Ecole Polytechnique Fédérale de Lausanne

Master's Thesis

Mode choice models with smartphone data

Student:
Michael Friederich

Supervisors:
Michel Bierlaire
Anna Fernández Antolín
Evanthia Kazagli
Marija Nikolic

Master's Degree thesis

in the

Transp-OR laboratory

June 2014

Abstract

Mode choice models with smartphone data

by Michael Friederich

Nowadays, the widespread use of smartphones and their ability to collect longitudinal data without increasing the burden of the traveler enable better monitoring, understanding and analysis of travel behavior. In this thesis, we make as a first attempt a mode choice model with smartphone data when data collection is passive. Our research consists in identifying and solving arising issues, due to the nature of the data, in order to derive a dataset suitable for mode choice analysis. The key components of the proposed methodology concern the detection of trips, activities and identification of the trip purpose based on smarthphone data, and common issues to mode choice modeling, such as the determination of the chosen mode and missing attributes of the unchosen alternative, are addressed as well. The derived dataset is further enriched by complementary datasets including socio-economic and meteorological information. Estimation results are consistent with intuition and relevant works from the literature, showing the feasibility and potential of using smartphones for mode choice analysis. A smartphone dataset collected in Lausanne is used to illustrate the issues and estimate the model.

Acknowledgements

First I would like to thank Evanthia Kazagli, Marija Nikolic and Anna Fernández Antolín for their advices, support and the time spent together with me for the project. Through patience and perseverance, our team manage to find the resources to overcome the difficulties of this beautiful but challenging master's thesis.

Furthermore, I am thankful to my supervisor Professor Bierlaire who believed in my motivation and accepted me within his team and prestigious laboratory. I would also like to thank all the people of the Transp-or laboratory for the beautiful working environment.

Finally I am grateful to all my classmates of the GCB31 where we studied every day during 5 months sharing the same fridge and coffee machine in a serious and relax atmosphere. I leave EPFL with great memories!

Contents

List of Figures

List of Tables

Abbreviations

BMI	**B**ody **M**ass **I**ndex
CATI	**C**omputer **A**ssited **T**elephone **I**nterviews
CASI	**C**omputer **A**ssited **S**elf **I**nterviews
EPFL	**É**cole **P**olytechnique **F**édérale de **L**ausanne
GIS	**G**eographic **I**nformation **S**ystem
GPS	**G**lobal **P**ositioning **S**ystem
GSM	**G**lobal **S**ystem for **M**obile communications
MAC	**M**edia **C**ontrol **A**ccess
NRC	**N**okia **R**esearch **C**enter
OFS	**O**ffice **F**édérale de la **S**tatistique
PAPI	**P**aper **A**nd **P**encil **I**nterview
PDA	**P**ersonal **D**igital **A**ssistant
SQL	**S**tructured **Q**uerry **L**anguage
UML	**U**nified **M**odeling **L**anguage
Wi-Fi	**Wi**reless - **Fi**delity
WLAN	**W**ireless **L**ocal **A**rea **N**etwork

Chapter 1

Introduction

1.1 Motivation

Understanding the travel behavior of people is essential to many areas from the field of transportation that is subject to the continuously increasing demand and saturation of infrastructures. For example, many researchers and policy makers are concerned with the issues regarding the mode choice decisions of the population, given that the changes in transit routes and schedules on ridership affect the individual travelers' mode choice and consequently the revenues and traffic congestion. The significant fare increase, for instance, may lead to a decrease of transit revenues which is not the desired outcome.

In such cases a simple application of a particular policy, without the previous study and understanding of the concrete problem might lead to some very costly trial and error solutions. One solution is therefore a comprehensive analysis of the complexity of the problem, based on the detailed data, followed by the development of mathematical modeling approaches that would serve as useful forecasting tools.

The proliferation of smartphones in the future, able to collect longitudinal data over years and to provide personalized travel information will have significant impacts on travel behavior understanding and give the opportunity to upgrade significantly transportation operations. With the progress on band-width connectivity, we will likely have instantaneous data collection over the entire population (Vautin and Walker, 2011).

Nowadays, the ability of smartphones to provide longitudinal data have already revolutionized the acquisition of information about traveler's behavior and made possible to stretch out data collection campaigns over several months. With such technologies, high resolution observations are passively collected while the participants' burden is reduced.

In this context, we have displayed true interest in exploring the prospects of using data from these powerful devices for travel behavior analysis. Furthermore, the lack of research exploring the challenges of these data for mode choice analysis tilted the balance toward mode choice analysis.

1.2 Objective

The purpose of this thesis is to investigate the feasibility of using smartphone data for mode choice analysis.

1. The core part of this thesis consists in the identification of a catalog of issues due to the specific nature of this data and the development of strategies to address them. These issues are identified in accordance with discrete choice modeling framework.

2. The development of a mathematical model in the context of discrete choice analysis that would explain mode choice behavior of Lausanne inhabitants going from home to work together with the estimation of the model and analysis by means of statistical tests.

The Nokia Lausanne data collection campaign (Nokia dataset) is utilized to illustrate the issues and serves as an input for the mode choice model.

1.3 Outline

The remainder of the master's thesis is composed of 6 chapters. The second chapter reviews the different data collection techniques with their strengths and limitations and provides a short overview of discrete choice models. Chapter 3 presents the available dataset taken from different sources: the smartphone data (e.g Nokia dataset), the results of the map-matching algorithm developed by Chen (2013) and MeteoSwiss archives

of the weather in 2010 and 2011. Chapter 4 identifies the issues to be solved to build a discrete choice model. In the same chapter, a list of processing solutions is proposed to address these issues. Chapter 5 finally proposes a mathematical model explaining mode choice behavior for inhabitants of Lausanne area and an analysis of the estimation results. The last chapter summarizes our research and identifies future directions for further analysis.

Chapter 2

Literature review

The chapter reviews the different data collection technologies used until now for travel behavior models. We present each technology with its strengths and limitations in section 2.1.

Section 2.2 goes through most commonly used discrete choice models and their applications.

2.1 Data collection techniques

In the late 90's, the advent of low cost sensors composing GPS [1] loggers and many mobile devices, has introduced a new alternative in the field of data collection as shown in Fig. 2.1. The following section discusses the reasons why GPS technologies has been supplementing conventional surveys and even more recently replacing them.

2.1.1 Conventional surveys

Household travel surveys have been for a long time the conventional form of collecting data in order to study travel behavior. The improvements of these surveying methods over time is considerable (Fig. 2.1): The earlier paper-and-pencil interview (PAPI) methods consisted in mail-out and mail-back surveys complemented by interviews in-home

[1] Global Positioning System (GPS): "space-based satellite navigation system that provides location and time information, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites" (Wikipedia)

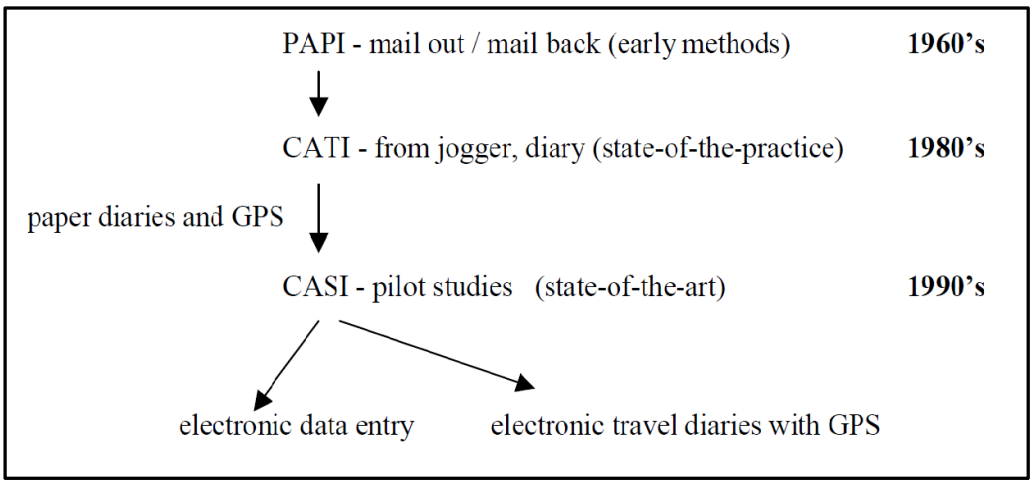

FIGURE 2.1: Evolution of travel survey instruments and data collection methods (Wolf et al., 2001)

of the participants. In the 80's, with the emergence of computers, mail-back retrieval of the data is replaced by computer-assisted telephone interviews (CATI) where the participant reports his answers by phone to a person in charge that directly stores them into a computer. In the 90's, (CATI) methods are replaced by computer assisted-self interview (CASI) where the user directly stores his responses into his personal computer (desktop, laptop) (Ortuzar and Willumsen, 1994) and (Wolf et al., 2001). Recently, web-based diary accessible from the computer or the smartphone have made their apparition (Vlassenroot et al., 2014).

Over time, researchers have argued the limits of these data collection processes. One is that many resources are required to collect even small amounts of data. Indeed, the survey, relying on paper or phone surveys, is time and cost consuming which is significantly restricting the size of the samples and the duration of the campaigns (Vautin and Walker, 2011). Even current web-based trip diaries, where respondents self-report their trips directly in the web, has not lower that much the burden because one has still to remember all the actions done earlier in the day .

A common bias in conventional surveys is the lack of details of the reported trips because of memory recall (e.g start or end time, duration, locations, distances and mode). It also happens that people do not have enough time to answer correctly the questions. Previous studies have shown that short trips, short activity stops and non home-biased trips are usually under reported (Yalamanchili et al., 1999). Surveys also suffer from

poor accuracy (Murakami et al., 1997), (Murakami and Wagner, 1999), (Stopher and Collins, 2005). It also happens that time and cost attributes have to be imputed with travel time data sources (e.g. Google maps, CFF [2], etc.) because of the poor reporting of the respondents (Atasoy et al., 2011). Furthermore, car users tend to underestimate travel time whereas public transport customers use to overestimate it (Dizaji, 2012).

In addition, understanding the regularity and the variability of individual travel behavior *over time* is a key issue in transportation behavioral research. And conventional surveys have so far restricted in-depth studies on long-term mobility patterns of individuals because of the limited longitudinal data available (Lasky et al., 2006). Main reasons of this deficiency, besides the high cost and effort required to extend the campaign, are due to the small amount of persons that would participate for long-term travel behavior studies which is consistent with the fatigue effects reported when the survey is longitudinal (Doherty et al., 2001), (Schönfelder et al., 2002). Actually Doherty et al. (2001) report that survey participation rates substantially decrease as the effort required to complete the survey increases. It is therefore rare to extend the length of the study over a few days.

As shown in Fig. 2.1 , GPS in the late 90's, that can automatically collect travel data and extract trip characteristics, permits to resolve some of the issues aforementioned and logically becomes a serious alternative to conventional surveys.

2.1.2 GPS technologies

Two sorts of GPS technologies emerge in the late 90's / beginning of the century: one is the in-vehicle sensor that captures the movement at the vehicle's level (Fig. 2.2a). This technology can be installed in all kind of vehicle (car, bus, metro, tram). The other one is the wearable GPS loggers that records the position and speed of the participant wherever he is.

Advantages of GPS data collection are numerous such as Schönfelder et al. (2002) expose:

- The reduction (in case of active monitoring) or even elimination (in case of passive monitoring) of respondents' burden which allow surveys to be very long with lower fatigue effects.

[2] Chemins de Fer Fédéraux (CFF)

(A) In-vehicle GPS

(B) 2003 Wearable GPS

FIGURE 2.2: GPS technologies

- Path's choices of the user is available which make GPS data very valuable for route choice models.
- The high resolution of spatial information with accuracies from 1 to 5 meters for GPS loggers. Plus the velocity of the records that are associated with a precise time stamp (Wolf et al., 2001).
- Data is generated and stored in digital format which allows direct processing or analysis.

Among the limitations of the in-vehicle sensor, one is that it can't be used for all travel modes, for example walking or biking is omitted. Hence, travel data collected from this devices can't be used for precise mode choice models. In addition to this important drawback, this traffic sensor captures travel behavior of a single vehicle rather than an individual which needs to be considered in disaggregated models where individuals are the basic units of observation (Dizaji, 2012). Nevertheless, in-vehicle systems have usually shown better performances than loggers because they have fixed positions inside the vehicle (for example the roof) and are provided with the vehicle power source (Wolf, 2004).

The wearable logger, including all the advantages aforementioned and specifically designed for transportation research, is logically suitable for more applications. Both passive and active GPS logging technologies can be designed for vehicle-based logging or for person-based logging, which enables capturing travel behavior on all travel modes. However, the use of this dedicated GPS loggers, able to collect highly accurate individual travel data at trip level, also brings about several challenges related to data collection

(Casas and Arce, 1999), (Draijer et al., 2000), (Battelle and Administration, 2000), (Wolf et al., 2001), (Casello et al., 2011):

- Problems related to the the distribution of the units to the survey participants.
- Problems with the transmission of the data from the PDA [3]/logger which requires usually a connection to a computer with internet. This constraint makes the data transmission non-automatic and therefore the data collection is not in real time. Battelle and Administration (2000) report failures of the cable that links the GPS to the data logger.
- The burden for the user related to always carrying the device with him for all trips. It is for example difficult to carry the device when biking.
- The importance to turn on/off the device at the right moment (e.g. issue with the real start or end of the trip).
- The time required for the GPS to locate the first signal sent by the user, commonly called "cold-start", which may be more than 5 minutes and that results in missing the beginning of the trip.
- Current loss of GPS signals due to improper GPS antenna on the respondent body and signal blockage inside vehicles such as buses and software bugs.

Furthermore, one of the core question that raises when discussing of passive automatically collected GPS data is whether it is possible to post-process it to obtain missing trip attributes such as the trip purpose, the stops/activities in between trips and a precise identification of trips' origin/destination which can turn out to be quite challenging knowing the aforementioned issues. Hence, algorithms have been discussed in the literature to transform position data to travel characteristics (Schüssler, 2010).

2.1.3 Smartphones

In recent years, the widespread adoption of smartphone technologies that are equipped with GPS sensors has provided the opportunity to resolve many limitations associated with traditional surveys and GPS wearable loggers. Smartphone data collection can be active or passive. In the first case, information on the trips is asked to the participants

[3] Personal Digital Assistant (PDA): "mobile device that functions as a personal information manager"

via an app. [4] interface (Fig. 2.3) (Vlassenroot et al., 2014), (Pereira et al., 2013). In case of passive monitoring, travel data is recorded without the need for the users to write the trip's information in a travel diary or app. interface. This advantage is essential in order to capture accurate longitudinal data without burdening the participant.

When GPS data collection is required, a software/app. is generally designed which is capable of recording the trip characteristics (position, time, speed, acceleration, etc.) of individuals and decides which sensor needs to be activated (e.g. GPS, Wi-Fi, GSM, Bluetooth [5]) depending on the state of the participant (e.g. stationary, moving)(Doyle et al., 2013),(Kiukkonen et al., 2010). This concept is called "phased sampling". The essential feature of the software is the battery management of the device that remains one of the most troublesome drawback of smartphones nowadays. Indeed, it is very important that the participant is not bothered by GPS data collection process that is high power consuming. For this reason, the software "allocates" time when the GPS can be activated for data collection which results in a sparse data. In addition to these "willful" losses, issues related to GPS signal losses, inside buses or trains when the user is not clost to a window, still remains. Depending on the objective of the research, it might be preferable to omit GPS recording and collect with other less energy consuming sensors of the smartphone. Eagle and Pentland (2006) demonstrates that Bluetooth records are a good source of data to measure information access, or to recognize social connections (activities, relationship) between different users.

Moreover, smartphone data is less accurate than devoted GPS devices. Watzdorf and Michahelles (2010) estimate the accuracy location of the GPS from 5 to 10 meters whereas the smartphone integrated GPS accuracy can deviate of several hundreds of meters. Wi-Fi data that relies on WLAN access points [6] location has a good accuracy from 30 to 50 meters, only it requires the availability of the registered wireless hotspot.

As the position data is automatically sent wirelessly to the dedicated server, smartphones overcome an important limitation of GPS loggers. Also, the real time recording app.

[4] application

[5] Wireless-Fidelity (Wi-Fi): "a local area wireless technology that allows an electronic device to exchange data or connect to the internet using radio waves"
Global System for Mobile communications (GSM): "a standard developed by the European Telecomunications Standards Institute (ETSI) to describe protocols for second generation (2G) digital cellular networks used by mobile phones. Is it the global standard for mobile communications with over 90% market share" (Wikipedia)

[6] Wireless local area network (WLAN): "a device that allows wireless devices such as smartphones to connect to a wired network using Wi-Fi, or related standards" (Wikipedia)

FIGURE 2.3: Illustration of the "MOVE" mobile application in the passive logging mode (left) and active logging mode (right), with a diary application (Vlassenroot et al., 2014)

doesn't suffer from on-board memory limitations because data transmissions with the server are frequent. Moreover, survey participants are much more likely to carry their cellular phone with them throughout the day than devices as GPS loggers that they don't usually bring with them.

Furthermore, smartphone are recent high-tech devices, therefore not all the population has access to these devices nor that they necessarily use mobile internet access. Also, not all the applications (in case of active monitoring) have the same ease of use which creates an additional difficulty for inexperienced smartphone users and may introduce biases (Fernee et al., 2012).

Without any additional information reported by the participant on the modes used and trip purposes, important post-processing is necessary to transform position data (e.g. records from the different sensors) for analysis and discrete choice model estimation.

Whereas in common paper based travel diary data, trip end is defined by the user's reporting of arrival time and destination, the automatic collection of movements and stops eventually yields ambiguous results. Indeed, considering short stops of only about one to five minutes, it is difficult to distinguish between stops due to experienced traffic condition (e.g. congestion, waiting at traffic lights) or transmission gaps and stops

for performing an activity (Wolf et al., 2000). Therefore many papers have proposed methodologies to varying levels of details to determine the beginning of trips, stops in between and the end of the trips for data collected from GPS devices (loggers or smartphones). Jong and Mensonides (2003) proposes a trip end detection based on the speed of the position data, the heading between two segments of trips and the duration of the signal loss. More recently, Schüssler (2010) separates the activity detection for activities with ongoing GPS recording and the other ones with signal loss. In the first case, activities are detected by two criteria: When the speed is lower than 0.01 meters/second for at least 2 minutes or when the density based clustering function finds 10 points with at least 15 points within a 15m radius or the latter density of points during 5 minutes. Otherwise, activities with signal loss are detected when the time difference between two consecutive points is at least 15 min. Once activities are detected, trips are defined as the time periods between these activities. Furthermore, accurate GIS (Geographic Information System) land use and address data can be used to infer trip purpose once trip detection step is done (Wolf et al., 2001), (Axhaussen et al., 2003).

Identifying the travel mode with smartphone data is also possible through the analysis of the user's speed distribution combined with its acceleration in case of active data collection (trip start and end are reported by the participants) (Nitsche et al., 2012). Reddy et al. (2010) and Chen (2013) have proposed a travel mode detection based on sparse GPS records and acceleration collected with smartphones. Chen (2013) also uses Bluetooth records and transport networks characteristics (e.g. railways, bus networks or bike lanes) to infer the map-matched trips. Hemminki et al. (2013) proposes a mode detection algorithm based on the accelerometer of the device made possible by an improved algorithm to detect the gravity component of acceleration. A big advantage of this approach is that it is not dependent on whether or not GPS records are available (e.g. when the user is moving underground, inside a station or while the traveler is in the train unable to stay close to the window). In comparison to Reddy et al. (2010), he finds a significantly better precision for stationary, bus and train detection and recall for train, metro and tram. [7]

[7] precision: "fraction of retrieved instances that are relevant" and recall: "fraction of relevant instances that are retrieved" (Wikipedia)

2.2 Mode choice models

This section describes discrete choice models, that have very broad applications, in the case of mode choice modeling. For this purpose, different discrete choice models can be utilized depending on the objective. We present below most commonly used forms of discrete choice models that are the multinomial logit and nested logit models.

2.2.1 Discrete choice analysis framework

Discrete choice models are used in transportation research among other numerous applications that include econometrics and marketing. In the context of discrete choice theory, the assumption is that an individual chooses one of the alternatives among all the possible alternatives available which is defined as the exhaustive choice set. A common example in transportation research is a person deciding which mode (car, bike, transit, etc.) he will take take to go working. Basically, the assumption is that a person (denoted by n) associates an unobserved utility U_{ni} to each alternative (denoted by i, with $i \in \mathcal{C}_n$ and $\mathcal{C}_n$ the choice set). This utility is not observable and is only known of the individual. In this context the utility can be expressed by

$$U_{ni} = \beta_{ni} \cdot x_{ni} + \epsilon_{ni} \tag{2.1}$$

where x_{ni} correspond to the observed variables: attributes of the alternatives (travel time, distance, cost, etc.) or the characteristic of the decision making agent (car availability, seasonal ticket, age, gender, number of persons in the household), the decision-making agent being the person, household or business making the choice. β_{ni} are unknown parameters to be estimated with the data. ϵ_{ni} captures the remaining factors that are not included as observed variables. This error term can come from various sources (preferences are not homogeneous, incomplete information about the attributes of the alternatives such as comfort, measurement errors, etc.). At this point, different forms of the model can be derived depending on the assumptions made on the density function of the error component which is presented further down.

The behavior of the user is utility-maximizing, we assume he is perfectly rational and therefore by weighting the positives and negatives, he chooses the alternative that provides the highest utility:

$$P_{ni} = Prob(\, U_{ni} > U_{nj}, \quad j \neq i \,) \tag{2.2}$$

where P_{ni} is the probability that individual n chooses alternative i. This assumption is commonly called the decision rule. For the estimation of the parameters ($\beta - values$ in equation 2.1), the researcher collects data from surveys to gather choices and observed explanatory variables of the decision-maker.

Parameters β_i are finally estimated with a function "called the maximum likelihood estimator" that maximizes the probabilities for all the observations as shown in equation 2.3 where N is the total number of observations.

$$\hat{\beta} \; = \arg\max_{\beta} \prod_{n=1}^{N} P_{ni} \tag{2.3}$$

2.2.1.1 Multinomial logit model

The multinomial logit model is the most common form of discrete choice model. An important advantage of this model is that its individual choice probability is expressed in a very simple and elegant form (see equation 2.4). Nevertheless, the model makes a strong assumption on the random error component. Indeed, the ϵ_{ni} is assumed to be independent and identically distributed (i.i.d) extreme value (EV).

$$P_{ni} = \frac{exp(\beta_i x_n)}{\sum_{j=1}^{J} exp(\beta_j x_n)} \tag{2.4}$$

where J is the total number of alternatives, β_i are unknown parameters to be estimated with the data and x_n are vectors describing the attributes of alternatives. The model is suitable for mode choice modeling in transportation research. In this case, the choice set is defined as a set of available travel modes for different kinds of trip purpose (e.g. business trips, leisure trips).

However, the model is not able to capture correlation among alternatives because of the assumption of independence on the disturbance term. In order to solve this issue, the nested logit model, that was first introduced by Ben-Akiva (1973), allows correlation over alternatives by partitioning the choice set into nests.

2.2.1.2 Nested logit model

As mentioned above, there are situations in which certain alternatives share important unobservable qualities that can't be captured by a multinomial logit. In these situations, using a nested logit model, with its ability to account for similarities between alternatives via partial correlation of the error terms, can improve the quality of the estimation (Silberhorn et al., 2006).

Considering a tree structure as exposed in Fig. 2.4, the choice probability P_{im} of alternative i within nest m results from the product of the marginal choice probability P_m (see equation 2.5 and 2.6) for nest m (Level 2) and the conditional probability $P_{i|m}$ (see equation 2.7) for alternative i within nest m (Level 1) where V_m is the deterministic part of the utility (systematic utility expressed as a function of observed variables), N corresponds to the number of nests and C_m corresponds to the choice set within nest m. The errors term of the marginal utility from Level 2 and the conditional utility from Level 1 are identically and independently distributed (i.i.d) extreme-value. Hence both levels contributes to the compound error term. The scale parameter that appears in the marginal probability ϕ (equation 2.5) and the logsum term denoted by IV_m (equation 2.6) is described as an inverse measurement of the correlation amongst alternatives within each nests. To be consistent with the utility maximization theory, this number must be between 0 and 1. If the value $\phi = 1$, the model is equivalent to the Multinomial logit model (Silberhorn et al., 2006).

$$P_m = \frac{exp(V_m + \phi IV_m)}{\sum_{n \in N} exp(V_n + \phi IV_n)} \tag{2.5}$$

$$IV_m = ln \sum_{j \in C_m} exp(V_{j|m}) \tag{2.6}$$

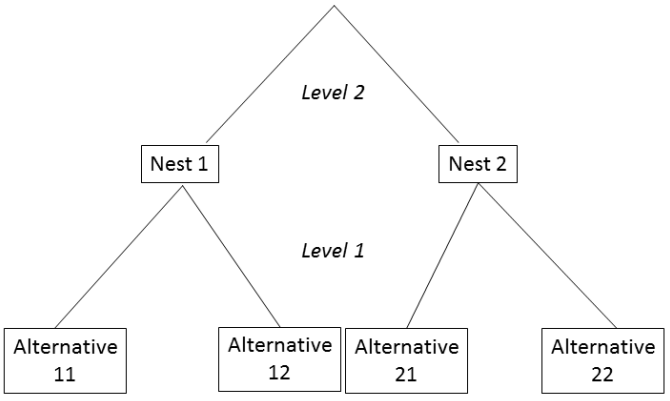

FIGURE 2.4: Tree structure of a nested logit model

$$P_{i|m} = \frac{exp(V_{i|m})}{\sum_{j \in C_m} exp(V_{j|m})} \tag{2.7}$$

The multinomial logit model and nested logit model have been used in particular to study the influence of weather conditions on individual's mode choice decisions. Sabir et al. (2010) analyzes the influence of changing weather conditions on mode choice decisions of individuals in the Netherlands. The multinomial logit model takes into account seasonal variables (summer, winter, etc.), meteorological variables (temperature, wind and precipitation) and socio-economics for different activity based trips (business trips, commuting trips, educational trips, etc.). The conclusion reached in this study is that people prefer biking in warm weather and under gentle wind conditions, whereas the car is preferred in opposite conditions. Moreover, public transport and walking alternatives are not significantly influenced by weather conditions.

Furthermore, Saneinejad (2010) studies the influences of temperatures, wind speed and precipitations on mode choices in the city of Toronto, Canada by specifying 3 discrete choice models. A multinomial logit model and two nested structures are therefore considered and evaluated: (i) the multinomial logit model evaluates all the possible alternatives (e.g. auto-driver, auto-passenger, transit, walk and bike) without creating nests (ii) the first nested logit has two nests: the *motorized* nest composed of *transit*, *auto-driver* and *auto-passenger* against the *non-motorized* nest composed of *walk* and *bike*. (iii) the second nested logit model keeps the *non-motorized* nest and considers the other alternatives

at the same *level*: *transit*, *auto driver* and *auto-passenger*. Statistical tests show finally that both nested structures are not suitable for modeling the impact of weather on mode choices. Moreover, the multinomial logit model shows that changes of temperature as well as precipitation have a large impact on the number of pedestrian and bicycle trips. It is therefore interesting to conclude from this example that the nested logit model is not necessarily preferable to the simpler multinomial logit model.

Finally, no evident research utilizing smartphone data in passive monitoring for mode choice modeling was found in the literature. Therefore in order to fill this important gap, this paper investigates the challenges of using this opportunistic source of data for mode choice analysis.

Chapter 3

Available data

3.1 Available smartphone data

In the late autumn 2008, the Nokia research center Lausanne (NRC) concludes that no public dataset is matching its internal research needs which has triggered the motivation to collect a comprehensive dataset gathering as much information as possible about people, places and the interaction between people and places. Two years later, the data collection campaign begins, involving a sample of 174 participants provided with smartphones and ending in September 2011, a little over 20 months. For more information on the Nokia data collection campaign than the one presented in this section, you can refer to Kiukkonen et al. (2010).

3.1.1 Participants and technology used

The main goal of the enrollment was to include a heterogeneous set of real life social networks in the population. Hence, an initial set of individuals was chosen from Lausanne, to serve as seeds for generating the sample. Then, these start nodes were encouraged to recruit their friends, colleagues, and family members eventually leading to a population comprising real life social networks with individuals from mixed backgrounds. Hence, position data of the sample is not limited geographically to Switzerland, even if we expect a majority of records to be located near Lausanne. Furthermore, all the participants had experience in using a mobile phone when they joined the campaign.

The phone used during the campaign is the Nokia N95 (see Fig. 3.1a) provided with the Symbian OS [1] and an integrated GPS (see Fig. 3.1b).

(A) Nokia N95 smartphone

(B) Integrated smartphone GPS (Texas Instruments GPS5300 NaviLink 4.0)

FIGURE 3.1: Nokia data collection campaign's device and integrated sensor

3.1.2 Data collection method

Smartphones of the surveys are equipped with a client software that aims to make the data collection invisible to the users while optimizing the ratio between the data collected and power consumed. Indeed, for a twelve months period campaign, participants don't have to be bothered in their day to day use of the smartphone while the device is collecting data. For this reason, the user has enough battery to use his device normally during the day and charges the batteries over the night.

The data collection software is based on a state machine that can run through 13 possible states. The range of parameters sampled as well as the sampling frequency varies over the state in order to increase the battery performance. States are changed based on the acceleration observed, WLAN, cellular network, GPS and the device state information (e.g. for instance, whether or not the battery is connected to a charger). More details

[1] Operating System (OS): "software that manages computer hardware resources and provides common services for computer programs" (Wikipedia)

on state transitions and data collected in each state is shown can be found in Kiukkonen (2009).

The data collected in each state is then stored to the device and automatically uploaded to the server twice a day via a known WLAN access point. If the upload is not successful, because the user is outside of Switzerland or because no known WLAN is detected, the client keeps trying to connect every two hours until the upload is successful. Anyway, the device can keep several weeks of data in its memory. The different data modalities collected with the client can be partitioned into 3 main categories:

- *Position data*: GPS (when available), WLAN access point information (when available), acceleration and cellular network information.
- *Social interaction data*: Call logs, short message logs and Bluetooth scanning results.
- *Media creation data*: Locations where images have been captured, video shot, music played. And information on Applications used with the Symbian OS.

Once uploaded in the server, data is anonymized by clearing real names of the users, phone numbers, calendar information, Bluetoot, WLAN MAC addresses [2] and acoustic data.

3.1.3 Data format

The data uploaded and anonymazed is then further processed by a parser that generates the SQL [3] (postgresql) database. The UML [4] of the dataset is shown in Appendix A. The "records" table is central as it stores the time information and user_id (e.g. number identifying the user) information for any record of the relational database. All other tables are related to this "records" table via the db_key integer (e.g. identifier of the record). For this project, we use the following tables: "gps", "accel", "gpswlan", "wlanrelation" and "wnetworks","btrelation", gsm and gsmcells.

[2] Media Access Control address (MAC address): "unique identifier assigned to network interfaces" (Wikipedia)

[3] Structured Query Language (SQL): "special-purpose programming language designed for managing data held in a relational database management system" (Wikipedia)

[4] Unified Modeling Language (UML): "a general purpose modeling language in the field of software engineering which is designed to provide a standard way to visualize the design of a system" (Wikipedia)

"gps" table provides information on the vertical and horizontal accuracy of the record, latitude, longitude, altitude coordinates, speed, heading and the time since the GPS boot. Wi-Fi tables are only providing the time when the WLAN is seen by the user and the coordinates if they are available.

3.1.4 Data amount

The quantity of data collected with the devices at the end of the campaign is large:

- 14 million GPS tags are recorded with different levels of accuracy within a range from 5 meters to 400 meters (see Fig. 3.2).
- 44 million Wi-Fi records are collected and 77 % of this amount is localized.

The first conclusion is that GPS, which is the most accurate positioning method, doesn't record as often as the WLAN port of the device due to its high power consumption. Even during trips or movements of the user, GPS recording is not always activated. Furthermore, when the device has consumed 75% of the battery's life autonomy, GPS is definitely turned off until the device is connected back to a charger. Hence, the direct result of these observations is that our GPS data is very sparse compare to Wi-Fi data.

Besides GPS, the less accurate localization technique WLAN triangulation can be used. By combining GPS and Wi-Fi data, we identified the position for 66% of the access points. The reason why some access points can't be localized (e.g. 34% of the Wi-Fi dataset) is because these coordinates couldn't be inferred from GPS. The procedure to localize access point is the following: the client tags automatically the known WLAN access points with GPS coordinates if the location information is available roughly simultaneously with the WLAN scanning results. Then, the Wi-Fi access point coordinates are set as the centroid of all the associated GPS points and stored in the "wnetworks" table. The accuracy of these coordinates depends mainly on the range of the Wi-Fi networks which is dependent on the bandwidth of the installation (higher in public places for example) as well as its position (e.g. outdoors or indoors). Watzdorf and Michahelles (2010) reports an accuracy between 30 meters and 50 meters when the precise location of the WLAN acces point is known. Hence in our case, WLAN access points have an accuracy from 30 meters to 50 meters plus the added accuracy of the GPS points used

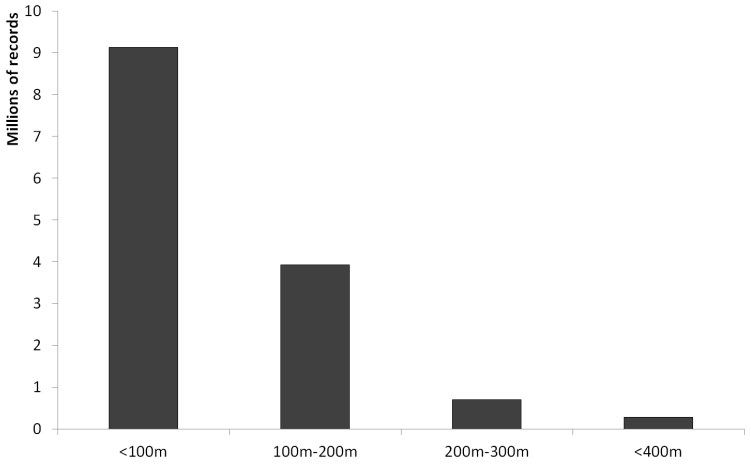

FIGURE 3.2: GPS accuracy (in meters)

to infer their locations. At this point, all the access point that can't be localized are removed from the dataset. This represents 10 million Wi-Fi records.

3.2 Demographic questionnaire

In addition to the consent forms, participants were asked during the enrollment to fill a demographic questionnaire providing information like age, gender, working status or the number of persons in the household. Some of these socio-economics statistics are provided in the table 3.1. We can see that the largest age group is from 22 to 38 which is relatively young and may explain why the sample is considered as advanced technologically as follow-up to the "segmentation" questionnaire (Kiukkonen et al., 2010). Furthermore, nearly 10 % of the participants haven't filled the questionnaire.

As Nokia data collection campaign was not conducted for mobility analysis, important socio-economics for travel mode choice modeling are missing (e.g. possession of a car license, number of car in the household, possession of a seasonal ticket, possession of a bike) for most of the participants (only 15 participants over the 158 have reported these socio-economics)

Due to the snowball sampling used for the enrollment of the participants, that took place in addition in the campus at EPFL [5], an important part of the sample is composed of EPFL students, PhD students or professors as well as researchers from the Scientific

[5] École Polytechnique Fédérale de Lausanne (EPFL)

Category	Sample
Working status	
Workers full time	53%
Workers part time	10%
Students	26%
Other employment status	11%
Age	
< 21	9%
22-38	80%
> 38	11%
Gender	
Male	61%
Female	38%
Not reported	1%
No survey data	9.20%

TABLE 3.1: Socio-economic characteristics of the users

Parc also located inside the university campus. Hence, This "working" population is not representative of the real Lausanne population.

3.3 Travel mode information

As discussed earlier in the report, different algorithms have been proposed in the literature to infer the travel modes used by the participants during their trips. In this project, we use the results of the map-matching algorithm executed on the Nokia dataset by Chen (2013). For any details on the algorithm, you can refer to the author's thesis.

The probabilistic map-matching algorithm identifies within a travel period the physical path and sequences of modes traveled. To do so, it uses the information collected via the sensors of the device: GPS when available or Bluetooth and acceleration when GPS is not available. The results of the algorithm are probabilistic meaning that for each trip, there is a set of candidate paths given with their probability to be the right one.

Furthermore, the algorithm provides an estimated length of the *arcs* composing the paths, an *arc* being a segment where there is no change of the travel mode.

The data format is displayed for one trip in Fig. 3.3 where the *user_id*, *trip_id* and *path_id* are the identifier of respectively the user, trip and path, the *begin_time* and *end_time* denote respectively the beginning and the end of the map-matched trip, the *mode* represent the sequences of modes traveled and finally the *likelihood* denotes the probability associated with the path. We decide for this project to retain only the most likely path of each map-matched trip. For example, in Fig. 3.3 we are only retaining *path_id* number 13 that represents the highest probability to be the right path.

user_id integer	trip_id integer	path_id integer	begin_time timestamp without time zone	end_time timestamp without time zone	mode character varying(30)[]	line_geom geometry	likelihood numeric
6271	15	13	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.27
6271	15	3	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.25
6271	15	1	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.21
6271	15	2	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.09
6271	15	7	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.04
6271	15	12	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.04
6271	15	8	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.03
6271	15	10	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.02
6271	15	9	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.01
6271	15	11	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.01
6271	15	6	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.01
6271	15	4	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.01
6271	15	15	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.01
6271	15	5	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.00
6271	15	14	2010-09-25 00:23:09	2010-09-25 00:49:29	{bike,bike,bike,bi	01050000	0.00

FIGURE 3.3: Different multimodal paths candidates for trip number 15

The dataset is composed of 15'148 map-matched trips and has detected with success the following travel modes: car, bike, walk, metro and bus (only train is missing).

3.4 Weather database

The data source is a weather database of MeteoSwiss recorded in Lausanne Pully's station at altitude 461 meters. The weather station is the nearest to Lausanne that has archived weather variables during the dates of the campaign (e.g. from January 2010 to September 2011). The weather variables are recorded in an hourly basis:

- The hourly maximum Gust peak (in meter/second) and the maximum ten minutes mean wind speed of the past 12 hours.

- The hourly total precipitation (in millimeters), the hourly maximum ten minutes total (in millimeters) and the total precipitation during 24 hours.
- The hourly mean temperature (in degree celsius).
- The hourly sunshine duration (in minutes).

Chapter 4

Building a mode choice model with the available data

Chapter 4 begins by identifying the issues to solve in order to derive a dataset suitable for mode choice analysiss and pursues with the strategies proposed to address these issues.

4.1 Identification of the issues

The project aims to estimate a multinomial logit model which framework is described in section 2.2.1.1. The difficulty associated with processing smartphone data for mode choice modeling in addition to the limitation of the available smartphone data is important and require an extensive processing. For this reason, we propose a basic model to apply the derived dataset.

The objective of the model is to analyze mode choice behavior of workers for their way to work and way back to home on weekdays (see Fig. 4.1).

4.1.1 Direct trips from home to work

Whereas in conventional surveys, trip start and end are reported by the respondent as well as the trip purpose, none of this information is available with smartphone data

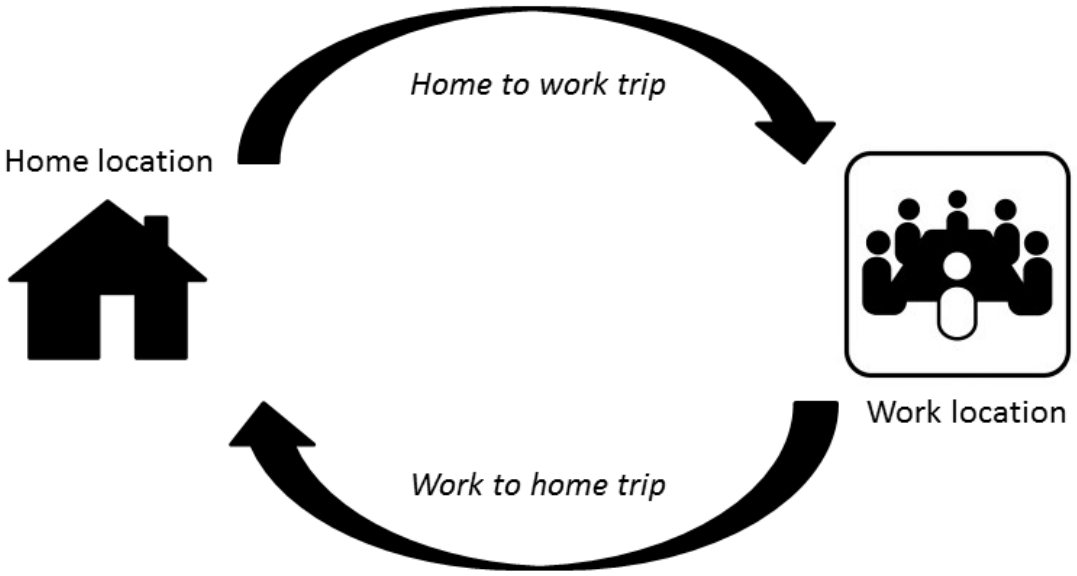

FIGURE 4.1: Activity based trips from home to work

when the data collection is passive. Hence first important issues to solve in order to detect trips from home to work are:

- Identify the trip purpose in order to focus on workers and their home to work trips (and not other activity based trips).
- Detect trips which means extracting time windows (e.g. trip departure time and arrival time) which requires the proposition of a trip detection algorithm.

In activity based approach, it is frequent to consider a "work tour" as the way to work plus the way back (Vij and Akshay, 2013). In this thesis, we won't consider the tour but the home to work trip as a single observation. Therefore, we assume that the mode choice on the way back from work is independent to the choice on the way to work (for the same user on the same day).

Furthermore, we focus on direct trips, which means we exclude from the estimation dataset trips including all sort of stops/activities (e.g. stops in the way to work or stops requiring a detour). Even though, it is still important to know which trips are including stops or activities to be able to omit them. Hence, we propose an algorithm able to detect stops made during the trips.

4.1.2 Detection of the travel mode

Travel mode information is not reported by the respondents contrary to data collected with conventional surveys. Hence, this information had to be inferred with the available data. The issue is solved in this project by retrieving the results of the map-matching algorithm that provides such information on the travel modes (Chen, 2013). The mode detection algorithm was launched on the entire Nokia dataset, so we will only extract the map-matched trips that are useful for our analysis (e.g. map-matched trips describing the travel modes of the inferred home to work trips).

4.1.3 Definition of the choice set

The choice set of the model is defined as the set of possible alternatives of the user when he goes to work. In this project, we assume mode choice is between 3 alternatives, *car* that accounts for car as the driver, passenger and taxis, *public transport*, that corresponds to trip by bus and metro, and *soft modes*, that correspond to trips by bike or walking.

At this step, we identify a first issue due to the difficulty of processing smartphone data, that is the missing train in the public transport alternative. This issue has a direct consequence on the choice set: for long trips, car is the only alternative available.

Another issue due to the available dataset (e.g. Nokia dataset) concerns missing socio-economics of the data (see 3.2). Hence, we can't differentiate choices that are due to a lack of alternative (e.g. a user without car and bike that chooses the public transport alternative) and real choices after consideration of the three alternatives. This issue will be easy to correct in future surveys by adding these questions in the initial demographic questionnaire. In our model, we propose to report these socio-economics when they are reported and to assume bike and car are available in the remaining cases (e.g. the user has a bike, a car and a driver license).

4.1.4 Determination of the chosen mode

Next issue is common to mode choice models, it concerns the determination of the chosen mode for a home to work trip. Therefore, we define for each trip a "main mode" which

corresponds to the travel mode used to cover the longest motorized distance. Hence, we relax the hypothesis that mode choice is dictated by the longest leg of the trip.

4.1.5 "Quality" of observations

Friederich et al. (2014) report a deficiency in the number of map-matched trips inferred due to the difficulty of detecting the mode with smartphone data. It results in parts of home to work trips without mode information. Therefore, we propose to associate to each trip a "quality measurement".

4.1.6 Assigning attributes to the chosen alternative

Main attributes of the alternatives in mode choice models are travel time, distance and cost.

Travel time is easily inferred with the departure time and arrival time of the trips. Distance needs to be imputed as it depends on the route choices of the participants. We use Google directions to estimate the distance of the trips assuming that the traveler always chooses the shortest path to go working.

Finally, cost is not available and can't be inferred rigorously because we miss the important socio-economic "possession of a seasonal ticket", required to infer the cost of the public transport alternative.

4.1.7 Assigning attributes to the unchosen alternatives

Next issue is also common to mode choice models in general, attributes of the unchosen alternatives are unknown and need to be imputed.

4.1.8 Summary of the processing issues

Table 4.1 presents the catalog of issues that are solved in next section. The table also specifies issues due to the way data was collected (e.g. "smartphone data") and issues common to mode choice models ("Mode choice models").

Issues	
Description	Due to
Identification of the trip purpose	Smartphone data
Missing train trips	Smartphone data
Trip detection	Smartphone data
Detection of the travel modes (retrieved data)	Smartphone data
Determination of the chosen mode	Mode choice models
Detection of stops/activities	Smartphone data
Missing travel mode information for parts of the trips	Smartphone data
Attributes of the unchosen alternative	Mode choice models

TABLE 4.1: Catalog of issues

4.2 Solutions to address the issues

This section addresses the issues identified and summarized in table 4.1.

4.2.1 Identification of the trip purpose

In order to identify the trip purpose, we propose to retrieve first home and work locations found by Buisson et al. (2013) in order to be able to focus then on trips starting from home and ending at work (respectively from work and ending at home).

4.2.1.1 Home and work clusters (retrieved data)

Discussion

To find home and work locations, Buisson et al. (2013) proposes a Density-Based spatial clustering applied on the Wi-Fi dataset. Even though GPS records could be used as well, a higher density of Wi-Fi records are collected in places like home or work because of the states transition of the mobile that are explained in Kiukkonen et al. (2010). Therefore the density function algorithm is more likely to detect home and work if we use Wi-Fi records.

Processing

Finding home and work location required 3 steps as explained by Buisson et al. (2013):

1. **Identify places of interest** [1]: For each user, most viewed access points (more than 100 times) are extracted assuming that more views are associated with places of interest and the opposite during trips.

2. **Apply the DB scan clustering**: The Density-Based spatial clustering is applied on the dataset and among all clusters found for each user, the two biggest one are extracted (e.g. the ones with the biggest sum of records for the access points inside the cluster). The assumption is that people spend the most of their time whether at home or work.

[1] Point Of interest (POI): "specific point location that someone may find useful or interesting. Most consumers use the term when referring to hotels, campsites, fuel stations or any other categories used in modern navigation systems" (Wikipedia)

3. **Assign the meaning home or work to these clusters**: For each of the two clusters found before, two ratios are computed: P_{work} corresponds to the number of views on weekdays and business hours (10 am to 4 pm) over the total number of views and P_{home} corresponds to the number of views on weekends and night hours (11 pm to 5 am) over the total number of views. The assumption is that people spend more time at home on weekends and night hours and more time at work during weekdays and business hours. Thus, if $P_{home} > P_{work}$ the cluster corresponds to the home location and if $P_{work} > P_{home}$, the cluster corresponds to the work location.

Conclusion

In the end, the results found by Buisson et al. (2013) in addition to our own conclusions are:

- 2 users have less than 2 recorded Access Points. (FAIL)
- 3 users have less than 2 clusters found. (FAIL)
- 44 users have two clusters but no meaning could be associated to them. (FAIL)
- 30 users are non working participants or from other unknown or not reported categories and therefore the work location is not meaningful. (FAIL)
- the home and work location is found for 95 students, workers and part time workers.

As a precaution, we won't use the results found for part time workers. We assume that the relevancy of the work location inferred for these users is too much dependent on the number of days they work per month. Therefore, 9 more users are excluded from the dataset.

In the end, we have 86 accurate home and work locations of students and full time workers. To study home to work trips, we need to extract home and work locations of the workers but not of the students as we don't consider going to university as a working activity. Hence over 86 workers and students, 53 full time workers are kept in the estimation dataset.

4.2.1.2 Home and work centroids and size of the clusters

Discussion

The home and work locations analyzed in the previous section are scatters of WLAN access points assigned with the meaning home and work. Furthermore, the number of Wi-Fi MAC addresses assigned to a cluster can vary from one to a big number. At this step, it is important to have an idea of the accuracy of our clusters and one way to do it is by measuring the size of the scatter of points representing the cluster.

We propose now to compute for each cluster the centroid with the k-means clustering, and the radius of the cluster, defined as the distance from this centroid to the farthest point of the scatter of points. The centroid is the best point of reference for the cluster and the radius is a good indicator of the accuracy.

Processing

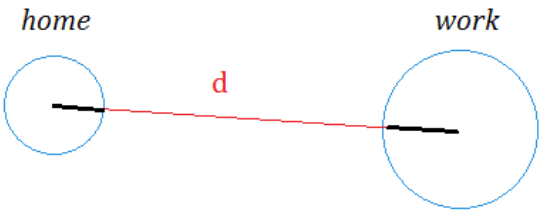

FIGURE 4.2: Home and work clusters definition and distance between the clusters

As presented schematically in Fig. 4.2, we compute the centroid, radius and distance between the centroids, for the 53 users provided with home and work locations. Results for the radius (in meters) and coordinates of the centroids (in the geographic coordinate system) are shown in Appendix B.

Conclusion

- The first type of error that could happen is that home and work of a user share the same location or are very close to each other meaning that we capture the same activity location. After the processing, no one of the 53 users are having this type of error.
- The second issue is in relation with the radius of the clusters. Typically for home and work locations, we assume that a radius lower than 200 meters results in a

good approximation of the activity location. As shown in the histograms of Fig. 4.3, a majority of clusters for both activities have a radius around 150 meters. More precisely, 10 home locations and 6 work locations have a radius higher than 200 meters. Thus, in the overall, the size of the clusters is relevant for home and work locations.

- For user 5974, the home address is reported, therefore we don't compute the radius. For few users, the radius is equal to 0 meter meaning that the cluster is composed of only one WLAN access point.

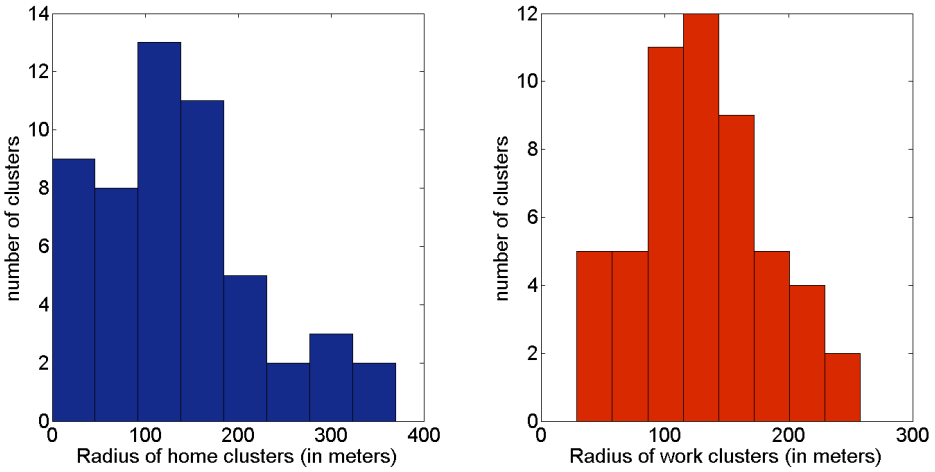

FIGURE 4.3: Histogram of the radius of home and work clusters

4.2.2 Missing train trips in the public transport alternative

Discussion

In previous chapters, we reported that transit trips are not available in our dataset. Hence, whenever a user is not located in an urban area and wants to do long trips (e.g. >10km), he has no choice than the car alternative. To address this issue, we propose to set the scope of the analysis to Lausanne area where workers have a real set of available alternatives to go working. We use below the results of the map-matching algorithm developed by Chen (2013) and presented in section 3.3.

Solution proposed

Public transport trips available in our dataset are those made by metro inside cities, and those by bus mainly inside cities or locally outside of the agglomerations. Hence,

we expect to observe a big majority of these public transport trips inside the Lausanne agglomeration or inside other big Swiss agglomerations. In Fig. 4.4, we plot origins and destinations of all the arcs traveled by bus or metro. As expected, origins or destinations are mainly located inside or close to big agglomerations that are 1. Lausanne, 2. Geneva, 3. Neuchâtel, 4. Bern, 5. Zurich and 6. Bâle.

Therefore the option is to reduce the scope of the analysis to these big agglomerations where users have a real choice between car, public transport modes *bus* and *metro*, and soft modes.

With this aim in mind, we still have to differentiate users that both live and work in the same big agglomeration and participants that live and work in different agglomerations. Indeed, for the second category of users the issue remains. In Fig. 4.5, home and work locations (in blue and red) are plotted for the sample of full time workers: In the end, the number of full time workers living and working in the same agglomeration is reported (per agglomeration): 1. Lausanne (27 users), 2. Bern (1 user) and 3. Geneva (1 user). The scope of the analysis is therefore reduced to the Lausanne agglomeration where we find the highest number of workers under this condition. In the end, our dataset is now composed of 27 full time workers for whom we can state there is a set of available alternatives other than car for their home to work trips.

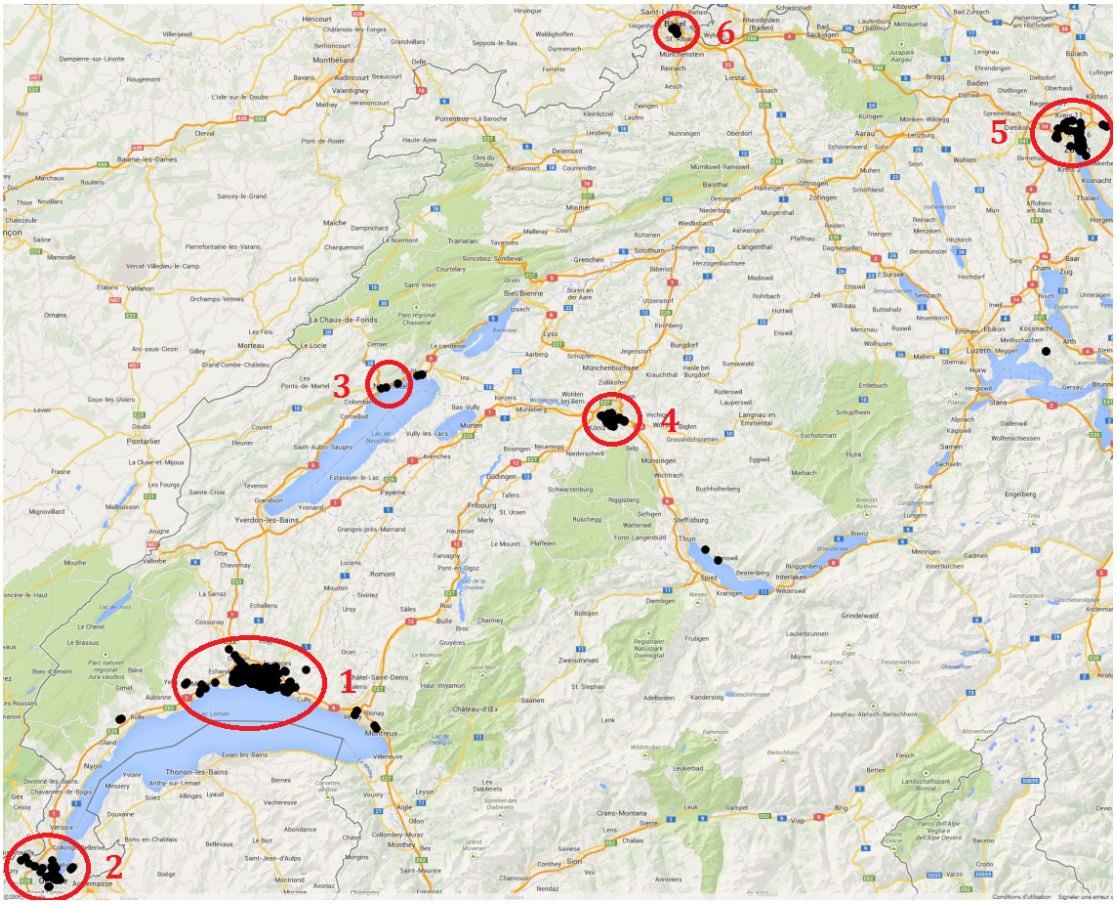

FIGURE 4.4: Origins and destinations for all the arcs traveled by bus or metro

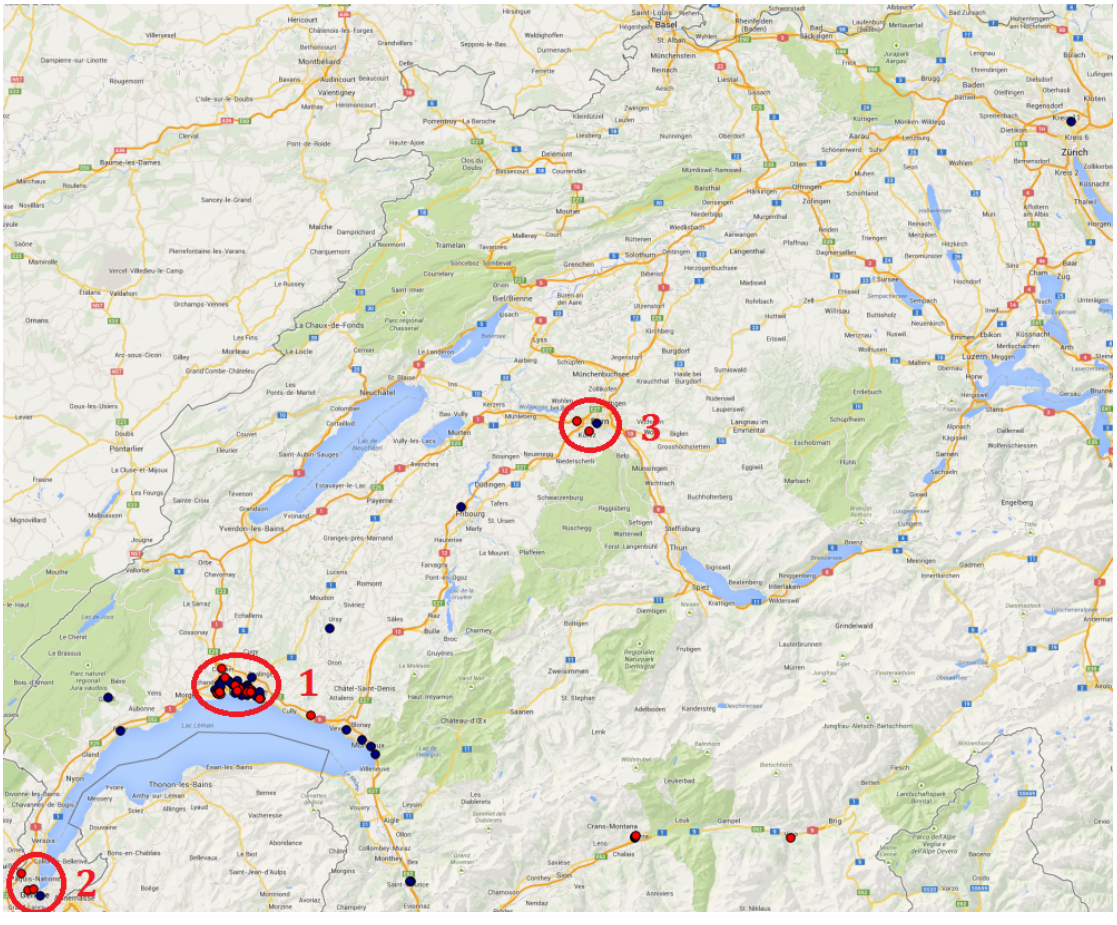

FIGURE 4.5: Home (in blue) and work (in red) locations for full time workers

4.2.3 Trip detection

Discussion

The thesis aims to conduct a mode choice analysis on home to work trips of the users. Therefore it is necessary to detect the time windows of their trips for this particular trip purpose.

Based on the acceleration communicated by the smartphone sensors and the position of the user, the client software detects these trip starts and ends and activates the GPS when the participant is in a trip. Therefore a trip detection based on the transition of states can be imagined. A first problem is that the transition of state from "stationary" to "in movement" is not instantaneous. Indeed, a middle state is required to detect whether or not the user is transitioning outdoors or just changing the location indoors and the transition from "stationary" to "in movement" can therefore take several minutes. Furthermore, the software decides to change the state when a certain threshold for the acceleration is reached which makes some situations ambiguous and explains reported "bugs" of the software as observed for the transition of states displayed in Fig. 4.6. In this situation, a trip detection based on the transition of states would detect 6 trips of less than 5 minutes in only one hour. Finally, when the battery of the phone is over, transition of states are not triggered anymore. For these reasons, the reported transition from "stationary" to "in movement" states won't be used as an indicator of trip start and end.

In this project, we use a trip detection algorithm based on the Wi-Fi records. The advantage of Wi-Fi data compare to GPS is that it is not relying on the client software decision or the battery of the device to be recorded. Hence, Wi-Fi is always recorded during the most frequent states, the only difference is the sampling interval. Furthermore, we analyzed that the accuracy of these records when they are localized is good and close from the GPS one. Finally, more WLAN access points are visible inside agglomerations and hence we expect to have an important number of records.

Processing

The algorithm used is following the steps:

db_key integer	user_id integer	time_stamp timestamp without time zone	type character varying(16)	state character varying(32)	reason character varying(128)
932871	5974	2010-03-18 17:30:45	state	stationary	not moving by gps
932897	5974	2010-03-18 17:39:46	state	indoor mobile	moving by accel
932905	5974	2010-03-18 17:40:46	state	outdoor mobile w:	gained fix
932971	5974	2010-03-18 17:47:47	state	stationary	not moving by gps
933020	5974	2010-03-18 18:05:47	state	indoor mobile	moving by accel
933030	5974	2010-03-18 18:06:47	state	outdoor mobile w:	gained fix
933057	5974	2010-03-18 18:09:51	state	stationary	not moving by gps
933061	5974	2010-03-18 18:10:51	state	indoor mobile	moving by accel
933075	5974	2010-03-18 18:11:51	state	outdoor mobile w:	gained fix
933362	5974	2010-03-18 18:19:56	state	stationary	not moving by gps
933366	5974	2010-03-18 18:20:56	state	indoor mobile	moving by accel
933379	5974	2010-03-18 18:21:56	state	outdoor mobile w:	gained fix
933404	5974	2010-03-18 18:24:56	state	stationary	not moving by gps
933424	5974	2010-03-18 18:30:56	state	indoor mobile	moving by accel
933434	5974	2010-03-18 18:31:56	state	outdoor mobile w:	gained fix
933465	5974	2010-03-18 18:34:56	state	stationary	not moving by gps
933480	5974	2010-03-18 18:38:56	state	indoor mobile	moving by accel
933490	5974	2010-03-18 18:39:57	state	outdoor mobile w:	gained fix

FIGURE 4.6: Bugs of the client software

1. **Identify home and work records**: Having home and work centroids and size of the clusters for each user, we detect Wi-Fi records collected inside "home" and "work" clusters. In the end, over 5'700'000 Wi-Fi records on weekdays, 4'600'000 are detected inside "home" and "work" clusters

FIGURE 4.7: Wi-Fi records (in green) inside home and work clusters (blue circles) are assigned the meaning home and work

2. **Extract time windows**: For each user, the procedure to detect a trip is the following: we extract the last record seen at "home" or "work" defined as the departure time of the trip and we look for the next record seen in a cluster ("home" if the departure time was from "work", "work" if the departure time was from "home") defined as the arrival time. The procedure starts with the oldest records of the user seen in the cluster "home" or "work" and is repeated until no more records of the user are available. For all the users, We find 7'000 trips.

FIGURE 4.8: Determination of the departure time and arrival time. Green points represent Wi-Fi records and blue circles represent clusters home and work

3. **Clean non consistent trips**: The following trips are removed from the dataset:

   - Departure and arrival are not the same day (595 trips).
   - More than 6 home to work trips in a day (815 trips).
   - Trips longer than 2h: We aim to capture direct home to work trips. Hence, the threshold corresponds to the maximum travel time that a user would take to cross by bike (e.g. the slowest alternative) the two farthest boundaries of the Lausanne agglomeration (390 trips).

Conclusion

In the end, we find 5'200 trips. In figure 4.9, the histogram of the departure times is shown. As expected more departures are happening during peak hours which is consistent for home to work trips on weekdays.

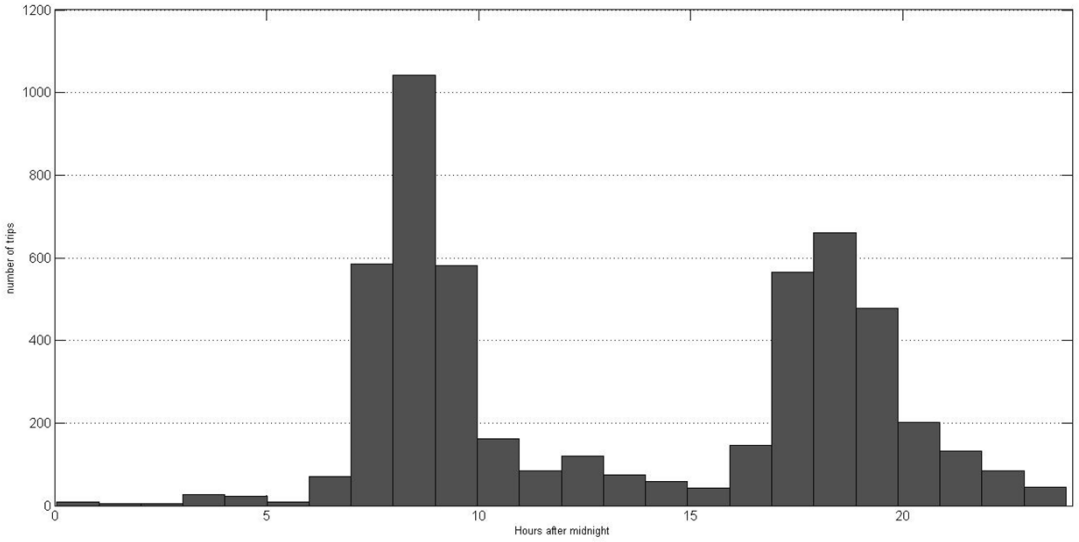

FIGURE 4.9: Departure times for all the trips inferred after the cleaning of the data

4.2.4 Detection of the travel modes (retrieved data)

So far, we have detected home to work trips of workers inside the Lausanne area. The issue is that no information on neither the path followed or the travel modes used during the trip is known. Although we are not interested in paths followed in this project, travel mode information is crucial as we plan to build a mode choice model.

For this reason, we retrieve the results of the map-matching algorithm developed by Chen (2013). As explained in Friederich et al. (2014), map-matched trips are generally segments of trips of a longer meaningful trip. It is rare for example to have one map-matched trip starting at home and ending at work. Typically, for a trip from home to work, it is usual to find several map-matched trips describing the sequence of travel modes. Hence, the procedure consists in retrieving all the map-matched trips inside our meaningful time windows. The path and the sequence of modes of the trips are therefore partially reconstituted which is schematically represented in Fig. 4.10.

In the end, we wish to have at least one map-matched trip describing the sequence of modes for each home to work trip. After the processing, over 5'200 trips, 697 trips have at least one map-matched trip included inside their time windows. One user doesn't have map-matched trips inside his home to work time windows. Therefore the sample is reduced from 27 to 26 users.

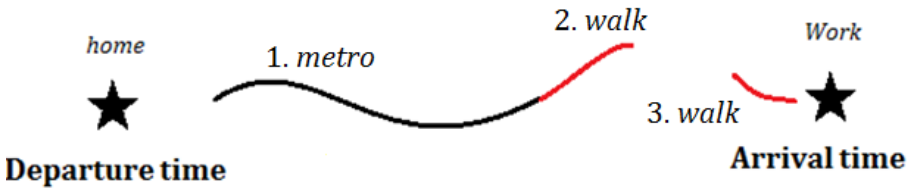

FIGURE 4.10: Sequence of travel mode used during home to work trips

4.2.5 Determination of the chosen mode

At this step, we have 697 home to work trips with travel mode information. We use the definition of the chosen mode presented in section 4.1.4.

To illustrate our point, we will be using Fig. 4.10 as an example. Furthermore, in order to get used with some terms, we precise that the ***trip*** in Fig. 4.10 is composed

of two ***map-matched trips***. The 1st ***map-matched trip*** is a ***sequence*** of two ***arcs*** {'metro','walk'} and the 2nd map-matched is only one ***arc*** {'walk'}. Furthermore, the algorithm estimates the lengths of the ***arcs*** but not their durations. Hence we only have the duration of the ***map-matched trips***.

The processing steps to identify the chosen mode are explained below:

1. For each trip, we compute the distances traveled with each travel mode. In Fig. 4.10, if we denote by d_1 and d_2 the estimated distances of respectively the first and second arc of the 1st map-matched trip and by d_3 the only arc of the 2nd map-matched trip, we have a total walking distance of $d_2 + d_3$ and a distance d_1 by metro.

2. For each trip, we sum the distances of the modes belonging to the same alternative (e.g. metro and bus belong to public transport alternative, walking and bike belongs to soft mode alternative, etc.). In Fig. 4.10, we have a total distance of $d_2 + d_3$ for the soft mode alternative and d_1 for the public transport alternative.

3. For each trip, the alternative with the highest distance is defined as the chosen mode. In Fig. 4.10, assuming that $d_1 \geq (d_2 + d_3)$ means that the public transport is the chosen alternative.

Ambiguities concerning the sequences of mode had to be treated manually, as reported in (Chen, 2013), because the algorithm has difficulties in differentiating bus and car under traffic conditions. For this reason, trips which sequences are car and bus changing over time are not feasible and need to be modified to only car or to only bus. Among our 697 trips, 5 trips are in this situation and in each case the distance traveled with one mode is always small compare to the other. Hence, we assume the correct mode is always the one inferred within the longest distance.

4.2.6 Detection of stops/activities

Discussion

Over 697 trips, it is important to know which trips are including activities or stops in order to remove them according to our objective to focus on direct trips. Hence, the issue here is to detect these stops inside our time-windows.

To do so, a first possibility is to use more clusters than the two biggest one found in Buisson et al. (2013). The algorithm found smaller clusters for each user that are possibly other points of interest. Thus, the idea would be to go though the same processing steps than in section 4.2.1 and 4.2.3, for these smaller clusters. Therefore we could see if the participant stops on his way to work in one of his inferred point of interest (e.g. gym, grocery, shopping, school for children, etc.)

The main limitation of this proposition is that we can't rely on the quality of these smaller clusters. Some of them are very close from home and work locations and may represent a place with a lot of passage of the user but are not a real point of interest. Others are only duplicates of the home and work location. Moreover, stops are not always places of interest of the users and therefore we might omit some stops with this method.

The proposed algorithm is close to a Density-Based spatial clustering (Ester et al., 1996) that adds a time parameter for the clustering in addition to the distance. It computes for each record (GPS and Wi-Fi) of a trip, the number of points that are within a radius of 150 meters and separated in time from at least 300 seconds. Moreover, we define the *point density* of a record, as the number of points around him under these two conditions. Hence, if the point density of a record is higher than 20, an activity is flagged. In Schüssler (2010) the radius threshold is set to 15 meters for records collected with GPS loggers. Knowing that the accuracy of these devices are more or less ten times more accurate than smartphones (5 to 10 meters for GPS loggers against 50 to 100 meters for smartphone sensors), we set a threshold of 150 meters. Furthermore, we assume 300 seconds is a good time threshold to detect stops. Finally, Schüssler (2010) defines a point density threshold of 15 points whereas we set it to 20 points. The reason of this choice is that we are considering more records (GPS and Wi-Fi) so we want to be more strict in the definition of this point density. For Schüssler (2010), the activity is flagged when at least 10 points have a point density of 15. In our case, we will impose a minimum time of 180 seconds inside the stop's cluster to assume the user made a stop.

The trip detection algorithm utilizes both Wi-Fi and GPS records. On the one hand, if we only consider GPS, we could miss stops happening when signal of the GPS is lost or willfully turned off by the client software. On the other hand, if we only consider Wi-Fi data, we omit stops when Wi-Fi records are not available. For example when the device

can't see a WLAN access point. Furthermore, we reported that Wi-Fi records had an accuracy close to the one of GPS records. For these reasons, we use both sensors that are complementary for the detection of stops.

Processing

The steps of the processing are the following:

1. **Extract available records**: For the 697 trips, we extract GPS and Wi-Fi records available inside the time windows. In the end, 175'000 records are extracted.
2. **Apply the algorithm**: The algorithm is launched 697 times (e.g. one time for each trip) with the following parameters: The distance between points is less than 150 meters, the time more than 300 seconds and the point density threshold is 20.
3. **Centroid and radius of the stops**: The centroid and radius of the cluster is computed for each stop according to the procedure employed in section 4.2.1.2.
4. **Duration of the stops**: Back to the records, in the same way than in section 4.2.3, the time spent inside a cluster is measured and extracted for each stop. Furthermore, we impose a minimum time threshold of 180 seconds inside a cluster to consider there is a stop.

Conclusion

In the end, the results are the following: Over 697 trips, 184 trips include one or more stops. During these 184 trips, a total of 234 different stops are detected and localized. If we take into account that a user can stop several times in the same place during the trip (e.g. for example if the user has to come back to get something he forgot), then we count 241 stops during these 184 trips.

We focus now according to our objective on direct trips. Hence, we omit 184 trips with stops from the estimation dataset. In the end, the estimation dataset is now composed of 513 trips.

To check the validity of these findings, we can have a look at the time recording in important states (e.g. stationary, outdoors, etc.) and for two situations (i) during the travel periods of the 697 trips that represent a total recording time of 400 hours

(e.g. when the user is not stopped according to our results) (ii) during the stops of the 697 trips that represent a total recording time of 100 hours (e.g. when the user is stopped). As the software triggers the recording states depending on different criteria of acceleration explained in Kiukkonen (2009), we can separate states that are more likely to be activated during travel periods in "blue" in the pies (outdoor state) and states that are more likely to be triggered during stops in "red" in the pies (outdoor without fix, stationary and urban stationary) (see Fig. 4.11 and Fig. 4.12).

In Fig. 4.11 and Fig. 4.12, we can see how long the phone has been recording in each state for both situations (e.g. (i) and (ii)).

We can reach the following conclusions:

- The outdoor state records when the user is in movement outdoors. In other words, this state is triggered when the user is traveling. Hence, it makes sense that the duration of the state forms the majority when the user is in the first situation. In the second pie, the percentage is low which makes sense as the participant is making a stop. Also, this percentage is not null because making a stop doesn't mean necessarily being stationary (For example if the user is shopping).

- The outdoor without fix is a mode where the GPS signal has been lost recently. According to (Zhang et al., 2010) who studied the situations where the Nokia N95 looses signal with the same integrated GPS, it is frequent to loose signal under concrete roofs or inside buildings whereas outdoors or in the country it is pretty infrequent. Hence, we can expect a high percentage in this state when the user is making a stop and probably going indoors which is consistent with the high percentage in the second pie.

- The percentage of the stationary increases in the second pie as expected.

- The known WLAN is triggered when a MAC address has been detected frequently enough. Typically, places where the participant goes frequently are places where we can expect a higher number of records within this state (e.g. grocery address, sport center, etc.). Here during stops, the percentage of this state remains very low which supports our initial assumption that stops are not necessarily points of interest.

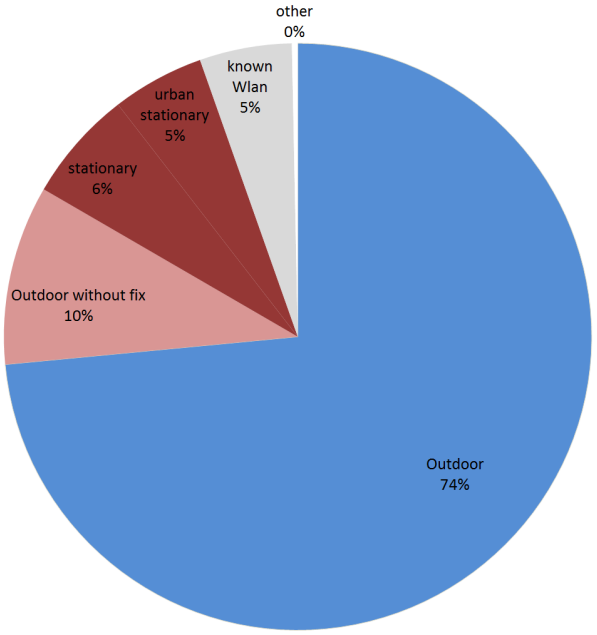

FIGURE 4.11: (i) Duration recording in important states when traveling (shares in %)

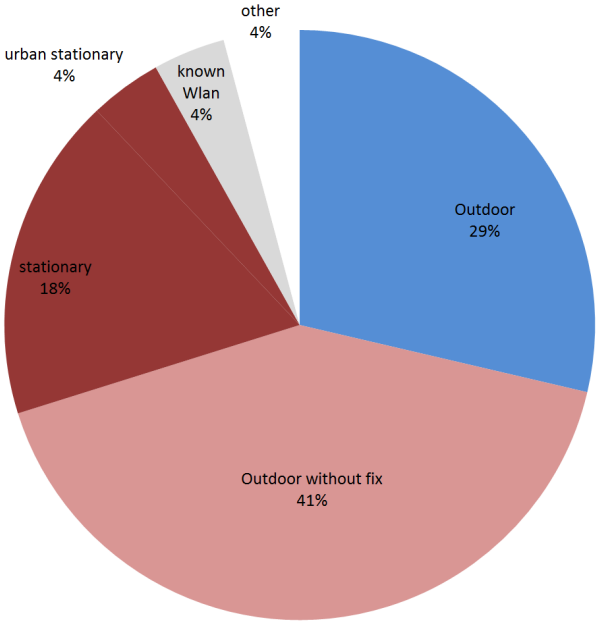

FIGURE 4.12: (ii) Duration recording in important states when stopped (shares in %)

4.2.7 Missing travel mode information for parts of the trips

Up to this point, we have identified the chosen mode with no regard that parts of the trips are actually missing. Figure 4.10 represents schematically a trip without stops where we can see these parts without travel mode information.

Hence, we measure the "quality" of the trip by defining the duration where the travel mode is known during the trip. To that end, we build a new variable called the *ratio_mode_time* that estimates this "accuracy". It is defined as the total duration where the travel mode is known over the duration of the trip. For one trip, the ratio is:

$$ratio\_mode\_time_{trip} = \frac{\sum_{j=1}^{N} t_j}{\text{Arrival } time \text{ - Departure } time} \tag{4.1}$$

where N is the number of map-matched trips during the trip and t_j is the duration of map-matched trip j.

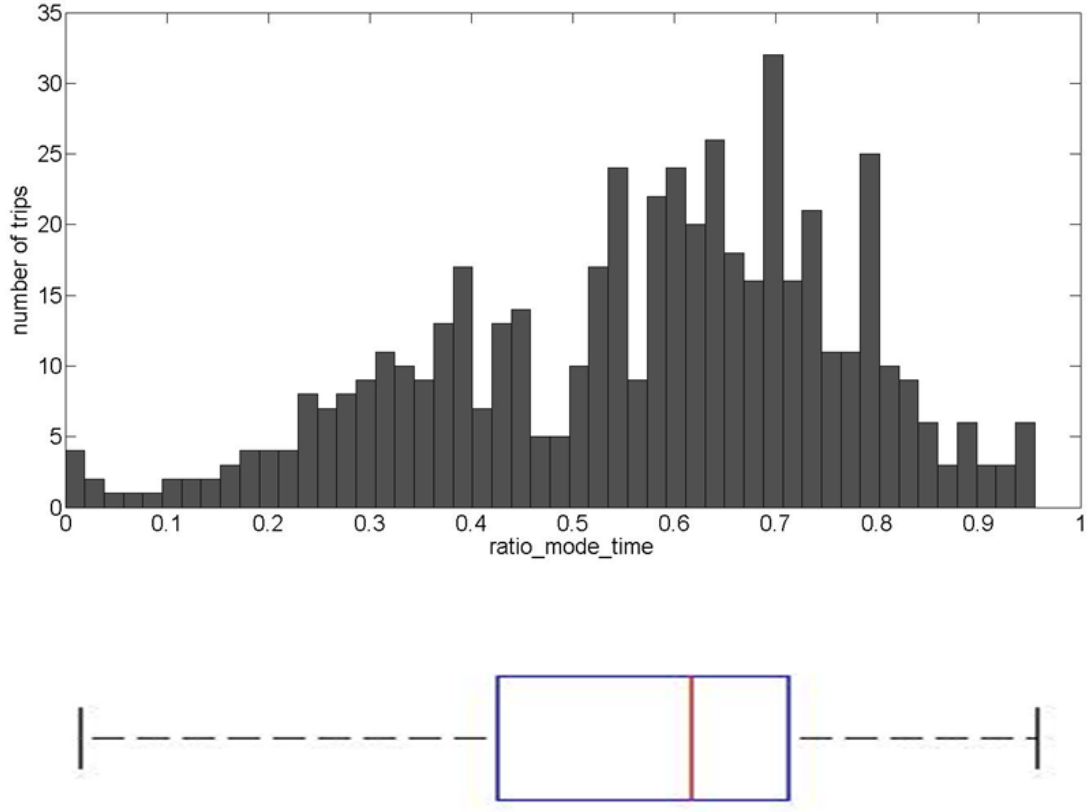

FIGURE 4.13: Ratio_mode_time histogram of the final trips' dataset

The histogram presented in Fig. 4.13 shows a median for all the trips a little above 0.6 meaning that for half of the trips we know the sequence of modes used for more than 60% of the trip duration. Therefore it is important to capture this relative "quality" of the observation in the model and give more weight to these observations close to 1.

4.2.8 Attributes of unchosen alternatives

For each trip, the attributes of the unchosen alternatives are not known and still have to be imputed. Travel time and distance are imputed manually with Google directions for the 26 users of the dataset. The same attributes are assigned for the way to work and for the way back as we assume the difference in time and distance is small between those trips. Furthermore these attributes are given for each user and do not depend neither on the time, day or month of the observation. Arbitrarily, we impute the attributes for the 1st February of 2010 at 09am.

Bike trips

For bike trips, Google directions makes a quite good estimation of the travel time between two locations. The algorithm indicates paths with designated bike lanes when they exist and tries to avoid areas with important climbing hills and drops. Moreover the estimation is made for a person with an average BMI [2] level pedaling at a normal speed. In the end, we impute the travel time and distance of the shortest path for the 26 home to work trips of our dataset.

Public transport trips

The Google directions site is not displaying the total distance (total distance for the combination of transport mode proposed plus the walking distance) so we have to implement a Javascript code based on the free Google directions API and described on the code report to impute it.

Car trips

For car trips, Google directions estimation of the travel time is not taking traffic into account. This is an important issue as we are modeling home to work trips that happen mainly during peak hours (see Fig. 4.9). For this reason, we apply the following procedure to include the traffic:

1. We impute the travel time and distance with Google directions (no estimation of the traffic).

[2] Body Mass Index (BMI): "measure of relative weight based on an individual's mass and height" (Wikipedia)

2. To model the traffic, we impute the same trips with Tom tom site that estimates the speed during peak hours based on millions of observations recorded from their in-vehicle GPS. The site estimates an additional time from 5 to 7 minutes in average in the Lausanne agglomeration which can be added to the initial Google directions estimation.

3. Finally, egress time to the vehicle is an important element to consider as the participant usually has to walk a bit before getting to his home and/or location. The solution we choose to estimate this duration is to analyze the real observations we have. Therefore, we filter all the trips where the main travel mode is the car with an important accuracy (e.g. ratio_mode_time higher than 0.7) and compute the average walking distance for the 20 trips found under these conditions. Hence, we find an average walking distance of 250 meters for these trips by car which corresponds to approximately 3 minutes assuming an average speed of 5 kilometers/hour.

In the end, we add 10 minutes for each home to work trip imputed with Google directions.

Chapter 5

Model specification and estimation

So far, our work has been focused in dealing with the challenges due to smartphone data. We assume now that the derived dataset is suitable for a mode choice analysis. Foremost, the objective is to manage estimating a consistent logit model which is already challenging.

5.1 Model expectations and behavioral hypothesis

The model expects answering to the questions with respect to workers' mode choice behavior when they go to work on weekdays. The analysis is conducted for home to work trips of Lausanne inhabitants, both living and working in the Lausanne agglomeration.

Intuition and literature on the topic have proved that distance and time are generally variables influencing the choice of one mode towards another. So we expect proving such influences of time and distance in our sample. Furthermore, it is assumed that older users have a disutility towards the soft mode alternative because of underlying comfort issues.

As explained in Chapter 2, smartphones have made possible to collect data over months. Hence we wish to analyze long term effects variable such as the effects of the seasons. We suggest that summer months have a positive effect on soft mode alternative and

inversely for winter months. Moreover, we assume that meteorology has an influence on the choice of soft modes. Indeed, previous studies have shown a decreasing share of bicycle trips in the Netherlands for temperatures under 0 °C whereas high temperatures up to 25 °C had positive effects on bike's utility (Sabir et al., 2010). Therefore both seasonal and meteorological effects on the use of soft modes are analyzed.

5.2 Model estimation

Maximum likelihood is used for the estimation. This standard technique estimates the unknown parameters based on the probability for the model to reproduce the whole sample as shown by equation 5.1 where k denotes the observation, i the alternative, n the individual , $P_{ni}^{k}\ (\beta, \gamma)$ is the probability to reproduce the observed choice, $\mathcal{K}$ corresponds to the number of observations, β are the unknown parameters associated with the utility function and γ are the unknown parameters associated with the multinomial logit model.

As explained in section 4.2.7, observations from our estimation dataset are not equally reliable. For some trips, the inferred choice might not be true when the identification of the chosen mode relies on a small part of the trip. With the creation of the *ratio_mode_time*, we are able to capture this reliability of the observation.

Hence, we associate a weight ω_k to each observation in order to adjust the relative importance of the observation according to its *ratio_mode_time*. The log-likelihood function is presented in equation 5.2 and the weights used are proportionally adjusted in order that their sum is equal to the number of observations $\mathcal{K}$ (Bierlaire, 2003).

$$\mathcal{L}^{*}(\beta, \gamma) = \prod_{k=1}^{\mathcal{K}} P_{ni}^{k}(\beta, \gamma) \tag{5.1}$$

$$\mathcal{L}(\beta, \gamma, \omega_k) = \sum_{k=1}^{\mathcal{K}} \omega_k \ln P_{ni}^{k}(\beta, \gamma) \tag{5.2}$$

The estimation is done by using the software package BIOGEME which allows for the estimation of such model (Bierlaire, 2003).

5.3 Model specification and estimation results

In the following paragraphs, PT corresponds to the public transport alternative and SM to the soft mode alternative.

Four models are specified and estimated within this section. The likelihood ratio test is used to test the model with the added variables (e.g. the unrestricted model) against the restricted one. It tests the hypothesis that the added parameters are equal to 0 and is defined as

$$-2(\mathcal{L}(\beta_R) - \mathcal{L}(\beta_U)) \tag{5.3}$$

where R denotes the restricted model and U the unrestricted one, $\mathcal{L}(\beta_R)$ is the final log-likelihood of the restricted model and $\mathcal{L}(\beta_U)$ the final log-likelihood of the unrestricted model. The test is χ^2 distributed to the number of restrictions $(\mathcal{K}_U - \mathcal{K}_R)$ of the model where $\mathcal{K}_U$ and $\mathcal{K}_R$ are the number of parameters of the unrestricted model, respectively the restricted model (Bierlaire, 2011). In the end, we test the null hypothesis at a 95% level of significance.

5.3.1 Base model - model with attributes of the alternatives

In the first model, we assume travel time is a factor influencing the choice of car and public transport. To test this idea, two time coefficients are introduced in PT utility $\beta_{TT_{PT}}$ and car utility $\beta_{TT_{car}}$ with TT_{PT} and TT_{car} the travel time of the PT alternative respectively, car alternative (see table 5.1).

Furthermore we assume the distance of the trip is an important factor influencing the choice of the SM utility: the shorter the trip the more attractive turns to be the SM alternative. Hence we add a distance coefficient to the SM utility $\beta_{distance\_SM}$ with $distance_{SM}$ the distance of the SM alternative (see table 5.1).

Values of the parameters are rescaled for numerical reasons so that the parameters are around 1 (Bierlaire, 2003). Hence time of public transport alternative and car alternative is divided by 10 and distance is converted in km. Results of the model are shown in table 5.2.

As expected, both alternative coefficients of time are significant and have negative signs, the larger absolute value for $\beta_{TT_{car}}$ indicates that people are more sensitive to time in case of private modes.

Moreover, $\beta_{distance\_SM}$ is also negative and significant which is in agreement with our behavioral hypothesis.

Furthermore, alternative specific constant of the soft mode alternative is significantly positive and higher than both PT and car constants. It means that, without consideration of variables, there is a natural turn towards the soft mode alternative.

5.3.2 Model 1 - adding socio-economics

Previous model only included variables that are attributes of the alternatives. Now we consider socio-economic dummy variable *senior* that corresponds to participants aged over 45 in the SM utility. The coefficient of this variable is denoted β_{senior} (see table 5.3).

The results are provided in table 5.4. As expected, participants aged over 45 have a higher disutility towards using the soft mode alternative. Moreover this variable is significant.

The likelihood ratio test consists in testing model 1 (e.g. unrestricted model) against the base model (restricted model). The test rejects the null hypothesis $\beta_{senior} = 0$ (see table 5.4).

5.3.3 Model 2 - adding seasonal variables

Both seasonal dummy variables *summer* that corresponds to the months of July, August and *winter* that corresponds to the months of December, January, February are added to the SM utility. These variables are respectively denoted β_{summer} and β_{winter} (see table 5.5).

Seasonal *summer* variable has a positive sign and is significant. This result is consistent as we can expect that being in summer is a motivation to use soft modes. However *winter* variable is not significant, which means cold months of the year are not affecting the mode choice of our 26 participants when they go to work.

The likelihood ratio test rejects the hypothesis: $\beta_{summer} = \beta_{winter} = 0$ (see table 5.6)

5.3.4 Model 3 - adding aggregated meteorological variables

In the previous model, we considered seasonal variables in our model and proved an influence of the *summer* variable towards using the SM alternative. For this reason we assume that aggregated meteorological variables could also have an influence on the use of this alternative.

Weather variables are associated to trips based on the assumption that the worker base his decision on the weather conditions prevailing the hour before departure time.

Two aggregated dummy variables are considered. (i) the dummy variable *good_day* that corresponds to a day with temperatures between 20 °C and 30 °C, no precipitation during 24 hours and a gentle breeze as the worst situation of wind speed (e.g. we set an hourly maximum 10 minutes wind speed of the past 12 hours lower than 5 meters/second).

(ii) the dummy variable *bad_day*, on the contrary, corresponds to a day with temperatures lower than 5 °C, precipitations during the last 24 hours and a strong breeze (over 14m/s).

These variables are added to the SM utility and their coefficients are respectively denoted: β_{good_day} and β_{bad_day} (see table 5.7).

The results show that *good_day* is not significant which means good weather conditions are not affecting the choice of the SM alternative (see table 5.8). Unfortunately, only one observation corresponds to a day with bad day conditions. That's why the value estimated of the *bad_day* is not relevant.

The likelihood ratio test that tests the hypothesis $\beta_{good\_day} = \beta_{bad\_day} = 0$ can't be rejected. Hence we reject the unrestricted model.

The same model is tested with different definitions of the variables *good_day* and *bad_day*, for example by changing the temperature value threshold or the wind value threshold. However, variables are still not significant with values of the t-test always close to 0. Furthermore, for all these situations, the likelihood ratio test never rejects the null hypothesis of β_{good_day} and β_{bad_day}.

5.4 Analysis of the results

The results found in the previous section are compared to the Swiss Microcensus 2010 (OFS et al., 2010) [1]. In this study, a questionnaire is submitted to Swiss workers to know what are the reasons motivating their mode choice when they go to work given that several mentions can be reported (see Fig. 5.1). The study has been conducted for 2835 tours where the soft mode is chosen, 6076 when private motorized mode is chosen and 1608 tours when the public transport alternative is chosen. Figure 5.1 shows the result of the questionnaire. The following conclusions are reached:

Soft mode alternative

- The main reason motivating workers to choose the soft mode alternative is a short distance from their home location to work location. In the base model, the results indicate that a shorter distance is indeed a reason motivating the choice of soft modes.
- In the survey, none of the respondents have mentioned the weather conditions as a reason motivating their choice towards the SM alternative. In model 3 we have reached the same conclusions, both aggregated meteorological variables *good_weather* and *bad_weather* are not significant and the likelihood ratio test did not reject the null hypothesis of both coefficients.
- "boardwalk" and "pleasure of traveling" represent together 10% of the reasons of choosing soft modes in the survey when participants have chosen SM alternative. Model 2 indicates that summer months have an influence on the choice of the SM alternative. Maybe, there is a link between the two observations which results is saying that summer months increase the pleasure of traveling more than other months of the year.
- Model 1 have indicated that participants aged over 45 are less attracted by soft modes. The reason underlying is that older people are more sensitive to the lack of comfort of an alternative. This reason is also confirmed with the category "comfort" in the table that is mentioned as a reason to choose the car and PT alternatives but not the SM alternative.

[1] Federal Statistical Office (OFS)

Car alternative and PT alternative

- The main reason motivating workers to use car alternative is the travel time. The results of the base model indicate that indeed travel time has an influence towards the choice of the car. The larger absolute value for the time coefficient in the car alternative compare to the time coefficient of the PT alternative indicates that people are more sensitive to time for private motorized modes, although this difference is small. The survey also reports that more participants are sensitive to the travel time when they travel by car than by PT (38% of the answers against 19% respectively).

- We assumed that distance was not a factor influencing the choice of car and PT alternatives. The questionnaire reaches the same conclusions with the category "short trip" that is not reported as a reason motivating neither the choice of the car or PT alternative.

- The main reason reported by the participants when they choose the PT alternative is the lack of alternatives. As we are missing the socio-economics possession of a car license or possession of bike/car in the household, we are not able to capture this element in the model.

5.5 Model conclusions

The results of the models are consistent with intuition and reach the same conclusions than OFS et al. (2010). Important variables such as distance and time are consistent as well as socio-economic variable age. Interestingly, whereas summer season motivates users towards the SM alternative, weather variables are not influencing the use of soft modes . We found rational explications of this difference with Sabir et al. (2010): The authors show effects of temperature and wind on the choice of the bike but not walking where he reports very small effects of the weather variables. Therefore, it may not be relevant to compare the results as we have grouped walking and biking in the soft mode alternative. Furthermore, their study was conducted in Netherlands and therefore we can expect different sensitivities to weather between Swiss and Dutch population. Also, their analyze include different activity based trips such as commuting trips, recreational

	Moyen de transport choisi		
	MD	TIM	TP
Disponibilité d'une voiture/d'un motocycle ou possession d'un abonnement	-	10,3	12,5
Manque d'alternatives	7,1	17,6	37,6
Temps de trajet	-	38,2	19,3
Trajet court	65,5	-	-
Coût	4,6	1,8	10,1
Plaisir de voyager	7,1	1,2	5,3
Motivations écologiques	5,5	-	6,7
Transport de bagages, objets encombrants	-	6,6	0,1
Confort	-	17,5	13,8
Raisons de santé	14,8	0,8	1,0
Conditions météorologiques	-	5,5	3,8
Disponibilité d'une place de stationnement à destination	2,6	0,9	9,1
Bonne offre TP	2,6	21,0	15,1
Promenade	3,0	-	-
Autres	15,5	18,2	15,0

Base: 2835 boucles (MD), 6076 boucles (TIM), 1608 boucles (TP)

FIGURE 5.1: Reasons of mode choice to go working (in %, several mentions are possible). MD corresponds to Soft mode, TIM corresponds to Private motorized mode ant TP to public transport (OFS et al., 2010)

and sport trips or educational trips whereas we are focused on home to work trips. Thus, we find more relevant the comparison with OFS et al. (2010) that has carried out the questionnaire for the same activity purpose and population, and reached the same conclusions.

Although the model is satisfying and reaches our expectations, we had to make assumptions to overcome some issues. First, we assumed that all the workers had a car, a bike and a driver license except for one user that reported he didn't have a car. Furthermore, we assumed that a linear weighting based on the *ratio_mode_time* of the observation would rigorously adjust its relative importance in the model. However, more subtle distribution might be more appropriate to adjust the importance of the observation with small *ratio_mode_time* even more penalized and high *ratio_mode_time* put forward.

Furthermore, by choosing the multinomial logit model, we assumed the model to be homoscedastic across individuals and to have equal variances across alternatives. Therefore we assume the choice of each observation to be independent of other choices. However, this assumption might not be true in some situations. Indeed, we assumed that the way

to work and the way back of a user were two independent observations. Although it is rare to observe the outbound and return of a user the same day, it happens 45 times and in this cases the mode choice of the return might be correlated to the mode choice of the way to work. Furthermore, the snowball enrollment sampling as reported in 3.2 creates natural social connections in the sample and 16 of the participants over 26 are working at EPFL. Therefore, mode choices of these users might be correlated in some situations (e.g. for example if the users work in the same laboratory).

TABLE 5.1: Specification table of the utilities of the base model

	MNL model		
	V_{PT}	V_{car}	V_{SM}
Alternative specific coefficients			
ASC_{SM}	-	-	1
ASC_{car}	-	1	-
Attributes of the alternatives			
$\beta_{distance\_SM}$	-	-	$distance_{SM}$
$\beta_{TT_{PT}}$	TT_{PT}	-	-
$\beta_{TT_{car}}$	-	TT_{car}	-

TABLE 5.2: Estimation results of the base model

Parameter number	Description	Coeff. estimate	std. error	t-stat	p-value
1	ASC_{SM}	1.95	0.798	2.44	0.01
2	ASC_{car}	0.0436	1.27	0.03	0.97
3	$\beta_{\text{distance\_SM}}$	-0.531	0.119	-4.46	0.01
4	$\beta_{\text{time\_PT}}$	-1.23	0.316	-3.90	0.00
5	$\beta_{\text{time\_car}}$	-1.24	0.495	-2.51	0.01

Summary statistics

Number of observations = 513

$$\begin{aligned} \mathcal{L}(0) &= -557.008 \\ \mathcal{L}(\hat{\beta}) &= -240.630 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 632.757 \\ \rho^2 &= 0.568 \\ \bar{\rho}^2 &= 0.559 \end{aligned}$$

TABLE 5.3: Specification table of the utilities of the model 1

	MNL model		
	V_{PT}	V_{car}	V_{SM}
Alternative specific coefficients			
ASC_{SM}	-	-	1
ASC_{car}	-	1	-
Attributes of the alternatives			
$\beta_{distance\_SM}$	-	-	$distance_{SM}$
$\beta_{TT_{PT}}$	TT_{PT}	-	-
$\beta_{TT_{car}}$	-	TT_{car}	-
Socio-economics			
β_{senior}	-	-	$senior$

TABLE 5.4: Estimation results of the model 1

Parameter number	Description	Coeff. estimate	std. error	t-stat	p-value
1	ASC_{SM}	0.963	0.827	1.16	0.24
2	ASC_{car}	0.332	1.17	0.28	0.78
3	β_{senior}	-1.87	0.409	-4.58	0.00
4	$\beta_{\text{distance\_SM}}$	-0.164	0.133	-1.23	0.22
5	$\beta_{\text{time\_PT}}$	-1.06	0.298	-3.55	0.00
6	$\beta_{\text{time\_car}}$	-1.22	0.453	-2.70	0.01

.

Summary statistics

Number of observations = 513

$$\begin{aligned} \mathcal{L}(0) &= -557.008 \\ \mathcal{L}(\hat{\beta}) &= -228.864 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 656.289 \\ \rho^2 &= 0.589 \\ \bar{\rho}^2 &= 0.578 \end{aligned}$$

Likelihood ratio test

Restricted model

Number of parameters = 5

Final log-likelihood = -240.630

Unrestricted model

Number of parameters = 6

Final log-likelihood = -228.864

$$\begin{aligned} -2[\mathcal{L}(\beta_R) - \mathcal{L}(\beta_U)] &= 23.532 \\ \chi^2_{0.95,1} &= 3.84 \end{aligned}$$

The restricted model is rejected

TABLE 5.5: Specification table of the utilities of the model 2

	MNL model		
	V_{PT}	V_{car}	V_{SM}
Alternative specific coefficients			
ASC_{SM}	-	-	1
ASC_{car}	-	1	-
Attributes of the alternatives			
$\beta_{Distance\_SM}$	-	-	$Distance_{SM}$
$\beta_{TT_{PT}}$	TT_{PT}	-	-
$\beta_{TT_{car}}$	-	TT_{car}	-
Socio-economics			
β_{senior}	-	-	$senior$
Seasonality variables			
β_{summer}	-	-	$summer$
β_{winter}	-	-	$winter$

TABLE 5.6: Estimation results of the model 2

Parameter number	Description	Coeff. estimate	std. error	t-stat	p-value
1	ASC_{SM}	0.883	0.842	1.05	0.29
2	ASC_{car}	0.380	1.17	0.32	0.75
3	β_{senior}	-1.89	0.408	-4.63	0.00
4	$\beta_{\text{distance\_SM}}$	-0.176	0.133	-1.32	0.19
5	β_{summer}	0.936	0.401	2.34	0.02
6	$\beta_{\text{time\_PT}}$	-1.05	0.303	-3.46	0.00
7	$\beta_{\text{time\_car}}$	-1.23	0.453	-2.71	0.01
8	β_{winter}	-0.0592	0.349	-0.17	0.87

Summary statistics

Number of observations = 513

$$\begin{aligned} \mathcal{L}(0) &= -557.008 \\ \mathcal{L}(\hat{\beta}) &= -225.101 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 663.815 \\ \rho^2 &= 0.596 \\ \bar{\rho}^2 &= 0.582 \end{aligned}$$

Likelihood ratio test

Restricted model

Number of parameters = 6

Final log-likelihood = −240.516

Unrestricted model

Number of parameters = 8

Final log-likelihood = −225.101

$$\begin{aligned} -2[\mathcal{L}(\beta_R) - \mathcal{L}(\beta_U)] &= 8.99 \\ \chi^2_{0.95,1} &= 7.528 \end{aligned}$$

The restricted model is rejected

TABLE 5.7: Specification table of the utilities of the model 3

	MNL model		
	V_{PT}	V_{car}	V_{SM}
Alternative specific coefficients			
ASC_{SM}	-	-	1
ASC_{car}	-	1	-
Attributes of the alternatives			
$\beta_{distance\_SM}$	-	-	$distance_{SM}$
$\beta_{TT_{PT}}$	TT_{PT}	-	-
$\beta_{TT_{car}}$	-	TT_{car}	-
Socio-economics			
β_{senior}	-	-	$senior$
Meteorological variables			
β_{good_day}	-	-	$good\_day$
β_{bad_day}	-	-	bad_day

TABLE 5.8: Estimation results of the model 3

Parameter number	Description	Coeff. estimate	std. error	t-stat	p-value
1	ASC_{SM}	0.974	0.834	1.17	0.24
2	ASC_{car}	0.376	1.18	0.32	0.75
3	β_{senior}	-1.88	0.411	-4.58	0.00
4	$\beta_{\text{bad\_day}}$	0.00	1.80e+308	0.00	1.00
5	$\beta_{\text{distance\_SM}}$	-0.164	0.134	-1.23	0.22
6	$\beta_{\text{good\_day}}$	0.389	0.506	0.77	0.44
7	$\beta_{\text{time\_PT}}$	-1.04	0.300	-3.48	0.00
8	$\beta_{\text{time\_car}}$	-1.22	0.452	-2.70	0.01

Summary statistics

Number of observations = 513

$$\begin{aligned} \mathcal{L}(0) &= -557.008 \\ \mathcal{L}(\hat{\beta}) &= -228.537 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 656.943 \\ \rho^2 &= 0.590 \\ \bar{\rho}^2 &= 0.575 \end{aligned}$$

Likelihood ratio test

Restricted model

Number of parameters = 6

Final log-likelihood = −228.864

Unrestricted model

Number of parameters = 8

Final log-likelihood = −228.537

$$\begin{aligned} -2[\mathcal{L}(\beta_R) - \mathcal{L}(\beta_U)] &= 1 \\ \chi^2_{0.95,1} &= 0.654 \end{aligned}$$

The restricted model is accepted

Chapter 6

Conclusion

6.1 Summary

Smartphones consist of a tremendous source of data collection provided with a rich set of sensors (GPS, Wi-Fi, Bluetooth, GSM, etc.) and an overall good recording accuracy for GPS and Wi-Fi. These devices have significantly reduced the cost and burden of both surveying companies and participants by way of a data that is automatically collected and stored in digital format. These improvements have enabled the recording of longitudinal data by stretching out data collection campaigns over several months without fatigue effects.

This project has investigated the challenges of using smartphone data for mode choice modeling. These challenges included the detection of trips, activities and identification of the trip purpose, the determination of the chosen mode and missing attributes of the unchosen alternative. Given that GPS data was sparse, because of energy saving issues of the device, more energy friendly Wi-Fi sensor, became a good alternative to GPS when it was not available. Moreover, much more Wi-Fi records were collected than GPS in places where the user goes frequently, which is especially the case for home and work. Therefore, the identification of the trip purpose and the trip detection that was relying on these home and work locations could be inferred only with Wi-Fi data. The detection of stops relying on the available records during trips required both GPS and Wi-Fi data. To obtain the travel mode used during the trips, we included the results of the multimodal map-matching algorithm proposed by Chen (2013). The difficulty

of detecting the mode with the sparse GPS dataset resulted in parts of trip where travel modes could not be inferred. The issue was therefore addressed by weighting the observations according to the percentage of the trip where the travel mode was known. Issues specific to the Nokia dataset that were the missing socio-economics, or the missing train trips, due to the unreported train network in the open street map of Lausanne, might not remain in future data collection campaigns but had to be overcome in this project too.

The issues identified were addressed in order to build a dataset suitable for the mode choice model developed. Besides the derived smartphone dataset, other data sources were utilized to impute missing attributes of the alternatives (Google directions), to add characteristics of the participants (demographic questionnaire) and to include long term effects of meteorology on the mode choice (MeteoSwiss weather archives).

Different hypothesis were tested in the process of building the model resulting in a model specification that gives results consistent with intuition and literature. Travel time when increasing had a negative impact on the choice of car and public transport alternative whereas distance when decreasing was a reason motivating participants toward the soft mode alternative. Socio-economic age was also consistent with our expectations: participants aged over 45 had a disutility towards the soft mode alternative which we assumed was due to the lack of comfort of this alternative compare to car and public transports. Summer season had a positive effect on the number of participants choosing the soft mode. Interestingly, we concluded that meteorological variables were not influencing the mode choices of workers for their home to work trips which was also reported in OFS et al. (2010).

In a context where smartphones are revolutionazing the way data is collected, we showed the feasibility and potential of using smartphone data in the context of discrete choice analysis.

6.2 Directions for future research

In this thesis, results of the multimodal map-matching from Chen (2013) are used to include travel mode information on the home to work trips. An interesting undertaking would be to use Hemminki et al. (2013) mode detection algorithm to infer the modes and

compare the results to the map-matched trips utilize in this project. The proposed mode detection requires to calculate gravity projections of vertical and horizontal acceleration, available in much more states than GPS, and estimated with the robust method proposed in Hemminki et al. (2013). Although the author reports a better precision and recall of the train detection with his algorithm, the android smartphone utilized in their study is more recent than the Nokia N95 of our data collection campaign. We wonder how the algorithm would work with a lower accuracy of the collected accelerations.

A possible addition to the model would be to add trips with stops. To do so, a threshold needs to be defined first in order to separate stops on the way to work and those that require a detour. On the one hand, for the stops on the way to work, we assume we can use the same level-of-service attributes (e.g. time, distance) than those home to work without additional stops. Finally, we represent in the model the presence or absence of intermediate stops by a binary variable. This a typical practice with activity-based approach (Vij and Akshay, 2013). On the other hand, for the stops requiring a detour, we would have to impute the attributes of the unchosen alternative. To do so, Google directions allows to add way points in the way from home to work. The main difficulty would be to automatize the imputation for all the trips.

Finally, the issues specified and the strategies addressed in this thesis remain the same if we focus on other activity based trips. Home and university clusters were found for 33 students. It is therefore possible to analyze home to university trips with the available data. Furthermore, Buisson et al. (2014) identifies more points of interest such as shopping or grocery address based on a probabilistic approach. More activity based trips can therefore be considered.

Bibliography

[1] B. Atasoy, A. Glerum, and M. Bierlaire. Attitudes towards mode choice in switzerland. Technical Report TRANSP-OR 110502, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, 2011.

[2] K.W. Axhaussen, S. Schönfelder, J. Wolf, M. Oliveira, and U. Samaga. 80 weeks of gps-traces: approaches to enriching the trip information. In *83rd Transportation Research Board Meeting*, 2003.

[3] Battelle and U.S. Federal Highway Administration. *Final report [electronic resource] : personal travel unit (PTU) development and initial testing : to Federal Highway Administration / by Transportation Division.* The Administration, 2000.

[4] M. Ben-Akiva. Structure of passenger travel demand models. Technical report, M.I.T, 1973.

[5] M. Bierlaire. Biogeme: a free package for the estimation of discrete choice models. In *Proceedings of the Swiss Transport Research Conference*, Ascona, Switzerland, 2003.

[6] M. Bierlaire. Statistical tests. Slides from the course "Decision-aid methodologies in transportation", École Polytechnique Fédérale de Lausanne, 2011. URL http://transp-or.epfl.ch/courses/decisionAid2011/slides/06-tests.pdf.

[7] A. Buisson, A. Danalet, and E. Kazagli. Identify user's locations of interest from smartphone wifi data. Technical report, École Polytechnique FÉdÉrale de Lausanne, 2013.

[8] A. Buisson, A. Danalet, and E. Kazagli. Individual activity-travel analysis based on smartphone wifi data. Technical report, École Polytechnique FÉdÉrale de Lausanne, 2014.

[9] J Casas and C.H Arce. Trip reporting in household travel diaries: A comparison to gps-collected data. In *78th Annual Meeting of the Transportation Research Board*, Washington, DC, 1999.

[10] J.M. Casello, A.Nour, K.C. Rewa, and J. Hill. An analysis of stated preference and gps data for bicycle travel forecasting. In *78th Annual Meeting of the Transportation Research Board*, Washington, DC, 2011.

[11] J. Chen. *Modeling Route Choice Behavior Using Smartphone Data*. PhD thesis, Polytechnic School of Lausanne, Lausanne, Switzerland, January 2013.

[12] R.T Dizaji. *Acquiring Multimodal Disaggregate Travel behavior Data using Smart Phones*. PhD thesis, University of Waterloo, 2012.

[13] S.T Doherty, N. Noel, M.L. Gosselin, C. Sirois, and M. Ueno. Moving beyond observed outcomes integrating global positioning systems and interactive computer-based travel behavior surveys. Technical report, Groupe de Recherche Interdisciplinaire Mobilitvironnement et SritRIMES) Drtement dAmgement Universitval, 2001.

[14] C.C. Doyle, F.C. Pereira, F. Zhao, I. Dias, H.B. Lim, M. Ben-Akiva, and C. Zegras. The future mobility survey: Experiences in developing a smartphone-based travel survey in singapore. 2013.

[15] G. Draijer, N. Kalfs, and J.Perdok. Gps as a data collection mehod for travel research. In *79th Annual Meeting of the Transportation Research Board*, Washington, D.C, January 2000.

[16] N. Eagle and A. Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006. ISSN 1617-4909. doi: 10.1007/s00779-005-0046-3. URL http://dx.doi.org/10.1007/s00779-005-0046-3.

[17] M. Ester, H. Kriegel, S. Jorg, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[18] H. Fernee, N. Sonck, and A. Scherpenzeel. Data collection with smartphones: experiences in a time use survey. Technical report, The Netherlands Institute for Social Research and Centerdata, Tilburg University, December 2012.

[19] M. Friederich, M. Nikolic, E. Kazagli, and M. Bierlaire. Exploration of smartphone users trips data to investigate travel behavior. Technical report, École Polytechnique Fédérale de Lausanne, 2014.

[20] S. Hemminki, P. Nurmi, and S.Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 13:1–13:14, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2027-6. doi: 10.1145/2517351.2517367. URL http://doi.acm.org/10.1145/2517351.2517367.

[21] R.D. Jong and W. Mensonides. Wearable gps device as a data collection method for travel research. Technical report, The University of Sydney and Monash University, February 2003.

[22] N. Kiukkonen. Technical report: Data collection campaign. Technical report, Nokia Research Center, December 2009.

[23] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS,',','), Berlin.*, 7 2010.

[24] T. Lasky, K. Yen, S. Donecker, K. Yan, T. Swanston, A. Adamu, L. Gallagher, M. Assadi, and B. Ravani. Development of vehicular and personal universal longitudinal travel diary systems using gps and new technology. Final Report UCD-ARR-06-12-31-01, AHMCT, UC Davis, 2006. URL http://ahmct.ucdavis.edu/pdf/UCD-ARR-06-12-31-01.pdf.

[25] E. Murakami and D.P. Wagner. Can using global positioning system (gps) improve trip reporting? In *Transportation Research Part C: Emerging Technologies 7 (2)*, pages 149–165, 1999.

[26] E. Murakami, D.P. Wagner, and D.M. Neumeister. Using global positioning systems and personal digital assistants for personal travel surveys in the united states. In *International Conference on Transport Survey Quality and Innovation*, Grainau, Germany, 1997.

[27] P. Nitsche, P. Widhalm, S. Breuss, and P. Maurer. A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. *Procedia - Social and Behavioral Sciences*, 48(0):1033 – 1046, 2012. ISSN 1877-0428. doi: http://dx.doi.org/10.1016/j.sbspro.2012.06.1080. URL http://www.sciencedirect.com/science/article/pii/S1877042812028169. Transport Research Arena 2012.

[28] OFS, Office fédéral du développement territorial (ARE), Collectivité de travail Planidea, and Haute école de Lucerne. *La mobilité en Suisse Résultats du microrecensement mobilité et transports 2010*. Office fédéral de la statistique (OFS), 2010.

[29] J. Ortuzar and L.G. Willumsen. *Modeling Transport, New York*. John Wiley and Sons, 1994.

[30] F. Pereira, C. Carrion, F. Zhao, C.D. Cottrill, C. Zegras, and M. Ben-Akiva. The future mobility survey: Overview and preliminary evaluation. In *Eastern Asia Society for Transportation Studies*, ISSN:1881-1132, page 31, Taipei, Taiwan, 2013.

[31] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Trans. Sen. Netw.*, 6(2): 13:1–13:27, March 2010. ISSN 1550-4859. doi: 10.1145/1689239.1689243. URL http://doi.acm.org/10.1145/1689239.1689243.

[32] M. Sabir, M.J. Koetse, and P. Rietveld. The impact of weather conditions on mode choice: Empirical evidence of the netherlands. Technical report, Vrije Universiteit, Amsterdam, 2010.

[33] S. Saneinejad. Modelling the impact of weather conditions on active transportation travel behaviour. Master's thesis, University of Toronto, 2010.

[34] S. Schönfelder, K. W. Axhausen, N. Antille, and M. Bierlaire. Exploring the potentials of automatically collected gps data for travel behaviour analysis - a swedish data source. *GI-Technologien für Verkehr und Logistik*, (13):155–179, 2002.

[35] N. Schüssler. *Accounting for similiraties between alternatives in discrete choice models based on high-resolution observations of transport behaviour*. PhD thesis, ETH Zurich, 2010.

[36] N. Silberhorn, Y. Boztuğ, and L. Hildebrandt. Estimation with the nested logit model: specifications and software particularities. SFB 649 discussion paper 2006,017, Berlin, 2006. URL http://hdl.handle.net/10419/25100.

[37] P. Stopher and A. Collins. Conducting a gps prompted recall survey over the internet. In *Transportation Research Board Annual Meeting, 84th*, Washington, DC, 2005.

[38] D. Vautin and J. Walker. Transportation impacts of information provision & data collection via smartphones. In *Transportation Research Board 90th Annual Meeting*, page 21p, Washington, DC, 2011.

[39] Vij and Akshay. Incorporating the influence of latent modal preferences in travel demand models. University of california transportation center, working papers, University of California Transportation Center, 2013. URL http://EconPapers.repec.org/RePEc:cdl:uctcwp:qt7ng2z24q.

[40] S. Vlassenroot, D. Gillis, R.. Bellens, and S. Gautama. The use of smartphone applications in the collection of travel behaviour data. *International Journal of Intelligent Transportation Systems Research*, pages 1–11, 2014. ISSN 1348-8503. doi: 10.1007/s13177-013-0076-6. URL http://dx.doi.org/10.1007/s13177-013-0076-6.

[41] S. Von Watzdorf and F. Michahelles. Accuracy of positioning data on smartphones. In *Proceedings of the 3rd International Workshop on Location and the Web*, LocWeb '10, pages 2:1–2:4, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0412-2. doi: 10.1145/1899662.1899664. URL http://doi.acm.org/10.1145/1899662.1899664.

[42] J. Wolf. Applications of new technologies in travel surveys. In *International Conference on Transport Survey Quality and Innovation*, Costa Rica, August 2004.

[43] J. Wolf, R. Guensler, and W. Bachman. Elimination of the travel diary: An experiment to derive trip purpose from gps travel data. In *Transportation Research Board 80th Annual Meeting*, Washington, D.C, January 7-11 2001.

[44] Jean Wolf, Directed Dr, and Randall Guensler. Using gps data loggers to replace travel diaries in the collection of travel data. In *Dissertation, Georgia Institute of Technology, School of Civil and Environmental Engineering*, pages 58–65, 2000.

[45] L. Yalamanchili, R.M. Pendyala, N. Prabaharan, and P. Chakravarthy. Analysis of global positioning system-based data collection methods for capturing multistop trip-chaining behavior. *Transportation Research Record: Journal of the Transportation Research Board*, pages 58–65, 1999.

[46] J. Zhang, B. Li, A.G. Dempster, and C. Rizos. Evaluation of high sensitivity gps receivers. In *2010 International Symposium on GPS/GNSS*, Taipei, Taiwan, October 2010.

Appendix A

UML of the postgreSQL dataset

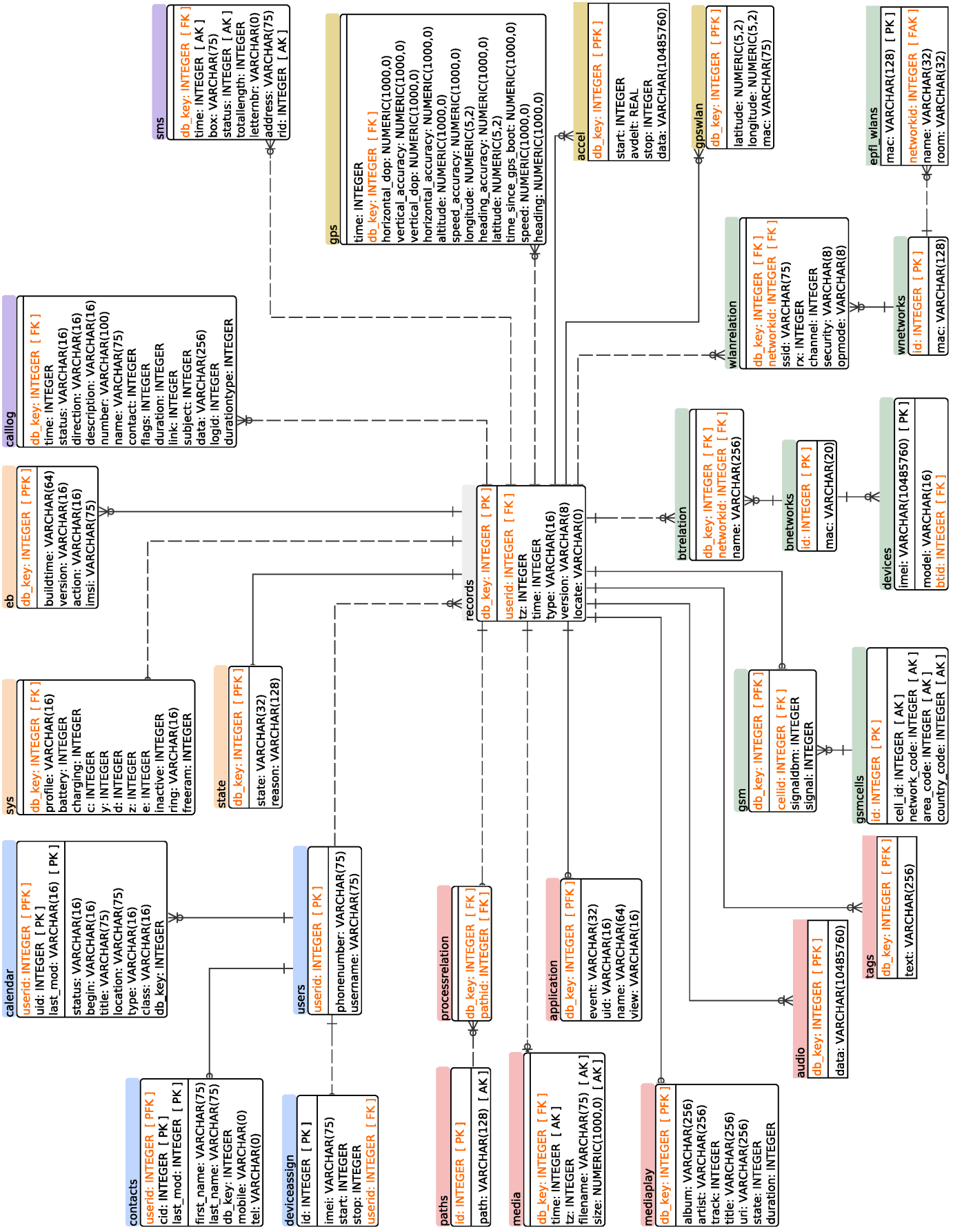

FIGURE A.1: UML of the postgreSQL dataset

Appendix B

Home and work locations

TABLE B.1: Home and work centroids and radius of the clusters (in m).

user_id	cluster_meaning	longitude	latitude	radius of the cluster (in m)
5924	home	7.523293	46.28882	258.0916967
5924	work	7.084619	46.1092	107.8152574
5936	home	6.895996	46.44344	150.6873497
5936	work	7.084849	46.10923	132.9488459
5937	home	6.630193	46.53033	83.30751895
5937	work	7.084576	46.10912	104.4747086
5939	home	6.819972	46.62052	0
5939	work	6.63733	46.52315	101.5179035
5940	home	6.311243	46.5105	0
5940	work	6.63736	46.52314	204.5098231
5943	home	6.631548	46.51545	70.61950876
5943	work	6.567333	46.5197	181.1386306
5944	home	6.620199	46.51614	126.4293523
5944	work	6.567345	46.51977	215.9088719
5947	home	7.076906	46.10508	198.1335933
5947	work	7.084777	46.10924	124.913205
5948	home	6.640886	46.52252	123.3479796
5948	work	7.084807	46.10927	127.9819943
5949	home	6.615003	46.5244	84.49520008
5949	work	6.566843	46.52009	194.4122018
5950	home	6.614928	46.52442	151.4867214
5950	work	6.562258	46.52039	28.89971401
5954	home	7.062187	46.1011	0
5954	work	7.084632	46.10914	110.2112184
5957	home	6.574886	46.53878	295.1086219
5957	work	6.613315	46.52554	154.5340233
5959	home	6.641249	46.54257	305.4413768
5959	work	6.637277	46.52066	159.4024397
5960	home	7.080495	46.10762	44.42996545
5960	work	7.084792	46.10916	127.5146904
5962	home	7.065362	46.09876	205.2689882
5962	work	7.084743	46.10922	121.3573434
5963	home	6.642327	46.51812	61.83491251
5963	work	6.60199	46.53107	69.1989583
5965	home	7.007197	46.21874	116.0794471
5965	work	7.084721	46.10924	118.7431129
5967	home	7.005435	46.21851	235.1900731
5967	work	7.525789	46.29166	75.29045296
5972	home	6.637967	46.52063	140.3921941
5972	work	6.126067	46.20481	77.7208435
5974	home	6.652215	46.51292	not computed
5974	work	6.562811	46.51767	102.0914176
5976	home	7.080507	46.10762	45.75048744
5976	work	6.579618	46.54238	170.2875748
5979	home	6.617231	46.51492	128.991973
5979	work	6.566166	46.51753	122.5939474
5980	home	6.630314	46.52923	104.3008325
5980	work	6.566464	46.51778	159.1926133
5987	home	7.078595	46.10541	126.7853388
5987	work	7.08477	46.10923	124.1993796

user_id	cluster_meaning	longitude	latitude	radius of the cluster (in m)
5990	home	7.076711	46.10495	178.7414235
5990	work	7.084679	46.10916	115.0070687
6001	home	6.590114	46.53502	92.77101157
6001	work	6.628465	46.51742	222.8794084
6002	home	6.598915	46.53169	227.7281886
6002	work	6.56186	46.51949	33.10932296
6007	home	6.598626	46.53103	168.588097
6007	work	6.561842	46.51947	32.67086825
6010	home	6.60658	46.53764	153.1955353
6010	work	6.566207	46.51752	125.2251507
6014	home	7.435741	46.94392	117.8579841
6014	work	7.390008	46.94708	109.6857032
6015	home	7.524108	46.28944	370.3446508
6015	work	7.883456	46.28807	244.2643556
6023	home	6.338247	46.45788	65.7088018
6023	work	6.562986	46.51732	108.5371361
6029	home	6.616809	46.52666	120.7703948
6029	work	17.94551	59.40032	140.5906313
6039	home	6.565987	46.53194	75.77976624
6039	work	6.566258	46.52069	257.8189184
6040	home	6.627647	46.52462	167.0383439
6040	work	6.570846	46.55641	173.0212644
6069	home	6.565487	46.5317	-
6069	work	6.56237	46.51772	112.1921382
6075	home	6.598813	46.53321	163.6585276
6075	work	6.568003	46.51724	98.73777306
6077	home	6.578304	46.53958	294.5302333
6077	work	6.566151	46.52068	161.1362298
6085	home	8.52889	47.41631	86.12744709
6085	work	6.562459	46.51768	125.7305892
6090	home	6.574169	46.53873	175.9824837
6090	work	6.56656	46.52071	227.5722718
6100	home	6.153617	46.19619	125.0451168
6100	work	6.111018	46.2322	153.0656097
6103	home	6.568285	46.52819	97.95541303
6103	work	6.566071	46.52071	156.9946214
6104	home	6.588065	46.53745	366.5524149
6104	work	6.629869	46.52109	86.22429006
6109	home	6.657904	46.51297	116.7243
6109	work	6.608978	46.52216	156.2557174
6168	home	6.858735	46.45919	73.07664078
6168	work	6.775881	46.48245	103.8285919
6170	home	6.658519	46.51974	1.77E-10
6170	work	6.66065	46.51006	169.0706973
6171	home	6.611955	46.52557	194.2100959
6171	work	6.632456	46.52082	49.3454159
6176	home	7.122516	46.81282	167.5506793
6176	work	6.660102	46.51009	108.749024
6179	home	6.916774	46.43377	30.75155588
6179	work	7.418922	46.93165	183.016408
6192	home	6.925578	46.42134	202.2540611
6192	work	6.631731	46.52037	177.7624978

user_id	cluster_meaning	longitude	latitude	radius of the cluster (in m)
6197	home	6.557596	46.52427	138.7554077
6197	work	6.139204	46.20636	67.21626735
6198	home	6.602138	46.51779	108.4227287
6198	work	6.638367	46.52048	46.64058587