Travel demand estimation and network assignment based on cellular network data

David Gundlegård, Clas Rydergren, Nils Breyer and Botond Rajna

Journal 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

Travel demand estimation and network assignment based on cellular network data

David Gundlegård*, Clas Rydergren, 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. Berlingerio 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).

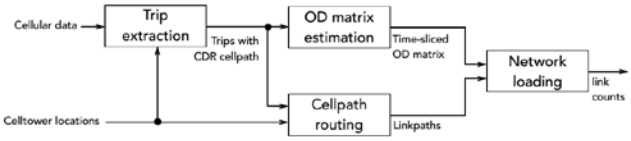

Figure 2: Overview of the processing pipeline.

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.

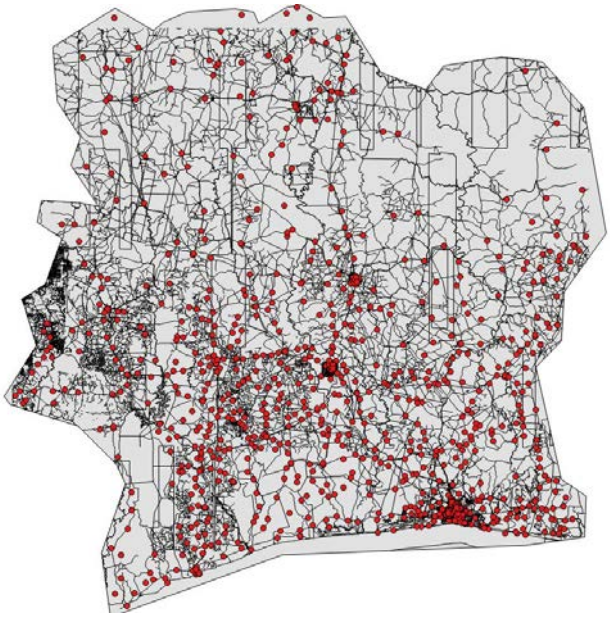

Figure 3: Antenna distribution (red dots) and road network from Open Street Map for Côte d'Ivoire.

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.

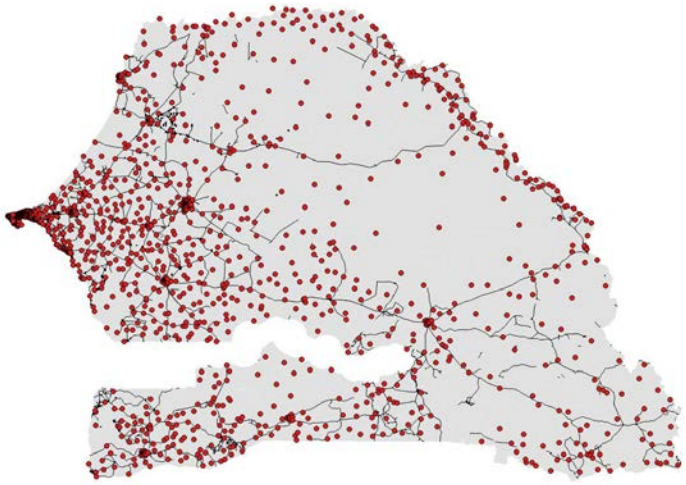

Figure 4: Antenna distribution (red dots) and road network from Open Street Map for Senegal.

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	Massa-chusetts (USA)	Trip extractio, 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	Massa-chusetts (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 (Bangla-desh)	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 temporarily 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.

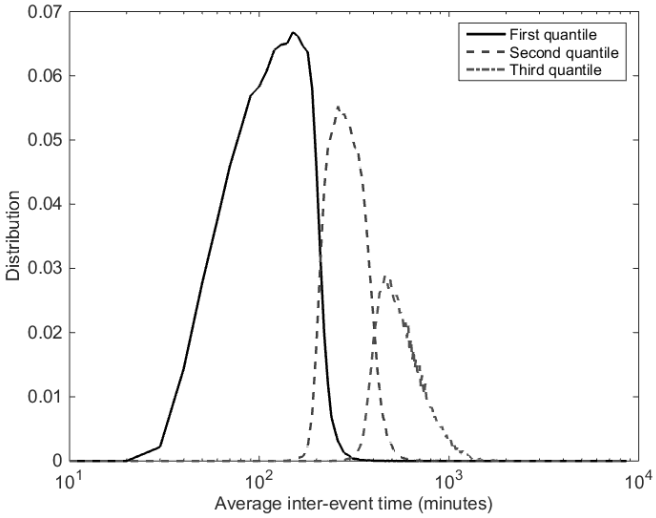

Figure 5: Distribution of average inter-event times for three quantiles for the Senegal dataset.

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.

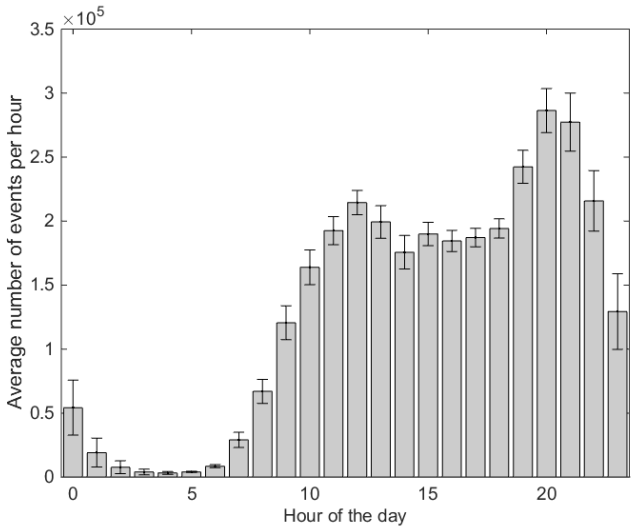

Figure 6: Average number of CDR data events per hour for the first two week period of data in the Senegal dataset. The error bars show one standard deviation.

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 south-east 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.

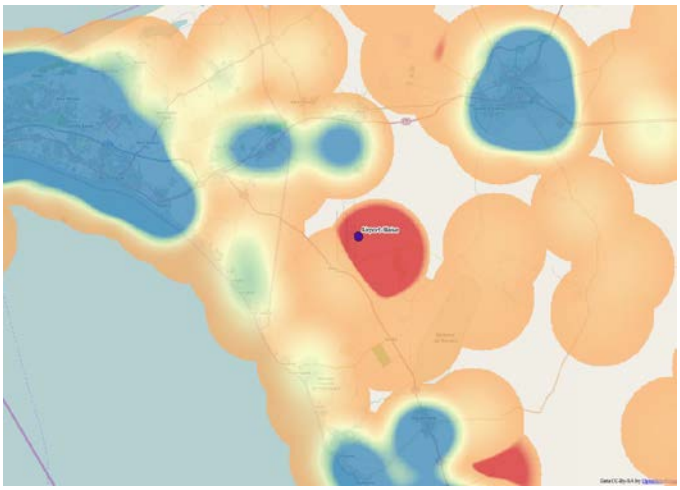

Figure 7: Heat map of difference of number of home locations and number of work locations. Blue indicates potential residential areas and red indicates areas with large daytime activity. The red area in the middle of the figure is an airport.

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, Berlingerio 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 != homebase and d(p_uak,homebase) < d_max and
                                  d(p_uak,homebase) > d_min)
19      trip_active = true
20      origin = homebase
21    end
22    if(p_uak != homebase and d(p_uak,homebase) > d_max)
23      trip_active = true
24      origin = p_uak
25    end
26    if(p_uak == workbase and d(p_uak,homebase) > d_max)
27      trip_active = true
28      origin = homebase
29      destination = workbase
30    end
31  else
32    if(p_uak != previous_trip_start(trip_set) and
          d(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_ua(k+1) exists)
40      if(p_uak == p_ua(k-1))
41        destination = p_uak
42      end
43    else
44      if(d(p_uak,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).

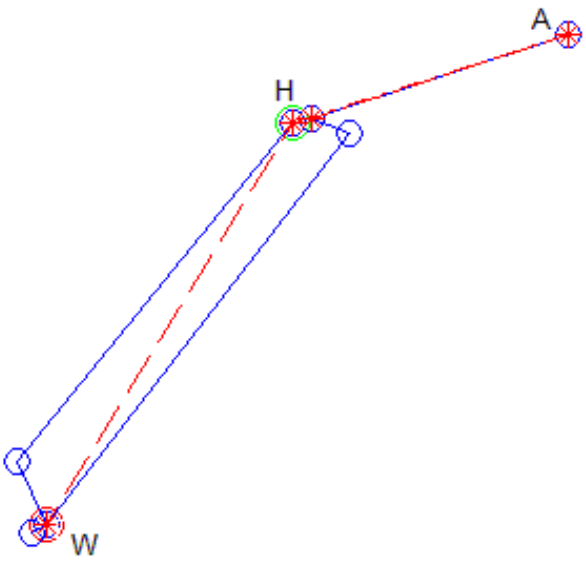

Figure 8: Example of trips generated for one user one day. Blue lines are trajectories and red lines are the generated origin-destination trips.

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).

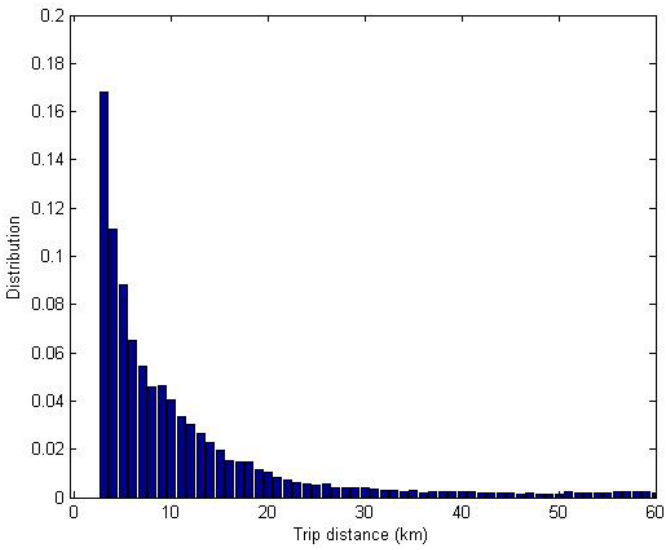

Figure 9: Trip distance distribution for the generated trips.

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'Ivorie 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.

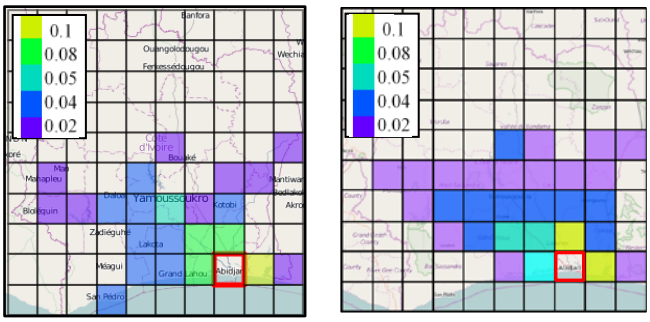

Figure 10: Left: Proportions of long distance trips that started between 9 am and 10 am and terminated in Abidjan. Right: Traffic demand proportions estimated by a gravity model based on population density and distance between zones terminated in Abidjan.

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.

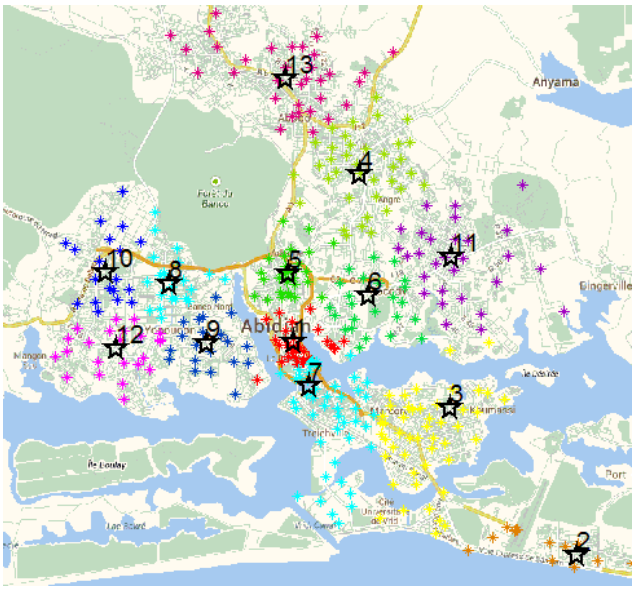

Figure 11: 13 clusters in Abidjan created using K-means clustering of antenna locations.

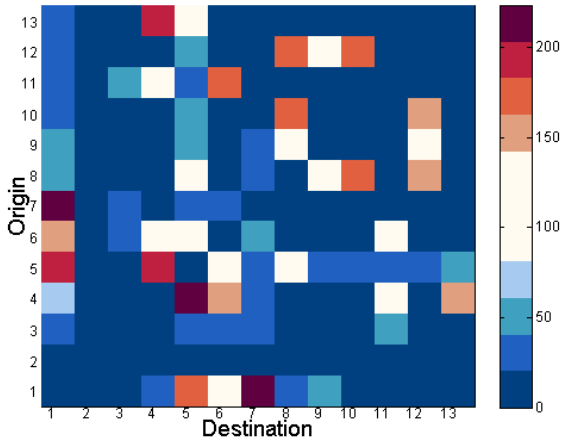

Figure 12: OD matrix proportion visualization for the different zones shown in Figure 11 based on t-OD trip extraction, i.e. each pair of antenna locations are considered as a trip.

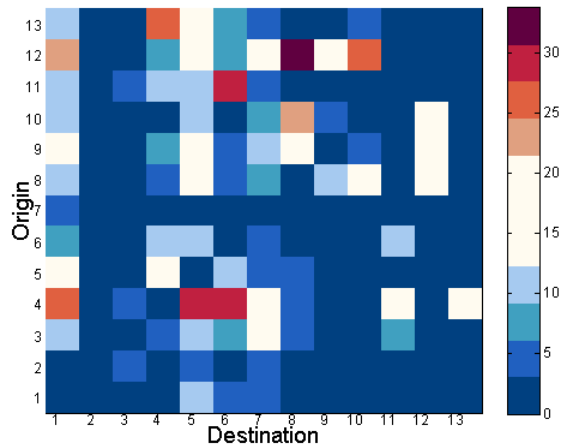

Figure 13: OD matrix proportion visualization for the different zones shown in Figure 11 based on static trip extraction, i.e. a trip occurs when a user has changed zone between predefined time periods.

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.

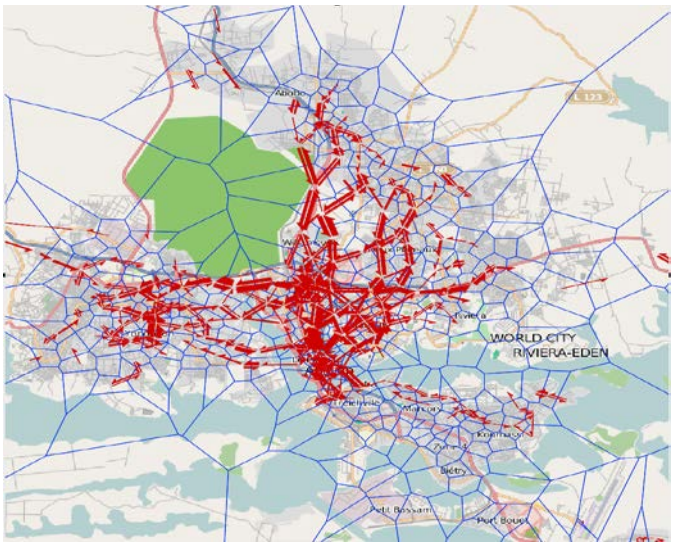

Figure 14: Cell travel flows between antennas represented with Voronoi polygons. The flow is aggregated for a two-hour period in Abidjan.

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.

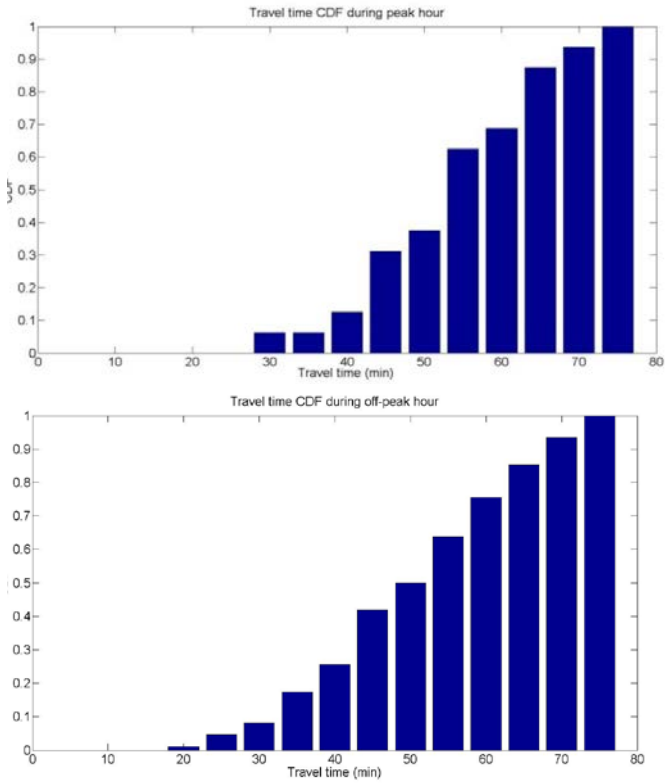

Figure 15: Top: Peak hour travel time measurement CDF between Abidjan city and Abidjan airport grouped in intervals of 5 minutes Bottom: Off-peak hour travel time measurement CDF between Abidjan city and Abidjan airport grouped in intervals of 5 minutes. Travel time measurements larger than 75 minutes 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

$$w_i = \frac{\min\{t_B - t_{AB}, t_i + 1\} - \max\{t_A, t_i\}}{(t_B - t_{AB}) - t_A}, \lfloor t_A \rfloor \leq t_i \leq \lceil t_B \rceil$$

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 $\lfloor t \rfloor$ and $\lceil t \rceil$ denote the rounding down and up to the nearest hour, respectively.

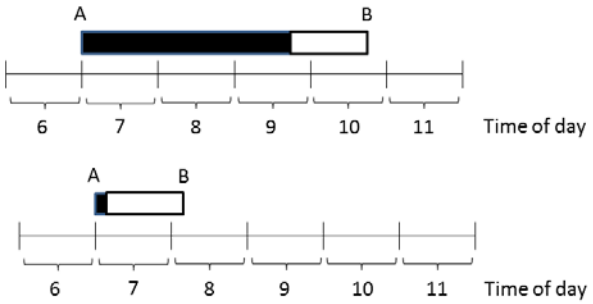

Figure 16: Assignment of trip weight to the time dynamic origin destination matrix. Top: large uncertainty due to a long time span $t_B - t_A$and short trip time. Bottom: little uncertainty due to short time span.

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 par. 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.

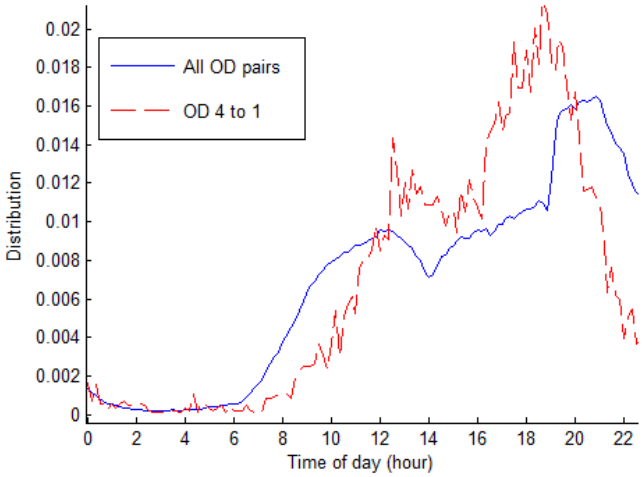

Figure 17: Temporal distribution of trips for OD pair from arrondissement 4 to 1, based on trips with an estimated average speed larger than 10 km/h.

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.

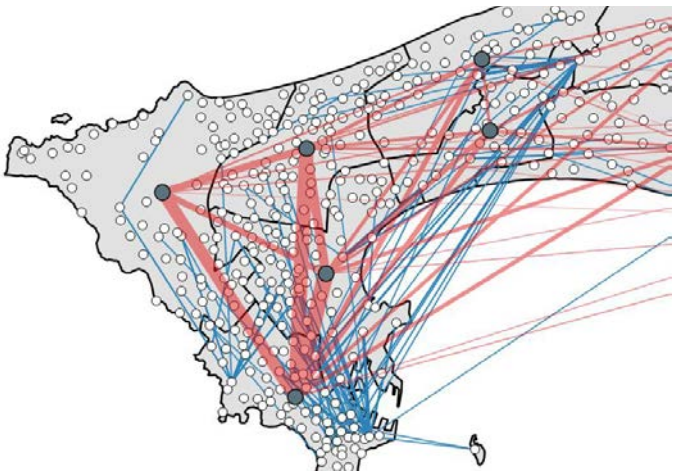

Figure 18: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands for the city of Dakar in red and antenna level demand in blue.

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.

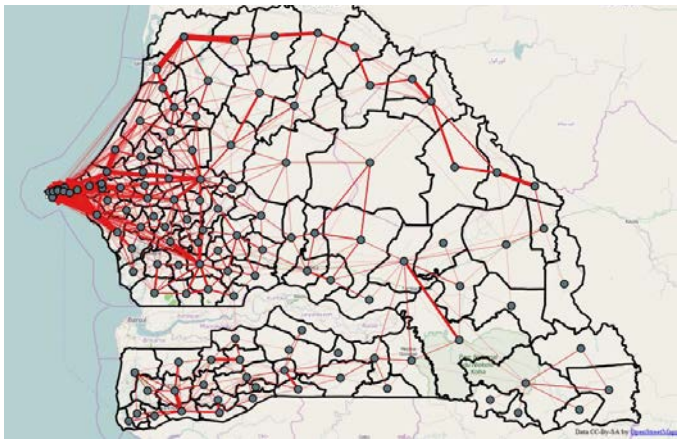

Figure 19: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands.

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.

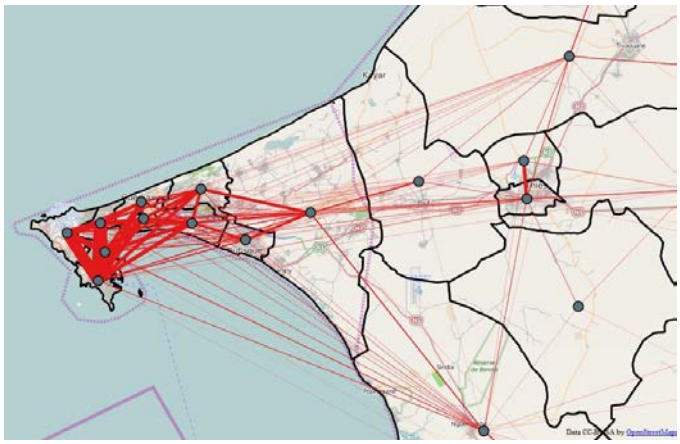

Figure 20: OD demand on the level of arrondissement for the first two week time period filtered on large OD demands for the Dakar area.

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.

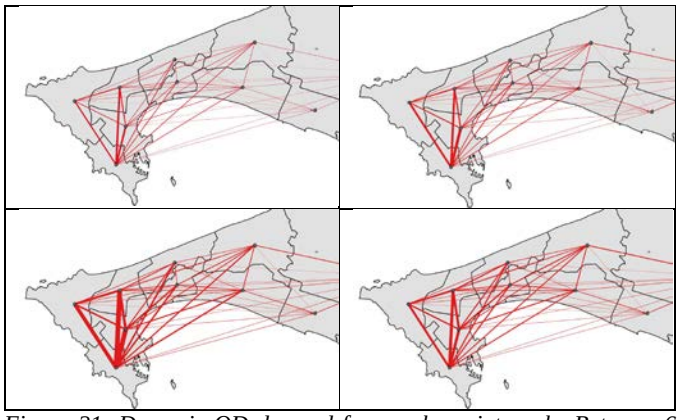

Figure 21: Dynamic OD demand for one hour intervals. Between 6 and 7 in top left, 12 to 13 top right, 18 to 19 bottom left and 22 to 23 bottom right.

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.

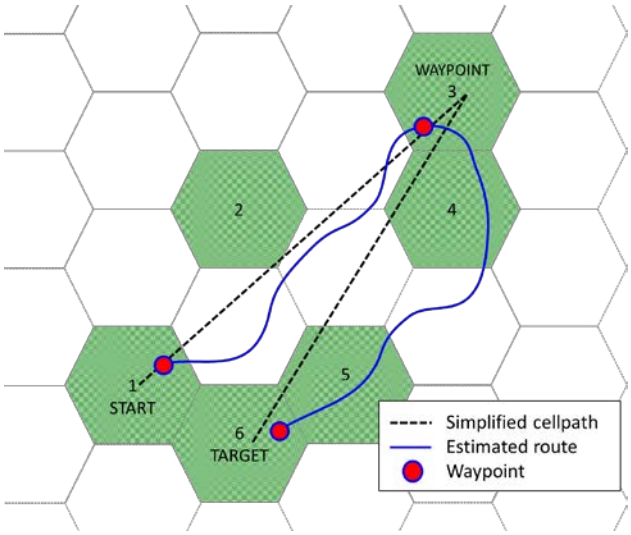

Figure 22: Illustration of route inference using the Lazy Voronoi Routing algorithm for the cellpath (1,2,..,6) using a combination of waypoints and shortest-path calculation with modified link costs for visited cells.

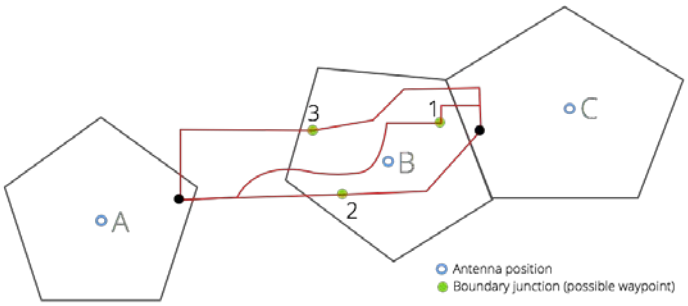

Figure 23: Optimal waypoint selection in cell B when coming from A and heading to C. The waypoint that gives the route (red) with the lowest cost is selected as optimal.

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).

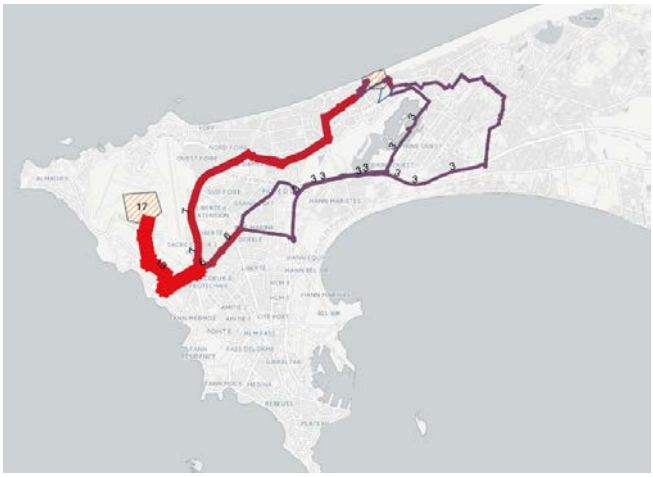

Figure 24: Route travel flow distribution for an example antenna pair (cell 17 to cell 307) in Dakar, based on assigned demand between 18:00-19:00 on a typical weekday according to probability of route choice, calculated using sequences with frequent sampling of antennas.

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. Rydergren, 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. Caceres, 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. Caceres, J. P. Wideberg, and F. G. Benitez, "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. Rydergren, J. Barcelo, N. Dokoohaki, O. Görnerup, and A. Hess, "Travel Demand Analysis with Differentially Private Releases", 2015. Netmob 2015: Mobile phone data for development.

D. Gundlegard 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. Ziemlicki, "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. Ziemlicki, 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.