# Clustering Weekly Patterns of Human Mobility Through Mobile Phone Data

Etienne Thuillier, Laurent Moalic, Sid Ahmed Lamrous, Alexandre Caminada

# Clustering weekly patterns of human mobility through mobile phone data

Etienne Thuillier[1], Laurent Moalic[1], Sid Lamrous[1], Alexandre Caminada[1]
[1]Univ. Bourgogne Franche-Comté, UTBM, OPERA, F-90100, Belfort, France

**With the rapid growth of cell phone networks during the last decades, call detail records (CDR) have been used as approximate indicators for large scale studies on human and urban mobility. Although coarse and limited, CDR are a real marker of human presence. In this paper, we use more than 800 million of CDR to identify weekly patterns of human mobility through mobile phone data. Our methodology is based on the classification of individuals into six distinct presence profiles where we focus on the inherent temporal and geographical characteristics of each profile within a territory. Then, we use an event-based algorithm to cluster individuals and we identify 12 weekly patterns. We leverage these results to analyze population estimates adjustment processes and as a result, we propose new indicators to characterize the dynamics of a territory. Our model has been applied to real data coming from more than 1.6 million individuals and demonstrates its relevance. The product of our work can be used by local authorities for human mobility analysis and urban planning.**

*Index Terms*—**Clustering algorithm, Human mobility, Mobile communications, Territory dynamics, User profiles**

## I. Introduction

The use of cell phone network data for urban and human mobility studies have been proposed and largely developed during the last decades [1],[2]. Cellphone networks were considered as a more reliable data source in comparison with household mobility surveys and road sensors. In this sense, the rapid growth of cell phone networks, the deep penetration and the even deeper daily usage of mobile phones in the world [3], genuinely made cellular network data an unavoidable candidate for the development of a large panel of decision support tools and human mobility simulation tools. But although these sources represented a new opportunity for crowd tracking and road traffic estimates, their reliability have been largely discussed [4],[5]. One of the main potential drawbacks with cellular network data is the high dependence on the users' behaviors: the more the users use their phones, the more data are generated. And the more one can leverage these data.

Initially, the reuse of cellular data with other goals than customer billing, was to exploit the users' locations to estimate the road traffic and to plan road network infrastructures. The relatively easy accessibility of cellular data attracted other research fields, and many studies using individuals' locations emerged. Thus, different aspects of cellular network data have been used for different researches [6]. Among them we find; studies on patterns of mobility [7],[8], large scale studies of urban mobility [9],[10],[11],[12] conception of human mobility models [13],[14],[15],[16], studies on land usage [17],[18], etc. Although the literature abounds with examples of cellular data usages, the suitability of such data to identify and characterize human mobility is discussed [19], and many inherent aspects of these data such as granularity and coarse location estimates, may disrupt mobility models. Recent works propose to merge different types of mobility data in order to avoid limitations of single view models [20], [21].

In this paper, we focus on using cellular data to understand the human mobility patterns observable on a territory. Our characterization covers three different aspects: a user profile geographical aspect (origin and destination), a user's temporal aspect (when do people move), and a flow intensity aspect. We propose a novel approach to address the human mobility understanding through mobile phone data. Firstly, we characterize the cellular network customers and their relation with the territory with a profiling process based on Call Detail Records (CDR). In [21] however, the authors argue that mobility models driven by CDR data are limited by the individuals' activity during the modeling time. To avoid this problem, we propose to cluster individuals' behaviors into weekly patterns of mobility. Our methodology then identifies 12 weekly patterns, and classifies more than 1.6 million individuals into these global patterns. In order to evaluate the relevance of using only mobile phone data, we compare our results with existing National Census and surveys. We use our results to dress population estimates, and we propose a temporal, geographical and community vision of a territory that can easily be verified with national data and reproduced in other places. Finally the contributions of our work are:

- We propose a user profiling algorithm, classifying individuals into 6 distinct groups from CDR data.
- We present a novel approach to extract weekly patterns of mobility from millions of CDR data.
- We propose 12 global weekly patterns, summarizing the mobility behaviors of 80% of studied individuals.
- We validate our approach with data from three national sources and show the consistency of our methodology.
- We propose new indicators of mobility which provide strong added-values for mobility analysis.

The paper is structured as follows. In the section II we propose to get back on some basis of cellular networks and cellular data. Section III presents the profiling process that we used to leverage cellular data. In section IV we show the results of our clustering experiment. In section V we propose

a comparative analysis with National Census data to validate our results and new insights on territory dynamics. The section VI concludes our results.

## II. Basic concepts

### A. Cellular network and CDR

Cellular networks are based on a single concept: to allow the exchange of information from one mobile device to another one through linked base stations. If the cellular network detects the geographical position of the mobile equipment, it can thus know the position of the equipment holder, human or machine.

Today the most worldwide used cellular network technology is the GSM (Global System for Mobile communications). A GSM network is composed of several base stations that transmit a radio signal into a geographical zone. Any mobile equipment within that area that receives enough field strength is connected to the network, and thus reachable for communications. The size of each cell (antenna coverage) depends on the network design and may vary from 300 meters wide (microcells in an urban environment) to 30 kilometers wide (macrocells in rural environment). All adjacent cells are overlapping, allowing a continuous connection to the network when the mobile equipment is moving. Many adjacent cells are grouped in zones identified by a Local Area Code (LAC).

For billing purposes, when a mobile equipment is used (call, messages, data, etc.), the mobile phone operators store the data about the mobile equipment: identification code, location, type of event, day and hour, service, duration of the connection, etc. Those records are called Call Detail Records (CDR). We may generalize and say that all CDR contain at least three basic information which are:

1) an IMEI (International Mobile Equipment Identity), which is unique and identifies a terminal
2) a timestamp, as day and hour
3) a cell Id, the network operator identity code for the cell

The combination of those three components is particularly useful for mobility studies since they represent an object in time and space. We easily understand that the combination of several CDR sharing the same IMEI represents the trajectories done by that mobile equipment during a time period.

### B. Challenges of using CDR data

CDR are registered according to mobile equipment events on the network [22],[23]. Thus, they are a pertinent indicator of presence in an area at a certain time, but their usage as a continuous indicator of presence is relative and depends on the frequency of the CDR. Indeed, CDR data highly depend on users' behaviors and cannot detect the presence of residents without cellphone data [20], [21]. This implies that their usage as a marker of mobility can be discussed. Moreover, a CDR links a mobile equipment to a certain spatial area, and the location of the mobile depends on the size of the cell which is space dependent and not strictly known. We can summarize the main problems of CDR:

1) Their number and frequency depend on the mobile equipment holder usage.
2) The location of the equipment depends on the size of the cell.

To counter these problems, many works propose specific tools to generate synthetic CDR data from actual mobile network data and from survey data with fair results [15], [16]. In [20] and [21] the authors propose to use multiple mobility data and to merge CDR data with complementary mobility information to model human mobility. They show great results when compared with the synthetic tool proposed in [15]. More generally, generation of synthetic data and multi-source data fusion are current research trends, especially in mobility analysis. These developments facilitate the access to trustful and relatively free mobility data.

Although CDR may present different problems, they also offer several advantages. The first one is the deep penetration of cellphone in the population, which makes them an omnipresent and reliable source of mobility data. [23] identifies a second advantage which is a passive or active data collection methodology. We can also argue that, unlike other kind of technologies (Bluetooth sensors, road magnetic sensors, road cameras, National Census, etc.), mobile networks are already widely spread on territories, and do not need specific equipment deployment and additional costs.

### C. Event category

We previously stated that a Call Detail Record corresponds to the storage of specific information such as mobile phone identification code, date, time, cell identification code, etc. Nonetheless, this definition is incomplete: a CDR corresponds to one interaction between the network operator and the mobile equipment. We count six different events types and three different event categories for any digital cellular system.

*Owned events*:

- Emission or reception of a phone call
- Emission or reception of a message

*Transition events*:

- Handover during a phone call
- Location area transition

*Induced events*:

- Every 3 minutes during a phone call
- Every 3 hours if none of the above events are done

## III. User profiling methodology

### A. data set description

For this research, we use a data set of real CDR consisting of more than 800 million records collected from a three week period (month of October 2014). These data contain about 1.6 million distinct IMEI with both national and international mobile phones (roaming). The data set was collected in a territory around Paris in a 130,000 inhabitants suburban area with buildings, houses, stores, companies including railway and highway infrastructures.

The construction of the data set was done as follows: we determined a geographical region on which we proposed to focus our study (approximately shaped as the administrative area of the studied territory of 70 $km^2$). We registered all the

network cells that overlapped this region in a subset that we call $S_1$. We also prepared a subset $S_2$ that contains all cells from the Ile-de-France province which includes the study area. Then, during the three week period, for each mobile phone which was at least detected once in $S_1$, we stored all the past and future events of this mobile phone within the limits of the three week period, and within the geographical limit of subset $S_1 \cup S_2$. With all these data, we are able to better understand the mobility flows of each individual that crossed this territory.

For privacy purpose the data set was separated according to the three different weeks, anonymizing all IMEI with three different keys, giving three one-week subsets. With the same privacy concern, all information about the type of events were erased, all IMEI with less than 10 records per week were erased, and international mobile phone records were kept only if the total number of distinct IMEI from that country was above 10. In the end each sub-data set on one week consists of more than 175 million of CDR which corresponds to more than 600,000 distinct IMEI. Each CDR is composed as follows: *<IMEI, Date, Time, Network cell Id, Roaming information, Country Code (if roaming)>*

Finally, we consider the period between two consecutive events as a period of presence or absence within the territory $S_1$. As due to the nature of CDR data, it is not possible to accurately calculate the presence period of an individual, we deliberately consider that between two consecutive events one individual is still attached to the last corresponding antenna (discrete model).

### B. User profiles

For local authorities, territory dynamics are closely linked to individuals that use local infrastructures for living, working, circulating, etc. Having the possibility to evaluate and quantify the usage made of its road network and public infrastructures, like car parks, access ways, malls, schools, is important for decision making processes. Many studies showed that CDR can be used to profile individuals that are present in a territory, or to qualify certain areas of a territory. In [24] the authors use a system based on a top-down and a bottom-up algorithm to classify individuals on four distinct profiles: Residents, Commuters, In transit and Visitors. We find the same approach in [25] where individuals are characterized according to their calling behaviors into Residents, Commuters and Visitors. In early works, [26] analyzed the daily rhythms of commuters in a territory by computing distances between home and work places. These distances often called radius of gyration are used as a mobility indicator for mobility studies [7],[8]. Then, although admitting that roaming information was not the most pertinent, [27] used CDR to evaluate the proportion of tourists in a territory, whereas [28] used phone calls, ticketing and online photos information to extract tourist statistics. Such individual profiling has also been developed in [10] which leverages cellular data to detect communities inside a city.

More recently, many works studied the daily profiles of individuals. [29] proposes to analyze the results of an activity-based travel survey by the clustering of the individuals according to their *weekday* and *weekend* activity patterns. [30] uses CDR and survey data from Paris and Chicago to detect 17 daily mobility profiles. These mobility profiles represent the daily patterns of 90% of individuals present on a territory. As well, [31] proposes to identify the mobility profiles by analyzing filtered CDR from Singapore.

In this paper, we are interested in the description and evaluation of the practices made of a territory by individuals. The chosen area covered by $S_1$ is particularly known for its numerous big companies and its access points to the city of Paris (train and roads). To best characterize this territory we present six different and exclusive daily profiles that are inferred from the CDR data set, and more specifically from the daily patterns of individuals within the territory:

1) *Resident Working in Zone* (RWiZ): a *RWiZ* is an individual that lives, sleeps and stays (work, study, etc.) in the territory covered by $S_1$ during the day
2) *Resident Working out of Zone* (RWoZ): a *RWoZ* is an individual that lives and sleeps in the territory $S_1$, but is out of the territory during the day
3) *Commuters* (C): a commuter is an individual that lives and sleeps outside but works or study in the territory $S_1$
4) *Multiple Single Transit* (MST): a *MST* is an individual that crosses the territory and appears irregularly several times during the day for more than one hour of daily presence
5) *One Single Transit* (OST): an *OST* is an individual whom the total consecutive daily presence does not exceed one hour
6) *Weekend* (WE): a *WE* is an individual that is pre-sent in the territory only during the weekend

Each of the listed profiles contains important information for the characterization of a territory. For example, the presence of residents (*RWiZ*, *RWoZ*) indicates that there are houses or residential areas within the territory. Thus, there is a need for administrative infrastructures (school, local authorities) but also supermarkets and night car parks. On the contrary, commuters will need rapid entrance and exit ways, but also day time car parks. People "in transit" (*MST*, *OST*) will want to quickly cross the territory, thus they may need secondary itineraries such as highways or trains. Finally *WE* will need accommodations for the weekend, easy access ways, and maybe tourist areas.

In addition to these patterns we set the index *A* for absent to all IMEI of the zone $S_1$ which are not detected during a given day. Absent will further be considered as the seventh category of profile even if it is not a real profile at all.

### C. Profiling algorithm

Our main concern for this work is to be able to characterize the individuals that are actors and users of territory infrastructures. According to the spatiotemporal pattern of each individual we propose to classify them into a distinct profile for each day of the week. A classification algorithm is used in [24], which uses temporal patterns of individuals to categorize them, but according to a whole week of data. The main problem with this approach is that it is generalizing and reducing individuals to a global definition. Unfortunately,

although human mobility patterns showed to be repetitive [8], they highly depend on the day of the week (part-time jobs, displacements on specific days of the week, special working hours, etc.).

In this work we propose to study individuals according to their daily patterns, and then over a whole week. For that we present an algorithm that determines the profile of each individual's days. We call $p_i^d$ the profile of type $i$ associated to a day $d$, and $w$ $(p_i^1, p_i^2, \ldots, p_i^7)$ the list of the seven profiles over the week. The profile classification is based solely on the temporal patterns by analyzing the presence and absence periods within the territory $S_1$. For that we determine six exclusive classification rules depending on the possible succession of events within a day. Note that this algorithm is deterministic and does a hard mapping; a fuzzy version will be checked further.

The rules of this algorithm are given by the function *DefineProfile*, and figure 1 illustrates our profile classification process. For each individual we gather the daily set of CDR and start the algorithm: each diamond corresponds to a True/False condition. In the figure, if a condition is verified, then the algorithm follows the bold line. Otherwise, if the condition is not verified it follows the hashed line until a profile is set.

**Function** `DefineProfile()`**:**
- **if** ***not is_in*** **then**
  - `/* Is present during the day (is_in) ? The individual is detected at least once in` $S_1$ `during the selected day */`
  - **return** ***Absent [A]***
- **if** ***is_we*** **then**
  - `/* Is present only on the weekend (is_we) ? The individual is detected only during the weekend. We define the weekend period as ]Friday 19:00 to Monday 06:00[ */`
  - **return** ***Week-end [WE]***
- **if** ***is_ost*** **then**
  - `/* Has only one transit (is_ost) ? The total daily presence does not exceed one hour (from the first to the last event of the day within` $S_1$`) */`
  - **return** ***One Single Transit [OST]***
- **if** ***is_mst*** **then**
  - `/* Has multiple transits (is_mst) ? There are several presence times in the day within` $S_1$`. Each of them do not last more than 1 hour, and they are spaced out of more than 3 hours */`
  - **return** ***Multiple Single Transit [MST]***
- **if** ***not is_res*** **then**
  - `/* Is resident (is_res) ? The individual is detected during the periods [00:00 - 06:00[ and [19:00 - 23:59[ within` $S_1$ `*/`
  - **return** ***Commuters [C]***
- **if** ***is_pres*** **then**
  - `/* Is continuously present (is_pres) ? All consecutive events within` $S_1$ `are spaced out of 3 hours maximum (there should be at least an event every 3 hours according to the location update operation of the network) */`
  - **return** ***Resident Working in Zone [RWiZ]***
- **return** ***Resident Working out of Zone [RWoZ]***

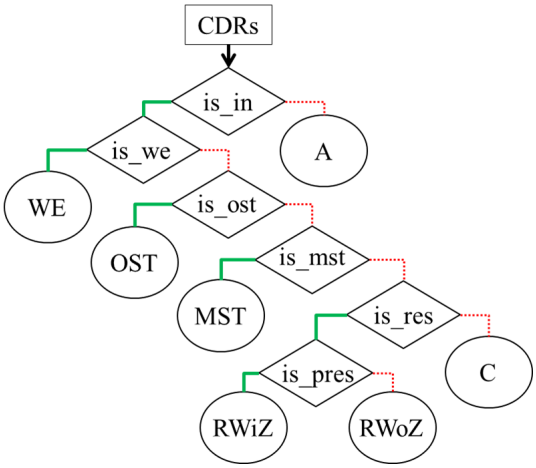


Fig. 1: The profiling algorithm

Additionally we propose some classification examples in figure 2. The hashed zones correspond to presence periods within $S_1$ during a specific day and for each example the corresponding profile is given. The last example is one possible representation of a WE profile for any Friday, while in addition the presence might be at any time on Saturday or Sunday.

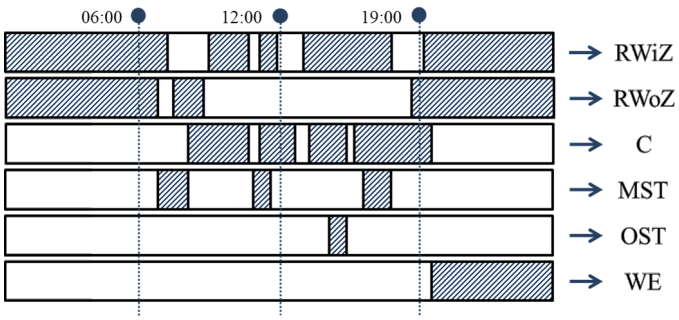


Fig. 2: Examples of profiling (Hashed boxes correspond to presence period within $S_1$)

### D. Users distribution

We use the profiling algorithm for each distinct IMEI of the data set. For that, the algorithm needs to survey all events done in one week by this IMEI. Once all events are stored in a chronological order, the different rules for the classification are computed, and for each day is attributed a unique profile. In the end all the IMEI of the data set are represented by a list of seven profiles, one per day.

We propose to study the distribution of the different profiles for each day. The figures 3 and 4 show the distribution of the profiles for all detected IMEI during the 21 days. For a better understanding figure 3 shows the accumulated curves of several profiles: *RWiZ* and *RWoZ* are summed up into "*Residents*", *OST* and *MST* profiles are summed up in "*Transit*". This was done in order to dress a general overview without surcharging the graph. First of all, the sum of the all the profiles' curves gives almost the same result between each week with a total variation of 0.658%. Thus, for the first week we count 534,820 distinct IMEI, 527,622 for the second week and 535,271 for the last week. As a matter of facts, it

is possible for an IMEI to be detected during one week and to be totally off the two other weeks, and thus not counting in the results. Consequently, the global attractiveness of the zone is very stable from one week to another. The intensity of the curves tells us that many users are regularly detected in the profile *Absent*. This does not mean that the algorithm did not manage to set a profile, but rather that those individuals are not present in the territory $S_1$ on this specific day. Then 60% of the detected IMEI are not present every day whatever the profile. The second main aspect is the repetitiveness of the curves over the three weeks, comforting the idea that human beings produce repetitive patterns, and that their behaviors are highly predictable.

Additionally, we observe a peak of individuals with profile *WE* during the weekend, but also a peak of *Absent* profiles, which means that a large part of individuals who are seen during the week are not seen at all over the weekend. Consequently we observe that there is a decrease in the *Commuters* and *Transit* profiles during the weekend. The result of these two observations is that this territory is a crossing and working area during the week, and this working force is not present during the weekend. Moreover, this territory attracts many visitors during the weekends; such visitors can be tourists, but also people that live and work outside of the territory like students returning home for the weekend. These *Weekend* individuals are present in almost the same proportions as the *Residents* accumulated (*In zone* and *Out of Zone*).

The figure 4 presents a detailed view of the figure 3, where each profile is clearly presented. We chose to not display the *Absent* profile for readability reasons. The first observation once again is the highly reproductive patterns for each profile. The curve representing the *RWoZ* profile is almost constant during the week except for Mondays and weekends where this profile is less represented. *RWiZ* profile on the contrary has an increasing curve all along the week with the highest level during the weekend. Similar end-patterns are observed for the *OST* and *MST* profiles, but in much less proportions.

Conclusions and interpretations on the users distribution are important for the comprehension of the dynamics of a territory. However, at this stage of our study we simply do a daily sum of all distinct IMEI for each user profile. This does not imply that the individuals reported within a profile during one day are the same the next day. On the contrary, individuals adopt different profiles during the week [8]. This leads to the next point that focuses on the "week patterns" of individuals.

## IV. Clustering experiment

### A. Clustering methodology

The previous graphs inform us on the distribution of the different profiles along three weeks of data but without taking into account the individuals' profiles changes from one day to another. We now wonder what is the repetitiveness of each individual within a profile. In this sense, we propose a totally new process to cluster the individuals according to their whole week of data, knowing that each day of the week has a specific human mobility behavior [8]. In [30] the authors showed that the human mobility patterns are similar over several days, but do not specifically study the patterns over one entire week. Thus they do not observe the multiple behaviors adopted by a same individual. Similarly, [29] presents *weekday* and *weekend* patterns without distinction between the days, and [31] reduces its data set to study averaged days and loses the features of each day. In our approach, we keep the features of each day, from Monday to Sunday, and characterize the territory through weekly patterns of human mobility.

The first step of this process is to identify the structures that can be used for the clustering. For that, we create a subset of data to run our clustering, and we propose to study the inherent limitations of the clustering. As we do not know in advance the structure of our data, we decide to use a *k-means* like approach. A *k-means* clustering algorithm is simple and easy to setup, and allows us to have a first look at our data without introducing bias. Note that different clustering approaches will be checked further. We do a cluster analysis of the population and extract the principal characteristics of these clusters. Then, we propose to analyze the structure of the resulting classes, and we link the different patterns with the dynamics of the studied territory. This methodology is based on the one proposed by [29] and [32].

#### 1) Representation of the individuals

**Vector for each day**

In our methodology we propose to give to each individual a profile for each day of the week. This means that each individual has seven profiles for one week. The profiles may be different each day, according to the relation of the individual with the territory. We propose here to consider each day of an individual as a binary vector of seven items where each item corresponds to a profile, and where the value of the item corresponds to the membership intensity of the individual to this profile. We call such a vector a *vector-day*. Logically for an individual the membership intensity range is composed of the two binary values $\{0, 1\}$. This means that for each vector-day only one item (i.e. one profile) of the vector can be set to 1, the others being put to 0. This furthermore implies that the total sum of the items values for a vector-day is 1. Figure 5 shows the representations of two vectors-day (Monday and Tuesday) for a random individual.

**Vector for each week**

We propose to also consider the week as a vector of seven items, where each item corresponds to a day of the week. We call such a vector a *vector-week*. Each day being represented by a vector-day with seven items, we can represent an individual's vector-week as a simple vector of 49 items. The total sum of all items of such vector has to give 7. For example, figure 6 shows the representation of an individual which is identified as *Commuters* during the week and *Absent* during the weekend. The advantages of using this kind of model, is that each individual is represented by a unique vector, which simplifies the computation and the classification of individuals.

#### 2) Production of a subset of data

In order to identify the similarities of the population individuals, we propose to sample the weeks by using a subset

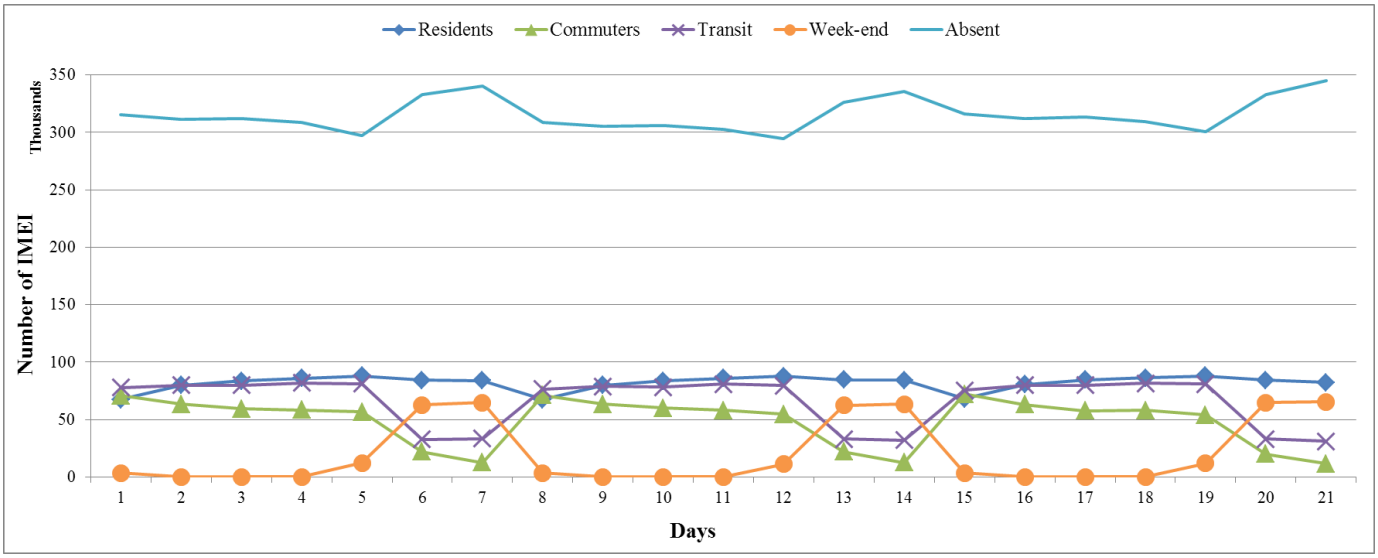


Fig. 3: Global overview of the profiles distribution

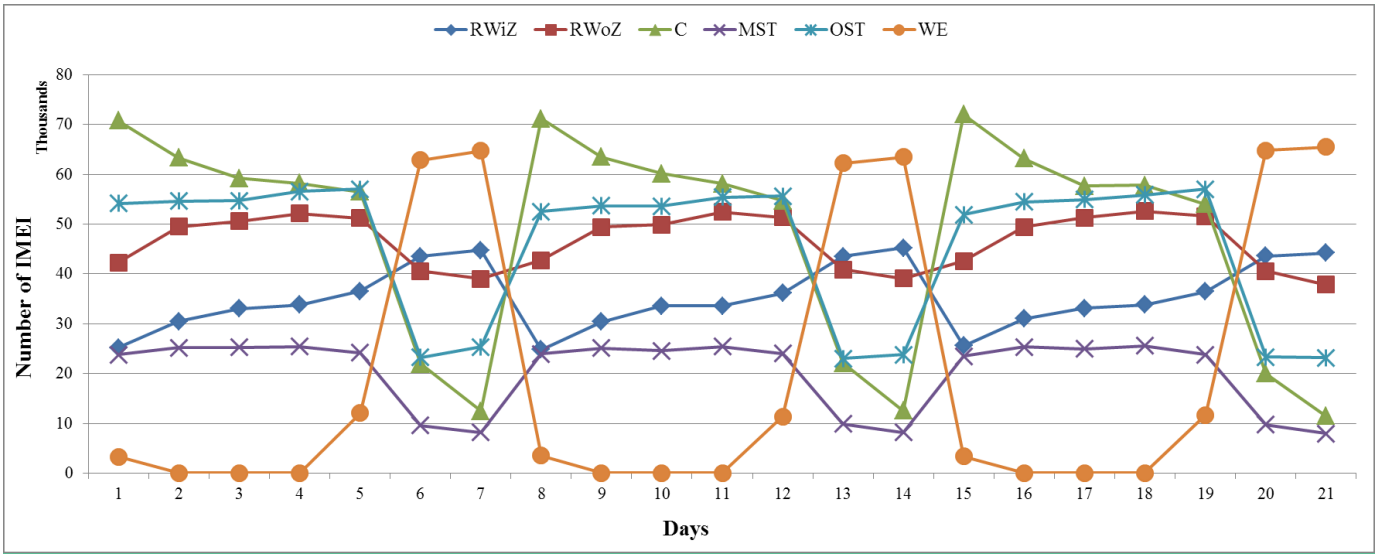


Fig. 4: Detailed view of the profiles distribution

| Day \ Profile | U | RiZ | RoZ | T | Ht | V | C |
|---|---|---|---|---|---|---|---|
| Monday | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Vector-day (Monday) | 0001000 | | | | | | |
| Tuesday | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Vector-day (Tuseday) | 0000100 | | | | | | |

Fig. 5: Representations of two vectors-day

| Day | Individual has profile | Corresponding vector day |
|---|---|---|
| Monday | C | 0000001 |
| Tuesday | C | 0000001 |
| Wednesday | C | 0000001 |
| Thursday | C | 0000001 |
| Friday | C | 0000001 |
| Saturday | U | 1000000 |
| Sunday | U | 1000000 |
| Vector week | 0000001000000100000010000001000000110000001000000 | |

Fig. 6: Representation of a vector-week

of individuals for each week. The sampling has proved its efficiency to allow a better generalization of the proposed classification and to avoid a useless overlearning of the data set [33]. For that we extract a panel of 10,000 random individuals (1.87% of the total data set) for each of the three weeks and run the cluster analysis on each of the three subsets. The main concern of working with a subset of data is the representativeness of the sample. In order to guarantee that the samples are representative, we propose to compare the profiles' distribution of each subset with the total profiles's distribution. In figure 7 we present two types of curves. The dotted lines represent the average number of profiles per day, for the three weeks. And the straight lines represent the

average number of profiles in the three subsets, multiplied by an adjustment factor i.e. the *Average number of individuals in a week* divided by *10,000*. For readability the profiles are grouped in four classes (*Residents*, *Commuters*, *Transit* and *Week-end*) and the Absent profile is not shown. On average, the number of wrongly estimated profiles for each day represents 1.46% of the data set. This is small enough to consider that our subsets are representative of the whole data set. Note that we repeated our cluster analysis with several random subsets and that we obtained similar results, for the representativeness of the subset, as well as for the derived conclusions of our clustering analysis which are developed later.

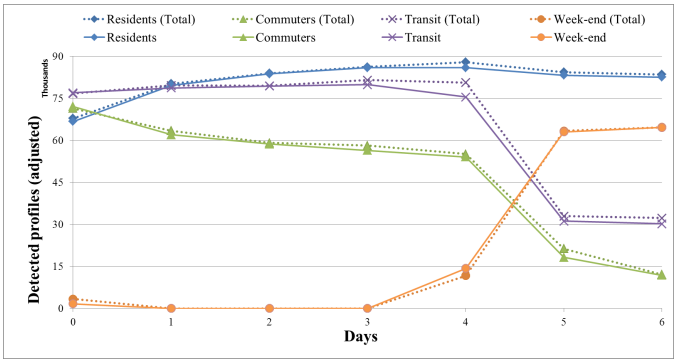


Fig. 7: Representativeness of the subsets

### B. k-means clustering

#### 1) Determination of k for k-means

Working with *k-means* has many benefits, but the main drawback in our case is to define the appropriate number $k$ of clusters. To find the best number $k$ we run several clustering tests where for each test we set the number $k$. To achieve that, we study a subset of 250 individuals chosen randomly that we cluster with $k$ ranging from 2 to 250. For each test of $k$ we compute the average silhouette of the solution obtained by the average of all silhouettes of the clustered objects. The silhouette is a good indicator of the quality of a clustering solution, and has been widely used in clustering analysis [34]. The silhouette of a clustered object shows if this object lies well or not within its cluster. It represents the distance ratio between the cluster it lies in, and the second-best possible cluster. The original formula is given by:

$$s(i) = (b(i) - a(i)) / \max(a(i), b(i)) \quad (1)$$

With $s(i)$ the silhouette of the object i, $b(i)$ the dissimilarity to the second-best cluster and $a(i)$ the dissimilarity to its own cluster.

Figure 8 shows the average silhouette obtained with a $k$ varying from 2 to 250 and 250 individuals. In the figure the dotted lines correspond to the best average silhouette obtained for each subset and for a specific $k$. The straight line corresponds to the average of the 3 subsets. We can observe that this last curve is plummeting when $k$ is small, then increasing to reach a maximum at one third of the abscissa limit and finally slowly decreasing. By looking at this curve, the best silhouette is obtained when $k$ is around 60 with 250 individuals.

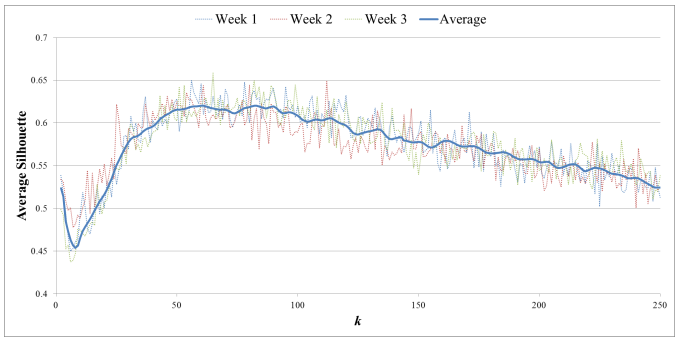


Fig. 8: Average silhouette for 250 individuals, k ranging from 2 to 250

However, we need a critical interpretation of these results. What this graph is showing is that the best average silhouette is obtained for a specific $k$. We have to understand what the silhouette of a cluster means. The dissimilarity of an individual is the relation between its own cluster and the next closest cluster in the solution. By increasing $k$ we increase the probability for an individual to be in the best suitable cluster and thus we increase the probability of having clusters made of very few and specific individuals. These individuals will surely present a good silhouette, but the figure 8 shows us that having a number $k$ too big decreases the general quality of the solution. To sum up, we need a number $k$ that allows a large number of possibilities, but small enough to have a good overall solution. Moreover, clustering with *k-means* means that we lately can classify individuals into distinct groups. A too large number of groups is difficult to analyze, and the different results would not have much signification. That is why, due to our preliminary test on 250 samples, we propose in our methodology to do the cluster analysis using three different numbers $k$ $\{20, 50, 70\}$, thus multiplying the experimental conditions, and avoiding local optimum problems. The resulting clusters are easier to interpret, and evaluating the relation between the individuals and the territory is simplified.

#### 2) Distance computation between clusters

The *k-means* algorithm is based on the proximity of an individual to the center of a cluster, often under the form of a distance. Here, all individuals are represented by a vector of 49 items holding binary information $\{0, 1\}$, that is why we consider that the best solution is to use Hamming distances between individuals [35]. Then, during the clustering process, the individuals have to determine the closest cluster. This is done by computing the Hamming distance between the individual and the centroids of the clusters. However, using Hamming distances raises drawbacks. In a *k-means* analysis the clusters are made of similar individuals, and it seems interesting to us to have the possibility of observing the small variations (i.e. the profiles' variations) within each cluster. In order to do that, we propose to use real values (instead of binary values) to represent the clusters' centroids during the

clustering process. Thus the centroid of a cluster is determined by computing the average of the vector-week of all individuals belonging to this cluster.

Then, for the computation of the distance between an individual and a cluster's centroid we slightly modify the vector-week of the centroid in order to obtain a corresponding Hamming vector. For that, for each vector-day of the cluster's vector-week we set the item with the highest value to 1, and the six others items to 0. Figure 9 shows the modification process of a vector-day of a cluster's centroid.

| | A | RWiZ | RWoZ | MST | OST | WE | C |
|---|---|---|---|---|---|---|---|
| Real vector-day (Monday) | 0.1 | 0.3 | 0.12 | 0.54 | 0.56 | 0.212 | 0.02 |
| Corresponding Hamming vector | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Fig. 9: Modification of the vector-day of a cluster centroid

To sum up, individuals are represented by a vector-week of 49 items with only one positive item every seven items, and a cluster is represented by a vector-week of 49 items with real values (i.e. within the interval [0, 1]), except during the distance computation process where a cluster is represented by the corresponding binary vector-week.

### C. Clusters extraction of individual profiles

In our methodology, we propose to do the *k-means* cluster analysis with three different values of $k$ {20, 50, 70}, and with three different weeks, which allows various clustering possibilities. This computation generates several groups from the subset of data, and we propose to observe if some particularities emerge from these groups. However, our main objective is to extract general trends and patterns of the relations between the individuals and their territory. To avoid too many combinations of clusters we set a process to detect the main ones that we explain below.

#### 1) Filtering of representative individuals

The first step is to limit the number of clusters in the output of the solution. We propose in our methodology to display only the clusters that hold more than 1% of the total number of individuals. Limiting the output possibilities means that the individuals that are not kept in the solution are not representative of a real pattern followed by many. A drawback here is that the output solution may contain a really reduced number of individuals. In theory, if $k-1$ clusters contain exactly 1% of the individuals, then the last cluster contains $N(1-((k-1)*1/100))$ individuals. This corresponds at worst to 81% of individuals for $k=20$, 51% for $k=50$ and 31% of all individuals for $k=70$. In order to study the potential effects of this drawback, we propose to count the number of individuals present in the output clusters, and to compare it to the real number of individuals in the subset (10,000). This allows us to guarantee that the clusters that hold more than 1% of the individuals can be considered as good indicators at the scale of the whole population. In Table I we present

TABLE I: Clustering results with several $k$

| $k$ | Global silhouette | Average number of kept clusters | Individuals kept (%) | Individuals lost (%) |
|---|---|---|---|---|
| 20 | weak | 13.7 | 96.4 | 3.6 |
| 50 | average | 16.7 | 86.44 | 13.56 |
| 70 | good | 16.7 | 82.34 | 17.66 |

different statistics about the three number $k$ chosen for the methodology. The number of kept clusters is the average number of clusters that hold more that 1% of the individuals for the three different weeks. We see that the rate of kept individuals changes according to the number $k$. This confirms our previous analysis on figure 8 where we acknowledged that increasing the number $k$ results in a decreasing quality of clustering. Moreover, the average correct clustering rate is almost 88%, which means that the individuals kept by the output clusters may be considered as good representatives of the subset. And we demonstrated that the subsets are themselves good representatives of the data of the three weeks.

#### 2) Detection of main-profile clusters

For this analysis, we run nine clustering sessions, one for each week {1, 2, 3} and for each $k$ {20, 50, 70} respectively. For each session we keep the best run according to the average silhouette, and from this run we keep the clusters with more than 1% of the data set. This gave an average of 15.7 clusters for each session (141 clusters in totality).

We remark that for each session the clusters present different patterns and behaviors, but the most remarkable particularity is that we encounter similar clusters in each of the nine sessions. Which is to say that even with a different week and a different $k$, many clusters are identical. We thus managed to group these clusters according to their similarities. For that, we compare the clusters two by two. We compute for the seven days the sum of the standard deviations between all corresponding profiles. If the average of these sums is lower than 0.05 (which we empirically set), then we link these two clusters with an imaginary edge. We can then map the clusters into a graph, and we finally compute for each "clique" of clusters an averaged cluster that we call a main-cluster. We then obtain 12 main-clusters from our analysis. On average, these 12 main-clusters hold 82.75% of the individuals of the subset. This means that in 8 over 10 cases an individual follows one of the 12 extracted patterns. Each diagram of figure 10 represents the different patterns displayed by the 12 main-clusters. On the abscissa are the different days of the week (from 1 to 7) and on the ordinates are the rates of the profiles (from 0 to 1). The values of the curves show the percentage of a specific profile within the main-cluster. It is thus totally possible to have multiple curves crossing each other's in a diagram. However the total sum of the curves for one day is always 1. The color scheme used is the same as in figure 4, *RWiZ* profile is in dark blue diamonds, *RWoZ* profile in red squares, Commuters profile in green triangles, *MST* profile in purple crosses, *OST* profile in light blue double crosses, *WE* profile in orange disks and *Absent* profile is in gray line.

Additionally, for each of these main-clusters we propose in Table II to count the number of occurrences within the nine

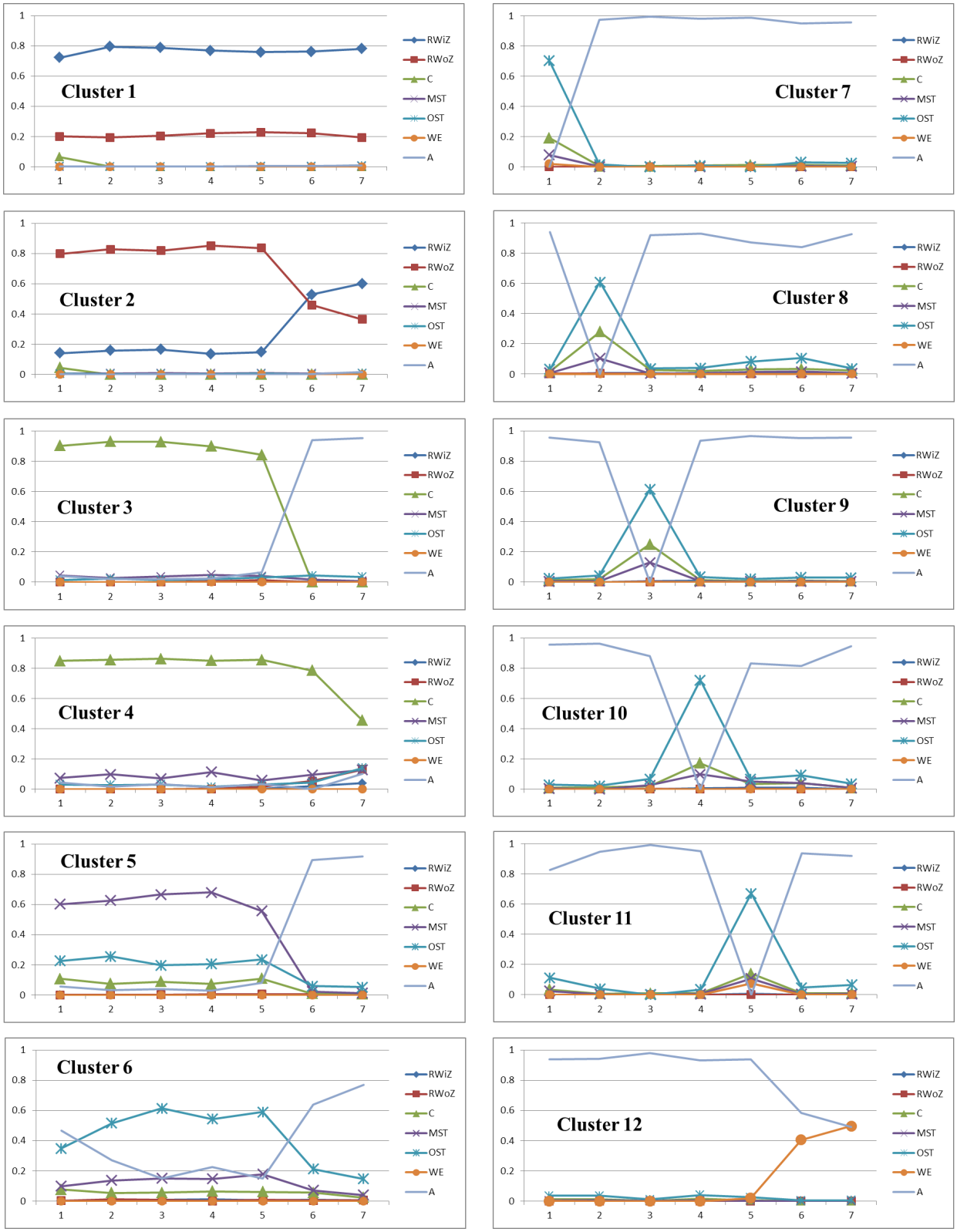


Fig. 10: The 12 main-clusters extracted from the nine clustering sessions

TABLE II: Number of individuals and occurences of the 12 main clusters

| Cluster | Number of individuals (over 10,000) | Occurences over the 9 sessions (%) | Profile type |
|---|---|---|---|
| 1 | 400.25 | 88.9 | Residents |
| 2 | 623.11 | 100 | Residents |
| 3 | 373.11 | 100 | Commuters |
| 4 | 164.57 | 77.8 | Everyday commuting |
| 5 | 225.88 | 100 | Everyday traveling |
| 6 | 160.66 | 100 | Everyday traveling |
| 7 | 744.2 | 55.6 | Punctual crossing |
| 8 | 665 | 66.7 | Punctual crossing |
| 9 | 889.62 | 88.9 | Punctual crossing |
| 10 | 794.66 | 77.8 | Punctual crossing |
| 11 | 852.83 | 77.8 | Punctual crossing |
| 12 | 2381.78 | 100 | WE visitors |

sessions and to compute the average number of individuals each main-cluster contains from the nine sessions. The occurrence column represents in percentage the number of clusters, from the nine sessions, used to create this main-cluster, i.e. it represents the size of the clique used to create this main-cluster. The occurrence shows the representativeness of the main-cluster inherent behavior given the initial conditions ($k$, week) and we see that the results are fair (from 55% to 100%). Moreover, as we present later, some of these main-clusters are related to the same profiles and we propose to classify these main-clusters into six different types which are easier to manage and to interpret. Eventually, by analyzing these main-clusters, we manage to extract the major trends of individuals with respect to their territory, which is the main objective of our approach to understand the characteristics of the territory.

### *D. Analysis of the 12 main clusters*

Our first observation is that 8,275 distinct IMEI (over 10,000) are represented by those 12 main-clusters. Each of these IMEI follows one of the 12 presented patterns; this indicates that there is a high tendency for individuals to reproduce identical patterns.

The different patterns within each cluster show that our previous statements about the filling and emptying of the territory are correct. For example, clusters 1 and 2 show that *RWiZ* and *RWoZ* are closely linked to each other, with a pattern on cluster 2 that tends to show that many individuals that were *Residents Working out of Zone* adopt an in zone profile during the weekend. This demonstrates that almost 60% of the residents that work outside during the week stay in their nearby environment during the weekend. Cluster 1 in the opposite is in majority composed of *Residents Working in Zone* that tend to stay in their nearby environment during the weekend.

Cluster 3 and 4 concern commuting patterns within the territory. We can observe in cluster 3 that individuals have a *Commuters* profile during the week, and are absent during the weekend. The individuals of cluster 4 are of type *Commuters* every day of the week, but this predominance tends to slightly fall during the weekend, allowing a larger variety of profiles on Sunday.

Clusters 5 and 6 show the dynamics of transit profiles, such as *OST* and *MST* individuals. Cluster 5 present individuals that are mostly of profile *MST* during the week and *Absent* the weekend, whereas in cluster 6 individuals with profile *OST* are also present during the weekend. These clusters inform us of the presence of many individuals that travel within the territory almost every day of the week, sometimes during less than one hour. It is important to remember here that the total time needed to cross the territory about one hour.

Clusters 7, 8, 9, 10 and 11 present the exact same dynamics, but each on a different day of the week. These clusters inform us of the presence of individuals that are absent of the territory during the week, except for a specific day where they cross the territory, and in majority for less than one hour. These clusters are interesting since they hold almost one third of the behaviors of the individuals detected within the territory. However, some of the individuals spend enough time in the territory to be classified as *Commuters* or *MST*. In these categories enter individuals like transporters or punctual industry or schools visitors. Finally, cluster 12 shows that almost all *WE* are present in the territory two consecutive days during the weekend, with a preference for Sunday.

Consequently, we propose a more advanced analysis for some of these clusters that we think are interesting for the characterization of the territory. We then focus on three of the different clusters above, cluster 2, 3 and 9. Cluster 2 shows a residential and working pattern with 623 detected individuals (over 10,000). During the week almost 83% of the cluster is considered as *RWoZ*, and 15% as *RWiZ*. During the weekend a convergence occurs, and the individuals tend to show an *in Zone* profile. This comforts us in the idea of a strong relation between these two profiles. Many of these individuals live in the territory, but work outside during the week (*out of Zone*), and when their economic activity is less important (i.e. during the weekend) they are detected in the territory as *RWiZ*. Note that in cluster 1, most of the time the individuals are *RWiZ* during the week and the weekend (76% of the cluster).

Cluster 3 shows a recurrent commuting pattern. This pattern totally corresponds to individuals that are employed in the territory but reside outside. The majority of these individuals have a commuting pattern every day of the week; they are mostly employees and students that live outside of the territory. Since they are not present during the weekend we can deduce that they do not work in touristic and commerce industry but in manufacturing, public services, schools... We can consider this cluster as the opposite of cluster 2 where its dynamic would be switched with the *RWoZ* flows. It is interesting to note that in cluster 4 the individuals are present during the weekend and so that they may work in touristic and commerce industry.

Cluster 9 groups 889 individuals (almost 9% of the sample) where *OST* profile is present on the territory on a specific day (Wednesday). This pattern corresponds to individuals that are only present on Wednesday, and if we look at the details of the cluster we can see that these individuals are at 61% classified in the *OST* profile, 25% in *Commuters* and 13% in *MST*. It means that 74% of individuals were present during periods of less than one hour, and 25% were present during all the day. We can interpret these dynamics as flows of individuals that wish to cross the territory without staying (transporters, Wednesday afternoon activities, etc.), which are at the border

of the zone, or which came inside the territory to visit industry sites or schools (the studied territory is a dynamic part of the South-West of Paris, with many companies and schools). We can also add to these individuals a large part of tourists since the territory is also known for its huge cultural heritage.

We finally propose a graphical visualization of these dynamics. Figure 12 represents the dynamics of the territory according to these three different patterns (clusters 2, 3 and 9). The territory is represented by the hexagon shape while each profile is represented by its letter within a color circle. The size of the circle represents the profiles' rates within the cluster for this particular day. The arrows represent the inherent dynamics of each profile, *OST* cross the territory, Commuters enter, and *RWoZ* go out of the territory. The length of the arrows is directly linked to the rate of the profile.

## V. Validation of the model

In this section we verify our conclusions about the dynamics of a territory with credible sources that already studied this territory. For that we use three distinct data sets, a National Census, a professional mobility survey and a scholar mobility survey. All these three surveys are coming from the INSEE, the French national statistic organism, and were done in a classical way, by interrogating people face to face. From the three surveys we extract several variables related to the demography and the mobility of individuals over the studied area.

### A. Data from national surveys

From the National Census we collect four demographic variables:

1) The total number of inhabitants (residents)
2) The number of residents within 15 and 64 years old
3) The working force (registered workers)
4) The non-working force (which includes students)

Additionally, we use the professional mobility and the scholar mobility databases from the INSEE to collect more information about workers and students within the zone. We call workers and students *active* individuals. From these mobility data sets we consider three more variables:

1) The number of individuals living and being active in the same city
2) The number of individuals living in the city but being active in another city
3) The number of active individuals coming from outside the zone

A summary of these seven variables is given in table 3. All the individuals of the studied territory are recorded by these variables and some of these variables are exclusive, which means they are used to categorize the individuals into distinct classes (working force and non-working force for example). Figure 12 illustrates the interlinking of these exclusive classes of individuals.

Moreover, we propose to study the similarities of these classes with some profiles from our methodology. For that we focus on the three main profiles that can be easily extracted from the seven variables of the survey: *RWiZ*, *RWoZ* and *Commuters*. The *RWoZ* is the sum of the active individuals going out of zone for their activity (work or school), and the *RWiZ* is the sum of the active individuals staying in zone for their activity, plus the inactive individuals. Finally the *Commuters* are individuals which live outside the zone, but who have their activity (work or school) within the zone. Figure 12 links our three profiles to the different classes of individuals and a summary is given in table IV.

TABLE III: National data about the zone

| Source | Inhabitants | Inhabitants between 15-64 | Working force | Non-working force and non-students |
|---|---|---|---|---|
| National census | 412,500 | 272,900 | 207,650 | 46,385 |
| **Source** | **Active from inside, staying in zone** | **Active from inside, going out of zone** | | **Active in zone, from outside** |
| Professional mobility survey | 98,550 | 85,950 | | 101,915 |
| Scholar mobility survey | 27,115 | 14,900 | | 18,315 |

TABLE IV: Summary of the different profiles present on the zone

| Category from national survey | RWiZ | RWoZ | C |
|---|---|---|---|
| Workers | 98,550 | 85,950 | 101,915 |
| Students | 27,115 | 14,900 | 18,315 |
| Non-working force and non-students staying in zone | 46,385 | / | / |
| Total | 172,050 | 100,850 | 120,230 |

### B. Adjustments on mobile phone data

#### 1) Size estimation of the profiles

In part 4 we clustered a subset of 10,000 individuals and then we obtained 12 main clusters (figure 10). Now we classify every individual of the full data set (527,622 individuals from week 2) into one of these 12 main clusters. This gives us the actual individuals distribution which will be used for comparative analysis with national survey data. Note that the upcoming results are similar for week 1 and week 3.

Firstly, we estimate the size of each profile. For that we multiply the average rate of each profile within a cluster (value between [0, 1]) by the actual size of this cluster, and we only use clusters where a profile is distinctly present. For example, the profile *RWiZ* corresponds to the number of *RWiZ* from cluster 1 and from cluster 2. A rate of 0.8 from cluster 1 gives 19,993 individuals and a rate of 0.2 from cluster 2 gives 12,292 individuals. Then the total and actual number of *RWiZ* individuals is 32,285. Table V presents the results of this classification for the 527,622 individuals of week 2. In addition the same calculation is done for other profiles *RWoZ*, *C*, *OST* and *MST*. Note that for the profiles *Commuters*, *OST* and *MST*, we also propose an estimate based on the results from *punctual crossing* clusters 7, 8, 9, 10 and 11, to track people coming for one day per week and to look at their intensity.

#### 2) Size adjustments from mobile network market

Our clusters are only representative of the individuals tracked by the network operator which provides the data. This

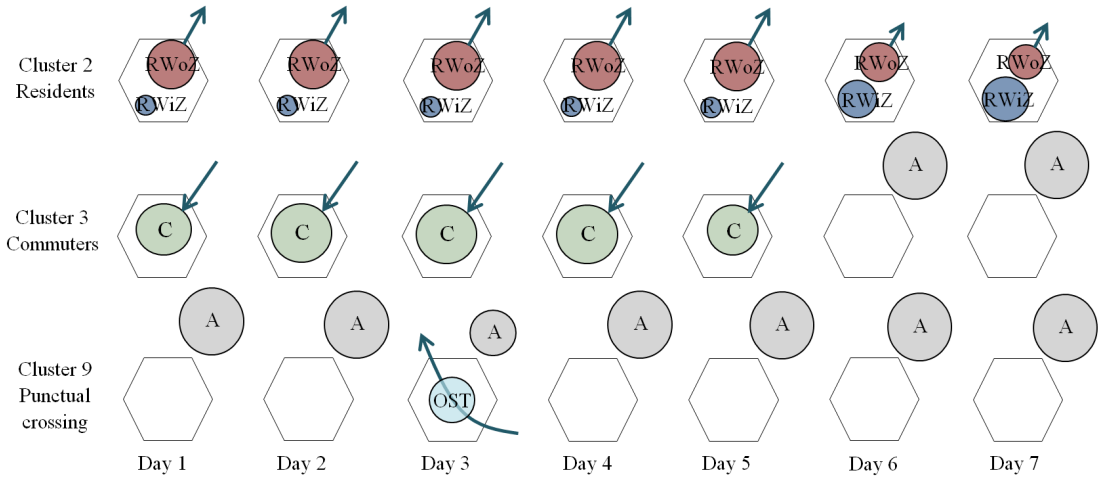


Fig. 11: The territory dynamics through three different patterns

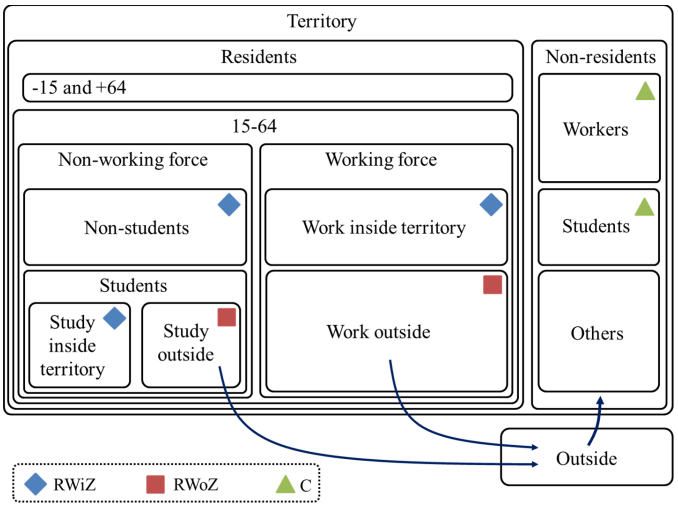


Fig. 12: Inhabitants class distribution within the territory

TABLE V: Results of the classification for 527,622 individuals

| Profile | From cluster 1 | From cluster 2 | Total | |
|---|---|---|---|---|
| RWiZ | 19,993 | 12,292 | 32,285 | |
| RWoZ | 5,898 | 37,164 | 43,153 | |
| Profile | From cluster 3 | From cluster 4 | Total | Average from clusters 7,8,9,10,11 |
| C | 23,897 | 11,066 | 34,963 | 13,148 |
| Profile | From cluster 5 | From cluster 6 | Total | Average from clusters 7,8,9,10,11 |
| OST | 11,400 | 2,669 | 14,069 | 31,618 |
| MST | 2,043 | 9,748 | 11,791 | 4,330 |
| Profile | From cluster 12 | | Total | |
| WE | 62,882 | | 62,882 | |

means that these numbers need to be adjusted to the total population of the studied area. At the time we obtained the data there were three main network operators in France. A commonly used rate to adjust cell phone data is about one third. We show that this rate is correct for our data set. From figure 3 we estimate the total number of residents to be 81,912. A simple division shows that 81,912 divided by 272,900 (inhabitants between 15 and 64 years old given by the National Census) gives 0.3 which is close to one third. In [21] the authors note that users without cellphones are not modeled in cellphone data mobility models, that is why we propose to use the number of inhabitants between 15 and 64 years old to estimate this rate because they are the most susceptible of owning a cell phone, and furthermore they are the individuals taken into account for the *active* part in national surveys. Table VI presents the adjusted estimations for each of the three profiles *RWoZ*, *RWiZ* and *C* compared with national survey data. Note that *Residents* adds *RWiZ* and *RWoZ*.

TABLE VI: Adjusted estimations for the 527,622 individuals

| Profile | Estimations (recurrent clusters) | Adjusted estimations |
|---|---|---|
| Residents | 75,438 | 251,460 |
| RWiZ | 32,285 | 107,617 |
| RWoZ | 43,153 | 142,843 |
| C | 34,963 | 116,543 |

### C. Comparative analysis

We propose a comparative analysis between our estimations and the data collected from the three national surveys over the profiles *Residents*, *RWiZ*, *RWoZ* and *C*. A short summary of the estimated values is given in table VII.

We see that the correlation rates vary from 58% to 97% according to the variables. The estimations on the total number of *Residents* and *Commuters* are close, which shows the relevance of the adjustment factor. However, on the profiles that involve a dynamic behavior (*staying in the zone* or going *out of the zone*) the correlation rates drop to 58 and 62%. The interpretation of such a difference is that for INSEE survey, many resident people declare to work in the zone attached to their company, but are actually often outside (taxi drivers, transporters, business travelers, managers, etc.). So when they are tracked by their mobile phone, they are tracked outside their residence and working area [36]. This is a strong added-value of the mobile phone tracking data. Furthermore the national surveys deliver estimations based on only one-day

TABLE VII: Comparison between mobile phone model and surveys data

| Profile | INSEE | Adjusted mobile phone data | Correlation rate (%) |
|---|---|---|---|
| Residents | 272,900 | 251,460 | 92.14 |
| RWiZ | 172,050 | 107,617 | 62.55 |
| RWoZ | 100,850 | 142,843 | 58.36 |
| C | 120,230 | 116,543 | 96.93 |

TABLE VIII: Adjusted estimations for the 527,622 individuals

| Profile | Estimations (recurrent clusters 3,4,5,6,12) | Adjusted estimations | Estimations (punctual clusters 7,8,9,10,11) | Adjusted estimations |
|---|---|---|---|---|
| C | 34,963 | 116,542 | 13,148 | 43,827 |
| OST | 14,069 | 46,897 | 31,618 | 105,393 |
| MST | 11,791 | 39,303 | 4,330 | 14,433 |
| WE | 62,882 | 209,607 | / | / |

survey and are limited by a standard deviation coefficient [37].

If we look at the *RWoZ* collected, we see that from the mobile phone data there are almost 42,000 more than from the INSEE survey, and inversely, 64,000 less in *RWiZ* category. Concerning the collection of static information about people (age, work, revenue, etc.), the face-to-face surveys are very confident, but concerning the collection of dynamic information about people (when they move, how long, etc.), the information are not so truthful and the error rate is up to 30% between what people say and what they really do. Here we see that deviation.

### D. Conception of new indicators

We show that our results are consistent with national data in terms of static behavior (residents and commuters). Now, we propose new indicators to qualify the dynamics of this territory. In addition to the *RWiZ* and *RWoZ* analysis, table VIII presents the adjusted estimations for the other profiles we detected with our methodology based on the analysis of mobile phone data.

Almost 47,000 *OST* and 39,000 *MST* are present daily on the territory. They are individuals that cross the zone, or that punctually come, like taxis or transporters. The number of estimated *WE* is also interesting; almost 210,000 individuals are present in the territory only during the weekend, which represents 33% of the population present at this time. This is notably due to the presence of many cultural and tourist attraction in the territory, but we can also put in this category many individuals that work or study outside during the week and come back home only for the weekend. Finally, we detect many individuals that are present only one day of the week. They are individuals that cross the territory, visit touristic or industrial venues, taxis or transporters. Thus, almost 164,000 individuals are present on the territory for less than one day, which represents almost 20% of the individuals present on the territory during the day. From those individuals, almost 105,000 (12%) stay in the territory for less than one hour.

## VI. Conclusions

Cellular networks are definitely omnipresent in our daily lives, and many applications for the understanding of human and urban mobility use these systems. Unfortunately, cellular data suffer from many drawbacks, coarse location estimates, user dependent data frequency, etc. as seen in [22] and [19]. Then using these data as a continuous presence indicator can lead to erroneous conclusions. Moreover, for privacy issues, all collected data is anonymized, and it is impossible to follow continuously the movements of individuals, and to link the individuals with a specific social category.

We proposed to leverage the human behavior and the predictive human patterns by analyzing cellular data from an aggregated and clustered point of view. The activity of the cellular network's users allowed us to understand the relation that individuals have with their territory.

We presented a new profiling algorithm that allows the classification into six different profiles of individuals, according to their usage of the cellular network. We analyzed the distribution of these profiles over a three week period of data and we presented a novel approach to extract the individuals' weekly mobility patterns.

We presented an innovative methodology that groups the individuals according to their profiles, and we extracted 12 patterns that hold the main mobility behaviors of individuals over a territory. Our results can be used by mobility modeling algorithms that require human mobility patterns like [15] or [16]. We compared our results with data from three National surveys and we proved that our methodology based on the analysis of mobile phone data was consistent and relevant. We then proposed several dynamic indicators inferred from this approach and we used them to dress a map of the dynamics of a territory. In particular, we saw that a lot of new and unpredictable information about mobility may be added to the face to face survey from mobile network data.

Our future works concern the evolution of the mapping algorithm to propose a fuzzy version and to compare the fuzzy distribution to the hard one. We also intend to work on the influence of the sample size to build the main-clusters and on the thresholds of similarities between the clusters. Furthermore, we intend to apply the method on a second region to improve and to generalize the process.

## References

[1] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting Places from Traces of Locations," in *Proceedings of the 2Nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, ser. WMASH 04. Philadelphia, PA, USA: ACM, 2004, pp. 110–118.

[2] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular Census: Explorations in Urban Data Collection," *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, Jul. 2007.

[3] "World Telecommunication Development Conference (WTDC-14): Final Report."

[4] G. Rose, "Mobile Phones as Traffic Probes: Practices, Prospects and Issues," *Transport Reviews*, vol. 26, no. 3, May 2006.

[5] C. Iovan, A.-M. Olteanu, T. Couronné, and Z. Smoreda, "Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies," in *Geographic Information Science at the Heart of Europe*, ser. Lecture Notes in Geoinformation and Cartography, D. Vandenbroucke, B. Bucher, and J. Crompvoets, Eds. Springer International Publishing, May 2013, pp. 247–265.

[6] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban Sensing Using Mobile Phone Network Data: A Survey of Research," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 25:1–25:20, Nov. 2014.

[7] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.
[9] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Roma," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, Mar. 2011.
[10] F. Alhasoun, A. Almaatouq, K. Greco, R. Campari, A. Alfaris, and C. Ratti, "The City Browser: Utilizing Massive Call Data to Infer City Mobility Dynamics," in *The 3rd International Workshop on Urban Computing (UrbComp 2014)*, New York, NY, Aug. 2014.
[11] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human Mobility Characterization from Cellular Network Data," *Commun. ACM*, vol. 56, no. 1, pp. 74–82, Jan. 2013.
[12] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr., and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, Jan. 2013.
[13] C. Joumaa, A. Caminada, and S. Lamrous, "Mask Based Mobility Model A new mobility model with smooth trajectories," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks and Workshops, 2007. WiOpt 2007. 5th International Symposium on*, 2007, pp. 1–6.
[14] K. A. Ali, M. Lalam, L. Moalic, and O. Baala, "V-MBMM: Vehicular Mask-Based Mobility Model," in *Networks (ICN), 2010 Ninth International Conference on*, 2010, pp. 243–248.
[15] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human Mobility Modeling at Metropolitan Scales," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. Low Wood Bay, Lake District, UK: ACM, 2012, pp. 239–252.
[16] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, and R. N. Wright, "DP-WHERE: Differentially private modeling of human mobility," in *2013 IEEE International Conference on Big Data*, Oct. 2013, pp. 580–588.
[17] J. Yuan, Y. Zheng, and X. Xie, "Discovering Regions of Different Functions in a City Using Human Mobility and POIs," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 186–194.
[18] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring Land Use from Mobile Phone Activity," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '12. New York, NY, USA: ACM, 2012, pp. 1–8.
[19] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility," *Journal of The Royal Society Interface*, vol. 10, no. 81, p. 20120986, Apr. 2013.
[20] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring Human Mobility with Multi-source Data at Extremely Large Metropolitan Scales," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '14. New York, NY, USA: ACM, 2014, pp. 201–212.
[21] D. Zhang, J. Zhao, F. Zhang, and T. He, "coMobile: Real-time Human Mobility Modeling at Urban Scale Using Multi-view Learning," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '15. New York, NY, USA: ACM, 2015, pp. 40:1–40:10.
[22] M. A. Bayir, M. Demirbas, and N. Eagle, "Discovering spatiotemporal mobility profiles of cellphone users," in *2009 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks Workshops*, Jun. 2009, pp. 1–9.
[23] Z. Smoreda, A.-M. Olteanu, and T. Couronné, "Spatiotemporal data from mobile phones for personal mobility assessment," *Transport Survey Methods: Best Practice for Decision Making*, 2013.
[24] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo, "Identifying users profiles from mobile calls habits," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '12. Beijing, China: ACM, 2012, pp. 17–24.
[25] M. Nanni, R. Trasarti, B. Furletti, L. Gabrielli, P. V. D. Mede, J. D. Bruijn, E. D. Romph, and G. Bruil, "MP4-A Project:Mobility Planning For Africa," in *Analysis of mobile phone datasets for the development of Ivory Coast*, ser. Mobility/Transport, V. Blondel, N. d. Cordes, A. Decuyper, P. Deville, J. Raguenez, and Z. Smoreda, Eds. netmob, May 2013, pp. 423–446.
[26] R. Ahas, A. Aasa, S. Silm, and M. Tiru, "Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 1, pp. 45–54, 2010.
[27] R. Ahas, A. Aasa, A. Roose, l. Mark, and S. Silm, "Evaluating passive mobile positioning data for tourism surveys: An Estonian case study," *Tourism Management*, vol. 29, no. 3, pp. 469–486, 2008.
[28] F. Girardin, F. Calabrese, F. Fiore, C. Ratti, and J. Blat, "Digital Footprinting: Uncovering Tourists with User-Generated Content," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 36–43, Oct. 2008.
[29] S. Jiang, J. Ferreira, and M. C. González, "Clustering daily patterns of human activities in the city," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 478–510, Nov. 2012.
[30] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, Jul. 2013.
[31] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore," *IEEE Transactions on Big Data*, vol. PP, no. 99, pp. 1–1, 2017.
[32] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting Space through Digital Signatures," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 78–84, 2010.
[33] R. Fergus, A. Zisserman, and P. Perona, "Sampling methods for unsupervised learning," in *Advances in Neural Information Processing Systems 17*. Neural information processing systems foundation, 2005.
[34] P. J. Rousseeuw, "Silhouettes - A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53–65, 1987.
[35] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, Apr. 1950.
[36] INSEE, "Definition of working place." [Online]. Available: http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/doc/definitions-rp.htm#def_lieu_travail
[37] ——, "Recensement de la population - La précision des résultats du recensement," INSEE, Tech. Rep., Aug. 2009.

**Etienne Thuillier** received the title of engineer in computer sciences (M.S. Degree) at Belfort-Montbéliard University of Technology (UTBM), France, in 2013. He is currently pursuing a Ph.D. degree in computer sciences at the OPERA team, UTBM. His research interest includes Intelligent Transportation Services technologies, and more generally human and urban mobility.

**Laurent Moalic** received the title of engineer in computer sciences (M.S. Degree) at UTBM, France, in 2003. He later received the Ph.D. degree in computer science from the University of Franche Comté in 2013. In 2004, he joined the Transportation Laboratory of UTBM, as a research engineer. His current research interests include operational research, combinatorial optimization, and human mobility modeling. Currently, he is the team leader for the French National Research Agency project, Norm-Atis, about new standards to develop advanced transportation information services.

**Sid Lamrous** is currently associate professor of computer science at UTBM. From 1996 to 1999, he obtained the postgraduate diploma in systems control (M.S. Degree) and the Ph.D in computer science at the University of Technology of Compiegne. From 2011 to 2013 he was head of engineering education by learning for 3 years. For 10 years he has been working on models and algorithms for combinatorial optimization, and stochastic modeling of mobility.

**Alexandre Caminada** is currently full professor of computer science at UTBM. He received the MSc research degree in Artificial Intelligence from the University of Paris XII/Paris VIII and the diploma in Computer Science from ESIEA Paris (M.S. Degree), then the Ph.D from the University of Montpellier II in telecommunication software engineering. From 1993 to 2004, he worked at the National Telecommunications Research Centre (CNET France) as head of a Unit Research on Wireless Optimisation and in 2004 he has been nominated as full Professor at UTBM. His personal research is about resources optimization and network performance modelling for mobile and wireless systems.