48
TOD morning peak	W	1.68	40.95	1.67	28.33	1.6
TOD evening peak	E	-0.79	-9.68	0	(fixed)	0
TOD evening peak	S	-0.464	-6.82	-0.972	-6.67	-0.629
TOD evening peak	W	1.32	34.47	0	(fixed)	0.263
TOD night	E	-1.18	-12.09	0	(fixed)	0
TOD night	S	-0.999	-12.91	-0.628	-3.89	-1.17
number of trip legs	S	0	(fixed)	-0.266	-6.4	0
number of trip legs	W	0	(fixed)	-0.376	-7.12	0
Activity duration	S			-0.00218	-10.49	
Activity duration	W			0.00466	31.33	



Trip based purpose inference model

$$V_{n,work} = -1.33 + 0.267 X_1 + \begin{bmatrix} 0.532 \\ 0.525 \end{bmatrix} \hat{X}_2 + \begin{bmatrix} 0.954 \\ 0 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} 0 \\ 0.662 \\ 0.310 \end{bmatrix} \hat{X}_4 + \begin{bmatrix} 2.28 \\ 1.68 \\ 1.32 \\ 0 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 0$$

$$V_{n,educ} = -1.78 + 0.324 X_1 + \begin{bmatrix} -0.779 \\ 0.849 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} -0.609 \\ 0 \\ 0 \end{bmatrix} \hat{X}_4 + \begin{bmatrix} 0 \\ 1.14 \\ -0.79 \\ -1.18 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 1$$

$$V_{n,shop} = -0.862 - 0.246 X_1 + \begin{bmatrix} 0.585 \\ 0 \end{bmatrix} \hat{X}_2 + \begin{bmatrix} 0 \\ -0.631 \\ 0 \end{bmatrix} \hat{X}_4 + \begin{bmatrix} 0 \\ -1.71 \\ -0.464 \\ -0.999 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 2$$

Tour based purpose inference model

$$V_{n,work} = -2.53 + 0.346 X_1 + \begin{bmatrix} 0 \\ 0.285 \end{bmatrix} \hat{X}_2 + \begin{bmatrix} 0 \\ -0.812 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} 0 \\ 0.527 \\ 0.431 \end{bmatrix} \hat{X}_4 + \begin{bmatrix} 1.88 \\ 1.67 \\ 0 \\ 0 \end{bmatrix} \hat{X}_5 - 0.376 X_6 + 4.66 * 10^{-3} X_7 \quad 6.1 \\ 3$$

$$V_{n,educ} = -2.38 + 0.467 X_1 + \begin{bmatrix} -1.61 \\ 0.849 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} -0.490 \\ 0 \\ 0 \end{bmatrix} \hat{X}_4 + \begin{bmatrix} 0 \\ 2.08 \\ 0 \\ 0 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 4$$

$$V_{n,shop} = \begin{bmatrix} 0.694 \\ 0 \end{bmatrix} \hat{X}_2 + \begin{bmatrix} -0.264 \\ 0 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} 0 \\ -0.614 \\ 0 \end{bmatrix} \hat{X}_4 + \begin{bmatrix} -2.17 \\ -1.52 \\ -0.972 \\ -0.602 \end{bmatrix} \hat{X}_5 - 0.266 X_6 - 2.18 * 10^{-3} X_7 \quad 6.1 \\ 5$$

Non-home-based purpose inference model

$$V_{n,work} = -1.24 + 0.131 X_1 + \begin{bmatrix} 0.0801 \\ -0.836 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} 2.13 \\ 1.60 \\ 0.263 \\ 0 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 6$$

$$V_{n,educ} = -3.07 + 0.179 X_1 + \begin{bmatrix} -1.02 \\ 0.534 \end{bmatrix} \hat{X}_3 + \begin{bmatrix} 0 \\ 1.98 \\ 0 \\ 0 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 7$$

$$V_{n,shop} = -1.55 + \begin{bmatrix} 0 \\ -1.48 \\ -0.629 \\ -1.17 \end{bmatrix} \hat{X}_5 \quad 6.1 \\ 8$$

With: V_{ni} = systematic utility of purpose i for trip n

X_1 = value of travel frequency

\hat{X}_2 = unit vector of concession=[EBS both]

\hat{X}_3 = unit vector of contract duration = [year month]

\hat{X}_4 = unit vector of mode = [tram metro multiple]

\hat{X}_5 = unit vector of departure time = [early morning morning peak evening peak night]

X_6 = value of number of trip legs

X_7 = value of activity duration

In these utility formulas, the dummies of categorical variables are presented as unit vectors. These consist of vectors with a 1 for the represented category and zeros for



the other categories. In case of the reference category, the vector is a null vector, with only zeros.

6.3.5 Interpretation of the individual model parameters

All the estimated utilities are relative to the purpose other. Moreover, the coefficients for categorical variables are relative to the reference category. The alternative specific constants correct the estimated utilities for unobserved factors. Since these are also relative to the estimated coefficients, they do not represent a specific utility or disutility.

The travel frequency has a positive relation to the utilities of the compulsory purposes and a negative effect on the utility of shopping, although this was not significant in the tour-based model. Since the effect of the travel frequency was expected to be similar for shopping and other, this complies with the expectations.

The attribute concession shows an increased utility for the purposes work and shopping for trips made in Waterland in the trip-based model. However, in the tour-based model, trips made in Waterland do not have a higher utility for the purpose work. Trips covering both concessions do have a higher utility for the purpose work. This indicates that travellers from Waterland that commute to Amsterdam are more likely to continue their trip with GVB, while shopping travellers do not.

In the trip-based model, year contracts have a positive effect on the utility of the purpose work, while it has a negative effect for the purpose education. In the tour-based approach, the effect of year contracts on the purpose work is not included due to high correlations with the activity duration and the frequency. Month contracts have a positive effect on the utility of the purpose education. This was expected with the high share of month contracts for educational trips that was shown in the travel analysis.

The different modes have several distinct effects on the utilities of purposes. For the purpose work, trips by metro relate to a higher utility. Also multiple modes are an indication for the purpose work, although less than the metro alone. Trips made by tram have a negative relation to the utility of educational trips and a similar effect is revealed for the metro on the utility of the purpose shopping. Although interpreting the effect of mode on travel purposes is less straightforward compared to contract duration and travel frequencies, the coefficients show high similarity between the trip-based and the tour-based approach.

The departure time, or Time Of Day, was one of the main explanatory variables identified in literature. This is confirmed by the estimated coefficients, which are the highest of all categorical variables. Regarding the purpose work, the highest effect of the departure time is found in the early morning. Although not many trips are made during this time of day, the share of work trips is very large. Also the peak periods result in an increased utility, but the effect is lower, especially in the evening peak. This can be attributed to the fact that the evening peak of commuters is more spread over time compared to the morning peak. Regarding the purpose education, return trips are generally made before the evening peak. This results in an increased utility for the morning peak and a decreased utility for the evening peak, relative to the day. In the tour-based approach, however, only the morning peak was significantly



different from the day. Since educational trips are less frequently observed in the night compared to other trips, a negative coefficient was expected. Regarding the purpose shopping, all coefficients are negative, indicating that most shopping trips are made during the day.

An increasing number of trip legs decreases the utility of the purpose work. This is counterintuitive as work trips make relatively more transfers. The attribute was included since it improved the model fit, but it is also correlated to the alternative specific constant. For interpretational reasons, it might be better to exclude the effect of the number of trip legs for the purpose work. The negative effect on the purpose shopping does comply with the observations from the travel analysis.

The activity duration was expected to provide additional explanatory value to the distinction between purposes work and education. However, the implementation of the activity duration as continuous variable resulted in an insignificant effect on the purpose education. Since work trips generally have longer activity durations than education trips and are more frequent, a positive effect on the utility of education would overestimate the influence of long activities for education. Consequently, the distinction between work and education requires the implementation of the activity duration as a categorical variable, similar to the rule-based purpose inference. Longer activities do have a negative effect on the utility for the purpose shopping. This was expected since shopping generally does not involve long activities.

Even though the distinction between work and education trips was limited by the continuous level of the activity duration, the implementation did results in a substantially higher fit statistic for the tour-based model. Therefore, it can be concluded that the activity duration explains a different effect on the travel purpose than any other available attribute.

The non-home-based purpose inference model only incorporates three attributes. The mode and the concession did not have a significant effect on the utilities of any purpose. The remaining attributes indicate very similar effects compared to the trip-based and tour-based models. The effect of a year contract was not significant in the 2008 dataset, but was in every other year. Since it only involves a small value, and thus has a limited effect, we included in the model.

The large number of estimated parameters results in difficulties to assess the influence of specific attributes on the final probabilities of purposes. In order to get a feeling of the probability distributions over purposes, Table 25 presents four fictional trips and their respective probability distributions according the three purpose-inference models. The four trips represent a common trip for a specific purpose.

The first example trip shows a very high probability for work, indicating that the model is well able to distinguish work trips. The second example, related to the purpose education, assigns the highest share to this purpose. However, the probability is closer to the probability of the purpose work. The tour-based model does assign a higher probability for education compared to the trip-based model. Hence the inclusion of the activity duration does result in an improved inference of the education purpose, even though it is not specified in its own utility function. Example trip three indicates that the trip-based model does not assign high probabilities to the purpose shopping, as it is still lower than the probability of other. The tour-based model does



result in a higher probability. The fourth example trip indicates that the purpose other can also obtain high probabilities, although not as high as work.

Table 25: Examples of travel purpose inference

<i>Attributes</i>		<i>trip 1</i>	<i>trip 2</i>	<i>trip 3</i>	<i>trip 4</i>
Frequency		4	5	1	1
Concession		Amsterdam	Amsterdam	Waterland	Amsterdam
Contract duration		Year	Month	None	None
Mode		Metro	Bus	Bus	Bus
Departure time		Morning peak	Morning peak	Day	Night
Trip legs		2	1	1	1
Activity duration		540	420	90	240
Model	Probabilities	$P_{1,i}$	$P_{2,i}$	$P_{3,i}$	$P_{4,i}$
Trip-based	$P_{n,work}$	91%	41%	24%	22%
	$P_{n,education}$	4%	52%	10%	5%
	$P_{n,shopping}$	0%	0%	25%	8%
	$P_{n,other}$	4%	8%	41%	65%
Tour-based	$P_{n,work}$	89%	37%	5%	15%
	$P_{n,education}$	5%	55%	6%	9%
	$P_{n,shopping}$	0%	0%	50%	15%
	$P_{n,other}$	5%	7%	40%	61%
Non-home-based	$P_{n,work}$	67%	33%	21%	23%
	$P_{n,education}$	6%	38%	3%	4%
	$P_{n,shopping}$	1%	1%	13%	5%
	$P_{n,other}$	26%	27%	63%	69%

The non-home-based model results in substantially more spread probabilities of purposes. The non-home-based dataset does contain a substantially larger share of trips with the purpose other, compared to the datasets used for the trip-based and the tour-based models. This is reflected by the relatively high shares of the purpose other in all examples. The less sharp probability distributions indicate that the non-home-based model cannot distinguish travel purposes with the accuracy of trip-based or tour-based purpose inference models.

6.3.6 Assessment of the predictive qualities

By applying the purpose inference models onto the WROOV data, an assessment of the predictive qualities of the models per purpose has been performed. Figure 33 and Figure 34 show the probability distribution of each purpose categorized by the purposes observed in the WROOV data, respectively for the trip-based model, the tour-based model and the non-home-based trips model. These graphs indicate that the purposes *work* and *other* obtain the highest probabilities within their own category. The purposes *education* and *shopping* obtain lower probabilities within their own category. For *education* trips, the average probability of *work* is higher than the



probability of *education*. For *shopping* trips the average probability of *other* is higher than the probability of *shopping*. Hence it can be concluded that the models perform well regarding the distinction of *work* and *other* trips, but lack the ability to distinguish between *work* and *education* and between *shopping* and *other*.

The travel analysis already indicated that these categories are correlated in terms of the key variables. The addition of the contract duration and the activity duration were expected to allow for a differentiation between work and education purposes.

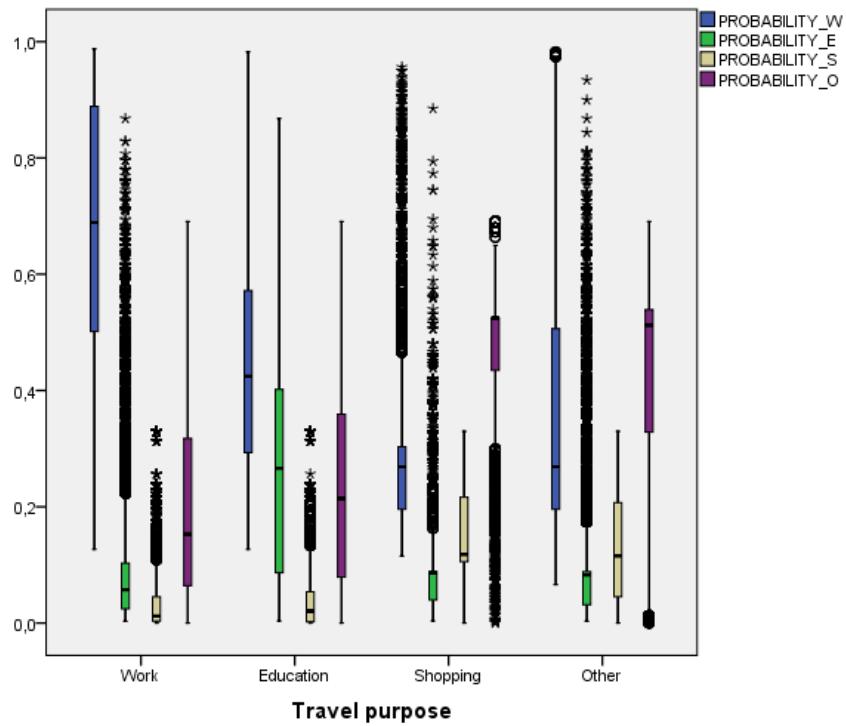


Figure 33: Probability distributions per purpose for the trip-based model



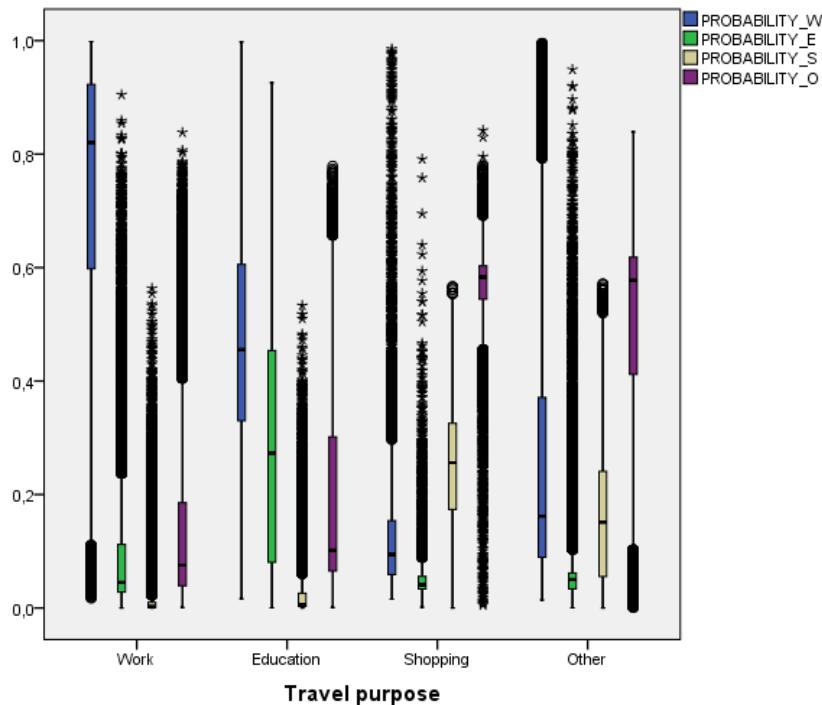


Figure 34: Probability distributions per purpose for the tour-based model

6.4 Identification of concession traversing transfers

Tests proved it was possible to add the alternative of a transfer at stops near train stations to the choice set, besides the zonal alternatives. However, this complicated the interpretation of the models, as the addition of a reference alternative added an additional value to the attributes. With the addition of the transfer alternative, the same attributes are to describe the chance a traveller originates or continues the travel in the train. Hence, the transfer to another system is positioned equal to the access or egress legs. While both processes include a trip leg, they do not describe the same phenomenon.

The method of a transfer alternative in the zonal allocation models would be interpretable with a nested logit structure. In that case, the zonal alternatives can be included in a nest, where the transfer alternative would comprise its own nest. Since this method requires a different structure of the logit models and an altered specification of the data set, this method has not been continued within this research. If future studies still have to deal with unavailability of data from one or several operators, this method might provide a potential solution.

Therefore, the identification of concession traversing transfers is based on the transfer shares found in the WROOV analysis (see Table 12 on page 59). The table presents the percentages of BTM trips using BTM as egress transport for a train leg at boarding stops near the specific train station. Similarly, the table indicates the shares of trips using BTM as access transport for a train leg at alighting stops. Using the tour-based definitions of home-end and activity-end, the percentages indicate the shares of travellers arriving in the study area by train, so using the train at the home-end, and the shares of travellers leaving the study area by train in order to perform an activity elsewhere, so using the train at the activity end of the observed BTM trip.



The tour-based definitions allow for a specification by location and, thereby, show a large difference between travellers coming to Amsterdam to perform their activities and travellers leaving Amsterdam to perform their activities elsewhere. For stations outside Amsterdam, the opposite is observed. This characterizes the attraction of the capital.

The WROOV data do not incorporate the train stations Amsterdam Science Park and Purmerend Weidevenne, since these stations did not exist at the time of data collection. As a result, transfer shares for these stations are not available. Therefore, the same transfer shares are applied as similar train stations, respectively the stations RAI and Wormerveer.

6.5 Conclusions regarding the model estimations

The estimation process of the logit models included the assessment of many model configurations in order to find the most suitable zonal allocation models and purpose inference models. This paragraph contains the conclusions of the assessment of the final model results. These conclusions on the models are also input for the final phase of this study: the matrix evaluation, which is presented in the next chapter.

6.5.1 Conclusions on the zonal allocation models

The catchment areas provide a suitable method for the alternative generation for the logit zonal allocation models. The distinction of the sizes of catchment areas by mode allows for including a larger set of relevant zonal alternatives. In general, the most relevant alternatives are selected in the choice set of the logit model as these are nearest to the used stop. However, the mode only explains part of the variation in access and egress distance distributions at stops. As a result, the selection of alternatives does not fully comply with the catchment areas of individual stops. This results in the inclusion of irrelevant alternatives, but more importantly, the exclusion of relevant alternatives. Additional fine-tuning of this method is considered possible by including attributes related to the level of service at the stop, like line frequencies and speeds, and the stop surroundings, like the stop density.

Regarding the attributes in the logit allocation models, the attribute stop density proved to be a stable estimator in the zonal allocation, in both the trip-based approach as the tour-based approach. However, this attribute is not directly related to activity locations, but can be interpreted as an indirect indicator of activity locations. It is not preferable to include indirect attributes in the model, but the attributes directly related to activity locations do not provide stable indicators in the model.

In the selections applied for the test data, filtering was based on catchment areas instead of indication of transfers. As a result respondents who entered a train station as origin or destination remained in the dataset. Since this only implies a very small share of the data, a slight overestimation of zones with train stations might occur. This overestimation is encountered in the attribute stop density, as train stations are generally served well by bus and light rail.

The tour-based approach clearly shows the difference in effect of the level of urbanization. Since addresses are the primary indicator of homes, and to lesser extent of activities, it is concluded that the tour based approach results in a more accurate allocation of home-zones. The allocation of activity zones is less accurate due to the instability of indicators of activity locations.



6.5.2 Conclusions on the purpose inference models

Five different attributes have a significant effect on the trip-based estimation of the travel purpose. The tour-based purpose inference includes two additional attributes: the activity duration and the number of trip legs. The activity duration is only included in the tour-based approach since it contains information about the relation between consecutive trips. The inclusion of the number of trip legs is due to correlations between attributes. With the inclusion of the activity duration, the effects of correlations have changed, resulting in a different model configuration.

The trip-based model performs well on the inference of the most frequently observed purposes work and other. The less frequently observed purposes education and shopping are less accurately predicted. This is caused by the correlation in key variables describing the compulsory purposes work and education, and on the other hand, the discretionary purposes shopping and other.

With the implementation of the attribute activity duration in the tour-based model, we expected a better distinction between the purposes work and education. However, the tour-based purpose inference model only shows a small accuracy increase. This can be attributed to the linear implementation. A categorical implementation of the activity duration is likely to improve the distinction between the purposes work and education.

The attributes have a different effect on distinct purposes. Therefore, purpose specific coefficients have been estimated in the models. This resulted in a large number of estimated coefficients in the model, even though the number of distinct attributes is moderate. Due to the large sample, overfitting of the model to the dataset was possible. However, the stability analysis showed that the model parameters are fairly constant over the years.

The application of the travel frequency might not be appropriate, as the distribution of travel frequency in the WROOV data does not comply with the distribution of travel frequency in the OV-chipkaart data. This is directly related to the over representation of contracts in the WROOV data that was observed in the quantitative comparison of the sources.

Land-use characteristics proved to be insignificant or instable in the logit estimation models. This can be attributed to several explanations. First, the land-use data are derived from a different time, namely 2010, where the travel data is collected between 2003 and 2009. Second, the land-use data are averaged around the used stop. The averaged characteristics do not necessarily comply with the land-use characteristics at the origin or destination, which is unknown. In their similar study, Chakirov & Erath also do not find a high explanatory value of land use characteristics.

Besides the better fit of the tour-based model, it is also preferable over the trip-based approach since it includes a higher level of behavioural richness. The tour-based approach has several qualitative advantages over the trip-based approach:

- Consistency of train transfer selection within tours;
- Application of different transfer shares at home-end and activity-end;
- Consistency between zonal allocation between consecutive trips in a tour (the destination of a trip matches the origin of the next trip);
- Addition of the highly relevant attribute of activity duration.





7 Evaluation of OD matrices

This chapter covers the evaluation of the resulting base matrices. The trip-based and tour-based models, described in the previous chapter, have been applied to the datasets of both WROOV and OV-chipkaart. In addition to these two logit modelling approaches, rule-based models have been applied. These models are based on a rule-based processing approach found in literature and adapted to fit the available data. In this chapter, we assess the qualities of these three model approaches by comparison of their resulting OD matrices.

The evaluation framework, which is described in paragraph 3.9, consists of linear regressions onto the matrix cell values. The specifications of the matrices allow for assessment of the matrices at different levels. The evaluation of the matrices at different zonal resolutions allows for the assessment of the zonal allocation models, where the distinction of the matrices by purpose allows for the assessment of the purpose inference models.

First, the procedure of the model applications is described (paragraph 7.1), in order to provide insight in the sophistication of each model. Second, the models are validated onto the WROOV data by comparison with the observed OD matrix (paragraph 7.2) in order to assess their predictive qualities. Third, in addition to the application onto WROOV data, we evaluate the differences between the model approaches by comparing the OD matrices constructed by application onto OV-chipkaart data (paragraph 0), in order to relate the quality to the required efforts of applying more sophisticated models. Fourth, the travel demand described by the different sources is compared (paragraph 7.4) in order to assess the durability of the estimated models. Finally, the conclusions of the evaluation are summarized (paragraph 7.5).

7.1 Procedure of model applications

In order to facilitate an unbiased comparison, adjustments have been made to the original data sets from WROOV and the OV-chipkaart. These adjustments are elaborated upon, before the model application procedure is expounded.

7.1.1 Creation of comparative data sets

In the qualitative comparison of the data sources WROOV and OV-chipkaart, dissimilarities have been observed in the coverage of the described system (see paragraph 3.4). In order to correct for these differences, selections have been applied to the OV-chipkaart data. Students are filtered based on their specific card type. Conversely, the filtering of tourists proved more difficult. By filtering the short term contracts, the datasets do become more similar regarding their coverage of travel products, but still do not describe the same travel in terms of key variables. The quantitative comparison shows that the OV-chipkaart data described more trips during off peak hours and more trips with shorter activities (seep paragraph 5.4).

Furthermore, the unavailability of data from adjacent public transport concessions results in the incapability to derive travel demand to zones outside the available service area. Therefore, transfers to the train network have been inferred based on transfer shares at individual train stations. In the WROOV data set, only trips without the indication of a train transfer are selected.



The total number of trips within the matrix differs between model approaches onto OV-chipkaart data due to different filtering of transfers to the train. Concession traversing transfers are inferred based on the transfer shares at train stations found in the WROOV data (see Table 12). In the trip-based approach, these shares are filtered randomly over trips boarding, respectively alighting, at the specific train stations. In the tour-based approach, filtering of concession traversing transfers at train stations is applied consistent between consecutive trips. This means that when a concession transfer is inferred at the activity-end of the trip, both the away trip and the return trip to that activity are filtered.

Table 26: Multiplication factors of OD cells for equal trip totals

<i>Data source</i>	<i>construction method</i>	<i>Trip total</i>	<i>Mean cell value</i>	<i>Standard deviation</i>	<i>Multiplication factor</i>
OV-chipkaart	Reference model	1.491.993	8,34	63,98	0,2148
	Trip-based model	1.587.888	8,87	39,99	0,2019
	Tour-based model	1.602.765	8,95	50,35	0,2000
WROOV	Observed	153.923	0,86	3,93	2,0826
	Reference model	156.611	0,87	5,15	2,0468
	Trip-based model	156.654	0,88	3,34	2,0462
	Tour-based model	158.718	0,89	2,93	2,0196

Furthermore, the reference models of both datasets result in different trip totals compared to the matrices constructed by the model applications due to the filtering of trips with origin or destinations with a mismatch onto the zonal grid. Some stops are projected on the wrong side of the zonal border, and in case there is no adjacent zone, i.e. water, it is filtered from the matrix. In order to compensate for the different trip totals, all OD matrices for the complete day and all purposes are corrected with a multiplication factor. The multiplication factors for OV-chipkaart matrices also include the aggregation of a week data into an average working day. Since the tour-based construction of OD matrices with OV-chipkaart data is considered to be the best representation of the real travel demand, all matrices are factored to that total. Matrices with distinctions per time of day or purpose are multiplied with the same factor as the corresponding complete matrix in order to preserve the differences between the model approaches and data sources. The multiplication factors are presented in Table 26.

7.1.2 Model application

The three applied approaches for the construction of purpose-specific OD matrices differ in complexity and, therefore, in the effort required to apply them.

The reference models require the least effort. The zonal allocation is a straight match between the used stops and zones. Because the coordinates of public transport stops and the zonal borders are derived from different systems, with a different coordinate system, slight errors are possible. However, these errors are generally negligible. Only in case the stop coordinates are on the waterfront, this can result in a mismatch. The purpose is inferred based on the activity duration and allocated to all trips in the tour.

The trip-based models require significantly more effort in their application. The allocation of origin and destination zones and the inference of the travel purpose are



based on the probabilities estimated by the logit models. This approach was chosen over a simulation procedure, which might result in biased allocations due to correlations between error terms. The allocation based on probabilities resulted in the spreading of one trip over multiple zones at the origin, multiple zones at the destination and four travel purposes. This procedure increases the data file size with a factor 35 and requires long computation times, especially with a large data set as the OV-chipkaart data.

The application of tour-based models requires the most effort, due to the consistency between trips in the same tour. This is especially complicated for tours with more than two trips. First, the selection of train transfers requires consistency. In case a transfer is inferred at the home or activity location, both the away trip and the return trip have been filtered. Second, zonal allocations for consecutive trips are consistent with the previous zonal allocation, with exception of the last trip in the tour, which destination is allocated to the same as the origin of the first trip: the home-zone. Travel purposes have not been inferred consistently within tours, so a tour can consist of trips with different purposes. It is possible to infer the travel purpose consistently within tours, similar to the zonal allocation. The purpose of the return trip can be handled in several ways. It can be allocated to the main travel purpose, for example based on the longest activity duration, or allocated to the specific purpose of being at home. The home purpose is not distinguished in this study, but the main travel purpose is also not identified. Therefore, purposes are inferred for individual trips.

The three approaches have been applied to both the WROOV data and the OV-chipkaart data. Together with the OD matrix observed by WROOV, this resulted in the seven differently constructed matrices presented in Table 26.

For the matrix evaluation, the entire matrix of the VENOM study area has been applied, although it contains many empty cells due to the structural boundaries of the available concessions. This is also visualized by the stop locations of the available concessions Amsterdam and Waterland in Figure 27.

7.2 Model validation on WROOV data

A ground truth of the real travel demand at the time of collection of the OV-chipkaart data is not available. The OD matrix observed by WROOV data can be considered as a valid representation of the travel demand. Although its accuracy is limited by the sample size and the data collection period is spread over seven years, it is the best representation of travel demand available. Therefore, the models have been validated by application onto WROOV data and comparing the resulting OD matrices with the OD matrix observed by WROOV. Five distinct specification of OD matrices have been compared for all three construction methods, on three levels of zonal resolution. Consequently, 45 comparisons have been made (see Appendix D). The resulting matrices of three approaches are assessed successively, in order of increasing complexity.

7.2.1 Rule-based OD matrices

The assessment of the rule-based approach on r^2 statistics of the linear regression shows a large deviation between the zonal resolutions. On the level of VENOM zones, which is the smallest grid in the assessment and the applied level of resolution in the VENOM model, the fit of the total OD matrix constructed with the rule-based approach is poor ($r^2 = 0.330$). With increasing zonal sizes, the fit increases. At the level of PC3 zones, the r^2 statistic of the total matrix approaches 1 ($r^2 = 0.981$), indicating a very



good fit. Consequently, we conclude that the access and egress legs have to be taken into account when describing the public transport travel demand at the resolution of VENOM zones or PC4 zones. When describing the travel demand at the resolution of PC3 zones, on the other hand, the influence of access and egress legs is insignificant regarding the allocation of origin and destination zones.

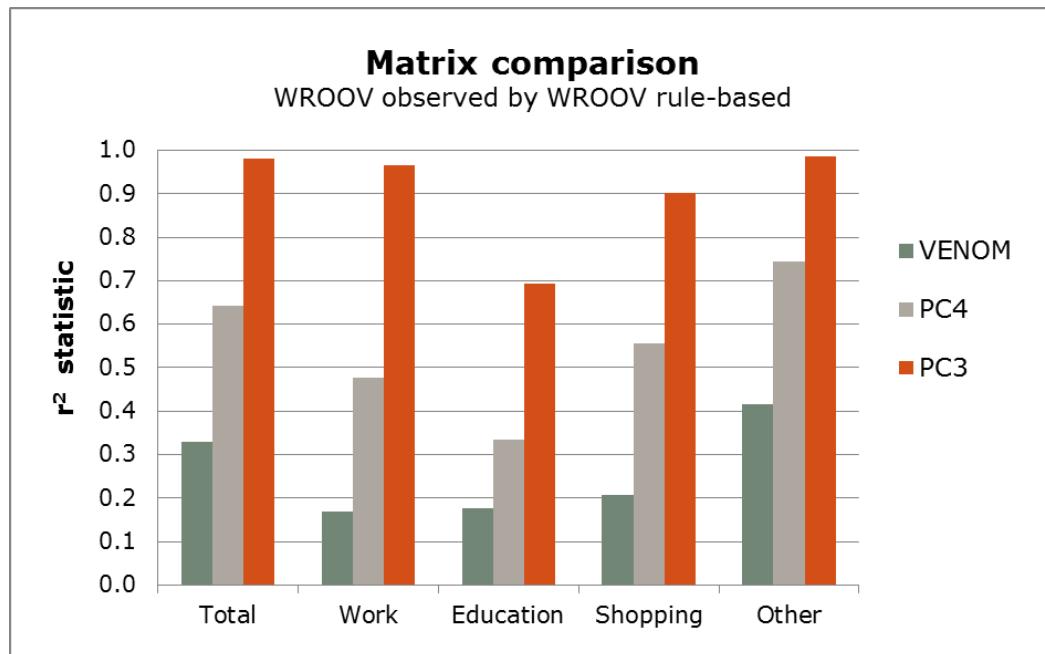


Figure 35: r^2 statistics of the rule-based model validation

The fit of purpose-specific matrices show a similar pattern as the total matrix, with low r^2 statistics at the VENOM zone level and increasing r^2 statistics for larger zones. However, the three specifically identified purposes *work* ($r^2 = 0.169$), *education* ($r^2 = 0.177$) and *shopping* ($r^2 = 0.206$) have a poorer fit than the non-specific purpose *other* ($r^2 = 0.415$). This indicates that the rule-based inference of the travel purpose does not capture the main features of the public transport travel demand.

7.2.2 Trip-based OD matrices

The trip-based approach can be considered as a more sophisticated method for the construction of purpose-specific OD matrices compared to the rule-based approach. The conversion from a stop-based matrix to OD matrix takes into account the access and egress legs by means of mode-specific catchment areas of public transport stops and allocates trip-ends to nearby origin and destination zones based on land-use attributes. The purpose inference is based on five specific trip characteristics, compared to one attribute in the rule-based approach.

The r^2 statistic of the total matrix constructed with the trip-based models shows a better fit ($r^2 = 0.626$) with the observed OD matrix compared to the rule-based approach at the VENOM zonal resolution. Also on the PC4 level, the trip-based models perform better than the rule-based models. On the PC3 level, the difference in r^2 statistics is very small ($r^2 = 0.990$).



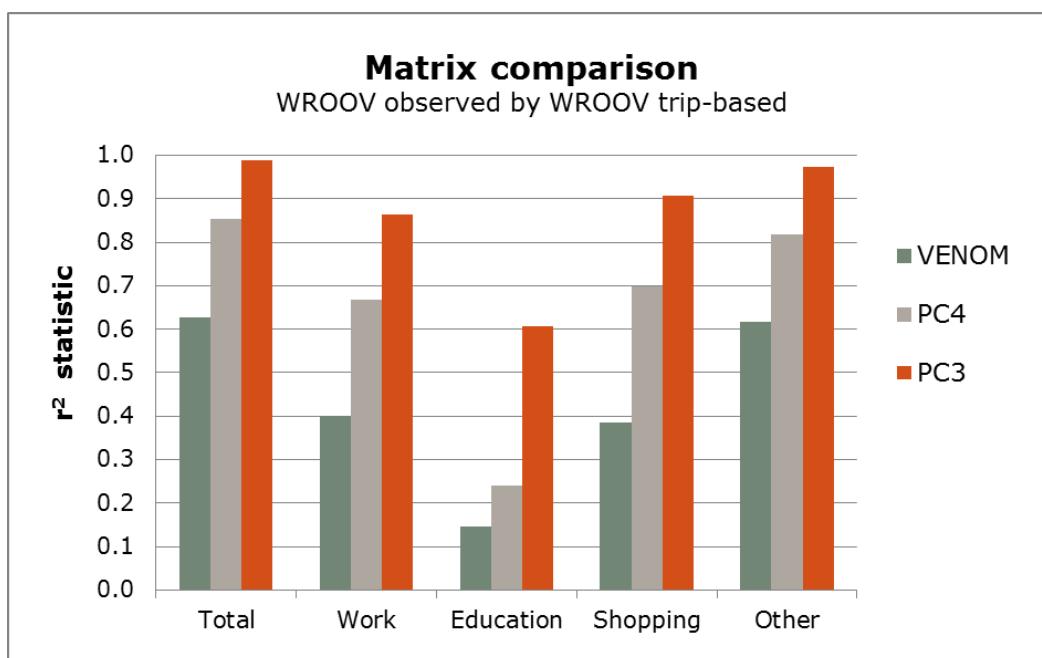


Figure 36: r^2 statistics of the trip-based model validation

Evaluating the purpose-specific OD matrices constructed with the trip-based approach, the r^2 statistics again show an inferior fit for the specific purposes *work* ($r^2 = 0.402$), *education* ($r^2 = 0.147$) and *shopping* ($r^2 = 0.384$), compared to the purpose *other* ($r^2 = 0.618$). The trip-based models perform better compared to the rule-based models on all purposes, except the purpose *education*, which has a slightly poorer fit. This was expected, since the model assessment indicated low probabilities for the purpose *education* (see paragraph 6.3.6). Conversely, the *shopping* matrix does improve compared to the rule-based approach, at a similar level as the *work* matrix, while the model assessment indicated low probabilities similar to the purpose *education*. Overall, we conclude that the trip-based approach performs substantially better than the rule-based approach.

7.2.3 Tour-based OD matrices

The tour-based models are based on the same allocation technique as the trip-based models, logit allocation, but take into account the relation between consecutive trips within tours. Specifically, the tour-based models specify trip-ends by their geographical location and allocate consecutive trips consistently. The tour-based purpose inference model incorporates the attribute activity duration, which is not available in the trip-based approach.

The r^2 statistic of the total matrix at the VENOM level is comparable to the trip-based approach, but slightly lower ($r^2 = 0.574$). This also holds for the total matrix at the lower levels of zonal resolution. This leads to conclude that the zonal allocation models do not perform better than the trip-based zonal allocation. This can be attributed to the lacking of land-use attributes in the tour-based zonal allocation models, which could possibly have translated the additional information of specific geographical locations into better allocation probabilities. However, these attributes are lacking since they have not been found stable indicators of home and activity zones (see paragraph 6.2.6).



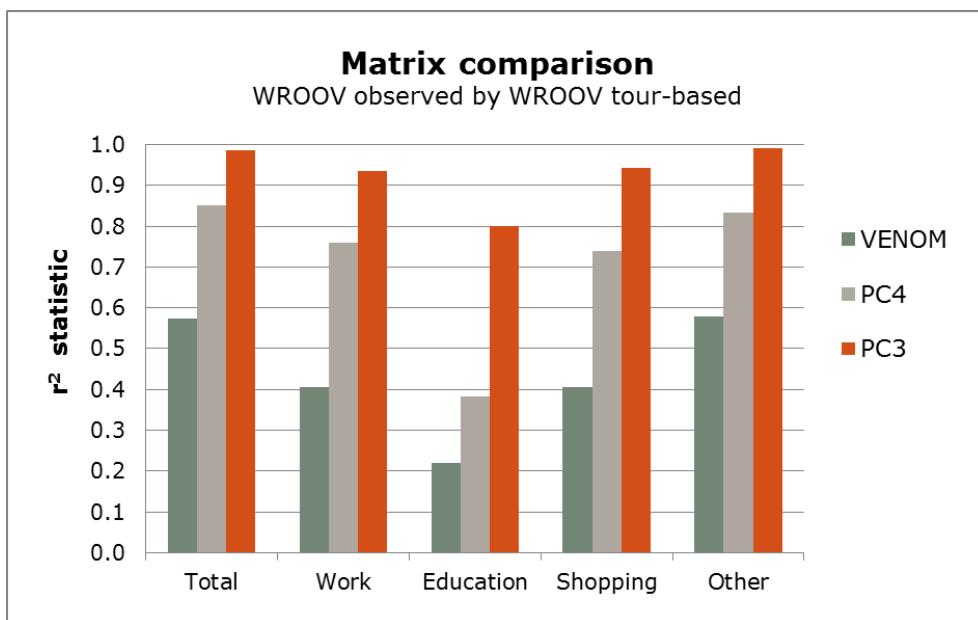


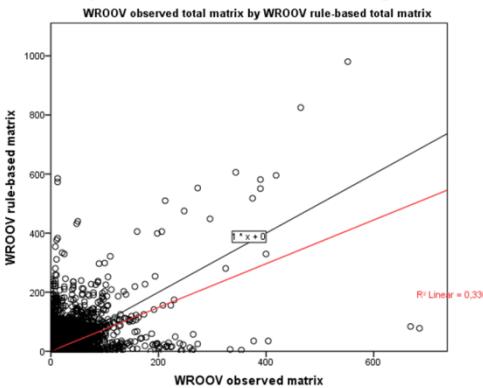
Figure 37: r^2 statistics of the tour-based model validation

Evaluating the purpose-specific OD matrices created with the tour-based models, we find similar fit statistics for the *work* matrix at the VENOM resolution ($r^2 = 0.405$), compared to the trip-based models. The matrices for the purposes *education* ($r^2 = 0.220$) and *shopping* ($r^2 = 0.404$) have a better fit than their trip-based counterparts, while the *other* matrix has a slightly lower fit statistic ($r^2 = 0.579$). At the PC4 level, the differences between the tour-based and the trip-based matrices are more apparent, in favour of the tour-based matrices. This can be attributed to the slightly lower accuracy of the zonal allocation in the tour-based approach, which is moderated at a lower zonal resolution. The tour-based other matrix also has a better fit than the trip-based other matrix at the PC4 level. Therefore, it can be concluded that the tour-based models perform best on the purpose inference. Nonetheless, the zonal allocation does not improve by the geographical distinction of trip-ends as might be expected.

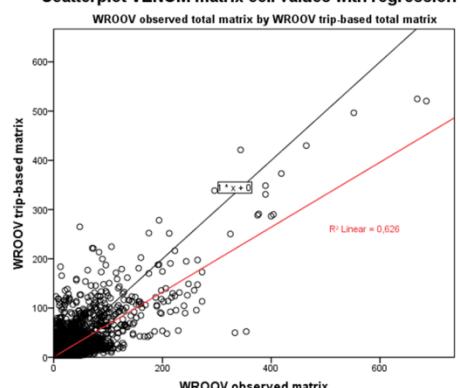
Figure 38: Model validation regression lines



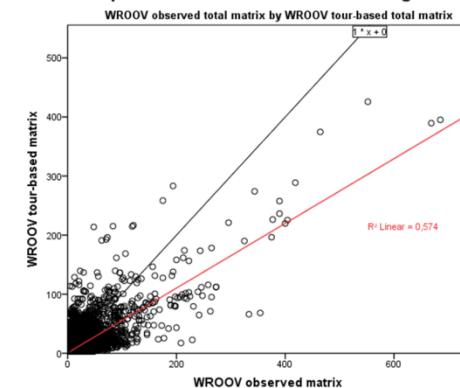
Scatterplot VENOM matrix cell values with regression line



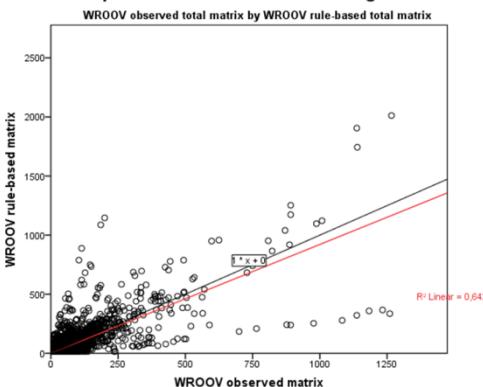
Scatterplot VENOM matrix cell values with regression line



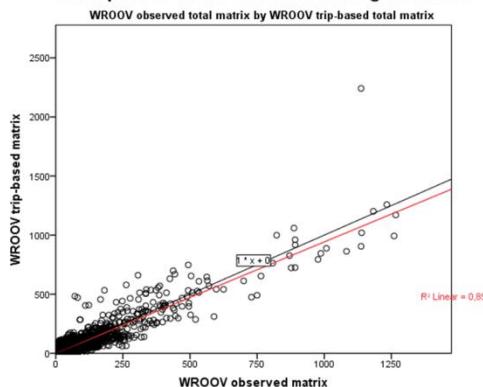
Scatterplot VENOM matrix cell values with regression line



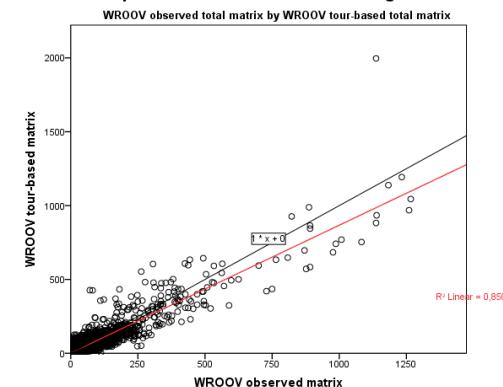
Scatterplot PC4 matrix cell values with regression line



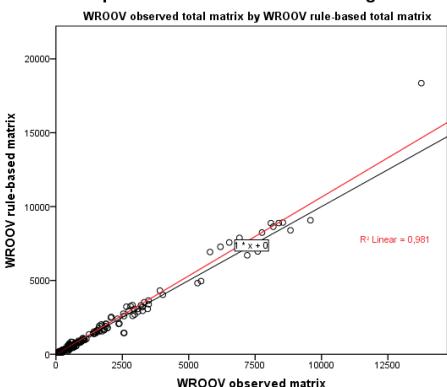
Scatterplot PC4 matrix cell values with regression line



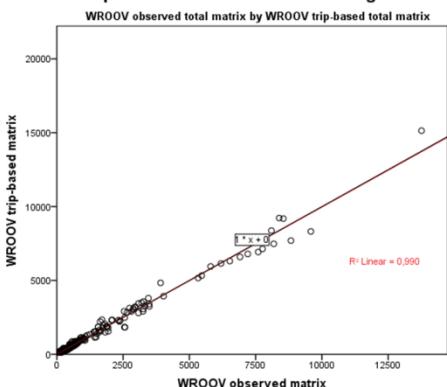
Scatterplot PC4 matrix cell values with regression line



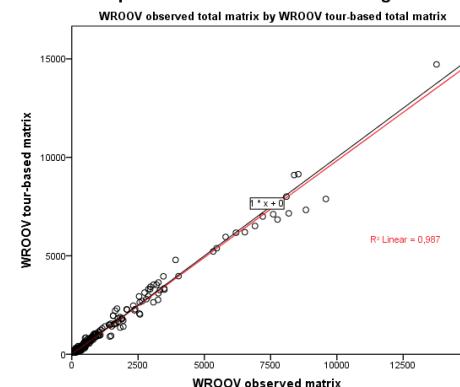
Scatterplot PC3 matrix cell values with regression line



Scatterplot PC3 matrix cell values with regression line



Scatterplot PC3 matrix cell values with regression line



7.3 Evaluation of differences between modelling approaches

In addition to the model validation onto WROOV data, the three model approaches have been applied to OV-chipkaart data in order to evaluate the differences between the OV-chipkaart OD matrices per model approach. Since there is no observed OD matrix available for the OV-chipkaart data, like there is for the WROOV data, we assess the differences between the matrices constructed by different approaches. In this comparison, the tour-based matrices have been used as dependent and are respectively estimated by rule-based and the trip-based matrices. The matrix comparisons for the model approach evaluation consist of five distinct matrices that have been compared on three levels of zonal resolution, resulting in 30 matrix comparisons (see Appendix D).

Evaluating the differences between the total matrices of the three approaches, we find large differences in the fit statistics. At the VENOM zonal resolution, the rule-based matrix has a very low fit with the tour-based matrix ($r^2 = 0.120$). The trip-based matrix has a better fit, but still deviates significantly from the tour-based matrix ($r^2 = 0.523$). At the level of PC3 zones, the difference between the rule-based and the trip-based approach is less apparent. The fact that even at the PC3 level, the matrices do not have r^2 statistics that approach the value of 1 indicates a large difference due to the filtering of concession traversing transfers.

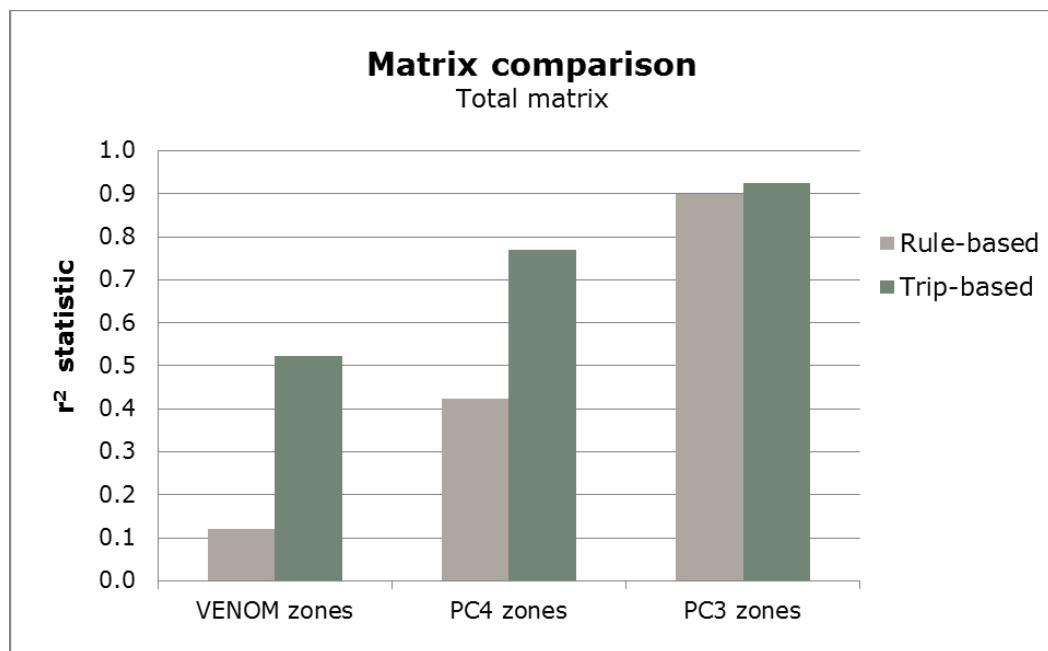


Figure 39: r^2 statistics of approach comparison of OV-chipkaart total matrices

Since a ground truth is lacking for the OV-chipkaart matrices, we have analysed the largest relations in the study area. A selection of the most frequent OD pairs, the highest values in the matrix, shows large dissimilarities between the model approaches. Even between the trip-based and the tour-based approach the selection of the highest values only show moderate comparability. Zones with train stations are very well represented in this selection of OD pairs. Hence, the influence of the selections at train stations is substantial.



The fact that, between the trip-based and the tour-based approach, the comparability of the purposes work and other is lower than the comparability of the purposes education and shopping was not expected. This might be related to the homoscedasticity assumption of linear regression. This assumption is not met by the matrix comparisons, as these show larger deviation for higher values (see Figure 38).

Especially at the resolution of VENOM zones, these deviations indicate that the current zonal allocation models do not fully capture the essence of access and egress behaviour.

7.4 Source comparison on travel patterns

The comparisons of the matrices from different sources consist of eight distinct matrices that have been compared on three levels of zonal resolution, resulting in 24 matrix comparisons (see Appendix D).

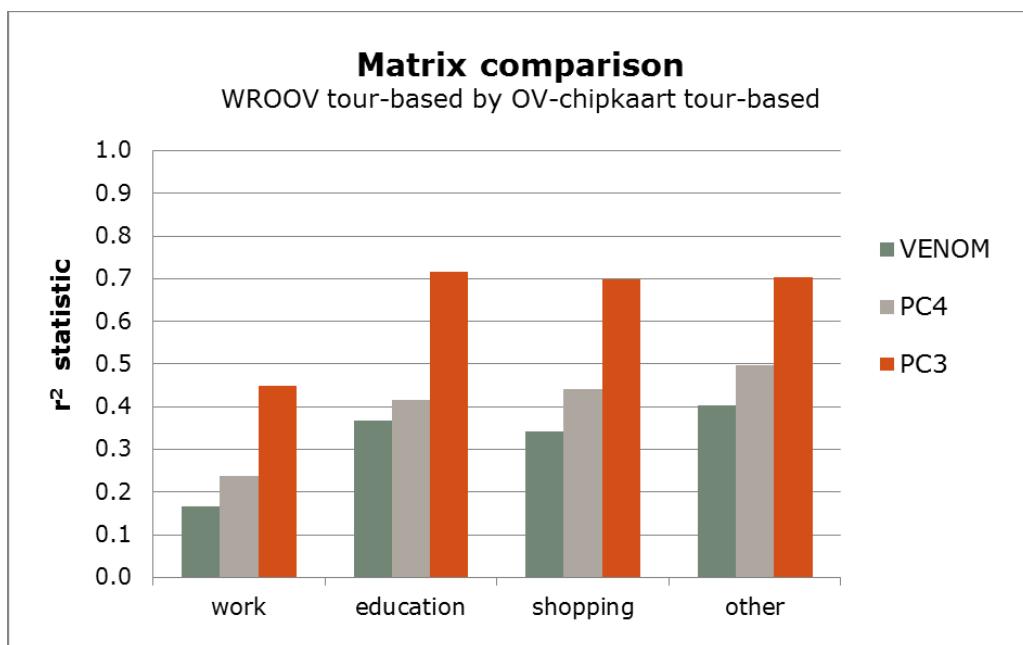


Figure 40: r^2 statistics of source comparison per purpose

The travel patterns described differs between WROOV and OV-chipkaart data. Explanations of these differences can be classified into four categories:

1. Differences in the target population;
5. The time gap between data collection;
6. Differences in the data samples;
7. Model inaccuracies.

In order to deal with differences of the first category, the differences in the target population, we have applied filters to create comparative data sets. Both students and tourist have been filtered from the OV-chipkaart data for the comparison. However, the accuracy of these filters is not perfect. The identification of students is accurate thanks to the specific card type, but the identification of tourists and concession-traversing transfers is less direct. Tourists are filtered based on short-term contracts,



but tourist may also travel with stored value. In that case, they cannot be distinguished and thus cannot be filtered. The identification of concession traversing transfers has been applied on an aggregated level per train station. The selection is based on transfers with the NS only, which leaves the transfers to the adjacent concessions of Amstelland-Meerlanden and Zaanstreek. As a result, the numbers of trips originating or terminating in the study area are overestimated.

Regarding the second category, differences due to the time gap, differences can be caused due to changes in the network and the provided transport services. Alterations of the travel supply influence the demand. In addition, the travel demand changes over time due to changes in the demography and travel behaviour. These changes have not been investigated within this research. With a minimal time gap of five years, their influence is considered small, but might not be negligible.

The third category, changes in the data sample, consists of a sample bias in WROOV and differences in the sample period. The WROOV surveys aimed at a description of travel demand for one year by means of multiplication factors. These factors have not been used in the analysis because of the uncertain influence of the many data selections. In addition, the WROOV sample consists of the stacked data of seven years. On the other hand, the OV-chipkaart data sample contains one week. Even though this week appears to be similar to the average work week, based on aggregated travel characteristics, it might be that this specific week describes different movement patterns between zones than the average work week.

The fourth category covers the model inaccuracies. As described in the previous chapter, the predictive qualities of the models vary.

7.5 Conclusions regarding the matrix evaluation

The tour-based models perform best on the model validation on WROOV data on almost all specifications of the OD matrix and all zonal aggregation levels and all assessment parameters.

The differences between tour-based models and trip-based models are small. The most notable difference occurs in the OD matrix with trips of the purpose education. For the OD matrix with trips of the purpose other, the trip-based approach even performs better on the r^2 statistic.

The rule-based approach generally performs significantly less on all assessment parameters compared to the trip-based and tour-based modelling approaches.

The influence of the zonal resolution heavily influences the r^2 statistic. The aggregation to larger zones diminishes the differences at higher resolution. However, the parameters a and b do not improve for zones in a lower resolution (i.e. larger zones). This is related to the higher mean of the cell values.

The education matrices overall have an inferior fit compared to other purposes. On the one hand, this was expected since the inference of the purpose *education* proved to be difficult with the available data. The WROOV data does not include college students, but only incorporates students of primary, secondary and vocational schools. Moreover, the locations of student places did not prove to be stable indicators of the purpose education. On the other hand, the predictive qualities are similar to the purpose *shopping*, while these matrices show higher comparability with the observed



WROOV matrix. This leads us to conclude that *educational* trips have very different spatial patterns than *work* trips, which are not captured by the trip-based model. *Shopping* trips, on the other hand, seem to have similar patterns to the purpose *other*.



8 Conclusions

The previous chapters all have contributed to answering the sub-questions of this research regarding their distinct topics. This chapter summarizes the main conclusions drawn from the different phases of this study, leading to the answer of the main research question.

The main goal of this study was the construction of purpose-specific OD matrices using public transport smart card data. We have investigated to what extent this is possible by enriching trips observed by smart card data with the required information: the origin zone, the destination zone and the travel purpose, based on data from the WROOV survey. The research was structured by dividing this main research question into four sub-questions, relating to the following subjects:

1. The identification of attributes correlated to the lacking information;
2. The transferability of information between the available sources by means of these attributes;
3. The quality of the enrichment models;
4. The evaluation of the constructed OD matrices.

First, we answer these four sub-questions, before we reflect on their accumulated conclusions in order to answer the main research question.

8.1 Relevant attributes

"Which travel characteristics are correlated to the information to be added to OV-chipkaart data and to what extent?"

8.1.1 Access and egress trip legs

First, we consider the allocation of origin and destination zones. In order to convert the stop-based matrices, which are derived from OV-chipkaart data, into OD matrices, the access and egress trip legs have to be considered, which are not observed by the OV-chipkaart.

Since we aimed for a generic conversion applicable for all stops, access and egress trip legs have been analysed in the WROOV data through their distances, as it was not feasible to assess the large number of stops individually. Previous studies indicated a relation between the access and egress distances to public transport stops and the level of service provided at these stops. However, neither the WROOV data nor the OV-chipkaart data contain information about frequency or operational speed of transit lines. Therefore the correlation of the mode has been investigated, as the level of service differs per mode.

The analysis verified that, on average, travellers cover longer distances during access and egress legs for trips by metro, and to lesser extent by tram, compared to trips by bus. Although the level of service deviates less strongly between the modes bus and tram compared to the level of service of the metro, travellers did cover larger distances at the home-end of trips to and from tram stops compared to bus stops. This difference was not observed at the activity-end. Considering attributes unavailable in OV-chipkaart data, we conclude that, at the home-end, the access and



egress distances depend mainly on the level of service. Conversely, on the activity-end, the access and egress distances depend less on the level of service and more on traveller characteristics. Moreover, the access and egress distances at the activity-end are correlated to the travel purpose.

8.1.2 Travel purpose

Considering the attributes related to the travel purpose, we have analysed the distribution of purposes over individual travel characteristics in the WROOV data. This analysis has shown that the travel purpose is strongly correlated to various travel characteristics, most notably to the activity duration, the departure time, the travel distance, the contract duration and the travel frequency. In addition, the travel purpose is also strongly correlated to the traveller characteristics unavailable in OV-chipkaart data, especially age, and to lesser extent gender.

The deviation of purpose shares over the departure time and the travel frequency distinguishes two clusters of purposes: the compulsory purposes work and education and the discretionary purposes shopping and other. The activity duration and the contract duration do show distinctions between the purposes work and education and to lesser extent between the purposes shopping and other.

Overall, we conclude that the analysed travel characteristics have large potential to describe the travel purpose. Especially the distinction between the compulsory purposes work and education and the discretionary purposes shopping and other is demonstrated by many attributes. In addition, the distinction between work and education is demonstrated by the attributes activity duration and contract duration. The distinction between the purposes shopping and other is less apparent via the investigated attributes.

8.2 Transferability of information

"How do the data sources OV-chipkaart and WROOV compare to each other?"

8.2.1 Qualitative comparison of the data sources

The qualitative comparison of the sources shows several influential dissimilarities. Predominantly, the data collection period of the WROOV surveys covers seven years between 2003 and 2009, where the OV-chipkaart data used in the analysis covers one week in 2014. This implies differences between the travel demand, which might have changed between 2009 and 2014. Moreover, the selected week of OV-chipkaart data might not be representable for the average working week. Even though the selection of the week was based on aggregated similarity with the full year average, circumstances, like weather conditions or maintenance, might influence the depicted travel patterns.

The passive collection of OV-chipkaart data results in lacking information for the construction of OD matrices. On the other hand, the passive data collection also facilitates a sample that approximates complete coverage of the entire public transport system. The only exceptions are fare dodgers and forgotten check-ins and check-outs. Despite this high potential coverage, only data from two public transport concessions were available for this study, resulting in unobserved transfers to adjacent concessions. Since these transfers result in different origins or destinations, this is a limitation to this study due to the unavailability of data. The consequences of this limitation on travel are discussed under sub-question 4 (paragraph 8.4).



On the other hand, the WROOV data do not cover the complete public transport system, but lack student cards and local tickets, which were mainly used by international tourists. Moreover, the WROOV surveys focus on bus and light rail and do not include train travel, although transfers to the train are indicated. Therefore, the WROOV data is not fit to estimate the lacking information for trips made by students and international tourists. This discrepancy in the coverage of both sources has been corrected by removing these travellers from the OV-chipkaart data. Students were easily identifiable by their card type but international tourists are identified indirectly, based on short term contracts, day and week tickets, since they are more likely to stay in Amsterdam for a short period.

8.2.2 Key variables

By comparing the available information in both sources, key attributes have been identified that are available in both sources. Since the WROOV survey is focussed on public transport, the data includes many attributes that are also available in OV-chipkaart data, resulting in a large set of potential attributes that can be used to describe the lacking information in OV-chipkaart data. Table 27 presents the key variables and their representation in the purpose inference models. The next paragraph discusses their appropriateness as model attributes

Table 27: Key variables and their representation in the purpose inference models

<i>Key variables between OV-chipkaart and WROOV data</i>	<i>Representation in the Unit/Categories</i>
<i>purpose inference models</i>	
Activity duration	Continuous Minutes
Travel frequency	Continuous Travels per week
Departure time	Categorical (5) Early morning Morning peak Midday Evening peak Night
Travel distance	Not included (kilometres)
Contract duration	Categorical (3) Year Month None
Fare	Not included (full fare Reduced fare Unlimited travel)
Mode	Categorical (4) Bus Tram Metro multiple
Concession	Categorical (3) Amsterdam Waterland Both
number of legs within trip	Continuous Legs per trip
number of trips within tour	Not included (Trips per tour)



8.2.3 Quantitative comparison

The quantitative comparison on these key-variables shows slight deviations between WROOV and the OV-chipkaart, i.e. after filtering of students (20%) and international tourists (9%) from the OV-chipkaart data. Regarding the modal shares, related to the access and egress distances, the WROOV data slightly overestimates the shares of multiple modes in one trip at the expense of the modes tram and metro. Regarding the attributes related to the travel purpose, the WROOV data underestimate trips with activities shorter than three hours and overestimate activities longer than nine hours. Moreover, the WROOV data underestimate short distance trips. These discrepancies verify this well-known phenomenon of underreporting of short trips in travel surveys. In addition to the overestimation of long activities, the WROOV data also overestimate the trip shares in peak hours at the expense of trips during off-peak hours and the higher travel frequencies. These three discrepancies are correlated to the higher share of long-term contracts in the WROOV data, compared to the OV-chipkaart data.

On the other hand, the OV-chipkaart data underestimate short activities. The distinction between short activities and transfers relies on rule-based processing. This study applied an enhanced framework of processing rules for the identification of short activities. Compared to the applied processing rules, the standard application of 35 minutes transfer time overestimates transfers with 22%. Nonetheless, the distribution of activity durations still indicates an underestimation of short activities.

This leads to the conclusion that the WROOV data over represents long term contracts compared to the current travel demand, and consequently, the purposes work and education. This can be caused by the data collection method or changes in the travel demand. Because the exact cause cannot be determined we have applied the described attributes in the model estimation despite their dissimilarities.

The key variables fare and number of trips within a tour have not been applied in the enrichment model, as they are not considered stable indicators of the travel purpose. Fare systems have been altered since the introduction of the OV-chipkaart and the WROOV data does not contain tours with more than two trips, while in the OV-chipkaart data 23% of the tours have more than two trips.

8.3 Quality of the enrichment models

"How can OD matrices by purpose and by time of day for the mode BTM be constructed using OV-chipkaart data?"

Three model approaches have been applied for the construction of purpose-specific OD matrices with OV-chipkaart data. The rule-based approach does not require additional data and can be applied based on simple assumptions. The trip-based and the tour-based models do require survey data in order to estimate the model parameters that represent the influence of the selected key variables. The tour-based models require more information than the trip-based models, as they incorporate the interaction between consecutive trips within tours. The trip-based approach describes travel between activities at origins and destinations, and thus defines trip-ends by their start and end points. The tour-based approach describes travel from the home-location back home. Hence, tours consist of one or multiple consecutive trips, depending on the number of activities performed. Consequently, trip-ends are classified by their location, either the home-end or activity-end, instead of origins and destinations.



8.3.1 Zonal allocation models

The rule-based zonal allocation allocates origins and destination based on stop locations. This method requires little effort and no additional data, but results in inaccurate allocations to zones in high resolution zonal grids.

The trip-based zonal allocation involves a logit allocation procedure. The available alternatives are selected based on the mode-specific catchment areas. Trips are allocated based on the probabilities of the available zonal alternatives. Separate models were estimated for the origin allocation and the destination allocation, but since 90% of the data consists of tours with two trips, the models are nearly equal. Three attributes proved to have a significant explanatory value to the origin and destination zones:

- The share of the catchment area in the zone;
- The stop density and;
- The level of urbanization.

The latter has a relatively low t-value and the stability analysis over the separate WROOV years showed relatively large deviations in the influence of the urbanization level. However, the level of urbanization was significant in every year and therefore included in the allocation models for both origin and destination.

The tour-based zonal allocation models deviate from the trip-based models in their definition of trip-ends. Due to the geographical distinction of home-ends and activity-ends, the influence of land-use attributes was expected to be greater compared to the trip-based model. However, this hypothesis was not verified, as land-use attributes number of jobs and student places were found to be unstable indicators of the activity zone. The distinction between home and activities instead of origins and destination did indicate the cause of the deviation in influence of the urbanization level in the trip-based models. At the home-end, the influence of urbanization level on the utility of zones tripled, while on the activity-end it was not found to be a stable indicator.

Several causes can be identified for the unforeseen lack of effect of land-use attributes. Primarily, the zonal data relate to the year 2010, the base year of the current VENOM model, while the travel data is collected between 2003 and 2009. Another possible explanation is that the land-use data do not differ enough between the available zones. Since the zones are relatively small, adjacent zones are likely to have similar aggregated land-use characteristics.

The stop-density does prove to be a stable indicator of home and activity zones, although this attribute is only indirectly related to trip production and attraction. Stop locations, and thus the stop density, are adapted to fit the travel demand, but also influence the demand. This interdependency of travel demand and stop density might result in overfitting of the model. Nonetheless, we included the stop density in the models as indicator of nearby activity locations, since its effect is not described by any of the other model attributes. As a result, the fit of the zonal allocation models is optimized, but, at the expense of the model durability.

8.3.2 Purpose inference models

Regarding the purpose inference models, the rule-based approach includes crude simplifications that do result in accurate shares per purpose when accumulating all trips, but at the level of OD pairs, this is not considered a suitable method.



The trip-based purpose inference model includes five trip attributes with specific influences on different purposes: the concessions travelled in, the contract duration, the travel frequency, the used modes and the departure time. The tour-based purpose inference model is similar to the trip-based model, but adds the attribute activity duration. The activity duration has a high explanatory value of the purpose *work*, resulting in a good fit of the tour-based purpose inference model.

The tour-based purpose inference model performs well on the estimation of the travel purposes *work* and *other*. However, the estimation of purposes *shopping* and *education* proves to be more demanding, these purposes cannot be identified with high accuracy by the estimated models. The purpose *education* is in several ways similar to the purpose *work*, but education does not generate as many trips. This complicates the identification of this less-frequent compulsory travel purpose. The same problem arises with the estimation of the purpose *shopping*, which is similar to the more-frequent purpose *other*. These clusters already emerged during the identification of relevant attributes, but with the inclusion of the activity duration and the contract duration a more accurate distinction between *work* and *educational* trips was expected. However, in the current form the purpose inference model does not perform well on the identification of the purposes *education* and *shopping*. The excellent fit statistic is mainly based on the accurate inference of the most frequent purposes *work* and *other*.

The quantitative comparison of key variables in the employed data sources also indicated dissimilarities between the descriptions of both sources regarding several attributes applied in the purpose inference model. Most notably, the distributions of contract duration and travel frequencies do not match. This might result in a model that is over fitted to the survey data and, consequently, overestimates the shares of *work* trips.

The logit purpose inference models do not incorporate land-used characteristics since these did not prove to be stable and significant attributes in the purpose inference. Land-use characteristics are often mentioned as explanatory variables of the travel purpose in the literature, but the most comparable study by Chakirov & Erath (2012) also indicates very little influence of land-use characteristics in a logit allocation model. Possible explanations for the insignificance of land-use attributes are the fact that these do not originate from the same year as the travel data. Furthermore, the land-use characteristics used in the model estimations are the aggregated values around the used stops. Since the actual origins and destinations are not observed in OV-chipkaart data, the aggregated land-use attributes might not comply with the actual values at the origin or destination zone.

Overall, we conclude that the models based on travel characteristics are well fit to distinguish between compulsory and discretionary trips, but do not contain sufficient power to accurately infer the purposes *education* and *shopping*.

8.4 Matrix evaluation

"How do base matrices created by different methods compare to each other?"



The three model approaches have been applied to WROOV data in order to compare the constructed matrices per approach with the observed WROOV matrix. In addition, the models have been applied to the OV-chipkaart data in order to compare the matrices constructed with the different approaches with each other.

8.4.1 Zonal allocation

The assessment of the zonal allocation models consists of the comparison of the constructed total matrices, referring to an average working day, with the observed WROOV matrix.

The comparisons of matrices constructed with WROOV data indicate an inferior fit of the rule-based matrices compared to the trip-based and tour-based matrices, which have similar fit statistics. The trip-based matrix even has a slightly better fit to the observed WROOV matrix than the tour-based matrix, which can be attributed to the inclusion of the attribute urbanization at both trip-ends.

Assessment of the matrices constructed with OV-chipkaart data shows large deviations between the rule-based and the tour-based approach. The matrix constructed with the trip-based models shows higher comparability with the tour-based matrix, but still lower than expected, taking into account their similarity in the zonal allocation.

Evaluating the differences between matrices at lower levels of resolution, we find that the added value of the probabilistic allocation of zones reduces at the level of PC3 zones. Hence, rule based processing of smart card data can provide suitable OD matrices for models with relatively large zones, comparable to PC3 areas. However, transport models with a higher level of resolution, like the VENOM model, are better served with the more deliberate approach of logit allocation models.

The fact that the trip-based and tour-based OV-chipkaart matrices have a relatively poor fit at the PC3 level indicates that these matrices are structurally different. Since the effect of access and egress distances at this level of zonal resolution is small, it can be concluded that the applied methodology of filtering trips with concession traversing transfers has a large effect on the travel demand described by these OD matrices.

8.4.2 Purpose inference

The assessment of the purpose inference models is based on the comparison of purpose-specific matrices. Compared to the generic zonal allocation models, the purpose inference models have an additional assessment option due to their specific alternatives: the comparison of the observed purposes in WROOV data with the probabilities calculated by the logit models. This comparison indicated a slight improvement in the accuracy of the tour-based model compared to the trip-based model.

When comparing the resulting matrices of the two approaches, this improvement is not visible at the level of VENOM zones, where the fit of the tour-based model is similar or poorer than the fit of the trip-based matrices onto the observed WROOV matrices. This can be ascribed to the differences in zonal allocation, which has to be taken into account when comparing OD matrices. However, at the level of PC4 zones, the tour-based matrices do indicate an improved fit with the observed matrices, especially for the purposes work and education.



Overall, the matrix comparisons verify the high accuracy of the inference of the purposes *work* and *other* and the lower accuracy of the inference of *education* and *shopping* purposes. Between the lower scoring purposes, the shopping matrix does indicate a better fit than the education matrix. This leads to the conclusion that education trips have more specific spatial patterns that deviate from commuting patterns, while shopping trips have a similar pattern as trips made for *other* purposes.

8.4.3 Durability of the method

In order to assess the durability of the method, the matrices constructed with WROOV data and OV-chipkaart data were compared. The total matrices are compared on the stop level, constructed with the rule-based approach, to eliminate the effect of access and egress legs and ensure an unbiased comparison. For the purpose-specific matrices, the comparison is made on the tour-based matrices, since this approach is the most accurate. The assessment based on linear regression shows a relatively poor fit, indicating that the described travel demand does not match. Especially, the work matrices have low comparability, which can be explained by the over-representation of work trips in the WROOV data.

While the influence of changes in travel demand may be small over time, the differences between the sources indicate that the models estimated on WROOV data have limited durability. The method of construction is renewable, but requires periodic updating of the survey data to re-appraise the model parameters.

8.5 Answer to the main research question

"To what extent can the travel purpose, origins and destinations of public transport trips derived from smart card data be inferred based on information from survey data, in order to construct purpose specific OD matrices suitable as base matrices in transport models?"

Accumulating the answers of the sub-questions, it can be concluded that the construction of purpose-specific OD matrices based on survey data generates added value to the representation of travel demand by OD matrices. The probabilistic allocation procedure outperforms the rule-based approach in the allocation to origins and destination zones. Especially on higher levels of resolution, like the zonal grid of the VENOM model, the improvement of the allocations is substantial. At a lower level of resolution, like the PC3 level, the influence of the distances covered during access and egress legs is reduced by the larger internal distances of the zones. Nonetheless, for regional transport models like the VENOM model, the logit zonal allocation method results in a significant improvement of the description of the travel demand by base matrices compared to a direct conversion of stops to zones.

In addition to the increased accuracy of the origins and destinations, the purpose inference based on survey data allows for an accurate distinction of trips between the purposes *work* and *other*. The less appearing purposes *education* and *shopping* are closely related to, respectively, the purposes *work* and *other*. Therefore the inference of these purposes is less accurate. Combining these two clusters does provide an accurate distinction between compulsory purposes and discretionary purposes. The attributes in survey data required to perform this method of enrichment include the home and activity zones, the used stops and modes, the departure times, the activity duration, the contract duration and the travel frequency.



The augmentation of the trip-based approach into the tour-based approach resulted in a more accurate inference of the travel purpose due to the inclusion of the attribute activity duration. In addition, the tour-based trip-end distinction, by home and activity side, resulted in increased effect of the level of urbanization at the home-end. Consequently, allocation of home zones is more accurate than the trip-based allocation to origins and destinations. On the other hand, the tour-based approach did not improve the allocation to activity zones, since land-use attributes jobs and student places did not prove to be stable indicators of activity zones. Overall, the tour-based contains more behavioural richness and is, therefore, preferable over the trip-based approach.

Due to limitations of the available data, the constructed purpose-specific OD matrices cannot readily be applied as base matrices in transport models. In order to realise comparable datasets, students and international tourists have been filtered from the OV-chipkaart data, since the WROOV survey does not cover these travellers. These travellers make up nearly 30% of the total travellers and therefore have to be considered in the assessment of the total public transport travel demand.

Another prominent limitation of this study is the unavailability of the OV-chipkaart data from adjacent concessions, resulting in unobserved transfers. The influence of these transfers on the OD matrix is substantial, since the origin or destination zone is not near the observed stop and many travellers transfer to the train network. This limitation is non-existent when data from all adjacent operators is available.

The applied methodology of constructing purpose-specific OD matrices based on smart card data shows great potential. During this research, an operational method has been built, which results in a more accurate description of public transport travel demand than previously available. However, restrictions of the available data and imperfections in the application have resulted in limited applicability and durability. With additional fine-tuning of this method and increased availability of the data, the method presented in this report can be enhanced to a fully applicable approach that can lead valuable improvements of the quality of public transport demand forecasts.



9 Recommendations

Based on the conclusions presented in the previous chapter, this chapter provides the recommendations to the various parties involved in this study. Two categories of recommendations are distinguished. First, recommendations for follow-up research are provided (paragraph 9.1). Subsequently, recommendations for the utilization of the results of this study are listed (paragraph 9.2).

9.1 Follow-up research

The recommendations for follow-up research are categorized by three subjects:

- Enhancements to the enrichment models;
- Expanding the method to obtain complete coverage of the public transport system;
- Related research topics that can build on this study.

9.1.1 Enhancements to the enrichment models

Reflecting on the applied methodology, we have found several issues that leave room for improvement. The enhancements are categorized into three subjects:

- Zonal alternative selection by means of catchment areas;
- Combined allocation of activity zones and corresponding purposes;
- Evaluation of the matrices.

Zonal alternative selection by means of catchment areas

The mode-specific catchment areas capture the main influence of the level of service at stops, but the distances covered during access and egress transport still show high deviation between stops of the same mode. We believe that the selection of zone alternatives in the logit zonal allocation can be improved by incorporating additional attributes related to the level of service: the frequencies and operational speeds at stops, instead of the indirect indication of the level of service via the mode. In addition, the relative location of the used stop to the locations of other stops on the same line provides additional insight in the access and egress behaviour. Adding these attributes in stop-specific catchment areas will improve the identification of the zonal alternatives and, consequently, the quality of the zonal allocation.

Combined allocation of activity zones and corresponding purposes

Regarding the probabilistic allocation of trips to the available zones, the inclusion of the stop density in the zonal allocation models should be re-assessed. Because its relative influence on the utility of alternatives is very large compared to its influence on the model fit, the inclusion of this attribute in the zonal allocation models is questionable. It is recommended to extend the search for alternative attributes that relate more directly to the trip production and attraction of the zones.

We recommend directing this search at magnification of the differences in land-use characteristics between adjacent zones. This can be achieved by subtracting the mean from the attribute values. In addition, the zonal allocation models could be estimated specifically per concession. The zonal structure in the highly urban concession Amsterdam has a higher resolution than the zonal structure in the more rural concession Waterland. Moreover, the mean values of land-use characteristics differ between Amsterdam and Waterland.



Another search direction comprises assessing the influence of the mismatch between the collection period of the travel data (2003-2009) and the reference year of the zonal data (2010). This can be done by selecting the WROOV data from the year 2004 and estimating the zonal allocation models with land-use data from 2004, which is the previous base year of the VENOM model. However, since the changes in land-use characteristics over this period are expected to be small, we recommend focussing on the magnification of differences in land-use characteristics.

Since the access and egress distances were found to be correlated to the travel purpose, we recommend estimating the activity zone with the corresponding purpose together in one model. The attributes jobs and student places directly refer to a travel purpose, respectively work and education. In the distinct purpose-inference models, these attributes did not prove valuable indicators, partly because the land-use data used did not necessarily comply with the origin or destination zone. Moreover, the presence of jobs in a zone might reduce the chance of other activities occurring in that same zone. These two issues can be solved by joining the purpose inference with the activity zone allocation. Therefore, the combined allocation of the activity zone and the purpose will have higher chance of indicating significant and stable relations. The model structure of such an integrated allocation has been set-up and successfully tested, but the optimization is complex and time-consuming.

In addition, the implementation of the attribute activity duration as categorical variable might increase its explanatory value for the purpose inference. In the current tour-based purpose inference model, the activity duration is implemented as a continuous variable, but very long activities are strongly related to the purpose work, diminishing the explanatory value of the activity duration for the purpose education.

Concluding, it is recommended to perform a re-appraisal of the home zone allocation model parameters and estimate a combined activity zone and purpose allocation model, based on a new survey that complies with the OV-chipkaart data sample.

With the knowledge acquired during this research, which consisted of many data handling procedures and conceptual ideas, the next optimisation round will require substantially less effort.

Evaluation of base matrices

Regarding the evaluation of the purpose-specific OD matrices, it is recommended to extend the evaluation by means of linear regression. We recommend assessing the comparability with the network performance. By assigning the resulting matrices to the network with a route-choice model, a direct comparison can be made with the network performance observed by OV-chipkaart data. This evaluation provides additional insight in the influence of differences between matrices on the network performance.

9.1.2 Complete coverage of the public transport system

In order to obtain complete coverage of the public transport system, the trips that have been filtered from the OV-chipkaart data, in order to comply with the coverage of the WROOV surveys, have to be taken into account. These trips include trips made by students and international tourists. In order to take into account the entire public transport system, we recommend initiating a new survey, in connection with OV-chipkaart data. Currently, travellers can review the transactions made with their OV-



chipkaart in an online overview¹³. Moreover, it is possible to add missed check-in and check-out transactions to this overview in order to reclaim part of the entry fare. In a similar way, travellers could supplement their travel purpose, origins and destinations. This requires very little time and effort and thus results in a low respondent burden. Moreover, such an online survey reduces processing times and costs.

A new survey also enables compatibility between the survey data and the smart card data. The WROOV surveys have provided a very large sample with all the information required for the enrichment of OV-chipkaart data. However, the travel patterns described by the WROOV data do not entirely correspond with the patterns described by the OV-chipkaart. This problem can only be solved by estimating the enrichment models on a representative sample of the available OV-chipkaart data.

Furthermore, renewal of the survey data is required at some point in time due to the decreasing value of information over time. Changes in the public transport system, concerning travel behaviour, demand and supply, develop slowly, but have to be taken into account. Since the WROOV surveys have been terminated after 2009, the durability of the estimated models is limited. Therefore, we recommend reassessment of the models with up-to-date, OV-chipkaart compatible survey data that cover the complete public transport system.

Similarly to survey data, smart card data require complete coverage for the construction of OD matrices. The problem with coverage of the OV-chipkaart is related to the fragmentation of data at different operators, serving specific public transport concessions. In order to construct fully applicable public transport OD matrices, the OV-chipkaart data is required from all operators with services in the study area. For the case study, these include additional data from Connexxion of the concessions Amstelland-Meerlanden and Zaanstreek and the data from NS regarding the national railways. It is advisable for public transport authorities to regulate the availability of OV-chipkaart data for research purposes, like this study, in the concessionary conditions. This case study has shown that the data from different operators can be coupled in order to obtain insight in the travel demand in the complete public transport system, beyond the structural boundaries of concessions. A complete overview of the transport demand generates added value for both public transport authorities and operators.

9.1.3 Related research topics

Smart card data provide many other opportunities for the improvement of public transport travel demand modelling than the construction of OD matrices. This study was focussed on the construction of OD matrices as first step of the incorporation of this new data source in transport modelling.

In the applied method, the construction of the OD matrix is completely separated from the construction of synthetic OD matrices by the model. This approach can also be turned around by calibrating the synthetic model based on the observed travel demand in OV-chipkaart data.

Another topic is the calibration of travel supply modelling by route choice models. Since the OV-chipkaart data also contain the exact route travelled, route choice models can be calibrated by assignment of the constructed OD matrices onto the network. This is related to the evaluation of the matrices by means of the network

¹³ Available via Mijn OV-chipkaart at www.ov-chipkaart.nl.



loads. When both the matrices and the route choice models have not been validated, it cannot be determined which of the two is related to potential dissimilarities. Therefore, we recommend first to optimize the construction of OD matrices, before calibrating route-choice models. Nonetheless, route choice model calibration is a very promising research topic that could build on this study.

9.2 Utilization of the results

When completed with the travel demand of students and international tourists, the constructed purpose-specific OD matrices can be used in travel demand studies in the VENOM study area, for example the influence of the new metro line in Amsterdam, leading to an improved description of the travel demand in the current situation. After implementation in the VENOM model, this will also lead to more accurate forecasts of travel demand in the forecast year.

Both the public transport authority SRA and the operators GVB and EBS can benefit from the increased insight in the relation of travel demand between of the respective concessions. By combining their OV-chipkaart data, all three parties have acquired insight in the previously concealed travel demand of beyond the borders of individual public transport concessions.

With the continuous flow of OV-chipkaart data, it is recommended to update the base year of the public transport OD matrices more frequently. The current renewing cycle of approximately five years can result in significant differences in travel demand, while an update based on OV-chipkaart data only requires a new application procedure of the enrichments models.

Moreover, the longitudinal character of the data collection can provide increased insight in the relation between average travel demand and peak demand. This can be done at two levels: the representation of the average working day and the diffusion of demand over the day. The analysis of the model on the average working day can be extended to a bandwidth analysis by representing quiet working days and busy working days, based on the yearly average.

The construction of purpose specific OD matrices with OV-chipkaart data allows for the specification of matrices by any desired specification of time. Trip generation and trip distribution models estimate synthetic matrices for a specific speak period or for 24 hours. Matrices constructed with OV-chipkaart data allow for selections based on the time-stamp, directly available from the data. Thereby, the construction of OD matrices with OV-chipkaart data provides increased insight in the distribution of the travel demand over the day.

The comprehension of travellers' motivations to travel provides valuable information for the operators that can be used in their policies of fare schemes and ticketing systems. However, it has to be taken into account that the estimation was based on survey data with higher shares of long term contracts, and thus higher shares of commuting travellers. In addition, it has to be noted that the logit allocation to zones still incorporates substantial deviations for individual WROOV matrix cells at the VENOM level, although it describes the travel demand significantly better than matrices constructed with the rule-based approach.



Bibliography

- Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *12th IFAC Symposium on Information Control Problems in Manufacturing*. Saint-Etienne: INCOM.
- Agard, B., Morency, C., & Trépanier, M. (2009). *Mining Smart Card Data from an Urban Transit Network*. Montréal: École Polytechnique de Montréal, Canada.
- Allos, A., Merrall, A., Smithies, R., & Fishburn, R. (2014). New data sources and data fusion. *European Transport Conference*. Frankfurt: AET.
- Alshalalfah, B., & Shalaby, A. (2007). Case Study: Relationship of Walk Access Distance to Transit with Service, Travel, and Personal Characteristics. *Journal of Urban Planning and Development*, 133(2), 114-118.
- Axhausen, K., Schönfelder, S., Wolf, J., Oliveira, M., & Samaga, U. (2003). 80 weeks of GPS traces: Approaches to enriching the trip information. *Institut für Verkehrsplanung und Transportsysteme*. Zürich: ETH.
- Bagchi, M., & White, P. (2005). The potential of public transport smart card data. *Transport Policy*, 12(5), 464-474.
- Barry, J. J., Freimer, R., & Slavin, H. (2009). Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. *Transportation Research Record: Journal of the Transportation Research Board*, 2112(1), 53-61.
- Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*. Cambridge: MIT press.
- Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models. *3rd STRC Swiss Transport Research Conference*. Ascona: STRC.
- Blythe, P. T. (2004). Improving public transport ticketing through smart cards. *Proceedings of the ICE - Municipal Engineer*, 157(1), 47-54.
- Blythe, P. T., & Holland, R. (1998). Integrated ticketing - Smart cards in transport. *IEE Colloquium: Using ITS in Public Transport and in Emergency Services*. London: IEE.
- Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285-297.
- Chakirov, A., & Erath, A. (2012). Activity Identification and primary location modelling based on Smart Card payment data for Public Transport. *13th International Conference on Travel Behaviour Research*. Toronto.
- Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice*, 44(10), 830-840.
- Chorus, C. G., Arentze, T. A., & Timmermans, H. J. (2008). A random regret-minimization model of travel choice. *Transportation Research Part B: Methodological*, 42(1), 1-18.
- Chu, K. K., & Chapleau, R. (2010). Augmenting transit trip characterization and travel behavior comprehension. *Transportation Research Record: Journal of the Transportation Research Board*, 2183(1), 29-40.
- Chu, K., & Chapleau, R. (2007). Imputation techniques for missing fields and implausible values in public transit smart card data. *11th World Conference on Transport Research*.
- Chu, K., & Chapleau, R. (2008). Enriching Archived Smart Card Transaction Data for Transit Demand Modelling. *Transportation Research Record*, 63-72.



- Cui, A., Wilson, N., & Attanucci, J. (2006). *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*. Cambridge: Massachusetts Institute of Technology.
- Deng, Z., & Ji, M. (2010). Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach. *Traffic and Transportation Studies*, 768-777.
- Department for Transport. (2014). *Transport Analysis Guidance unit M3.1 Highway Assignment modelling*. Londen: Department for Transport.
- Devillaine, F., Munizaga, M. A., & Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, 2276(1), 48-55.
- Dinant, J. M., & Keuleers, E. (2004). Data protection: Multi-application smart cards: the use of global unique identifiers for cross-profiling purposes-Part II: towards a privacy enhancing smart card engineering. *Computer Law & Security Review*, 20(1), 22-28.
- Djukic, T., Hoogendoorn, S., & Lint, H. V. (2013). Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index. *Transportation Research Board 92nd Annual Meeting* (pp. 13-4851). Washington DC: Transportation Research Board.
- Farzin, J. M. (2008). Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil. *TRB Annual Meeting CD-ROM*.
- Goeverden, C. v. (n.d.). *Multimodality in the Netherlands 2004 - 2009*. Delft: KIM.
- Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia-Social and Behavioral Sciences*, 138, 557-565.
- Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H., & Attanucci, J. P. (2013). Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2343(1), 17-24.
- Hofmann, M., & O'Mahony, M. (2005). Transfer journey identification and analyses from electronic fare collection data. *Intelligent Transportation Systems* (pp. 34-39). IEEE.
- Hofmann, M., Wilson, S. P., & White, P. (2009). Automated identification of linked trips at trip level using electronic fare collection data. *TRB Annual Meeting CD-ROM*.
- Jang, W. (2010). Travel time and transfer analysis using transit smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, 2144(1), 142-149.
- Jun, C., & Dongyuan, Y. (2013). Estimating Smart Card Commuters Origin-Destination Distribution Based on APTS Data. *Journal of Transportation Systems Engineering and Information Technology*, 13(4), 47-53.
- Kieft, S., Herder, J., & Pieters, M. (2013). Openbaar Vervoer Matrices in VENOM. *Colloquium Vervoersplanologisch Speurwerk (CVS)*. Rotterdam.
- Kieft, S., Linden, T. v., Bedem, J. v., & Scholten, M. (2014a). *VENOM2013 Basismatrices 2010*. Amsterdam: City Region of Amsterdam (SRA).
- Kieft, S., Linden, v. d., Bedem, v. d., & Scholten, M. (2014b). *VENOM2013 Basisprognoses 2030*. Amsterdam: Stadsregio Amsterdam (SRA).
- Kim, K. (2014). Discrepancy Analysis of Activity Sequences. *Transportation Research Record: Journal of the Transportation Research Board*, 2413(1), 24-33.



- Kim, K., Oh, K., Lee, Y. K., Kim, S., & Jung, J. Y. (2014). An analysis on movement patterns between zones using smart card data in subway networks. *International Journal of Geographical Information Science*, 28(9), 1781-1801.
- Kuhlman, W. (2014). *Onderzoek verrijking OV chipkaart data met informatie uit de WROOV onderzoeken*. Zoetermeer: Panteia.
- Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, 46, 179-191.
- Lee, S. G., & Hickman, M. (2013). Are Transit Trips Symmetrical in Time and Space? *Transportation Research Record: Journal of the Transportation Research Board*, 2382(1), 173-180.
- Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, 6(1-2), 1-20.
- Lee, S., Hickman, M., & Tong, D. (2012). Stop Aggregation Model: Development and application. *Transportation Research Record: Journal of the Transportation Research Board*, 2276(1), 38-47.
- Lianfu, Z., Shuzhi, Z., Yonggang, Z., & Ziyin, Z. (2007). Study on the method of constructing bus stops OD matrix based on IC card data. *Wireless Communications, Networking and Mobile Computing*.
- Liao, C. F., & Liu, H. X. (2010). Development of Data-Processing Framework for Transit Performance Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2143(1), 34-43.
- Ma, L. C., Banerjee, P., Lai, J. H., & Shroff, R. H. (2008). Diffusion of the 'Octopus' Smart Card E-payment System: A Business and Technology Alignment Perspective. *International Journal of Business and Information*, 3(1).
- Ma, X.-I., Wu, Y.-J., Wang, Y.-h., Chen, F., & Liu, J.-f. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C*, 36(1), 1-12.
- McGowen, P., & McNally, M. (2007). Evaluating the potential to predict activity types from GPS and GIS data. Washington: Transportation Research Board 86th Annual Meeting.
- Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. W. (2014). Trip Purpose Identification from GPS Tracks . *Transportation Research Record: Journal of the Transportation Research Board*, 2405(1), 16-23.
- Morency, C., Trepanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3), 193-203.
- Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24, 9-18.
- Munizaga, M. A., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, 44, 70-79.
- Nassir, N., Khani, A., Lee, S. G., Noh, H., & Hickman, M. (2011). Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. *Transportation Research Record: Journal of the Transportation Research Board*, 2263(1), 140-150.
- Nes, R. v. (2002). *Design of multimodal transport networks: A hierarchical approach*. Delft: Delft University Press.
- Nijenstein, S., & Bussink, B. (2014). Verkenning kwaliteitsverbetering OV met multimodale OV-chipkaart data (in Dutch). Eindhoven: CVS.
- Oort, N. v., Drost, M., & Brand, T. (2014). Betere OV prognoses met anonieme OV-chipkaartdata (in Dutch). Eindhoven: CVS.



- Ordóñez, S. A., & Erath, A. (2013). Estimating Dynamic Workplace Capacities by Means of Public Transport Smart Card Data and Household Travel Survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, 2344, 20-30.
- Ortúzar, J. d., & Willumsen, L. G. (2011). *Modelling transport - fourth edition*. Chichester: Wiley.
- Pelletier, M., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C*, 19(4), 557-568.
- Pollard, T., Taylor, N., & van Vuren, T. (2013). Comparing the Quality of OD Matrices in Time and Between Data Sources. Frankfurt: The Association for European Transport.
- Reddy, A., Lu, A., Kumar, S., Bashmakov, V., & Rudenko, S. (2009). Entry-Only Automated Fare-Collection System Data Used to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles. *Transportation Research Record: Journal of the Transportation Research Board*, 2110(1), 128-136.
- Rijkswaterstaat. (2012). *Documentatie Groeimodel 2011 Deel 1*. Rijkswaterstaat.
- Robinson, S., Narayanan, B., Toh, N., & Pereira, F. (2014). Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies*, 49(1), 43-58.
- Seaborn, C., Attanucci, J., & Wilson, N. H. (2009). Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record: Journal of the Transportation Research Board*, 2121(1), 55-62.
- Shen, L., & Stopher, P. R. (2013). A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies*, 36, 261-267.
- Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, 16(3), 350-369.
- Train, K. E. (2009). *Discrete choice methods with simulation* (2nd ed.). New York: Cambridge university press.
- Translink. (2015). *Jaaroverzicht Translink 2014*. Utrecht.
- Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*, 11(1), 1-14.
- Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *TRB Annual Meeting CD-ROM*.
- Wang, W., Attanucci, J. P., & Wilson, N. H. (2011). Bus Passenger Origin-Destination Estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, 14(4).
- Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board*, 1768(1), 125-134.
- Yue, Y., Lan, T., Yeh, A. G., & Li, Q. Q. (2014). Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*, 1(2), 69-78.
- Zhao, J., Rahbee, A., & Wilson, N. H. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5), 376-387.



Zhou, J., Murphy, E., & Long, Y. (2014). Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. *Journal of Transport Geography*, 41(1), 175-183.





A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow

Eduardo Graells-Garrido
Telefónica I+D
Santiago, Chile

Diego Saez-Trumper
EURECAT
Barcelona, Spain

ABSTRACT

Nowadays, travel surveys provide rich information about urban mobility and commuting patterns. But, at the same time, they have drawbacks: they are static pictures of a dynamic phenomena, are expensive to make, and take prolonged periods of time to finish. However, the availability of mobile usage data (*Call Detail Records*) makes the study of urban mobility possible at levels not known before. This has been done in the past with good results—mobile data makes possible to find and understand aggregated mobility patterns. In this paper, we propose to analyze mobile data at individual level by estimating *daily journeys*, and use those journeys to build Origin-Destiny matrices to understand urban flow. We evaluate this approach with large anonymized CDRs from Santiago, Chile, and find that our method has a high correlation ($\rho = 0.89$) with the current travel survey, and that it captures external anomalies in daily travel patterns, making our method suitable for inclusion into urban computing applications.

1. INTRODUCTION

Travel surveys provide information about urban mobility and commuting patterns, mainly through Origin-Destiny matrices derived from them. These matrices allow urban planners and policy makers to understand travel patterns in urban mobility. Such surveys are usually collected once per decade, due to their expensiveness (both in time and monetary costs). Moreover, they are limited in some ways: travel surveys are static pictures of a dynamic phenomena and, due to its sample size, they are limited to big areas (either administrative or designed).

In this paper, we propose to use mobile data used for billing, which indicates a subset of the antennas a mobile device has connected through the day, as well as the corresponding timestamps. We use these digital footprints to build extended travel diaries. Travel diaries are the basic elements of travel surveys, but we extend them through daily journeys, as we detect not only trips, but also other “non-trip” activities. From these daily journeys we build Origin-Destiny (OD) matrices, at a fraction of the expenses needed to build OD matrices from travel surveys.

Our main contribution is a method to detect these disaggregated daily journeys using *Call Detail Records* (CDRs). Our approach is based on *graphical timelines* and computational geometry algorithms, which are applied having in mind transport-based rules regarding trip duration and distance. We use an anonymized CDR dataset from one of the largest telecommunications company in Chile, with a market share of 38.18% as of June 2015. Chile, being one of the developing countries with highest mobile phone penetration, is a good candidate for analyzing urban mobility using CDRs—for instance, there are 132 mobile subscriptions per 100 inhabitants.¹ Particularly, we focus on Santiago, its capital and most populated city.

To evaluate our results, we compare our predicted urban flow (in the form of an OD matrix) with the last travel survey for Santiago, performed during 2012–2013. In terms of OD pairs, we obtain a very high correlation ($\rho = 0.89$), indicating that our method recognizes the urban flow on the city. We apply our methods at different days, and find that, in addition, we detect how urban flows change in the presence of unexpected conditions. This, jointly with the disaggregated nature of our method, has potential for applications in urban computing [13], discussed at the end of this paper.

2. RELATED WORK

The estimation of OD matrices, and thus, urban mobility patterns, is not new. The most basic way of estimating those patterns is by travel surveys, but also other methods have been developed. A common method is based on traffic counts [2], but the massive availability of other kinds of information has allowed to estimate such matrices in other ways, for instance, by using smart-card passive data [9] and, as in this paper, mobile data [1, 4, 5, 7, 11].

The work on mobile datasets has included both theory and practice. From a theoretical point of view, the analysis of how predictable humans are [5, 11]. A practical work has been the prediction of transient OD matrices by analyzing transitions of connections between cell antennas. These transitions are the basis of many methods, which employ techniques from optimization and temporal association rules [4], to transport-based rules to accept or discard transitions [1, 7]. Our approach is also based on antenna transitions. However, the main difference with previous work is that we reconstruct individual *daily journeys*, an extended version of the travel diaries used to build travel surveys. These journeys are built using a geometric approach based on *graphical timelines* [12], which are time-space diagrams displaying timelines according to time and distance covered. On these timelines we estimate the “turning points” of the daily journey, which serve to differentiate daily activities into trips and non-trips. These timetables are regularly used in transport to

¹<http://www.subtel.gob.cl/estudios-y-estadisticas/telefonia/>

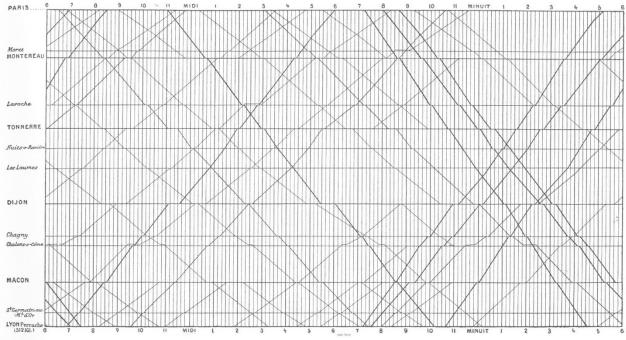


Figure 1: Graphical timetable of a train schedule, by E.J. Marey, 1885.

display and analyze schedules, as well as phenomena associated with vehicle behavior. For instance, Figure 1 shows a train schedule designed in 1885 by Étienne-Jules Marey. On it, passengers can see arrival and departure times, as well as the duration of stops and the velocity of trains. Transport planners can study vehicle behavior—since all vehicles share the same origin, if we were visualizing their true trajectories instead of the scheduled ones, *vehicle bunching* can be recognized immediately [8].

3. METHODS

Our methods consider *mobile user traces* estimated from anonymized CDR data. CDRs are comprised of logged data extracted from cell-phone antennas, and are used to bill customers. They contain the following events: *calls*, *SMS* (text messages), and *data events* (triggered every 15 Megabytes or every 15 minutes of an active connection). The following features are common between all events, and thus are considered for analysis: the anonymized user ID, the antenna ID, and time of the day. The antenna ID is used to determine a likely position for the user at the corresponding time of the day. Note that this position is assumed to be the same for all phones connected to the same antenna, *i.e.*, we do not perform triangulation based on signal strength with nearby antennas.

Problem Definition and Proposal. The problem we propose to solve is the estimation of a diary of activities for a day, using CDRs from mobile data. A daily journey J for user u is defined as follows:

$$J_u = \{(A_i, (t_{iO}, t_{iD}), (p_{iO}, p_{iD}))\}.$$

Where A_i is an activity, p_i (and t_i) are the positions (and times) associated to the start/origin and end/destination of A_i . Activities can be of types *trip*, *non-trip*, and *unknown*.

To solve this problem we propose a two-step algorithm: first, we define the candidate turning points of a day, where turning point is a moment in the day, at a specific position, where the user started to perform an activity (and, by definition, ends performing a previous activity). The second step is to detect activities based on the candidate turning points.

Graphical Timetables and Turning Points. For all users, we build the following vector:

$$\vec{u} = [(t_0, p_0), (t_1, p_1), \dots, (t_n, p_n)],$$

Where each element in \vec{u} corresponds to an event in the CDR events of u in a day, with the corresponding timestamp t and the antenna position p . These vectors can be transformed into graphical timetables (see Figure 1), where the x-axis is the elapsed time during the day, and the y-axis is the traveled (accumulated) distance from

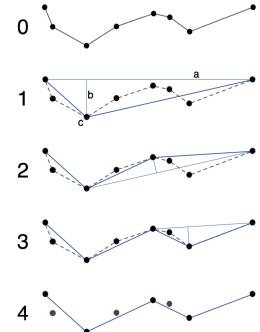


Figure 2: Illustration of the Ramer-Douglas-Peucker algorithm. Source: Wikipedia Commons.

the starting point. The different groups of segments present in the timetable define the activities in J_u .

However, due to the potential amount of CDR events, several contiguous segments could be related to the same activity. To tackle this, we propose to simplify the timetable using the Ramer-Douglas-Peucker line simplification algorithm [3]. This algorithm, depicted on Figure 2, starts with the extremes of the trajectory, and iteratively adds vertices according to their distance to the candidate simplified line. The vertex with greater distance (or error) is added, and the process is repeated until the error is less than a given tolerance. This results in a simplified vector \vec{u}_s , which contains candidate “turning points” of daily activity. Those candidate points define the potential activity segments that will end defining J_u .

Activity Classification. The second step is to identify which contiguous segments in the line defined by the points of \vec{u}_s can be merged into a single activity A_i . The most direct approach is to merge a series of horizontal (or nearly horizontal) segments. However, a continuously moving user (and thus, with non-horizontal segments) is also feasible. Take the example of a taxi driver, whose primary activity during the day is working as a driver. Although these activities are valid, they are not relevant for an OD matrix, as they are not trips in the individual sense. Thus, we define the following classifications: *unknown*, *non-trips*, and *trips*.

Unknown are segments where the total distance covered is greater than 100 kilometers. In those cases we cannot distinguish between trips and unknown situations. For instance, the mobile number is associated to a vehicle (*e.g.*, a taxi) or another vehicle/device. While these are indeed displacements in time and space, they do not fall on the daily journey concept we study in this paper. Another case is when users switch to WiFi networks, and thus disappear from the event log for a long period of time. For instance, users who switch to the wireless network of their work place might not be detected there, and the next logged event could appear after working hours. On those cases, we cannot identify trips reliably. To avoid this scenario, in some cases the displacement between events has been limited to specific time windows (*e.g.*, 10 minutes and 1 hour [7]).

Non-trips are segments that are not *unknown*. The criteria to assign this type to an activity is based on the covered distance and time. On the one hand, the antenna density of origin/destiny locations of the activity is considered to determine a minimum distance that cannot be attributed to signalling changes (this is discussed further in the next section). On the other hand, some activities involve displacements (*e.g.*, working/studying in a big campus), but the speed of movement is much slower than when performing a trip. Thus, if there is a distance displacement, but the time is greater than 180 minutes, we still consider a non-trip activity.

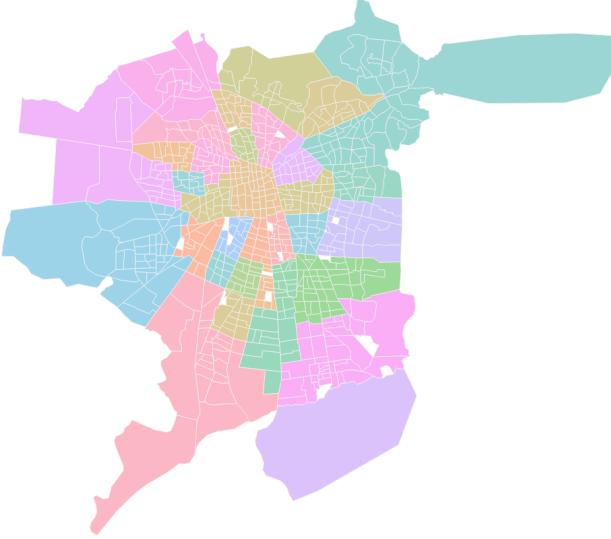


Figure 3: The 752 zones from the OD Survey zoning. Colors indicate the 35 municipalities under consideration.

Finally, *trips* are segments that are not *unknown* and are not *non-trip*, but, in addition to the 180 minute rule, we enforce a minimum duration of 15 minutes for a segment, due to the granularity of the CDR data. Any other activity that does not fall into the rules defined above is considered *unknown*.

Activity Merging. Having assigned an activity to each segment built from \vec{u}_s , we proceed to merge contiguous segments that have the same activity. Two or more segments are merged into an activity A_i by considering the first time and position in the segment as origin, and the last time and position in the segment as destination. We also consider an additional case: when two *trip* activities surround a *non-trip* activity, and the duration of the latter is lesser or equal than 15 minutes, its activity is changed to *trip*. This scenario corresponds to situations when users in public transport make a connection, or users in vehicles face congested traffic. In this way, after merging all activities, the daily journey J_u is built.

4. CONTEXT AND DATASET

We work with a dataset from Santiago, the capital of Chile. Santiago is a city with almost 8 million inhabitants, and it has an integrated public transport system named Transantiago. The Metropolitan Area of Santiago is composed of 35 independent administrative units named municipalities. We work with this set of municipalities, depicted on Figure 3.

Santiago 2012 Travel (OD) Survey. The Santiago 2012 travel survey (*ODS* hereafter) was performed during 2012–2013.² It took almost one year to finish, and contains 96,013 trips (from 40,889 users). The information of trips is obtained through the travel diaries fulfilled by the surveyed persons. The ODS, used to define public policy related to public and private transport in the city, as well as general urban mobility, is performed every 10 years due to its costs and its difficulty.

The survey considers other municipalities outside the area, as well as cities in other regions, due to the characteristics of the survey procedure. Additionally, the survey also defines a zoning of the city, with 752 zones within the considered municipalities. Each zone

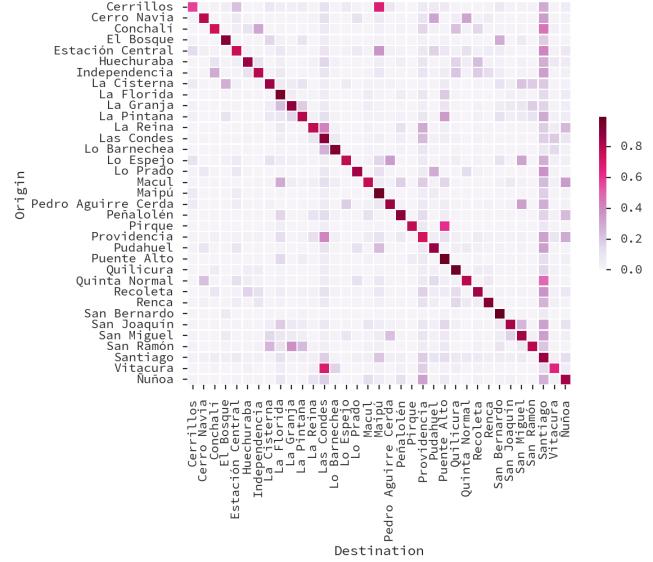


Figure 4: Distributions of OD Survey 2012 trip data from Santiago, Chile. The matrix has been L2 normalized on rows (origins).

intends to control for land-use and population density. Even though each trip in the survey is associated to a zone and a municipality, the survey is only representative at municipal level. At its current granularity, the data available is not enough to calculate reliable mobility patterns at zone level.

In this paper, without losing generality, we focus on the 51,819 trips performed on working days, from 22,541 users, performed on the inner 35 municipalities of the Metropolitan Area. We aggregated trips according to municipality into an OD matrix, shown in Figure 4. One can see that the most common trips are intra-municipalities, but there are still municipalities that tend to receive more trips than others. This is because most of the commercial and working land-use is on the municipalities of *Santiago*, *Providencia*, *Las Condes* and *Vitacura*. Note that the municipality of *Santiago* is at the center of the Santiago Metropolitan Area. In the rest of this paper, when we mention *Santiago* we refer to the Metropolitan Area.

In terms of trip variables, Figure 5 shows the distributions of trip start time, trip duration, and approximated trip distance (*i.e.*, euclidean distance). One can see that trip start time follows an expected pattern of two high peaks (one in the morning and one in the afternoon), with a third smaller peak at lunch time. With respect to trip duration, the mean duration is 41 minutes. Note that the self-reported nature of the survey is evident, due to the several peaks present on the distribution in 15 minute periods (*e.g.*, 30 and 45 minutes). Finally, with respect to travel distance, the mean euclidean distance is 6.05 kilometers.

Mobility Data. We analyze mobile data using CDRs from one of the largest telecommunications company in Chile. The CDR data contains events for all Mondays and Tuesdays of June 2015.

In the 35 municipalities under consideration, the company has 12,936 antennas, with 98% of the zones having at least one antenna. Figure 6 displays the antenna territorial density. Note that the antenna distribution is not homogeneous on the city, nor at any level. For instance, antenna distribution is correlated with the ODS, considering aggregated destinations at municipal level ($\rho = 0.91$, $p < 0.001$).

To avoid artifacts in the trip detection introduced by connectivity changes due to signal strength (*i.e.*, mobile phones looking for the best signal) and signal balancing (*i.e.*, mobile antennas pointing

²<http://www.sectra.gob.cl/biblioteca/detalle1.asp?mf=3253>.

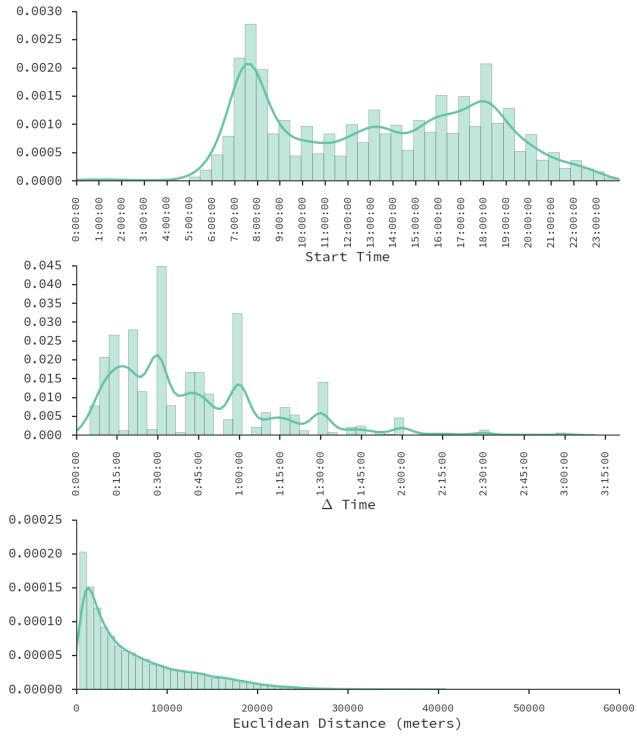


Figure 5: Distributions of EOD Survey 2012 trip data.

devices to connect to other antennas due to saturation), we estimated a distance matrix D_{z_i} for all antennas in each zone. Then, we defined that the minimum distance for an activity to be considered a trip between zones z_o and z_d is:

$$d_{\min} = \max\{Q(D_{z_o}), Q(D_{z_d})\}.$$

Where Q is the quantile function (by manual experimentation we have found that 0.8 is a good value). The mean value of $Q(D_z)$ is 732 meters (minimum of 45 m., and maximum of 14.1 km.).

While we do not disclose the exact number of users in each day due to confidentiality and commercial issues, Figure 7 displays the distributions of event frequency and hourly entropy for all days in the dataset. In the left chart, each dot is a minute in a specific day. The frequency encodes the fraction of events that the dot contains per day. One can see that the distribution of frequency of events can be approximated by a cubic linear regression, with a higher frequency of events in the afternoon.

The right chart of Figure 7 displays the distribution of user entropy with respect to hours of the day, for each day. This entropy is defined as the Shannon entropy of user u :

$$H_u = - \sum p_{i,u} \ln p_{i,u}.$$

Where $p_{i,u}$ is the probability that user u has a CDR event in the i th hour of the day. The purpose of estimating this entropy is to have a measure of diversity with respect to time for each user. Thus, we discard users who do not have enough diversity to be able to estimate their journeys, or that have too much diversity to be normal users (*e.g.*, they could be SIM cards associated to machines). We discarded users in the first quartile ($H_u < 0.4$) and in the last decile ($H_u > 0.9$).

We apply our methods to this dataset on the following section. To be able to compare with the ODS, we employ a similar sample size of $N = 100,000$ randomly selected mobile users (before filtering by entropy).

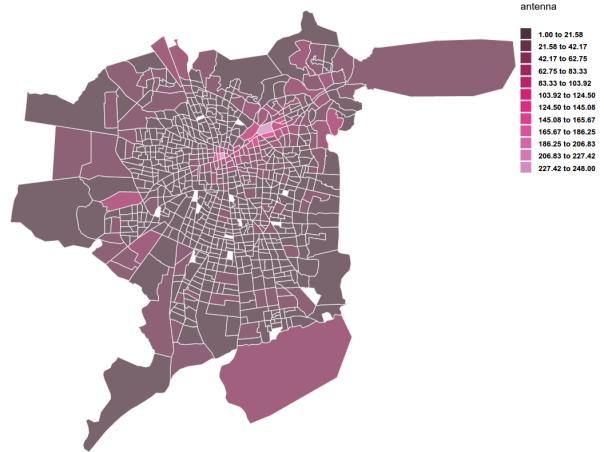


Figure 6: Zoning of Santiago from the OD Survey. Colors indicate antenna density in each zone.

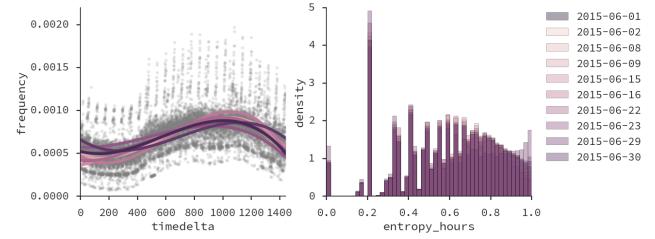


Figure 7: Distributions of CDR event frequency (left) and entropy with respect to hours of the day (right).

5. DAILY JOURNEYS AND OD MATRIX

We applied our method to generate daily journeys J_u to the CDR dataset. Figure 8 shows the result of randomly selected samples from the dataset. Each colored line is a different user, and the grey line underneath each colored line is the original timetable before simplification. The remaining “turning points” are rendered in purple, with a bigger size for easy identification on the image. Those points define the activity segments which we classify as *unknown*, *non-trip* and *trip* according to the definitions provided earlier.

After estimating the daily journeys from the graphical timetables, we discarded users without trip and non-trip activities. Table 1 shows the final number of users considered (after filtering by entropy, and after filtering without valid trips/non-trips). From these activities, we built a transient OD matrix for each day, using the initial and final positions of each trip, which were assigned to their corresponding municipalities. Since we estimated matrices for many days, we averaged the number of trips for each OD pair of origin/destination municipalities (m_o, m_d). Figure 9 shows the resulting matrix.

We also estimated the trip variables analyzed before for the ODS. Figure 10 shows the kernel density estimations of each distribution for each day. One can see that, overall, the distributions are mostly similar for all days, with the following exceptions: start time distribution is different on June 29th, and trip duration distribution is different on June 8th and 9th. On June 9th there was a strike in the Santiago public transport system, which explains partly the increased trip time in comparison to the other days. Additionally, on June 29th the semi-final of the latin-american soccer championship *Copa América*, where Chile was a contender, was played at 8pm. One can see that the number of trips in the afternoon was much higher than in the morning, making the morning peak to shrink in

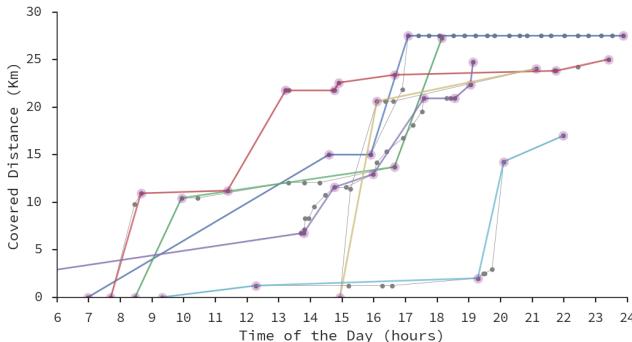


Figure 8: Sample graphical timelines of mobile user traces with our method applied, including the RDP line simplification algorithm.

Date	N	# Trips	Mean Time (m)	Mean Distance (km)
2015-06-01	37,037	56,148	78.38	7.68
2015-06-02	37,622	57,764	77.78	7.74
2015-06-08	34,272	50,017	83.86	7.80
2015-06-09	34,611	50,778	83.03	7.77
2015-06-15	36,274	53,642	79.45	7.68
2015-06-16	36,804	56,046	78.16	7.68
2015-06-22	34,288	49,434	80.05	7.78
2015-06-23	36,239	54,983	77.62	7.71
2015-06-29	22,770	31,963	74.78	7.50
2015-06-30	35,181	52,597	78.94	7.66

Table 1: Number of users and trips for each day analyzed, as well as mean trip duration and mean euclidean distance.

the density estimation. Moreover, the distribution highlights that the afternoon peak was earlier than usual, and that there was a night peak after the soccer match.

Table 1 shows the number of users (and their trips), the mean trip duration, and the mean euclidean distance of trips for each day. Mean times vary within 74.78 and 83.86 minutes, and mean distances vary within 7.5 and 7.8 kilometers. One can see that both distance and time are over-estimated in comparison to the ODS (mean time 41 minutes; and mean distance 6.05 kilometers). The differences in time could appear due to the latency in antenna changes and to CDR event granularity (which happens every 15 minutes in the case of data connections). The differences in distance could be explained by the approximation of each individual position to the corresponding antennas. However, even though the means are different, the distributions have similar shapes to those from the ODS, as displayed on Figure 10.

Comparisons with ODS. A key question is how much different our results are with respect to the ODS. First, we estimated the Spearman rank-correlation between our results and the ODS at municipal level, obtaining $\rho = 0.89$ ($p < 0.001$). The correlation is very high, which means that our averaged matrix reflects the flow of people in the city very well. To compare to what extent our result is good, we refer to a 2013 OD matrix estimated from the public transport smart-card data [9].³ This matrix has a correlation of $\rho = 0.3$ ($p < 0.001$) with the ODS, a result that indicates that the ODS captures a diversity of trip modes, not only public transport. We also tested what happened if we did not consider the matrix diagonal (*i.e.*, without intra-municipal trips), and we observed that we maintain our correlation, while the public transport OD increased ($\rho = 0.32$). This makes sense—short trips are less likely to use public transport (*e.g.*, if the destiny is at walking distance).

³<http://www.dtpm.cl/index.php/2013-04-29-20-33-57/matrices-de-viaje>

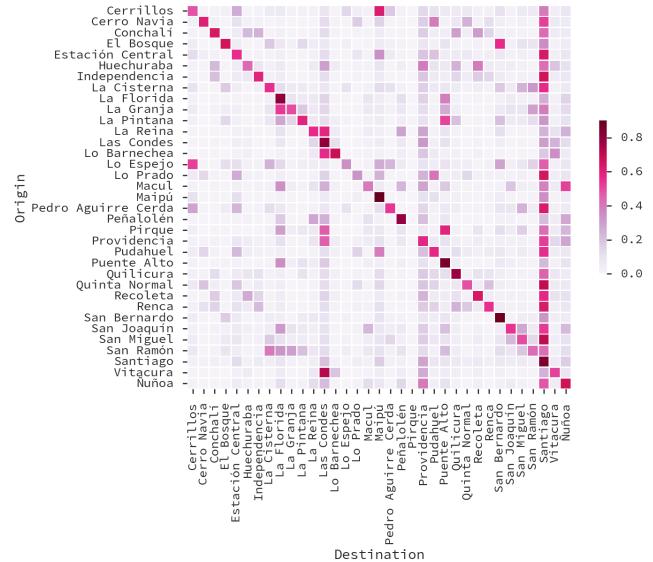


Figure 9: Distributions of CDR trip data from Santiago, Chile. The matrix has been L2 normalized on rows (origins).

Finally, we visually compare whether the distributions of start time, trip duration and trip distance are similar to those of the ODS. To do, Figure 12 shows the *Cumulative Density Functions* of these variables. We observe that, in terms of start time and euclidean distance, the CDFs are very similar, although in terms of trip duration there is a noticeable difference, as mentioned earlier.

6. DISCUSSION AND CONCLUSIONS

In this paper we predicted OD matrices from daily journeys built using mobile data. We did so by proposing a geometric approach based on graphical timelines, and found that, using a sample from mobile data connectivity, we were able to reconstruct the OD pairs in a city, at municipal level, with a very high correlation. Moreover, we were able to estimate start time, trip time and trip distance, in a way that resembled the ODS results, with exception of trip time, which needs calibration to account for the delay in antenna change with respect to the moment in which each trip started.

Implications. On the one hand, our algorithm is very simple and can work with streaming data if what matters is the number of trips. Consider the day in which a soccer match was played, and peak hours shifted. Reportedly, the transport authorities did not account for this shift, and instead they only created ad-hoc routes in public transport.⁴ Thus, our results could support urban computing applications [13] which need almost real-time transport data. On the other hand, our approach complements the ODS in important ways. The ODS is performed every 10 years, but, as with any static picture, it does not capture rich context-dependant dynamics of the city. Conversely, our approach is more dense, as it can be applied even at daily scale to observe differences that the ODS does not. In this paper we have shown that, even working with a sample of anonymous data from mobile data, it is possible to reconstruct part of the ODS, as well as finding diverging days from the typical patterns, at a fraction of its costs. This can be used to measure, for instance, what are the effects in urban flows caused by transport measures like road space rationing.

Limitations and Future Work. Our geometric algorithm, with arguably reasonable transport-based constraints, does not consider

⁴<http://www.mtt.gob.cl/copaamerica/santiago>

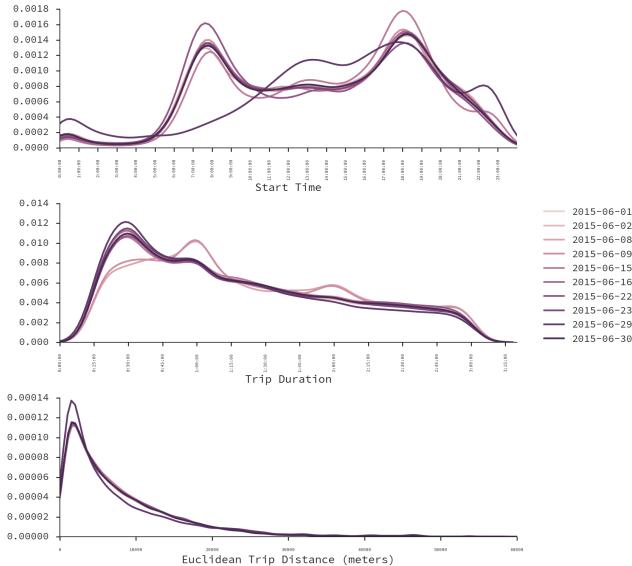


Figure 10: Distributions of CDR trip data for each day.

information that could be useful in determining the turning points of the day (*e.g.*, land-use or census information). Additionally, the distance and time estimation need to be corrected using scaling factors. Then, in addition to address our limitations, the main line of research for future work will be the characterization of non-trip activities. One way of performing such characterization is through the analysis of land-use derived from mobile data analysis [6, 10].

Acknowledgements. We thank Oscar Peredo, José García and Pablo García for valuable discussion.

References

- [1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. “Origin–destination trips by purpose and time of day inferred from mobile phone data”. In: *Transportation Research Part C: Emerging Technologies* (2015).
- [2] Ennio Cascetta. “Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator”. In: *Transportation Research Part B: Methodological* 18.4 (1984), pp. 289–299.
- [3] David H Douglas and Thomas K Peucker. “Algorithms for the reduction of the number of points required to represent a digitized line or its caricature”. In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 10.2 (1973), pp. 112–122.
- [4] Vanessa Frias-Martinez, Cristina Soguero, and Enrique Frias-Martinez. “Estimation of urban commuting patterns using cellphone network data”. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. ACM. 2012, pp. 9–16.
- [5] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. “Understanding individual human mobility patterns”. In: *Nature* 453.7196 (2008), pp. 779–782.
- [6] Eduardo Graells-Garrido and José García. “Visual Exploration of Urban Dynamics Using Mobile Data”. In: *Proceedings of Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information: 9th International Conference, UCAmI 2015*. Springer International Publishing. 2015, pp. 480–491.

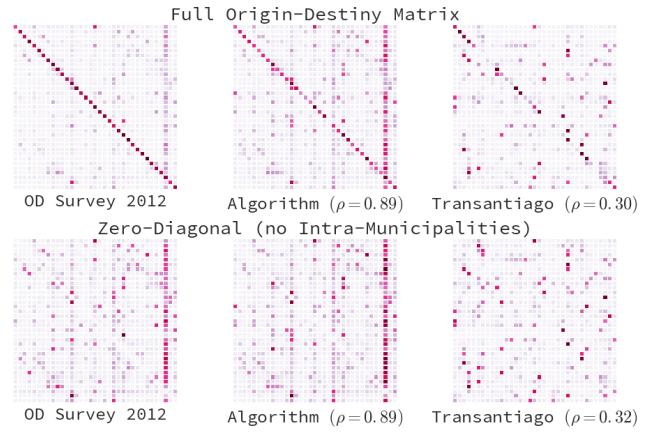


Figure 11: Comparison of OD matrices: OD Survey, our method, and public transport [9].

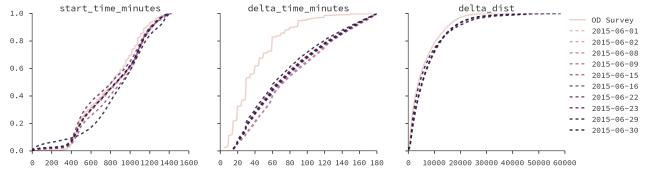


Figure 12: Comparison between OD Survey and CDR-based data.

- [7] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. “Development of origin–destination matrices using mobile phone call data”. In: *Transportation Research Part C: Emerging Technologies* 40 (2014), pp. 63–74.
- [8] Luís Moreira-Matias, Carlos Ferreira, João Gama, João Mendes-Moreira, and Jorge Freire de Sousa. “Bus bunching detection by mining sequences of headway deviations”. In: *Advances in Data Mining. Applications and Theoretical Aspects*. Springer, 2012, pp. 77–91.
- [9] Marcela A Munizaga and Carolina Palma. “Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile”. In: *Transportation Research Part C: Emerging Technologies* 24 (2012), pp. 9–18.
- [10] Kentaro Nishi, Kota Tsubouchi, and Masamichi Shimosaka. “Extracting land-use patterns using location data from smartphones”. In: *Proceedings of the First International Conference on IoT in Urban Space*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2014, pp. 38–43.
- [11] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. “Limits of predictability in human mobility”. In: *Science* 327.5968 (2010), pp. 1018–1021.
- [12] Edward R Tufte. *Envisioning information*. Graphics Press, 1993.
- [13] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. “Urban computing: concepts, methodologies, and applications”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.3 (2014), p. 38.

How close the models are to the reality? Comparison of Transit Origin-Destination Estimates with Automatic Fare Collection Data

Ahmad Tavassoli¹, Azalden Alsger¹, Mark Hickman¹, Mahmoud Mesbah¹

¹School of Civil Engineering, The University of Queensland, Brisbane, Australia

Email for correspondence: a.tavassoli@uq.edu.au

Abstract

There is a consensus on the importance and value of automatic fare collection (AFC) data in analysing different aspects of public transport. As such combining other data sources such as the General Transit Feed Specification (GTFS) can greatly improve the quality of the analyses and ultimately provide a better understanding of public transport performance. This paper presents a methodology for data processing and analysis to acquire a public transport Origin Destination (OD) matrix. The case study uses a very large dataset on passenger boarding and alighting of all three transit modes, namely bus, rail and ferry, in South-East Queensland (SEQ). The OD trip matrices are estimated for both the AM and PM peak periods for five weekdays. Also, the estimated public transport demands for the same periods from the SEQ strategic transport model (SEQSTM) are employed. This approach enables not only the comparison of OD matrices over time to determine changes in travel patterns but also investigates the similarity between the demands from the SEQSTM procedure and those from AFC data. A number of statistical measures, namely R^2 , GEH, %RMSE and an eigenvalue-based measure, are utilized to determine the level of similarity of these OD matrices. The results highlight the similarity of the trip pattern between four workdays (Mon-Thu). However, trip patterns on a Friday are slightly different from the other weekdays, particularly in the PM peak period. Also, the demand from SEQSTM for both time periods is not analogous to any of the AFC patterns.

Key words: Smart card scheme; Public transport, OD matrices, OD matrix similarity

1. Introduction

Smart card data are increasingly used for transit network planning, passengers' behaviour analysis and network demand forecasting. The primary advantage of using smart cards, in addition to their original use as a valuable payment option, is to provide a high quality and plentiful source of information for transit agencies and researchers (Pelletier et al., 2011). In addition, smart card data can be used to better understand passenger travel behaviour and measure trip habits (Lee and Hickman, 2011; El Mahrsi et al., 2014), improve strategic planning and manage the demand through the network (Frumin, 2010; Sun et al., 2012) and estimate missing information such as alighting locations and OD trips (Gordon et al., 2013; Alsger et al., 2015). The accuracy level of smart card data greatly influences the extracted information and the successful estimation of OD matrices (Pelletier et al., 2011).

An OD matrix is an important input to transport models to assess new transport policies. There have been a few attempts to evaluate and assess the accuracy and similarity of OD matrices, using a number of statistical measures. Ye et al. (2012) used the Chi-squared test as a goodness-of-fit measure to compare synthetic matrices. The Chi-squared test ignores the correlations between cells and deals with cells independently. Alsger et al. (2015) used

the Geoffrey E. Havers (GEH) statistic to evaluate the accuracy level of a set of estimated matrices with a base OD matrix.

Djukic et al. (2013) presented the Mean Structural SIMilarity (MSSIM) as a more appropriate comparator of matrices. This method compares OD matrices as images based on pixels equating to individual OD cells. The authors showed a degree of correlation between the neighbouring cells (pixels) just as in images. Later, Day-Pollard and van Vuren (2015) investigated the comparison of OD matrices based on the MSSIM and other comparison techniques, namely R^2 , GEH and RMSE. The authors concluded that the MSSIM approach requires further refinement for use with OD matrices. Ruiz de Villa et al. (2014) introduced a measure for comparing OD matrices, Wasserstein metric, unlike the methods that only based on the cell by cell comparison. The suggested method is based on the topology of the network and considering travel time between all OD pairs. However, this approach is impracticable for large networks due to the huge number of calculations required.

The accessibility and quality of data required for evaluation of estimated OD matrices have usually been a big challenge. In the current research, a unique smart card (automated fare collection, or AFC) dataset, known as GoCard and obtained from TransLink¹, is used to evaluate the estimated OD matrices. The important advantage of this dataset is that it includes both boarding and alighting times and locations for each passenger of the public transport services that comprise buses, trains and ferries.

In addition to the experimental data, this paper compares the results with a synthetic regional transport model. The South-East Queensland Strategic Transport Model (SEQSTM) is a four-step strategic transport model developed by the Queensland Department of Transport and Main Roads (TMR). This model is developed in the EMME/4 modelling platform² to serve as a long-range planning tool. The model was already calibrated and validated by TMR (TMR, 2011). The model is comprised of 1394 traffic zones, and this zoning system is used for estimating OD flows based on the AFC data. This model takes advantage of a mode choice model that includes seven modes: car driver, car passenger, walk to public transport, park and ride, kiss and ride, cycle and walk. The transit type includes three main modes: bus, rail, and ferry. The model forecasts demand for a 24-hour period and applies fixed time-period factors to allocate trips to the AM peak, inter-peak, PM peak and off-peak. Demand is segmented by eight resident trip purposes at trip generation, trip distribution and mode choice stages (Hunkin, 2009). The transit demand was obtained by aggregating all demands from different trip purposes after the mode choice step. The results of mode choice were calibrated and validated using the SEQ travel survey on the base year (2011) by TMR (Joycey and Ryan, 2008).

The objective of this paper is threefold:

- to investigate on the level of accuracy of the AFC data;
- to evaluate the travel pattern changes over time by comparing OD matrices based on AFC data on multiple days and in different periods of time; and,
- to compare the travel pattern obtained from the two sources of AFC and SEQSTM in both the AM and PM peaks.

The remaining sections are organised as follows. The next section explains the research methodology, comprising the data description, the preparation and cleaning procedure, the trip-chaining method, and the OD estimation algorithm. The results of the OD estimation matrices for different weekdays are provided in the third section. These results are then used to conduct the similarity analysis and evaluate the accuracy of these matrices for different days and also for the SEQSTM using different measurements. Finally, conclusions and suggestions for future work are presented.

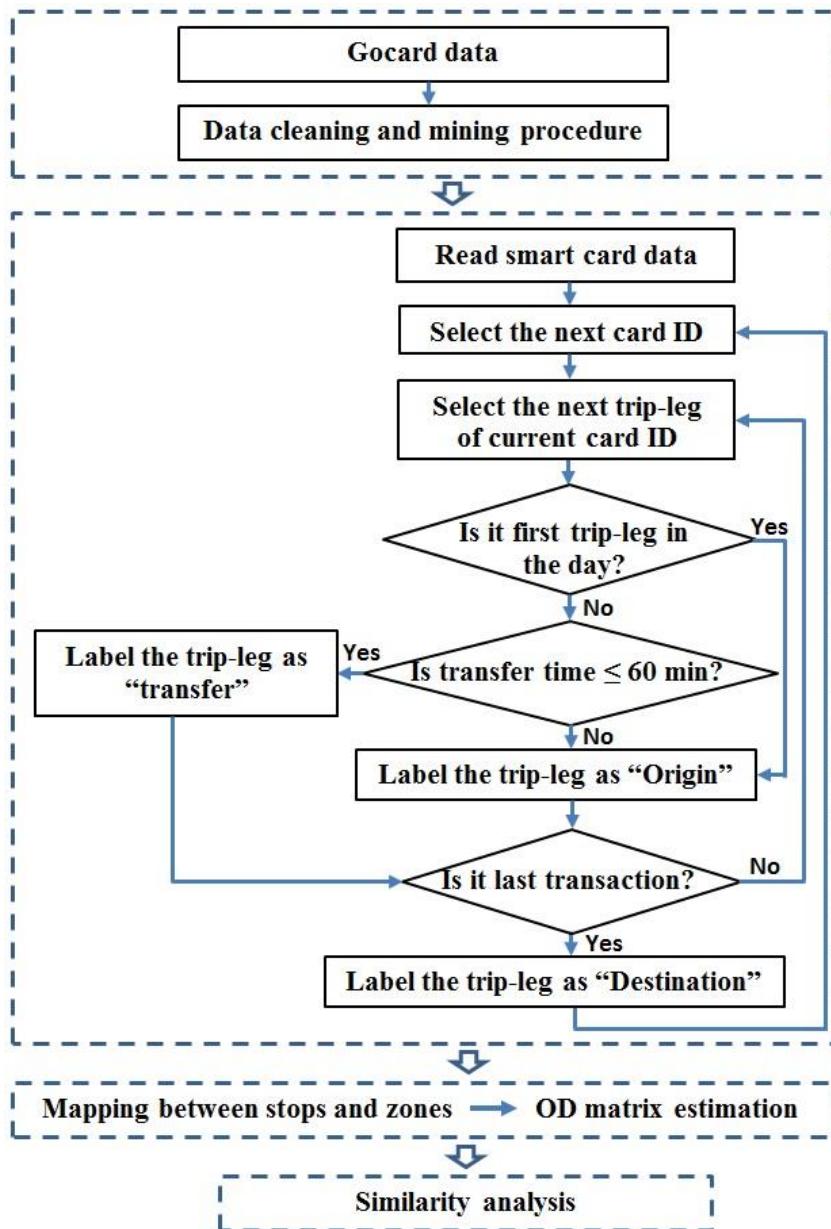
¹ The public transport authority of South-East Queensland (SEQ), Australia

² A commercial software package that is distributed by INRO in Canada

2. Methodology

Our study framework is shown in Figure 1. This framework encompasses four stages, namely: 1) AFC data processing, 2) application of a trip-chaining method, 3) OD matrix estimation, and 4) similarity analysis.

Figure 1: Study framework



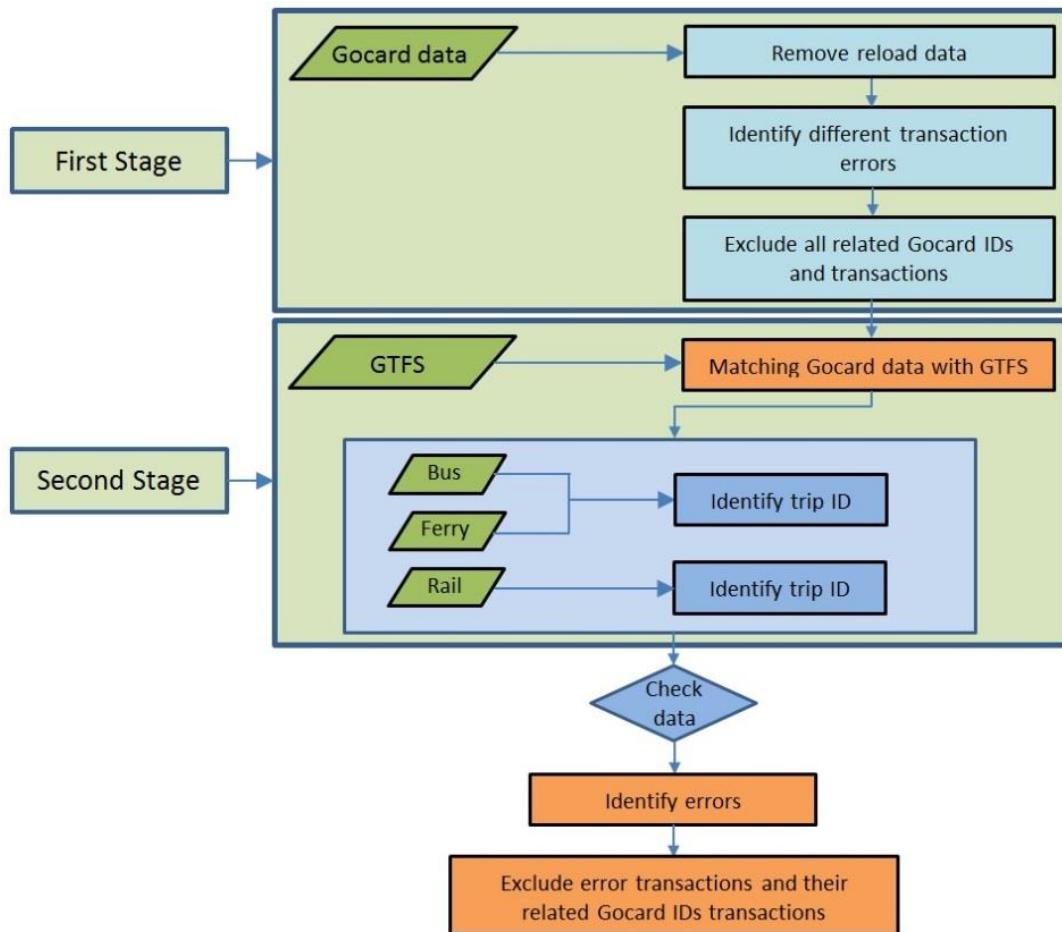
2.1. AFC data

Stage one relates to the GoCard dataset and the process of data cleaning. Robinson et al. (2014) highlight that the level of accuracy of AFC data may vary, and the data are affected by various types of errors. These errors may affect the accuracy of individual journeys and trip chains. In this regard, data validation and journey validation as described in Pelletier et al. (2011) are essential to ensure the quality of data for the purpose of this research. To perform the data cleaning, a framework is proposed including two stages as shown in Figure 2. Since the nature of data usually contains some errors caused by system failure or human error, the data are filtered with some transactions excluded, such as duplicate transactions and transactions when no boarding or alighting stops are recorded. In addition, all reload

transactions, which are related to GoCard credit top-ups and are not transactions related directly to public transit use, are excluded from the data.

The second stage involves mapping the GoCard data into the GTFS network. This stage facilitates our investigation of the validity of boarding and alighting stops and the associated route and direction. The approach includes finding all possible trip-IDs for each trip leg based on the boarding and alighting stops in conjunction with their associated times. Then, the best trip-ID is selected based on minimising the differences between the scheduled and actual transaction times calculated for both boarding and alighting stops. Ultimately, transactions with errors in defining origin or destination stops and/or times are identified and excluded from the analyses.

Figure 2: Framework of GoCard data mining and cleaning process



If any transaction of a card ID is excluded, the rest of the transactions for that card ID are also excluded for the given day, as the transactions on a single day must be in sequence to be chained into a tour for the given day.

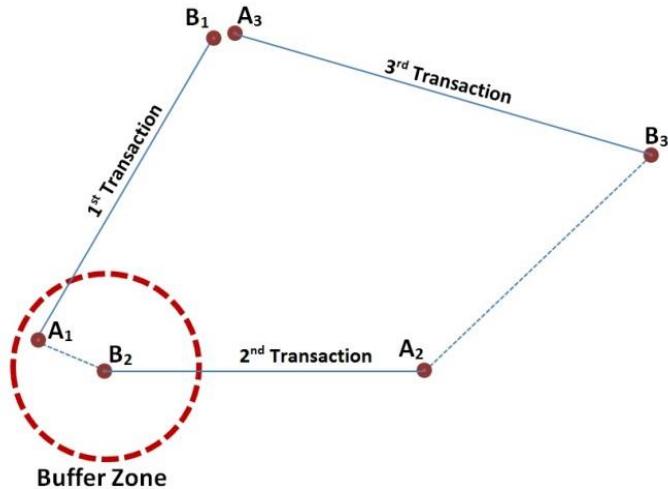
2.2. Trip-chaining method

The main purpose in the trip-chaining method is to connect transactions of a passenger to infer full passenger journeys. Figure 3 shows an example of the trip-chaining method.

A passenger has his first transaction from first boarding (B_1) to first alighting (A_1), and then walks to the next boarding stop to start the second transaction from second boarding (B_2) to second alighting (A_2). To complete a passenger's travel sequence, allowable transfer time has to be assumed. This time threshold is used to merge transactions into a single journey. If the transfer time exceeds this threshold, the next boarding is a new OD trip (Munizaga et

al., 2014; Alsger et al., 2015). It should be noted that the time between (A_1) and (B_2) could be short, with enough time only for a transfer, or long, with enough time for an intermediate activity.

Figure 3: Example of the trip-chaining method



Note: The first boarding transaction in a day is identified by B_1 and the first alighting transaction as A_1 , the time between B_1 and A_1 is the in-vehicle time. (Adapted from: Alsger et al., 2015)

The main function of this method is to detect transfers, so that trip-legs can be merged to obtain journeys. In this regard, the first transaction of the day is the boarding for the first trip-leg, for a unique card ID. The remaining trip-legs for the same card ID will be transfers if the transfer time is less than the allowable transfer time (threshold). The allowable transfer time is set as 60 minutes, to be compatible with TransLink's 60-min transfer time allowance (Alsger et al., 2015). If the transfer time exceeds this value, the alighting location of the prior trip-leg is the destination of the passenger's journey, and the next transaction is the boarding location of a new journey. If an alighting transaction is the last transaction of the day for the current card ID, another card ID is chosen, and the algorithm continues searching to create passenger journeys.

2.3. OD matrix estimation

The next step is to estimate the OD matrix from the passengers' journeys based on the GoCard dataset. In this process, stop-to-stop OD journeys should be converted to zone-to-zone OD journeys. For this purpose, the same traffic analysis zones (TAZs) in the SEQSTM are utilized as the level of aggregation. A journey from any stop located within the TAZ is counted as a journey originating in that TAZ; similarly, a journey ending at any stop within a TAZ is counted as a journey destined for that TAZ. Using the same procedure, OD matrices can be estimated for different days of a week and also different time periods in each day (AM peak and PM peak).

2.4. Similarity analysis

Different measurements are utilized to compare OD matrices and determine the level of similarity over time and between sources. These measures are described below.

2.4.1. R-squared (R^2)

The R-squared (R^2), as one of the most commonly and widely used (Washington et al., 2011), is a statistical measure of how close the data are to the fitted regression line, and it is used for comparing between origin-destination pairs of two OD matrices. R^2 values range

from 0 to 1, with higher values indicating less difference between OD matrices. A general formula for calculating R^2 is:

$$R^2 = 1 - \frac{SS_E}{SS_T} \quad (1)$$

$$SS_E = \sum_i (OD_{i,j}^1 - OD_{i,j}^2)^2 \quad (2)$$

$$SS_T = \sum_i (OD_{i,j}^1 - \bar{OD})^2 \quad (3)$$

where: $OD_{i,j}^1$ is the pair i,j of the first demand matrix and $OD_{i,j}^2$ is the pair i,j of the second demand matrix \bar{OD} is the mean of the OD^1 pairs.

Along with considering higher value of R^2 as a higher level of similarity, the regression line should be close to a 45-degree line through the origin. In this condition, the coefficient of the line should be closer to one and the intercept should be closer to zero. The lower and greater coefficient values indicate the tendency of the pattern to overestimate or underestimate values in the reference OD matrix.

2.4.2. Geoffrey E. Havers (GEH) statistic

The GEH statistic is used to evaluate the level of closeness between origin-destination pairs of two OD matrices. The GEH is applied to every pair in the two matrices, with a GEH of less than 5 indicating a good fit (Hollander and Liu 2008). The GEH formula is:

$$GEH = \sqrt{\frac{2(OD_{i,j}^1 - OD_{i,j}^2)^2}{OD_{i,j}^1 + OD_{i,j}^2}} \quad (4)$$

Then, the percentage of OD pairs that have a GEH equal to or less than 5 is calculated to indicate the level of closeness between two OD matrices.

2.4.3. Root Mean Square Error (RMSE) and Percent Root Mean Square Error (%RMSE)

The root mean squared error (RMSE) and accordingly the percent root mean squared error (%RMSE) are used to evaluate the closeness of the matrices. The %RMSE is where the variability of the demand is most evident: if two demand matrices were identical, the %RMSE would be equal to zero. The (RMSE) and (%RMSE) are:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (OD_{i,j}^1 - OD_{i,j}^2)^2}{N}} \quad (5)$$

$$\%RMSE = \frac{RMSE}{\left(\frac{\sum_{i=1}^N OD_{i,j}^1}{N} \right)} \times 100 \quad (6)$$

where: $OD_{i,j}^1$ is the pair i,j of the first demand matrix and $OD_{i,j}^2$ is the pair i,j of the second demand matrix.

2.4.4. Eigenvalue-based measure (EBM)

This measure is based on the concept of eigenvectors and is introduced for comparison of OD matrices in this study. OD_1 is similar to OD_2 if there exists a matrix P such that $OD_2 = P^{-1}OD_1P$.

$$\begin{aligned}
 \det(OD_2 - \lambda I) &= \det(P^{-1}OD_1P - \lambda P^{-1}P) = \det(P^{-1}[OD_1 - \lambda I]P) \\
 &= \det(P^{-1})\det(OD_1 - \lambda I)\det(P) \\
 &= \det(P^{-1})\det(P)\det(OD_1 - \lambda I) = \det(P^{-1}P)\det(OD_1 - \lambda I) \\
 &= \det(OD_1 - \lambda I)
 \end{aligned} \tag{7}$$

OD_1 and OD_2 have the same characteristic polynomial and therefore the same eigenvalues. On this basis, two similar matrices have the same eigenvalues (Ford, 2014). Based on this approach, two OD matrices are similar if their eigenvalues are the same. The sum of absolute error (SAE) is a promising technique to determine the closeness of model results to the actual data and has been used in a number of studies (Chowdhury and Saha, 2011; Purdy, 2012). This technique is employed in this study to establish a measure to show the similarity of ODs by comparing vectors of the eigenvalues of the OD matrices; the lower the value, the greater is the similarity.

$$EBM = SAE(\text{eig}(OD_1) - \text{eig}(OD_2)) \tag{8}$$

where: OD_1 and OD_2 are the demand matrices, and $\text{eig}(\cdot)$ is a vector containing the eigenvalues of a square matrix.

3. Data description and analysis

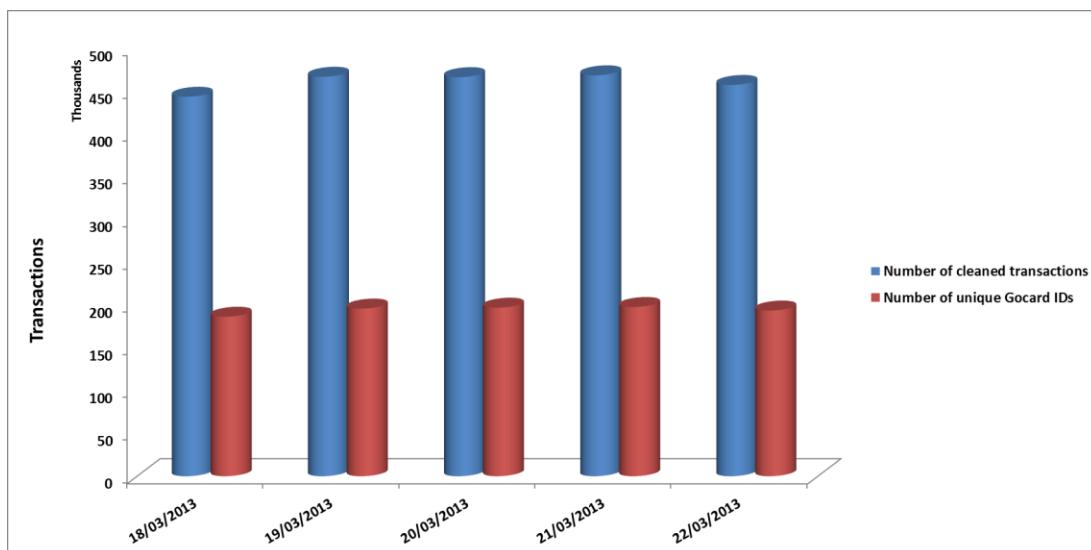
The GoCard dataset for five weekdays is analysed over the SEQ network, considering all modes, namely bus, rail, and ferry, for Monday 18 March to Friday 22 March 2013. These weekdays were selected as there was no public holiday and had normal weather conditions. In the SEQ network, a transaction record is generated each time a passenger boards and alights. Each transaction contains information comprising: the operation date, run, route, direction, ticket number, smartcard ID, boarding time, alighting time, boarding stop and alighting stop. However, transferring activities are not directly obtained.

Table 1 presents the results of the data in different stages of this study as described in Section 2. First, the general data description is given, including the total number of transactions per day before cleaning, the reload transactions, and the available trip legs per day after excluding these errors. Then, a summary of these errors is given for each day: the situations that boarding or alighting locations are missing or are not recorded, boarding or alighting times are missing, the boarding stop location is the same as the alighting stop, and boarding time is later than or equal to the alighting time. The next step in data cleaning is identifying and excluding stop mapping errors where the algorithm could not find a match between the stops in both the GoCard data and GTFS.

The next part of Table 1 presents the number of trip legs after performing the data cleaning, the number of errors in trip legs in the second stage, the number of treated trip legs, and ultimately the number of available trip legs after all data cleaning. The results indicate that about 68% of the initial data can be used for the current research. This available GoCard data is large enough to provide a reliable sample size, considering the fact that more than 82% of transit trips are made by GoCard users in SEQ (Moore, 2015). After applying the trip-chaining method to obtain passengers' journeys, the last section of the table shows the information related to the number of journeys per day and also the unique number of GoCard IDs related to these journeys. Figure 4 shows the total number of cleaned transactions and the corresponding number of GoCard IDs for the selected weekdays.

Table 1: Data description and statistics in different stages of the study

Description	Date				
	18/03/13 Mon	19/03/13 Tue	20/03/13 Wed	21/03/13 Thu	22/03/13 Fri
General data description					
Total transactions	609,509	635,841	628,479	633,106	613,481
Number of reload transactions	16,382	17,034	16,999	17,621	15,921
Available trip legs per day	593,127	618,807	611,480	615,485	597,560
General errors					
No boarding stop	18,005	18,683	16,353	16,916	14,915
No alighting stop	11,931	11,698	11,780	11,430	10,504
No boarding time or alighting time	29,936	30,381	28,133	28,346	25,419
Boarding time >= alighting time	2,776	2,846	3,025	2,836	3,090
No boarding stop or alighting stop	29,936	30,381	28,133	28,346	25,419
Boarding stop = alighting stop	12,017	12,242	12,035	11,852	12,596
Stop mapping errors					
Stop mapping error in boarding	12,223	12,684	11,989	12,231	11,559
Stop mapping error in alighting	11,988	12,297	11,680	11,939	11,296
Trip legs					
Number of available trip legs after first data cleaning	443,087	466,438	465,988	468,119	456,842
Number of errors in trip legs	45,836	46,651	44,973	46,601	44,395
Number of treated trip legs	6,306	6,635	6,324	7,109	7,813
Number of available trip legs after all data cleaning	395,633	418,823	418,098	418,424	409,296
Ratio of available data to all data	66.7%	67.7%	68.4%	68.0%	68.5%
OD trips information					
Number of Go Card IDs	182,799	192,501	193,433	194,151	190,187
Number of journeys	324,091	343,000	342,821	373,796	334,287

Figure 4: Number of cleaned transactions and GoCard IDs for the selected weekdays

OD matrices for weekdays based on the passengers' journeys are then generated. The next section introduces the results of the estimation of the OD matrices for both weekdays and for the SEQSTM.

3.1. OD matrices estimation

Transit travel demand mostly follows a non-uniform time-of-day distribution and includes two main peak periods, AM peak and PM peak. To understand the daily behaviour of demand, journeys were aggregated based on the start time during each 15-min interval. Figure 5 shows the public transport time-of-day demand based on the Go Card data for weekdays.

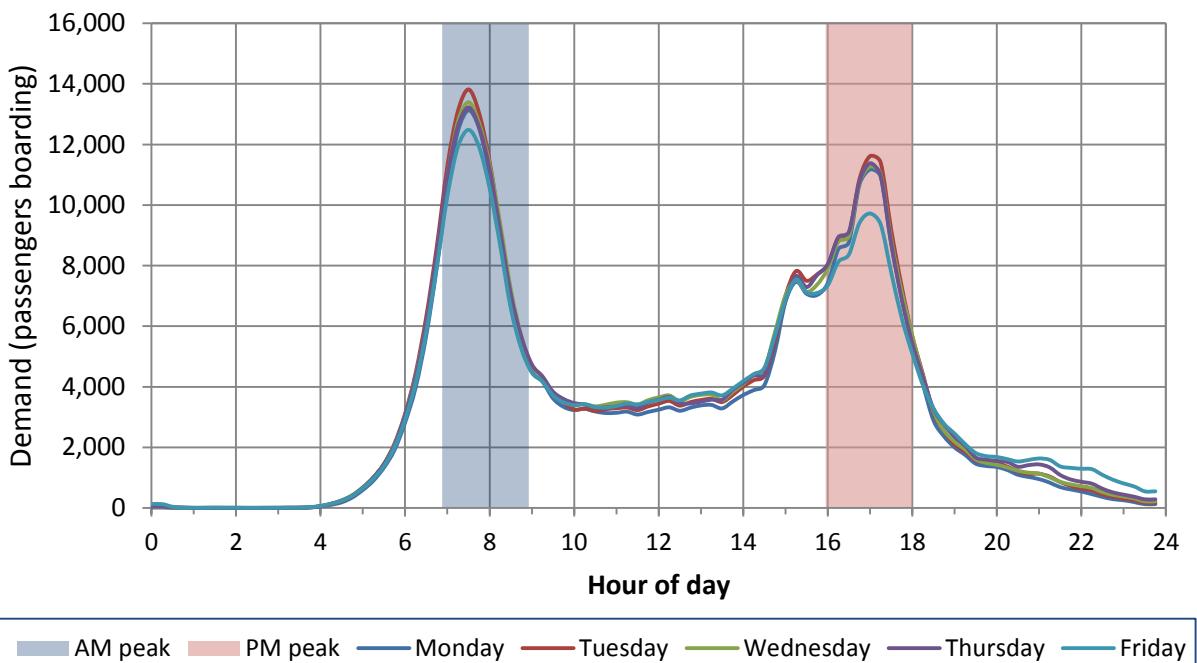
Since demand during each time unit is averaged over 15-min intervals, this could cause variation of demand. As a result, there is a potential risk of overestimating or underestimating the demand profile. The moving average technique is employed to minimize the effect of this variation and to smooth the demand profile. The moving average of three sequential demands can be calculated as:

$$MDS_t = (DS_{t-1} + DS_t + DS_{t+1}) / 3 \quad (9)$$

where

- DS_t = the moving average of three sequential speeds;
- DS_{t-1} = the demand at one time-interval before t ,
- DS_t = the demand at one time-interval at t ,
- DS_{t+1} = the demand at one time-interval after t ,

Figure 1: Illustrative Time-of-Day Variations in Transit Demand for weekdays



As can be seen from Figure 5, all weekdays follow a similar trend including morning and afternoon peak periods. The afternoon peak is lower and more broadly spread out compared with the morning peak. In addition, the demand on Friday is slightly different from that of

other weekdays and lower in both peak periods. To use similar peak periods as those in the SEQSTM, AM peak and PM peak are defined as 7:00AM-9:00AM and 4:00PM-6:00PM, respectively. Based on the start time of the journeys, OD matrices were calculated for weekdays and both AM and PM peak periods. In addition, public transport OD matrices from SEQSTM for the two peak periods are extracted. A summary of statistics of all OD matrices are shown in Tables 2 and 3.

Table 2: Summary statistics of demand matrices for AM peak

	Mon	Tue	Wed	Thu	Fri	SEQSTM
Total demand (journeys)	80,259	84,020	83,305	82,246	77,333	210,976
Number of pairs with non-zero demand	15,922	16,507	16,469	16,496	15,978	1,865,955
Maximum demand (journeys)	587	548	560	588	513	790
Average of demand (journeys)	5.04	5.09	5.06	4.99	4.84	0.11
STD of demand (journeys)	18.40	18.34	18.33	18.13	17.37	2.40

Table 3: Summary statistics of demand matrices for PM peak

	Mon	Tue	Wed	Thu	Fri	SEQSTM
Total demand (journeys)	73,733	76,872	75,099	75,323	66,409	139,476
Number of pairs with non-zero demand	14,560	15,100	14,957	15,047	14,440	1,865,953
Maximum demand (journeys)	562	549	561	580	437	203
Average of demand (journeys)	5.06	5.09	5.02	5.01	4.60	0.07
STD of demand (journeys)	18.31	18.22	17.84	17.92	15.44	1.00

The results indicate that the demand based on the GoCard data in the PM peak always is lower than that during the AM peak. Friday has the lowest demand and Tuesday has the highest demand among weekdays. On average, the number of OD pairs with demands more than zero are about 1000 pairs more in the AM peak. Nonetheless, the maximum demand, average demand, and STD of demand are very similar, comparing both peak periods.

Comparing the demands from GoCard data and SEQSTM in both peak periods reveals that the SEQSTM demand is significantly higher than the GoCard weekday demand in both the AM and PM peaks. This might be due to the fact that during the data cleaning about 32% of the data was excluded from the analysis. In addition, there were some passengers who used paper tickets rather than the GoCard and therefore were not considered in the analysis. On this basis, all demand matrices were normalised according to their total demand to analyse the similarity between matrices. Furthermore, the fairly low average demand and STD of demand, and the high number of number of OD pairs with demand greater than zero in the SEQSTM, indicate that there are many OD pairs with demand less than 1 journey, particularly for the PM peak. This may cause big discrepancies between the GoCard weekday demand and the SEQSTM demand. In this regard, analysis was performed using different threshold measures including 0.5, 1, 2 and 3 trips, in order to identify the impacts of demand pairs with a low value in the results. The results show that "Number of pairs with non-zero demand" decreased to 51,360, 28,653, 15,009 and 9,918 pairs in the AM peak, respectively, for the four thresholds. In PM peak, this measure decreased to 43,337, 22,310, 10,629 and 6,719 pairs, respectively, for the four thresholds.

4. Similarity analysis

The statistical measures from Section 2.4 are utilised to assess the similarity between demand matrices of the GoCard and the SEQSTM in AM and PM peak periods. The results of the analysis based on these measures are presented in Table 4 and Tables 5 for the AM and PM peak periods, respectively. The GEH measures for almost all comparisons are more than 99%, indicating a fairly good similarity, unlike the results from the other measures. This might be related to the scale of the demand pairs. On this basis, the GEH statistic was determined not to be a suitable measure for this analysis and was excluded from the considered measures for comparison. It is only indicative when a threshold is defined for the ‘negligible’ OD demand; for an example of this approach, see Alsgaer et al. (2015). As discussed in the previous section, different threshold measures were used to analyse the demand from the SEQSTM. The results indicate that in all conditions in both the AM peak and the PM peak periods, transit travel patterns from the SEQSTM are not similar to the weekday demand from the GoCard. Accordingly, the results in this section are based on the not excluding any demand from the SEQSTM.

The demand matrices are almost the same in the AM peak on weekdays, as the R^2 measure is more than 0.97 and the coefficient of the lines are quite close to one (more than 0.910) and the constants are close to zero (less than 0.27), as shown in Tables 4a and 4b. In addition, the %RMSE measures are within the same range and the EBM measures follow a similar pattern. However, the OD matrix on Friday is slightly different from the other weekdays, as the R^2 is lower on Friday, the coefficients and constants have more distance from the ideal measures on Friday, and also %RMSE and EBM measures are higher on Friday. This suggests different timing and scale of demand in the Friday PM peak period, as some people may leave work early or may prefer to do a social activity before going home.

Comparing normalised demand matrices between the GoCard and the SEQSTM indicates that transit travel patterns from the SEQSTM are not similar to the weekday demand from the GoCard in both AM and PM peaks. R^2 measures are clearly lower compared to the same measure between weekdays, with a fairly large distance of the coefficients and constants from the ideal measures. On average the coefficient is about 0.2 away from one and the constant is about 1.15.

The EBM measures are also about 250,000 on average for similarity between the SEQSTM matrix compared to an average of 65,000 for weekdays in the AM peak period as shown in Table 4. The same trends can be seen in the PM peak period. The %RMSE measure is also confirming the dissimilarity of the SEQSTM matrix with the weekday matrices based on the GoCard data. This suggests the need to re-evaluate and calibrate the demand within the strategic transport model with the demand from actual comprehensive datasets.

Table 4: Results of similarity measures between demand matrices for AM peak ***a) R² measure**

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1	0.98	0.97	0.97	0.97	0.21
Tue		1	0.98	0.97	0.97	0.19
Wed			1	0.98	0.98	0.20
Thu				1	0.98	0.20
Fri					1	0.21
STM						1

b) Parameters for the best fitting line **

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1 0	1 0.27	0.99 0.23	0.99 0.2	0.93 0.21	0.20 1.60
Tue		1 0	0.98 0.04	0.97 -0.002	0.91 0.03	0.19 1.57
Wed			1 0	0.977 0.04	0.92 0.06	0.20 1.56
Thu				1 0	0.93 0.08	0.20 1.58
Fri					1 0	0.21 1.58
STM						1 0

c) %RMSE

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	49	51	52	53	351
Tue		0	49	50	55	346
Wed			0	47	49	348
Thu				0	47	349
Fri					0	345
STM						0

d) Eigenvalue-based measure

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	63,793	60,072	63,984	67,975	247,866
Tue		0	70,292	68,605	70,135	249,276
Wed			0	60,181	67,891	248,125
Thu				0	68,802	249,120
Fri					0	248,580
STM						0

* Demand of weekdays based on Go card data, STM: public transport demand from SEQSTM

** The top value is coefficient of the line, the below value is intercept of the line

Table 5: Results of similarity measures between demand matrices for PM peak ***a) R²**

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1	0.98	0.96	0.97	0.94	0.3
Tue		1	0.96	0.97	0.95	0.29
Wed			1	0.96	0.94	0.29
Thu				1	0.95	0.3
Fri					1	0.28
STM						1

b) Parameters for the best fitting line **

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1 0	0.99 0.24	0.96 0.33	0.97 0.25	0.81 0.58	0.21 1.14
Tue		1 0	0.95 0.17	0.97 0.09	0.80 0.43	0.21 1.12
Wed			1 0	0.98 0.12	0.93 0.42	0.22 1.11
Thu				1 0	0.94 0.45	0.22 1.11
Fri					1 0	0.25 1.05
STM						1 0

c) %RMSE

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	49	61	57	72	335
Tue		0	58	54	67	332
Wed			0	64	72	329
Thu				0	65	331
Fri					0	312
STM						0

d) Eigenvalue-based measure

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	89,445	62,214	86,293	101,385	138,536
Tue		0	91,617	91,788	100,427	138,665
Wed			0	86,210	99,204	140,525
Thu				0	75,658	141,131
Fri					0	148,862
STM						0

* Demand of weekdays based on Go card data, STM: public transport demand from SEQSTM

** The top value is coefficient of the line, the below value is intercept of the line

5. Conclusions

Automated fare collection systems have been widely used in public transport and have provided very large datasets. One of the main challenges of using such data is its accuracy level. This study makes use of the AFC system in SEQ, Australia, which has extensive records of passenger boardings and alightings by all transit modes. On this basis, this study presents the errors at different stages of data cleaning and presents the quality of such data for journey estimation.

From another perspective, AFC systems are a rich source of information for many transport planning applications. The proposed methodology in this study utilises the AFC data to characterise passenger journeys in order to estimate the OD matrices of transit passengers in weekdays during both AM peak and PM peak periods. In addition, a traditional four-step model, the SEQSTM, is used to compare transit OD demand in both the AM peak and PM peak periods.

Comparing demand matrices provides essential information about passenger travel patterns. This comparison may also help to avoid unnecessary surveys by suggesting the use of similar available data. This study introduces a new measure for OD matrix similarity, an eigenvalue-based measure (EBM), along with the other established statistical measures of R^2 , GEH, and %RMSE. The results show that the proposed measure has good performance in terms of indicating the level of similarity between matrices. This study performs the similarity analysis between weekdays demand matrices based on GoCard data and also with the SEQSTM demand in both peak periods.

The results show that the AM peak has slightly higher demand compared to PM peak for all weekdays, and the demand fluctuations are greater across days in the PM peak. Also, the Friday demand is slightly different from other weekdays (Monday to Thursday) in the PM peak. Furthermore, the SEQSTM has larger demand compared to the GoCard weekday demand. These findings highlight that the public transport travel OD matrix according to the SEQSTM is distinct from that of the GoCard data.

Further research needs to be conducted to investigate on the similarity of the transit demand on weekends. Also, it is recommended that the effects of adverse weather on transit demand and passengers' travel behaviour be examined.

Acknowledgements

The authors acknowledge TransLink (a unit of TMR covering all of Queensland, Australia) for providing the data for this research. We also acknowledge the Department of Transport and Main Roads (TMR), Transport Strategy and Planning Branch for providing access to the South-East Queensland Strategic Transport Model (SEQSTM). The research in this paper is partially supported by the Australian Research Council through a Discovery Early Career Researcher Award (grant number DE130100205) and also by the ASTRA (Academic Strategic Transport Research Alliance) Chair at the University of Queensland.

References:

- Alsgar, A. A., Mesbah, M., Ferreira, L. & Safi, H. (2015) Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix. *Transportation Research Record: Journal of the Transportation Research Board*, 2535, 88-96.
- Chowdhury, S. & Saha, P. (2011) Adsorption Kinetic Modeling of Safranin onto Rice Husk Biomatrix Using Pseudo-first-and Pseudo-second-order Kinetic Models: Comparison of Linear and Non-linear Methods. *CLEAN–Soil, Air, Water*, 39(3), 274-282.
- Day-Pollard, T. & Van Vuren, T. (2015) When are Origin-Destination Matrices Similar Enough? *Transportation Research Board 94th Annual Meeting*.
- Djukic, T., Hoogendoorn, S. & Van Lint, H. (2013) Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index. *Transportation Research Board 92nd Annual Meeting*.
- El Mahrsi, M. K., Etienne, C., Johanna, B. & Oukhellou, L. (2014) Understanding passenger patterns in public transit through smart card and socioeconomic data: A case study in rennes, france. *ACM SIGKDD Workshop on Urban Computing*.
- Ford, W. (2014) *Numerical linear algebra with applications: Using MATLAB*, Academic Press.
- Frumin, M. S. (2010) Automatic data for applied railway management: passenger demand, service quality measurement, and tactical planning on the London Overground Network, Massachusetts Institute of Technology.
- Gordon, J., Koutsopoulos, H., Wilson, N. & Attanucci, J. (2013) Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board*, (2343), 17-24.
- Hunkin, P. (2009) *Critical review of transport modelling tools* Sinclair Knight Merz (SKM), viewed 28/09/2016,
https://bitre.gov.au/publications/2009/files/cr_001_Review_of_Transport_Modelling_Tools.pdf.
- Joycey, A. & Ryan, M. (2008) *BSTM Multi-Modal Model Development, Base Year Validation, Working Paper 16*, Queensland government- Queensland transport main roads, Queensland, Australia.
- Lee, S. & Hickman, M. D. (2011) Travel pattern analysis using smart card data of regular users. *Transportation Research Board 90th Annual Meeting*.
- Munizaga, M., Devillaine, F., Navarrete, C. & Silva, D. (2014) Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, 44, 70-79.
- Pelletier, M.-P., Trépanier, M. & Morency, C. (2011) Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4), 557-568.
- Purdy, J. E. (2012) A Theoretical Development and Simulation-Based Comparison of Four Parameter Estimation Methods for the Spatio-Temporal Autologistic Model with Emphasis on Maximum Generalized and Block Generalized Pseudolikelihood, PhD thesis, The University of Montana, Missoula, MT.
- Robinson, S., Narayanan, B., Toh, N. & Pereira, F. (2014) Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies*, 49, 43-58.
- Ruiz De Villa, A., Casas, J. & Breen, M. (2014) OD matrix structural similarity: Wasserstein metric. *Transportation Research Board 93rd Annual Meeting*.
- Sun, L., Lee, D.-H., Erath, A. & Huang, X. (2012) Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. *Proceedings of the ACM SIGKDD international workshop on urban computing*. ACM.
- TMR (2011) South East Queensland Strategic Transport Model (SEQSTM). SEQSTM_MM_v1 ed. Brisbane, Australia, Transport Strategy And Planning Branch, Department of Transport and Main Roads.

How close the models are to the reality?

Comparison of Transit Origin-Destination Estimates with Automatic Fare Collection Data

- Washington, S., Karlaftis, M. G. & Mannering, F. L. (2011) *Statistical and econometric methods for transportation data analysis*, Boca Raton, FL, CRC Press.
- Ye, X., Cheng, W. & Jia, X. (2012) Synthetic Environment to Evaluate Alternative Trip Distribution Models. *Transportation Research Record: Journal of the Transportation Research Board*, (2302), 111-120.



N/LAB

Best Practices and Methodology for OD Matrix Creation from CDR data

Prepared by:

Dr James Goulding

N/LAB, University of Nottingham.

Tel: +44 (0) 7730 559203

Email: james.goulding@nottingham.ac.uk

Executive Summary	3
2. Introduction to CDR Data	4
3. Opportunities and Limitations of CDR data for OD matrix creation	5
4. Overview of OD matrix creation from CDR data	8
4.1 Comparison to Traditional Approaches	9
4.2 Summary of OD Matrix generation process	9
5. Repurposing CDR Data for OD-Matrices: Theoretical challenges and solutions	10
5.1 Population Level-data Completeness and Scaling	11
5.2 Individual-level Data Completeness and Hidden Movement	14
5.2.1 Sparse Temporal Frequency Challenges	14
5.2.2 Update Frequency Bias	16
5.2.3 Bias due to Heterogeneity In Call Rates	17
5.2.4 Single Network Activity Issues	18
5.3 Location Precision Issues and False Movement	19
5.3.1 Limited precision	20
5.3.2 Non-Uniform BTS Density	21
5.3.3 Changes in BTS Operation	22
5.3.4 BTS service region overlap	23
5.4 Lack of directly recorded movement metadata	24
5.5 Privacy and Ethical Implications of Using Call Detail Records	25
6. An Agenda for Future Research	26
ANNEX 1: Physical Implementation of Origin-Destination Matrix Generation from CDR data	28
ANNEX 2: Additional Map Generation from CDR data	34

1. Executive Summary

This section considers the potential for development of micro-indicators of economic development as derived from Call Data Records (CDRs), detailing proposed methods for their creation and validation. This is joined by a critical assessment of the state-of-the-art in OD Matrix creation, the theoretical underpinnings of that process and the challenges that emerge. Finally we present an agenda for future research and development that may serve as foundation for overcoming challenges in using CDR data for mobility analysis, and to fully realise the extensive opportunities it opens up.

The exemplar data underpinning this report comes from a representative sample of approximately 10% (512,039 individuals) of anonymized mobile phone users active in the Dar es Salaam region, shared for purposes of mobility analysis by a major Tanzanian network operator for the year 2014. Section 2 introduces the nature of CDR, with Section 3 going on to discuss the opportunities and limitations of using such datasets for transport planning. Novel data streams such as CDRs allows for unprecedented insights at a scale otherwise impossible using traditional mobility analysis techniques such as Road Side Interviews. However, this potential is set against two key challenges - those of a technical nature and the absence of clear ethical and regulatory frameworks at National and International levels.

To contextualize those opportunities and challenges, Section 4 introduces the general methodology behind creation of OD Matrices from CDR data. Section 5 then provides a detailed breakdown of the theoretical and practical issues that must be addressed. Each of these limitations affects a different part of the data collection and analysis life cycle. Some of these shortcomings can be alleviated during the pre-processing stage, while others are remedied during subsequent analysis. These are drawn together into best practices of using CDR for the creation of OD Matrices, with an emphasis on how coordination with traditional techniques is key to providing optimal solutions in the future.

Section 6 extends these conclusions, providing a future research agenda that lays the foundation for overcoming identified limitations in the use of CDR data for the assessment of mobility.

Finally, in Annex 1, we provide a full, detailed breakdown of the technical process of generating OD Matrices via CDRs, along with aggregated exemplar results (which preserve both individual and commercial privacy). Here we also discuss techniques that are able to produce mobility maps, population/activity maps and sub-journey summaries. Additionally, we document the process of producing a web interface to outputs. Code snippets and further technical documentation found on the DECS OD Matrix Github repository¹ (for more extensive results and output maps please see the *Mobility Report* document).

¹ <https://github.com/DECS-UK/OD-Matrices>

2. Introduction to CDR Data

Call Detail Records (CDRs) are metadata (data about data) that capture subscribers' use of their cell-phones — including a timestamp, subscriber identification code and, at a minimum, the location of the phone tower that routed the call for both caller and receiver. Large operators standardly collect over six billion CDRs per day. Note that the scope of the term CDR has expanded extensively beyond the meaning implied by its original acronym ("Call Detail Records"). In current usage CDR data refers not only to calls, but to all network events (made by cell phones, tablets, etc.) which the operator records. This includes, but is not limited to, calls, sms, data usage sessions and mobile money transactions². The exact form of a single CDR is dependent on the type of network event and data retention policies within the operator.

At a basic level, the CDR data that underpin OD matrix generation are expected to consist of: a timestamp corresponding to when the event took place; an anonymized identifier representing the individual initiating the event; and a unique identifier for the cell tower providing the network service to that initiator. This cell tower is formally referred to as a *Base Transceiver Station* (BTS), which is designated a unique identifier and has an associated physical location recorded. An example of such a record is shown in Table 1.

Table 1: Structure of data common across all network event types for Call Detail Records (CDRs). In this case the record is limited to a timestamp, an anonymised identifier for the event initiator (i.e. caller) and a cell coverage area identifier.

Initiating subscriber ID	Unique BTS ID	Timestamp
jggsmit13227abc	12038097523	14-03-2014 00:01:12

When the network event *type* involves two individuals (i.e. is not a data session) then the CDR will additionally contain a further two identifiers: one for the receiving individual and one for the BTS which provides the service to the receiver. Finally, the record will typically be accompanied by additional fields specific to the type of network event. Examples include call duration, sms length, data session duration and mobile money spend. Importantly, a CDR record only contains metadata about the network event but *never any of the content transmitted as part of the exchange*.

CDR's are automatically generated by network operators for billing, network management and maintenance purposes (allowing monitoring of network usage and performance³). However, significant potential also exists in the repurposing of these records for use in other application areas. This report focuses on one such use case - repurposing anonymized and aggregated records to understand the mobility patterns occurring across a city at a fine grain level of detail

In actuality, some form of location management of handsets by network operators is standard practice if they are to provide optimal service provision. Such tracking allows operators to direct

² Bias correction for these data streams is subsequently discussed in Section 5 (in theory) and in Annex 1 (in practice)

³ D. Maldeniya, S. Lokanathan, and A. Kuramage. Origin-destination matrix estimation for sri lanka using mobile network big data. In Proc. of the 13th International Conference on Social Implications of Computers in Developing Countries, 1–10, 2015.

incoming calls to the appropriate BTS in the network with optimal speed. Location management is generally undertaken by one of the following three policies:

- **Never-update:**

Location information is not collected passively but rather in a just-in-time fashion, with all cells being 'pinged' to find out the appropriate cell to direct an incoming call to;

- **Always-update:**

The handset informs the network whenever it is moving into a new cell. While there is no paging cost herein, networks can get quickly overwhelmed by the frequent updates;

- **Location-area-update:**

This approach is a combination of the previous two, with BTS grouped together into Location Areas (LA). The operator is informed when a user moves to a different LA and the cells in a subscriber's LA are pinged when an incoming call is directed.

In the United States and the European Union, choice of which policy to utilize is also influenced by law that mandates operators keep track of handsets so as to provide emergency services with location approximations in emergency scenarios. Under enhanced 911 in the US⁴ and enhanced 112 in the EU⁵ operators have to locate users within a 50m radius in 67% of cases and 150 meters in 95% of cases. As the imposed accuracies are not achievable through BTS-only positioning, new techniques for tracking handsets have been developed. Operators can use network-centric cellphone positioning⁶ or device-centric cellphone positioning⁷ to effectively triangulate handsets. With never-update the prior can be prohibitively expensive to network operators in emerging economies due to the higher load imposed on the network, while with always-update the latter is restricted to smartphones with GPS capabilities.

The CDR datasets that underpin this report are not part of this mechanism, and fall under the Never-update policy remit. Thus, they only indicate a subscriber's location when that subscriber initiates or receives a network event. While data generated through more advanced tracking techniques described above could potentially improve the accuracy of insights generated via CDRs, they are unlikely to be available in emerging economies (due to both privacy issues and the prohibitive costs involved). Thankfully this lower fidelity, specifically the increase in sparsity due to the never-update policy, is of less concern due to the mass expansion of mobile phone usage in regions such as East Africa over the last decade. Network events logged within CDRs now occur at such scale and frequency across all demographics that their analysis alone is generally sufficient to generate detailed patterns of movement and mobility across a population. CDRs, however, should not be considered as a complete replacement for traditional approaches, with CDR data augmenting and significantly minimising the effort required to be spent on traditional approaches. The requirement for traditional approaches in parallel to CDR analysis is discussed in detail within this report, particularly with regard to bias correction and the inference of modality and trip purpose.

⁴ J. Spinney. Mobile positioning and lbs applications. *Geography*, 88:256–265, 2003.

⁵ C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33:727–748, 2006.

⁶ Existing network capabilities are used to triangulate handsets based on time, angle and distance measurements generated from signal strength and handover information

⁷ Where measurements and calculations are done locally within the handset

3. Opportunities and Limitations of CDR data for OD matrix creation

The use of CDR data for mobility analysis is of increasing interest to the transport planning community. Traditionally, OD matrices and associated transportation insights have been generated from Road Side Interviews (RSI). Due to logistics and costs, however, most manual counts involve a relatively small sample size occurring at any given location over a time period of a few hours⁸. Transport surveys, which have also proven an effective tool to collect data on travel trends, are equally costly to undertake, can be quickly outdated and potentially susceptible to observer biases and reporting errors^{9,10,11}. Survey based methods of this nature represent significantly human intensive and logically challenging operations.

Recently such endeavours have been augmented through data mining efforts within the public transport system (i.e. smart card access and ticketing to entrance and exit points) and more broadly using roadside cameras and Automatic Number Plate Recognition (ANPR). These approaches aim to increase the sample size (coverage) of the data collected. However, they are limited to a subset of transportation type in the former¹² case and constrained by the feasibility to deploy and maintain a city-wide camera based infrastructure in the latter.

The potential to use CDR data to underpin mobility analysis significantly advances the trend away from time consuming and/or resource intensive techniques. Repurposing CDR data provides several advantages: unparalleled **scale** (in terms of sample size); **coverage** (in terms of observed spatial area and modes of transportation); **spatial granularity** (in terms of the precision of origin/destination units that can be outputs) and **temporal fidelity**¹³. This occurs at orders of magnitude reduced infrastructural and human resource costs.

Generation of OD Matrices from this relatively new form of data naturally has both strengths and weaknesses in comparison to conventional approaches. The coverage, precision and eliminated collection cost that CDR analysis promises must be considered in the light of limitations on what can be derived from such data. Data that has not been collected for the sole purpose of transport transport planning increases the risk of a range of **biases** that may occur within the data - biases that need careful assessment and adjusting for. Identification of **short trips** also becomes difficult, as does identification of **mode of transport** when they cannot be directly surveyed. Such problems - inherent to passive data collection - stand in contrast to the active nature of surveys and other traditional approaches, which allow one to directly query transport motivations (such as understanding mode-choice).

⁸ F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. A system for real-time monitoring of urban mobility using cell phones: A case study in rome.IEEE Transactions on Intelligent Transportation Systems, 12(1):141–151, 2011.

⁹ R. M. Groves. Nonresponse rates and non-response bias in household surveys.Public Opinion Quarterly,70(5):646–675, 200.

¹⁰ J. C. Herrera, S. Amin, A. Bayen, S. Madanat,M. Zhang, Y. Nie, Z. Qian, Y. Lou, Y. Yin, and M. Li.Dynamic estimation of od matrices for freeways and arterials. Technical report, Institute of Transportation Studies, 2007.

¹¹ J. C. Herrera, D. B. Work, R. Herring, X. Ban,Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: the mobile century field experiment.Transportation Research Part C: Emerging Technologies,18(4):568–583, 2010.

¹² I.e. formal public transport systems where smart cards are in operation. In Dar es Salaam, for example, this would not include the extensively used bus (Dala Dala) network which operates on a semi-formal basis and loosely regulated basis

¹³ The impact of these notions is discussed in much further detail in Section 5

These challenges evidence the fact that, while there are immense benefits of using CDR data at such scale, there is still a valuable place for traditional approaches in the analysis process. Indeed, with CDR-driven matrices providing a cornerstone for mobility analysis, traditional techniques can then be used to augment results with increased parsimony and precision, fine tuning and ground truthing OD matrices. In order to illustrate the advantages and characteristics of both CDR and traditional Road Side Interview techniques, a comparison is provided in Table 2.

In the generation of OD Matrices for the Dar es Salaam region of Tanzania we have also identified specific technical challenges that must be carefully accounted for in the use of CDR data in order to realize their potential. These challenges and associated solutions are enumerated in depth in section 4, with Annex 1 documenting a practical and technical description of how these solutions were implemented.

Table 2: Comparison of Call Data Record and Roadside Interview Data for supporting the generation of Origin-Destination Matrices, adapted from Tolouei et al. (2015)¹⁴ and extended via the analysis described in section 4 onwards.

Issue	Roadside Interview (RSI) Data	Call Detail Record (CDR) Data
Type of data source	Cross-sectional (sample for a single period, normally days)	Longitudinal (cross-sectional data collected over an extended period, normally a year)
Collection Approach	Active (for purposes of Transport Planning)	Passive (for network operation purposes)
Sampling Rate	Very Low (for any given road 10-20% of vehicles, with the number of roads sampled limited by available resources and time-scales)	High (for the whole city, the percentage of population covered by network provider. This sample size is vast compared to other technologies)
Processing Costs	High (RSI are a significantly labour-intensive process, and data collection alone can take months to coordinate successfully)	Low (Once processes are in place, analysis of passive CDR data can be performed in a few days using appropriate Big Data technologies)
Variation Observed	Spatial Variation of Trips (Assumes stationarity in behaviour)	Spatial and Temporal Variation of Trips (Allows for non-stationarity in behaviour)
Data Bias	Potential for Response Bias (minimizable via careful survey design and sampling strategy)	Potential for Bias towards the profile of subscribers compared to the full population (minimizable via careful analysis of the demographics of the subscriber base)
Extrapolation of Sample	Independent and well understood (combining count data with statistical modelling where journeys traverse more than one sample site)	Dependent and under continued development (requires ground truthing data on how mobile phone user's behaviour is representative of the whole population)
Coverage of Purpose	Limited (Focussed to main roads and long trips)	Very High (All purposes apart from micro-trips covered)
Identification of Purpose	Straightforward (Via a survey question in the interview)	Complex (Requires inferring via modelling assumptions or rule-based domain knowledge)
Coverage of Mode	Limited	Very High

¹⁴ SR Tolouei, P. Alvarez, and N. Duduta. Developing and Verifying Origin-Destination Matrices using Mobile Phone Data: The LLITM Case, Proceedings of the 2015 European Transport Conference.

	(Focused on Private Motorized Transport)	(Covers all modes of transport used)
Identification of Mode	Straightforward (Via a survey question in the interview)	Complex (Requires inferring via modelling assumptions or rule-based domain knowledge)
Geographical Scope	Limited (Only movement intercepted by screenlines / cordons can be assessed)	Extremely Wide (In theory all movements outside of short trips occurring within a single BTS catchment area can be assessed)

4. Overview of OD matrix creation from CDR data

We now move on to the process of generating OD matrices from raw CDR data. At its heart, this task requires us to process the mass set of digital traces left behind by the population as they traverse their daily lives. From those traces we must attempt to reconstruct the journeys that they undertook over that period. While the number of people represented within the dataset is immense, each individual is only represented by a sparse sample of the locations they actually visited¹⁵. Therefore, technically we must transform a vast number of sparsely sampled event series (i.e. expressing each individual's location history only when a network event occurs) into a mass set of 'journey' constructs, all of which must meet pre-defined validity conditions. This set can then be aggregated, scaled and interpolated to produce our final OD matrices.

Unlike traditional analyses, the nature of this data mining approach forces us to first provide rigid, formal definitions of exactly what we mean by the terms *origin*, *destination* and *journey*. Here, both origins and destinations are subcategories of the overarching concept of a '*stop*'. A stop is defined as a set of contiguous network events that occur at the same location, over a minimum period of time. This notion is parameterized to ensure we have sufficient confidence that any stop we have detected is not a transient location, but actually a location that the individual has actually settled in.

An algorithm must consequently be developed to exhaustively mine each person's event series for such stops. Once achieved, the algorithm must next detect pairs of contiguous stops which occur at different locations and hence reveal movement. This pair can then be designated as a *journey*-the initiating stop becoming the journey's *origin*, and the concluding stop as the journey's *destination*. A further filtering algorithm can then be utilized to process the resulting journeys in order to refine results and ensure that each journey fulfills certain strict conditions (ensuring, for example, that we have sufficient confidence that the journey did not go via any *other* stops and hence represents a true direct trip, or ensuring that the generating individual was not exhibiting spurious or outlying behaviour - such as might be exhibited by the traces of bus drivers)¹⁶.

Resultant journey data is then transformed (via aggregation and counting) into an intermediary Origin Destination matrix. This intermediary only reflects the subset of journeys that we have been able to identify in our sample population (specific users of that network operator). Consequently, this matrix must still be scaled to estimate the behaviour of the population as whole. The scaling process is performed as follows: first, the journey set is processed so as to produce a model of all movements occurring across the region at specific time periods. This allows computation of traffic counts at a set of key locations at those time periods. A scaling factor can then be calculated so as

¹⁵ Discussed in detail in Section 5.2

¹⁶ Discussed in detail in Section 5.3

to minimize the error between these *computed counts* and *ground-truth* counts directly observed at those test locations.

Once scaling has been achieved, interpolation to a desired zonal representation can finally be undertaken, further protecting both individual and commercial privacy¹⁷ - and producing our final output OD matrices ready for practical usage in transport planning.

4.1 Comparison to Traditional Approaches

Prior to progressing to best practices within this process, it is important to consider the key differences between a CDR based approach compared to its traditional RSI based counterpart. As detailed above, CDR-based OD matrix creation requires the reconstruction of a population's journeys (and associated metadata) from 1. A sparse sample of their location history¹⁸; and 2. Any available external information such as transport infrastructure and ground-truthing counts.

This is in contrast to OD matrix creation from RSI which focuses on an extremely small sample of the population - but which is able to obtain fully declared journeys and metadata. For CDR such metadata has to be inferred. Traditional approaches also relinquish necessity for strict definitions of concepts such as *origin*, *destination* and *journeys* (*which are inherently nebulous*). When OD matrices are formulated from roadside interviews, such concepts may be left somewhat subjective and/or based on criteria from the individual's broad context. Nevertheless, in both cases the resultant OD matrices indicate movement from one location to another.

4.2 Summary of OD Matrix generation process

Based on a fixed definition of origins, destinations and journeys, we can now enumerate the high level process that must be undertaken when constructing OD matrices from CDR data:

1. Data Cleansing:

Raw data must be converted into network event series. At this point both any BTS and spatial regions can be merged if appropriate (e.g. to deal with intermittent towers, or towers located in such close proximity that the location of user cannot be distinguished);

2. Stop Identification:

A series of potential *origins* and *destinations* must then be extracted from each event stream. These are typically mined under a combined definition as *stops* at the individual level, with stop's being sub-categorized as *transient* or *semi-permanent*. Stops are parameterized by the number of contiguous network events required in the stop, the minimum duration time as a whole, and the maximum inter-event time;

3. Journey Generation:

For each individual, their resulting series of stops must then be converted into <origin,destination> journey pairs. At this point an appropriate score for confidence in the efficacy of the pair as a direct journey can also be attributed to it;

4. Journey Cleansing:

Any individuals with outlying behaviour must then be removed to reduce bias in the dataset. Journey's which do not meet some predetermined confidence threshold may also be removed at this point;

¹⁷ For more information see Section 5.5 *Privacy and Ethical Implications of Using Call Detail Records*

¹⁸ Sparsity here refers to the sparsity of location samples observed for each individual (for further discussion see Section 5). This is distinct from the sampling of the individuals themselves from the overall population (the challenges of which are discussed in section 5.1).

5. Metadata Tagging:

Surviving journeys may then be algorithmically tagged with metadata (such as the journey's purpose or mode of transport) via further inference. This inference is based on further processing of the CDR data or integration of external domain knowledge;

6. Intermediary OD Matrix Generation:

The resultant journey set can then be filtered dependent on the task at hand before aggregating and counting to produce an appropriate OD matrix representation. This filtering may be temporal (e.g. weekends), spatial (specific wards) or via metadata categories (such as commutes);

7. Scaling:

Raw OD Matrices must then be scaled in order to correctly extrapolate the number of detected journeys to the full population. By necessity this process requires some external ground-truthing data for validation;

8. Final OD Matrix Generation:

Finally, results can be interpolated into a specific geospatial representation. This serves to protect both individual and commercial privacy, with matrices being projected to a disjoint set of geographical regions (*zonal units*) that underpin the final output OD matrix and match the task at hand.

While this process is conceptually straightforward and produces OD matrices of unprecedented scale and fidelity, the repurposed nature of CDR data means that its properties can present a range of technical challenges in many of these steps. These technical challenges and their respective solutions, along with definitions of origins, destinations and journeys they admit are discussed in Section 5. Full details of how theoretical solutions are implemented in our final approach are provided in Annex 1.

5. Repurposing CDR Data for OD-Matrices: *Theoretical challenges and solutions*

CDR data contains significant information from which to build high fidelity origin destination matrices at a previously unseen scale. Not designed for this purpose, however, the properties of the CDR data (i.e. data is only recorded about subscriber events rather than movement directly) are not optimal and require careful processing. In addition, as data is pertaining to individuals, there holds the potential for raising privacy concerns if not handled correctly. Below we identify the challenges which face OD matrix generation via CDR data (which can broadly be grouped into five categories), and then proceed to break them down in more detail along with best practice solutions in the following subsections.

Challenges of using CDR Data for OD Matrix Generation

5.1 Population-level data completeness and scaling

CDR data is restricted to both the subset of the population which utilize cellular devices and those that choose the network provider providing the CDR data. This introduces potential biases in the population being modeled, and scaling issues in understanding how to accurately extend to represent the whole population.

5.2

Individual-level data completeness and hidden movement

CDR data only contains events when an individual initiates or receives a network event. As such movement is not directly observed, and must be inferred from individual event streams - which are themselves *sparse* (depending on network usage patterns and loyalty). This introduces technical challenges pertaining to journey reconstruction and how to handle issues of "hidden movement" in individual streams.

5.3

Location precision and false movement

CDR data does not record individual's locations exactly, but rather as a non-precise proxy via the location of the infrastructure delivering the service. This introduces various challenges including: 1. How to handle the limited precision afforded by CDR (in comparison to positioning technologies such as GPS); 2. How to integrate different granularities occurring at geographical level, occurring due to varying local infrastructure density; 3. How to accommodate for precision that varies temporally due to infrastructure issues/upgrades; and 4. How to handle issues relating to overlapping service areas, handovers and load balancing mechanisms (which can produce "false movement").

5.4

Lack of directly recorded movement metadata

In contrast to traditional surveys or RSIs, repurposed CDR data cannot directly capture many of the movement metadata, such as transport mode, route and/or motivation for undertaking a journey. To solve this, methods are required for computationally inferring and verifying those metadata, strategically augmenting this analysis utilizing alternative data sources and more traditional interview and surveys.

5.5

Privacy

Due to the private nature of the data encoded in CDR's such repurposing must only occur on data where user identifiers have been pre-anonymised by the network provider as recommended by the Groupe Special Mobile (GSMA), the mobile telecom industry body, in their emergent guidelines on the use of CDR data¹⁹. This is a necessary, but not sufficient, condition to addressing privacy concerns. Specifically, additional steps need to be undertaken to prevent the re-identification of individuals through individual specific patterns within the data and external information. This is discussed, along with the solution undertaken in this work, in section 5.5 (along with a practical implementation of the solution in Annex 1).

5.1 Population Level-data Completeness and Scaling

Challenge: CDR data contains records for all subscribers of the network operator providing the data, this however is not a complete sample of the population. This is caused by competing operators and lower than 100% population uptake of cellular services. This results in reduced datasets with respect to the population. Moreover, each CDR dataset is likely to be biased towards some sector of the population. Operators often target to different population groups, and those without cellular service often fall into lower socio-economic and/or age group demographics. This means care must be taken not to bias results towards the specific demographic of users

¹⁹ GSMA. 2014. GSMA Guidelines on the protection of privacy in the use of mobile phone data for responding to the Ebola outbreak. Retrieved March 20, 2016 from http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-Ebola-outbreak_-October-2014.pdf

subscribing to the network provider's services. As a consequence all processing results must be considered as intermediary in their raw form, offering relative, rather than absolute, origin-destination counts until scaling occurs.

Discussion & Solution: If no systematic bias exists in then scaling to the whole population can be approached by a range of measures which are discussed below (although, as will be expanded upon, we will ultimately recommend the use of ground truth camera-based measurements in conjunction with known statistics on modality ratios to determine the correct scaling factor for accurate journey count estimates for the whole population).

Research in the area of OD matrix scaling has focused on either re-purposing existing statistics (e.g. census figures) or on the strategic sampling of ground truths (at key transport locations) fusing the more traditional approaches such as road-side interviews and traffic counting into the process. One data-driven approach to scaling that has been explored in the literature is to estimate a count for residents in each region from the CDR data, and to compute a scaling factor that matches actual population sizes in those regions (commonly referred to as expansion factor) against known population statistics²⁰. This scaling factor is then used as a surrogate to scale the OD matrices counts.

While such an approach may be appealing if up-to-date population statistics exist, this is rarely the case in developing countries (and particularly not at the spatial/temporal granularity we desire). The use of this sort of surrogate scaling factor of could then actually introduces a risk of further error - the assumption that resident counts estimated from CDR data will reflect true population levels is clouded by issues of cell phone demographic usage^{21,22}.

Another alternative approach has used *mode choice probability, vehicle occupancy and usage ratios* to determine scaling factors. However, a heavy reliance on the availability of accurate traffic and demographic data in a high spatial resolution remains²³. Equivalently, cellular penetration, mobile phone non-usage and vehicle usage have been used in an attempt to calibrate scale factors in computer simulations. While this approach to scaling has shown promise when used in a micro-simulation for Dhaka²⁴, transferability from the microsimulation model of traffic points to real-world scenarios in other study areas has not been tested in detail. Driver decisions which are modelled in such simulations in particular may be less accurate in rapidly urbanizing cities such as Dar es Salaam.

As a **best practice**, we therefore recommend the identification and ground truth sampling of multiple important location points over the region, examining both different times of day and repeating on multiple days. This approach enables not only a scaling factor to be computed, but additionally provides an opportunity for confidence measures to be developed (Figure 1 shows the points taken

²⁰ S. Colak, L.P. Alexander, B.G. Alvim, S.R. Mehndiratta, and M.C. Gonzalez. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Record: Journal of the Transport Research Board, 2526: 126-135, 2015.

²¹ L. Alexander, S. Jiang, M. Murga, and M.C. Gonzalez. Origin Destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies, 58:part B, 240-250, 2015.

²² M. G. Demisse, et al. Inferring origin-destination flows using mobile phone data: A case study of senegal. In Electrical Engineering / Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016 13th International Conference on , 2016.

²³ P. Wang, et al.. Understanding Road Usage Patterns in Urban Areas. Scientific Reports, 2, 1-6. 2012.

²⁴ Md. S. Iqbal et al. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C 2014.

in Dar es Salaam). This is achieved by computing scaling factors for each ground truth sample point and considering their variance. In the practical implementation detailed in Annex 1, this approach was undertaken using video camera based traffic counting at multiple traffic sample points. Note that such an approach leads to OD matrices for vehicular journeys counts, with external modal statistics required and/or more complex surveying required to extrapolate to full modality OD matrices. There is *much potential* to automate this work via computer-vision and deep learning algorithms. When combined with deployment of a small number of fixed cameras would provide sophisticated scaling factors which vary over time and geographical region (see section 6: “An agenda for future research”).

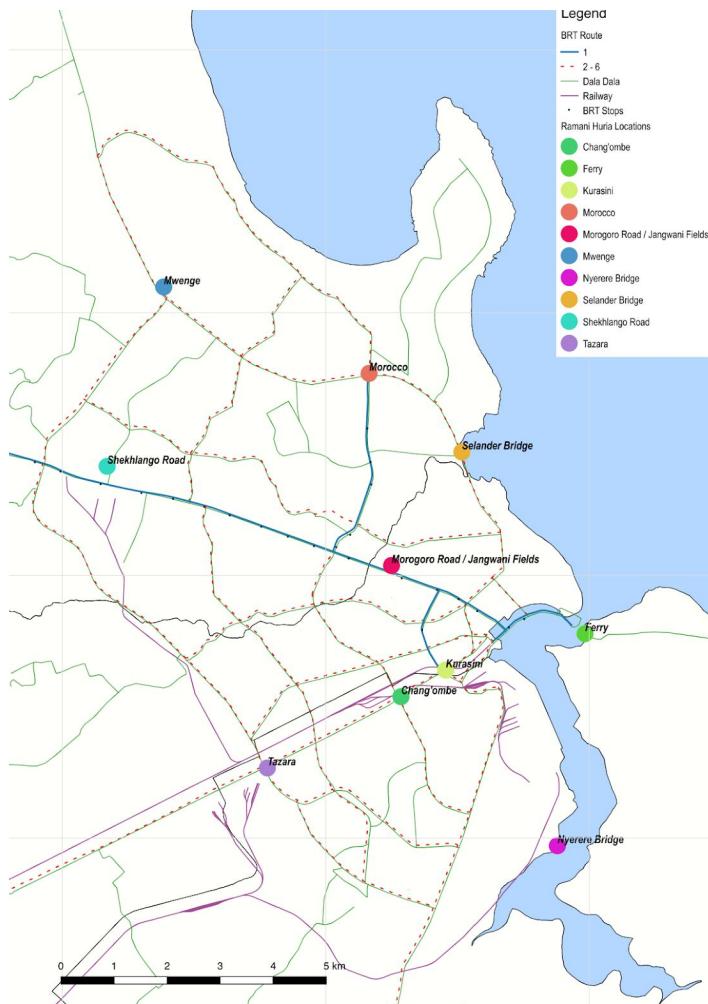


Figure 1: Example Identification of key flow points within the Dar es Salaam Region. In order to perform scaling, ground truth data was obtained at each of these points, with cameras collecting footage across multiple days at three daily time periods (morning, day and evening). Traffic counts could then be produced and cross-referenced with model estimates to produce a set of robust scaling factors.

Unlike scaling factors, issues of population bias can be complex to accurately resolve. In urban spaces in Tanzania, where mobile phone penetration has been reported to be close to 92%^{25,26}, bias in uptake may (relatively safely) be assumed to be negligible and left uncorrected. Operator bias is then required to be assessed based on the known coverage of the operator in the area of interest and a qualitative assessment to the viability of using the operator's CDR undertaken. Alternatively OD matrices could be computed assuming no systematic bias and a confidence measure for the

²⁵ Ministry of Health, Community Development, Gender, Elderly and Children - MoHCDGEC/Tanzania Mainland, Ministry of Health - MoH/Zanzibar, National Bureau of Statistics - NBS/Tanzania, Office of Chief Government Statistician - OCGS/Zanzibar, and ICF. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016. 2016, Dar es Salaam, Tanzania: MoHCDGEC, MoH, NBS, OCGS, and ICF. Retrieved December 18, 2016 from <http://dhsprogram.com/pubs/pdf/FR321/FR321.pdf>.

²⁶ Compared to a 79% penetration rate in Tanzania overall (Tanzania Communications Regulatory Authority. Quarterly Communications Statistics Report. April-June 2016 Quarter. (<https://www.tcar.go.tz/images/documents/telecommunication/CommStatJune16.pdf>)

scaling factor computed as discussed above. Low confidence in such a measure indicates error within the system potentially indicating the presence of such bias. **Best practice** must therefore currently focus on the acquisition of a provider with sufficiently unbiased population coverage (until improved geo-demographic intelligence can be sourced in emerging economies). This must be coupled with comprehensive measurement of scaling factor error to ensure fit-for-purpose and confidence. Initial studies have shown this approach to be successful with, for instance, mobility models in Kenya that have been developed from CDR data producing similar insights as manually constructed ground truth data²⁷.

5.2 Individual-level Data Completeness and Hidden Movement

In an ideal world, all of a population's movement would be available for aggregation into a mobility analysis. However, this is not the case in CDR data, which only record an individual's location when a network events occur (whether calls, sms, data sessions, mobile money transactions). This results in a sparse stream of events, containing a subset of each individual's actual behaviour. This in turn leads to several issues that must be ameliorated, of which four of the most pressing will be broken down as follows:

- ❑ *Section 5.2.1: "Sparse Temporal Frequency Challenges"* considers how to handle the issues which arise around sparsity in general, most central of which is "hidden movement";
- ❑ *Section 5.2.2: "Update Frequency Bias"* addresses the fact that some individuals' event series will also be far more sparse than others, introducing the danger that OD matrices will be biased towards those people with high usage patterns;
- ❑ *Section 5.2.3: "Bias due to Heterogeneity in Call Rates"* considers that activity in certain regions may also introduce biases - areas with forced waiting time, for example, (i.e. at ferry terminals) may result in overestimations of stop frequencies, and this needs accounting for;
- ❑ *Section 5.2.4: "Single Network Activity Issues"* considers that subscribers may have multiple SIM cards acting over multiple networks, resulting in a single network provider's CDR data only containing a biased lens into an individual's movement.

5.2.1 Sparse Temporal Frequency Challenges

Challenge: Observing movement accurately requires the continuous, precise sampling of an individual's location (e.g. as within sensors such as GPS) at low temporal intervals. In contrast, CDR data is only recorded when a network event takes place. Depending on an individual's usage pattern the time between these events could range from a few seconds to multiple hours, resulting in significant amounts of *unobserved* movement (often referred to as "hidden movement"). As such, two sequential records involving a change in location may involve any number of unobserved sub-journeys and/or periods of non-movement in between them of which we have no indication. Furthermore, a lack of samples reduces the certainty as to whether an individual has actually *stopped* in a location when we see a network event occurring at that location (i.e. are we truly observing a journey origin or destination when we examine a network event's location, or merely some intermediary point within their travels?).

²⁷ A. Wesolowski, N. Eagle, A.M. Noor, R.W. Snow, and C. O. Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. *Interface (Journal of the Royal Society)*, 10(81), 2013.

Discussion & Solution: The sparsity in recorded locations in CDR data means that not all stops in an individual's daily trajectory can necessarily be identified. Thankfully, the sheer scale of the dataset render hidden movement in a particular individual less of a problem - the key is that we have enough samples across the population to fill in the gaps, and this is where CDR data is at its strongest. In order to construct valid OD matrices, existing approaches therefore (correctly, due to sample size) focus entirely on reduction of *false positives* when detecting stops. Such an approach has the effect of actually increasing the sparsity of sequential events from which to identify journeys, but as will be discussed - this is an acceptable concession in any individual stream. Equally the sparseness of events prevents the identification of any journeys with either the origin, destination or both missing from the CDR data. Yet again ensuring that false positives (false movement) are eliminated is essential. With enough data, and under the assumption that phone usage across stops is not systematically biased at different locations (see subsection 5.2.3: *Heterogeneity in Call Rates*), then such loss in individual series is ameliorated by the sheer scale of the dataset. More pressure is necessarily then applied to determining an accurate population scaling factor (see Section 5.1), but this is entirely achievable with thorough ground truthing - and we are left with confidence that the journey's detected represent complete trips, and not sub-journeys.

An alternative solution to the issue of hidden movement is to integrate the fact that human mobility is known to be highly predictable - it is now well documented that humans tend to follow similar patterns over time, such as commuting from home to work^{28,29,30}. This fact can be leveraged by viewing a set of days (say all weekdays in a month) as a repeated *sample* of behaviour over a single day and modelling behaviour statistically as a counting process. Assuming each individual follows a similar daily pattern over this period of temporal aggregation, aggregation of this type directly can significantly decrease the sparsity of the events. Such an approach will increase and affirm the presence of regular journeys at a higher granularity (in terms of sub-journey information) at the expense of uncommon ones. While this is an extremely promising solution, research into the full effect of the modelling technique and optimal periods for aggregation is still required - and hence not undertaken in this work.

The above considerations, however, generate the following best practices. When mining for journey's the definitions provided below will provide the most robust outputs when we are availed with data at scale:

Best Practice "Journey" Definition - with respect to an OD matrix, a journey should refer to a high-confidence movement from one stop to another. The focus should be on ensuring efficacy of origins/destinations and defending against mis-casting of intermediate stops as true destinations. This policy enforces a strict answer on questions such as "if someone is travelling from one city to another, but stop for lunch, is this two trips or one?", or "Does it make a difference if this is a sandwich at a roadside petrol station while fueling the car with fuel or involves stopping in the CBD and catching up with friends?". This policy mandates that all efforts ought be made to ensure sub-journeys are *omitted* in any raw OD matrix - unless a transient matrix is actively sought.

²⁸ C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968), 1018–1021, 2010.

²⁹ R. A. Becker, R. Caceres, K. Hanson, J.M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp 11), 2011.

³⁰ G. Smith, R. Wieser, J. Goulding, and D. Barrack. A refined limit on the predictability of human mobility. *Pervasive Computing and Communications (PerCom)*, 2014 IEEE International Conference on. IEEE, 2014.

Setting high thresholds for acceptable journey identification means that we may, as a consequence, miss valid stops and hence. This raises the danger of mis-classifying movement as direct, when in reality it was composed of two or more individual journeys ("misinterpreted movement"). To combat this, our definition of a journey must also therefore enforce a temporal bounding. This requires an upper temporal limit, above which the likelihood we have missed a stop becomes unacceptable. This is accompanied by a lower limit, below which movement would have been unfeasibly fast. For how confidence and temporal thresholds can be physically implemented please see Annex 1.

Best Practice "Stop" Definition - commonly a stop is considered as a location at which an individual spends some minimum threshold of time and can therefore be assumed to have 'settled'. This *stop based approach*^{31,32} typically defines a location either as a single BTS or within a fixed radius and/or a parameterized set of nearby towers (due to service provision overlap - see section 5.3.4). Note, that if the minimum time threshold is set to zero, OD matrices will degenerate to what is known as a *transient approach*^{33,34} containing all sub-journeys - and won't truly describe movements between origins and destinations at all *per se*. Elimination of static trips/false positives is therefore the priority.

A balance is required here however - as we extend the minimum time events must span to be deemed a valid stop, we introduce the risk that an individual may have left and returned within that period ("splitting the stop"). To combat this a stop definition must also integrate the notion of a maximum *inter-event time* permissible between sequential network events at the same location. Thus, a stop is strictly defined around two thresholds: 1. the minimum amount of time at least two contiguous events occurring at the same location must span; and 2. a maximum inter-event time below which it is considered unlikely there will have been unobserved movement (i.e. where the individual will have left the location and come back). Again, physical implementations of these requirements can be seen in Annex 1.

Final Remark: While not as pronounced in emerging economies, there is an increasing shift away from SMS and phone calls toward mobile data usage. Assuming data events are logged, this transition will likely aid the generation of OD matrices, reducing the sparsity of individuals logs with the average inter-event time generally being lower compared to SMS and cellular calls^{35,36}, as data applications typically request data at significantly higher frequencies. Finally, if possible, movement events (between tower locations) could be logged by the data provider. While these are not recorded for billing and hence not typically available (as is the case in the dataset underpinning this work) working with the data providers to expose this information would eliminate many of the discussed challenges.

³¹ P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M.C. Gonzalez. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*, 2, 1-6. 2012.

³² Md. S. Iqbal, C.F. Choudhury, P.Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C*, 40, 63-74, 2014.

³³ P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M.C. Gonzalez. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*, 2, 1-6. 2012.

³⁴ Md. S. Iqbal, C.F. Choudhury, P.Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C*, 40, 63-74, 2014.

³⁵ F. Calabrese. The Geography of Taste: Analysing Cell-Phone Mobility and Social Events. *Proceedings of the International Conference on Pervasive Computing*, 22-37, 2010.

³⁶ C. Chen, L. Bian, and J. Mac. From traces to trajectories: How well can we guess activity locations from mobile phone traces?. *Transportation Research Part C*, 46, 2014.

5.2.2 Update Frequency Bias

Challenge: As CDR data is only created when network events take place, users with a high network activity (and consequently their journeys) are over-proportionally represented within the data set^{37,38}. Without further adjustments, this over-representation can significantly affect the representativeness of any predictions based on CDR data alone³⁹.

Discussion & Solution: Two options exist to deal with the issue of *update frequency bias*. The first is the identification and removal of over-proportionally represented individuals identified as outliers⁴⁰. Alternatively, journey contributions per individual to each OD matrix can be normalised by weighting contribution counts per individual by either 1. their total number of identified journeys or 2. the scale of their network activity. While this will artificially reduce intermediary OD matrices, this will have more impact on the distribution of journeys rather than total journey counts (due to the scaling approach implemented). More research is currently required in the field to understand the impacts of such weighting individual contributions to OD matrices, and thus **best practice** recommendation is currently to use the former solution - outlier removal (as we do in Annex 1). While there appears much promise in the (potentially superior) weighting approach further research efforts are required (as detailed in section 6: "A future research agenda").

5.2.3 Bias due to Heterogeneity In Call Rates

Challenge: The generation of network events by individuals may be either positively or negatively biased within certain spatio-temporal regions. This may artificially increase/decrease the probability of an origin or destination being identified *in comparison* to other areas. This may introduce errors when scaling factors are used to generalise the observed results to the population, over emphasising journeys to zones where high network usage occurs. This bias can cut both ways - an example when negative bias may occur is in locations such as home or work where landlines may be used in favour of mobile connections. Conversely, an example of positive bias is within waiting areas such as bus or ferry terminals network activity where subscribers are potentially using their phones at higher rates to 'kill time'.

Discussion & Solution: Negative areal bias due to reduced mobile usage is typically considered less of a concern within emerging economies and hence **best practice** recommendation is, currently, to leave this unadjusted. This is due to the very low number of landline available in emerging economies (With only an estimated 2% penetration across Sub-Saharan Africa⁴¹) and 'mobile first' usage patterns. Sub-Saharan Africa mobile connections have been shown to be almost 50x greater than landline connections⁴².

³⁷ T. Couronne, A.-M. Olteanu, and Z. Smoreda.. Urban mobility: Velocity and uncertainty in mobile phone data. In Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, 1425–1430, 2011.

³⁸ T. Couronne, Z. Smoreda, and A.-M. Olteanu. Chatty mobiles: Individual mobility and communication patterns. In NetMob 2011, 2011.

³⁹ J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabasi. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 41(22):1–11, 2008.

⁴⁰ H. Wang, F. Calabrese, G.D. Lorenzo, and C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call records. In 13th International IEEE Conference on Intelligent Transportation Systems, 318–323, 2010.

⁴¹ Pew Research Center. Cell Phones in Africa: Communication Lifeline. 2015, April 15. Retrieved December 10, 2016 from <http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/>

⁴² GSMA. Gauging the relationship between fixed and mobile penetration. Technical report, GSMA, 2014.

In contrast, **best practice** recommendation is to actively adjust for positive areal bias via ground truth information, identified in data sources such as public transport maps and/or ferry port and airport locations in the calculation of the scaling factor⁴³. Currently these adjustments must be guided by heuristics, but there is much potential for research to automate this and generate a dynamic scaling map (varying scaling factors both geographically and temporally) across an extent. Additionally, outlying regions (in terms of exceptional activity across all individuals in the CDR data) can be identified and investigated. Adjustments could also be made to journey counts via custom scaling factors through a small number of targeted observations using more traditional approaches such as interviews, surveys or (automated) traffic counting.

5.2.4 Single Network Activity Issues

Challenge: Any CDR dataset sourced from a single operator is necessarily limited to a subset of the population (and consequently a specific demographic distribution) who have self-selected to be subscribers to that service. Additionally, with the advent of multi-sim devices, many subscribers may be using multiple operators, with each operator subsequently only receiving a sample of overall network events biased towards certain usage situations (e.g. calls on one sim, data on another). Such a situation is not uncommon to mobile users in emerging economies. Motivating factors for multi-sim use include: 1. The ability to offset some of the issues associated with limited network coverage when travelling; 2. The mitigation of network down time, which is a particular concern in rural areas within emerging economies; and 3. The reduction of costs through the use of different SIM cards for calls to different user groups^{44,45} utilizing cheaper intra-network call tariffs and/or different SIMs for different network services (e.g. Data vs voice and sms). The bias introduced in assuming that mobility analysis generated from Single Network Activity is representative of overall behaviour can be broken down into three main challenges: 1. The systematic loss of network events due to usage behaviour; 2. The temporary loss (or gain) of network events due to the CDR operator's network (or their competitors) becoming temporarily unavailable; and 3. The systematic, per individual, loss of events in a given region as they change providers to maintain/improve service.

Discussion & Solution: Systematic occlusion of network events due to usage behaviour can be considered to be a manifestation of individuals with increased sparse temporal frequency bias and low update frequency bias. Correction for these biases is therefore often sufficient as described in section 5.2.1. If the occlusion is geographically systematic (i.e. individuals changing providers in a given region for personal and/or network reasons) then the resultant bias is similar to - and can hence be addressed in a similar fashion to - heterogeneity in call rates (Section 5.2.3). Specifically, areas of known reduced operator coverage, or operator penetration, can be selected for examination and manual correction. Additionally analysis of low usage areas, compared to known population statistics and network operator penetration rates can be done, again to select areas for examination and manual correction based on the deployment of small numbers of targeted more traditional observations.

⁴³ Md. S. Iqbal, C.F. Choudhury, P.Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C, 40, 63-74, 2014.

⁴⁴ K. Ding. The Disintegration of Production: Firm Strategy and Industrial Development in China, chapter The specialized market system: the market exploration of small businesses, pages 149–176. Edward Alger Publishing Limited, Cheltenham, UK, 2014

⁴⁵ A. M. Desai and E. Forrest. E-Marketing in Developed And Developing Countries: Emerging Practices, chapter Mobile Marketing: The Imminent Predominance of the Smartphone, pages 97–115. Business Science Reference, Hershey, PA, USA, 2013

Non-systematic events with large scale impact, such as regionwide network downtime, however require specific investigation and potential manual adjustment. Due to their global and significant impact on the CDR data for a given region these are easily identified. Corrections then may include omission of the temporal/spatial period with global rescaling to the OD matrices to account for the missing data, data imputation or another bespoke adjustment depending on the severity and spatio-temporal area affected and the regularity of the problems.

Final Remarks: Many of the discussed single network activity biases are significantly more pronounced in areas where operator coverage is reduced and infrastructure less reliable. In urban areas with good coverage and reliable infrastructure adjustments may not even be required. Finally, while not discussed, these biases are obviously less of an issue in the (atypical) situation where multiple operators have contributed to the CDR dataset underpinning the mobility analysis.

5.3 Location Precision Issues and False Movement

A key consideration when working with CDR data is that it does not specifically record the location of each device that it services, but rather the base transceiver station (BTS) that provided the service. It is therefore the physical location of the BTS station that is used as a proxy for an individual's location within CDR data. Each BTS is surrounded by a region within which it is able to provide network service. Depending on the density of the BTSs these may overlap.

While not always the case, in general subscribers will be serviced by their closest (strongest signal) BTS⁴⁶. Following discussion with network engineers and analysing with them the complexities of cell operation, **best practice** recommendation is that subscribers should be assumed to be within the "voronoi" catchment region surrounding the BTS handling their network event. A voronoi region (or cell) for a particular BTS consists of all points closer to that BTS than any other. The combination of every voronoi region is referred to as a voronoi map. An example of how such voronoi regions physically manifest is provided in Figure 2.

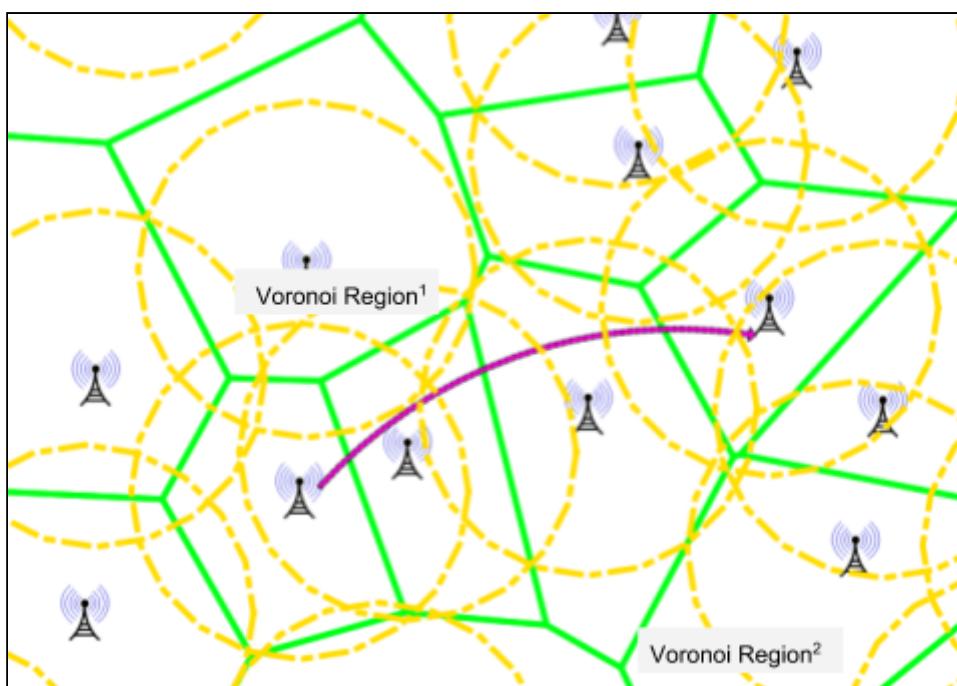


Figure 2. Potential BTS catchment area (yellow) contrasted with a Voronoi cell model, which reflects the most likely location region of a user. An example movement path of a user is illustrated in purple.

⁴⁶ P. McGuigan. GPRS in Practice: A companion to the specifications. Chapter: Operation in the Physical Layers. Wiley

While this assumption maximises the likelihood of correctly positioning a subscriber when constructing OD matrices (but certainly doesn't guarantee a correct positioning), several challenges remain, which can be categorized as follows:

- ❑ *Section 5.3.1: "Limited Precision"* examines the issues that arise due to the non-precise, areal nature of these voronoi locations;
- ❑ *Section 5.3.2: "Non-uniform BTS density"* addresses the fact that the size of each voronoi cell varies significantly across the extent of both city regions, and countries as a whole (depending on infrastructure density and broadcast capacity) - a factor which must be taken into consideration during analysis;
- ❑ *Section 5.3.3: "Changes in BTS operation"* examines how to deal with the fact that cell activity is highly dynamic, with cells are continually being added to and removed from the network. Malfunctions and consequent repairs are commonplace, and ignoring this issue can lead to incorrect analysis;
- ❑ *Section 5.3.4: "BTS service region overlap"* addresses issues which arise due to mechanisms put in place by operators to improve network performance - namely load balancing across towers with overlapping service areas. These dynamics introduce further uncertainty with regard to the recorded location and must be considered within the process of OD creation;

5.3.1 Limited precision

Challenge: The location precision of a user at the time of a network event is limited to the granularity provided by the BTS infrastructure and its geospatial layout across the extent.

Discussion & Solution: This is not a challenge per-se, but rather a fundamental characteristic of the data. Depending on the data collected by network operators, there are opportunities for geolocating handsets through available signal strength information^{47,48,49,50}. However, this information is not required for billing purposes and as such very rarely recorded within CDR data. Within urban areas, however, there are typically sufficient BTSSs in existence for us to derive quite fine-grained OD matrices. The density, indeed, is often greater than the granularity required for the zonal units that will underpin the OD matrix itself. Moving into rural areas, however, has the potential to result in OD matrices based on quite low-grained spatial quantization.

We recommend as **best-practice** that when interpolating from a voronoi cell representation to the representation used by the final output matrix (we use a grid based representation in Annex 1 for example), that interpolation occurs proportionately based on building counts - or even more preferably floor space coverage (due to the growing number of multi-story buildings being built in urban environments). While tower catchment areas can cover wide areas, calls predominantly come from areas where the human population congregates, and this centres round physical

⁴⁷ Most studies have relied on a data set provided by AirSage, a US based company using its proprietary Wireless Signal Extraction technology to anonymise, aggregate and analyse mobile phone signal data from multiple network operators to predict real-time traffic speeds and travel times. Similar companies include IntelliOne in the US, ITIS holding in the UK, Delcan in Canada and CellInt in Israel

⁴⁸ F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating Origin-Destination flows using mobile phone location data. *Pervasive Computing*, 10(4), 2011

⁴⁹ S. Jiang, G.A. Fiore, Y. Yang, J. Ferreira Jr, E. Fazzoli, and M.C. Gonzalez. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In *Proceedings of UrbComp'13*, 2013.

⁵⁰ D. Gundlegård, C. Rydbergren, N. Breyer, and B. Rajna. Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95, 29-42, 2016.

infrastructures such as buildings and transport. This gives us an invaluable heuristic for pinpointing activity geospatially with greater granularity (a physical implementation is discussed in more detail in Annex 1). An example of this is illustrated in Figure 3, which shows building centroids in the port region of Dar es Salaam. Such data is invaluable in preventing network event activity being incorrectly attributed to grid-cells which predominantly contain natural features (in this case water)..

Final remark: Moreover, and as will be discussed in Section 5.5, a level of coarseness with regard to the spatio-temporal quantisation is not necessarily detrimental to analysis, and may actually be required in order to ensure privacy preserving results.

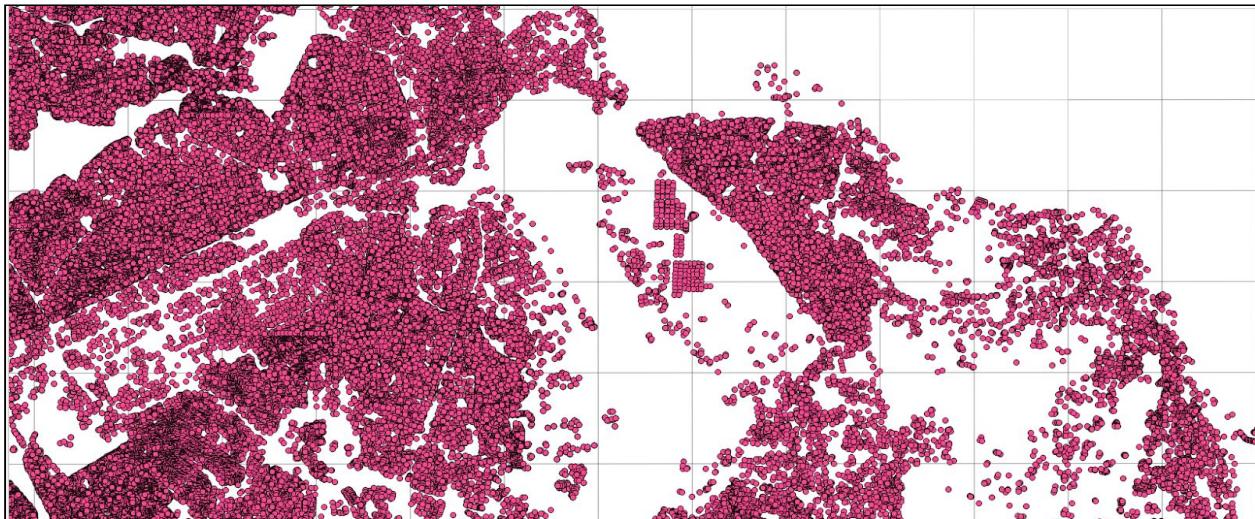


Figure 3. Example of Building Location data in Dar es Salaam Region, which provides an invaluable heuristic for performing informed interpolation of activity between voronoi regions and grid cells.

5.3.2 Non-Uniform BTS Density

Challenge: Urban spaces generally see a higher level of network activity, leading providers to operate more BTS in those areas. Due to a lower population density and associated lower levels of activity network operators therefore tend to operate fewer BTS in rural areas. This can result in BTS being located multiple kilometres apart compared to densely populated urban areas which can have a median distance of a few hundred meters between BTSs, if not less. The varying density in BTS coverage impacts the accuracy of descriptive, predictive and prescriptive insight generated outside of densely covered urban areas.

Discussion & Solution: Dealing with OD matrices with origin and destination regions of vastly varying sizes is difficult. Should one account for the size of an area when considering movement to and from it? Should there be understanding of where in a region movement is occurring from? Do variations in spatial regions affect decision making processes? Following discussions with transport experts in the UK, when constructing raw OD matrices we recommend as **best practice** the use of equally sized zonal units (which become origins or destinations respectively). Such a policy is recommended due to the ease on human interpretation that results⁵¹.

⁵¹ From a transport perspective within the UK zonal units are often constructed to equalise the number of trips generated per zone. Such an approach masks many of the discussed problems, with BTS density typically correlated with population and therefore journey density. The approach also implicitly addresses privacy concerns (see section 5.5). However, such an approach does introduce additional complexity in interpretation, requiring both size and value (colour) of the zonal units to be taken into account when interpreting the results (visualisation).

However, interpolation can also be misleading if not implemented with care. For example, in rural areas where there is coarse BTS granularity, and consequently voronoi cells larger than the zonal units being used, there is a danger of some proportion of journeys being attributed to unpopulated areas (This would be dangerous for example in the Mbezi ward, at the North West edge of the Dar es Salaam region illustrated in Figure 4). The computation of additional probabilistic origin/destinations on top of the derived OD matrices based on known building locations (i.e. from crowdsourced datasets⁵² or from automatically identified buildings from satellite or drone imagery) is therefore particularly valuable (as detailed in section 5.3.1). In regions where there are limited areas of population, this has the potential to provide significantly better insights (although, these approaches have only been trialed within urban areas as part of this work). Understanding how CDRs can be used and relevant factors for developing rural OD matrices is currently out of scope for this work, but is a topic that provides many opportunities for future research.

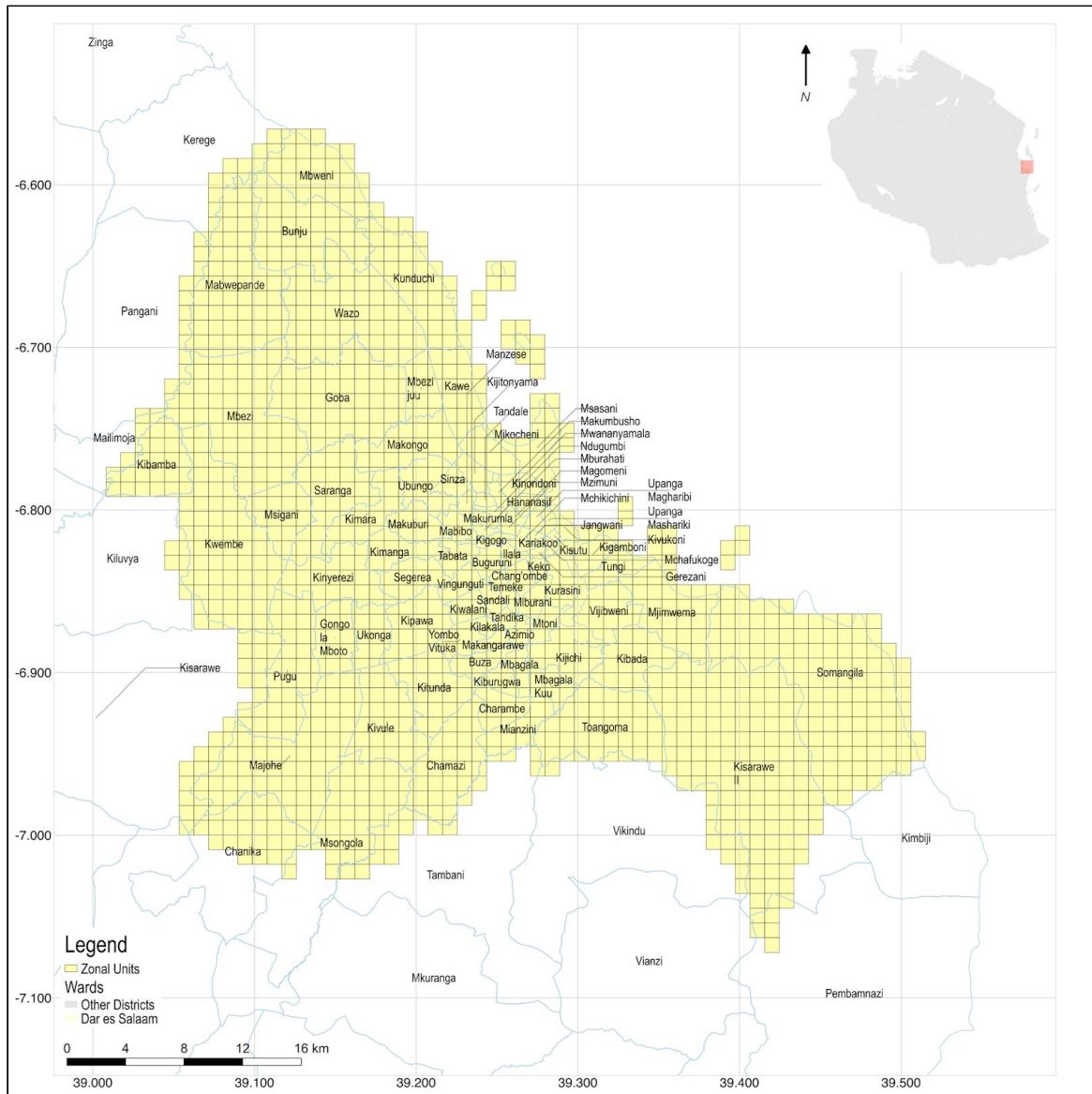


Figure 4: Example of a grid-zonal unit representation - specifically the 1km grid used for mobility analysis of Dar es Salaam.

⁵² In Dar es Salaam such information exists due to the Ramani Huria (www.ramanihuria.org) flood resilience project

Finally, we also recommend as **best practice** the merging of BTS activity in high density areas (in Annex 1 we use a 75m cordon, below which any adjacent BTS are merged). This is due to a deleterious reduction in confidence in high density areas, which are subject to greater load balancing, that the BTS's associated voronoi region accurately reflects a user's true position⁵³.

5.3.3 Changes in BTS Operation

Challenge: BTS infrastructure can change significantly over time due to malfunctions and/or the installation of BTSs to meet demand. Such alterations alter voronoi cell construction (see Section 5.3) and directly affect the precision of location, with nearby towers often taking over service in the case of malfunction. Depending on how the BTS locations have been mapped to form regions within a OD matrix this can introduce significant error, with the same device location being mapped to different locations within the OD matrix. Equally the addition of BTS means that locations that were never previously reported become active at the expense of others. As such the reported locations should not be used directly as origin/destination points as they usage within the OD matrix will change systematically with the introduction and malfunction of BTS.

Discussion & Solution: In order to account for changes in BTS operation, **best practice** is to enforce that the origin/destination regions (zonal unit representation) formed as part of the OD matrix layer must be independent from the BTS locations. If only minor BTS changes are observed within the analysis timeframe then the issue can be addressed by a simple one-to-one mapping after removing BTS that do not meet an operational threshold (i.e. number of days active) and redistributing their events. Otherwise a more complex approach is required. This involves attributing journeys for a given BTS voronoi region (computed based on active BTS at the time of the journey) proportionally to all OD regions for which they overlap, based on some measure of overlap. Example measures include area, or if available, external information such as building counts. Note that the generation of such proportional mapping is often required anyway, with the generated origin/destination regions often desired to be different to the BTS regions for both commercial sensitivity, privacy (see Section 5.5) and or interpretation (i.e. visualisation by wards) reasons.

5.3.4 BTS service region overlap

Challenge: Depending on the locations of the BTS infrastructure two or more BTS service regions may overlap, within which either BTS could provide network service to a subscriber. In general subscribers will be serviced by their closest (strongest signal) BTS and therefore **best practice** is to assume (in lieu of external data) that subscribers are located within the voronoi region of the BTS handling their network event. However, this is not always the case, with users potentially being transferred to a nearby BTS able to provide service in order to balance load on the mobile network or for other service quality reasons. As location is based on the BTS information contained in a CDR log, this effect leads to the mis-identification of a user's location, often referred to as "false movement". These rapid *handovers* occur without the knowledge of the user and can appear in CDR data in the same way as physical movement. Moreover these non-movement handovers can occur at high and varying rates across the extent, potentially generating a non-trivial amount of data that (without consideration) could be incorrectly interpreted as movement - and heavily bias OD matrices to over-represent local journeys. This is particularly true in urban areas of high BTS density.

⁵³The detection of mode and purpose, and the resulting complications due to the data characteristics are discussed in Section 5.4

Discussion & Solution: An effective way to address issues of false movement is within the stop identification stage of the OD matrix creation. **Best practice** recommendation is to address this issue by preventing artificial stops being identified when such events are occurring. As detailed in section 5.2.1, this is achieved by the requirement for a stop to consist of a minimum two or more consecutive co-located events separated by a minimum (parameterised) time period. This is also part of the mechanism utilized to address sparse temporal frequency bias, sharing the same caveats and justification. Most previous studies have employed a minimum temporal filter of 10 minutes between records to account for false displacement^{54,55}, however determining this threshold from ground-truthing is the optimal approach if possible. In addition to this, we also recommend application of a low *minimum inter-stop time* (set to <2 minutes in the implementation in Annex 1), which directly removes handovers from being considered journeys (n.b. if this minimum inter-stop threshold is set to zero, we will again return to a transient rather than a stop based OD matrix).

An alternative solution is the exclusion/inclusion of surrounding BTSs as part of the stop detection process (as these are the towers where load balancing will occur). Due to the non-uniform BTS density (see Section 5.3.2) such an approach is only applicable in densely served environments, however, and has generally only been attempted with geolocated CDR data with a spatial filter between 300m and 1km based on accuracy of device collecting location data⁵⁶. Specifically an exclusion approach would simply ignore events from surrounding towers when looking for a subsequent co-located consecutive event to fulfil the definition of a stop, while an inclusion approach would count surrounding BTSs as being co-located^{57,58,59}. The impact of this approach, however, currently requires further investigation.

5.4 Lack of directly recorded movement metadata

Challenge: Traditional OD matrices constructed via Roadside Interviews (RSI) or other survey based approaches typically contain additional meta-information such as trip purpose, vehicle type/transport modality and vehicle occupancy. This data is not explicitly available from CDR data.

Discussion & Solution: Despite not being directly available, it remains possible that in the future algorithmic improvements will allow meta-information to be inferred from journey data via the integration of assumptions and/or additional data sources. The best methods to do this are necessarily specific to the metadata required - so no specific best practice can be provided for all potential metadata. However, even when inference is employed, a general **best practice** is to perform that analysis in a supervised fashion and not solely rely on heuristics (such as movement speed). This means obtaining a sample of ground truth metadata via roadside interviews and/or

⁵⁴ F. Calabrese, G.D. Lorenzo, L. Liu and C. Ratti. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *Pervasive Computing*, 10(4), 2011.

⁵⁵ Md. S. Iqbal, C.F. Choudhury, P. Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C*, 40, 2014.

⁵⁶ Most studies have relied on a data set provided by AirSage, a US based company using its proprietary Wireless Signal Extraction technology to anonymise, aggregate and analyse mobile phone signal data from multiple network operators to predict real-time traffic speeds and travel times. Similar companies include IntelliOne in the US, ITIS holding in the UK, Delcan in Canada and CellInt in Israel

⁵⁷ F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating Origin-Destination flows using mobile phone location data. *Pervasive Computing*, 10(4), 2011

⁵⁸ S. Jiang, G.A. Fiore, Y. Yang, J. Ferreira Jr, E. Fazzoli, and M.C. Gonzalez. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In Proceedings of UrbComp'13, 2013.

⁵⁹ D. Gundlegård, C. Rydbergren, N. Breyer, and B. Rajna. Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95, 29-42, 2016.

camera-based analysis at key locations from across the region being analysed (with the number of interviews and locations deployed on a best effort basis to achieve a feasible accuracy cost trade-off). Nonetheless this approach remains an open machine learning research task.

In order to provide *trip purpose* metadata, **home** and **work** locations must first be inferred (and indeed, this can be achieved in a relatively robust fashion). This enables the subsequent identification of journeys between home, work and other locations⁶⁰. Inference of these location types can be done jointly based on measures of occupancy frequency and consistency (BTS usage entropy) during hours known to correlate with times people are at home and work^{61,62}. While such inferences can never be 100% accurate (i.e. night workers will be incorrectly inversely labeled), this approach provides insights of high utility at limited cost when performed at scale.. Note that further refinements could be undertaken (i.e. to identify night workers based on known areas of night work), however, this attempts to return the definition of home and work to a more informal vaguely defined concept which may undermine explicit insights, and this we recommend maintaining a focus on day and night BTS modes for each anonymised individual.

Provision of transport *modality*, in contrast, requires the CDR data to be augmented with additional information. Here inference is complex, and requires extensive further research when performing on CDR data. Hence we recommend a **best practice** of targeted sampling of locations for roadside interviews to determine the proportion of different modalities. This can then be linked to estimated traffic counts over the same road/path segment based on a given OD matrix by utilizing routing algorithms over external transportation network data⁶³. These, in turn, then provide a scaling factor at each location from which to arrive at population values. Error can hence be measured and mitigated with increasing confidence in the number of locations (repeatedly) sampled on a best effort basis trading-off confidence and cost. Again in the future the requirement for roadside interviews may be reduced, and instead replaced with computer vision algorithms utilizing fixed roadside cameras and/or satellite or drone imagery.

5.5 Privacy and Ethical Implications of Using Call Detail Records

Challenge: CDR data contains significant information regarding an individual's behaviour. The creation of OD matrices from such data is therefore required to be done in such a way to as not violate individual's right to privacy including, but not limited to, meeting any legal or regulatory obligations.

Discussion & Solution: In emerging economies limited ethical, legislative and regulatory frameworks exist, if at all. Certainly, **best practice** recommendation is to use the guidelines⁶⁴ from the GSMA as a

⁶⁰ Further location types have been considered as part of activity-based approaches but found to be causing ambiguous relationships between mixed-land use and activity types. Trip-based (i.e. stop or transient) are generally chosen over activity-based approaches due to the aforementioned relationship and complexity of implementation involved. Further reading for activity-based approaches can be found in Transportation Research Board. Activity-based travel demand models: A primer. Strategic Highway Research Program 2015.

⁶¹ S. Phithakkitnukoon, Z. Smoreda, and P. Olivier. Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data, PLoS ONE, 7(6), 2012.

⁶² M. Dash, H.L. Nguyen, C. Hong, G.E. Yap, M.N. Nguyen, X. Li, S.P. Krishnaswamy, J. Decraene, S. Antonatos, Y. Wang, D.T. Anh, and A. Shi-Nash. Home and workplace prediction for urban planning using mobile network data. Mobile Data Management (MDM), 2014 IEEE 15th International Conference on, pp. 37–42, 2014.

⁶³ For instance from openstreetmap in Dar es Salaam

⁶⁴ Developed following the Ebola epidemic in Western Africa, the guidelines are specifically for the usage of CDR data in "exceptional circumstances". However, it is the experience of the DECs team that these guidelines are the de facto guidelines used regardless of situation. The guidelines are available from:

fallback to this situation (most network operators are GSMA members). Of relevance to the process of OD matrix creation are the guidelines that require *only* network operators have access to non-anonymised CDR data: that analysis will not be undertaken that could single out identifiable individuals and that the resultant output is non-sensitive.

The first of these guidelines is met without compromise, as OD matrix creation does not require the identity of any individual at any time, only aggregated movement patterns. One concern may be that in some cases the sparsity of the data may enable the movement patterns to be re-attributed to a specific individual using external information. This concern is ameliorated by both the granularity of the data, location granularity is broadly and implicitly limited to an area covering significantly more than one person due to BTS placement, and the vetted nature and agreement of the individual's undertaking the analysis to not attempt such re-attribution. This latter requirement on processing additional satisfies the second requirement.

The final requirement, that the resultant output is non-sensitive requires additional checks on all produced OD matrices and their associated metadata. Here, our best practice recommendation is use of k-anonymity, a common method to producing non-sensitive output. Specifically, *k*-anonymity, ensures that any individual cannot be distinguished from at least $k - 1$ other individuals⁶⁵. Since OD matrices do not store information about individuals, but rather counts of journeys between if you know the time someone left a zonal unit, to preserve k-anonymity you need to ensure $k - 1$ other people left the area over the period being considered to preserve privacy⁶⁶. Now, as these technologies develop, presented periods will likely increase in granularity, with k-anonymity placing constraints on what can be exposed to ensure confidence that privacy is being maintained. This is of course always an arbitrary choice - others would no doubt be content with a much lower *k* value, however given the exploratory nature of this work and general novelty we advise to err on the side of caution.

regions this involves ensuring that counts within each cell are greater than $k \times m$, where m denotes the maximum number of journeys for any given individual. I.e. that certain journeys could not be attributed back to any single $k - 1$ people as the only people undertaking such journeys. This represents an easily computable upper-bound to preserving privacy in this fashion. However, for large temporal periods this may result in significant amount of data being omitted. An alternative, when available and a significantly large temporal period is reported, is to utilize known population statistics of the regions⁶⁷, ensuring that the population in each region is greater than k . While users can potentially be identified through their preferred location visits, this risk is significantly reduced when using CDR data due to the coarse spatial granularity involved⁶⁸. Traditional approaches to developing OD matrices in the UK typically enforce 12-15 journeys per region to ensure privacy, suggesting a similar value for *k* would be appropriate.

<http://www.gsma.com/mobilefordevelopment/country/global/gsma-guidelines-on-the-protection-of-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-ebola-outbreak>

⁶⁵ B. Gedik, and L. Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing*, 7(1), 1-18, 2008

⁶⁶ Note, in the outputs of the accompanying mobility report a whole year is used as the aggregate period, so the minimum granularity we have to address is "weekends" (representing 104 days out of 365 days in the year). The amount of journeys for even the zonal unit with the lowest movement is ~1000 vehicular journeys, providing a k-anonymity of ~1001, giving extensive confidence in privacy preservation.

⁶⁷ In the implementation part of this work this is approximated by building counts as no reliable population estimates were available. In this case *k* represents a requirement that journeys can not be reattributed to individuals frequenting/residing in $k - 1$ buildings.

⁶⁸ Y. de Montjoye, C.A. Hidalgo, M. Verleysen, and V.D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, vol. 3, 2013.

6. An Agenda for Future Research

Through identifying the state of the art in OD creation via CDR data, further avenues of research have been identified that will push mobility analysis into its next stages of effectiveness, especially in emerging economies. These are listed below (although are not laid out in any order of priority).

Promising Research and Development areas

6.1 Country/Region-wide Origin Destination Matrices

As demonstrated in this work, research into using CDR for OD and mobility analysis primarily occurs in cities. However, it has become clear in this analysis that it would only take limited research to expand these processes to produce mobility maps for **all of the large urban areas of Tanzania** - and indeed the whole country, including **rural areas**. Additionally, it was observed that a highly significant number of journeys passing through Morogoro represented destinations at industrial zones within the city, indicating analysis of country-wide **industrial transport flows** may also become possible.

6.2 Dynamic Generation of Demographic Maps

National statistics Institutions across the developing world are often poorly resourced, with insufficient capacity to meet the demands of rapid urbanisation and population growth. The effect of this for cities such as Dar es Salaam, means that demographic information is sparse if available, though commonly is non-existent or not temporally accurate. Even when it is collected it is likely to be rapidly out of date. It is our strong belief that **CDR data can reveal these demographics** (when combined with state of the art machine learning techniques, but which are readily available), extending OD mobility maps to **full demographic information maps**, that are data driven and hence updateable at low logistical cost every quarter. Examples of the new (neo) forms of demographic informations that this could add to the advanced mobility insights gleaned here, are insights into: **household expenditure** (through mobile money); estimates of **health and educational levels; employment estimates**; dynamic analysis of **unplanned urban expansion**; predictions of **population growth**.

6.3 Advanced Mobility Analysis Via Computer Vision/Drone Technology

Ground truth measurements sampled across the road network are essential for correctly scaling the generated OD matrices and correcting for bias. Collected by traditional techniques these are time consuming and only a minimal number are taken. However, such counts can be done semi-automatically through the huge advances recently made in **machine vision** (and “deep learning”). Employment of **cameras/drones** in combination with these computation techniques will allow *significantly* more samples to be taken on an ongoing basis. This increase in sample size and temporal granularity would enables the generation of much higher confidence levels in resulting matrices - and provides the opportunity for low cost **continually updated** matrices. Such analysis would also then significantly improve **transport mode analysis** when combined with CDR data, as well as supporting demographic analyses detailed in 6.2, as well as further infrastructural issues (such as road condition analysis).

6.4 Advanced Mobility Insights via Topic Modelling

While CDR data allows for extensive developments in the generation of OD matrices, the extra fidelity it brings can actually complicate issues of interpretation and visualization (for example in Dar, prior analyses have concentrated on around 10 super-regions and flows between them, which made sense given the precision and size of the datasets being used). Topic Modelling is a relatively new technique used to summarize complex datasets for purposes of analysis. It is based on a mathematical approach called Latent Dirichlet Allocation that, if applied to OD matrices, would **reveal underlying movement trends and key transport patterns** occurring across the extent that were: 1. Previously lost inside the data, even when

contained in OD matrices; 2. Directly translatable into transport policy insights; 3. Able to provide new visualizable descriptions of the city.

6.5

Financial Flow Analysis

CDRs can include data on Mobile Financial Services, with associated location. Accordingly, the mobility of money across a city can now be identified. From this point, it would be possible to identify not just the 'home' areas of the population, but also where money is spent - and vitally how that **money flows** around the city, the region, and whole country.

ANNEX 1: Physical Implementation of Origin-Destination Matrix Generation from CDR data

Having introduced the nature of CDR data, and the opportunities and challenges that emerge in using it as a cornerstone for mobility analysis, we now present a practical and technical description of how we implemented the approach (and specifically addressed those challenges for the Dar es Salaam case study region). The technical process of how OD Matrices were generated from CDRs data is described in 7 steps below (each step corresponding to the high level summaries that were listed in section 4.2):

1. Data Preparation:

- a. **Raw Data Collation:** The first stage of analysis required that 5 network event tables were extracted from the raw CDR data tables shared for purposes of mobility analysis by the network provider. A high performance SQL instance is sufficient for these purposes:

- incoming call events;
- incoming sms events;
- outgoing call events;
- outgoing sms events;
- data events;

As the region under consideration was Dar es Salaam, each table either listed all events where the initiating tower was located in the Dar Region (outgoing events) or the receiving tower was in Dar es Salaam (incoming events). Note that this means individual communications can appear twice, but with a different anonymized subscriber id in each case (the initiator and recipient respectively). Any subscriber who does not visit Dar es Salaam at least once is filtered out.

- b. **Master Event Table:** An SQL union of these tables (UNION ALL) was then taken to produce master network event table, keyed by the anonymized subscribed id and a timestamp.
- c. **Distinct Subscribers:** A list of distinct subscriber ID existing in this combined table could then be extracted - this list represents all anonymous subscriber ids observed as being active in Dar es Salaam at least once over the period (totalling just over half a million individuals in our selected sample).
- d. **Data Cleansing:** At this point both any BTS and spatial regions were merged as follows:
 - i. **Proximity Merging:** In order to deal with the intermittent BTS any towers not active for at least than 300 days (82% of the year) had their activity with the nearest fully active tower, based on euclidian distance (this is a direct implementation of the solution detailed in section 5.3.2);
 - ii. **Proximity Merging:** In order to deal with sets of BTS located in such close proximity that the location of user cannot be distinguished the activity of all towers within 75m of each other was merged with that of the most active tower in the set. (this is a direct implementation of the solution detailed in section 5.3.3);
- e. **Event Series Extraction:** Just over half a million event series were then produced by filtering the master event table for each of the distinct anonymized subscribers occurring.

2. Stop Sequence Generation

- a. **False Movement reduction:** At this point any events occurring at a new tower that occur within a duration, s , of a prior event at a different tower are assumed to be load balancing artifacts and

removed from the dataset. This was set at a very conservative value of 2 minutes to ensure protection of false movement (a direct implementation of the solution discussed in section 5.3);

- b. **Stop Identification:** A distributed python script was then used to convert each event series into a sequence of valid ‘stops’. A stop is a set of at least k contiguous events which are all recorded at the same BTS over a period longer than duration, d . The maximum gap, g , between events (known as the *max_inter_event_time*) is also specified (n.b. hours between 1am and 6am are not included in this gap time as they reflect sleeping periods where minimal network activity occurs). The parameterizations used in the Dar es Salaam case study are listed in Table 3:

Item	Symbol	Value	Description
minimum number of events	k	2	The minimum number of consecutive events required within a given region for the segment to be considered a stop.
minimum permissible duration	d	10 mins	The minimum permissible duration between the maximal and minimal times of a set of events with the same tower_id to consider them a stop.
maximum inter-event gap	g	4 hours	The maximum time between any two network events before they are considered to be non-consecutive events due to the high probability of unobserved movement.
minimum inter-event gap	s	2 mins	The minimum time permitted between events. This is set to prevent detection of false movement from identified stops, due to handovers or other network load balancing.

Table 3: Parameterization of the stop identification process for mobility analysis of the Dar es Salaam region.

- c. **Confidence Assessment:** All stops were committed to an SQL table along with a value representing our confidence in each stop’s efficacy. This confidence score was calculated as follows one minus the ratio of the largest event gap to the whole duration period of the stop.

3. Journey Generation

- a. **Journey Identification:** Stop sequences were then converted into a set of journeys for each subscriber. A journey is defined as 2 contiguous ‘stops’ that are separated by a period of at least t_{min} but no more than t_{max} . The value for t_{min} ensures that false movement hasn’t slipped through the net(a double check of the solution discussed in section 5.3). Setting t_{max} ensures that we are unlikely to have reconstructed a journey that, in actually, is missing a midpoint destination.
- b. **Confidence Thresholding:** Journey’s which do not meet a predetermined confidence threshold may also be removed at this point. Confidence is determined as the mean of the respective confidences attributed to the journey’s origin stop and destination stop - and in this instance was enforced to be greater than 0.1. Note this favours journey’s whose origins and destinations contained more events occurring over a longer period.

4. Journey Cleansing

- a. **Outlier Removal:** The dataset is then filtered to remove individuals with outlying behaviour (i.e. disproportionately high network usage) who would hence skew the representative nature of the results due to the disproportionately large number of journeys attributed to them (this is a direct implementation of the solution discussed in section 5.2.2). The threshold for outlying behaviour, i_{max} , is subjective. Following observation of the distribution of journey frequencies over the year, a value of 4000 was selected. We hence filtered out all subscriber who were detected as having undertaken

more than 4000 journeys over the year (over 10 per day - one would expect many of these users would be driving public transport).

5. Metadata Tagging

- a. **Day/Night mode detection:** In order to detect journey purpose metadata, and allow us to categorize journeys as either: **Home Based Work (HBW); Work Based Home (WBH); Home Based Other (HBO); or Non Home Based (NBO);** the CDR data was processed to identify subscribers where we could, with high confidence, designate their Home and work based BTS. This was achieved by identifying the tower that they used the most at night and in the day respectively (night/day modes).
- b. **Filtering:** Once modes were obtained for each subscriber, those who did not meet a sufficient threshold of observations, or whose normalized entropy⁶⁹ (essentially their variance in tower usage) was too high were discarded. This was parameterized as detailed in Table 4 below:

Item	Symbol	Value	Description
min observations	o	50	The minimum number of observations that we must have had of the individual at the designated mode BTS.
maximum normalized entropy	H	0.5	A representation of the spread of towers used by a subscriber over the specified period. A high entropy means the individual is seen at a higher variation of towers, which in this instance is undesirable as we are less confident as to which is the person's true home or work location.

Table 4: Parameterization of the day and night mode identification process for journey purpose tagging

- c. **Tagging:** Finally all origins and destinations for just under 200,000 subscribers in our sample who survived this process were tagged with a home or work cell as appropriate, hence allowing categorization of a subset of journeys into HBW, WBH, HBO and NBO types.
- d. **Aggregation:** This process additionally allowed us to visualize the distribution of residences and workplaces across the extent. This is illustrated in Figure 5, however far more detailed renderings and population statistics are supplied in the accompanying report "Mobility Insights in Dar es Salaam".

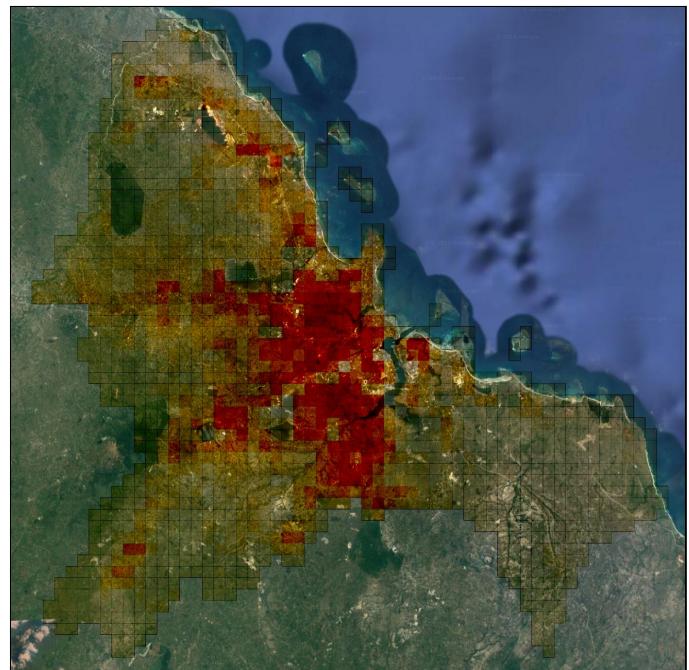


Figure 7. Visualization of the number of homes detected in each cell, across the Dar es Salaam case study extent.

⁶⁹ normalized entropy provides an indication that the subscriber didn't settle in one place for the majority of their time and therefore we are unable to predict that the most common location was their home/work cell with sufficient confidence.

6. Intermediary OD Generation

- a. **Journey Aggregation:** Finally all anonymized subscriber information was completely removed from the entire data, and a union taken of all sets of journeys across the subscriber base. This produced our entire journey set for the period examined.
- b. **Matrix Generation:** From this journey set a corresponding OD matrix was created (via a simple process of counting the number of journeys identified between each BTS). The input journey set used was also varied through filtering to allow various forms of analysis: time of day, day of week, time of year, minimum journey duration, minimum journey distance or journey purpose (based on work, home or other), etc. However, at this point journey counts are restricted to our sample size.

7. Scaling

In order to convert our intermediary, relative OD matrices into absolute counts representing the whole population, a process of scaling has to be implemented. This process as implemented in the Dar es Salaam case study is described below:

- a. **Obtaining Ground Truth Traffic counts:** Raw OD Matrices must then be scaled in order to correctly extrapolate the number of detected journeys to the full population. By necessity, this process requires some external ground-truthing data. This was performed using video camera based traffic counting at the multiple traffic sample points listed below (please see Figure 1 for a map of these points).



Figure 8. Chosen survey points

Figure 9. Example setup of traffic camera

The points of Figure 8 were chosen due to their key locations on arterial roads in consultation with the Dar es Salaam Master Plan team with attendees of Dar es Salaam City Council and JICA. The collection of videos was conducted in a manner similar to that identified in Figure 9.

- b. **Calculating Scaling Factors:** In order to determine a scaling factor ground truth traffic counts were taken at ten different locations across Dar es Salaam (incoming and outgoing traffic for 8 locations,

incoming only for 2 with an average of 6 hours over three periods in the day, for multiple days). All locations selected were major roads. For each location a scaling factor was computed as follows:

- i. Computing the route between every region pair using the OpenStreetMap road network information modified to include the Kigamboni ferry. For each region routing was undertaken from the closest routing point to the center of the region. If the route taken included the observed road segment then the journeys between the regions as indicated by the OD matrix were attributed to the estimated journey count. In order to get a robust estimate of the scaling factor the average journey count for a day (24 hrs) over the year was taken;
- ii. The observed data was extrapolated to an estimated count over the 24 hour period and scaled up to represent all journeys (in contrast to just motorised transport) using a factor based on the proportion of motorised to non-motorised transport reported in the Road Side Survey of 2007 conducted by JICA., in combination with their reported statistics for passenger ratios for different modes of transport.
- iii. A scaling factor was then computed as the multiplier required to adjust the OD estimated observation count at each location to the figure based on the observed data. A global scaling factor was then computed as the mean of these;

8. Output OD Generation

- a. **Determining a Zonal Unit Representation:** Finally, results could be interpolated into the specific output geospatial representation. This serves to protect both individual and commercial privacy, with matrices being projected to a disjoint set of geographical regions (zonal units) that underpin the final output OD matrix and match the task at hand. To move from voronoi cells to grid ‘areas’ first a grid of dar es Salaam was settled upon. This grid contained cells of 1km^2 , covering a wide spatial extent and broadly covering areas where buildings or transport infrastructure evidence human population activity. Any grid-cells that were generally determined “building less” were deemed safe to be removed from analysis and/or merged into external regions.
- b. **Fine Tuning the Representation:** The grid was clipped to the shape file of the country (providing a smooth coastal outline), and clipped to only incorporate wards with the Dar es Salaam administrative region. A selection of 30 external zonal units were then also manually crafted to allow a non-grid addition to the zonal unit representation for: 1. Main through routes in the Dar es Salaam region; and 2. To reflect representative BTS across the rest of the country. Much of this process occurred within the QGIS environment, resulting in production of the regional-representation.
- c. **Interpolation:** conversion of journey counts from the voronoi cell representation to this new hybrid 1km^2 grid + external zonal unit representation occurred as follows. First the grid was “sharded”, layering the grid and voronoi representations on top of each other and ‘cutting out’ areas that intersect (i.e. the shards). This produces a grid/voronoi cell ‘cover’ mosaic, as illustrated in Figure 11 (where darker colours reflect the amount each component of a grid cell is representative of its parent voronoi cell. The darker the component is the more it is ‘sucking up’ the BTS’s data).

A weight can then be attributed to each shard, reflecting how much of its parent voronoi cell it represents. This weighting could be based on a number of methods:

- the proportion of spatial intersection;
- the proportion of the voronoi cell’s overall building counts the shard intersects with;
- the proportion of the voronoi cell’s overall floor space the shard intersects with;

In our analysis we use building space to calculate a shard’s weighting factors (as per the solution recommended in Section 5.3.1).

For scalar statistics (such as number of residences estimated in a cell), a figure for each grid-cell can then be calculated by 1. first estimating a value of the statistic by multiplying the count for the shard’s

parent voronoi cell by its weight; and 2. adding up the resulting values for all shards contained in the grid cell. However, for journey interpolation, things are slightly more complicated. A value for the number of journeys between each <origin-cell, destination-cell> pair can be calculated by:

- i. Estimating all <origin-shard, destination-shard> counts, by multiplying the journey count *between* both shards parents by *both* shards weightings;
- ii. Summing these counts for every shard contained in the origin and destination grid cells;

Once calculated any cell to cell route can be constructed in exactly the same manner, working out the shard to shard weighted (multiplied) journey counts it contains, and adding them together. This produces a new OD matrix for each voronoi, in grid-cell format, as illustrated in Figure 12.

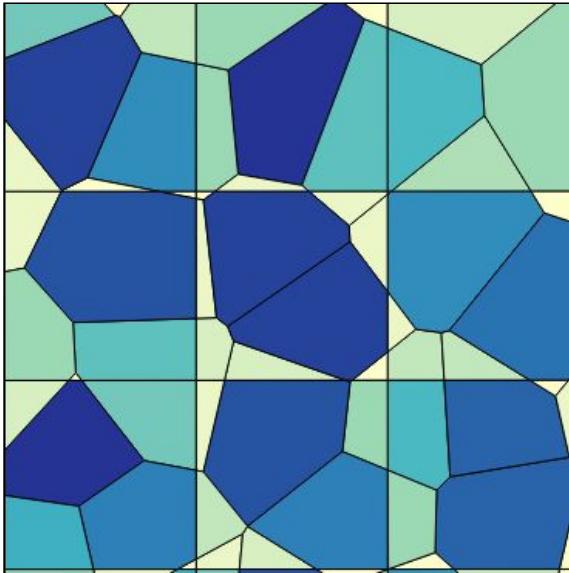


Figure 10: A synthetic example of Sharding to allow interpolation between voronoi cells and grid cells. Here the shade of the shard reflects the percentage of buildings the shard covers relative to the whole voronoi cell the shard intersects with. This provides a weighting that can then be used to interpolate with.

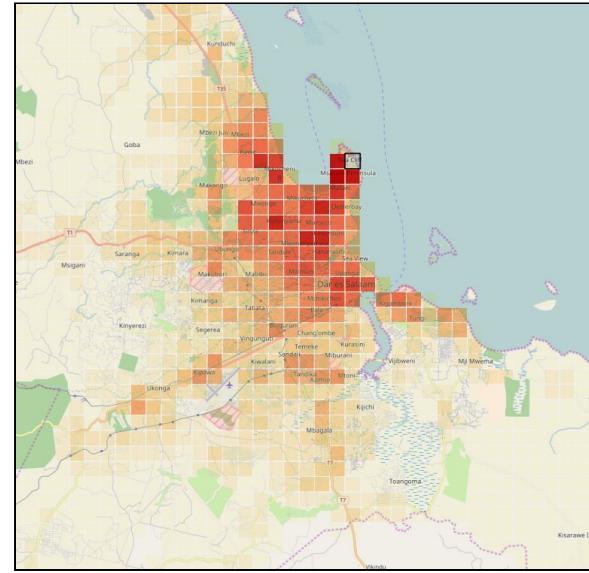


Figure 11: An example of an OD matrix once it has been interpolated into the grid-based zonal unit format (via the shard format intermediary illustrated in Figure 6). This OD map represents destinations of journeys leaving the masaki peninsula area of Dar es Salaam.

- d. **Collation:** by applying this interpolation process we produced 9 different OD maps for each of the 1527 zonal grid-units (a total of ~14,000 maps). Each analysis visualizes 9 different OD maps for the focus zonal unit, incorporating representations of:

- inbound journeys;
- outbound journeys;
- commuting patterns to and from the zonal unit;
- temporal breakdowns;
- summaries of inbound traffic from external regions;
- impact on transport flows via routing densities.

A much richer description of these results is provided in the accompanying report “Mobility Insights in Dar es Salaam”.

ANNEX 2: Additional Map Generation from CDR data

Following the methods of core OD Matrix generation described in Annex 1. Numerous insights and maps can be subsequently created from the data.

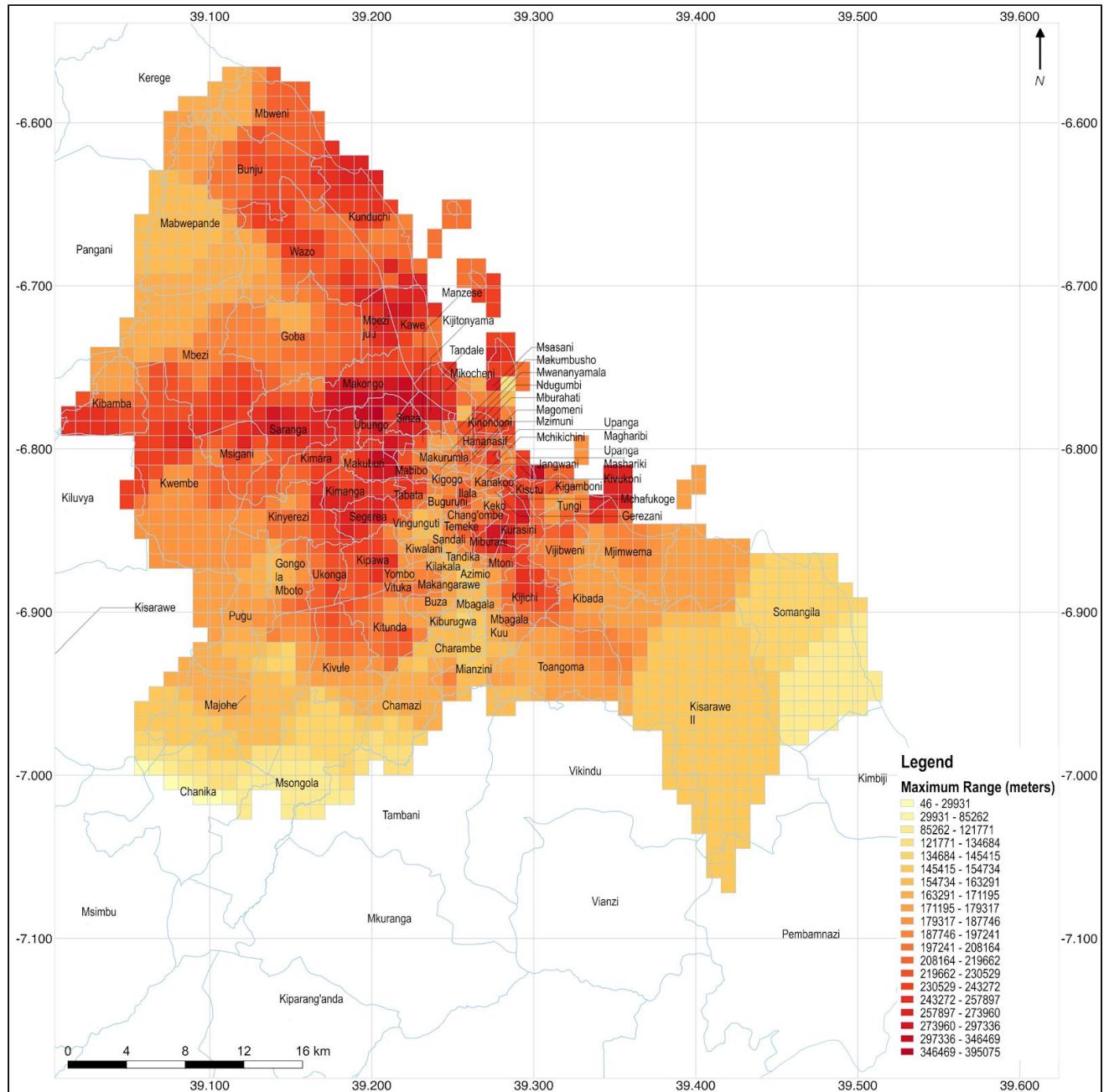


Figure 10: Shows the Mobility Across Dar es Salaam.

Mobility Map Generation

The median and maximum range of individuals whose homes were identified within each BTS's catchment area. These are then aggregated at the BTS level and converted into a grid-cell representation to produce two different forms of mobility maps.

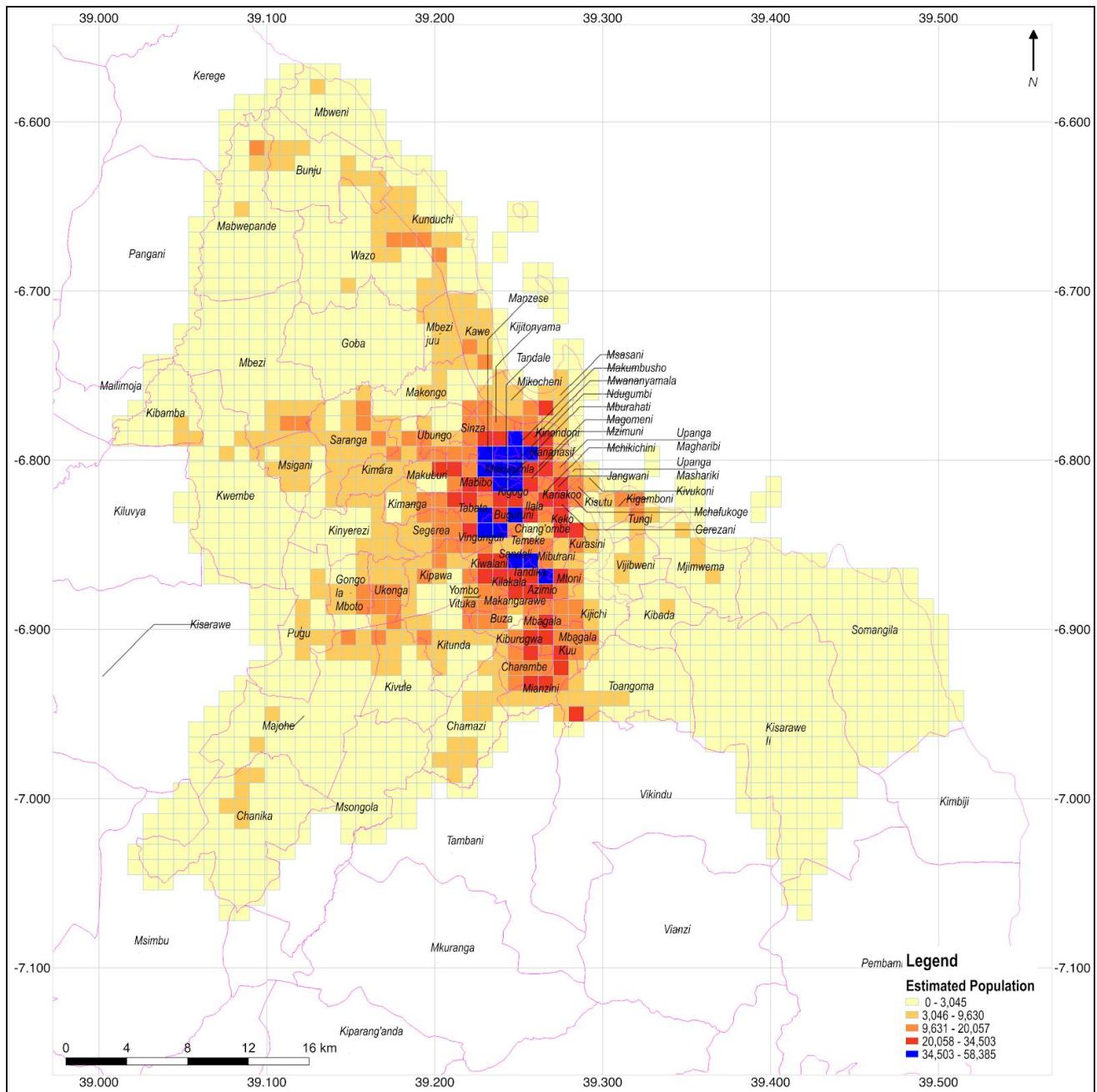


Figure 11: Population statistics derived from a CDR information

Population and Activity Map Generation

SMS and Call events give an indication of the population of an area. However, so do the number of homes attributed to each tower. Counts for each of these can be aggregated at the BTS level and converted into a grid-cell form to form activity and population maps respectively.

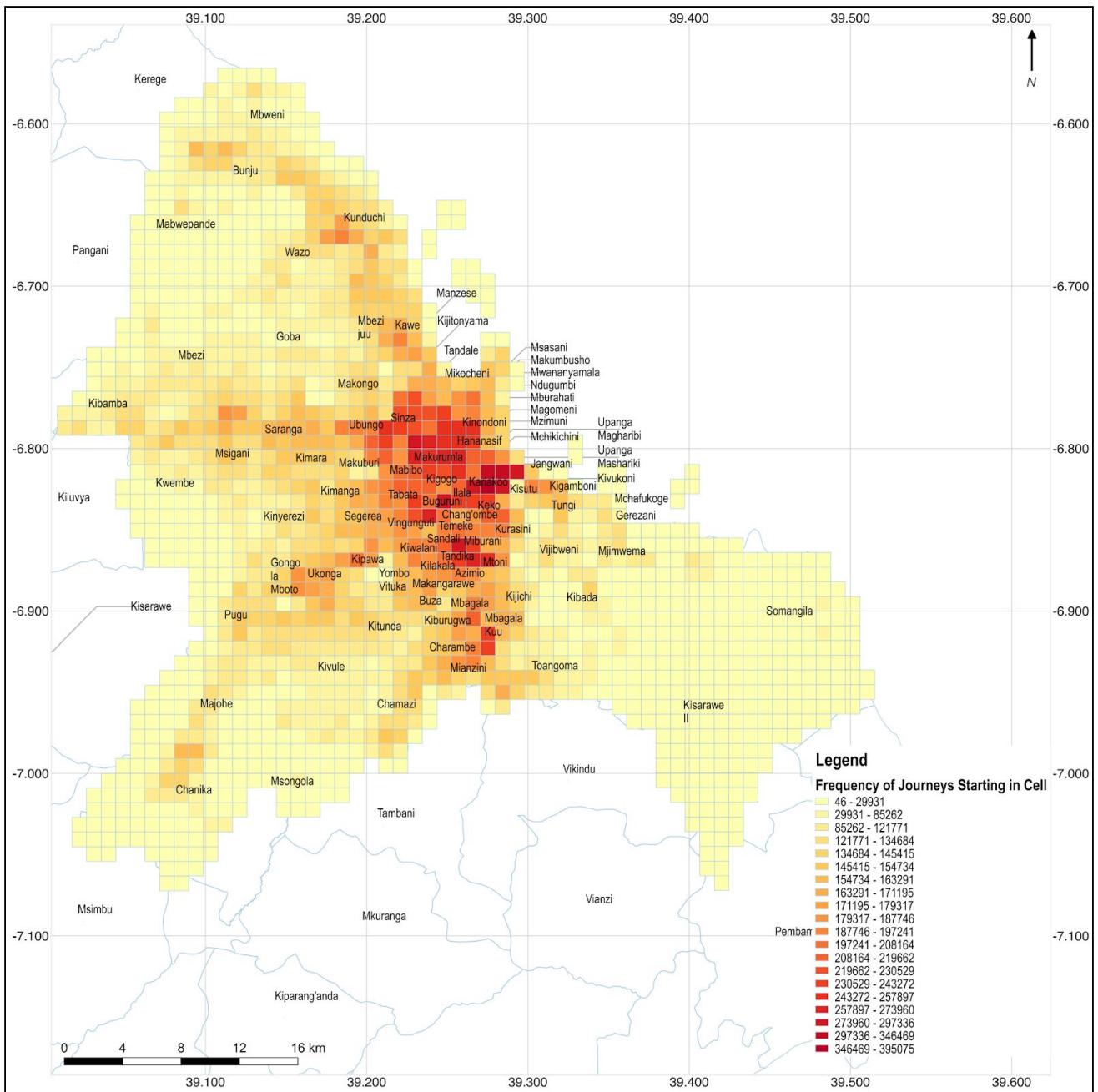


Figure 12: The Frequency and Flow of Journeys for Each Zonal Unit

Through Flow Map Generation

By examining the number of different individuals who appear in each zonal unit, or even better the entropy of each zonal unit in terms of those individuals (hence detecting transience), you can distinguish between which areas are static (in terms of the people who are active there) and which have high throughput to produce variations of through-flow maps.



*Estimating Origin-Destination flows using
opportunistically collected mobile phone location data
from one million users in Boston Metropolitan Area*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation	Calabrese, Francesco, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. "Estimating Origin-Destination Flows Using Mobile Phone Location Data." IEEE Pervasive Computing 10, no. 4 (April 2011): 36–44.
As Published	http://dx.doi.org/10.1109/mperv.2011.41
Publisher	Institute of Electrical and Electronics Engineers (IEEE)
Version	Original manuscript
Accessed	Sun Jan 20 11:59:11 EST 2019
Citable Link	http://hdl.handle.net/1721.1/101623
Terms of Use	Creative Commons Attribution-Noncommercial-Share Alike
Detailed Terms	http://creativecommons.org/licenses/by-nc-sa/4.0/



senseable city lab:::

Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area

Francesco Calabrese, *Member, IEEE*, Giusy Di Lorenzo, Liang Liu, Carlo Ratti

Abstract

In this paper, we present an algorithm for the analysis of opportunistically collected mobile phone location data to estimate a population's travel demand in terms of origins and destinations of individual trips. Aggregating the trips from millions individual mobile phone users in the Boston Metropolitan area, we show that the estimated Origin-Destination flows correlate well with the US Census estimates at both the county and census tract levels. Moreover, compared to traditional census survey data, our estimations allow capturing weekday and weekend patterns as well as seasonal variations. These features could make methods for Origin-Destination flow estimation based on opportunistically collected mobile phone location data a critical component for transportation management and emergency response.

I. INTRODUCTION

Origin Destination (OD) matrices represent one of the most important sources of information used for strategic planning and management of transportation networks. A precise calculation of OD matrices is an essential component for enabling administrative authorities to optimize the use of their transportation networks, not only for the benefit of users on their daily journeys but also with a view to the investments required to adapt these infrastructures to envisaged future needs. Traditionally, urban planning and transportation engineering rely on household questionnaires or census and road surveys conducted every 5-10 years and develop methodologies for OD matrices estimation. This approach has two main drawbacks:

- the process involved in the calculation of an OD matrix, from the initial data-gathering to the exploitation of the first results, is lengthy and may take years to only get a snapshot of the travel demand;
- the collected data has shortcomings both in terms of spatial and temporal scale.

Sensor-based OD estimation methods have also been developed in the past few years, making use of street sensors such as loop detectors and video cameras together with traffic assignment models. Analogous methods have been developed using probe vehicles, where vehicles traces are used as data sources [1], [2]. Those methods are, however, limited by the fact that models are often underdetermined because the number of parameters to be estimated is typically larger than the number of monitored network links [3].

On the other hand, the wide deployment of pervasive computing devices (e.g. mobile phone, smart cards, GPS devices and digital cameras) provide unprecedented digital footprints, telling where people are and when they are there. In former projects, different methodologies for detecting the presence and movement of crowds through their digital footprint (flickr photo, mobile phone logs, smart card record and taxi/bus GPS traces) were developed, see for instance [4]–[6]. This fine grained analysis can potentially make a big leap in terms of understanding the use of space and daily commuting flows for the purposes of urban mobility planning and management. Thus, it is no surprise that the idea of using mobile phones to monitor traffic conditions is not new. A fair number of studies relating to this matter have been published in recent

Francesco Calabrese and Giusy di Lorenzo are with the IBM Research, Ireland. Liang Liu and Carlo Ratti are with the SENSEable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 02139 Cambridge, MA (e-mail: fcalabre@ie.ibm.com).

years. Bolla et al. [7] presented a model for estimating traffic by means of an algorithm that calculates traffic parameters on the basis of mobile phone location data. A case study was developed in Rome for real time urban monitoring using aggregated mobile phone data to monitor traffic and movement of vehicles and pedestrians [8]. Cayford et al. analyzed the main parameters to be taken into account, namely precision, metering frequency and the number of localizations necessary to achieve accurate traffic descriptions [9]. Several companies worldwide, including ITIS Holdings (Britain), Delcan (Canada), CellInt (Israel), as well as AirSage and IntelliOne (USA), have begun developing commercial applications of mobile phone based traffic monitoring.

With the specific goal of measuring origin-destination flows, different mobile phone signaling datasets have been considered and simulated to evaluate the feasibility of estimating trips. Initial work was done by [10] using billing data, consisting of cell phone tower information every time a phone received or made a call. In [11] the authors used mobile phone positions every two hours to infer trips. In [12] the authors studied the use of location updates to infer mobile phone movement. In [13] the authors used cell phone tower handover information acquired every time, during a call, a phone switches a tower it is connected to. In the latest effort, [14] estimated the daily OD demand using simulated cellular probe trajectory information (extracted from location updates, handover, and transition of Timing Advance values) and tested the methodology via the VISSIM simulation.

Though these results show great potential for using cellular probe trajectory information as a means to estimating travel demand, all methods have several shortcomings before they can be put into practice. Indeed, as mentioned in [14], field tests are needed for the following reasons:

- real coverage areas of cellphone towers are very different from the simulated ones, and vary from urban to rural areas;
- validations of methods to determine origin and destination of trips should be performed using real individual mobility data;
- real mobility and calling patterns should be included in the analysis, as they crucially influence the performance of the methods;
- existing OD matrices should be used as ground truth to verify the correctness of the estimated results.

In this paper, we design a methodology that makes use of opportunistically collected mobile phone location data to estimate dynamic OD matrices. We address all above concerns using a real mobility and calling dataset from 1 million mobile phone users. We use the Boston Metropolitan area as a case study and validate our methodology using census survey data for both county and census tract levels [15]. Both the methodology developed and the data precision and amount are thus far novel and unique to our knowledge.

The paper is structured as follows. Section II describes the mobile phone dataset considered. Section III describes the OD estimation method. Section IV shows the application of the method to a real case study in the Boston Metropolitan area, and comparison of the estimated OD matrices with Census commuting flows. Section V shows some new potentials for dynamically updated OD matrices. Finally, discussion and conclusion are given.

II. MOBILE PHONE DATASET

The considered dataset consists of anonymous location measurements generated each time a device connects to the cellular network, including:

- when a call is placed or received (both at the beginning and end of a call);
- when a short message is sent or received;
- when the user connects to the internet (e.g. to browse the web, or through email programs that periodically check the mail server).

In the remainder of the paper we will call these events *network connections*. These events represent a superset of the ones contained in the Call Details Records, previously considered in [10], [16]. In this research we have been able to analyze 829 million mobile location data for 1 million device collected

by AirSage¹. Not only the id of the cell tower the mobile phone is connected to was available, but also an estimation of its position within the cell is generated through triangulation by means of AirSage's Wireless Signal Extraction technology. Each location measurement $m_i \in M$ is characterized by a position p_{m_i} expressed in latitude and longitude and a timestamp t_{m_i} .

In order to infer trips from these measurements, we first characterized the individual calling activity and verified whether that is frequent enough to allow monitoring the user's movement over time with a fine enough resolution. For each user we measured the interevent time i.e. the time interval between two consecutive network connections (similar to what was measured in [16]). The average interevent time measured for all the whole population was 260 minutes, much lower than the one found in [16] (500 minutes) as we are also considering mobile internet connections. Since the distribution of interevent times for an user spans over several temporal scales, we further characterized each calling activity distribution by its first and third quantile and the median. Fig. 1 shows the distribution of the first and third quantile and the median for all users available into the dataset. The arithmetic average of the medians is 84 minutes (the geometric average of the medians is 10.3 minutes) with results small enough to detect changes of location where the user stops as low as 1.5 hours.

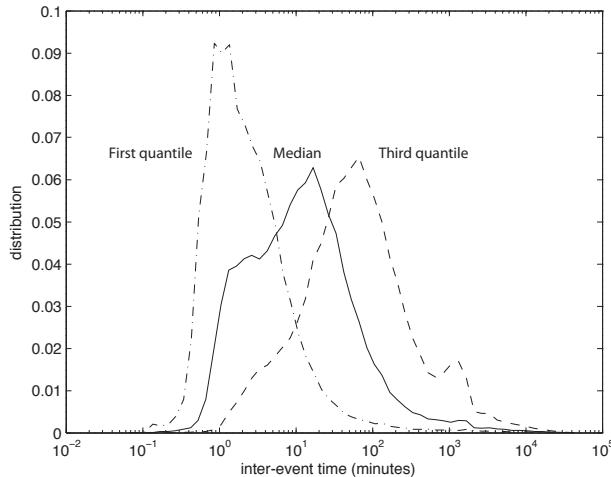


Fig. 1. Characterization of individual calling activity for the whole population. Median (solid line), first quantile (dash-dotted line) and third quantile (dashed line) of individual interevent time.

Mobile phone-derived location data has lower resolution than GPS data: internal and independent testing suggests an average uncertainty radius of 320 meters, and a median of 220 meters. Moreover, at some peak usage periods additional locational error may be introduced when users are automatically transferred by the network from the closest cellular tower to one which is further away but less heavily-loaded.

III. ORIGIN-DESTINATION ESTIMATION METHOD

The procedure for estimating dynamic OD matrices is composed of two steps: trips determination and origin-destination estimation.

To alleviate the effects of localization errors and event-driven location measurements on the determination of individual trips, we propose the following method: we apply a low-pass filter with a resampling rate of 10 minutes to the raw data, this follows an approach tested with data from Rome, Italy [8]. In addition, since lesser localization errors might still generate fictitious trips, we adapt a pre-processing step employed in the analysis of gps traces, which uses clustering to identify minor oscillations around a common location. In more detail, the approach employed to handle locational errors and identify meaningful locations in a user's travel history can be understood as follows:

¹<http://www.airsage.com/>

- We begin with a measurement series $M_s = \{m_q, m_{q+1}, \dots, m_z\} \in M^{z-q-1}$, $q > z$, derived from a series of network connections over a certain time interval $\Delta T = t_{m_z} - t_{m_q} > 0$.
- We define an area with radius ΔS – in this case, $1km$ to take into account the localization errors estimated by AirSage – such that

$$\max distance(p_{m_i}, p_{m_j}) < \Delta S \quad \forall \quad q \leq i, j \leq z$$

- All the consecutive points $p_j \in M_s$ for which this condition holds can be fused together such that the centroid becomes a ‘virtual location’ ($p_s = (z - q)^{-1} \sum_{i=q}^{i=z} p_{m_i}$, the centroid of the points) that is the origin or destination of a trip.
- Once the virtual locations are detected, we can evaluate the stops (virtual locations) and trips as paths between users’ positions at consecutive virtual locations. Each trip $trip(u, o, d, t)$ is characterized by user id u , origin location o , destination location d and starting time t .

Section IV presents some statistics on the trips estimated using the proposed method comparing it with reference statistics, showing how the method performs well in estimating trips in our case study.

Once trips are extracted, the procedure to derive Origin-Destination flows is the following:

- 1) The geographical area under analysis is divided into regions: $region_i$, $i = 1, \dots, n$.
- 2) Origin and destination regions, together with starting time are extracted for each trip of each user $trip(u, o, d, t)$.
- 3) Trips with the same origin and destination regions are grouped together at different temporal windows tw e.g. weekly, daily, hourly:

$$m(i, j, tw) = \sum_{o \in region_i, d \in region_j, t \in tw} trip(u, o, d, t).$$

The result is a three-dimensional matrix $M \in \Re^3$ whose element $m(i, j, tw)$ represents the number of trips from origin region i to destination region j starting within the time window tw . The potentials of using adaptive time windows will be shown in Section V-A.

IV. CASE STUDY IN THE BOSTON REGION AND COMPARISON WITH CENSUS COMMUTING FLOWS

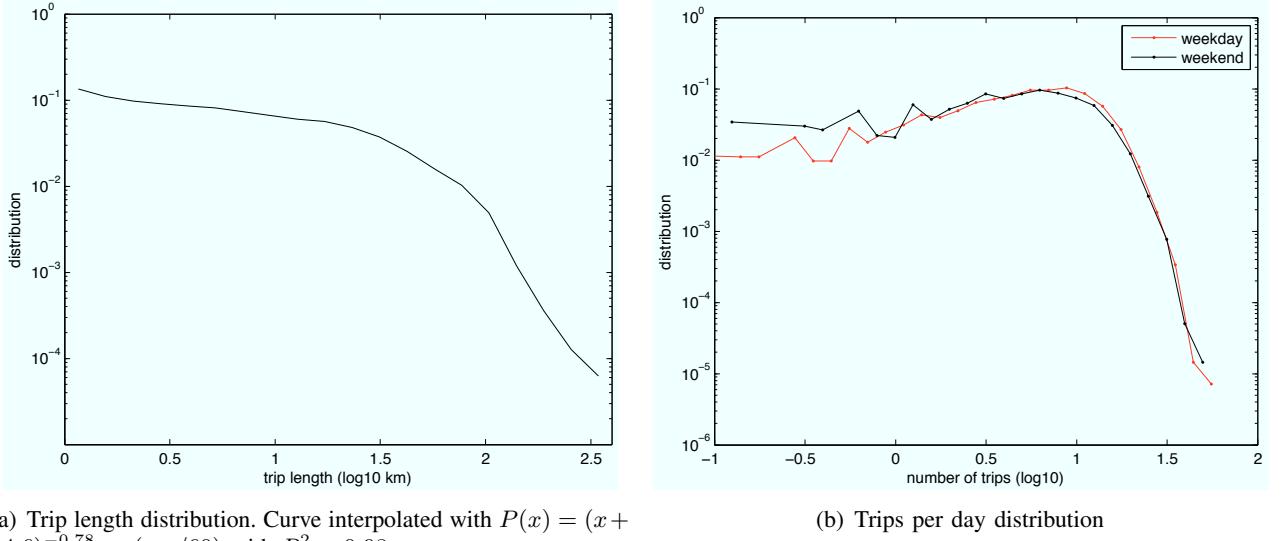
In this section we study the effectiveness of the methodology in a real case study in the Boston region. Based on the area covered by the mobile phone locations dataset, we analyzed the movements among areas in 8 counties in east Massachusetts (Middlesex, Suffolk, Essex, Worcester, Norfolk, Bristol, Plymouth, Barnstable) with an approximate population of 5.5 million people. To simplify the analysis, we extracted traces for 25% randomly selected users among the available ones.

A. Characterization of trips

As a first analysis we studied the trip length distribution (see Figure 2(a)), showing that trips range from 1 to 300 Km. We determined the trip length x by calculating the Euclidean distances among trip’s origin and destination. The distribution is well approximated by $P(x) = (x + 14.6)^{-0.78} \exp(-x/60)$ with $R^2 = 0.98$, which confirms what was found in [16]. The slightly different coefficients found in this case could be attributed to the different built environment in Europe and US, see [17]. To check the plausibility of our segmentation of the trajectory in trips, we compute same statistics computed on the number of individual trips per day. The distribution over the whole population is shown in Figure 2(b), separating weekday and weekend trips. We obtain an average of 5 trips per day during the weekday, and 4.5 during the weekend. This number is reasonable when compared to the US National Household Travel Survey² which evaluated this number to be between 4.18 during weekdays and 3.86 during weekends³.

²<http://nhts.ornl.gov/>

³The sources of differences can be associated to several reasons, including the several years of difference between when the two datasets have been collected, and the fact that NHTS is based on a sample over all US population, so not focused on the behavior of people in the Boston Metropolitan area.



(a) Trip length distribution. Curve interpolated with $P(x) = (x + 14.6)^{-0.78} \exp(-x/60)$ with $R^2 = 0.98$.

(b) Trips per day distribution

Fig. 2. Statistics on the detected trips.

To evaluate whether we have sampling biases in our data, we computed the home locations distribution estimated from the mobile phone data, and compared it with data from the US 2000 Census. To detect the home location, we first group together geographic regions that are close in space, creating a grid in space where the side of every cell is 500 meters. For each cell we evaluate the number of nights the user connects to the network in the night time interval while in that cell, and select as a home location the cell with the greatest value⁴.

To validate the home location distribution, we then compared it with population data from the US 2000 Census, at the level of the census tract [18]. In the selected 8 counties, we have 1171 distinct census tracts, with populations ranging from 70 to 12 thousand people (on average 4705), and an area ranging from 0.08 to 203 km^2 (on average 10.8 km^2). The census tract population estimated using mobile phone users' home locations scales linearly with the Census population, as shown in Figure 3(b), corresponding to an average 4.3% of the population being monitored.

B. Characterization of OD flows

To validate the accuracy of the OD matrices produced using the mobile phone traces, we used the most recent Tract-Tract Worker Flows dataset from Census Transportation Planning Package [15]. CTPP is a special tabulation of responses from households completing the Census long form. It is the only Census product that summarizes data by place of work and tabulates the flow of workers between home and work.

The *Tract-Tract Worker Flows* data shows the number of workers in each tract of work by tract of residence. Workers are defined as people age 16 years old and over who were employed and at work, full time or part time, during the Census reference week (generally the last week of March). The data contains the number of workers in the flow who were allocated to tract, place, and county of work.

Given the two levels of granularity (tract and county) available in the CTPP dataset, we computed our OD estimates at two levels of aggregation. Since commuting flow generally accounts for two trips (home to work and work to home), we considered undirected flows between two locations to compare our OD

⁴The considered night time interval is 6pm-8am and has been defined considering the statistics available in the American Time Use Survey, <http://www.bls.gov/tus/charts/work.htm>

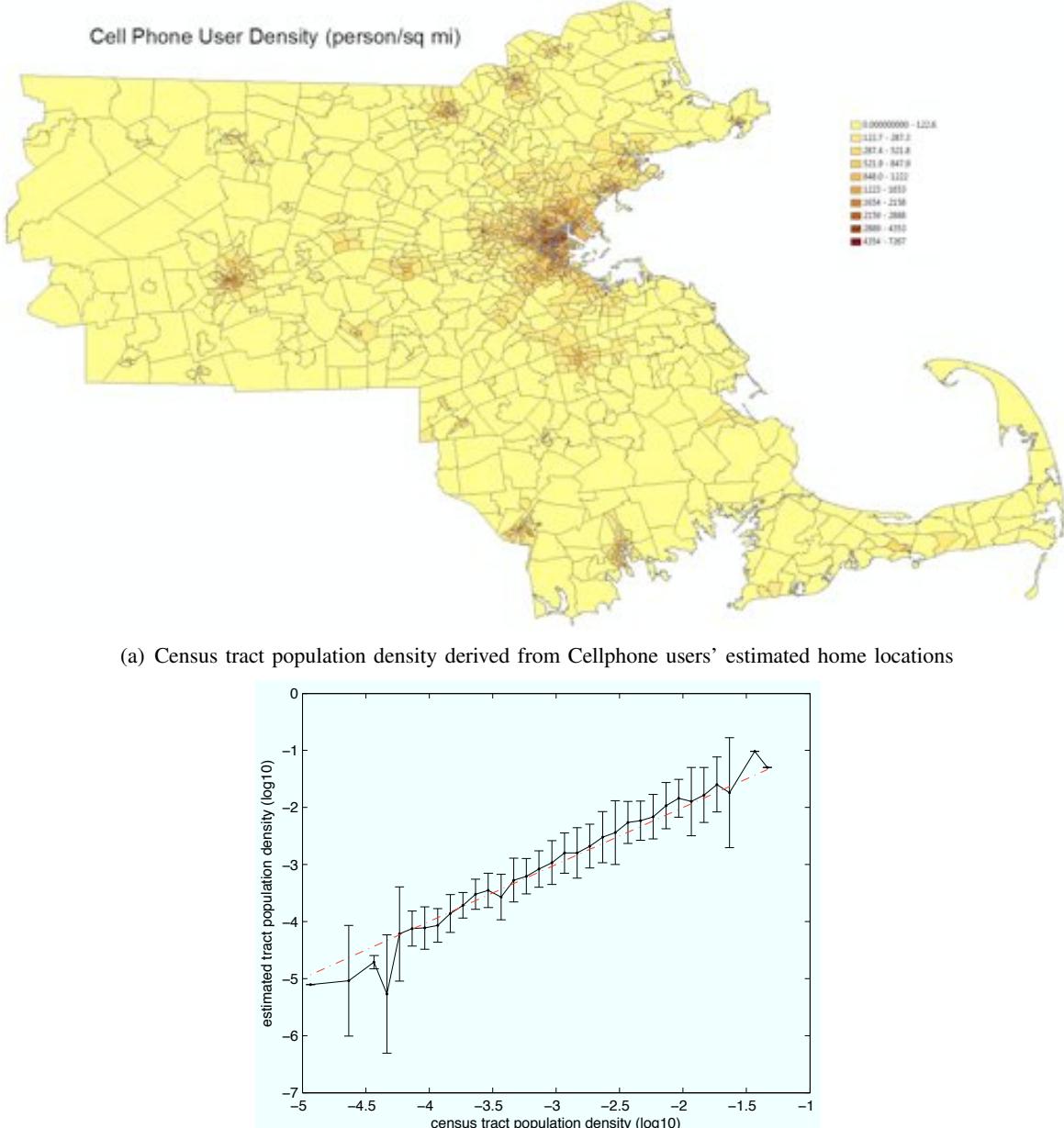


Fig. 3. Census tract population density derived from US 2000 Census compared to Cellphone users' estimated home locations density

estimations. For each granularity, we computed the average daily number of trips:

$$m_{All}(i, j) = \frac{K_{All}}{\#\text{days}} \sum_{tw=\text{day}} (m(i, j, tw) + m(j, i, tw)),$$

$$i = 1, \dots, n, \quad j = 1, \dots, i - 1,$$

where K_{All} is a scaling factor we use to compare them with the Census estimations.

Moreover, since according to the definition, the census dataset includes only commuting trips, we evaluated the average daily number of trips made only on weekdays mornings (6-10am) from the estimated

home to estimated work location⁵:

$$m_{WM}(i, j) = \frac{K_{WM}}{\#\text{weekdays}} \sum_{tw=wm} (m(i, j, tw) + m(j, i, tw)), \\ i = 1, \dots, n, \quad j = 1, \dots, i - 1,$$

where K_{WM} is a scaling factor. Finally, we also considered the well known and widely used gravity model [19] to compare our predictions with:

$$m_{Gravity}(i, j) = K_G \frac{P_i \cdot P_j}{d_{i,j}^2}, \quad i = 1, \dots, n, \quad j = 1, \dots, i - 1,$$

where K_G is a scaling factor, and $d_{i,j}$ is the Euclidean distance (in kilometers) between the centroids of the regions. The results at the county level are shown in Figures 4(a). The plots correspond to models which minimize the least square errors, using: $K_{All} = 16.9$ for the prediction made with the average number of trips in a day m_{All} ; $K_{WM} = 71.4$ for the prediction made with the average number of trips on weekday mornings m_{WM} , and $K_G = 58.4$ for the gravity model $m_{Gravity}$.

Correlations show very encouraging results, with $R^2 = 0.59$ for the gravity model, $R^2 = 0.73$ for the prediction made with all trips, and the best result $R^2 = 0.76$ for predictions made considering only weekday morning trips. The resulting high correlation shows that the estimated OD matrices are able to resemble very well OD matrices generated using completely different information.

Using the best model m_{WM} , we compared our results with the tract level census data. At this level, noise is more evident (see Figure 4(b)), but still we can see on average a very good linear relationship between census estimation and our estimation. $R^2 = 0.36$ in this case, which is however very high compared to the $R^2 = 0.10$ of the gravity model⁶. The relatively low value of R^2 compared to the county level analysis is partially due to the fact that the relationship seems less linear for cases when the census estimates less than 10 trips from tract to tract. This might be explained by the fact that census flows are estimated from a subsample, that might result in very small numbers for particular pairs of census tracts. Moreover, census estimates were not available for the same year as the mobile phone data, and origins and destinations of trips might have slightly changed (at this high level of spatial detail) between the two monitored periods.

We note that the scaling factor K_{WM} used for the last model m_{WM} corresponds to a share of monitored trips which is about 1.4% compared to the census estimations. This factor can be explained by the percentage of mobile phones selected (about 4.3%) and by the calling activity which is not very high in the morning. Other elements such as the fact that we are monitoring not only commuting flows might explain the remaining difference. Estimating K_{WM} allows to extrapolate the ODs computed using the mobile phone data to the whole population.

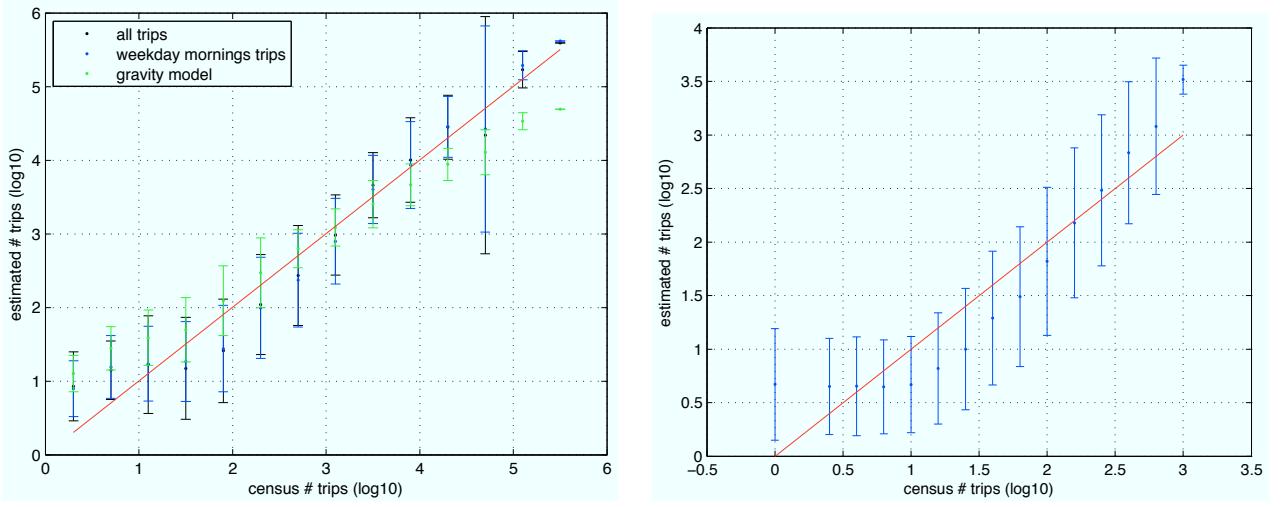
V. NEW POTENTIALS

Origin-destination flows data estimated through census surveys have the following limitations (see [15]):

- The decennial census monitors "usual" days to avoid local or regional anomalies such as transit strike or severe weather, on a single sampling day. However, this tends to hide the less common uses, such as telecommuting once every two weeks or carpooling once a week due to the ever-changing life and work patterns.
- According to the definition, the census dataset does not include non-work trips, and modelers have to develop relationships between work and non-work trips.
- The census data is based on a fixed point "snapshot" approach, and so transportation planners can only interpret data over geographic space, rather than over time.

⁵The work location has been estimated as the most frequent stop area on weekday morning 8-10am.

⁶We have also evaluated more sophisticated gravity-like models by optimizing the d exponent and substituting the populations with the total estimated number of trips outgoing or incoming an area, but have still obtained $R^2 < 0.3$.



(a) County level. All trips, weekday morning trips and gravity model.

(b) Tract level. Weekday morning trips

Fig. 4. Comparison between mobile phone and census OD estimates. Error bars are showing one standard deviation from the average.

Compared with traditional census data, our methodology to detect OD matrices from mobile phone traces has several advantages:

- It can capture the weekday and weekend patterns as well as seasonal variations.
- It can capture work trips and non-work trips, which is essential for trip chaining and activity based modeling.

For these reasons, they could then be used to complement traditionally generated OD matrices providing a very fine grain spatial-temporal patterns of mobility.

In the following subsections, examples of these potentials are shown.

A. Temporal analysis

While the census gives only a static information about origin-destination flows, the OD matrices derived from mobile phone data allows us to appreciate the differences in travel demand over time. Figure 5(a) shows the total *daily* travel demand for 3 different weeks in October 2009. A weekly pattern clearly appears in the travel demand, with the minimum over weekends (especially sundays) and a maximum over fridays. Moreover, Figure 5(a) shows a particular change in travel demand in the second monday (day number 9 in figure), corresponding to Columbus Day. For a better look at this pattern, we plot the *hourly* travel demand for Columbus Day compared to the other mondays (see Figure 5(b)). We clearly see a higher travel demand in the first 2 hours of the day, followed by lower demand from 4 to 9, and from 12 to 20, due to the holiday.

B. Spatiotemporal analysis

Our methodology can capture very fine grain OD matrices in both spatial and temporal scale, essential data for understanding transport demand and transport modeling especially during special events. For example, Figure 6 compares the incoming flows toward the Boston Baseball stadium Fenway Park. We compare two different days: Sunday October 11th where the local baseball team the Red Sox played against the Angels in a postseason game, and an average sunday without events. As it can be seen from the figures, we are able to capture the increasing incoming flow due to the special event, both in terms of new origins of trips, and in volumes of flow. Further studies with the same dataset have also shown regular spatial patterns of attendee origins based on the type of event, information that would be very valuable for event management [5].

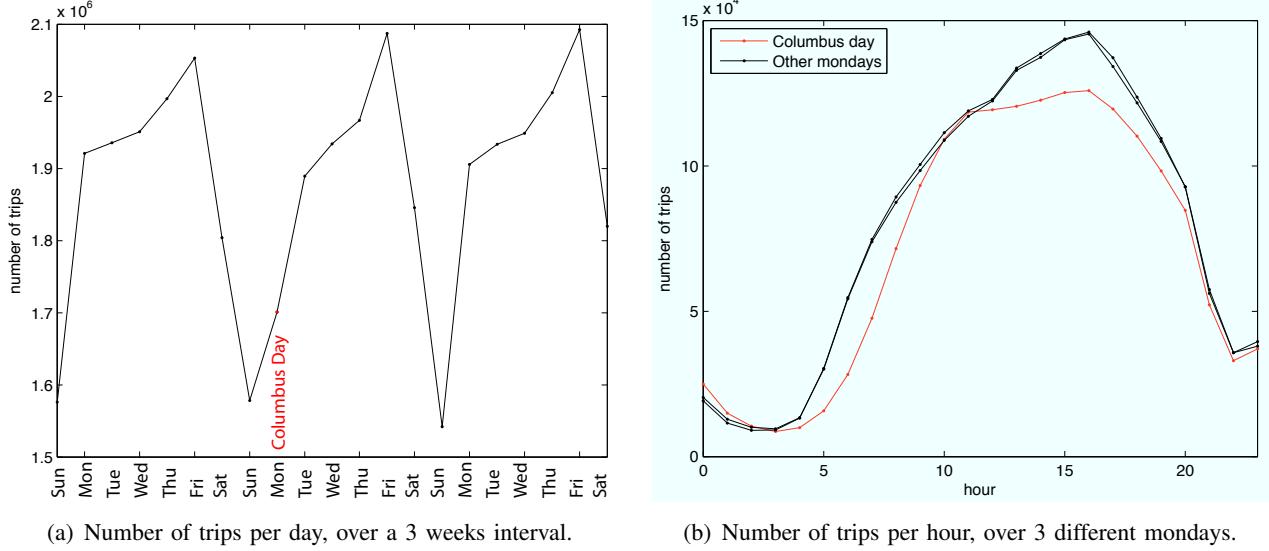


Fig. 5. Temporal variation in the number of trips.

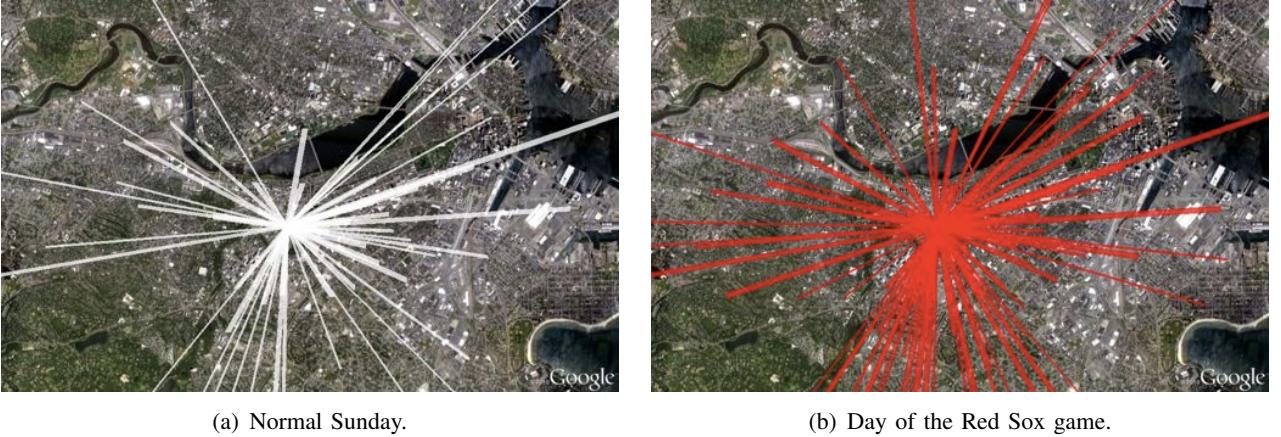


Fig. 6. Incoming trips in the Fenway Park area. Flow volume is represented by the thickness of the line.

VI. DISCUSSION AND CONCLUSION

As shown in this study, pervasive datasets such as mobile phone traces provide rich information to support transportation planning and operation. Meanwhile, some related limitations should also be addressed when applying these datasets in mobility analysis. A crucial parameter to take into account is the localization error, which limits the minimum size of the regions that can be considered. Other elements that can affect the statistical results include: 1) the market share of the mobile phone operator from which the dataset is obtained, 2) the potential non-randomness of the mobile phone users (e.g. teenagers), 3) calling plans which can limit the number of samples acquired at each hour or day, 4) number of devices that each person carries. Moreover, due to the fact that the considered dataset is event-driven (location measurements available only when the device makes network connections) the connection patterns of users can affect the possibility to capture more or less trips. This last limitation could be solved by continuous location readings from GPS devices, which would however require the users consent. An hybrid approach could be envisioned, integrating both event-driven and continuous location measurements, as the current method can be easily generalized to different datasets with different spatio-temporal resolutions. Nonetheless, the analysis performed on the inter-event time, the spatial distribution of mobile phone users, and comparisons with census estimations confirm that the mobile phone data represent a reasonable proxy for human mobility.

Apart from reproducing data derived by means of expensive census surveys, our methodology to detect OD matrices from mobile phone traces has several advantages: 1) It can capture the weekday and weekend patterns as well as seasonal variations. 2) It can capture work and non work trips. 3) It can produce real time, continuous OD matrices which can capture the very fine grain spatialtemporal patterns of urban mobility.

Future work will involve reproducing the analysis for other cities, in order to understand which parameters influence the scaling factors to be used to extrapolate the ODs computed using the mobile phone data to the whole population. The research output will give transport planners an automatic and systematic way to understand the dynamics of daily mobility in a real complex metropolitan area.

REFERENCES

- [1] X. Zhou and H. S. Mahmassani, "Dynamic origin-destination demand estimation using automatic vehicle identification data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 105–114, 2006.
- [2] S. Baek, Y. Lim, S. Rhee, and K. Choi, "Method for estimating population od matrix based on probe vehicles," *KSCE Journal of Civil Engineering*, vol. 14, no. 2, pp. 231–235, 2010.
- [3] M. L. Hazelton, "Some comments on origin-destination matrix estimation," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 10, pp. 811 – 822, 2003.
- [4] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive Computing*, 2008.
- [5] F. Calabrese, F. Pereira, G. DiLorenzo, and L. Liu, "The geography of taste: analyzing cell-phone mobility and social events," in *International Conference on Pervasive Computing*, 2010.
- [6] D. Quercia, G. DiLorenzo, F. Calabrese, and C. Ratti, "Mobile phones and outdoor advertising: Measurable advertising," *IEEE Pervasive Computing*, vol. 10, no. 2, pp. 28–36, 2011.
- [7] R. Bolla and F. Davoli, "Road traffic estimation from location tracking data in the mobile cellular network," in *IEEE Wireless Communications and Networking Conference*, vol. 3, 2000, pp. 1107 –1112.
- [8] F. Calabrese, C. Ratti, M. Colonna, P. Lovisolo, and D. Parata, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.
- [9] R. Cayford and T. Johnson, "Operational parameters affecting use of anonymous cell phone tracking for generating traffic information," in *Institute of transportation studies for the 82th TRB Annual Meeting*, 2003.
- [10] J. White and I. Wells, "Extracting origin destinationn information from mobile phone data," in *Road Transportation and Control*, 2002.
- [11] C. Pan, J. Lu, S. Di, and B. Ran, "Cellular-based data-extracting method for trip distribution," *Journal of Transportation Research Board*, pp. 33–39, 2006.
- [12] N. Caceres, J. Wideberg, and F. Benitez, "Deriving origin destination data from a mobile phone network," *Intelligent Transport Systems, IET*, vol. 1, no. 1, pp. 15 –26, 2007.
- [13] K. Sohn and D. Kim, "Dynamic origin-destination flow estimation using cellular communication system," *Vehicular Technology, IEEE Transactions on*, vol. 57, no. 5, pp. 2703 –2713, sept. 2008.
- [14] Y. Zhang, X. Qin, S. Dong, and B. Ran, "Daily o-d matrix estimation using cellular probe data," in *Transportation Research Board Annual Meeting*, 2010.
- [15] CTPP, "Us department of transportation, census transportation planning products," <http://www.fhwa.dot.gov/ctpp/>, 2010.
- [16] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
- [17] L. Liu, F. Calabrese, A. Biderman, and C. Ratti, "The law of inhabitant travel distance distribution," in *European Conference on Complex Systems*, Warwick, UK, September 2009.
- [18] MassGIS, "Census 2000 tracts datalayer description," <http://www.mass.gov/>, 2010.
- [19] J. Anderson, "A theoretical foundation for the gravity equation," *The American Economic Review*, vol. 69, pp. 106–116, 1979.



Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement

Françoise Bahoken, Ana-Maria Olteanu-Raimond

► To cite this version:

Françoise Bahoken, Ana-Maria Olteanu-Raimond. Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement. ICC'13 - 26th International Cartographic Conference, Aug 2013, DRESDEN, Germany. ICC'13 - 26th International Cartographic Conference, 15p, 2013. <hal-01011987v2>

HAL Id: hal-01011987

<https://hal.archives-ouvertes.fr/hal-01011987v2>

Submitted on 26 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths

The effect of spatiotemporal filtering on flow measurement

Françoise Bahoken^{1,2}, Ana-Maria Olteanu Raimond^{3,4}

¹ Paris-Est University, French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR),

² UMR 8504 Géographie-Cités, Paris, France

³ Orange Labs, Sociology and Economics of Networks and Services department, Paris, France

⁴ French Mapping Agency, Cogit Laboratory, Saint-Mandé, France

Abstract. In the past few years the mobile phone data are considered as a useful complementary source of information for human mobility research. In this paper, we focus on the computation of the Origin Destination matrix using mobile phone data. First, a new approach of OD matrix design which takes into account the spatiotemporal heterogeneities of mobile phone data is proposed. Second, some analysis allowing measuring the effect of spatio-temporal filtering on the OD matrix results are carried out.

Keywords: Human migration, Flow mapping, Origin-Destination matrix, Mobile Phone Data, Spatiotemporal Filtering

1. Introduction

Origin-Destination (OD) matrices are traditionally used in transportation, urban planning engineering and human migration studies. Matrices result from the sum of individual movements which are produced in a time interval on a given space. Their design has been estimated using a wide range of different approaches which can be grouped into three categories: directed, undirected and alternative methods. The statistics obtained using directed or undirected methods focused on the retrospective information gathering on travel, or individual questionnaires, are very sensitive to errors due to

the adopted methodology. Thus, alternative methods have been developed, focused on the markers of migrations, such as GPS or mobile phones data.

In this paper, we are interesting in designing OD flows matrix using mobile phone data as an alternative of traditional data. We assume that the individual daily movements can be captured from spatiotemporal traces of mobile phones. These traces are not the exact path of the user, but an estimation of it, derived from the mobiles' defined positions in space and time. One of the advantages of using mobile phone traces lies in the space paradigm shift : OD flows can then be generated and studied on a (pseudo) continuous basis, that is to say in fine spatial and temporal resolutions instead of being available on a discrete spatial (administrative units) and temporal ways (defined period). An OD matrix that reasonably reflects temporal distribution is often indispensable for applications ranging from short-term planning to within-day traffic control/management. Moreover, mobile phone data can be automatically collected at relatively low cost and presenting an important sample of users.

The paper is structured as follows. In the next section the state of art is presented. In Section 3 mobile phone data are briefly described. Section 4 first introduced the proposed approach to design OD flows matrix using mobile phone data and second the effects of spatiotemporal filtering are discussed. Finally, section 5 concludes and suggests some directions to future work.

2. State of art

The main issue of classical OD matrix methods (unless for registries) is that they are partial and require a complex statistical treatment in order to estimate the OD flows matrix. In fact, survey approaches implies that the OD matrix represent a snapshot of the commuting patterns over time at a selected spatial scale. Moreover, surveyed data are expensive and sometimes they are likely to be out-of-date. It is for all these reasons that alternative approaches have been developed. They focus on the use of markers of migrations in order to reconstruct individual movements and the resulting aggregated OD matrix. One way is to use the files subscriptions to a service provider like electricity, water or telecommunications furniture. Another possibility is to use mobile phone data.

In recent years, mobile phones have become one of the main sensors of human mobility at a large scale. Generally, there are two main approaches to model human mobility from mobile traces: trip-based, where aggregated data are used (Gonzalez et al. 2008, Sevtsuk & Ratti 2010) and activity-based when individual data are considered (Reades et al. 2007, Ahas et al. 2008 and Gonzalez et al. 2008, Olteanu Raimond et al. 2012).

A variety of studies having as specific goal to infer OD flows matrix using sensors data such as mobile phone and GPS were carried out in recent years. Friaz-Martinez et al. (2012) proposed a method that generates commuting OD flows matrix based on temporal variation of association rules using aggregated mobile phone data. Another approach consists on identifying moving and staying points (Byeong-Seok et al. 2005). If the mobile phone is staying on the same base station over a pre-defined threshold time, then it is consider that trip of the user is finished and an OD flow is generated. This method is very sensitive to the recording rate, and it should be applied only if the recording rate is constant. Calabrese et al. (2011) first computed stops and trips in individual trajectories and then flows for each trips of each user are extracted. Flows are aggregated by origin-destination (i.e. predefined regions) and by temporal window.

Flows can also be defined such as a path starting (the first point of the path) in the origin region and ending (the last point of the path) in the destination region (Caceres et al. 2007, Giannotti et al. 2011). A time window is thus necessary to be defined if fine temporal analysis must be carried out.

In this context, we propose a static OD flows matrix computation which means that linked flow data exist only for one time period. Such an approach is inspired from (Caceres et al. 2007, Giannotti et al. 2011). Our method, described in section 4.1, consists on taking into account the spatio-temporal heterogeneities characterizing mobile phone data.

3. Mobile phone data

Each mobile phone operator collects and store for a given period customers' mobile phones activities for billing or for technical measurements purposes. This type of collection is called "passive collection", since recordings are made automatically. They are three mainly types of mobile phone data collected using the passive collection: Call Detail Records (CDR) data, Probes data and Wi-Fi data. In this paper, only CDR data are described, since these data are used to validate our approach. For more information about mobile phone data, see (Smoreda et al. 2012).

CDR data are cell phone billing records, where location information (cell id) is automatically generated at the moment of communication: call's start (in/out-coming) and SMS (in/out). The records contains the following attributes: i) the anonymised SIM card identifier; ii) the antenna identifier; iii) the base station location; iv) the record type (call in/out, SMS in/out) and v) the time of communication activity (timestamp).

In this study the location of mobile phone users is limited to the base station location. The main advantages of CDR data are: the big mass of located data for a long period of time and for a very large spatial extends (e.g. country level). Moreover, records represents all operator' clients, not only a sample of users. The disadvantage is due to the heterogeneity of records (only when a communication occurs).

4. OD Matrix Design and the Effects of spatiotemporal filtering

In this section, we first describe the OD matrix approach. Second, the effect of spatiotemporal filters is analyzed by consider a real case study.

4.1. OD matrix design

Let's consider a spatiotemporal trajectory T_k composed by a set of n consecutive points, noted $T_k = \{p_1, p_2, \dots, p_i, \dots, p_{n-1}, p_n\}$, where :

$p_i = (x_i, y_i, t_i)$ is a record point having a spatial location (x, y) at moment t ,

$t_1 < t_2 < \dots < t_{n-1} < t_n$,

$i = 1..n$ represents the number of points composing the T_k .

Definition: For each trajectory T_k , there is a flow between the areas (i,j) , noted $F(i,j)$ if the two following conditions are satisfied :

(1) $p_1 \in i \text{ OR } (p_2 \in i \text{ AND } p_1 \in \text{neighbors list of } p_2)$

AND

(2) $p_n \in j \text{ OR } (p_{n-1} \in j \text{ AND } p_n \in \text{neighbors list of } p_{n-1})$

Figure 1 shows an example of our flow path definition. Let's consider two distinct trajectories (T_1 and T_2) belonging to two distinct users. Considering the below definition, the trajectory T_1 can be consider such as an OD flow between the origin (i) and the destination (j) : at least two points of the path are related to different spatial units of the study area and the flow can be design from the crossing borders phenomena. On the opposite, the trajectory T_2 cannot be consider as an OD flow between (i,j) since the first condition (1) is not met : only one point of the trajectory is included in a spatial area and crossing border phenomena do not exist.

The distinction between the two trajectories is due to spatiotemporal filters which affected the number of migrants between (i,j) and, consequently, the number of migrations or paths. It depends, on the one hand, on the tem-

poral filter -which defines position along time- and on the other, on the spatial filter -which defines the size of the areas.

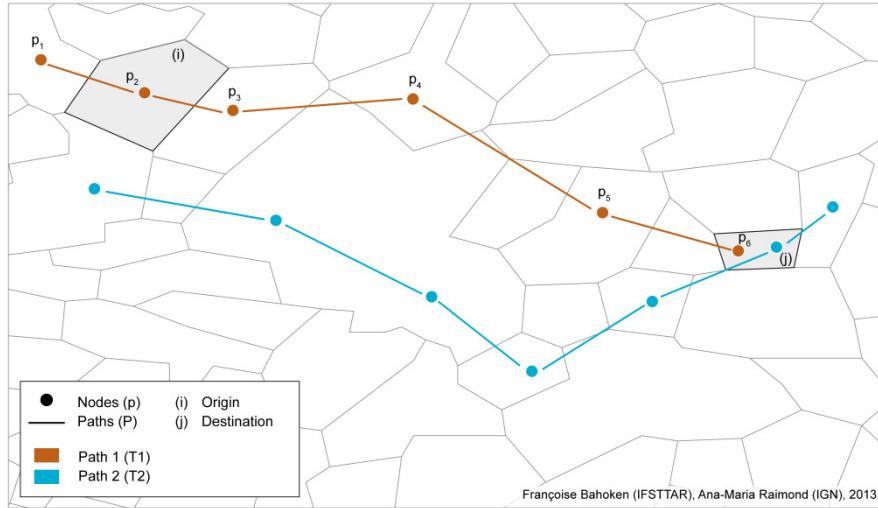


Figure 1. Spatial definition of flow path.

4.2. Data sets presentation

Three datasets are used: the CDR data, the French municipalities zoning system and the French urban nodes in *Picardie Region*.

The first dataset is the CDR data containing all mobile phone calls and SMS collected from BTS towers located in the study area during six weeks from 1st of September to 15 October 2007. BTS locations are used to build Voronoi polygons representing the antennas coverage (see Figure 2 and 4).

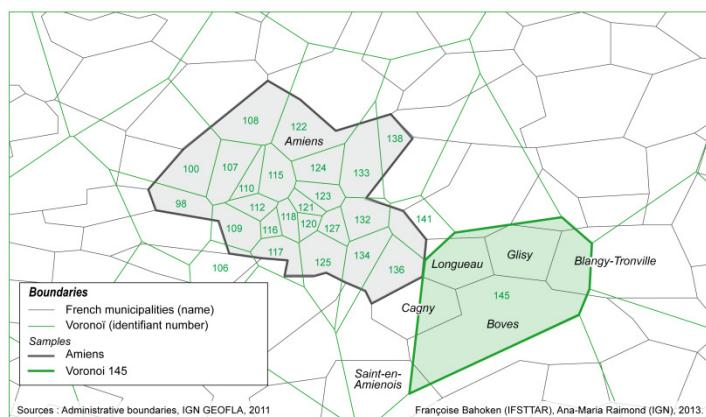


Figure 2. Voronoi polygons in the study area.

As we can see in Figure 2, the Voronoi polygons are very heterogeneous from a spatial point a view. For example, the municipality of Amien is characterized by small Voronoi areas, since the Voronoi “145” covers partially or completely five municipalities (e.g. *Boves*, *Longueau*, *Cagny* and *Blangy Tronville*). The dataset accounts 10,145,916 users.

Figure 3 shows the temporal distribution of mobile phone events aggregated by 60 minutes for one week. Some activity peaks can be observed at 1pm, and 7pm for weekdays and 1pm, 8pm and 9pm for weekend. These peaks are related to lunch break, commuting coherent with the French daily activity pace. We notice that the distribution of events is less important during the weekend except during the night between 1 am and 5 am.

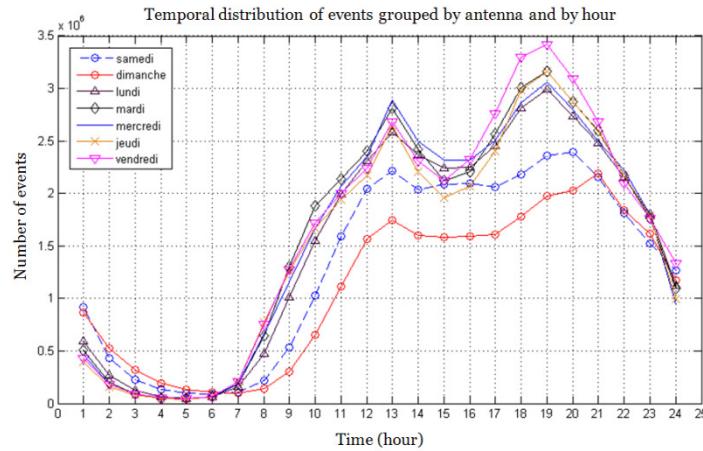


Figure 3. Temporal distribution of events during a weekday.

The second dataset contains the French municipalities zoning system areal data (polygons) representing urban areas (see Figure 4). An Urban Area (UA) represents a set of municipalities (the French *communes*) requiring several conditions: the UA must be in one piece and without enclave, composed by an urban center and must have i) more than 10,000 employments and ii) rural municipalities surrounding that have at least 40% of the resident population which is employed in the urban center (INSEE 2003).

Finally, the third dataset represents urban nodes system (polygons). A node is a group of municipalities in one piece and without enclave, consisting i) of an urban center from 5,000 to 10,000 employments and ii) rural municipalities with at 40% of the resident population which is employed in the urban center (INSEE 2003). It should be notice that the selected nodes correspond to the Central municipality of large urban cores defined in the Urban area zoning system.

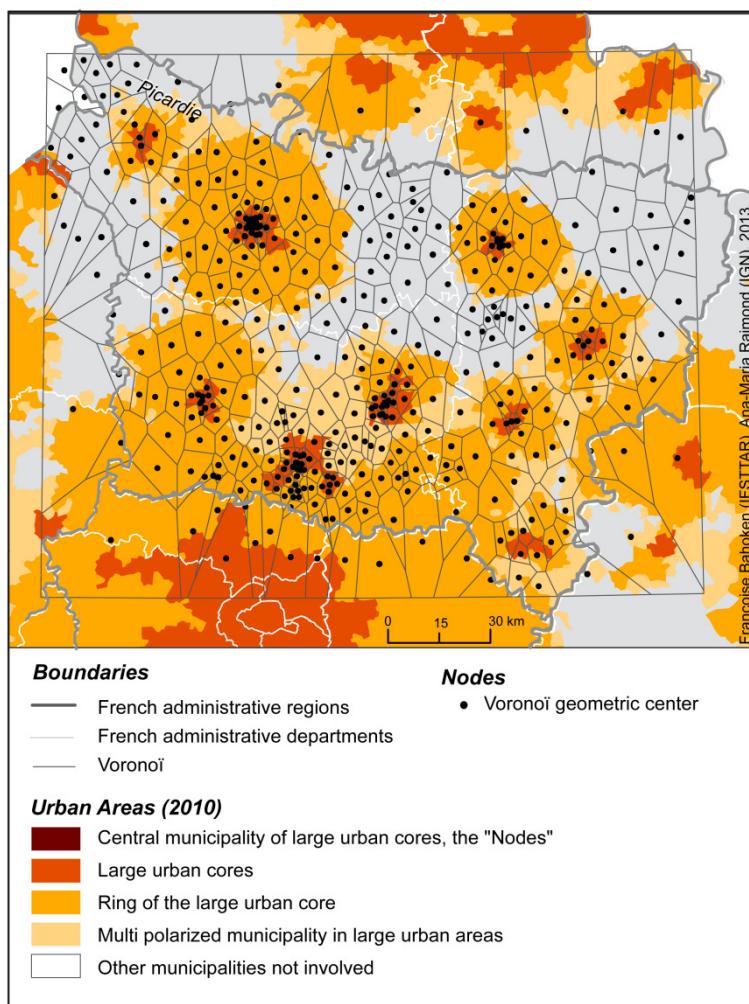


Figure 4. Urban Area zoning system (2010) and Node zoning system in The Picardie Region.

4.3. The effect of temporal filtering

As we noticed earlier, an OD matrix can be computed in defined time window interval. We are interesting to measure the quantity of links and flows which is lost when temporal filtering is applied. The number of links measures the interaction between places. The more, the number of links is the more relevant the interaction between places is. The number of flows quantifies how strong the interaction between two places is.

Generally, time window intervals are defined according to a specific goal such as flows analysis during the peaks hours, slack periods or work-days flows analysis (Lavielle 2008). According to this, different time window intervals were applied: 24h-, [6am - 10pm], [4pm - 8pm]. Thus, the temporal filtering generates three datasets. The OD matrix method was applied to each dataset at Voronoi polygons level. The results are then aggregated in four classes: Weekday class contains respectively links and flows calculated for weekdays in 24 hours time interval. Morning and Evening weekday classes represents the links and the flows computed for weekdays rush hours, respectively in [6am – 10pm] and [4pm - 8pm] time window intervals. Finally, Week-end class contains links and flows computed for week-end in 24 hours time interval. Table 1 shows the loss information (in terms of links and flows) after time filtering. Notice that *total* represents information before applying the following temporal filters (weekday, morning weekdays, evening weekdays and week-end).

As we can see in Table 1, generally the effect of temporal filtering is more important for the quantity of flows based on the total.

For example, morning and evening weekday's classes lose more than 65% of flows. However, the most effective temporal filter is the one that keeps significant values, of flows while reducing the number of small links, which help to confuse the message. From this point of view, it is then interesting to observe that when a 24 hours filter is applied for weekdays only 2% of areas interaction is lost since, 36% of flows are lost. Thus, selecting links that occur during the weekdays, allow keeping 64% of the total flux values, which is relevant and leads to a loss of 2% of the number of links that is not substantial. The 24 hours filtering is the one which keeps the part of the most relevant information (links and flows).

	Percentage of Loss Information	
	Links (%)	Flows (%)
Total	0%	0%
Weekday	2%	36%
Morning Weekdays	24%	65%
Evening Weekdays	23%	71%
Week-end	19%	64%

Table 1. Loss information account due to the temporal filtering.

4.4. The effect of spatial filtering

In order to study the effect of spatial filtering we focused on commuting flows for weekdays. Combining morning and evening flows, sub-matrices at several spatial levels allows us to generate a complete matrix representing daily flows. Daily flows are computed at three different scales (e.g. the Voronoi, the urban area and node scales) using the approach described in Section 4.1. The Voronoi scale is the most detailed scaled that it can be consider, knowing the spatial resolution of mobile phone data.

We notice that the spatial filtering has no apparent effect between the Voronoi and UA scales i.e. that no information is lost when merging Voronoi to UA. The spatial filtering effect is however relevant for node spatial scale (55% of flows are lost when passing from Voronoi scale to node scale). Nevertheless, the effect of spatial filtering can be measured by observing the distribution of inter and intra flows. Therefore to measure the change of the importance of zones, both inter zonal and intra zonal flows are analyzed. Let's remember that the elements of the OD matrix, noted F_{ij} , represent the number of trips (flows) from one origin area i to one destination j during a determined time interval. The OD flows matrix (F) is asymmetric if $i=j$, then $F_{ij} \neq F_{ji}$.

Table 2 described the percentage of inter and intra flows considering the spatial scale. We notice that at Voronoi scale the proportion of flows is quite similar between inter and intra flows. The most relevant spatial effect is noted at node scale when 97% of flows are inside the nodes, the interaction between nodes being weak (3%).

	Inter Flows (%)	Intra Flows (%)
Voronoi	46%	54%
UA	15%	85%
Node	3%	97%

Table 2. OD matrix results at different scales.

In Figure 5 obtained flows at Voronoi scale are represented.

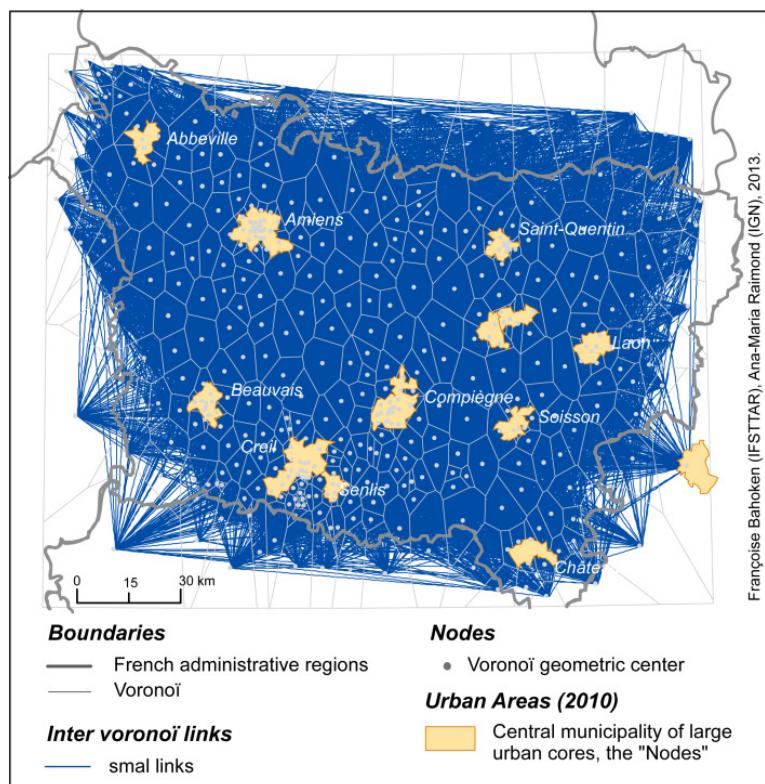


Figure 5. The total amount of flows at Voronoi Scale –example of spaghetti effect.

Mapping such flows requires the use of two different filtering procedures due to the spaghetti-effect (see Figure 5). The first one focuses on reducing the number of ties represented that we called the graphical filtering and the second one, on the aggregation procedure which is, in reality, a spatial filtering procedure based on the flow values (Bahoken, 2012). The spatial distribution of antennas determines a very specific Voronoi coverage: antennas are numerous and aggregated in urban and peri-urban areas and scattered in rural one (see Figure 5). This distribution prohibits the definition of partitions based on the relevance of

the value of flows. Therefore, we proposed an analysis based on inter Voronoï links' length, but not weighted by the number of links involved in the first place.**Erreur ! Source du renvoi introuvable.** shows the normal Quantile-Quantile diagram of link's distance. It compares the observed distribution with a Gaussian. The point are not on the first bisector so that the distribution do not follow a standard Gaussian distribution law (the standard deviation is 37.6).

Figure 6. Normal distribution of links length (km).

The median value of the Voronoï links' length is 75 km, the minimum is 0.6 km and the maximum is 208 km. Thresholds can be defined to observe the distribution of flow values according to the distance. We have chosen to present, by way of illustration, the following three thresholds based on quartiles (first quartile, median, third quartile). Three classes of links are obtained: small links- distance smaller than 44 km, medium links- distance between 44 and 53 km) long links- distance greater than 99 km (see Figure 7).

The analysis of the distances between the antennas shows that areas having a relevant interaction (i.e. the number of flows is important) correspond to urban areas. We also notice that the small links are inside the urban core which corresponds to our nodes and the medium links polarized the urban areas. The *Amiens* Node has a central position in terms on in links.

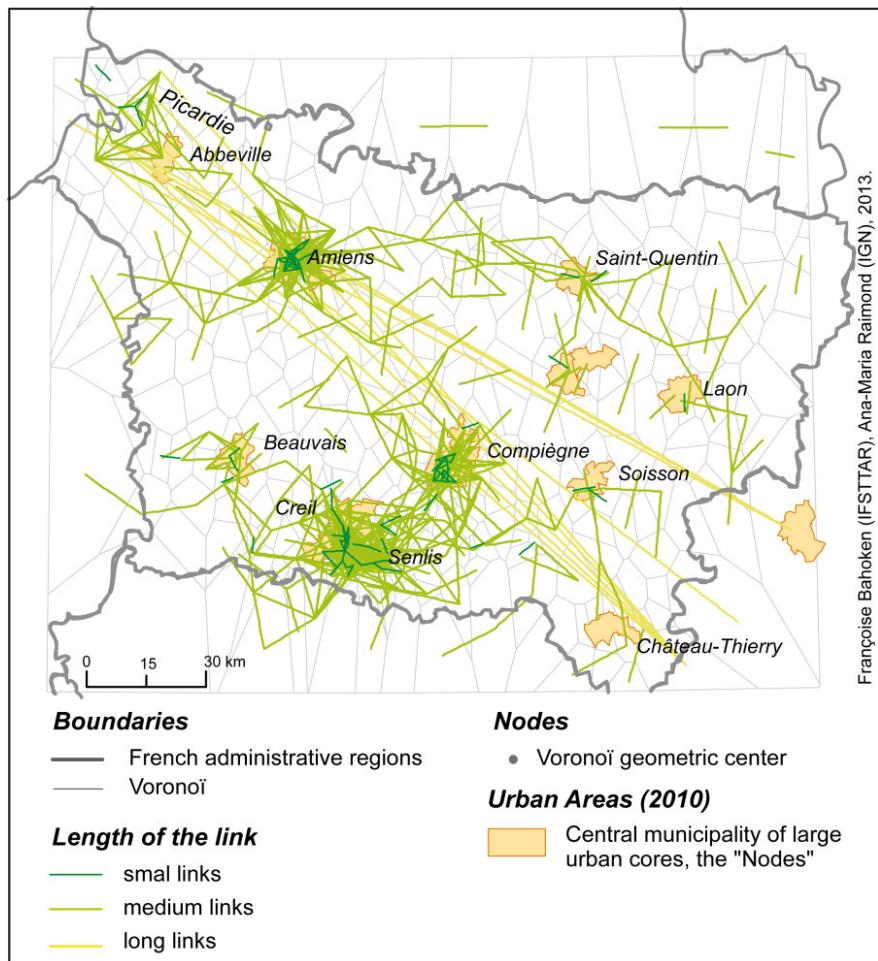


Figure 7. Flows representation according to the link's length at Node scale.

Figure 8 shows the flows representation at Node scale. It should be noted that, at this scale, there is no spaghetti effect and all bilateral flows can be represented.

The value of the flows represents the quantity which has moved between two nodes, for a week. We can note that the most important flows occurred between *Beauvais* and *Tergnier* then between *Senlis* and *Compiègne*. It is therefore at the heart of the *Picardie* region that produces the most important exchanges. Finally, it should be noted that this result obtained at the poles scales does not highlight the centrality of *Amiens* that was observed at the level of Voronoï. This assumption illustrates the instability of the statistical results when changing spatial scale, it deserves to be deep-

ened by a more analysis of the flow values, but it was not the goal of this paper.

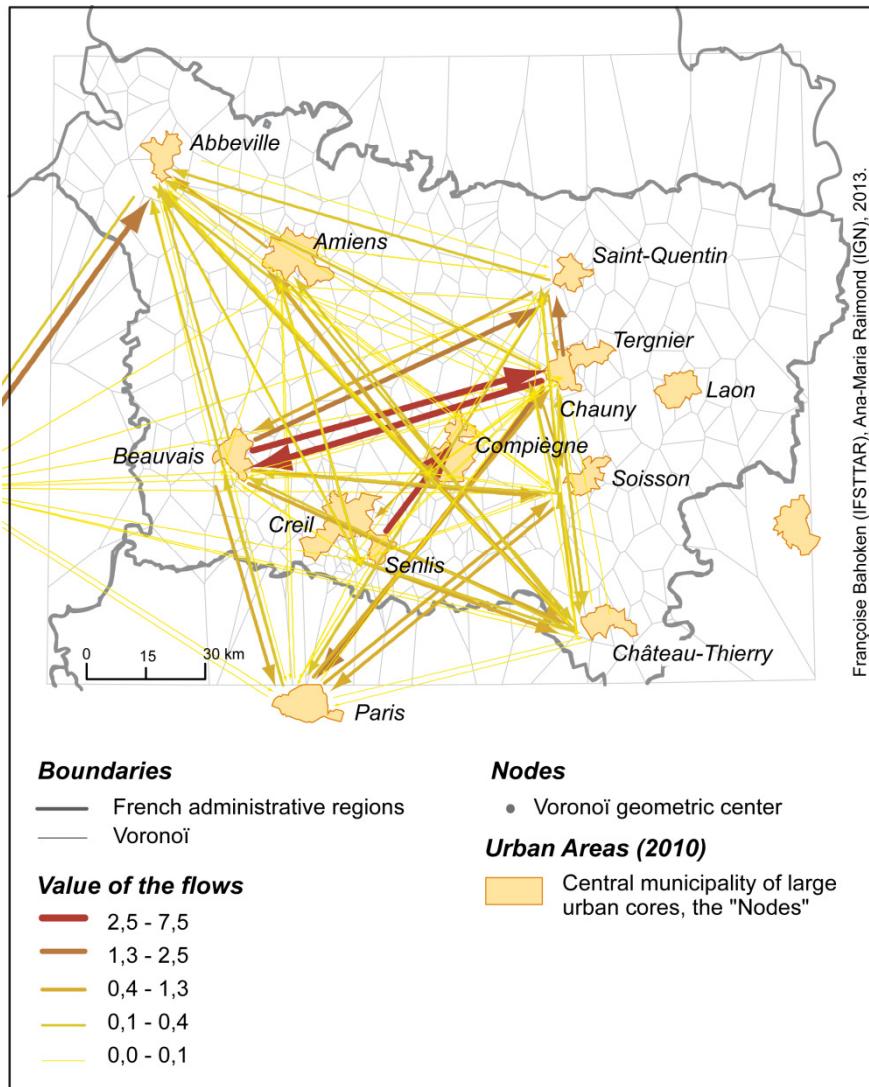


Figure 8. Flows representation at Node scale.

5. Conclusion

In this paper we focused on OD matrix design using mobile phone data and observed the loss of information associated with the different filtering procedures. A new approach allowing taking into account the mobile phone data spatiotemporal heterogeneities is then proposed. Since the computation of OD matrix depends on both temporal and spatial filtering, an analy-

sis of the effects of different spatiotemporal filtering was carried out. The results show that the more the temporal window is small the more the information in terms of links and flows is lost. Concerning the spatial filtering, flows were aggregated at three different scales: Voronoi, urban area and node. We observed that the loss of information is no relevant when flows are aggregated from Voronoi scale to UA scale and it is relevant (55%) when flows are aggregated at poles nodes. It is important to notice that the spatial filtering has relevant consequences on the results when inter and intra flows are distinguished.

In order to respond to privacy issues, few conditions were respected: all records were anonymized, no individual demographic or socioeconomic data were added to mobile phone data, the commuting flow detection were made at the spatial aggregate level and finally, the information presented is always aggregated.

References

- Ahas R, Aasa A, Roose A, Mark, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29: 469-486
- Bahoken F (2011) Comparison of functional regionalization's of the world: a methodological study based on Intramax procedure, in European Colloquium on Quantitative Geography Proceedings, Athens, pp. 647-654
- Bahoken F (2012) Application du raisonnement logique à la cartographie des flux, Proceedings of SAGEO, International Colloquium of Geomatics, Liège, Belgique, 7-9/11/2012
- Byeong-Seok Y, Kyungsoo C (2005) Origin-destination estimation using cellular phone as information. *Journal of the Eastern Asia Society for Transportation Studies* 6:2574–2588
- Caceres N, Wideberg J, Benitez F (2007) Deriving origin-destination data from a mobile phone network, *Intelligent Transport Systems, IET* 1(1): 15-26
- Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area, *IEEE Pervasive Computing*, 10(4):36-44
- Frias-Martinez V, Soguero C, Frias-Martinez E (2012) Estimation of Urban Commuting Patterns Using Cellphone Network Data, *ACM SIGKDD Int. Workshop on Urban Computing*, Beijin, China
- Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data, *The VLDB Journal — The International Journal on Very Large Data Bases*, 20(5):695-719

- Grasland C, Bahoken F, Beauguitte L, Pion G, Van Hamme, G (2009), Toolbox for flows and network analysis (Methodological Paper). Deliverable D.5.1. Euro-BroadMap.Vision of Europe in the World. Small or medium scale focused project FP7-SSH-2007-1
- González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns, *Nature* 453:779–782
- Holland, S.C., Plance, D.A., 2001, Methods of mapping migration flow patterns. *Southeastern Geographer*, 41(1): 89-104
- Insee (2011) Base communale du zonage en aires urbaine 2010, URL : http://www.insee.fr/fr/methodes/default.asp?page=zonages/aires_urbaines.htm (verified, 2011, 05, 09)
- Olteanu Raimond AM, Couronne T, Fen-Chong J, Smoreda Z (2012) Le Paris des visiteurs, qu'en disent les téléphones mobiles ? Inférence des pratiques spatiales et fréquentations des sites touristiques en Ile-de-France. *Revue Internationale de la Géomantique*, 3:413-437
- Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular Census: Explorations in Urban Data Collection. *IEEE Pervasive Computing* 6:30-38
- Sevtsuk A, Ratti C. (2010). Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60
- Smoreda Z, Olteanu-Raimond AM, Couronné T (2013) Spatiotemporal data from mobile phones for personal mobility assessment, In Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), *Transport Survey Methods: Best Practice for Decision Making*, Emerald Group Publishing, London
- Tobler, W., 1987, Experiments in migration mapping by computer, *American Cartographer*, 14:155-163

Travel demand estimation and network assignment based on cellular network data

David Gundlegård, Clas Rydergren, Nils Breyer and Botond Rajna

Journal Article



N.B.: When citing this work, cite the original article.

Original Publication:

David Gundlegård, Clas Rydergren, Nils Breyer and Botond Rajna, Travel demand estimation and network assignment based on cellular network data, COMPUTER COMMUNICATIONS, 2016. 95(), pp.29-42.

<http://dx.doi.org/10.1016/j.comcom.2016.04.015>

Copyright: Elsevier

<http://www.elsevier.com/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-134086>



Travel demand estimation and network assignment based on cellular network data

David Gundlegård*, Clas Rydberg, Nils Breyer and Botond Rajna

Department of Science and Technology, Linköping University

Abstract

Cellular networks' signaling data provide means for analyzing the efficiency of an underlying transportation system and assisting the formulation of models to predict its future use. This paper describes how signaling data can be processed and used in order to act as means for generating input for traditional transportation analysis models. Specifically, we propose a tailored set of mobility metrics and a computational pipeline including trip extraction, travel demand estimation as well as route and link travel flow estimation based on Call Detail Records (CDR) from mobile phones. The results are based on the analysis of data from the Data for development "D4D" challenge and include data from Côte d'Ivoire and Senegal.

Keywords: mobility analytics, travel demand estimation, traffic modeling, mobile phone call data, cellular network data, call detail records, intelligent transport systems

I. INTRODUCTION

The use of cellular network signaling data has the potential to fundamentally change how we can analyze the efficiency of a current transportation system, estimate transport models, and predict future transportation use. By mapping the cell phone data to the transport infrastructure it becomes possible to estimate the current use of the transport system. From the results of such estimations, suggestions for improvements to the existing transport system can be generated. The outcome would be more efficient mobility and, in the long run, increased economic growth. Furthermore, in developing countries the cellular networks can provide a much better coverage than traditional sensor infrastructure for traffic and transport. Therefore, this type of data will be very important to generate decision support information for large infrastructure investments.

Investments in transport infrastructure have been identified to have a positive effect on the economic growth. Since large transport infrastructure investments are very costly, it is important to make careful analysis of the cost-benefit-ratio for each potential investment. The use of mobile phone data for planning of transport infrastructure has been shown to have great potential (see e.g. Berlingero et al., 2013 and Blondel et al., 2013).

One benefit of using cellular network data over traditional sensors, like link counts and manual travel surveys, is a much better spatial coverage. However, the ubiquity of the data together with the relatively easy and fast deployment, once efficient software has been developed, makes it possible to also perform studies that have a temporal component. Examples include before and after studies to evaluate the effect of transportation investments as well as trends with several

different types of resolution in time, e.g. days, weeks, months or years.

In travel demand estimation based on cellular network signaling data we get direct observations of the generated trips and the distribution of trips for a large sample of the population. Dynamic origin and destination matrices can be constructed using techniques for assigning trips into time periods.

Cellular network data gives a possibility for a much better understanding of dynamic travel patterns, which has a large number of different applications within traffic and transport management, analysis and decision support. However, the data source has several key characteristics that are different from traditional data sources and these characteristics needs to be carefully handled while processing the data for estimation and prediction purposes. Unlike fixed infrastructure systems for data collection, cell phone signaling data is not bounded by any transport mode or any specific spatial region. This makes it possible to analyze the travel demand and travel times independent of travel mode.

A. Aim

The aim of this article is to outline the potential of mobile cellular network data with focus on Call Detail Records (CDR) in the context of mobility, transport and transport infrastructure analysis. We describe how mobile phone data can be processed to enter in traditional transportation analysis models and a modified methodology for handling the different steps in travel demand estimation and network assignment.

B. Contribution

A key outcome of the article is a set of mobility metrics, based on the concepts of trajectories, trips and cellpaths that can be estimated using the present type of CDR data. Based on these metrics, we present new algorithms for dynamic demand and route choice estimations as well as some potential applications for this type of data in Côte d'Ivoire and Senegal, applicable also to other regions where the same type of data is available.

Travel demand analysis for transportation planning is traditionally performed using the classical four-step model, which divides the problem into 4 different sub-problems: trip generation, trip distribution, mode choice and finally route assignment (see Figure 1). From cellular network data we get direct observations of combined trip generation and trip distribution, and to some extent also route choice, for the users in the data set, but the poor resolution in time and space in CDR data causes problems to relate antenna movements to physical movements. The poor resolution in time and space is even more problematic in the last two steps, mode choice and route choice. A key component in this paper is to present a set of tools that

* Corresponding author: David Gundlegård, david.gundlegard@liu.se, Linköping University, 601 74 Norrköping, Sweden.

enables efficient use of CDR data for understanding mobility from a transportation planning perspective.

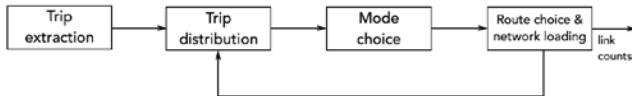


Figure 1: Overview of the traditional four-step model.

To be able to make analysis on route choice, temporal demand characteristics as well as travel times, we have decoupled these parts from the travel demand estimation, i.e. the trip extraction. All trips are used in the travel demand estimation and different subsets of trips are used for different parts of the processing pipeline, depending on their spatiotemporal characteristics. For route choice we have filtered trips that have good resolution in space and for temporal demand analysis as well as for travel time estimation we have filtered trips with good resolution in time. Due to the large amount of trips in the whole data set, we can still get enough observations to enable also analysis of dynamics that is rarely captured in the majority of user trajectories. The processing pipeline from the raw CDR data and cell tower locations to the link travel flows are illustrated in Figure 2.

In this paper, the process of scaling demand data to be representative for the full population of an area is not discussed. A discussion of techniques for such upscaling can be found in e.g. Jiang et al. (2015).

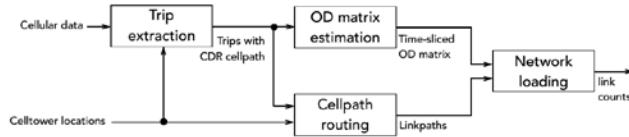


Figure 2: Overview of the processing pipeline.

C. Outline

The rest of the paper is organized as follows. In Section II the background to the studied cases of Côte d'Ivoire and Senegal are presented. In Section III, a review of the previous work on demand and flow estimation from cellular network data is given. In Section IV trip definitions are discussed and a new procedure for extracting trips from CDR data is presented. Section V presents a set of mobility metrics tailored for transport analytics based on CDR data. In Section VI a new procedure for generating time dynamic origin destination matrices from trips are given. Section VII covers the technique of assigning travel flow to routes and links. Section VIII provides a discussion of the results and section IX concludes the paper.

II. BACKGROUND

The analysis in this paper is based on the two data sets provided by the mobile operator Orange in the two research challenges; Data for Development (D4D) - Côte d'Ivoire in 2012/2013 and D4D - Senegal in 2014/2015 (Blondel et al., 2013, de Montjoye et al., 2014). The mobility data consists of timestamps, antenna IDs and user IDs. The positions of the calls are identified according to the connected antenna. The position

of each antenna is given as the longitude and latitude, slightly blurred to obfuscate sensitive information. The coverage area of each antenna is approximated by the corresponding Voronoi cell.

It should be noted that these datasets contain data from call data records only, i.e. a limited subset of the mobility data that is available in different interfaces of the cellular networks. An overview of other types of data that can be collected from the cellular network, for example location updates, handover events or measurement reports, is given in Gundlegård and Karlsson (2006).

A. Côte d'Ivoire dataset

Côte d'Ivoire is located in the west of Africa and has about 19 million inhabitants. The city with the largest number of inhabitants is the city of Abidjan. Abidjan is located at the coast in the south east part of the country.

The data was collected during a period of 150 days between Dec. 1st 2011 and Apr. 28th, 2012. This period covers 2.5 billion calls and SMS messages. This dataset has several subsets where each subset is a user trajectory table, in which the positions of the connected antennas are described for 50.000 users during a two-week period. There are ten two-weeks periods altogether, where IDs are changed for each period. An overview of the road infrastructure, as presented in the Open Street Map, is presented in Figure 3, where also the distribution of the mobile antennas is shown, represented by the red dots.

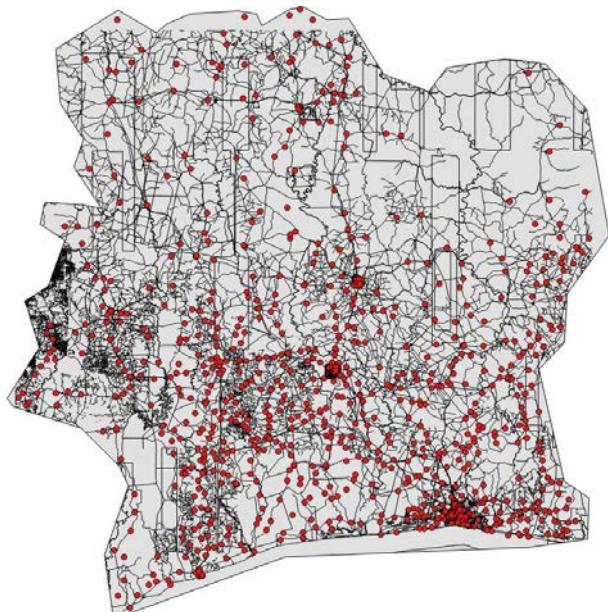


Figure 3: Antenna distribution (red dots) and road network from Open Street Map for Côte d'Ivoire.

B. Senegal data set

Senegal is located in the west of Africa and has about 12 million inhabitants. The capital of the country is Dakar in the far west part of the country and close to the Atlantic Ocean. Dakar has 1.1 million inhabitants, with about 2.7 million inhabitants in the urban area close to the city.

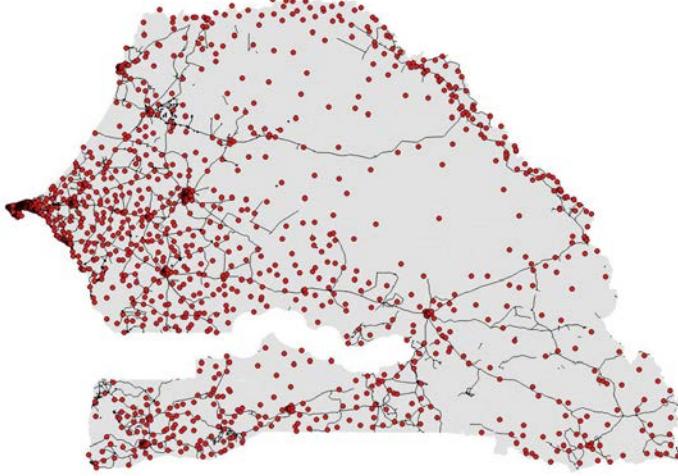


Figure 4: Antenna distribution (red dots) and road network from Open Street Map for Senegal.

The data is collected between January 1, 2013 and December 31, 2013. The data used in this paper consists of 1666 antenna locations (see Figure 4) and mobility data on a rolling 2-week basis for a year for about 300,000 randomly sampled users.

III. PREVIOUS WORK

The use of cellular network data to understand mobility patterns has been studied almost since these networks became widely available. A meta-study presented by Steenbruggen et al. (2013) contains studies in the field from as early as 1994. As algorithms evolve and processing large-scale data becomes easier, cellular network data is on the way to become a natural complement to expensive travel surveys and observations that are typically only available for a much smaller sample than cellular network data (Becker et al., 2011). The information that can potentially be estimated from cellular network data includes not only the travel demand and traffic flows, but also metrics like the daily range of travel (Becker et al., 2011) or the home and workplaces of the users (as in Alexander et al., 2015, Gundlegård et al., 2015 and Isaacman et al., 2011), which can be interesting for analyzing commuting patterns.

There are different kinds of data that can be acquired from cellular networks. Many studies are using CDRs, which only occur, when users are actively using the phone, while others had access to location area updates that are occurring more frequent and independently of the calling behavior of the user (see Table 1 for an overview of recent studies). Some studies, such as Shad et al. (2012), collect data on the phone side instead of the network side, which is potentially more detailed, but requires additional software to be installed on the mobile unit.

In order to extract the movements relevant for traffic analysis from the raw cellular network data, most studies perform some kind of trip extraction partitioning the raw data into stationary sections and sections of movement. Due to the fact that cellular network data can contain a lot of noise, there is no obvious definition of what a movement/trip is. Therefore, trip extraction algorithms vary a lot among different authors. Several studies like Iqbal et al. (2014), Ming-Heng et al. (2013), Gundlegård et al. (2015), Sohn et al. (2006) use a time-window during which a

continuous movement has to be detected in order to filter out cell-switching noise between neighbor cells, which can occur even if the user did not physically move. Another widespread concept is to merge subsequent locations in a user's trajectory if they are spatially close, see for example Alexander et al. (2015), Leontiadis et al. (2014), Shad et al. (2012), Toole et al. (2015) and Calabrese et al. (2011).

Table 1: Recent studies using cellular network data for traffic analysis.

Paper	Dataset	Location	Major contributions
Gundlegård et al. (2015)	CDRs (D4D dataset)	Senegal	Trip extraction, challenges of cellular network data
Shad et al. (2012)	LAC/Cell ID recorded on 100 phones, 9 months	Worldwide	Estimating position from LAC/Cell ID, clustering locations
Calabrese et al. (2011)	CDRs, 1M users	Massachusetts (USA)	Trip extraction, scaling with census data
Fillekes (2014)	CDRs and GPS traces	Estonia	Validation of map-matched CDR trajectories using GPS
Zang et al. (2011)	CDRs, 3 month, 25 million users	USA	Differential privacy to preserve personal integrity of users
Larijani et al. (2015)	Location area updates for Paris	Paris (France)	Detection of subway segments, O/D flows
Doyle et al. (2011)	One week of CDRs, 2009	Ireland	Trip extraction, mode detection
Becker et al. (2013)	CDRs for 5% of subscribers, 62 days	Los Angeles and New York (USA)	Daily range of travel estimation, home- and work location estimation
Ming-Heng et al. (2013)	AirSage position data based on CDRs	Kansas city (USA)	Trip extraction, data filtering, O/D flows
Alexander et al. (2015)	CDRs, 2M users, 60 days, spring 2010	Boston (USA)	Trip extraction, home-/work estimation, trip scaling, O/D flows
Hoteit et al. (2014)	AirSage position data based on CDRs, one day 2009	Massachusetts (USA)	Comparison of trajectory interpolation methods, localization of popular places
Iqbal et al. (2014)	CDRs, 6.9M users, 1 month and traffic counts at 13 locations	Dhaka city (Bangladesh)	Trip extraction using time window, trip scaling using traffic counts, O/D flows

Using the extracted trips, an origin-destination matrix (OD-matrix) containing the travel demand between each pair of zones can be computed. The travel demand can be given in different forms. The most obvious is to simply aggregate trips that start and end at the same zones as done by Calabrese et al. (2011), Larijani et al. (2015) and Ming-Heng et al. (2013). This gives an estimation of the number of cellphone users of the operator that provided the data that are travelling. While this might be good enough to understand how the travel demand distributes relatively between different OD-pairs, it doesn't give an absolute estimate of the travel demand for the whole population. Alexander et al. (2015) estimate the total travel demand in terms of the number of people travelling using scaling factors obtained from census data. A third way of expressing travel demand is in terms of the number of vehicles (see Caceres et al., 2007, Iqbal et al., 2014, Toole et al., 2015), which especially is interesting for the comparison with road traffic counts. While Caceres et al. (2007) use a "cell-phone per vehicle equivalent" computed using the market share of the operator and population statistics, Iqbal et al. (2014) use a micro-simulation to obtain a scaling factor per individual OD-pair and Toole et al. (2015) rescale trips using census data. As travel demand varies over the day and during the

course of a week, Calabrese et al. (2011), Ming-Heng et al. (2013) among others use time-sliced OD-matrices. As origin and destination zones the cells defined by the base-stations can be used as in Larijani et al. (2015). For the purpose of comparison with other data, some authors like Calabrese et al. (2011) convert the travel demand to be between Traffic Analysis Zones (TAZs) instead of between cells.

Several attempts have been made to reconstruct the specific travel mode and route that a user took for a trip in order to perform a traffic assignment providing the flows on each link of the transportation network. The travel mode classification is challenging given the low temporal and spatial resolution of CDRs. However, for example, the extraction of subway travel can be done reliably given that subway tunnels are being served by dedicated base stations (Larijani et al., 2015). Above ground the geographical shape of routes (Doyle et al., 2011) or characteristics like different travel times among modes can be used (Wang et al., 2010, Sohn et al., 2006) to classify the mode of travel.

To infer the route for road-bound traffic Fillekes (2014) used several map-matching techniques as they are typically used with GPS data. However, these methods perform poorly with spatially and temporally sparse CDRs. Leontiadis et al. (2014) proposes to calculate a shortest-path using lowered link costs for the links inside cells that the user connected to during the trip to make the route more likely to pass through these cells. Another approach used by Wu et al. (2015) and Fiadino et al. (2012) is to fetch a predefined set of alternative routes for each OD-pair and select the route that has the highest spatial similarity with the cellpath (the cells that a user connected to during a trip). Tettamanti et al. (2012) estimate link flows assisted by classical traffic assignment methods. Other studies use Bayesian classification (Gundlegård et al., 2009) and neural networks (Demissie et al., 2013).

IV. TRIP DEFINITION AND EXTRACTION

The challenge in using CDR data for travel demand analysis is to infer travel behavior based on a set of sparse space-time tuples with a spatial resolution limited to the antenna deployment density. Let us call all the space-time tuples available for one user a trajectory, the aim is to infer the physical movements for this user based on the sparsely sampled trajectory. For travel demand analysis we are interested in turning the user movements into a finite set of trips, so that we can aggregate different user's trips and get an understanding of the demand to travel between different geographic areas. Hence, we are interested in turning the trajectories into trips, which are then aggregated in space and time to describe a travel demand. Depending on how we define a trip, or even more important, how we design the process of extracting trips from a trajectory, the travel demand can be very different. The process of turning the trajectory into a set of trips is here referred to as trip extraction and a trip is simply defined as a movement between spatially separated user activities. Although the trip extraction process is central when using CDR data for travel demand estimation, the literature on the topic is relatively unexplored.

A. Mobility characteristics

The resolution in space and time of user location sampling is a key component in determining which type of mobility analysis that can be made with the dataset. To enable comparison of the results based on this dataset we have calculated average inter-event statistics for the Senegal dataset, see Figure 5, which can be compared with Figure 1 in Calabrese et al. (2013). Calabrese et al. analyze data that not only include call and SMS connections, but also connections to the Internet over the cellular network. They report an arithmetic average of 84 minutes for the medians (corresponds to the blue group). They conclude that the average of 84 minutes allows the detection of changes in locations where the user stops for as little as 1.5 hours. The corresponding values for our dataset is an arithmetic average for the medians (blue group) of 308 minutes which indicates that it would be possible to detect stops which are about 5 hours and longer.

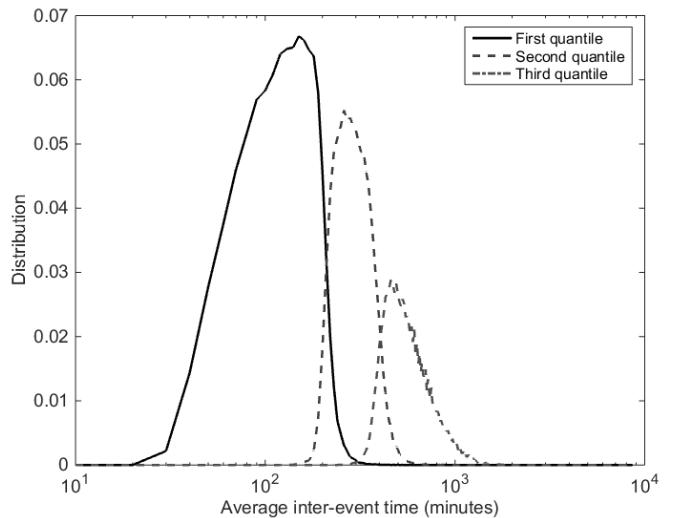


Figure 5: Distribution of average inter-event times for three quantiles for the Senegal dataset.

A problem when analyzing cellular network data based on user activities, e.g. SMS and phone calls, is the time bias in the samples; typically, users have a tendency to make fewer phone calls early in the morning. Figure 6 shows the total average number of events per hour, together with error bars showing one standard deviation, over the day. It can be seen that there is much more phone activity late in the evening compared to early in the morning. This impose a problem for estimation mobility, since the possibility of detecting mobility is lower for a small number of events. This becomes an important problem to take into consideration when scaling up results from the data set. The problem can be reduced by using time-dependent travel demand scaling, but this requires access to dynamic scaling data.

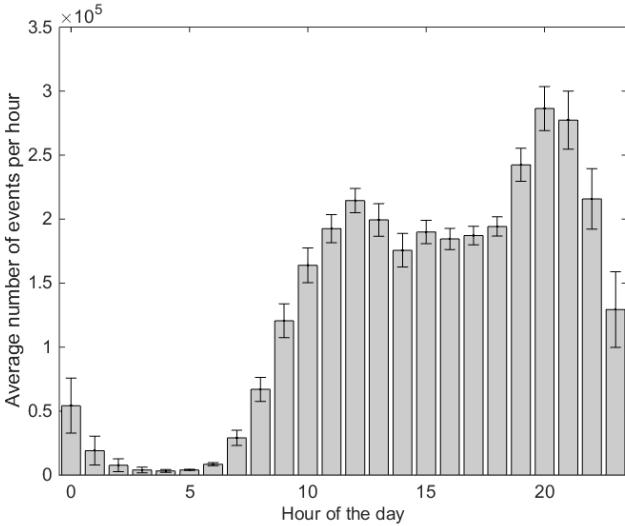


Figure 6: Average number of CDR data events per hour for the first two week period of data in the Senegal dataset. The error bars show one standard deviation.

B. Home and POI

Calling activity is correlated to the users' points of interests (POI). Therefore, it is feasible to estimate the home location of users based on a sequence of location observations, see e.g. Dash et al. (2014). Since POIs, especially home and work, are very important for a user's trip generation and distribution we have used the estimated home and work location of users as input to the trip generation. We have used the call events to estimate the home and work location of users, based on the frequency of calls from different locations during daytime and during nighttime. The home and work location are identified as locations with a minimum distance of three kilometers, and not belong to neighboring antennas in the Voronoi graph. By aggregating home and work locations for all users, we have a technique for identifying residential and industrial or public areas, which can be useful in developing countries where census data can be poorly updated. In Figure 7 a heat map of the difference between home and work locations is shown. Blue indicates more home locations than work locations, and red indicates more work locations than home locations. The red area in the middle of the figure is the International Blaise Diagne airport, located southeast of Dakar. The red area in the lower part of the figure most likely indicates an industrial area located along Route Sindia-Thies.

Since trips generally are generated from residential areas to industrial or public areas in the morning and the opposite in the evening, this kind of map can directly give a better understanding of trip production and attraction, compared to population statistics only.

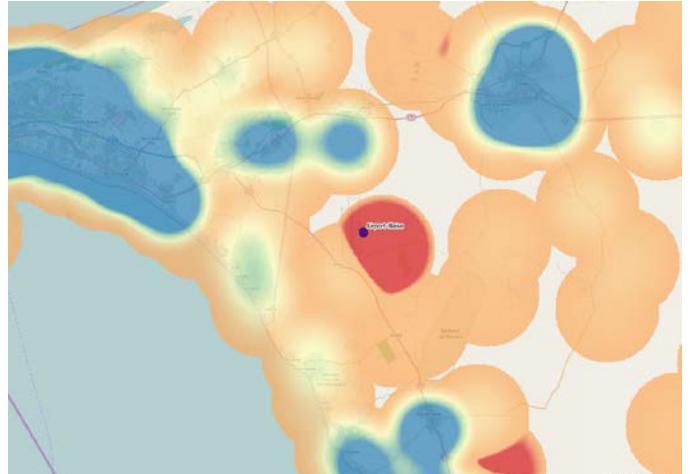


Figure 7: Heat map of difference of number of home locations and number of work locations. Blue indicates potential residential areas and red indicates areas with large daytime activity. The red area in the middle of the figure is an airport.

C. Extracting trips

In order to analyze travel demand and mobility, individual movements need to be identified and aggregated. In this section we define three ways of describing movement; *trajectories*, *trips* and *cellpaths*. Trajectories are the set of space-time tuples available for one user. Trips are here related to movements between activities and are only defined by start and end location, referred to as origin and destination. A cellpath is the sequence of connected cells for a given trip.

Trips are here referred to as movement between activities, hence trip extraction is equivalent to identifying different activities or stops in the movement. Several papers (e.g. Wang et al. 2012) define each antenna event as an activity, and a trip as a movement between these places. This very basic form of trip extraction typically generates a large number of very short trips. Wang et al. (2012) described OD observations based on this trip extraction as transient OD observations (t-OD). Other, more realistic, activity proxies have some kind of criteria for stops (e.g. Calabrese et al. 2011, Berlingero et al. 2013), which are mainly temporal and/or spatial thresholds for detecting stops. The trips are often also filtered by thresholds in inter-event time in order to be suitable to aggregate in different time intervals for dynamic OD estimation, which also removes a large number of physical trips.

When the temporal dynamics of the demand is less important a static trip extraction can be used. A common approach is to use predefined periods of time, (e.g. {10pm-7am} and {9am to 4pm}), where each can be associated with a "Home" zone and a "Work" zone, respectively. The originating or terminating zone of the user is calculated by its most common position during the predefined time period and a trip is detected if the originating zone is different from the terminating zone.

The advantages of generating OD matrices based on this definition is that we can capture a large part of the trips from home to work and that very few artefact trips caused by antenna oscillations are generated. Trips can be detected even if the sampling is very sparse, e.g. a "home sample" in the evening or

night and a “work sample” in the day is enough to capture a trip. However, travels within the relatively large spatial and temporal thresholds will not be detected.

In order to study travel demand, it is important to capture as many movements as possible from the CDR data, even with poor resolution in time and space. In this paper this is done using assumptions on travel behavior related to predefined POIs (here, home and work location) and by relaxing the constraint on inter-event time compared to standard trip definitions. The relaxed constraint on inter-event time requires a new methodology to aggregate the trips into different time slices in dynamic OD estimation, which is described in more detail in Section VI.

In the POI-based trip definition we assume that all movement start from the home location in the morning and end in the home location in the evening, unless the user’s distance to home is larger than a threshold value d_{max} . Furthermore, a threshold value, d_{min} , is used as a minimum movement distance to identify the start of a trip as well as snap the origin or destination location to any of the user’s POIs. One of the rationales for this trip definition is that it is relatively easy to estimate the home location of a user, given that the user has a sufficient number of events during the study period.

The POI-based algorithm for generating trips is divided into three functions, which are presented in Algorithms 1a, 1b and 1c. The *main()* function (Algorithm 1a) loops through all available CDR events of each user for each day. To begin with, the algorithm scans through the events and calls *detect_trip_start()* (Algorithm 1b) to detect if a trip start condition is fulfilled. In algorithm 1b and 1c, the function *d* computes the distance between two antennas. The trip start condition is fulfilled if the distance from the ending point of the previous trip (line 33) or from the home position in case of the first trip of a day (line 19) exceeds d_{min} . As long as a trip is active, *detect_trip_end()* (Algorithm 1c) is invoked for every event to check if the trip has ended. The trip is ended if the user arrives at home or at work (line 37 and line 45, respectively) or, alternatively, if two subsequent events have the same position (line 41). When a trip has ended, *main()* repeats the same procedure and tries to detect the next trip start of the user by calling *detect_trip_start()*.

main()

```

1 for each user  $u$ 
2   for each day  $a$ 
3     for each CDR event  $k$ 
4       let  $p_{uak}$  = position for event  $k$ 
5       if(trip_active == false)
6         trip_active = detect_trip_start()
7       end
8       if(trip_active == true)
9         trip_ended = detect_trip_end()
10      end
11      if(trip_ended)
12        store_trip()
13      end
14    end
15  end
16 end

```

Algorithm 1a: Main function for the trip generation.

detect_trip_start()

```

17 if (trip_set empty)
18   if( $p_{uak} \neq$  homebase and  $d(p_{uak}, \text{homebase}) < d_{max}$  and
       $d(p_{uak}, \text{homebase}) > d_{min}$ )
19     trip_active = true
20     origin = homebase
21   end
22   if( $p_{uak} \neq$  homebase and  $d(p_{uak}, \text{homebase}) > d_{max}$ )
23     trip_active = true
24     origin =  $p_{uak}$ 
25   end
26   if( $p_{uak} ==$  workbase and  $d(p_{uak}, \text{homebase}) > d_{max}$ )
27     trip_active = true
28     origin = homebase
29     destination = workbase
30   end
31 else
32   if( $p_{uak} \neq$  previous_trip_start(trip_set) and
       $d(\text{previous\_trip\_start(trip\_set)}, p_{uak}) > d_{min}$ )
33     origin = previous_trip_start(trip_set)
34   end
35 end

```

Algorithm 1b: Function for detecting trip start.

detect_trip_end()

```

36 if( $p_{uak} ==$  workbase or  $p_{uak} ==$  homebase)
37   destination =  $p_{uak}$ 
38 else
39   if( $p_{uak(k+1)}$  exists)
40     if( $p_{uak} == p_{uak(k-1)}$ )
41       destination =  $p_{uak}$ 
42     end
43   else
44     if( $d(p_{uak}, \text{homebase}) < d_{max}$ )
45       destination = homebase
46     else
47       destination =  $p_{uak}$ 
48     end
49   end
50 end

```

Algorithm 1c: Function for detecting trip end.

Figure 8 shows an example of generated trips for a specific user for a specific day. Blue circles are antennas in the trajectory for this user, the home location is marked by H , the work POI is marked by W and an additional location is marked by A . The location A is identified by two consecutive calls referring to the same antenna, here taken as an indication of an activity at this location. The trips generated in this case are 1) from H to W , 2) from W to H , 3) from H to A and 4) from A to H . Note that the fourth trip (A to H) is generated even if the trajectory does not end in H for the specific day. This corresponds to the generation of an activity profile, $HWHAH$, as discussed in Liu et. al. (2015).

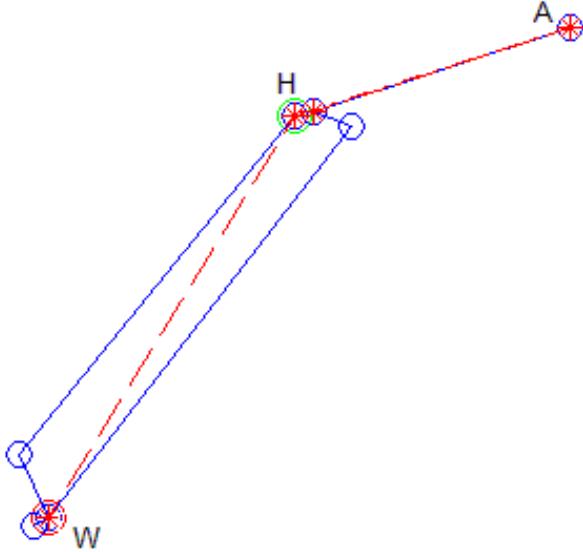


Figure 8: Example of trips generated for one user one day. Blue lines are trajectories and red lines are the generated origin-destination trips.

The number of trips generated using this trip definition in Senegal is approximately 0.7 trips per day and user, with d_{min} set to 3 km and d_{max} set to 100 km. This can for example be compared to the number of trips generated by adding a 60-minute temporal threshold and 5-antenna spatial threshold (Gundlegård et al, 2015), which is approximately 0.06 trips per user and day. Note that the sparse sampling in time causes a large amount of trips being discarded for the latter type of trip definition.

The distance distribution for the trips generated is shown in Figure 9. It can be noted that the number of trips tend to follow the decay of the distance with a negative exponential; similar to what is common in gravity models for trip distribution (Wilson, 1967).

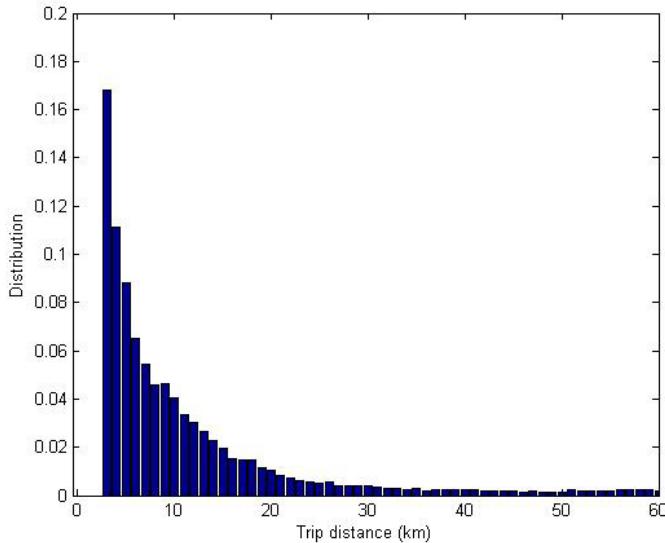


Figure 9: Trip distance distribution for the generated trips.

V. MOBILITY METRICS

In this subsection, we define five types of mobility metrics that can be estimated from CDR data: *static travel demand*, *dynamic travel demand*, *cell travel flow*, *route travel flow*, *link travel flow* and *travel times*. Dynamic travel demand and link travel flow is further discussed in Section VI and Section VII, respectively.

Since the spatial resolution of CDR data is relatively coarse, it is challenging to classify trips by individual vehicles. Therefore, we use the term travel flow to indicate that it is an aggregate of the number of devices (proxy for number of people) that travel between different cells, ODs or on specific links.

A. Static travel demand

Static travel demand is based on the static trip extraction described in Section IV.C and is suitable to reflect a commuter travel demand. Due to the static approach with only two time periods, the trip generation process is relatively easy and robust.

Traditional techniques for finding static origin-destination matrices include statistical models, entropy-based models and full travel surveys, which also lack a detailed temporal component. CDR data can be fused with these traditional observations, and be used to improve the quality of the output from the techniques. The demand data is used as input to models that predict the transport behavior in more detail, for example, how the demand is split on different travel modes. This is normally done using discrete choice models. These models also require mode choice data in order to be estimated. Choice data may also be possible to infer from CDR data. If this is the case, the data can be fused with observed choice data, and therefore contribute to an improved output from the choice model.

In order to benchmark the static traffic demand calculated from the CDR dataset an independent way of estimating the demand is used. A classical way of estimating traffic demand is to use a gravity model where the trip attraction between different zones are modeled based on standard parameters such as population density, distance and travel cost. Also explaining factors like socio-economic characteristics and land-use can be integrated in the model. The number of trips T_{ij} between two zones i and j can be computed as (Wilson, 1967):

$$T_{ij} = k \frac{O_i D_j}{d_{ij}^2}$$

where:

O_i = total trip origins at i

D_j = total trip destinations at j

k = adjustment factor

d_{ij} = distance from zone i to zone j

In our case we have total trip origins and total trip destinations proportional to the origin and destination population density, respectively. The impedance for a pair of zones is a function of the distance d_{ij} between the zones.

Figure 10 (right) shows the gravity model estimation distribution of traffic demand terminating in Abidjan from the

whole Côte d'Ivoire using a grid structure. Clear similarities can be seen when comparing the gravity model output with the estimates based on cellular network data shown in Figure 10 (left). Based on this we can conclude that in comparison to a well-established method to estimate traffic demand the cellular network data does at least not seem to contain any larger structural bias. However, the gravity model output is very rough and static by nature and the cellular network data can most likely improve traffic demand estimates dramatically compared to that.

Except for benchmarking, the gravity model can also be combined with cellular network data. The cellular network data can be used to estimate trip production, attraction and impedance parameters as well as socioeconomic factors. For example the travel times estimated in this paper (Section V.F) are typically a better impedance variable than the distance. In order to estimate trip productions with reasonable accuracy, relatively detailed information about cell phone penetration rates and usage is required.

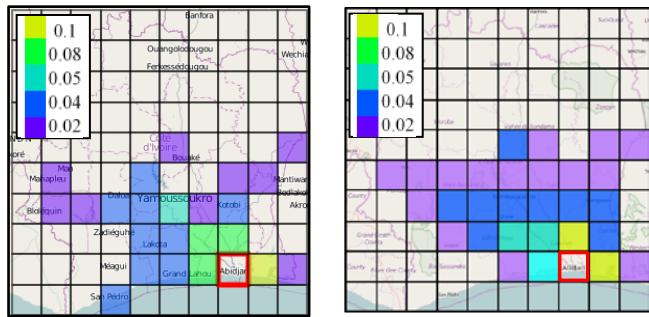


Figure 10: Left: Proportions of long distance trips that started between 9 am and 10 am and terminated in Abidjan. Right: Traffic demand proportions estimated by a gravity model based on population density and distance between zones terminated in Abidjan.

Figure 11 shows 13 different OD zones created using K-means clustering of antenna locations in Abidjan. K is chosen to reflect the number of TAZs in the city to enable comparison of results. Figure 12 and Figure 13 exemplify the importance of using a suitable trip extraction method, the figures show the number of trips between the different zones in Figure 11 in a static context for two different trip extraction methods. Figure 12 is based on trip extraction for transient OD whereas Figure 13 is based on static OD with only two time periods. Note that both the amount of trips as well as the spatial pattern is quite different. Typical for transient OD estimation is that more trips are generated between neighboring zones, which in reality might be trips that just pass through a zone.

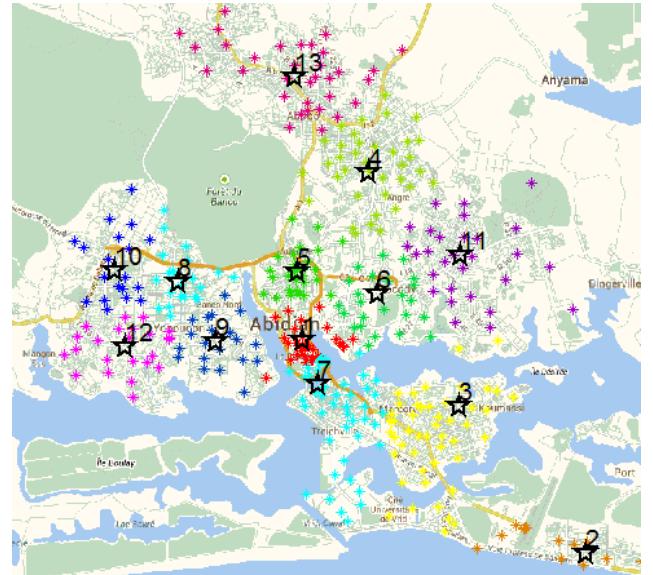


Figure 11: 13 clusters in Abidjan created using K-means clustering of antenna locations.

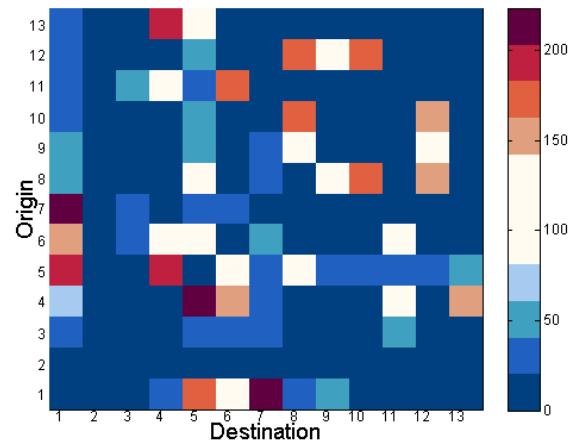


Figure 12: OD matrix proportion visualization for the different zones shown in Figure 11 based on t-OD trip extraction, i.e. each pair of antenna locations are considered as a trip.

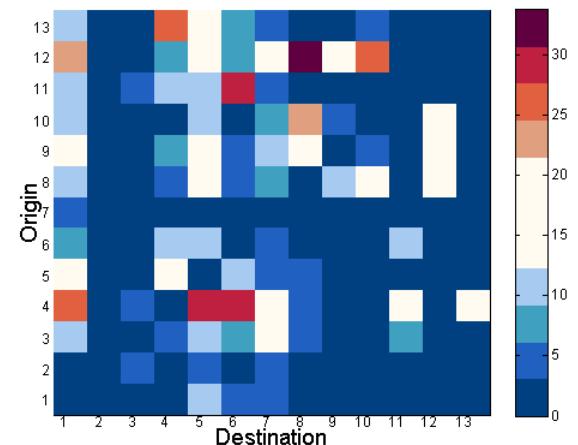


Figure 13: OD matrix proportion visualization for the different zones shown in Figure 11 based on static trip extraction, i.e. a trip occurs when a user has changed zone between predefined time periods.

B. Dynamic travel demand

Another way of defining the OD matrices is not to use a static definition of the times when a user is considered as being stationary, in order to try to capture as many trips made by the users as possible and make a more realistic representation of the travel demand. This mobility metric requires trip extraction with a better resolution in time compared to the static trip extraction.

The advantage of using this definition for calculating the OD matrices is that we do not need to make any assumptions on the travel times or habits of the users, e.g. such as time of day that they travel or when they are at home or at work. This type of mobility metric is suitable when a higher resolution in space and time is required to separate travels, for example when we want to capture travels to other activities than work, for example travels to shopping, daycare etc.

The traditional techniques for dynamic OD estimation require different input data to improve the temporal component, such as information from road traffic counts and travel time measurements from traffic cameras. All these measurements and models are possible to combine with the cellular network data, which potentially can provide reasonable accuracy also on dynamic OD estimation.

The methods developed in this paper for dynamic OD estimation together with results from Dakar are described in detail in section VI.

C. Cell travel flow

Travel demand captures how many users that travel between two zones in a certain time interval. By removing the stationarity requirement of users in the travel demand description of Section V.B and reducing the size of the zones, we are moving towards a metric that describes how many users pass between two cells during a specific time interval; hence we call this a cell travel flow metric. Note that there is no connection to the underlying transportation infrastructure here, the flow is only based on the movements between different antennas in the cellular network.

This type of analysis can be made directly on the trajectories or on trips after the trip extraction process. Doing this analysis directly on the trajectories typically generate a lot of flow due to antenna oscillations caused by radio resource management functions in the cellular network or poor cell coverage representation, which does not represent the physical movements of users in a good way. This noise can to a large part be removed if a suitable trip extraction method is used before the analysis and the cellpath of each trip is used as input.

We have applied the generalization and aggregation approach described in Andrienko and Andrienko (2012) for aggregating the mobile phone call data into cell travel flows between generalized places defined around the networks' antennae positions. This is performed in a sequence of steps as follows. First the trajectories are extracted from the mobile phone call data: The calls received/Performed by each user are ordered chronologically into a sequence of calls representing the trajectory of this user in space. Second generalized places are extracted by using the antenna positions as seeds around which Voronoi polygons are generated. These polygons define the set of places that the explored area is divided into. The trajectories are then aggregated into moves between pairs of places by

defining transitions between them, and counting the number of transitions present. Figure 14 shows a visualization of cell travel flows for the city of Abidjan.

If we reduce the spatial resolution of the zones, i.e. aggregate antennas into larger zones, while still not requiring any stationarity to separate trips, we get travel flows between zones. The difference compared to the travel demand is that travels that pass through a specific zone without any stop is counted as flow in and out of the zone, which can be compared to the transient OD metric.

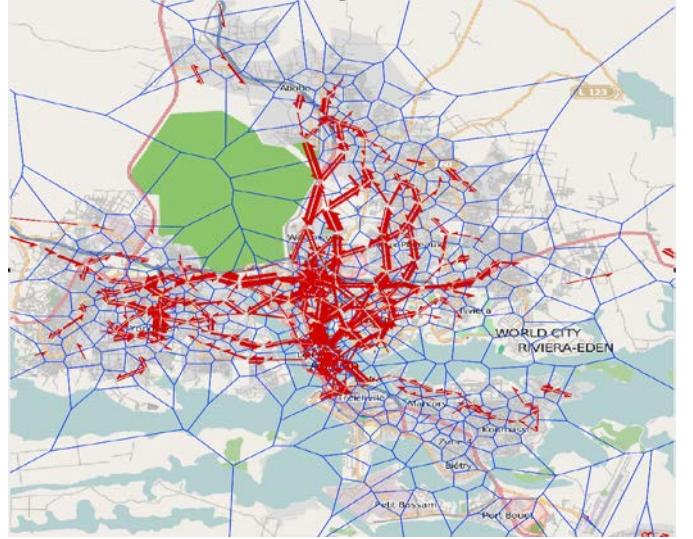


Figure 14: Cell travel flows between antennas represented with Voronoi polygons. The flow is aggregated for a two-hour period in Abidjan.

D. Route travel flow

With knowledge about the underlying transportation infrastructure, it is possible to map the trips to routes between ODs using the cellpaths. The routes could be specified for different travel modes, e.g. car, bus or train. The route mapping is preferably done on the trips data set, since each trip typically belong to one (main) travel mode. Due to the sparse spatiotemporal resolution of the CDR data, this is a quite challenging task, but it is possible to utilize the large amount of data available to generate reasonable results. The methodology and results for this is described in detail in section VII.

E. Link travel flow

By aggregating the different route flows, we can also get a link travel flow, which is simply the sum of all route flows that pass a given link. This mobility metric is well known in the transportation community and can be observed using a number of different sensors, e.g. radar sensors or loop detectors, which makes it possible to validate results relatively easy. Also this mobility metric is further described in section VII.

F. Travel times

A final mobility metric that we have analyzed for the CDR data sets is travel time, which is traditionally utilized extensively

to understand and assess the transportation system state and quality. Due to the fact that we can only obtain location in terms of antenna positions and that samples are limited to when the user is active, the measurements contain large errors in both space and time domain. The space domain errors limit us to measure travel times for travels that are of a minimum length, and the length requirement is dependent on which relative travel time (average speed) error that can be accepted. The error in time due to sparse sampling limits us to draw conclusions of the minimum travel time instead of the full travel time distribution. However, the minimum experienced travel time is also very useful and a good indicator of travel quality.

By dividing the travel time into two parts, one caused by distance and type of transport infrastructure and one caused by queuing delay we are able to identify parts of the network that are congested. We can do this for example by separating the measurements into peak hour measurements and non-peak hour measurements. By comparing the cumulative distribution function of travel time measurements for the two time periods, where unreasonably high travel times are filtered out, it is possible to identify a travel delay metric.

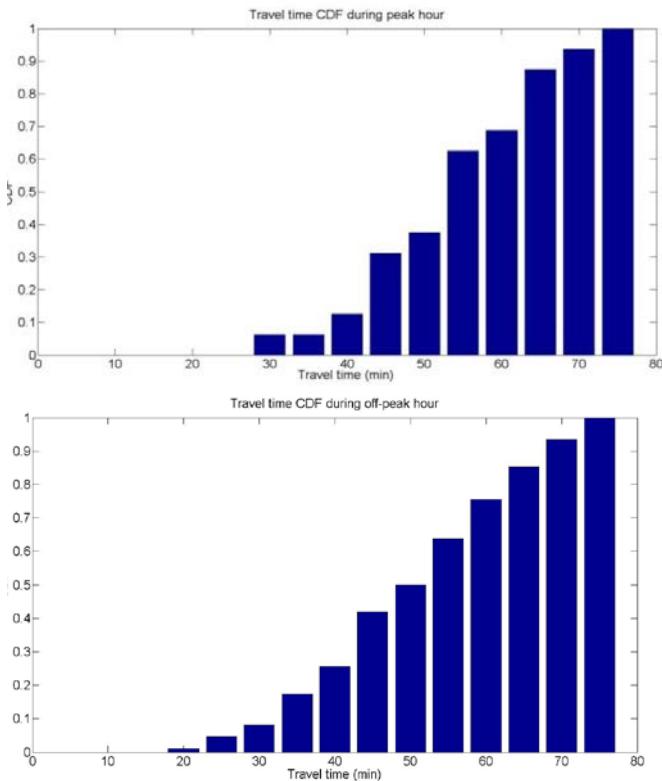


Figure 15: Top: Peak hour travel time measurement CDF between Abidjan city and Abidjan airport grouped in intervals of 5 minutes. Bottom: Off-peak hour travel time measurement CDF between Abidjan city and Abidjan airport grouped in intervals of 5 minutes. Travel time measurements larger than 75 minutes are not included.

We have estimated the travel time distribution between Abidjan city and Abidjan airport during off-peak hours (9-16, 18-06), see Figure 15 (bottom) and during peak hours (7-9, 16-18), see Figure 15 (top), for a time period of six weeks. By

comparing the two CDFs we notice that the minimum travel time is 10 minutes longer for peak hours, indicating that the minimum travel time increases approximately 10 minutes due to congestion in the road network. By combining the two CDFs with travel flow estimates, it is also possible to express aggregated delay metrics like total queuing delay per route and time period.

VI. DYNAMIC TRAVEL DEMAND ESTIMATION

The travel demand is one essential input to models for transportation analysis. The travel demand is normally described in an origin-destination matrix. Given a division of a geographical area and a division of the area into zones, the origin-destination matrix describes the number of trips from each pair of zones, e.g. from zone A to zone B for each pair (A, B). The origin-destination matrix describes the demand in a given time interval, for example one hour. Normally, the origin-destination matrix describes the number of trips that starts at zone A during the specified time interval, going to zone B.

Cellular network data is interesting from demand modelling perspective, since we can get direct observations of the travel demand for all transport modes, see Angelakis et al. (2013). Input for generating a time-sliced OD-matrix are the trips generated by the trip definition described in Algorithm 1a-c. These trips give direct observations of trip generation and distribution for the sample of users in the data set.

Previous work is mainly focused on estimating t-OD directly. In this paper we separate the demand estimation from both the temporal and spatial (route) distribution of the demand. This way we can extract behavior from subsets of the data where the accuracy in at least one of the aspects is high, and apply to a larger sample.

The spatial resolution of the data set is limited by antenna density. The antenna density is strongly correlated to population density and hence we get a better spatial resolution of trips in areas with denser population. However, the main problem when generating travel demand from CDR data might not be the spatial resolution, but rather the overlapping coverage of antennas, which makes the standard Voronoi representation of cell coverage a poor representation. This problem becomes worse in areas where macro cells with large transmission power in elevated positions are used for coverage and micro cells with low transmission power are used for capacity. We try to cope with the antenna oscillations by only considering trips longer than a minimum distance d_{min} and not consider trips between antennas that are Voronoi neighbors.

Since users are sampled only during phone activity in terms of calls and SMS, there is a large uncertainty in the temporal domain for the start and end of each trip. Since we want to include as many trips as possible to get a good estimate of the travel demand, we need to include trips with poor temporal resolution. We assign each trip to a time period according to the probability of the trip being started in each time period.

For an individual that makes a trip as defined by the trip definition, corresponding to a CDR at location A at 7:00 and a CDR at location B at 10:45, the contribution to the demand matrices will be computed as follows. First, we estimate a travel time based on the Euclidean distance from A to B and a travel

speed based on prior knowledge the road network (here a fixed value of 50km/h). Let us, as an example, assume that the distance between A and B is 50 kilometers, then we deduce that the trip has started sometime between 7:00 and 9:45. By assigning equal probability to all start times during this time interval, the contribution from this specific trip will be 1/2.75 to the demand matrix holding the demand from 7:00-8:00, 1/2.75 to the demand matrix holding the demand from 8:00-9:00 and 0.75/2.75 to the matrix holding the demand from 9:00-10:00, for the element representing the travel relation A-B. The trip weights assigned is illustrated in Figure 16. The weight for time slice i (hour) is given by the expression

$$w_i = \frac{\min\{t_B - t_{AB}, t_i + 1\} - \max\{t_A, t_i\}}{(t_B - t_{AB}) - t_A}, [t_A] \leq t_i \leq [t_B]$$

where t_i is the clock time (decimal) of slice i , t_{AB} is the estimated time of going from A to B, t_A is the known clock time at the start location, t_B is the known clock time at the end location and $[t]$ and $[t]$ denote the rounding down and up to the nearest hour, respectively.

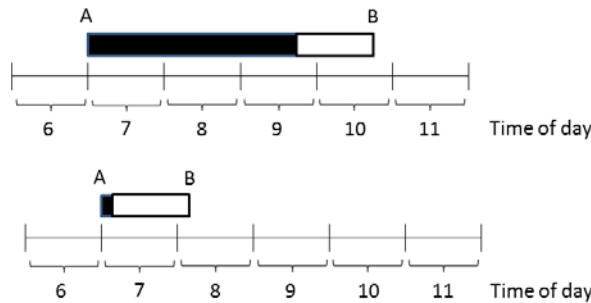


Figure 16: Assignment of trip weight to the time dynamic origin destination matrix. Top: large uncertainty due to a long time span $t_B - t_A$ and short trip time. Bottom: little uncertainty due to short time span.

In order to get a potentially higher temporal resolution for trips, we have further analyzed trips that has a small difference in estimated travel time based on origin and destination location compared to the timestamps of the start and end observations. Due to the large data set it is still possible to get a large number of travels in each OD pair. Figure 17 shows the distribution of start times for all travels (blue) and for one specific OD pair (red). The specified trip definition in combination with this filtering of well-defined start times indicates that there is a peak in travels that start around 12 and 21. However, one should note the strong correlation with the number of events shown in Figure 6, indicating a bias due to bias in location sampling. The above weighting is modified to use the distribution of start time, normalized by the number of events, replacing the uniform probability distributions, and therefore taking into account more information about trip departure times.

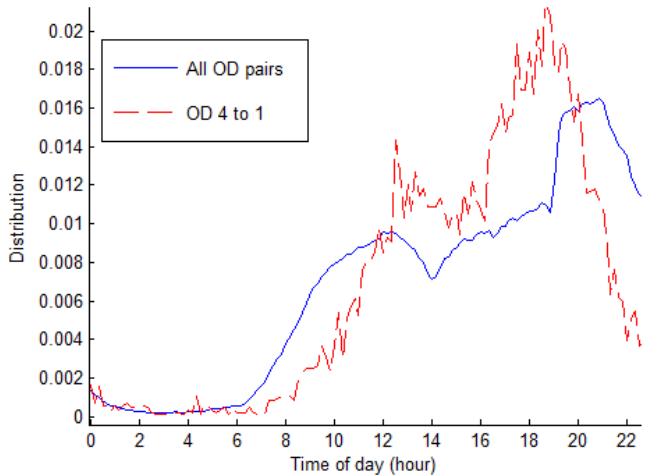


Figure 17: Temporal distribution of trips for OD pair from arrondissement 4 to 1, based on trips with an estimated average speed larger than 10 km/h.

This type of weighted OD matrices has been calculated for both antenna level and an arrondissement level for Senegal. In Figure 18 both antenna level (blue) and the arrondissement level (red) OD is shown for the city of Dakar, filtered for the pairs with largest number of trips. Due to the large number of antenna pairs, it is difficult to see any general trends in the visualization for antennas, however, at least in this example, it is easier to capture in the arrondissement level OD.

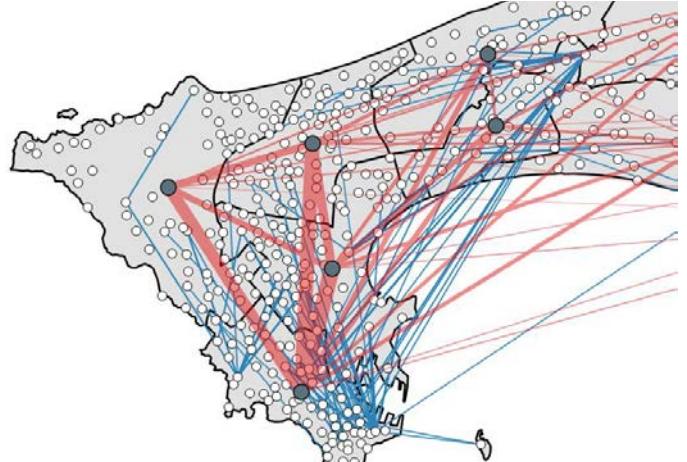


Figure 18: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands for the city of Dakar in red and antenna level demand in blue.

In Figure 19 arrondissement level OD is shown for the whole country. It can be seen that most of the travel demand is located in the Dakar area and along the north border of the country.

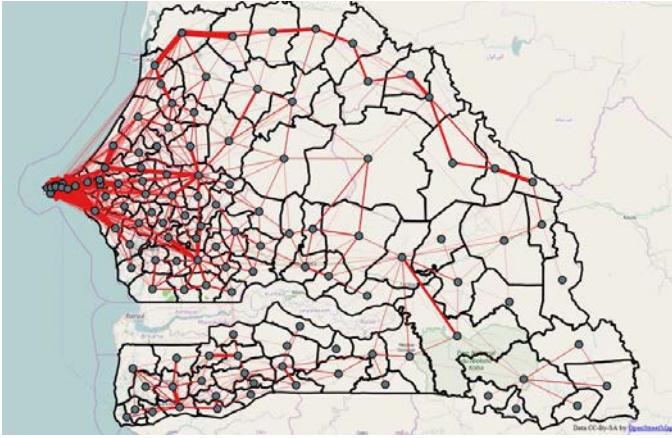


Figure 19: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands.

In Figure 20 the arrondissement level OD is shown for the Dakar region and it can be seen that most of the trips are made within the city, but Dakar also attracts trips to and from the larger cities in the region.

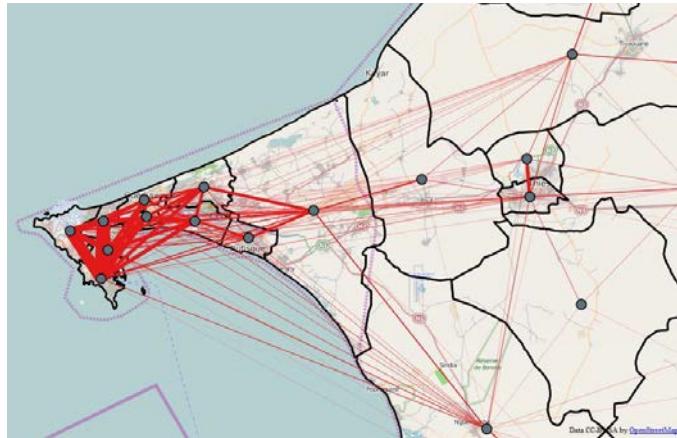


Figure 20: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands for the Dakar area.

In Figure 21 the resulting dynamic OD demand is shown for the arrondissement level and for four one hour time intervals. If we discard the direction of travel, we can see that the travel pattern proportions are relatively similar throughout the day.

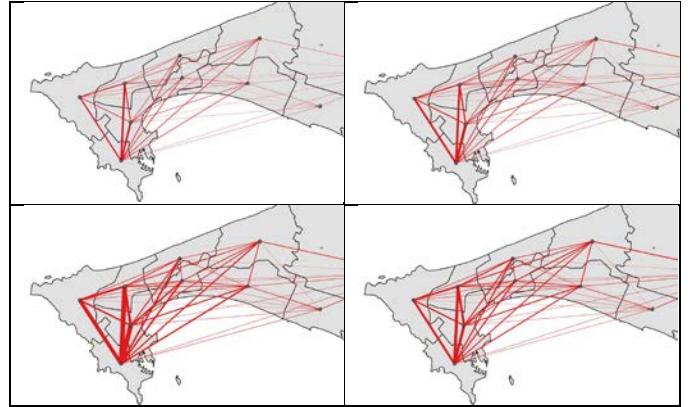


Figure 21: Dynamic OD demand for one hour intervals. Between 6 and 7 in top left, 12 to 13 top right, 18 to 19 bottom left and 22 to 23 bottom right.

VII. ROUTE AND LINK TRAVEL FLOW ESTIMATION

The transport route choice can be studied by filtering out a subset of the trips that are well defined in space. We have filtered out trips with a minimum number of visited cells in the cellpath and assigned them to routes in an algorithm we call “Lazy Voronoi Routing”:

1. Choose a start and end intersection in the road network in the first respectively last cell (based on the distance to the second respectively second last base station)
2. Segmentize the cellpath: Apply the Ramer-Douglas-Peucker algorithm for line-simplification (Douglas et al., 1973) to the line that connects all visited base stations (dashed line in Figure 22) and split the cellpath in each cell, where a point is kept in the simplified line (cell 3 in Figure 22).
3. Select an optimal waypoint in every cell that has a split point (see step 2) according to Figure 23.
4. For each segment of the cellpath calculate the shortest-path between the previous and next waypoint using lowered link costs for all links inside visited cells. The estimated route is the union of all segment routes (blue line in Figure 22).

Using this strategy makes the resulting route likely to follow the visited cells, while, however, not forcing it to enter every single cell, which would cause unrealistic routes due to noise in the data and the inaccurate cell coverage modeled by the Voronoi cells. Adding waypoints at the extreme points of the cellpath ensures that the route always follows the cellpath roughly. Calculating a shortest-path without the waypoint in Figure 22 would, even with lowered link costs for the visited cells as proposed by Leontiadis et al. (2014), lead to a shortcut route directly from cell 1 to cell 6, which is certainly not the route that the user took given the other cells in the cellpath.

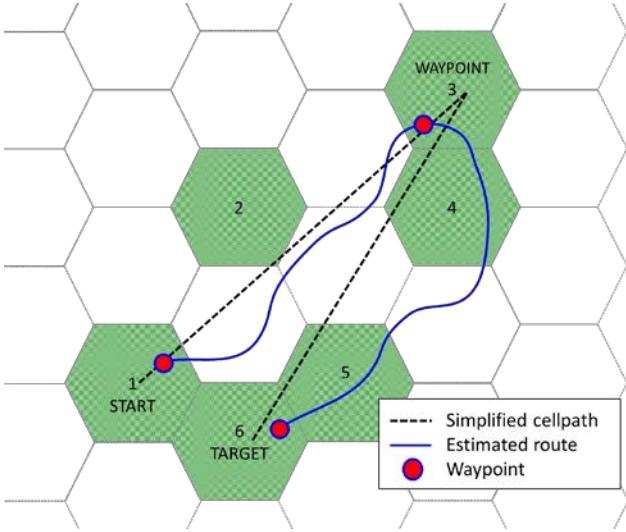


Figure 22: Illustration of route inference using the Lazy Voronoi Routing algorithm for the cellpath (1,2,...,6) using a combination of waypoints and shortest-path calculation with modified link costs for visited cells.

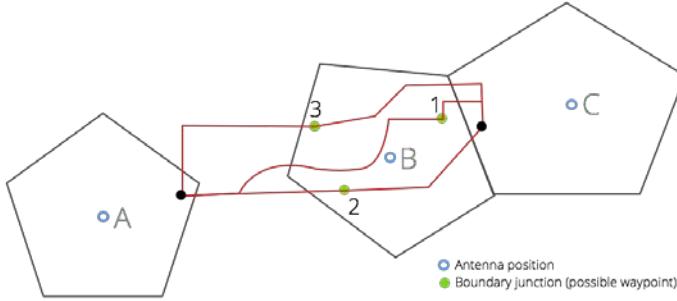


Figure 23: Optimal waypoint selection in cell B when coming from A and heading to C. The waypoint that gives the route (red) with the lowest cost is selected as optimal.

Calculating a route for every distinct cellpath in each OD-pair weighted by the number of occurrences of the same cellpath, gives a route probability that can be used together with the travel demand between the specific antennas to estimate a travel flow distribution on the computed routes for the specific OD pair (see example in Figure 24). By summing all the OD route flows that pass a given link, we can also get an estimate of the flow on that link (see Figure 25).

For route choice estimation based on spatially sparse CDR data most of the trips will have very few or no intermediate cells in the cellpath. This means that the route assignment will rely heavily on information about the road network structure and is often, due to lack of a calibrated model for traffic assignment, simplified with a static shortest-path assignment. During rush-hours the route choice due to congestion can differ significantly from the shortest path. However, the filtering of spatially well-defined trips reduces the proportion of trips that is assigned using the shortest path and can give a better estimate of the route flow proportions compared to the case where all trips are used as input to the estimate. A possible improvement would be to combine this approach with a route choice generated using a classic traffic assignment model (Tettamanti et al., 2012). Note that the algorithm for generating route flow proportions also has

potential for generating choice sets as input to more advanced route choice models (see Bekhor et al. 2006).

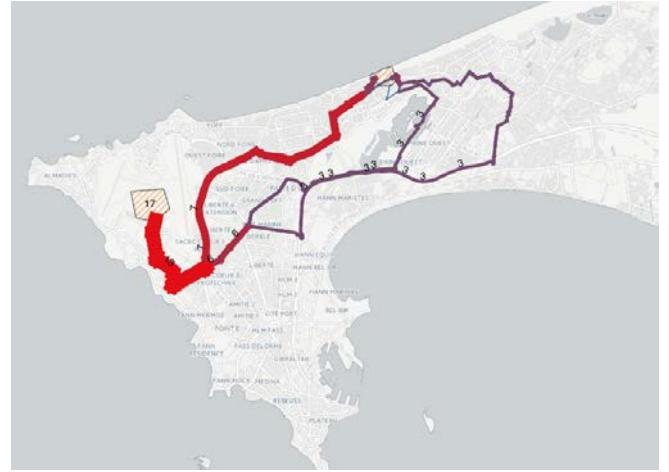


Figure 24: Route travel flow distribution for an example antenna pair (cell 17 to cell 307) in Dakar, based on assigned demand between 18:00-19:00 on a typical weekday according to probability of route choice, calculated using sequences with frequent sampling of antennas.

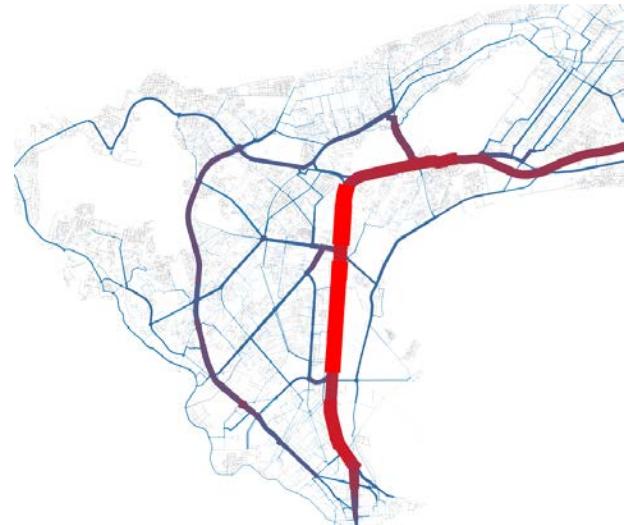


Figure 25: Link travel flows estimated between 8:00 and 9:00 for a part of Dakar (red = high flow). Gray links were not assigned any flow. All traffic was assigned to the road network despite the actual mode of travel.

VIII. DISCUSSION

The travel demand and route flows is a crucial input to infrastructure and transportation planning and is traditionally estimated using census data, travel surveys and models for trip generation and trip distribution together with static traffic models for route assignment. The travel surveys include detailed travel patterns for a small percentage of the travelers and are relatively expensive to collect, hence the travel surveys are typically updated with a very low frequency. Furthermore, the traditional demand estimation problem, based on selected traffic counts is severely underdetermined and the estimates include a lot of uncertainty. However, cellular network signaling data

enables direct observations of trips and to some extent route choice for a large number of travelers in a cost efficient way and this will change our understanding of human mobility and travel demand fundamentally.

In order to be able to build and adapt the transportation infrastructure efficiently, it is crucial to have reasonable estimates of the traffic demand. The travel demand from cellular networks is capturing all types of travel modes, which enables also public transport planning or integrated road and public transport planning, which will be an important area of development in the near future. Since the traditional way of estimating travel demand depends on census data that lacks a temporal component and static models for trip generation, travel demand dynamics has not been studied in great detail. Most of the efforts have been made related to road traffic demand, where dynamic demand estimation has been performed by fusing static demand with sensors that has high temporal resolution, e.g. traffic counts in the road network. Furthermore, road traffic counts are only measuring vehicles and not travelers, which for some applications are less suitable.

The travel flows that are estimated from aggregated movements with higher spatial resolution compared to the travel demand enables an understanding of how the traffic demand is distributed in the transportation network and how it varies over time for different parts of the network. Based on this information it is possible to, for example, make better decisions on where in the network to make sure the infrastructure is maintained properly and where to improve public transport.

The travel time estimates give a possibility to identify parts of the transportation network which has poor infrastructure, limited public transport or a transportation network that is not well adapted to the traffic demand. This can for example be improving public transport service or measures to spread out the travel demand over a longer period of time during the day.

In developing countries, the cellular network is typically much more developed than the traffic and transport sensor infrastructure. However, the traffic situation can be really problematic and the need for well-informed traffic planning decisions is large. Together, this makes cellular network signaling data for traffic planning especially interesting in these countries.

IX. CONCLUSION

In this paper we have demonstrated how to estimate and visualize different types of mobility metrics in both national- and city wide aggregation levels. These mobility metrics can be used to identify different types of bottle necks of the transportation infrastructure, which can be used as input in order to determine where infrastructure investments should be made in order to improve transportation efficiency.

The spatial and temporal resolution that is possible to achieve with cellular network signaling data depends on the cellular network infrastructure, but also on which interface in the cellular network the data is collected from as well as any preprocessing that is made on the data. CDR data based on SMS and call activities, which is the most commonly used type of data, typically suffers from a relatively poor temporal resolution,

which needs to be compensated for in the processing pipeline of travel demand and travel pattern analysis.

In the travel demand estimation from cellular network signaling data we get direct observations of combined trip generation, trip distribution and, to some extent, route choice for a sample of the population. However, the suggested concept of decoupling the travel demand estimation process from more detailed spatial and temporal analysis gives the possibility to design trip extraction algorithms that capture a larger part of actual trips that are made. Furthermore, dedicated algorithms for temporal distribution of demand, travel times as well as route choice can be designed based on a subset of trips with spatiotemporal characteristics suitable for the specific task. New dedicated algorithms for temporal distribution of demand, travel time estimation and route choice are implemented in this paper, but the principle holds also for mode choice. The paper demonstrates the importance of the algorithms for trip extraction, temporal demand distribution, travel time estimation and route choice for accurate travel pattern analysis, using a large scale CDR data set.

Several of the mobility metrics that are estimated in this paper, such as dynamic travel demand and route choice, are of special interest to the transportation community since traditional sensors cannot be used to observe them. Instead, extensive research has been performed to use models to estimate and predict these metrics. However, the models rely on basic assumptions that may not always be valid and can also contain a very large set of model parameters that are difficult to calibrate. For example, many route choice models rely on the assumption that each user has perfect knowledge of the traffic situation and requires volume-delay functions for each link in the network.

Cellular network signaling data will change how we understand travel demand dynamics and human mobility in general. In developing countries, the cellular network is typically much more developed than the traffic and transport sensor infrastructure, which will make it an extremely valuable source of information for strategic, tactic and possibly also for operational planning of transportation networks. Efficient algorithms and models that utilize the characteristics of the underlying cellular network data will have a large potential in improving transportation and environmental quality in many large cities in the world.

ACKNOWLEDGEMENTS

This work was supported by the Swedish Governmental Agency for Innovation Systems (VINNOVA).

REFERENCES

- L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data", 2015, Transportation Research Part C: Emerging Technologies, pp. 240 – 250.
- N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures", 2012, Information Visualization, pp. 3-24.
- V. Angelakis, D. Gundlegård, C. Rydberg, B. Rajna, K. Vrotsou, R. Carlsson, J. Forgeat, T. H. Hu, E. L. Liu, S. Moritz, S. Zhao, and Y. Zheng, "Mobility modeling for transport efficiency: Analysis of travel characteristics based on mobile phone data", 2013. In: Netmob 2013: Mobile phone data for development.

- R. A. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: Using cellular network data for urban planning", 2011, IEEE Pervasive Computing, pp. 18-26.
- R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data", 2013, Communications of the ACM, pp. 74- 82.
- S. Bekhor, M. E. Ben-Akiva, M. S. Ramming, "Evaluation of choice set generation algorithms for route choice models," Annals of Operations Research, volume 144, 2006, pp. 235-247.
- M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M. Sbodio, "AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data", 2013, pp. 663-666.
- V.D. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, S. Zbigniew, "Mobile Phone Data for Development, Analysis of mobile phone datasets for the development of Ivory Coast (D4D Book)", 2013, <http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf>
- N. Cáceres, J. P. Wideberg, and F. G. Benítez, "Deriving origin-destination data from a mobile phone network", 2007, IET Intelligent Transport Systems, pp. 15-26.
- F. Calabrese, M. Diao, G. D. Lorenzo, J. F. Jr., and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example", 2013, Transportation Research Part C: Emerging Technologies, pp. 301-313.
- F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating Origin-Destination Flows Using Mobile Phone Location Data", 2011, IEEE Pervasive Computing, pp. 36-44.
- M. Dash, H. L. Nguyen, C. Hong, G. E. Yap, M. N. Nguyen, X. Li, S. P. Krishnaswamy, J. Decraene, S. Antonatos, Y. Wang, D. T. Anh, and A. Shi-Nash, "Home and Work Place Prediction for Urban Planning Using Mobile Network Data", 2014. In: Mobile Data Management (MDM), 2014 IEEE 15th International Conference on, vol. 2, pp. 37-42.
- M. G. Demissie, G. H. de Almeida Correia, and C. Bento, "Intelligent road traffic status detection system through cellular networks handover information: An exploratory study", 2013, Transportation Research Part C, pp. 76-88.
- J. Doyle, P. Hung, D. Kelly, S. McLoone, and R. Farrell, "Utilising mobile phone billing records for travel mode discovery", 2011. In: ISSC 2011.
- P. Fiadino, D. Valerio, F. Ricciato, and K. Hummel, "Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data", 2012, pp. 66-80.
- M. Fillekes, "Reconstructing Trajectories from Sparse Call Detail Records", Master's thesis, 2014, University of Tartu.
- D. Gundlegård and J. M. Karlsson, "Route classification in travel time estimation based on cellular network signaling", 2009. In: Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on, pp. 1-6.
- D. Gundlegård, C. Rydberg, J. Barcelo, N. Dokoochaki, O. Görnerup, and A. Hess, "Travel Demand Analysis with Differentially Private Releases", 2015. Netmob 2015: Mobile phone data for development.
- D. Gundlegård and J. M. Karlsson, "Generating Road Traffic Information from Cellular Networks - New Possibilities in UMTS", 2006. In: ITS Telecommunications Proceedings, 2006 6th International Conference on, pp. 1128-1133.
- S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data", 2014, Computer Networks, pp. 296 – 307.
- M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin-destination matrices using mobile phone call data", 2014, Transportation Research Part C: Emerging Technologies, pp. 63-74.
- S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data", 2011, pp. 133-151.
- S. Jiang, J. Ferreira JR, M.C. González, "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data:A Case Study of Singapore", UrbComp'15, August 10, 2015, Sydney, Australia.
- A. N. Larijani, A.-M. Olteanu-Raimond, J. Perret, M. Brédif, and C. Zieliński, "Investigating the Mobile Phone Data to Estimate the Origin Destination Flow and Analysis; Case Study: Paris Region", 2015, Transportation Research Procedia, pp. 64-78.
- I. Leontiadis, A. Lima, H. Kwak, R. Stanojevic, D. Wetherall, and K. Papagiannaki, "From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data", 2014. In: Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, pp. 121-132.
- F. Liu, D. Janssens, J. Cui, G. Wets, "Building workers' travel demand models based on mobile phone data", 2015, <http://hdl.handle.net/1942/18883>
- W. Ming-Heng, S. D. Schrock, N. V. Broek, and T. Mulinazzi, "Estimating dynamic origin-destination data and travel demand using cell phone network data", 2013, International Journal of Intelligent Transportation Systems Research, pp. 76 – 86.
- Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Zieliński, and V. D. Blondel, "D4D-Senegal: The Second Mobile Phone Data for Development Challenge", 2014, CoRR
- S. A. Shad and E. Chen, "Precise Location Acquisition of Mobility Data Using Cell-id", 2012, CoRR.
- T. Sohn, A. Varshavsky, A. LaMarca, M. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. Griswold, and E. de Lara, "Mobility Detection Using Everyday GSM Traces", 2006, pp. 212-224.
- J. Steenbruggen, M. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities", 2013, GeoJournal, pp. 223.
- T. Tettamanti, H. Demeter, and I. Varga, "Route choice estimation based on cellular signaling data", 2012, Acta Polytechnica Hungarica, pp. 207-220.
- J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "The path most traveled: Travel demand estimation using big data resources", 2015, Transportation Research Part C: Emerging Technologies, pp. 162 – 177.
- H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records", 2010. In: Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, pp. 318-323.
- P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, "Understanding road usage patterns in urban areas", 2012, Scientific reports.
- A. G. Wilson, "A statistical theory of spatial distribution models", 1967, Transportation research, pp. 253-269.
- C. Wu, J. Thai, S. Yadlowsky, A. Pozdnoukhov, and A. Bayen, "Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization", 2015, Transportation Research Part C: Emerging Technologies.