ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
TRANSPORTATION AND MOBILITY LABORATORY (TRANSP-OR)

MASTER THESIS

Presented for obtaining the
Civil engineering master diploma

By: Amélie Buisson

Individual Activity Travel Analysis based on smartphone WiFi data

Professor: Michel Bierlaire

Supervisors: Evanthia Kazagli and Antonin Danalet

Academic Year:2013-2014

Abstract

The mobility patterns of the population are the basis of most analyses in the transportation field. We aim to extract these patterns from smartphone traces. The following thesis proposes a Bayesian approach based on smartphone location records, land use information and schedule data to understand the activities that people daily perform. We investigate two alternatives. The prior is either based on schedule data or based on land use information.

We test the algorithm on the smartphone WiFi traces provided by Nokia. They have been obtained from people who live around the Leman Lake, essentially in the region of Lausanne. The Swiss Federal Statistical Office (FSO) and OpenStreerMap (OSM) provide the schedule and the land use information. The results show that the prior based on schedule data is not convenient: the importance of the individual behavior decreases, and the activity *Home* comes up too often. The results are much better when the prior depends on the land use that surrounds a users' location of interest. In addition, we have successfully extracted recognizable mobility patterns, particularly for the activities *Home* and *Work*.

Contents

Chapter 1

Introduction

1.1 Background and motivation

Today, the population growth and the increasing number of vehicles generate congestion on the transportation networks. These networks can only provide a good level of service if they are efficiently managed, according to the anticipated demand. The problem is that the demand is not easy to estimate because mobility is a derived-demand. People don't move for the pleasure of moving but in order to reach a location where they have something to do. Consequently, people tend to develop mobility patterns according to the activities that they have to perform during the day. The aim is to identify these patterns, analyse them and develop strategies in order to better operate the transportation network.

Several tools permit to collect information about people's mobility. Cameras, loop detectors and surveys are widely used, but they have some limitations. The two first gather aggregated information about specific points of the network. Surveys collect disaggregated data but they are expensive and time consuming. That's why the use of technologies such as smartphones emerges to gather the needed mobility information. These devices are very convenient: they are owned by a large majority of the population and they are able to continuously send information through wireless networks. Most of the smartphones hold a GPS sensor that continuously records a person's location. One limitation is that the sensor often takes time to turn-on when the trip starts, and often stops before the trip ends. Therefore, the origins and destinations are not included in the traces. Moreover, the GPS does not work indoor. This doesn't enable to follow people in great infrastructures like shopping centres.

We can overcome these drawbacks by tracking WiFi traces from users' smartphone. The use of internet brought cellphones to have access to the WiFi access points, identified with a unique mac address. Since we know the position of these access points, we are able to locate people according to WiFi traces. This is a fingerprint method because the origins and destinations are identified as sets of access points. This is feasible because the urban WLAN coverage is sufficient.

1.2 Literature review

People follow mobility patterns (Calabrese et al. (2013), Gonzalez et al. (2008)). Several approaches have been developed to identify them at the individual scale. Since 1974, the Swiss FSO realizes five-years surveys to gather information about people's mobility. The first surveys consisted in a written investigation and the sample was not larger than 18'000 people (OFS (2012)). The computer assisted phone interviews came in 1994. In 1998, Bowman (1998) developed a choice model where the activity scheduling decision is based on the travels characteristics (e.g., distance, travel time, cost). Then, the emergence and democratization of new technologies as GPS and smartphones has open the field to new possibilities (Hasan et al. (2013)). They permit to obtain continuous location records of large samples.

Human travels are characterized by a high degree of temporal and spatial recurrence. Few places are frequently visited by individuals (Gonzalez et al. (2008)). These places must be identified before understanding which activity is performed there. Many algorithms have been designed for continuous traces like GPS, or footprint data like GSM and WiFi traces (Ashbrook and Starner (2003), Marmasse and Schmandt (2000), Kang et al. (2004), Hightower et al. (2005)). Places are generally identified as a set of consecutive traces included in a predefined time-space window (Li et al. (2008)).

Once the locations where people perform activities are known, the place meaning must be inferred (Miller (2014)). Jiang et al. (2013) associate a place to an activity type according to the destination characteristics such as the location, land use or population density. These characteristics are linked with a table where the probabilities of human activities for different land use types have been extracted from the Massachusetts travel survey. Phithakkitnukoon et al. (2010) design a Bayesian approach to compute the probability that a location corresponds to an activity, based on the neighboring Points Of Interest (POIs).

This last approach is further extended by Danalet et al. (2014) who take the POIs' attractivity into account. A main difference between Danalet et al. (2014) and the previous works is also that they computes the probabilities on sequences of activities instead of studying locations independently.

In the work of Schönfelder et al. (2003), the identification of trip purposes relies on land use data and temporal characteristics of the activities, deduced from a national travel survey (the most likely activity purpose is revealed according to the individual characteristics and the time of day). Nevertheless, the temporal and the spatial informations are not merged together. The purpose of a trip is determined according to the land use data. The schedule data is only used if no clear POI is found. In addition, the inference doesn't rely on a probabilistic approach but on straightforward rules. Some behaviors are so characteristics of the population that can be used as rules to identify an activity with a satisfying accuracy. Calabrese et al. (2013) identify home as the place where people spend the highest number of nights, between 6 p.m. and 8 a.m.

The recent work of Montini et al. (2014) explores the ability of random forests to determine a trip purpose. They work on one week smarpthone data collected from 156 respondents. The sample is too small to highly extend the personalisation of trip purpose detection. The results show that personal-specific classifiers guaranty the best accuracy. It corresponds to the age, income, education level, marital status or the distance between home and work of individuals.

In the following paper, we propose to merge several ideas explored in the previous works. We aim to determine the activity performed by a user at a location, based on the surrounding attractivities and people's schedule. These two sources of information are merged together through a Bayesian approach.

1.3 Objective

We aim to determine the mobility patterns of a population as the result of an individual analysis based on smarpthone WiFi data. The work is subdivided in several tasks. First, we aim to identify the places of interest of each person's daily life. Besides, we want to determine the activity that the individual performs in each place. To this aim, we design a Bayesian approach based on WiFi traces, land use information and schedule data. Finally, we will aggregate the results and extract the corresponding mobility patterns.

We test the algorithm on the traces of the Lausanne Data Collection Campaign. Chapter 2 describes this campaign and the complementary data used for the needs of our work. Section 3 develops the mathematical approach used to determine people's locations of interest and the probability that it corresponds to a given activity. Two alternatives are proposed. Sections 4.2.1 and 4.2.2 show the results for each alternative. The contribution of the land use data in the results is evaluated in Section 5.1. Chapter 6 summarizes the results and proposes directions for future work.

Chapter 2

Data

In their everyday life, people visit several locations where they perform different activities. To identify these locations and their semantic meaning, we build an algorithm based on several inputs: WiFi traces, people's time budget and land use data. WiFi traces are available from the Lausanne Data Collection Campaign (LDCC) which is presented in this chapter. They are essential because they reflect people's movements and allow us to identify the locations of their daily life. The time spent, the frequency of visits and the visiting hours of each location can be deduced from the WiFi traces, and used to identify mobility patterns.

Once the locations are identified, we infer their meaning based on land use information and data regarding people's schedule. The allocation of time in Switzerland is obtained from the microsensus *Swiss mobility and transportation 2005* (OFS, 2007). The Swiss land use data is obtained from several sources. We gather information from OSM and from two surveys by the Swiss FSO: *the households and population census 2010* (Zaugg et al., 2012) and *the enterprise and buildings census 2005* (Zaugg et al., 2007).

2.1 Nokia Dataset

2.1.1 Lausanne Data Collection Campaign

WiFi traces are available from the LDCC launched by the Nokia research centre of Lausanne. The campaign involved 200 people from the surroundings of the Leman Lake between 2009 and 2011. The sample has been built on a viral process (snowball sample) to include people with different backgrounds. A seed of students engaged their colleagues, family and friends, who engaged their own relatives and so on (Laurila et al., 2012). The sample consists in 65 % of male participants and 35% of female, with a large majority of adults between 22 and 33 years old. Students, employees and non employed respondents represent respectively 26%, 63 % and 8% of the participants (Kiukkonen et al., 2010).

People have been offered a smartphone (Nokia N95) equipped with multiple sensors. The device was continuously recording GPS, cellular network and WLAN information. Due to battery constraints, sensors are not turned-on at the same time. When the mobile recognized WLAN access points (WLAN APs), the *known WLAN* state occurred and the GPS sensor was stopped (Kiukkonen et al., 2010). In such cases, WLAN was used instead of GPS to locate people. The WLAN APs were associated with coordinates by matching GPS and the WLAN APs information when both are roughly recorded simultaneously. Each AP had been assigned to the centroid of all the GPS points recorded when the AP was visible (Montoliu and Gatica-Perez, 2010).

The data collected with smartphones was combined with a survey answered from twenty participants. They provided their home, work and their first and second grocery addresses. This sub-sample is essential to validate the results obtained from the algorithm.

2.1.2 WLAN data format

WLAN records of 179 people are available. This represents an amount of 590'000 APs, 68% associated with coordinates (Fig. 2.1). The WLAN information is stored in two tables: wnetworks and wlan_relation. The wnetworks table gives information about each AP's id, machine address, latitude, longitude and geometry (Table 2.1). The wlan_relation table links this data with the users. Each time a user's smartphone sees an AP, a measurement $\hat{m}_{i,j}$ appears in the table, specifying the user id i, the time stamp t and the id of the observed AP n (Table 2.2). As a smartphone can see multiple APs at the same time, several measurements can have the same entries i and t but different values for n. The chronologically ordered set of measurements related to a participant i is written $\hat{m}_{i;1:J}$ $=(\hat{\mathrm{m}}_{i;1},\hat{\mathrm{m}}_{i;2}...\hat{\mathrm{m}}_{i;J-1},\hat{\mathrm{m}}_{i;J})$ $=((\hat{\mathrm{t}}_{i;1},\hat{\mathrm{n}}_{i;1}),(\hat{\mathrm{t}}_{i;2},\hat{\mathrm{n}}_{i;2})...(\hat{\mathrm{t}}_{i;J-1},\hat{\mathrm{n}}_{i;J-1}),(\hat{\mathrm{t}}_{i;J},\hat{\mathrm{n}}_{i;J}))$ where J is the total number of measurements for user i. The distribution of the data is shown in Fig. 2.2. The number of APs and the number of distinct timestamps vary through the respondents.

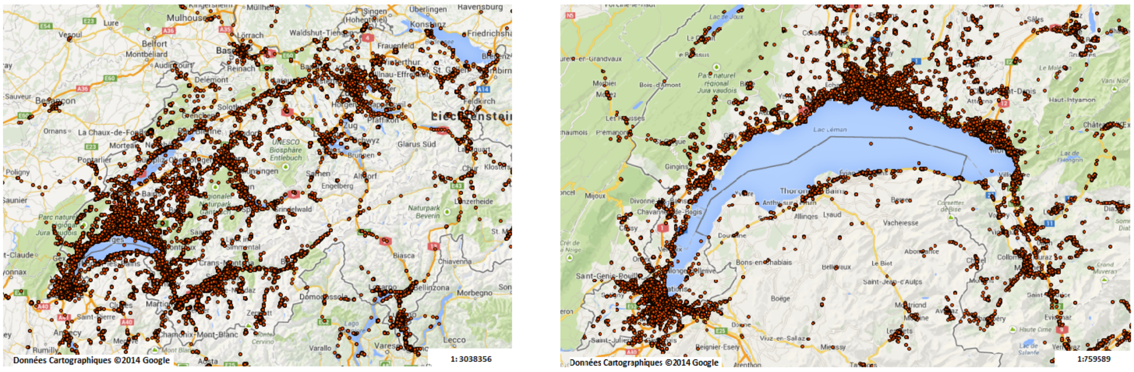
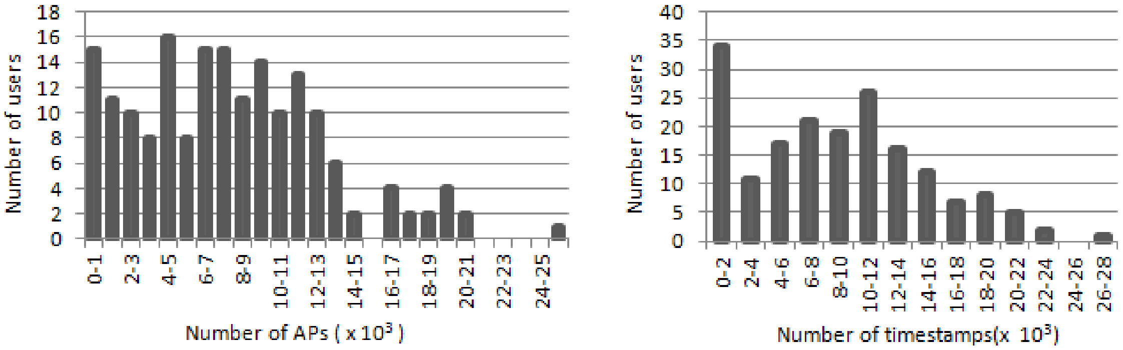

Données Cartographiques ©2014 Google
1:3038356
Données Cartographiques ©2014 Google
1:759589

Figure 2.1: WiFi APs associated with coordinates.

Table 2.1: Sample of information stored in wnetworks

WNETWORKS				
id	mac_adress	longitude	latitude	geom
9800	000fcc-a47b3ae63f 0b62efb9eebca38cc e4bcdafe0b6b15236c f8bc8826e6a87afc922	6.640873337	46.510085	0101000020E6 1000006E2C9D 1941901A400F 971C774A414740
9801	000fcc-d3b296fa85f ac9576f8872a0b3ec d58700d6d80ee1c817 b5d23ca60700133359	6.640382511	46.5102937	0101000020E6 10000092B3D6 6EC08F1A406B 6DD04D51414740



Source: Buisson 2013

Figure 2.2: WLAN records distribution.

Table 2.2: Sample of information stored in wlan_relation

WLAN_RELATION		
user_id	timestamp	networkid
6086	01.06.2010 07:44	9800
6086	01.06.2010 07:46	9800
6086	01.06.2010 07:46	9801

2.2 Microsensus Swiss mobility and transportation (2005)

People's schedule describes the activities they do during different hours of the day. It is the second input for the algorithm. We assume that people follow mobility patterns and that they perform an activity or another according to the time of the day. Hence, the hour of the day reveals what someone is likely to do. If we extract the visiting hours of people's locations from the WiFi traces and compare it to people's schedule, we should be able to understand what activity is satisfied in each location.

One thing we are interested in is $\mathrm{P}(t|a_k)$, the probability to be time t given that a specific activity a_k is performed. We compute it based on the *microsensus Swiss mobility and transportation 2005*. This microsensus is one of the five-year surveys done by the Swiss FSO to gather information about people's mobility. In 2005, phone interviews involved 31 950 households and 33 390 people. All the participants are Swiss residents older than 6 years old. After giving information of the household structure, one respondent (two if the household's size exceeds 4 persons older than 6 years old) details the activities she has performed during a day of the year.

As representativeness of the sample is a main concern of such a campaign, households have been randomly chosen with a minimum of 1 000 interviews for each of the seven great regions of Switzerland (list NUTS2 of Eurostat), and a draw proportional to their population. Moreover, weights have been allocated to each household. This is necessary to avoid considering people from crowded regions more representative of the Swiss population than they really are, or on the contrary, avoid to under-represent some groups of the population who could be more difficult to join like young singles (Gindraux, 2007). Respondents' individual weights have been computed by multiplying the household weight and the inverse of the probability that the respondent would be chosen (number of people in the household over the number of interviewed persons). Later, these weights have been corrected based on age, nationality and gender.

We have the detailed schedule of 19 089 people for a specific weekday of the year (this day is spread over all the year through the respondents). The start, the duration and the nature of each activity they have performed has carefully been recorded. In the microsensus, a set of thirteen activities was defined as shown in Table 2.4. Similarly to Pas (1982) and Bowman (1998), we consider that people mostly perform one of the four activities defined in Â (a_1: *Home*, a_2: *Work*, a_3: *Leisure*, a_4: *Shopping*). The thirteen activities of the microsensus are allocated as follows: *Home* (11), *Work* (2, 3, 6, 7), *Shopping* (4, 5), *Leisure* (8). The activities 1, 9, 10, 12, 13 of Table 2.4 are left out because they are not exclusively related to an activity of Â and they occupy a negligible part of people's timetable.

Taking people's weights into account, we have computed $\mathrm{P}(a_k|t)$ for each a_k of Â and each time t as:

$$P(a_k|t) = \frac{people_{a_k;t}}{people} \tag{2.1}$$

where $people_{a_k;t}$ is the number of people doing activity a_k during time t, and *people* is the total number of respondents. We decide to work with discrete values of t and compute $\mathrm{P}(a_k|t)$ for each hour of the day. The curves are shown in Fig. 2.3.

The algorithm doesn't operate with $\mathrm{P}(a_k|t)$ but with $\mathrm{P}(t|a_k)$. We obtain it by normalizing the previous curves to 1, in accordance with Eq. 2.2. We obtain the probabilities of Fig. 2.4. These curves only represent weekdays because the census doesn't give information for week-ends. On saturdays and sundays, we set $\mathrm{P}(t|a_k) = \frac{1}{24}$ for each hour of the day.

$$\int_{\mathrm{t}} \mathrm{P}(t|a_k)dt = 1 \tag{2.2}$$

We also compute the probability $\mathrm{P}(a_k)$ to perform an activity a_k as:

$$\mathrm{P}(a_k) = \frac{\sum_{i=1}^{q} duration_i(a_k)}{\sum_{k=1}^{4}\sum_{i=1}^{m} duration_i(a_k)} \tag{2.3}$$

where q is the total number of recorded episodes related to a_k and $duration_i(a_k)$ is the duration of the considered episode i for the activity a_k (Table 2.3). We notice the high value of $\mathrm{P}(\mathit{Home})$ and the small value of $\mathrm{P}(\mathit{Work})$. Considering that people work 42 hours per week in Switzerland, we could expect $\mathrm{P}(Work) = \frac{42}{7\times24} = 0.25$. Nevertheless, this does not include people who partially work or does not work at all. Elderly people, children, housewives or non-employed must be taken into account. This is why $\mathrm{P}(\mathit{Work})$ is lower than 0.025.

Table 2.3: Probability $\mathrm{P}(a_k)$ to perform an activity a_k

a_k	**P(a_k)**
Home	0.715
Work	0.198
Shopping	0.014
Leisure	0.067

We use the microsensus of 2005 but the raw data of the MTMC 2010 is available. This latter would better fit to the Nokia data collected between 2009 and 2011 but it requires to spend time on it to extract manageable schedule data. Moreover, a quick comparison between the reports of 2005 and 2010 reveals few changes in the travelled distances according to trips purposes (Fig. 2.5). The part of the daily travelled distance for shopping has increased from 11.3% to 12.8% and the same tendency is observed for work and formation. On the contrary, the leisure displacements show a decrease of 4.5% (OFS, 2012).

Table 2.4: Activities in the microsensus

	Activity	Description	Â
1	Mode change, parking		x
2	Work		Work
3	Education		Work
4	Shopping	Travel for shopping, supply and receive benefits (eg the doctor.)	Shopping
5	Personal business, consumption of services	Post office, bank, medical treatment, etc.	Shopping
6	Work related activities	To set a personal business (meeting, visit an agent)	Work
7	Work related travel		Work
8	Leisure activity		Leisure
9	Child dropping		x
10	Accompanying, doing something for someone else	Ex. Accompanying parents to the airport, accompanying children	x
11	Returning home, current accommodation		Home
12	Others		x
13	Border crossing		x

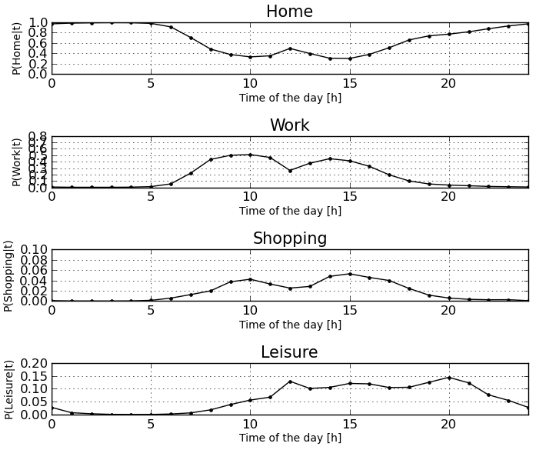


Figure 2.3: Probability to perform an activity for each hour of the day, weekdays (MTMC 2005).

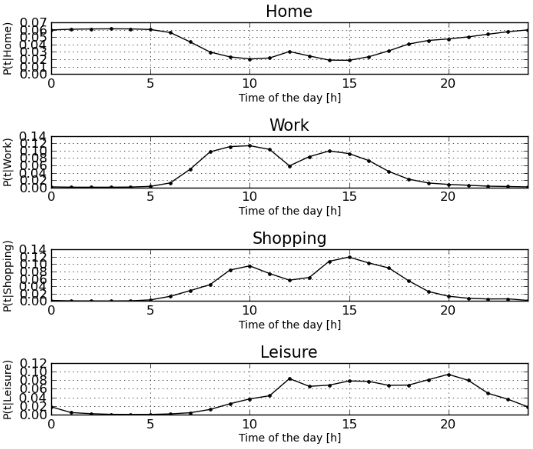


Figure 2.4: Probability to be hour h knowing that we perform an activity, weekdays (MTMC 2005).

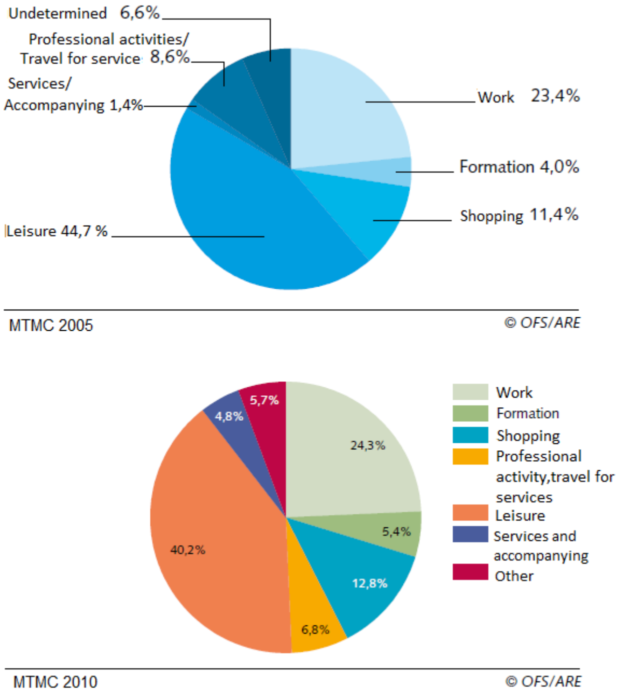


Figure 2.5: Part of the daily travelled distance according to trips purposes.

2.3 Land use Data

The schedule data of Swiss people reveals that they mainly perform the four activities listed in Â. These activities are done in specific locations called Points of Interest (POIs). It can be households for *Home*, work buildings for *Work*, parks and museums for *Leisure* and commercial centres for *Shopping*. For the needs of the work, we need to collect the attractivity of the POIs related to each activity of Â. Danalet et al. (2014) define the attractivity as *"a model of aggregated occupation per POI"*. They express it as a number of persons. It is the number of residents for *Home*, the number of workers for *Work*, and the capacity for *Leisure* and *Shopping* POIs (e.g., the number of seats for a restaurant). The following chapter describes how the attractivities are obtained for each of the four activities.

2.3.1 Home

The attractivity for *Home* comes from *the households and population census 2010 (STATPOP)* of the Swiss FSO. This latter is based on municipalities and cantons' administrative registers of people, from which the data is collected each 31 of December. The census gives detailed information about the population age, gender, nationality and other socio-economic information for each cell of 100 × 100 meters. For our needs, we take the total permanent resident population of each hectare. This includes Swiss people but also people holding the C, B, Ci, L, F or N permit. Few people whose address is unknown (0,67%) have been associated to the geographic centre of the city (collective hectare). To avoid any bias in the results, some detailed studies may need to spread these people uniformly through the municipality. We don't spread these records and we assume that their number is too small to be significant on our results.

2.3.2 Work

We find the attractivity for *Work* in the *the enterprise and buildings census 2005 (RE)* from the Swiss FSO. It includes all the economic branches of the secondary and third sectors: industry, handicraft, services, self-employed professions, public administrations, social insurances and non-lucrative organizations. From this census, we have taken the equivalent number of full-time jobs for the sectors 2 and 3. All the employees working more than 6 hours per day have been taken into account, their activity being payed or not. This data is aggregated over cells of 100 × 100 meters from the same grid than the one used in STATPOP.

In Table 2.4, education is associated to the activity *Work*. Because the LDCC sample has been built from a group of EPFL students and does not involve children, we manually considered a uniform distribution of 8'000 and 12'000 students respectively on the EPFL campus and the UNIL campus. These numbers come from the websites of the two universities.

2.3.3 Shopping

The category *Shopping* contains commerces and services. The detailed establishments taken into account are listed in Table 2.5. Their number is given for each hectare by the *enterprise and buildings census* (2005). The economical activity is attributed based on the general catalogue of economical activities of 2002 (NOGA) and the major activity of establishments (the one that requires the most employees). The buildings location is determined with the federal register of buildings and housings (RegBL). Few establishments that have not been associated to a building of the RegBL have been located at the centre of the municipality (collective hectare).

The census only provides a number of establishments. We describe how we convert it into an attractivity (number of persons) in Chapter 4.

Table 2.5: Land use data in the activity type *Shopping*

Commerce	Services
Trade, maintenance and repair of vehicles, gasoline stations	Land, air and water transports; auxiliary transport services
Intermediate trade and wholesale	Post and telecommunications
Retail trade	Financial intermediation, insurances and auxiliary activities
	Financial intermediation, insurances and auxiliary activities
	Real estate activities
	Machines and equipments location
	IT activities
	Public administration, defence, social welfare
	Health, veterinary affairs and social activities
	Other services

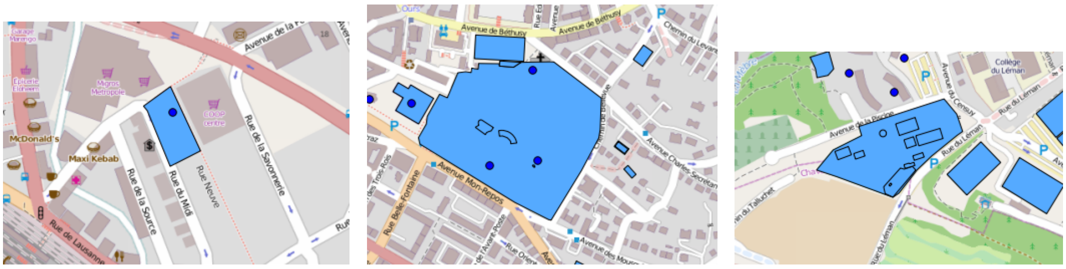
2.3.4 Leisure

According to the MTMC, leisure activities include (by order of importance) visits, gastronomy, outdoor activities, sport, sporting events as viewer, medicine/wellness, cultural events/recreational facilities, companies, trip/vacation, unpaid work, shopping, church / cemetery, recreation at home (e.g., among neighbours) and eating out.

From *the enterprise and buildings census 2005*, the number of hotels and restaurants is known for each hectare. Because other points of interests related to *Leisure* may be less dense and occupy bigger areas (e.g., stadiums for sport or parks for visits), it is interesting to have their exact location and their shape when provided. Hence, the OSM data holding the tags listed in Table 2.6 has been downloaded. In OSM, a point of interest can be represented by a point or a polygon (corresponding to the building shape). From it, several combinations are possible. Sometimes, a POI is represented by a polygon containing a point (at its centroid), a polygon containing several points (e.g., fountains in park) or a polygon containing several polygons (e.g., stadiums). This is illustrated in Fig. 2.6.

When we compute the attractivities for *Leisure* in a given area, few problems appear due to the OSM representation of the POIs. Let's give an example. In OSM, some stadiums are simple polygons whereas others are polygons containing other polygons that represent all the detailed sport grounds of the sport center. In the first case, the sport center is counted as 1 feature whereas it might correspond to 30 features in the second case. To simplify the data, we deleted all the features (points or polygons) contained in polygons.

The comment is the same as for *Shopping*. The census and OSM only provide a number of establishments. We describe how we convert it into an attractivity (number of persons) in Chapter 4.



(a) POI represented by a point and a polygon (b) The fountains of a park are represented as points (c) A sport center represented by a polygon containing multiple polygons

Figure 2.6: POIs features in OSM.

Table 2.6: OSM tags included in *Leisure*

cooking_school	spa	race_track	swimimng_pool
wellness	atterissage_parapente	internet_cafe	music_venue
cinema	cooking_school	kneipp_water_cure	scout_center
brothel	casino	dog_training	castle
community_centre	community_center	community_hall	concert_hall
fountain	gallery	garden	hunting_stand
library;social_facility	library	nightclub	sauna
scout_centre	stripclub	swingerclub	theatre
youth_centre	youth_club	alpine_hut	artwork
boat_rental	brothel	cinema	club
gallery	gambling	gambling_hall	hunting stand
stripclub,brothel	meeting_point	meetingpoint	music_school
parkbench	pique-niques	sauna	solarium
golf_course	horse_ riding	ice_rink	pitch
sport	sports_centre	stadium	swimming_pool
swimming_pool; ice_rink	swin_golf	tennis	track
fitness	fitness_centre	fitness_station	gym
swimming_hall	dive_center	dancing_school	gym
dojo	fitness	fitness_center	fitness_trail
sports_centre	swimming_pool	gym	Leichtathletik
snow_park	sport	sports	sports_centre
arts_centre	bicycle_rental	boat_rental	
boccia_course	Sportzentrum_Bärenmatt	sport	sports
ski_school	swimming_pool		

Chapter 3

Probabilistic inference of people's activity episodes sequences

Location records, schedule data and attractivities are the input of the algorithm. It proceeds in two successive steps. First, it determines the locations of interest of each user based on the traces. These locations correspond to places where daily activities are performed. The second part of the algorithm aims to determine the nature of these activities by means of a Bayesian approach. We define a set Â (a_1, ... , a_k, ..., a_K) of activity types. The output of the algorithm consists in the probability of activity type for each location. The following paragraphs develop the mathematical basis of the approach and its implementation.

3.1 Definitions

A measurement is defined as $\hat{m} = (\hat{x}, \hat{t}) \in \mathbb{R} \times \mathbb{R}$, where $\hat{x}$ is the measurement location (e.g., an access point) and $\hat{t}$ is the measurement timestamp.

The set of measurements related to a user i is defined as $\hat{m}_{i,1:J} = (\hat{m_{i,1}}, ..., \hat{m_{i,j}}, ..., \hat{m_{i,J}}) = ((\hat{x}_{i,1}, \hat{t}_{i,1}), ..., (\hat{x}_{i,j}, \hat{t}_{i,j}), ..., (\hat{x}_{i,J}, \hat{t}_{i,J}))$ where J is the total number of measurements related to the user.

A location where a user i performs an activity is identified by a group of $\hat{x}_{i,j}$ that constitute a cluster. The associated domain of data relevance $DDR \in \mathbb{R} \times \mathbb{R}$[1] is the area that includes the potential real positions of the user when he generates a measurement in the cluster (the DDR location is the cluster location).

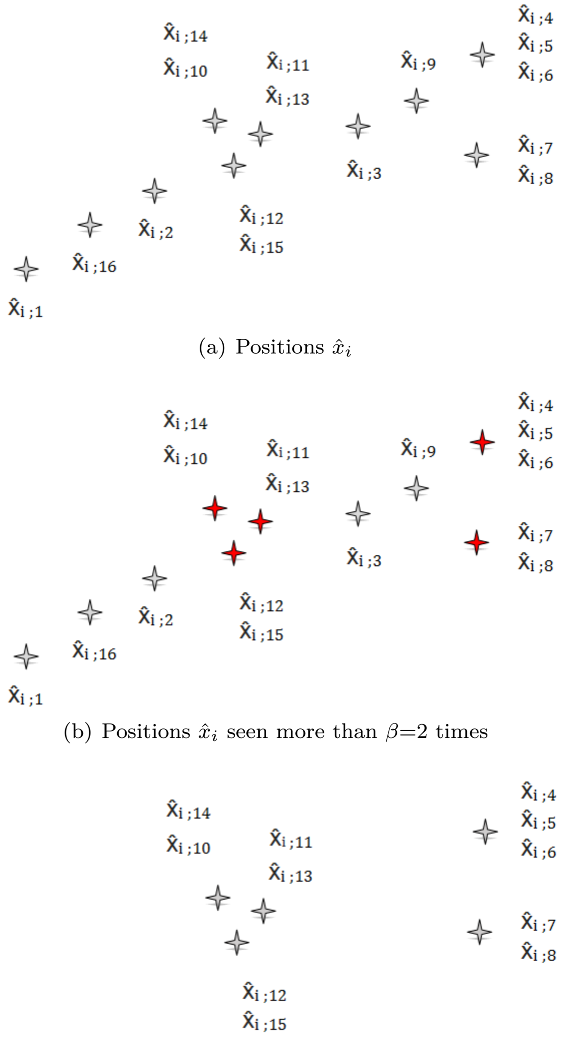
The time-extended domain of data relevance DDR^t defines the set of activity episodes performed by the user i in DDR. The episodes $DDR^t_s = (DDR, t^-_s, t^+_s)$ are chronologically ordered with the index s $\in [1 : J]$, where $\hat{t}^-_s$ and $\hat{t}^+_s \in R$ are the start and end time of the activity episode.

3.2 Identification of the individual locations of interest

The locations where a user i performs activities are deduced from the traces. The $\hat{x}_i$ seen less than β times are removed, assuming that they don't correspond to a location where an activity is performed (e.g., they have been seen when moving). This operation is illustrated in Fig. 3.1. The remaining $\hat{x}_i$ are clustered with a DBSCAN (Density-Based Spatial Clustering of Applications

[1] The Domain of Data Relevance is a notion defined by Bierlaire and Frejinger (2008)

with Noise) algorithm (Ester et al. (1996)). The algorithm explores the neighbour of each position $\hat{x}_i$ within a radius ε. If more than α points $\hat{x}_i$ are found, a cluster is generated and $\hat{x}_i$ is a *core point*. If a *core point* contains an other *core point*, they belong to the same cluster. If a position $\hat{x}_i$ is included in the neighborhood of a *core point*, it joins the cluster as a *border point* . Otherwise, $\hat{x}_i$ is considered as an *outlier* and doesn't belong to any cluster. Fig. 3.2 illustrates the mechanism of DBSCAN.

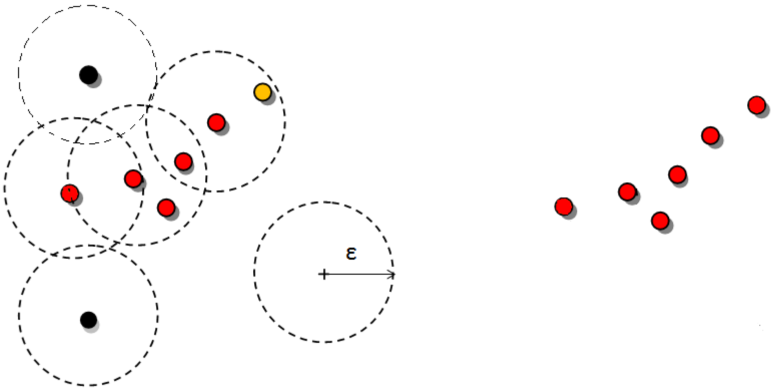


(a) Positions $\hat{x}_i$

(b) Positions $\hat{x}_i$ seen more than β=2 times

(c) Input for the DBSCAN algorithm

Figure 3.1: Constitution of the input for the DBSCAN algorithm.$\beta = 2$.



(a) Initial set of points

(b) The cluster

Figure 3.2: DBSCAN algorithm. Red: core points, Yellow: border points, Black: outliers. $\alpha = 2$.

3.3 Inference of a cluster's meaning

We aim to compute the probability $\mathrm{P}(a_k|DDR^t)$ that an individual with a smartphone generating traces performs the activity a_k, given the time-extended domain of data relevance DDR^t. It can be decomposed by means of the Bayesian theorem:

$$\mathrm{P}(a_k|DDR^t) = \frac{\mathrm{P}(DDR^t|a_k) * \mathrm{P}(a_k)}{\mathrm{P}(DDR^t)} \tag{3.1}$$

$$P(a_k|DDR^t) = \frac{\mathrm{P}(DDR^t_1, ..., DDR^t_S|a_k) * \mathrm{P}(a_k)}{\mathrm{P}(DDR^t_1), ..., \mathrm{P}(DDR^t_S)} \tag{3.2}$$

$$P(a_k|DDR^t) = \frac{\mathrm{P}(DDR^t_1, ..., DDR^t_S|a_k) * \mathrm{P}(a_k)}{\sum_1^K \mathrm{P}(DDR^t_1, ..., DDR^t_S|a_k) * \mathrm{P}(a_k)} \tag{3.3}$$

where $\mathrm{P}(DDR^t_1, ..., DDR^t_S|a_k)$ is the likelihood and $\mathrm{P}(a_k)$ is the prior of the Bayesian approach.

3.3.1 Likelihood

The likelihood $\mathrm{P}(DDR^t_1, ..., DDR^t_S|a_k)$ is the probability to observe the activity episodes 1 to S in DDR, knowing that they correspond to the activity a_k. The schedule data permits to compute the probability $\mathrm{P}(t|a_k)$ to be time t knowing that an activity is performed, for each time of the day. The likelihood is the integral of $\mathrm{P}(t|a_k)$ on the time intervals $[\hat{t_{1-}};\hat{t_{1+}}]$, ..., $[\hat{t_{S-}};\hat{t_{S+}}]$. This represents the time-of-day preference. A DDR that contains a restaurant is more likely to be the location where the individual really performs an activity if DDR^t occurs during lunch break than if it occurs in the middle of the afternoon.

3.3.2 Prior

Prior based on time budget

The prior $\mathrm{P}(a_k)$ is the probability that the extended domain of data relevance DDR^t corresponds to the activity a_k. We can compute it based on people's schedule. The more people spend time for an activity over the day, the more a user's daily location is likely to correspond to this activity (Eq. 2.3).

Prior based on land use

We can also compute the prior based on the land use that surrounds the cluster. A cluster is a set of measurement locations $\hat{x}_i$ and each of them can be the real location of user i when he performs an activity in the cluster. Their position being more or less accurate, we assume that each $\hat{x}_i$ can be located in a radius R around its current coordinates. Based on this assumption, each location $\hat{x}_i$ corresponds to one of the POIs contained in R. The DDR is the union of the domains R of the positions $\hat{x}_i$ that constitute the cluster.

The more establishments in the DDR permit to perform an activity a_k, the more DDR^t is likely to correspond to this activity. The prior becomes contextual as it depends on the neighbor of the positions $\hat{x}_i$. Based on time budget, the prior does not depend on the cluster, it is always the same.

A POI in DDR can be more or less attractive, that is more or less likely to be the real location where the user performs the activity. Based on the definition provided by Danalet et al. (2014), we define the attractivity as a number of persons: the number of workers for *Work* and the number of residents for *Home*. It is more difficult to have a number of persons for activities like *Shopping* and *Leisure* but the number of establishments can be mutiplied by average capacities.

Finally, the probability that a DDR^t corresponds to an activity a_k is computed as shown in Eq. 3.4

$$P(a_k) = \frac{\sum_{POIs \in DDR} attractivity(a_k)}{\sum_{k=1}^{K} \sum_{POIs \in DDR} attractivity(a_k)} \tag{3.4}$$

If we don't have any attractivity for the DDR, we set P(a_k) to $\frac{1}{K}$ for each a_k of Â.

Chapter 4

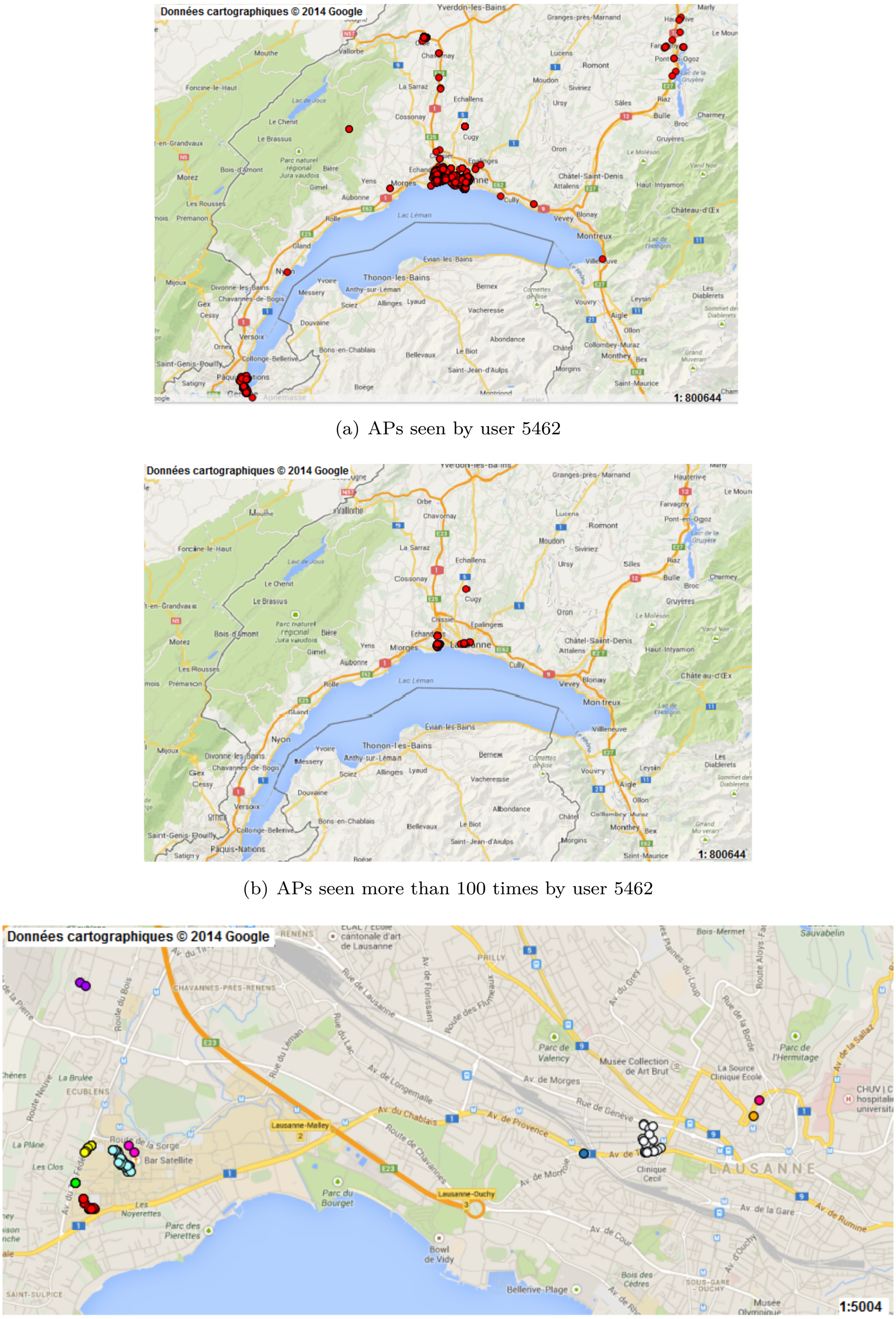
Case study based on Nokia data

4.1 An individual example

We test the algorithm presented above on the WiFi traces provided by Nokia. The schedule data comes from the microsensus *Swiss mobility and transportation 2005*. It allows us to compute $P(t|a_k)$ for each hour of the day and to define the set of activities Â (*Home*, *Work*, *Shopping* and *Leisure*). The land use data is obtained from OSM , *the households and population census 2010* and *the enterprise and buildings census 2005*. With the formulas presented in Section 3.3, four probabilities are computed for each cluster of each user. The relative difference between these probabilities measures the certainty about the activity performed by the user in a cluster. In this section, we illustrate the overall process that provides the activity performed in a location from WiFi traces.

4.1.1 Places of interest for the user 5462

We chose to illustrate each step of the algorithm for the user 5462. It begins with the identification of the locations where he performs activities. Fig. 4.1 shows how the APs have been filtered with a threshold of $\beta = 100$ timestamps. From a cloud of APs that covers all the region of Lausanne (4.1(a)), we understand that the user daily stops in the center of Lausanne, Echandens, Ecublens and Cugy (4.1(b)). The APs that have been kept are clustered with the DBSCAN algorithm. We set $\varepsilon = 1000m$ and $\alpha = 1$. Nine clusters are visible on 4.1(c) where we mapped the area of Ecublens and Lausanne. All the APs that hold the same colour belong to the same cluster. We observe that the number of APs and their dispersion vary from a cluster to another. In general, they correspond to different quarters of the city but if we look at Lausanne, we see that some clusters are quite close. This is the case for the pink and the orange clusters. The possibility that several clusters are related to the same activity is possible. It would be the case if someone would park not exactly at its destination location.



(a) APs seen by user 5462

(b) APs seen more than 100 times by user 5462

(c) Places of interest for user 5462

Figure 4.1: DBSCAN output. User 5462.

4.1.2 Likelihood

For each cluster of the user, we compute the probability that the associated activity episodes correspond to *Home*, *Work*, *Leisure* and *Shopping*. To this aim, we need to compute the likelihood based on people's schedule and WiFi measurements. An example is given for the most visited cluster of user 5462.

We have a set of J WiFi traces related to the user 5462. We take the J_c measurements of user 5462 that hold a value n related to an AP of the most visited cluster. They prove the presence of the user during two time intervals on January 1$^{\text{st}}$: from 01:00:12 to 7:10:30 and from 20:00:00 to 22:30:34. January 1$^{\text{st}}$ is a weekday, so we integrate the curves of Fig. 2.4 over the two time intervals to obtain the likelihood for each activity. Fig. 4.2 shows that the biggest likelihood corresponds to the activity *Home*.

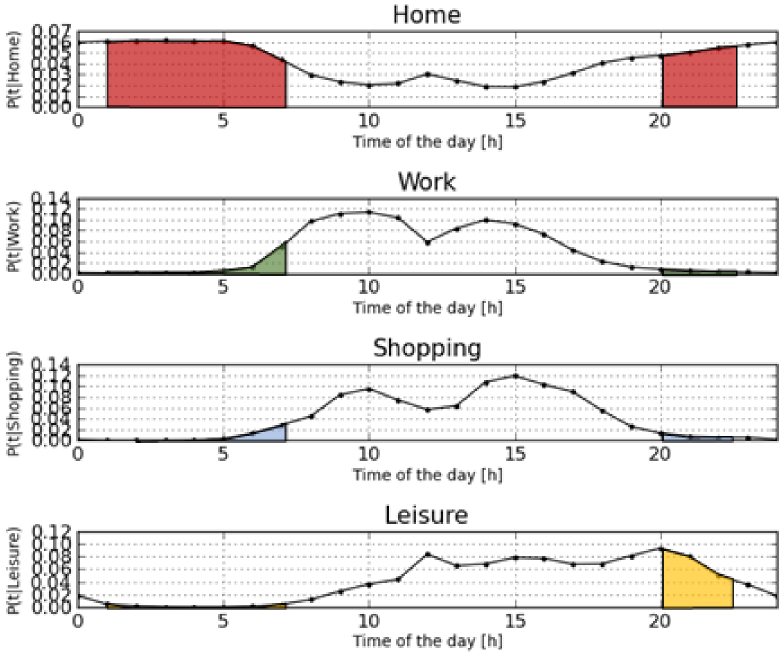
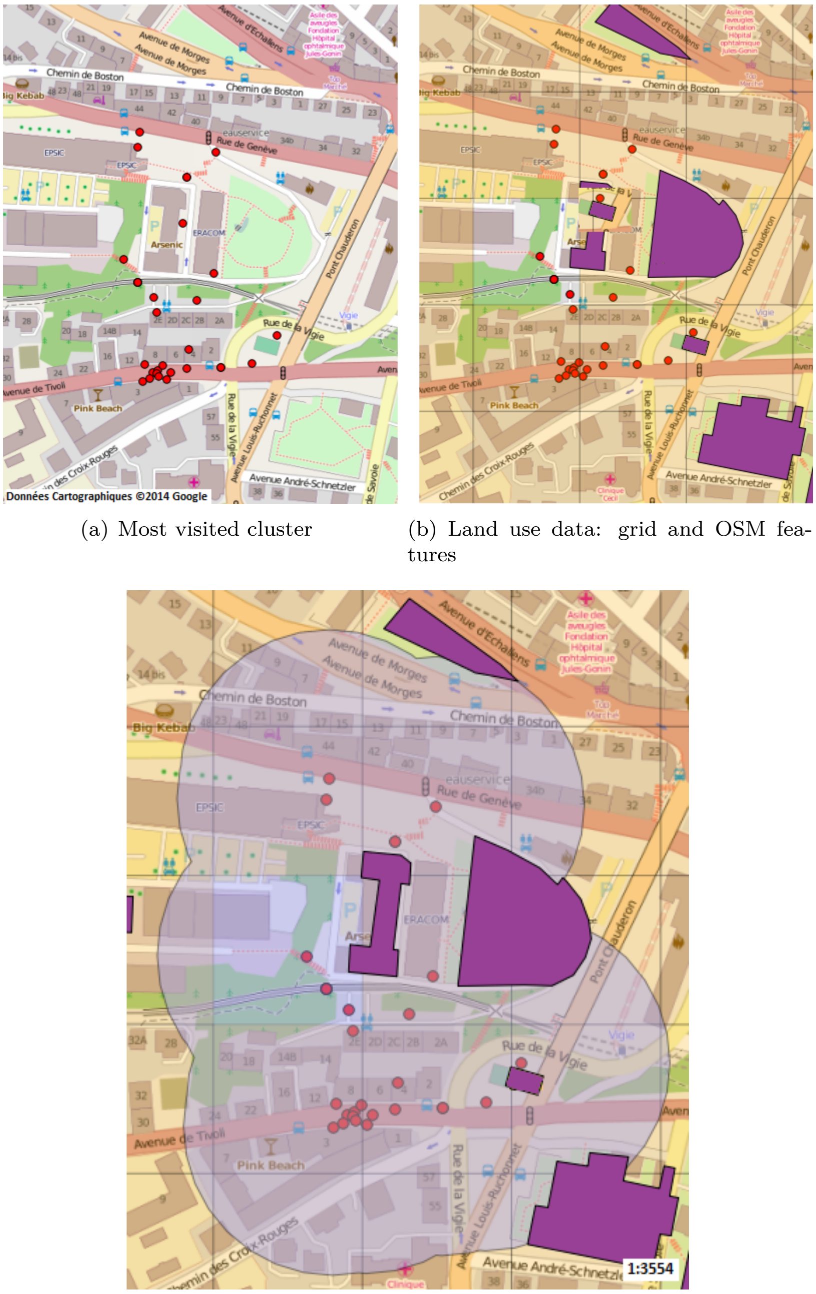


Figure 4.2: Likelihood for each activity on January 1$^{\text{st}}$.

4.1.3 Prior

The prior is computed based on the APs' location and the neighboring land use. We generate the domain DRR of the clusters and collect the attractivities in DDR for each activity. We show the procedure for the most visited cluster of the user 5462.

Blunck et al. (2011) conducted an experiment to estimate the GPS data accuracy when obtained with a Nokia N97. They proved that the error is lower than 60 meters for more than 90 % of the GPS points, and always lower than 100 meters. Therefore, we assume that each AP of the cluster can be located in a radius R of 100 meters around its current coordinates. The union of all the domains R of the cluster constitutes the DDR in which we compute the attractivities (Fig. 4.3).



(a) Most visited cluster

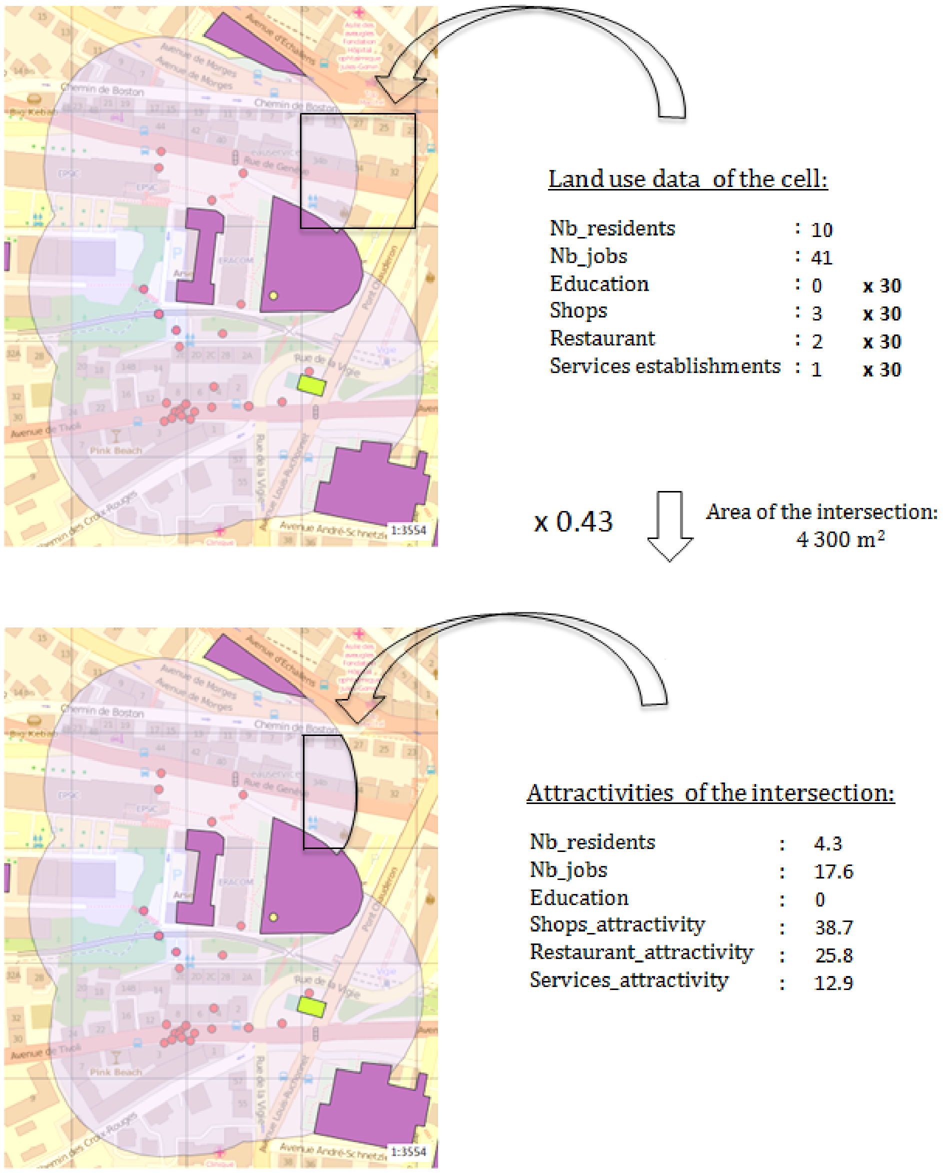
(b) Land use data: grid and OSM features

(c) *DDR*

Figure 4.3: DDR of the most visited cluster. User 5462.

When the land use data comes from OSM, each establishment is represented by a feature. Postgresql provides the number of OSM features that intersect the DDR. In Fig. 4.3, 5 features contribute to the attractivity of the activity *Leisure*.

With aggregated data, the attractivities are computed by intersecting the *DDR* with the hectares (22 in our case). The data from the Swiss FSO hectares are weighted by the area ratio between the intersection and the hectare. By this, we assume a uniform distribution of the land use data provided by the census. Number of residents and workers in the hectare is not modified. Number of buildings in the hectare (e.g., restaurants) is multiplied by the capacity (e.g., 30 persons) to define an attractivity. An example of these calculations is given in Fig. 4.4. The schema is repeated for all the hectares and the results sum up together to obtain the attractivities in the *DDR* (Table 4.1).



Land use data of the cell:

Nb_residents	:	10	
Nb_jobs	:	41	
Education	:	0	x 30
Shops	:	3	x 30
Restaurant	:	2	x 30
Services establishments	:	1	x 30

x 0.43

Area of the intersection:
4 300 m^2

Attractivities of the intersection:

Nb_residents	:	4.3
Nb_jobs	:	17.6
Education	:	0
Shops_attractivity	:	38.7
Restaurant_attractivity	:	25.8
Services_attractivity	:	12.9

Figure 4.4: Calculation of the attractivities.

Table 4.1: Attractivities in DDR

nb_residents	744	*Home*
nb_jobs	949	*Work*
education_att	0	*Shopping*
shops_att	717	
services_att	0	*Leisure*
restaurant_att	100	
leisure_att (OSM)	150	
sum	2660	

$$P(Home) = \frac{744}{2660} = 0.28; P(Work) = \frac{949}{2660} = 0.36 \tag{4.1}$$

$$\mathrm{P}(Shopping) = \frac{717}{2660} = 0.27; \mathrm{P}(Leisure) = \frac{250}{2660} = 0.09 \tag{4.2}$$

4.1.4 Probabilities

$\mathrm{P}(a_k|DDR^t)$ is computed based on the likelihood defined in Section 3.3.1 and the prior defined in Section 3.3.2. It is computed for each day of observation and averaged for each activity a_k of Â. The averaging does not take into account major changes in people's behavior over time. Such changes (e.g., a move) can be detected from WiFi traces, as shown in Section 4.2.2.

4.2 Test on the LDCC WiFi traces

We test the algorithm on 179 participants of the LDCC. We do it progressively, beginning with the two most visited places of each person and three probabilities : $\mathrm{P}(Home|DDR^t)$, $\mathrm{P}(Work|DDR^t)$ and $\mathrm{P}(Other|DDR^t) = \mathrm{P}(Shopping|DDR^t) + \mathrm{P}(Leisure|DDR^t)$. These probabilities are computed with both prior and likelihood based on people's schedule. Then, we work with the prior based on land use and the probabilities $\mathrm{P}(Home|DDR^t)$, $\mathrm{P}(Work|DDR^t)$, $\mathrm{P}(Shopping|DDR^t)$ and $\mathrm{P}(Leisure|DDR^t)$ are computed for all the clusters of the users.

4.2.1 Prior based on people's schedule

Results for the whole population

In Section 3.3.2, we propose two approaches to compute the prior. The first is based on people's time budget. Consequently, the prior is common through people and clusters. We obtained the values for $\mathrm{P}(a_k)$ from *the swiss microsensus mobility and transportation 2005* (Table 2.3). We run the algorithm on the two most visited clusters of each individual. We want to test if these clusters mostly correspond to *Home* and *Work*. We compute three probabilities for each DDR^t: $\mathrm{P}(Home|DDR^t)$, $\mathrm{P}(Work|DDR^t)$ and $\mathrm{P}(Other|DDR^t)$.

Over the 179 respondents, 5 have seen less than two APs during the study. We don't consider these users in the analysis as it is clearly not feasible to evaluate their daily activities with so little information. Three persons have seen more than 2 APs but DBSCAN only returns one cluster. We don't go further for these users either as we expect to identify at least two clusters in people's life: *Home* and *Work*.

When we look at the probabilities obtained for the 171 remaining persons, we make a striking observation. A large majority of the places are most likely to be *Home*. That concerns 312 places over 342. The other 20 places are most likely to be *Work* (Fig. 4.5). We observe that the probabilities P($Home|DDR^t$) of the 312 clusters are high, with an average of 0.588. We don't have such high values for the 20 clusters that tend to be *Work*. Over the 171 persons, the probability that the two clusters correspond to the same activity is equal to 0.43. The probability that they correspond to the same activity *Home* is equal to 0.36. We are far from our expectations as we could not infer *Home* and *Work* places for all the users.

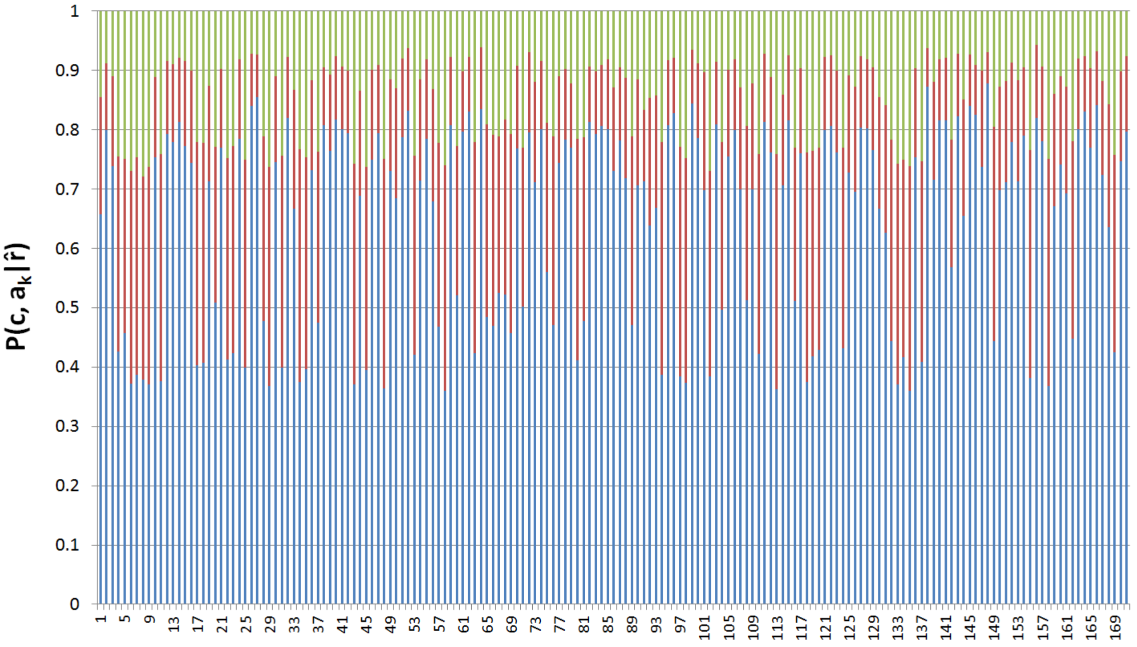


Figure 4.5: P($a_k|DDR^t$) for the two most visited clusters. Blue: a_k = *Home*; Red: a_k = *Work*; Green: a_k = *Other* .

Nevertheless, these results are coherent if DBSCAN has generated several clusters around the users' home address.

Results with known truth

Twenty people gave the addresses of their home, work, first and second grocery. For each respondent, we verify that:

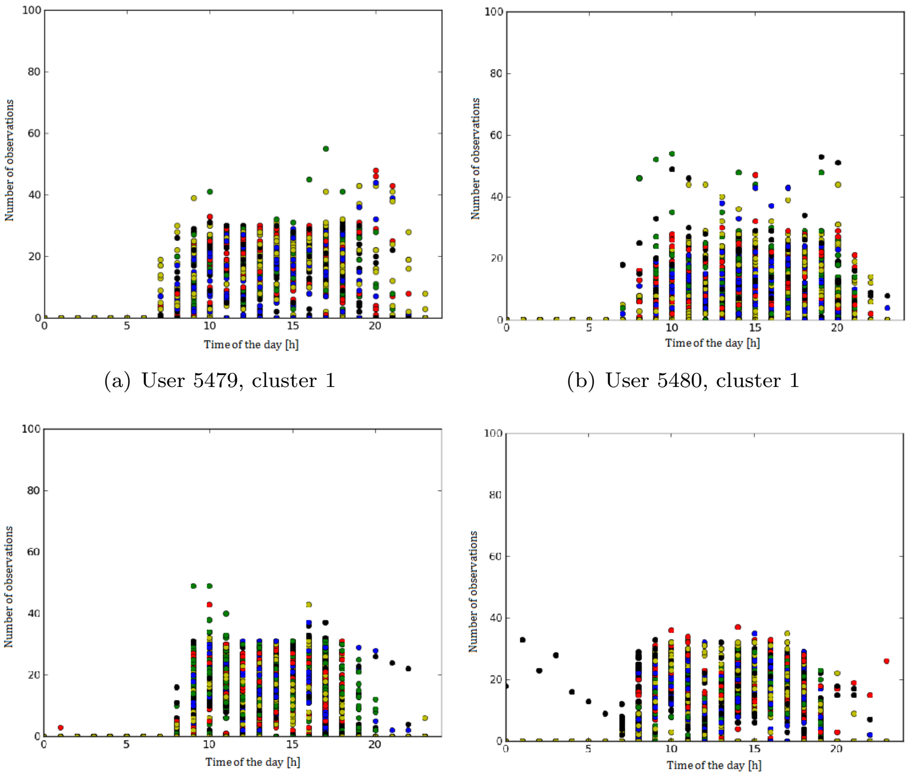
- one of the two most visited clusters corresponds to an address of the respondent
- the highest probability P($a_k|DDR^t$) corresponds to the activity announced by the user at the address.

We summarize the results in Table 4.2. In most cases, one of the two most visited clusters corresponds to the given home or work address. We note that the algorithm associates the activity *Home* to 9 work places. The associated cells are shaded (Table 4.2). Looking at user 5480, the algorithm proposes the activity *Home* with a probability of 0.72. The above signify that the prior might not be appropriate. The activity *Home* seems to come out too often.

Table 4.2: Results for the two most visited places with the prior based on people's schedule

user_id	cluster	Located at the address:	P($Home\|DDR^t$)	P($Work\|DDR^t$)	P($Other\|DDR^t$)	Highest probability
5988	1	Work	0.36	0.38	0.26	*Work*
	2	Grocery 2	0.59	0.29	0.12	*Home*
6094	1	Work	0.36	0.38	0.26	*Work*
	2	Home	0.75	0.14	0.11	*Home*
6093	1	Work	0.42	0.33	0.25	*Home*
	2	Home	0.80	0.10	0.096	*Home*
5462	1		0.79	0.11	0.08	*Home*
	2	Work	0.44	0.30	0.26	*Home*
5976	1	Home	0.73	0.15	0.12	*Home*
	2	Work	0.37	0.40	0.23	*Work*
5974	1	Work	0.36	0.39	0.25	*Work*
	2	Home	0.72	0.13	0.15	*Home*
5942	1		0.51	0.26	0.23	*Home*
	2	Home	0.76	0.15	0.09	*Home*
5937	1	Home	0.77	0.15	0.08	*Home*
	2	Work	0.37	0.35	0.28	*Home*
5947	1	Home	0.78	0.14	0.08	*Home*
	2	Work	0.39	0.35	0.26	*Home*
5960	1	Work	0.40	0.35	0.25	*Home*
	2	Home	0.78	0.11	0.09	*Home*
5451	1		0.65	0.19	0.14	*Home*
	2		0.71	0.14	0.14	*Home*
6066	1		0.36	0.40	0.24	*Work*
	2	Work	0.34	0.40	0.26	*Work*
6061	1	Work	0.42	0.33	0.24	*Home*
	2	Home	0.76	0.15	0.09	*Home*
6027	1		0.79	0.11	0.10	*Home*
	2	Home	0.77	0.12	0.11	*Home*
5965	1	Home	0.82	0.11	0.07	*Home*
	2	Work	0.37	0.39	0.24	*Work*
6031	1	Home	0.78	0.12	0.09	*Home*
	2	Work	0.37	0.38	0.25	*Work*
5479	1	Work	0.43	0.33	0.24	*Home*
	2	Home	0.84	0.09	0.07	*Home*
6030	1	Home	0.73	0.14	0.13	*Home*
	2	Work	0.37	0.39	0.24	*Work*
6005	1	Home	0.77	0.14	0.09	*Home*
	2		0.76	0.13	0.11	*Home*
5480	1	Work	0.46	0.29	0.25	*Home*
	2	Work	0.72	0.18	0.10	*Home*

The hourly number of measurements is plotted in Fig. 4.6 for selected clusters. The plots correspond to work places that the algorithm tends to associate to *Home* but we observe patterns that clearly correspond to a work activity. Then, the fact that these clusters are associated to *Home* does not come from the likelihood but from the prior. With a value of 0.715 for the activity *Home*, the impact of the likelihood on the results diminishes. It is hence reasonable to think of an alternative prior.



(a) User 5479, cluster 1

(b) User 5480, cluster 1

(c) User 5947, cluster 2

(d) User 5960, cluster 1

Figure 4.6: Hourly number of measurements for each day. Monday: green, Tuesday: red, Wednesday: black, Tuesday: blue, Friday: yellow.

4.2.2 Prior based on land use

Results with known truth

When the prior is deduced from the peoples' time budget, some clusters located at a work address are associated to *Home* even if the WiFi traces reflect working hours. We found out that this is due to the high value of P(*Home*). Hence, we suggest to compute the prior based on the attractivities (see Section 3.3.2). For each of the twenty respondents whose addresses are known, we determine:

- the clusters that correspond to an address of the respondent
- if the highest probability $\mathrm{P}(a_k|DDR^t)$ of these clusters corresponds to the activity announced by the user.

Table 4.3: Results with the prior based on land use data

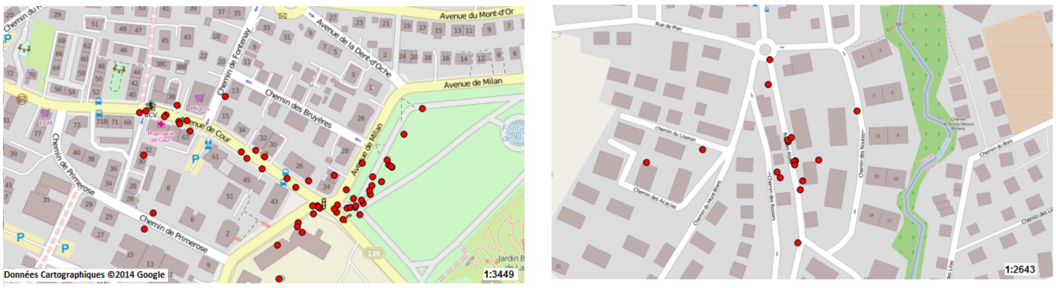
user_id	cluster	Located at address:	Highest prob(activity)	Highest prob(value)
5988	1	Work	W	0.73
	2	Grocery 2	W	0.55
	37	Grocery 2	W	0.69
6094	1;3;4;6;16	Work	W;W;W;W;W	0.88;0.98;0.72;0.85;0.93
	2	Home	H	0.56
6093	**1**	Work	**W**	0.86
	2	Home	H	0.84
	2	Grocery 1	H	0.84
5462	7	Home	H	0.82
	2;5;6;8;9	Work	**W**;W;W;W;W	0.80;0.84;0.90;0.88;0.79
5976	1;8	Home	H;H	0.85;98
	2	Work	H	0.40
5974	1;4;5;7 9;16	Work	W;W;W;W W;W	0.88;0.94;0.90;0.98 0.61;0.79
	2;10	Home	H;H	0.75;0.95
5942	4;17	Work	W;W	0.99;0.95
	2;3	Home	H;H	0.49;0.66
	11;14	Grocery 1	S;S	0.41;0.45
5937	1	Home	H	0.68
	2	Work	**W**	0.70
5947	1	Home	H	0.43
	2	Work	**W**	0.70
	5;20	Grocery 1	L;S	0.37;0.39
	41	Grocery 2	S	0.72
5960	**1**	Work	W	0.72
	2	Home	H	0.89
6066	2	Work	W	0.72
6061	**1**;4;7;11;15 16;17;18;19;20	Work	**W**;W;W;W;W W;W;W;W;W	0.86;0.84;0.97;0.96;0.87 0.93;0.84;0.72;0.96;0.91
	2;3;5;14	Home	H;H;H;H	0.58;0.36;0.74;0.84
	13	Grocery 1	S	0.47
6027	2;15	Home	H;H	0.87;0.99
5965	1	Home	H	0.64
	2	Work	W	0.74
	6	Grocery 1	S	0.42
6031	1	Home	H	0.49
	2;3;7;11;14	Work	W;W;W;W;W	0.89;0.87;0.94;0.94;0.81
	1	Grocery 1	H	0.49
5479	**1**;6;10;16;27;28	Work	**W**;W;W;W;W;W	0.91;0.90;0.60;0.84;0.99;0.88
	2;3	Home	H; H	0.81; 0.96
6030	1	Home	H	0.42
	2;6;35	Work	W;W;W	0.99;0.35;0.81
	1	Grocery 1	H	0.42
	16	Grocery 2	S	0.47

6005	1	Home	H	0.56
	3;5;7;8	Work	W;W;W;W	0.82; 0.87;0.99;0.85
5480	**1**;**2**;3;14	Work	W;W;W;H	0.90;0.80;0.85;
	5;7	Home	H;H	0.92;0.90
	9	Grocery 1	S	0.67

The results are summarized in Table 4.3. The second column of the table enumerates the clusters that are near to the address given in the third column (the clusters are ordered according to their number of observations). In accordance with the clusters enumeration, the fourth and five columns show the activity that the clusters are the most likely to be and the corresponding value of the probability. We denote with bold letters the places that are shaded in Table 4.2. We notice that they are now associated to the activity *Work*.

An overall view of the results proves that the algorithm works well. Nevertheless, we can observe a few singularities. Cluster 2 of user 5976 is located at the work address given by the respondent. The cluster was likely to correspond to *Work* with the first algorithm but now, it is more likely to be *Home*. It doesn't come from the likelihood because the WiFi traces reflect a work activity. We suppose that the prior does not have optimal values because the *DDR* of the cluster is too big. On Fig. 4.7, we can see that the APs are spread over several streets. This implies that useless attractivities are taken into account in the prior calculation. This problem also comes up for the cluster 1 of users 6031 and 6030. The cloud of APs is so big that the cluster corresponds to two addresses given by the user: home and the first grocery.

We also notice that we don't always find places at the four addresses given by the users. It is the case for 21 addresses, mainly for the groceries. The absence of clusters can be due to values chosen for the DBSCAN parameters. It is also possible that people may have not provided their real first and second most important grocery addresses. Few times, these addresses were not provided at all by the users. We should also consider the fact that people might had moved. These addresses were not provided at all by the users. An example of such a case is illustrated in Fig. 4.8.



(a) User 6030, cluster 1

(b) User 5976, cluster 2

Figure 4.7: Large clusters.

Results for the whole population

The twenty persons who provided their addresses represent 11 % of the sample. We now proceed with the results for all the participants. First, we compute the probability to perform an activity during each hour of the day. For each day and each participant, we look at the successive places that are visited. Each cluster has a probability to be either *Work*, *Home*, *Shopping* or *Leisure*. If a cluster c is visited during an hour h, $\mathrm{P}(a_k|h)$ is increased by $\mathrm{P}(a_k|DDR^t)$.

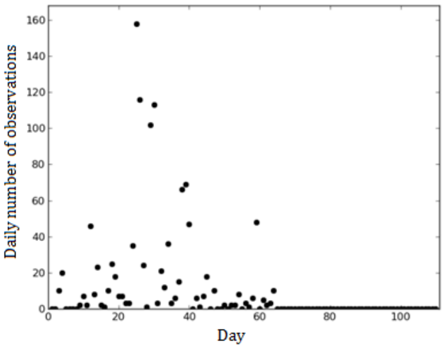


Figure 4.8: Total daily number of timestamps in WiFi traces.

$$P(a_k|h) = \frac{\sum_{i=1}^{179}\sum_{d=1}^{D}\sum_{e=1}^{E} e_{c,h} \times P(a_k|DDR^t)}{\sum_{i=1}^{179}\sum_{d=1}^{D} d} \tag{4.3}$$

where D is the total number of days covered by the WiFi traces for the person i, E the total number of activity episodes performed by i on day d, $e_{c,h}$ equals one when the user i is seen in c during hour h and $\mathrm{P}(a_k|DDR^t)$ is the probability that c corresponds to activity a_k. We obtained the curves shown in Fig. 4.9. We recognize the patterns that we deduced from the microsensus *Swiss mobility and transportation*, particularly for *Home* and Work. Hence, we can claim that the resulting curves make sens and that the algorithm gives coherent meanings to the places. We notice the peak for *Shopping* at 11 p.m. In Switzerland, shops and service are not open so late in the evening. It would be interesting to check if it would not correspond to shops located near to home addresses.

Moreover, we determined the daily activity sequences of people based on the algorithm output. For each day and each participant, we chronologically order the visited places and note the activity that the cluster is the most likely to be, according to the four computed probabilities. The most observed daily patterns are {*Home*}, {*Home*, *Work*, *Home*} and {*Home*,*Shopping*} with 9 475, 3 280 and 2 211 days respectively over a total of 38 258 days.

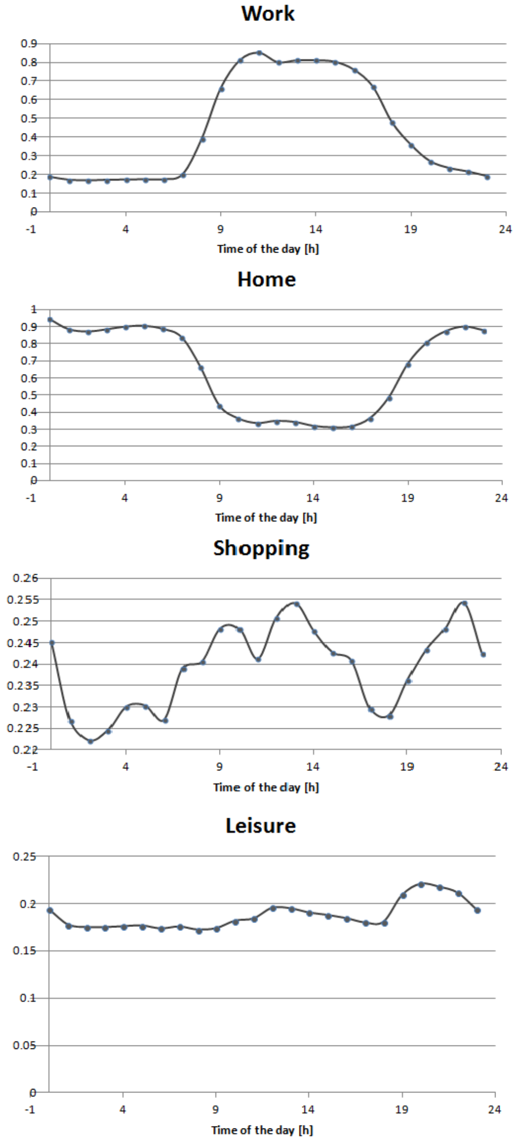


Figure 4.9: Probability to perfom an activity a_k according to the hour of the day based on the algorithm output (weekdays).

Chapter 5

Sensitivity analysis

5.1 Contribution of the land use data

The land use data comes in the second version of the algorithm as an alternative for the prior that was computed based on people schedule. To quantify the real benefit of the land use data, we run the algorithm as if we had absolutely no information about it, that is the prior is always equal to $\frac{1}{4}$. Fig. 5.1 illustrates the results obtained when the land use data is available for the 6 most visited locations of interest for few users, while Fig. 5.2 illustrates the results obtained with a constant prior. In general, the distribution of the probabilities over the activities changes when we consider the attractivities that surround a cluster. The reliability in the activity performed somewhere becomes much higher than if the land use data is ignored.

This is evident in Fig. 5.3. The likelihood of the Bayesian approach is sufficient to delineate the mobility patterns but the additional land use information emphasises them and reveals their real extent.

Table 5.1 presents the likelihood of each user computed when the prior is based on land use and when it is constant. The second column enumerates the clusters that correspond the most to each address given in the third column (H=Home, W=Work, S=Shopping). The likelihood is the multiplication of the corresponding probabilities. We observe that the likelihood is always lower when the attractivities of a cluster are ignored. We note the high likelihood of the user 5976, 6027 and 5474. They underline the high certainty of the results. The users 5937 is an example where the algorithm successfully determined the activity but with lower probabilities. The user 5988 is an example where the grocery address was badly inferred. It explains the lower value of the likelihood compared to the user 5480 which is also computed as the product of three probabilities.

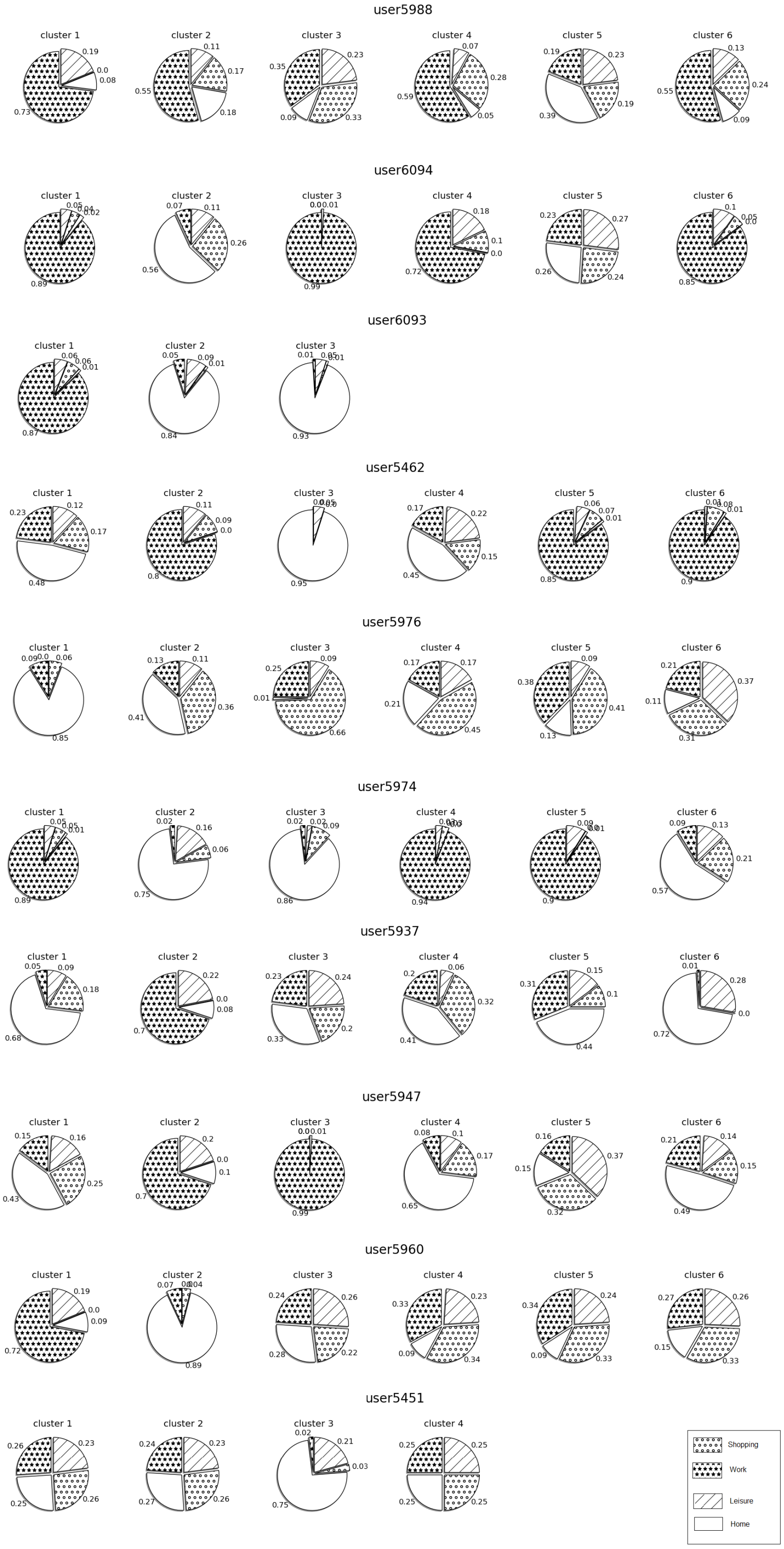


Figure 5.1: Output with prior based on the land use data.

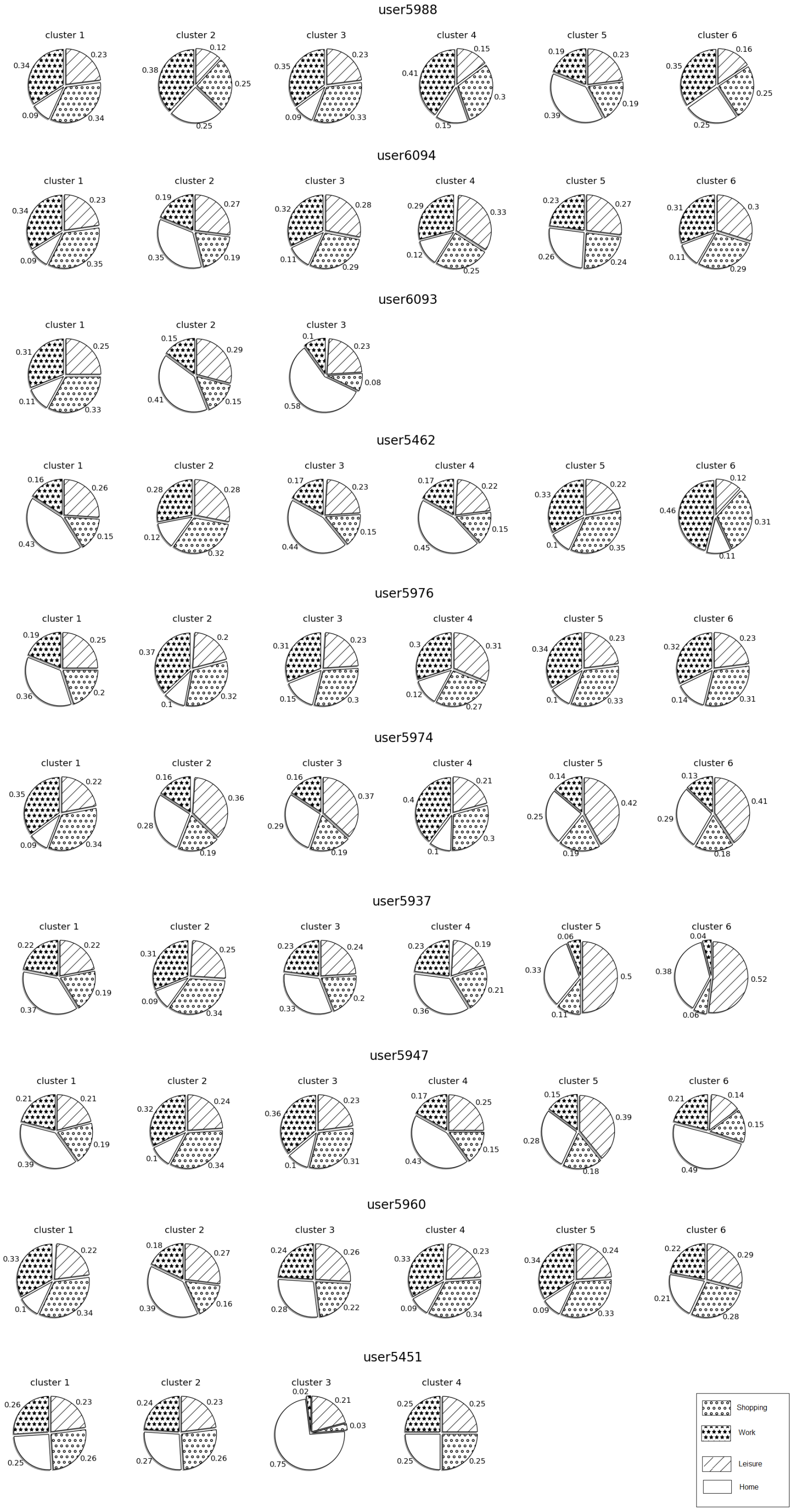


Figure 5.2: Output without land use data.

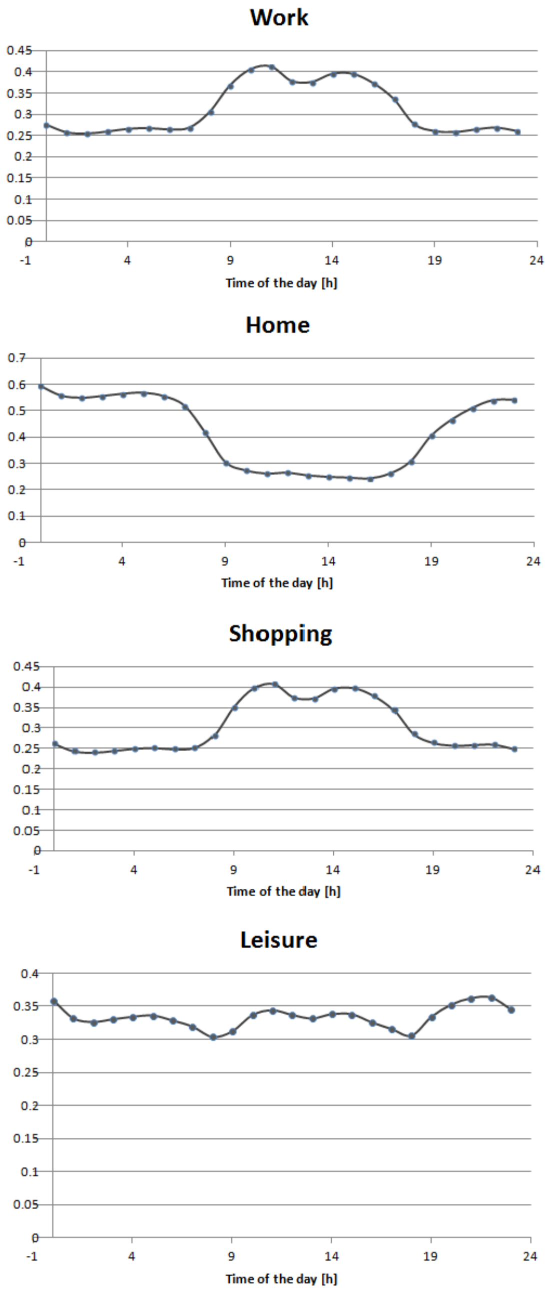


Figure 5.3: Probability to perfom an activity a_k according to the hour of the day based on the algorithm output (weekdays). The prior is constant and set to $\frac{1}{4}$

Table 5.1: Likelihood for each user

user_id	clusters	addresses	Likelihood(land use data)	Likelihood(constant prior)
5988	1,37	W,S	0.152	0.0933
6094	2;3	H,W	0.549	0.114
6093	1,2,2	W,H,S	0.0103	0.0091
5462	6,7	W,H	0.738	0.174
5976	2;8	W,H	0.127	0.043
5974	2,4	H,W	0.711	0.112
5942	3,4,11	H,W,S	0.272	0.0351
5937	1,2	H,W	0.476	0.117
5947	1,2,20,41	H,W,S,S	0.0864	0.00711
5960	15,1	H,W	0.676	0.0577
6066	2	W	0.722	0.355
6061	14,7,13	H,W,S	0.387	0.0639
6027	2	H	0.867	0.370
5965	1,2,6	H,W,S	0.198	0.0409
6031	1,7,1	H,W,S	0.0899	0.0205
5479	3,1	H,W	0.871	0.148
6030	1,2,1,16	H,W,S,S	0.0424	0.00726
6005	1,7,1	H,W,S	0.141	0.0176
5480	5,1,9	H,W,S	0.549	0.0354

5.2 Localization accuracy

In Section 3.3.2, we defined a radius R around each AP of a cluster. This radius represents the uncertainty of the APs'coordinates. To measure the impact of this radius on the results, we tested five values of R: 100 meters, 80 meters, 60 meters, 40 meters and 20 meters. The results appear in Table5.2. The likelihood changes with the radius R. We cannot observe a general tendency. For some users, the likelihood increases and for others, it decreases. The global observation is that the changes are not radical. The values stays close to the ones obtained for the initial run. Then, the localization accuracy is not a sensitive parameter. It would be interesting to look at the sensitivity of the clusters' size.

Table 5.2: Likelihood according to the radius R

user_id	Likelihood ($R = 100m$)	Likelihood ($R = 80m$)	Likelihood ($R = 60m$)	Likelihood ($R = 40m$)	Likelihood ($R = 20m$)
5988	0.152	0.153	0.155	0.164	0.169
6094	0.549	0.537	0.521	0.163	0.156
6093	0.0103	0.0039	0	0	0
5462	0.738	0.756	0.740	0.875	0.866
5976	0.127	0.119	0.109	0.108	0.117
5974	0.711	0.654	0.880	0.864	0.849
5942	0.272	0.118	0.123	0.117	0.117
5937	0.476	0.468	0.451	0.431	0.417
5947	0.0864	0.079	0.075	0.084	0.012
5960	0.676	0.683	0.685	0.696	0.720
6066	0.722	0.671	0.629	0.591	0.572
6061	0.387	0.394	0.382	0.448	0.163
6027	0.867	0.883	913	0.992	1
5965	0.198	0.199	0.232	0.238	0.271
6031	0.0899	0.092	0.088	0.082	0.022
5479	0.871	0.860	0.849	0.837	0.800
6030	0.0424	0.043	0.0416	0.038	0.031
6005	0.141	0.140	0.133	0.134	0.024
5480	0.549	0.559	0.560	0.551	0.476

Conclusion and further work

We designed a Bayesian approach to infer activity sequences and deduce the corresponding mobility patterns. Two alternatives are possible. The prior is either based on people's timetable or based on the land use data. The results show that the Bayesian approach works better when the prior is based on the land use data (the likelihoods are significantly higher) and that the results are good. We locate and correctly infer the semantic meaning of the addresses that few users provided. The high values of the likelihood also reveals that the inference is highly reliable. Moreover, the mobility patterns extracted from the results are clearly recognizable. This is particularly true for the activities *Home* and *Work*. The pattern for *Shopping* and *Leisure* also make sens. Finally, the most recurrent chains of activities are the ones that we could reasonably expect. Once more, it proves that the algorithm performs well.

Nevertheless, the work can be improved and extended. The first task would consist in improving the clustering algorithm. As we have seen, few clusters are so extended that they include several addresses. Then, we can imagine to include the frequency of visits to the clusters into the Bayesian approach. Finally, the set of activities $\hat{A}$ could be split into more details activities. This requires to have the needed information available.

Acknowledgements

I want to thank Mr. Bierlaire who allowed me to work in the laboratory TRANSP-OR of the Federal Polytechnic School of Lausanne (EPFL). He guided and encouraged me when presenting my results.

I particularly appreciated to work with Evanthia Kazagli and Antonin Danalet. Their involvement by my sides has been precious. I really appreciated their availability and their willingness to help me. I have worked in a very good environment.

I am grateful to Nicolas Richart of the computational solid mechanics laboratory and to Mathieu Martin who spent few hours to install the necessary softwares on my personal computer.

Bibliography and References

Ashbrook, D. and Starner, T. (2003). Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users, *Personal Ubiquitous Comput.* **7**(5): 275–286.

Bierlaire, M. and Frejinger, E. (2008). Route choice modeling with network-free data, *sportation Research Part C: Emerging Technologies* **16**(2): 187 – 198.

Blunck, H., Kjærgaard, M. and Toftegaard, T. (2011). Sensing and Classifying Impairments of GPS Reception on Mobile Devices, *in* K. Lyons, J. Hightower and E. Huang (eds), *Pervasive Computing*, Vol. 6696 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 350–367.

Bowman, J. L. (1998). *The Day Schedule Approach to Travel Demand Analysis*, PhD thesis.

Buisson, A. (2012). Potential of mobility learning from smartphone wifi data.

Buisson, A. (2013). Identify user's location of interest from smartphone WiFi data.

Calabrese, F., Diao, M., Lorenzo, G. D., Jr., J. F. and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C: Emerging Technologies* **26**(0): 301 – 313.

Danalet, A., Farooq, B. and Bierlaire, M. (2014). A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures, *Transportation Research Part C: Emerging Technologies* **44**: 146 – 170.

Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *in* E. Simoudis, J. Han and U. Fayyad (eds), *Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Portland, Oregon, pp. 226–231.

Gindraux, M. (2007). La mobilité en suisse 2005, rapport technique: plan d'échantillonage, taux de réponse et pondération, *Technical report*, DFI, OFS, DETEC, ARE.

Gonzalez, M., C.A., H. and Barabasi, A.-L. (2008). Understanding individual human mobility patterns, *Nature* **453**: 479–482.

Hasan, S., Schneider, C. M., Ukkusuri, S. V. and González, M. C. (2013). Spatiotemporal Patterns of Urban Human Mobility, *Journal of Statistical Physics* **151**(1-2): 304–318.

Hightower, J., Consolvo, S., LaMarca, A., Smith, I. and Hughes, J. (2005). Learning and Recognizing the Places We Go, *Proceedings of the 7th International Conference on Ubiquitous Computing*, UbiComp'05, Springer-Verlag, Berlin, Heidelberg, pp. 159–176.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira, Jr., J., Frazzoli, E. and González, M. C. (2013). A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities, *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, ACM, New York, NY, USA, pp. 2:1–2:9.

Kang, J. H., Welbourne, W., Stewart, B. and Borriello, G. (2004). Extracting Places from Traces of Locations, *Proceedings of the 2Nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, WMASH '04, ACM, New York, NY, USA, pp. 110–118.

Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. and Laurila, J. (2010). Towards rich mobile phone datasets: Lausanne Data Collection Campaign, *Proc. ACM Int. Conf. on Pervasive Services (ICPS,',','), Berlin.*

Laurila, J. K., Gatica-Perez, D., Aad, I., J., B., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J. and Miettinen, M. (2012). The Mobile Data Challenge: Big Data for Mobile Computing Research, *Pervasive Computing.*

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W. and Ma, W.-Y. (2008). Mining user similarity based on location history, *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, ACM, New York, NY, USA, pp. 34:1–34:10.

Marmasse, N. and Schmandt, C. (2000). Location-Aware Information Delivery with ComMotion, *Proceedings of the 2Nd International Symposium on Handheld and Ubiquitous Computing*, HUC '00, Springer-Verlag, London, UK, UK, pp. 157–171.

Miller, H. (2014). Activity-based analysis, *in* M. M. Fischer and P. Nijkamp (eds), *Handbook of Regional Science*, Springer Berlin Heidelberg, pp. 705–724.

Montini, L., Rieser-Schüssler, N. and Axhausen, K. W. (2014). Personalisation in multi-day gps and accelerometer data processing, *4th Swiss Transport Research Conference*, Monte Verità, Ascona.

Montoliu, R. and Gatica-Perez, D. (2010). Discovering Human Places of Interest from Multimodal Mobile Phone Data, *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, ACM, New York, NY, USA, pp. 12:1–12:10.

OFS (2007). La mobilité en suisse, résultats du microrecensement 2005 sur le comportement de la population en matière de transports, *Technical report*, OFS, ARE.

OFS (2012). La mobilité en suisse, résultats du microrecensement mobilité et transports 2010, *Technical report*, OFS, ARE.

Pas, E. I. (1982). Analytically derived classifications of daily travel-activity behavior: description, evaluation, and interpretation, *Transportation Research Record* **879**: 9–15.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R. and Ratti, C. (2010). Activity-aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data, *Proceedings of the First International Conference on Human Behavior Understanding*, HBU'10, Springer-Verlag, Berlin, Heidelberg, pp. 14–25.

Schönfelder, S., Ethz, I., Samaga, U., Dresden, T. U., Schönfelder, S., Samaga, U. and Dresden, T. U. (2003). Where do you want to go today?- More observations on daily mobility, *STRC 03 Conference paper Session Mobility* .

Zaugg, H.-U., Siegenthaler, C. and Humbel, R. (2007). Recensement fédéral des entreprises, secteurs secondaire et tertiaire, *Technical report*, DFI, OFS (Infrastructure statistique section géoinformation).

Zaugg, H.-U., Ullman, D. and Spahn, D. (2012). Statistique de la population et des ménages (STATPOP) dès 2010, *Technical report*, DFI, OFS, DTE.