On the Properties of Human Mobility

Michela Papandrea[a,*], Karim Keramat Jahromi[b], Matteo Zignani[b], Sabrina Gaito[b], Silvia Giordano[a], Gian Paolo Rossi[b]

[a] *NetLab, ISIN-DTI, SUPSI, Manno, Switzerland*
[b] *Computer Science Department, Universitá degli Studi di Milano, Italy*

Abstract

The current age of increased people mobility calls for a better understanding of how people move: how many places does an individual commonly visit, what are the semantics of these places, and how do people get from one place to another. We show that the number of places visited by each person (Points of Interest - PoIs) is regulated by some properties that are statistically similar among individuals. Subsequently, we present a PoIs classification in terms of their relevance on a per-user basis. In addition to the PoIs relevance, we also investigate the variables that describe the travel rules among PoIs in particular, the spatial and temporal distance. As regards the latter, existing works on mobility are mainly based on spatial distance. Here we argue, rather, that for human mobility the temporal distance and the PoIs relevance are the major driving factors. Moreover, we study the semantic of PoIs. This is useful for deriving statistics on people's habits without breaking their privacy. With the support of different datasets, our paper provides an in-depth analysis of PoIs distribution and semantics; it also shows that our results hold independently of the nature of the dataset in use. We illustrate that our approach is able to effectively extract a rich set of features describing human mobility and we argue that this can be seminal to novel mobility research.

Keywords: Human mobility, visited points, pervasive computing

1. Introduction

In recent years we have witnessed a rapid increase of people mobility as the world population has become more interconnected and has begun relying on faster transportation methods, simplified connections and shorter commuting times. Unveiling and understanding human mobility patterns has become a

[*] Corresponding author
Email addresses: michela.papandrea@supsi.ch (Michela Papandrea), karim.keramat@unimi.it (Karim Keramat Jahromi), matteo.zignani@unimi.it (Matteo Zignani), sabrina.gaito@unimi.it (Sabrina Gaito), silvia.giordano@supsi.ch (Silvia Giordano), gianpaolo.rossi@unimi.it (Gian Paolo Rossi)

crucial issue in supporting decisions and prediction activities when managing the complexity of today's social organization. In this, novel mobile communications technologies play a fundamental role. With such mobile technologies it is now possible to collect data about human habits and behavior all day long. Nowadays, people always carry their mobile phone with them. So, either in the form of Call Detail Records (CDRs) or with specialized apps [22], [25], people's mobility data can be collected from mobile phones. Therefore, in the recent years, researchers have devoted considerable effort to collecting and studying human mobility patterns [7] and have applied their understanding to a variety of critical problems ranging from disease spreading [2], urban planning, smart and green transportation to network infrastructure [37, 14], economy and marketing [30], and mobile network services [13]. Nonetheless, despite the advances in communications technologies and other important achievements, human mobility still represents an open and challenging research issue. In practice, the mobility pattern of each individual consists of the sequence of locations s/he visited. These locations and their correlations represent the core block of any modeling research and any activity aimed at understanding human mobility. Even though visited locations underlie all works in this field, their features remain largely unknown. This is due mainly to the fact that they have been considered as points in an area and social aggregation places, without anchoring spatial features to the behavior of each single user.

This paper, which represents an extension of our previous works [31, 44], aims to fill the gap by providing a general framework for dealing with modeling locations from a per-user perspective. Also, it paves the way towards enabling the semantic interpretation of locations to be overlaid on their spatial distribution.

First, we introduce the notion of user's Points of Interest (PoIs) along with the methodology to extract them from different types of data. Then we provide both a metric to measure the importance of PoIs for a person and a methodology to classify them in terms of: *(i)* Most Visited Points (MVPs), the places that a person visits most regularly, e.g. home and work locations; *(ii)* Occasionally Visited Points (OVPs), locations of interest for the user but visited just occasionally; and *(iii)* Exceptionally [1] Visited Points (EVPs), which correspond to seldom visited locations. This classification allows us to define a human mobility profile where the number of locations per each class and the time spent there are the characterizing attributes. We further study how people move across PoIs and PoI classes, enriching the knowledge derived from classification with the spatial as well as the temporal dimensions of mobility. The proposed classification and the PoIs and user features provide the basis for understanding human behavior by extracting the semantics of visited places. In line with similar works [10, 23, 15, 33], we used a heuristic approach for the semantic analysis and experimented it on a large dataset containing mobility patterns of hundred thousands of people in a metropolitan area.

[1] We use the adverb 'exceptionally' as a synonym for rarely, seldom.

The paper supports its findings by extensively validating results on four different datasets. The first two datasets contain Call Detail Records of phone activities of a large mobile operator. The third dataset is mainly composed of trajectories (parts of a continuous mobility trace), while the last one consists of continuously sampled location data. The first two datasets have different characteristics in terms of spatial and temporal distribution of the visited places w.r.t the other two databases. By showing the validity of our approach throughout datasets with sometimes antithetical properties, we demonstrate the independence of our results w.r.t. a specific setting, and we are able to extract a deeper understanding of human mobility.

As a result of this work, some interesting properties about human mobility emerge. In fact, it turns out that people visit many locations in their life, but they have a *very small number of preferred locations (MVPs)* which are visited daily (e.g., home, work place), and *a higher, but still limited, number of locations of interest (OVPs)* which are visited with a lower frequency (e.g., gym, favorite restaurant, parent's house). We spend more than 50% of our time in MVPs. This indicates that those points are *the ones that best represent and characterize our lives.* On this basis, we propose an algorithm to identify home and work places which leverages the relevance of a place for a specific person and outperforms other algorithms in terms of semantic accuracy.

By analyzing the transition rules between PoIs, we find that, in *contrast with commonly accepted assumptions, the decision to move between two places is not taken on the basis of the geographical distance, but according to the relevance individuals ascribe to them and to the travel time between places.* Also, we show that the transition rule based on relevance follows the same distribution law independently of the mobility scenario.

The key contributions of our mobility framework can be summarized as follows:

- a novel per-user mobility analysis that highlights the following key properties:
  - people visit regularly just few places where they spend most of their time;
  - people also spend a significant amount of time in places they only visit once;
  - people commute between places based on their temporal distance and not the spatial distance;
  - HOME and WORK places are in the set of few places mostly visited, and, as such, the relevance R is a fundamental feature for their semantic identification;
- a classification of visited locations (PoIs) that enables the above mentioned analysis;
- a classification of users, based on how people move across PoIs and PoIs classes, derived from our mobility analysis;

- a semantic understanding of human behavior based on our mobility analysis;
- a thorough experimental validation on datasets with different properties.

The comprehension and the modeling of human mobility patterns play a key role in the design of protocols and forwarding strategies in contact-centric network infrastructure. These novel results can change how mobility is analyzed and modeled. Indeed, we argue that, to produce more realistic mobility traces, a mobility model needs to consider *i)* the new classifications introduced herein, and *ii)* the new features, their relationships and their different laws. This work could impact several computer and communications areas such as: localization [28, 29], where our results indicate that a person's location can be predicted in the set of MVPs with a probability higher than 0.7; social interaction studies and data offloading [32] [16], as people tend to meet more frequently people with some MVPs in common and the latter characterize the single individual's mobility; human mobility modeling [41], as mobility can be described in terms of regular movement among MVPs and OVPs and extemporarily EVPs; recommendations [26] as people can get recommended places close to their MVP and not far in time from their current location.

2. Related Work

Nowadays smartphones have an important role in capturing various behavioral aspects of users, ranging from how the device is used across different contexts to analyzing the spatial, temporal and social dimensions of everyday life through sources such as GPS, call and text logs, Internet access and Bluetooth logs. These data can be used in many areas, from urban planning, predicting and controlling epidemic infection diseases to planning and optimization of wireless and infrastructure-less communication systems. Fundamentally, these applications require the comprehension and recognition of predictable mobility patterns. To gain a better understanding of the dynamics involved in mobility, many experiments, based on different detecting technologies and performed in various locations, have been conducted. Most of them have been made available in the public repository CRAWDAD [1]. Among these datasets we focus on GPS-based traces as they allow us to precisely determine the geographical positions of users. In this study we also compare mobility data from cellular network towers with the GPS positioning. We made this choice to highlight similarities and differences among the mobility habits, due to the different detecting technologies usually adopted to study them. That results in a heterogeneous set of data which require different pre-processing techniques to get a uniform representation through which we deal with the analysis. For the above reasons this work relates to different research topics.

Significant location extraction. Part of our work, which involves GPS data, has been devoted to detecting the significant locations of a user. Many authors have suggested different extraction methods [8, 18, 20, 39, 38, 40] based on clustering algorithms. Ashbrook *et al.* [8] have proposed a two-step method to infer the significant locations. In the first step, the loss of the GPS signal is used as an indicator of interesting locations because it likely corresponds to buildings or indoor points. In the second step these points are clustered into locations using a variant of the k-means algorithm. In the clustering procedure, round clusters with a given radius are initially placed at k chosen points, and iteratively they move to a denser area, until no further increases in the point density is observed. Since the loss of the GPS signal serves as the main clue to identify significant locations, main buildings are found; however, other types of interesting locations where the signal is available, such as outdoor places, may be lost. Furthermore, rather than detecting locations with an arbitrary shape, they retrieve only circular locations. On the contrary, we apply a clustering method able to find arbitrary shape clusters, independently of an a-priori number of places.

Hariharan and Toyama [18] proposed an approach that uses time information to distinguish significant places. From the raw traces they identify a contiguous sequence of GPS points within a distance d and for a period t adopting a variation of an agglomerative clustering algorithm. They called these areas 'stays'. Since their algorithm is computationally expensive (the identification of a stay requires the distance between all pairs of coordinates within a specified time window to be computed after every new location measurement) we choose a more computationally efficient algorithm that neglects the temporal information since the GPS traces have been recorded with a fixed sample rate.

Kang *et al.*[20] proposed a method, suitable for resource-limited mobile devices, that computes incrementally significant locations. Their time-based approach clusters the stream of incoming location coordinates along the time axis and drops those clusters where little time is spent. In particular, the algorithm compares each new GPS point with the previous coordinates in the current cluster; if the stream of coordinates is far from the current cluster a new location is detected. The authors validate their algorithm with localization data inferred from RF(radio frequency)-emissions of known base stations. Since the main goal of the method is portability on mobile devices, authors did not investigate the trajectories of multiple users.

Finally, to overcome the k-means limitations, a series of density-based approaches have been proposed. Zhou et *et al.*[40] proposed a density- and join-based clustering algorithm called DJ-Cluster to infer significant locations. The dense points are those with at least a certain number of other points lying within a distance of their neighborhood. Relaxing the DBSCAN conditions on reachability, the clusters are formed from a set of dense points, which are density-joinable: *i.e.* the neighborhood of the dense points shares a common point. A further preprocessing procedure, which removes GPS points corresponding to limited movements, is introduced to improve the performance of the algorithm. The experimental results indicate great improvements in terms of both recall and precision w.r.t. those obtained from the k-means algorithm. A similar approach has been adopted by Zheng *et al.*[39], [38]. They applied a density based clustering algorithm (OPTICS [5]) to extract significant locations in order to infer transportation modes and to predict users' preferred locations. Our definitions, which inherit preferred locations and the extraction algorithm, are inspired by the above methods.Nevertheless, in comparison with these works, we propose a more general definition of stay-location that enables us to consider temporal reappearances at the same place.

Statistical analysis of mobility. Spatial mobility patterns have been analyzed in different disciplines, from physics to pervasive computing. Works from the physicists' community focus on concepts from statistical mechanics and thermodynamics. Their main goal is to identify what kind of diffusion process is able to best reproduce human mobility. For these reasons they analyze the displacement and the length of movements, searching for evidence of sub- or super-diffusive processes. On the contrary, works from computer science focus more on human mobility properties, which can be exploited in the deployment of different services (from opportunistic networks to link prediction in location-based social networks).

In their seminal work Brockmann *et al.*[9] investigated human traveling statistics by analyzing the circulation of banknotes in the United States. Based on a huge dataset of over a million individual displacements, they found that the distribution of the traveling distances decays as a power law, indicating that trajectories of bank notes are similar to Lévy flights. Secondly, they showed that the probability of staying in a confined region (pause time distribution) is characterized by a long tail leading to a sub-diffusive process.

Gonzalez *et al.*[17] also focused on distances covered by people. In particular they analyzed mobile phone users for a six-month period in a large area. They found that the distribution of the distance between two consecutive calls is well approximated by a truncated power-law. Moreover, each individual tends to return to a few frequented locations with high probability.

Rhee *et al.*[33] were the first to deal with the statistical properties of human mobility using GPS traces. By analyzing GPS traces collected on a campus they reported that bursty hot spot sizes play an important role in causing the heavy-tail distribution of distances in human walk. They show that visit points are clustered and that pause time distribution in hot spots follows a truncated Pareto.

A recent study cast some doubts on the power law distribution of the distance as a universal feature of human mobility. In fact Noulas *et al.*[27] focused on human mobility patterns in a large number of cities. Mobility data have been retrieved from mobile location-based social services. They first observed that mobility, when measured as a function of distance, does not exhibit universal patterns. By contrast, considering another variable, they obtained more general results for all cities. Precisely, they discovered that the probability of transiting from one location to another is inversely proportional to a power of their rank, *i.e.* the number of intervening opportunities between them.

Other works investigate characteristics other than distance. For instance, Song *et al.*[35] studied the predictability of human trajectories derived from the estimated entropy of the mobile phone data. The predictability is centered around 93% over a large population, independently of the size of the area covered by individuals' mobility or other demographic factors. Probably, the high predictability is obtained based on low resolution positioning data since the average size of a 'location' is roughly 3 km^2. For higher resolution positioning data such as the GeoLife dataset, Lin and Hsu [23] showed that a high predictability is still present at fine spatial/temporal resolutions. However, they observed an invariance between the predictability and spatial resolution. In other words, we cannot obtain a high prediction accuracy and spatial precision simultaneously.

Kim *et al.*[21] used Access Point (AP) log data to extract information about users' movements and pause times but they did not care about location distances in computing users' transition probabilities. They found that pause time and speed distributions follow a log-normal distribution and that the directions of movement follow the direction of popular roads and walkways on the campus showing a symmetry across 180 degrees.

Home/workplace recognition from cellular network data. A great effort has been devoted to the assessment of the visited locations, trying to assign a particular meaning to each of them. Among the different problems in the evaluation of the location semantic, we focus on the detection of home and work places from cellular network data, based on the frequency of daily visits, a.k.a. relevance. To solve the aforementioned issue, Isaacman *et al.*[19] have proposed a technique based on clustering and regression to identify important places then assign them a semantic such as home and work. By contrast, Csaji *et al.*[12] have combined principal component analysis with clustering to robustly identify home and work places. Finally, Arai and Shibasaki [6] have proposed a methodology for the estimation of home and work locations based on time windows. After recognizing important places according to the length of stay and frequency of visits, they base the home/work identification on core hours at home/work. Most of the approaches require knowledge of the tower position (GPS or place names), but this information is not always available. So the strategies and methodologies proposed in above literatures are not applicable in our case.

An identification method not founded on knowledge of the tower positions has been presented by Alhasoun *et al.*[4]. In their work they identify the places where each user is more active (call) by dividing a day into daytime and night. Home is the most active place during the night window, while work is the most active location during the day. Apart from being time window dependent, the method does not consider regularity in visiting places as the main feature defining home and work. However, it is commonly accepted that most users regularly visit and commute between home and place of work on workdays. Thus, solely the number of activities is not a good indicator for home and work, since users may make a burst of on-phone activities in places which are not frequently and regularly visited.

In [15] the authors analyzed call and Bluetooth logs of approximately a hundred users for a duration of nine months in order to identify a structure in the daily life routine of mobile users. They attempted to quantify the amount of predictable structure in an individual's life using an information entropy metric. They expected people with low-entropy lives to be more predictable across all time scales. By using the discovered patterns and contextualized proximity information extracted from Bluetooth logs, they proposed a model for identifying location and activities.

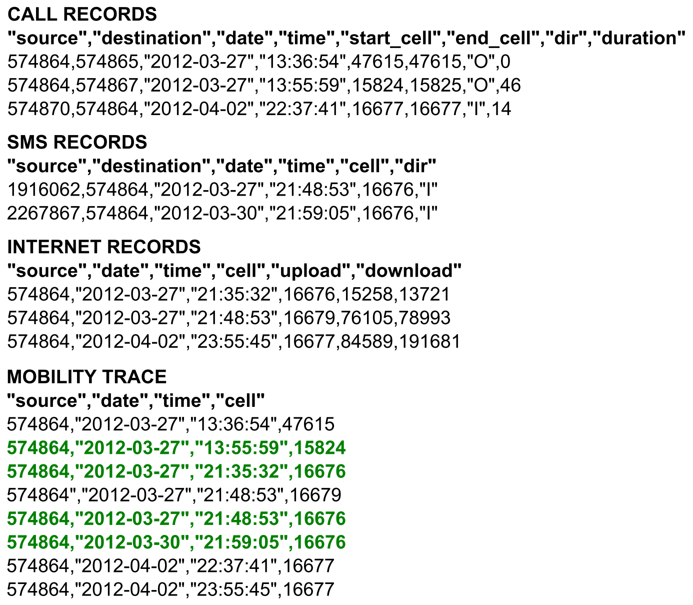
3. Datasets

Since smartphones are carried by people, they can capture movement patterns and behavioral aspects of their human carriers [22]. These mobile devices enable the development of data collection tools to record various behavioral aspects of users, ranging from how the device is used across different contexts to the analysis of spatial, temporal and social dimensions of users' everyday lives, through sources such as GPS, call and SMS logs and Internet accesses.

In our paper we exploit all these data in order to highlight mobility features common to different scenarios and geographical areas. Specifically, we performed our studies over four different datasets. The first two datasets are Call Detail Records of smartphones collected by a mobile operator. The third dataset is mainly composed of trajectories, while the fourth consists of continuously sampled location data - with both sets collected by means of GPS technology. The first two datasets have different characteristics in terms of spatial and temporal distribution of the visited places w.r.t the other two databases. We will discuss each dataset in greater detail in the next sections. By showing the validity of our approach in different types of datasets, we demonstrate the independence of our results from the dataset characteristics. So, the novel features and properties we are able to derive in this work are independent of the analyzed scenario.

3.1. Call Detail Records datasets

In our research we used two smartphone datasets collected in the metropolitan area of Milan, Italy. This type of dataset, known as Call Detail Records, is collected automatically by the cellular network operators for billing purposes. The first dataset includes 17 sampling days (May 1st to 17th, 2013) and covers the whole metropolitan area, i.e. the city of Milan and surrounding districts; the second includes 67 days (March 26th to May 31st, 2012) and is limited to the city proper. When a user makes a call, sends a text message or accesses the Internet, the user id, the cell id of the handling towers, and also the date and time of established contacts are all recorded. In Figure 1 we report a small sample for each kind of recorded activity accompanied by a mobility trace that comes from combining the CDR entries. One of the advantages of this dataset with respect to other datasets [17, 10, 19, 3, 12] is the chance to leverage the Internet access data for purposes of mobility pattern analysis [4]. Although CDRs are rich sources for studying and analyzing human activities in different fields, they have two significant drawbacks as to providing location information. Both the spatial and the temporal granularities of CDR data are quite coarse.



CALL RECORDS

"source","destination","date","time","start_cell","end_cell","dir","duration"

574864,574865,"2012-03-27","13:36:54",47615,47615,"O",0

574864,574867,"2012-03-27","13:55:59",15824,15825,"O",46

574870,574864,"2012-04-02","22:37:41",16677,16677,"I",14

SMS RECORDS

"source","destination","date","time","cell","dir"

1916062,574864,"2012-03-27","21:48:53",16676,"I"

2267867,574864,"2012-03-30","21:59:05",16676,"I"

INTERNET RECORDS

"source","date","time","cell","upload","download"

574864,"2012-03-27","21:35:32",16676,15258,13721

574864,"2012-03-27","21:48:53",16679,76105,78993

574864,"2012-04-02","23:55:45",16677,84589,191681

MOBILITY TRACE

"source","date","time","cell"

574864,"2012-03-27","13:36:54",47615

574864,"2012-03-27","13:55:59",15824

574864,"2012-03-27","21:35:32",16676

574864","2012-03-27","21:48:53",16679

574864,"2012-03-27","21:48:53",16676

574864,"2012-03-30","21:59:05",16676

574864,"2012-04-02","22:37:41",16677

574864,"2012-04-02","23:55:45",16677

Figure 1: The format and a small sample of the call, SMS and Internet records. The last sample reports a mobility trace that combines the locations given by call, SMS and Internet records associated to a random user. Bold and green entries highlight the problems related to the temporal sparsity of CDR traces.

```
<Latitude, Longitude,0, Altitude,Date, String Date, String Time>
39.92203,116.34871,0,210,39784.0075347222,2008-12-02,00:10:51
39.922057,116.348755,0,213,39784.0075925926,2008-12-02,00:10:56
39.922098,116.348828,0,215,39784.007650463,2008-12-02,00:11:01      a)
39.89851,116.378975,0,113,39784.0175462963,2008-12-02,00:25:16
39.898462,116.379057,0,87,39784.0176041667,2008-12-02,00:25:21
39.898443,116.379141,0,95,39784.017662037,2008-12-02,00:25:26

40.004361,116.322342,0,84,39745.15625,2008-10-24,03:45:00          b)
40.004616,116.322365,0,139,39745.162962963,2008-10-24,03:54:40
```

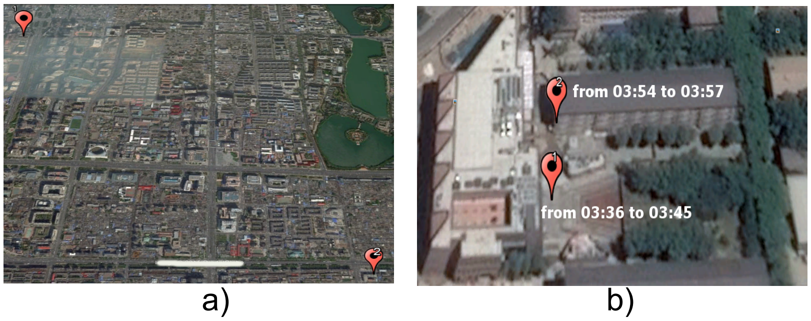


Figure 2: On the top a sample of a GPS trace. The records in (a) capture a movement between two PoIs not registered by the GPS device maybe due to the loss of signal (metro stops are close to PoIs). The map (a) shows the two locations. The path from 1 to 2 followed by the user is missing due the loss of the signal. In (b) we report a temporal gap concerning a user located at the same PoI. The user stays in position 1 for 9 minutes, then, after 9 minutes, s/he reappears in the close position 2.

Spatially, CDRs are accurate only up to the granularity of cell towers spacing, which varies from a few hundred meters in urban areas to several kilometers in rural areas. Moreover, in our datasets the cell position is not available (see Figure 1). Temporally, CDRs are generated only when phones are actively involved in a voice call, text message or Internet access. For instance, in Figure 1 we report a temporal gap on the same day (first green lines) and a 4-day long period (last green lines). From here on in, we denote the 17-day dataset as CDR-17 and the 67-day one as CDR-67.

3.2. Trajectories Dataset

We used the trajectories dataset collected in the GeoLife project and released by Microsoft Research Asia [38]. The dataset consists of a collection of GPS coordinates related to the movements of 178 people in a period of over 4 years. In the Microsoft experiment, people are equipped with GPS loggers or GPS-phones. Overall the dataset provides 17,621 trajectories with a total distance of 1,251,654 kilometers and a total duration of 48,203 hours. For purposes of our analysis, which is centered on the PoIs visited by the users during their daily lives, the most interesting characteristic of this dataset is its temporal and spatial fine granularity: namely, 91% of the GPS trajectory are recorded with a dense representation, every 1∼5 seconds or every 5∼10 meters per location sample.

```
<time,lat,long,altitude,accuracy,bearing,speed,locationProvider>
1335373194,46.023305,8.9171651,35.0,-1
1335373254,46.0233026,8.9171292,33.0,-1
1335373280,46.023382753204515,8.917032727886811,35.0,0.0
1335373313,46.0246694,8.9391406,2117.0,-1
1335373318,46.02354950628856,8.917070262296813,45.0,0.0
1335373395,46.027593795768375,8.918707672846919,40.0,0.0
1335373439,46.029563367280744,8.919618808404593,35.0,-1
1335373503,46.03534893257362,8.923844976591528,35.0,-1
1335373574,46.04323042666389,8.925787832537443,30.0,0.0
1335373629,46.0492995775597,8.926464822059568,30.0,0.0
1335373689,46.0575549649983,8.93063408585873,45.0,0.0
```

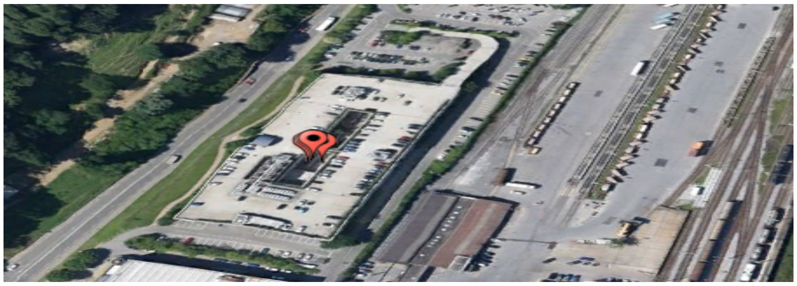


Figure 3: The format and a sample taken from the continuous mobility dataset. Besides the position we have information about the accuracy of the measurement and the technology leveraged to measure the position (Android Location Provider). In the map we visualize the first five lines of the sample.

However, the dataset has been built for the transportation prediction task, and thus does not directly characterize places. For this reason we developed a methodology to extract the places visited during the day, as briefly introduced in paragraph 3.4 and explained more in detail in Appendix A.2. In Figure 2 we report and visualize on the map two small samples taken from a user's trajectory. They illustrate two typical issues which will be further discussed in the next section.

3.3. Continuous Mobility Dataset

Although GeoLife represents the most reliable dataset publicly available, even after pre-processing its nature remains trajectory centered, and it differs from a continuously sampled dataset. The main difference between the trajectories and the continuous datasets consists of the fact that the first one contains only location samples related to movements among PoIs, while the second one also includes location data collected while visiting PoIs. For a clearer idea of the difference between the two types of dataset, we can think about how the mobile device collects the data: while collecting traces for a trajectory dataset, the user starts the location sampling as soon as he/she starts traveling on a path to a certain destination, and he/she stops the sampling as soon as he/she reaches the desired location; by contrast, while collecting continuous location data, after starting the sampling application on the mobile phone (in our continuous mobility dataset the sampling-start is automatic, performed by a background process at the phone bootstrap), it never stops unless the phone gets switched off. As opposed to the Microsoft one, which is a large dataset collected in a metropolitan area, we collected a dataset of continuously sampled coordinates locally in a small city environment during users' daily routine. We performed an experiment to collect traces over a time period of 20 days, from a group of 12 users [29]. The data collection system has been installed on the primary mobile phone of the users, to ensure they continuously carry it with them. The mobile phone sampling service performs a location reading every 60 seconds. The location information is provided by the Android OS Localization Manager, which queries both GPS and Network (WiFi or UMTS) Providers, so ensuring a continuous localization both outdoors and indoors. A sample of the resulting mobility trace is shown in Figure 3, where, in addition to the geographic position, we report other information such as the speed, the bearing and the accuracy of the measurement. The service runs continuously, collecting data 24/7 in the best of cases, for the whole duration of the experiment. For reasons of privacy , we gave the users the option of pausing the service manually. Thus, the collected data may present some holes rather than running non-stop 24/7.

3.4. PoI Extraction

In Appendix A we describe how we prepared our data to obtain a homogeneous description of people mobility. For a variety of reasons, each dataset needed to be pre-processed firstly in order to get the useful information and to make the users' traces fit for our purposes and analyses, and secondly to reconduct all the datasets to a unique representation, i.e. a sequence of temporal annotated Points of Interest (PoIs).

Given the different nature of the employed datasets, the characteristics of a PoI change slightly with respect to the analyzed data. Yet, its main meaning remains the same: namely, it is a place or area which is visited by a user. For the CDR datasets, a PoI is identified by a cell where a user is performing an on-phone activity (e.g., call, SMS, Internet access). However, for the Trajectory dataset, a PoI is identified by a place where the user is either standing still (data gap between consecutive trajectories) or an area within which the user is moving very slowly. Similarly, for the Continuous dataset, a PoI is identified by a high density of sampled location data. This corresponds to a standstill activity on the part of the user or to slow movements within a limited area. More details about the PoIs extraction methodology are presented in Appendix A.

The characteristics of the four datasets after the different pre-processing phases have been summarized in Table 1. The following analysis of the mobility behaviors is going to be based on the pre-processed datasets.

4. Relevance

We adopt a single user viewpoint to measure the importance of a PoI for a specific user. In particular, we are interested in evaluating the relevance of a

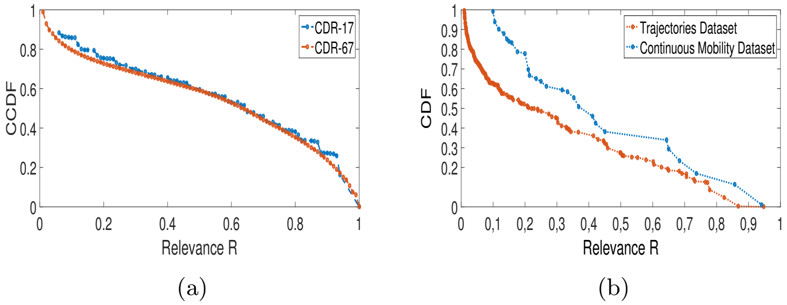
Table 1: Summary about the four datasets: cardinality of the datasets before and after the pre-processing, the number of days each trace spans at least and the number of visited PoIs.

Datasets	Number of Users		Number of Days	Number of POIs
	Before Preprocessing	After Preprocessing		
CDR-17	1,291,416	543,085	17	12,898
CDR-67	734,149	17,400	67	5,398
Trajectories	178	21	20	3120
Continuous	12	7	14	115

place in the user's daily mobility. The *relevance* R of a PoI P for a user u is defined as:

$$R(P,u) = \frac{d_{\text{visit}}(P,u)}{d_{\text{total}}} \tag{1}$$

where $d_{\text{visit}}(P)$ is the number of days a given PoI P has been visited (one or more times) by the user u and d_{total} is the total number of sampling days, *i.e.* it is the fraction of days the user has visited this PoI. Thus, $R(P,u)$ represents the probability that the user u visits the PoI P on any one day. We choose the day as temporal metric as it represents the fundamental time window when considering life routine of individuals. By means of the relevance we can capture how likely it is that an individual will move towards a place or return to it according to his/her tracking history.



(a)

(b)

Figure 4: Cumulative Distribution Function (CDF) of the relevance. In (a) the relevance distributions in the CDR datasets. In (b) the relevance distributions in the Trajectories and Continuous Mobility datasets.

The relevance distributions obtained from all traces are shown in Figure 4. CDR-17 and CDR-67 datasets, shown in Figure 4a, exhibit the same behavior, where a huge number of PoIs are visited only a few times, while some other PoIs are visited quite frequently (almost daily) and have a very high value of relevance. The median values are approximately 0.65 across datasets accounting for a highly regular pattern of PoI visits. A more pronounced trend characterizes the relevance distributions in the GPS traces, as reported in Figure 4b. Here we measure a lower value of the medians, which implies a higher number of places scarcely visited. Despite the fact that datasets are very different in nature, these results are very similar, thus confirming the generalizability of the relevance metric.

5. Relevance classes

People visit several PoIs per day, but different places play different roles in their lives. We propose the following PoI taxonomy organized in three classes, where each class accounts for places with different importance and semantic values in the user's daily life. As the importance of a place for a user is revealed by the frequency with which s/he happens to visit it, we resort to using relevance to measure it.

- **Mostly Visited PoIs (MVP)**: locations most frequently visited by the user. We can easily infer their semantic meaning, and associate them to home location and work place.
- **Occasionally Visited PoIs (OVP)**: locations of interest for the user, but visited just occasionally, such as the favourite place locally for hanging out with friends.
- **Exceptionally Visited PoIs (EVP)**: rarely visited PoIs.

The evaluation of the PoIs' relevance allows us a straightforward per-user identification of these three classes, as will be described in the following section. But simply by examining the aggregated relevance distribution shown in Figure 4 we can assign most of the probability distribution to the multitude of EVPs with very low relevance. Meanwhile, the first set of points expresses the few albeit highly relevant MVPs. The central part of the distribution contains OVPs.

5.1. Relevance class detection algorithm

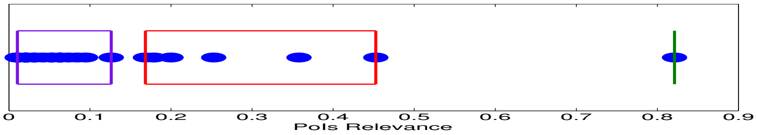
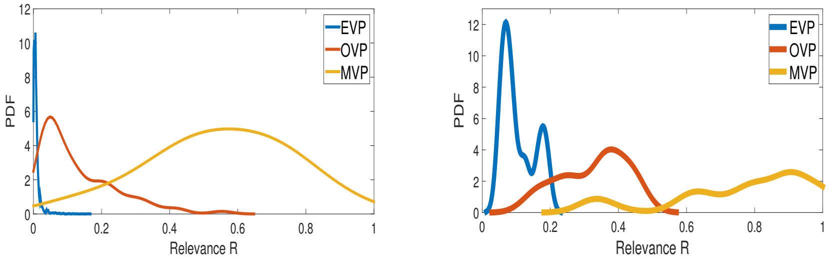


Figure 5: Three classes of relevance in a sampled user

Although the described classes of PoIs and their meanings are shared among all users, the relevance class bounds we use to identify them could be different on a per-user basis and cannot be fixed *a priori*. This argument advocates a clustering algorithm that adaptively adjusts according to the single user's mobility pattern. In particular, we adopt an unsupervised approach which groups the PoIs of a single user based on the PoI relevance and maximizes their separability. To this end we have chosen the k-means algorithm. To avoid the problem related to the initial choice of the centroids, we run 10 replicas of k-means with



(a)

(b)

Figure 6: Probability density function estimated through KDE (kernel density estimation) of the relevance in each class. The ordinates of EVP and MVP functions have been rescaled by a factor of 8 and 4, respectively, for a better visualization. In (a) and (b) the distributions for the trajectories and the continuous mobility datasets, respectively.

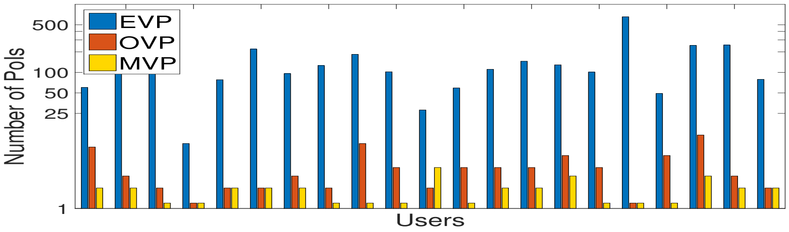
different initial seeds and choose the partition that minimizes the within-cluster sums of point-to-centroid distances, thus maximizing the separability. We run k-means with $k = 1, 2, 3$, then we assign to the user the number of relevance classes corresponding to the value of k with the best clustering performance, by choosing the value k which maximizes the silhouette separability. In Figure 5, as an example we show the result of the k-means, with $k = 3$, clustering on a sampled user. The EVP class (first box on the left) covers the range from 0.01 to 0.12, the OVP (central box) spans the range from 0.16 to 0.46 and the MVP class (first box on the right) contains only one PoI with relevance 0.82. In GPS datasets the best separability is achieved by $k = 3$ for nearly all users; however, the mobility captured by the CDR datasets is more varied and not every user satisfies the above classification.

In this section, we apply the class detection algorithm described above on the PoIs derived from the different datasets and analyze the obtained classes to extract their features.

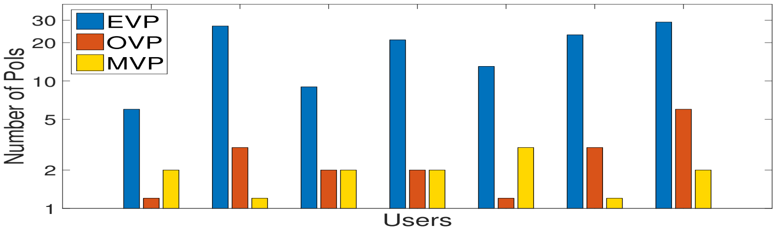
5.1.1. Trajectories and continuous mobility datasets

For each user, we apply the k-means algorithm (as explained in paragraph 5.1 for nearly all users the best separability is achieved by $k = 3$) to classify the related PoIs in three main classes of relevance (4) and over these classes we study three main features: *(i)* the number of PoIs which reside within each class of relevance, *(ii)* the percentage of time spent in each class and *(iii)* the average time of the visits to the PoIs of the classes.

The adoption of a clustering algorithm for detecting the three relevance classes allows us to adaptively select their bounds and avoid the choice of fixed thresholds. In fact, the application of a clustering algorithm best suits the diverse human mobility patterns and mitigates the spatio-temporal heterogeneity which characterizes the trajectories dataset. However the clustering of the relevance for each single user could generate overlappings among the classes of different users. For instance, relevance values which belong to the OVP class for



(a)



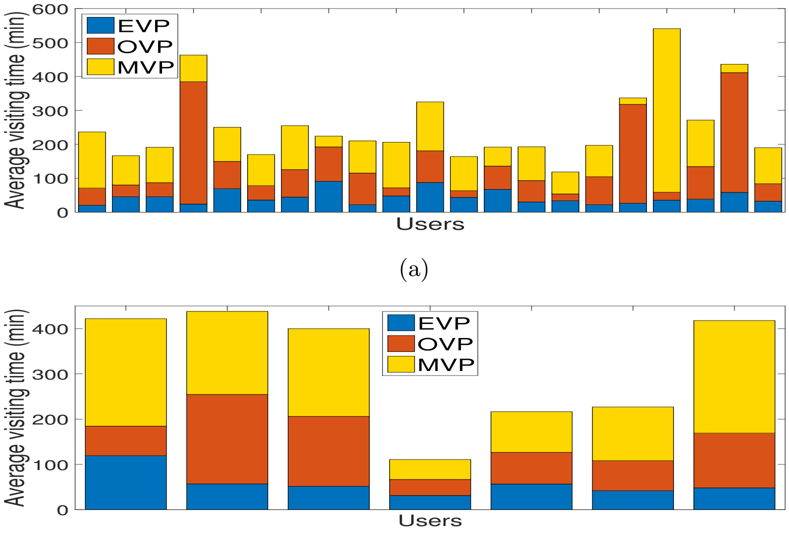
(b)

Figure 7: In (a) and (b) the number of PoIs per class of relevance, for each user. (a) reports the users in the trajectories dataset, while (b) the users in the continuous mobility dataset. In both figures y-axis is in logarithmic scale.

a user could correspond to the MVP class for another user. To verify whether that marginally happens, in Figure 6 we report the probability density function of the relevance for each class, obtained by kernel density estimation (KDE). We note that the three distributions are separable in both datasets. This suggests that the classes boundaries are similar among the users.

In Figure 7 we represent the per-user number of PoIs associated to each class of relevance. In Figure 7a we can observe the pronounced difference between the number of EVPs and the PoIs belonging to the other two classes of relevance (OVPs and MVPs) in the trajectories dataset: this is evidence of the fact that a user has the habit of visiting many new locations, but visits very few of them on a regular basis. By focusing on the classes OVPs and MVPs it turns out that the number of OVPs is limited and its average value is 4.19; also for the MVPs the number per user is limited, and its average value is 1.76. As expected, each user has a very small number of preferred locations (MVPs) which are visited daily (e.g., home, work place), and a higher yet still limited number of locations of interest (OVPs) which are visited with a lower frequency (e.g., gym, favorite restaurant, parent's house). As we note in Figure 7b the same behavior has been observed, with a few exceptions, in the continuous mobility dataset. In this dataset the average number of MVPs is similar (1.8) to the trajectories dataset, while the average number of OVPs is lower, due to a shorter observation period.

Figure 8 shows the average visiting times in the PoIs, grouped according to their class of relevance, and extracted from the trajectories and the continuous mobility datasets. From the figures we observe that for all users the average EVP visiting time is very limited and on average lower than one hour in both datasets. As for the OVP and MVP visiting times, the scenario is more faceted since the average visiting time for these classes depends on the mobility behavior of the user. In the trajectories dataset (see Figure 8a) some of the users tend to spend a long time in their MVPs, while other users have very long visit times in OVPs. Otherwise, in the continuous mobility dataset the behaviors are more pronounced as users usually spend more time in the MVPs. However, by considering the PoIs classification, we can see that MVPs and OVPs are equally relevant to the user, even if MVPs are visited more frequently than OVPs. Instead, EVPs are locations that are not really important to the user; they are where (according to the figure) s/he spends on average a shorter span of time.



(a)

(b)

Figure 8: Average visiting time per class of relevance in the trajectories and in the continuous mobility datasets.

In Figure 9 we represent a cumulative measure of the percentage of the total time each user spends visiting PoIs belonging to the three different classes of relevance. According to this figure, a user tends to spend half or more than half of the total time in the MVPs and the rest of the time is almost equally distributed between the EVPs and the OVPs.

5.1.2. CDR datasets

Smartphone traces differ from GPS datasets in many respects, as discussed in Section 3, both meaning and characteristics of PoIs extracted from these

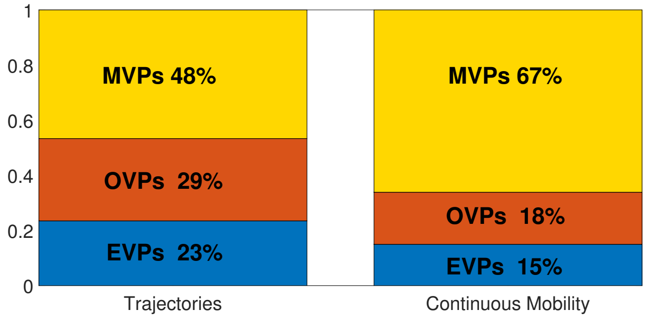


Figure 9: Percentage of the visiting time, per class of relevance, in both GPS datasets.

datasets are radically different, especially with reference to the relevance classes. First of all, the spatial granularity of PoIs is wider in smartphone data than in GPS data. In the former case, an urban PoI coincides with a cell tower and approximates a hexagon with a few hundred meters side. When a PoI is extracted from the GPS trajectory (see Section Appendix A) it approximates a circle with a radius of 60m. Consequently, a PoI extracted from a CDR dataset could actually aggregate other PoIs. This would require the finer grain of the GPS to emerge. For instance, a cell-based PoI could aggregate workplace and coffee shop or home and nearby stores. Moreover, the CDR datasets only record the cell where the user is performing a phone activity. As a result, the number of visited PoIs that can be extracted from a phone call dataset is smaller than the one obtained from trajectory datasets.

Users with fewer than 3 PoIs have been discarded: nevertheless, they represent only 1.53% and 0.01% of the users in the 67- and 17-day CDR traces, respectively. For all of the other users, we apply the k-means algorithm, as explained in paragraph 5.1. While in the GPS datasets for nearly all users the best separability was achieved by $k = 3$, in the CDR datasets the aggregation of PoIs in broader cells led to different results. For many users, PoIs clusterization according to their relevance achieves better performance when two (k-means with $k = 2$) or one (k-means with $k = 1$) classes are considered. Thus we consider three groups of users, each characterized by the number of relevance classes achieving the best performance in PoIs k-means clustering. The distribution of users among these groups is reported in Table 2. Only for about one third of users, those belonging to group 3, it is possible to identify all three classes of PoIs: MVP, OVP, EVP.

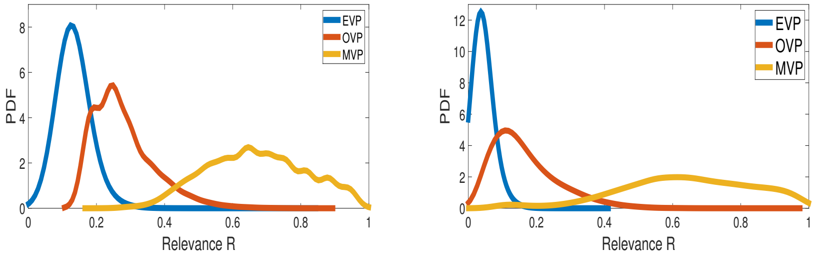
As mentioned above, the difference of k-mean algorithm output is due mainly to the spatio-temporal nature of CDR traces. For this reason, we limit our discussion to the 3-relevance class group.

In Figures 10a and 10b we show the distributions of the relevance characterizing MVPs, OVPs and EVPs in CDR-17 and CDR-67, respectively. In both CDR datasets, the relevance distributions reveal the high level of separability of the relevance classes. Besides, MVPs relevance is much higher than EVP and

Table 2: Users' distribution among groups identified by the number of mined relevance classes.

Group	Percentage of Users		Distinct visited cells	
	CDR-17	CDR-67	CDR-17	CDR-67
1	25.16%	18.42%	11,534	2,509
2	46.37%	47.6%	11,689	2,845
3	26.94%	33.97%	11,425	2,643

OVP ones, accounting for places actually visited very frequently and regularly, versus the two other classes which are visited occasionally and exceptionally.



(a)

(b)

Figure 10: Probability density function estimated through KDE (kernel density estimation) of the relevance in each class. EVP and MVP functions have been resized for a better visualization. Classes are separable.

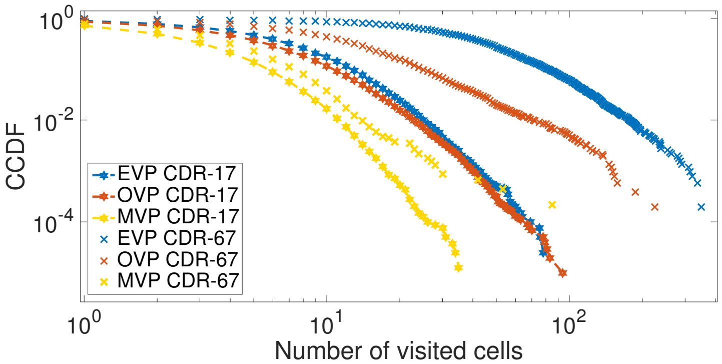


Figure 11: Distributions of number of distinct visited cells in group 3 in the different relevance classes.

In Figure 11 we represent the distribution of the number of distinct visited cells per user for each relevance class. In both cases, EVP and OVP distributions exhibit a heavy-tail behavior, while the MVP class covers a larger interval of relevance values. This result matches the location preference property in human mobility observed in [17, 36]. Moreover, we observe that the per-user number of distinct visited places increases when moving from 17- to 67-day CDR traces, with the consequence that the number of visited PoIs grows over time.

Finally, we enhance the generalizability of the feature of relevance class throughout different datasets by analyzing the percentage of PoIs lying in the 3 classes, as reported in Figure 12. The behavior is quite similar for all datasets. Most points belong to the EVP class; there are very few MVPs, while OVPs account for a number of places similar to the MVPs class.

We can therefore conclude that the classification we identified in terms of relevance at the beginning of this section (MVPs, OVPs, EVPs) is generally significant, since the distribution of the per-user number of PoIs associated to each class of relevance is similar across datasets with very different characteristics. We have shown that, independently of the dataset characteristics, the points visited by people fall mainly in the EVP class. However, most of the people spend most of their time in MVPs or OVPs; many of them can be found more than half of the time in MVPs.

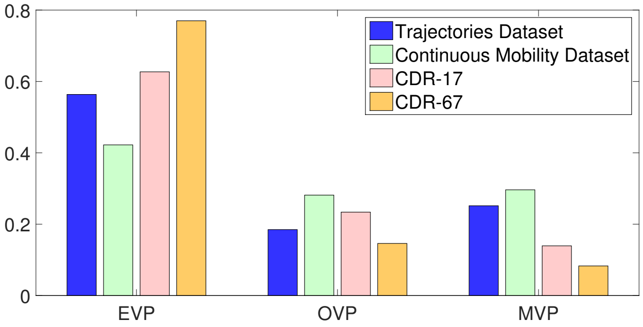
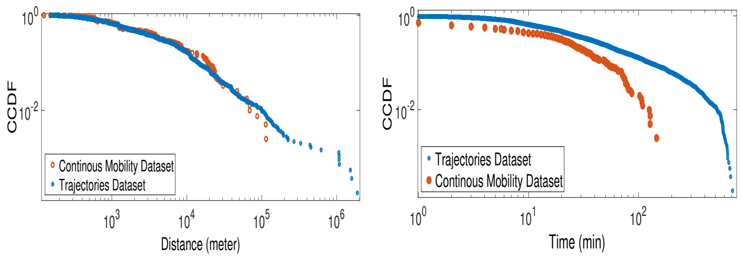


Figure 12: Percentage of PoIs in the relevance classes.

6. Time Distance versus Spatial Distance

All mobility studies and models in literature are based on the geographic distance between places: they assume that this is what underlies people's reasoning when moving. On the other hand, all services supporting human mobility - Google Maps, for instance - recognize that to a great extent people give priority to saving time. In fact, beyond the geographic distance, they compute the distance timewise between places for different modes of transportation. This is all the truer in cities where many different transportation systems offer people the opportunity to a minimum amount of time they need to get around town. Urban transportation systems per se are designed to minimize travel time by leveraging time-based and isochrone maps.

We aim to fill the gap between research studies and real-world mobility by analyzing the spatial and temporal distances between PoIs and the degree of correlation between them. This analysis is preliminary to the studying of the PoIs transition rules, since geographic distances, commuting time and PoIs relevance classes come into play in the decision process of the next PoI to be visited by individuals. The spatio-temporal features correlation requires a high level of accuracy. That's why we limit our analysis to GPS-based datasets. They provide a very high level of precision about the position, while the CDR-based data have coarse granularity and, in our case, the location of the cellular towers is unavailable.



(a)

(b)

Figure 13: (a) Complementary cumulative distribution function of the distance between consecutive PoIs for both datasets.(b) Complementary cumulative distribution function of the transfer time between consecutive PoIs for both datasets.

6.1. Geographic Distance

We measure the geographic distance between the departure PoI D and the arrival PoI A by considering their centroids and adopting the haversine formula to incorporate the Earth curvature. Some works in the literature [33], [17] have shown that the distance traveled and the radius of gyration follow a Pareto distribution with an exponential cut-off due to the spatial limits of human mobility and suggest that human movements can be modeled by a Levy-walk process. As evident in Figure 13a, we qualitatively observe the same kind of distribution in both datasets up to different geographic limits (longer tail in the GeoLife Project dataset). Consequently, these results are a further validation of previous works where only the spatial distance is considered for describing mobility of human beings [17].

6.2. Transfer Time

Taking inspiration from real life and from studies in urban planning, we do not limit our analysis to geographic distance. Rather, we observe that distance can also be expressed in terms of transfer time, *i.e.* the time needed to move from departure PoI D to arrival PoI A. The transfer time distribution of the dataset, as shown in Figure 13b, is also a power-law with a cut-off but it smooths the long tail of the geographic distance distributions. Specifically, whereas in the spatial case both distributions have the same trend except in the tail, if we consider the transfer time, we see that people behave differently. In fact the

Dataset	ρ
Continuous Mobility Dataset	0.4
Trajectories Dataset	0.1

Table 3: Pearson Correlation Coefficient (ρ) between geographical distances and transfer times on the Trajectories and Continuous Mobility Datasets

cut-off values are totally different; one and a half hours circa in the continuous dataset, and 4-5 hours in the GeoLife dataset.

The impact of this observation is fundamental as it suggests that time and space do not always match and are not always proportional. In particular, they do not match whenever long geographic distances are considered. We argue that the shorter tail in the time distribution is due to the fact that, in contrast to geographic distance distribution, in the time transfer analysis there are fewer occurrences of events far from the mean. It is unusual to spend more than a few hours in commuting between PoIs, while it is not unusual for the PoIs to be far from one another yet connected by fast transportation media.

6.3. Time Transfer and Geographical Distance Correlation

In our daily lives, we decide to move towards a particular place if we have enough time; by contrast, the current mobility analysis is driven only by the geographic distance. This dichotomy derives from the implicit assumption that time and distance are strictly related. Although this is roughly true on small scales, we find that the same does not hold in full when the mobility extends to, for instance, metropolitan or regional areas. To shed light on this aspect of human mobility we have computed the Pearson correlation coefficient between geographic distances and transfer times on both datasets, defined as:

$$\rho(tt, \Delta r) = \frac{\sigma_{(tt,\Delta r)}}{\sigma_{tt} * \sigma_{\Delta r}} \tag{2}$$

where $\sigma_{(tt,\Delta r)}$ is the covariance between the temporal and the geographic distances respectively, σ_{tt} and $\sigma_{\Delta r}$ indicate their standard deviations.

As shown by Table 3, when applied to the continuous mobility dataset, the Pearson coefficient is equal to 0.4. This indicates a small/medium degree of correlation; however, if we consider the GeoLife dataset it is equal to 0.1, meaning that the two quantities are not correlated. The above results indicate that in wider areas the adoption of different commuting strategies decreases the proportionality between the transfer time and the distance, typical of movement in small regions. Moreover they strengthen the difference between time and the geographic gap when measuring the distance among PoIs. To highlight this difference we show in Figure 14 the relation between geographic distance and transfer time. Considering a displacement typical of the urban/metropolitan area, we observe that the average transfer time has a sub-linear trend that accounts for the increasing speed of the different forms of transportation adopted to contract the geographic distances. This observation corroborates the intuition that temporal and spatial metrics capture different distances as the latter contracts the former. In particular these two factors should be considered separately whenever we study their impact on the human decisions involving the choice of the next destination.

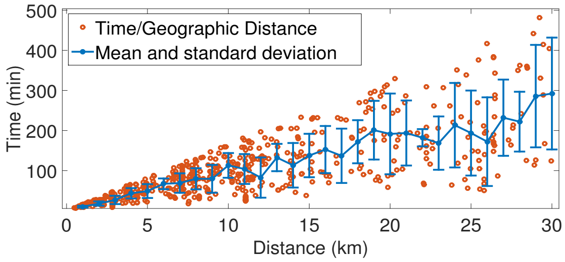


Figure 14: Relation between the traveled distance and the transfer time. Red dots denote the sample extracted from the GeoLife dataset and the blue line represents the mean trend (error bars correspond to the standard deviation).

Once the features characterizing the PoIs and the movement among them are illustrated, we aim to understand how they affect people's commuting between PoIs; in particular we want to measure the impact of the aforementioned features on the choice of the arrival PoI. Let us consider the transfers between the two PoIs D and A. Each transfer is characterized by the geographic distance between the two PoIs, the transfer time, the class of relevance of departure PoI D and the class of relevance of arrival PoI A. Given the relevance class of a destination, first we study the geographic distance or the transfer time a user is willing to spend. Second, we characterize the mobility among relevance classes exploring the probability of passing from class to class.

7. Transition rules

The human decision to move from one point to another emerges from a complex decision making process that is influenced by a variety of human and contextual behaviors. To improve the understanding of this process, we want to measure the impact of relevance, distance and time on the chance to get to a given arrival PoI A.

We start by investigating the impact of the geographic distance on the destination's selection process. To this end, we specifically analyze human behavior for the three relevance classes, EVP, OVP and MVP and we group the distance values in 500 m bins. As shown in Figure 15a and 15c where the joint probability of distances and classes is depicted, the behavior is very similar in both trajectories and continuous datasets. In all three relevance classes of destination we note a non monotone decrease of the visiting probability with a non negligible probability that people move also toward more distant PoIs, as predicted by a Levy-walk process and indicated by some peaks of brighter color in the right part of Figures 15a and 15c.

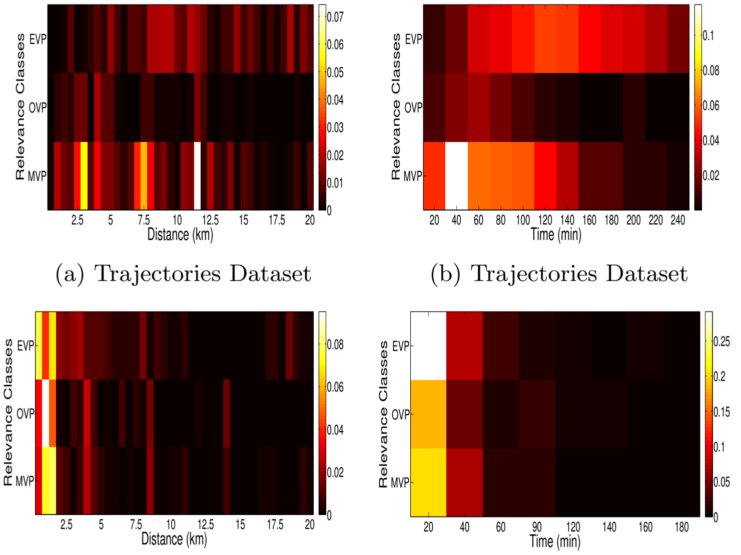
A different behavior can be observed when we consider the transfer time instead of the geographic distance. The visiting probability in the OVP and MVP is monotonically decreasing (color blurs from white to dark brown) with the temporal distance and reaches values close to zero according to different cut-off values, as shown in Figures 15b and 15d. This demonstrates that the transfer decision process of individuals is driven by the time they need to get to a place, as people are prone to focus on saving time. This observation advocates the paradigm shift in the analysis of human mobility we observed in Section 6: *the amount of time, not the distance, is the main parameter governing human decisions about movements.* Furthermore, although non monotone, the transfer time trend in the EVP is much smoother than in the geographic case. In particular, we can say that people who want to visit EVPs are willing to spend more time to reach these places, as the highest probabilities shift to 2-3 hours. This is due to the fact that a technological component affects human mobility, too, as people use different transportation means for different scales of distance. When people move in small areas, as in the continuous mobility dataset and in the right part of Figure 14, the commutation times do not differ much w.r.t different types of transportation. By contrast, when we consider a large dataset, the commutation times are highly affected by the means of transport.

Finally, the impact of the class of relevance of the departure PoI is independent of the scale of the scenario when we analyze the conditional probability to move from a PoI in a class c_1 to a PoI in a class c_2. As we can note in comparing Figure 16a and Figure 16b, both GPS-based datasets present the same characteristic despite the different geographic areas they span. Even if the conditional probabilities are heavily affected by the great number of EVPs, people commute to/from OVPs from/to MVPs, *i.e.* occasionally visited locations such as pub or free time spaces are related to home/work places (most visited PoIs). Clearly, even if people have to cover longer distances, they keep on moving between the places they frequent the most (MVPs: home and work), and some other OVPs (e.g. gym), and distance affects only the transitions to EVPs.

CDR traces present contrasting results. In Figures 16c and 16d the conditional probabilities of moving among the relevance classes in CDR-17 and CDR-67, respectively, are depicted. As shown in Figure 16c, we observe that the most probable movements occur between the same classes, *i.e.* the relevance class of the destination will likely be the same class as the departure location. Otherwise, movements among different classes are less probable. The scenario and the mobility habits change in the CDR-67 dataset. In this case (see Figure 16d), as in the GPS datasets, people mainly commute to/from MVPs from/to OVPs.

8. Semantic Analysis

We have established that the locations visited by people can be classified in terms of their relevance as well as the rules that characterize the mobility



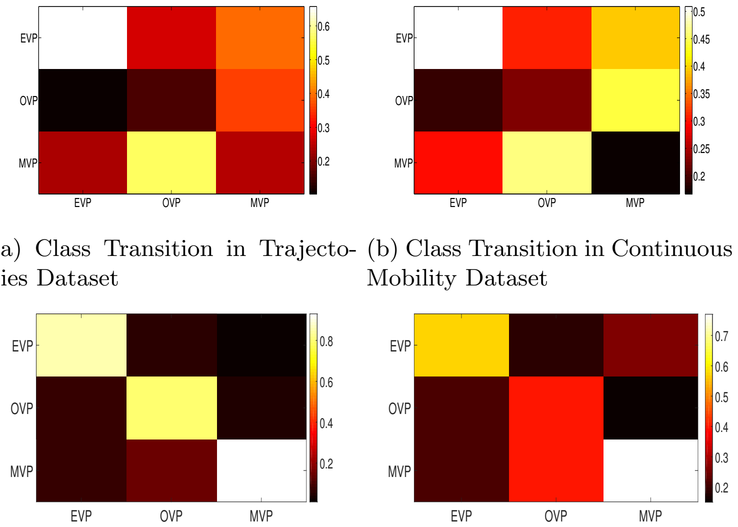
(a) Trajectories Dataset (b) Trajectories Dataset

(c) Continuous Mobility Dataset (d) Continuous Mobility Dataset

Figure 15: 15a) and 15c): joint probability distribution of the distance between consecutive PoIs and the relevance classes, $P(x \leq \Delta r < x + \delta, class = C)$. According to the heat bar, yellow and white squares represent higher probability. As regards distance we adopt 500 meter bins from 0 to 20 km (δ = 500m). 15b and 15d: joint probability distribution of the transfer time between consecutive PoIs and the relevance classes $P(x \leq tt < x+\delta, class = C)$. According to the heat bar, yellow and white squares represent higher probability. In this case, we adopt 20 min bins from 0 to 4 hours (δ = 20min).

between them. However, it is also important to understand the semantic value of such locations so as to better define human mobility. In particular, Home and Work are the most meaningful locations in human life. They are both characterized by a set of features, not shared with other places visited by a user. First of all they are the places people visit more frequently and regularly than others. This characteristic is fully measured by the relevance R described in the previous sections.

Therefore we decide to exploit R to identify home and work among all visited places. Specifically, places belonging to the class of most visited places (MVP) are the natural candidates for work and home identification as they have the highest relevance, as shown in Figures 10a and 10b. Beyond this main measure, a set of other features can help identifying home and work. Considering that these are the places where people spend the bulk of their lives, it is also reasonable to assume that they are the places where people perform the highest number of contact activities. Thus, we introduce a feature to quantify this aspect. Finally, to distinguish between home and work, we argue that, on average, people rarely spend most of the night at their workplace; therefore, we take into account the initial time of on-phone activities. The overview of the recognition strategy is



(a) Class Transition in Trajectories Dataset

(b) Class Transition in Continuous Mobility Dataset

(c) Class Transition in CDR-17

(d) Class Transition in CDR-67

Figure 16: 16a and 16b: transition probability among relevance classes. Each square represents the conditional probability to move from a PoI in a class c_1 to a PoI in a class c_2, *i.e.* $P(C_{new} = c_2 | C_{old} = c_1)$. On the x-axis the conditioning variable C_{old} and the on the y-axis the conditioned variable C_{new}.

presented in Figure 17, and it is mainly based on the relevance of a location. In the figure we represent only the values of the relevance which identify the MVP class for a given user.

We then apply this strategy to the two CDR datasets, as the two other datasets present a smaller number of users (which is statistically less significant). Furthermore, CDR traces are more demanding for such an analysis. In fact, as already mentioned, the CDR traces do not ensure a continuous tracking. So, it happens that some locations are not recorded regularly. Also, the position of a cell is not always a correct match w.r.t. the real user location, e.g. in the case of a ping-pong effect between two very close cells [34]. For this reason, CDR traces are perfect for illustrating that only the relevance is not sufficient to identify a location, and that we need to add some further features for assigning a meaning to the visited places.

```
Data: $\mathcal{L}$ = list of the locations visited by the user $u$
$H, W = null$;
$\mathcal{H} = heapify(\mathcal{L})$;
while $\mathcal{H}.size > 0$ do
    $L \leftarrow \mathcal{H}.extract\_max()$;
    switch $R(L, u)$ do
        case $R(L, u) \geq HighRR$
            if $H = null$ then $H \leftarrow L$;
        end
        case $R(L, u) \in [MediumRR, HighRR)$
            if Start time of contact activities during the NIGHT then
                if $H = null$ then $H \leftarrow L$;
            else
                if $W = null$ then $W \leftarrow L$;
            end
        end
        case $R(L, u) \in [LowRR, MediumRR)$
            if Start time of contact activities during the DAY then
                if $W = null$ then $W \leftarrow L$;
            end
        end
    endsw
end
```

Algorithm 1: Home/Work Place Recognition

As evident in Figure 17, we identify three relevance intervals where we can look for home and work candidate locations. If a location belongs to the red interval (High RR- on the right), it becomes the HOME. If more than one place have the same highest relevance due to the ping-pong effect, we recognize as HOME the place where most of the user's activities occur, discarding the other locations in High RR from the candidates set for work recognition. But as aforementioned, CDR traces are not punctual, so potentially the HOME location may not appear in the High RR interval. In this case, we can have a situation where HOME and WORK both have medium relevance (Medium RR-orange middle interval). Consequently, we need to introduce a further feature: the starting time of contact activities. We distinguish between night and day time. With this new feature, identifying contacts starting at nighttime, we again classify the highly ranked location as the HOME location. Otherwise, if it starts during day, we identify it as the WORK location. For low relevance (Low RR - on the left) home identification becomes less stringent since these users are very likely to live outside the city and come into town only for work purposes, so we identify only the WORK location. This is further detailed in algorithm 1. The algorithm receives a list of locations and builds the heap $\mathcal{H}$. In the heap, locations are primarily ordered by their relevance and by the number of

Dataset	HOME	WORK
CDR-17	$37093/80143 \approx 46.28\%$	$62258/80143 \approx 77.68\%$
CDR-67	$2577/4578 \approx 56.3\%$	$3383/4578 \approx 73.9\%$

Table 4: Percentage of recognized home/work locations.

activities on the part of user u in case of relevance equality. At each iteration the algorithm extracts and removes from the heap the maximum element and assigns it to the right relevance interval depicted in Figure 17. In the end the variables H and W contain the home and work whereas they are detectable.

The CDR traces we analyze are related to the urban area of Milan, which is why we consider the time interval 8 A.M. to 8 P.M. as day time. Similarly, from the relevance distribution, we can classify a point of interest as a location with high relevance when $RR >= 0.9$, *i.e.* being at home for at least 90% of the days. Medium relevance corresponds to $0.8 <= RR <= 0.9$, which means visiting a location at least $5-6$ days per week. We classify the relevance of a location as low if $0.65 <= RR <= 0.8$, which corresponds to 5 working days and also possible holidays. Otherwise the information is not significant. Also, the start time of the activities provides a semantic for distinguishing between home and work in the case of medium relevance: home if it is between 8 P.M. and 8 A.M. (when people are expected to be at home), work in all other instances.

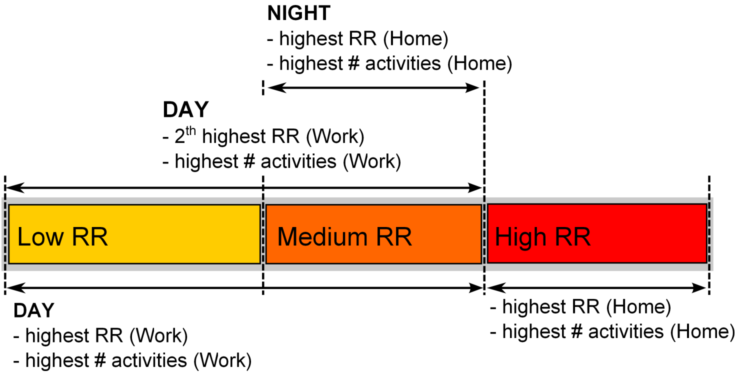

NIGHT
- highest RR (Home)
- highest # activities (Home)
DAY
- 2th highest RR (Work)
- highest # activities (Work)
Low RR
Medium RR
High RR
DAY
- highest RR (Work)
- highest # activities (Work)
- highest RR (Home)
- highest # activities (Home)

Figure 17: Home/Work place recognition process

In Table 4 we report the number of users for whom the algorithm is able to recognize the home and work locations. Overall we analyze 80,143 and 4,578 users belonging, respectively, to CDR-17 and CDR-67. Our methodology assigns a home location to 37,093 (46.28 %) and 2,577 (56.3 %) users, a work location to 62,258 (77.68%) and 3,383 (73.9%) ones, respectively. For users with low relevance in visiting MVP places, it is not possible to recognize their home/work places. Since a ground truth for the home/work detection does not exist, the goodness of the recognition algorithm is only partially verifiable. As already mentioned in Section 3, we exploit the billing mechanism to get an approximation of the ground truth. In particular the billing system records an Internet CDR every day at midnight indicating the position of the user. The most visited location on weekdays at midnight can be reasonably expected to correspond to the home location. Since the billing system is operator-dependent and undocumented in most cases, we have decided not to include this heuristic in the detection algorithm. Rather, we employ it in the evaluation. Keeping this setting, we measure a true positive rate equal to 0.83 in CDR-67, which is a good performance for the home detection task.

Dataset	HOME		WORK	
	Cell Level	Area Level	Cell Level	Area Level
CDR-17	83%	91.2%	69.73%	76.6%
CDR-67	83%	91%	56.5%	77.74%

Table 5: Conformity percentage of recognized Home/Work Places between Alhanson and Relevance based approaches

In addition we want to show that the relevance is of paramount importance and that our approach, where the main criteria is relevance, has some advantages compared to similar approaches that use different criteria. For that reason, we compare our algorithm to the one proposed in Alhansoun *et al.*[4] which uses only the highest number of total contact activities in day and night windows, to recognize home and work locations. The true positive rate of Alhansoun's algorithm for the home detection task is 0.63 in CDR-67, lower than the rate obtained by our algorithm. In Table 5 we observe that there is 83% match of recognized home places between the two approaches. For work places, the percentage drops to 69.73% and 56.5%, respectively, in CDR-17 and CDR-67 traces. If we consider the spatial granularity of a tracking area (which covers several nearby cell towers) instead of a single cell tower, the percentage of conformity between home places increases to 91.2 and 91, and the percentage between work places increases to 76.6 and 77.74 in CDR-17 and CDR-67. The differences in the recognized home and work places between our approach and the one presented by Alhasoun *et al.*[4] are due to the poor correlation between number of contact activities in a place and its relevance.

Figure 18 depicts the distributions of relevance of places recognized as work places by Alhasoun's approach [4], which are different from the places we recognize as work places. We observe that the majority of the work places recognized

Approach	**Dataset**	**Relevance Range**		**Number of recognized**	
		Home Places	Work Places	Home Places	Work Places
Relevance Based	**CDR-17**	0.80-1	0.65-0.90	37093	62258
	CDR-67	0.80-1	0.65-0.90	2577	3383
Alhasoun	**CDR-17**	0.42-0.88	0.27-0.93	80143	80143
	CDR-67	0.47-0.97	0.31-1	4578	4578

Table 6: Differences in the results among relevance-based and Alhanson approaches

by the approach described in [4] have low relevance, as shown in Table 6, although they have the highest total number of contact activities (since they get recognized). This means that most of these work places are not visited regularly by users; they do have, however, the highest number of on-the-phone activities. Also, places that have relevance higher than 0.9 can rarely be work places, since it is very unlikely that people went to work almost every day throughout the duration of the collected datasets. Therefore, we can conclude that our approach based on relevance allows to reduce the number of errors induced by the nature of CDR traces. Table 6 indicates the differences among the results obtained by the two approaches and highlights the relevance bounds which characterize home and work places extracted by Alhasoun's approach.

In the case of using GPS or WiFi datasets (high temporal continuity) the approach would be similar to what is discussed above; all the same, pause time duration would be used instead of the number of contact activities.

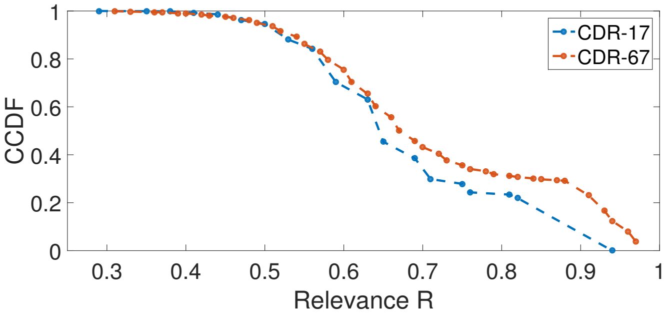


Figure 18: The CCDF distributions of the relevance of the places recognized as work places by Alhasoun's approach but not identified as work places in our approach.

9. Conclusion and Future Work

In this work we have taken a fresh look at the concept of location. We have proposed a general framework for extracting, characterizing, and classifying the points of interest of each individual according to their relevance for her/him. We have also proposed suitable metrics and algorithms to describe the semantic values of locations and the commuting rules among them.

Our key observations are as follows:

- individuals are regularly drawn to a limited set of locations where they spend most of their time;
- they also spend a significant amount of time in locations they only visit once;
- people commute between places based on temporal distance - not spatial distance - factors;

- HOME and WORK are among the most frequently visited locations, and, as such, the relevance R is a fundamental feature for their semantic identification.

These observations hold true across different datasets with completely different properties.

Based on above observations, we have derived a mobility framework where we are able to classify PoIs, the users and the way they move along PoIs, as well as the semantic meaning of PoIs. We have validated our framework with extensive experimental work.

These novel methods and results can change the way mobility is analyzed and modeled: we argue that, to produce more realistic mobility traces, a mobility model needs to consider (i) the new classifications of PoIs introduced, and (ii) the new features, their relationships and their different laws. Similarly, in localization activity, such laws can enormously simplify the prediction of the next location. In [29], the use of PoIs classification allows us to enhance the prediction (transition predictability) by a factor of 49% after fewer than 3 weeks of learning, while considerably reducing the costs. Finally, our framework successfully and powerfully combines social and physical characteristics, so it can serve as a basis for social analysis of mobile complex networks. This can be used, for example, in Recommendation Systems for Location Based Social Networks [26], where the next location can be recommended based on the class of locations that a user has already visited as well as on his/her own social history.

References

[1] Crawdad repository. `http://www.crawdad.org/`.

[2] R. Agarwal, V. Gauthier, M. Becker, T. Toukabrigunes, and H. Afifi. Large scale model for information dissemination with device to device communication using call details records. *Computer Communications*, 59:1–11, 2015.

[3] R. Ahas, S. Silm, O. Jarv, E. Saluveer, and M. Tiru. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1):3–27, 2010.

[4] 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 *SIGKDD International Workshop on Urban Computing*, UrbComp'14, 2014.

[5] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD'99, 1999.

[6] A. Arai and R. Shibasaki. Estimation of human mobility patterns and attributes analyzing anonymized mobile phone cdr: Developing real-time

census from crowds of greater dhaka. In *Proceedings 2nd AGILE PhD School 2013*, AGILE'13, 2013.

[7] N. Aschenbruck, A. Munjal, and T. Camp. Trace-based mobility modeling for multi-hop wireless networks. *Computer Communications*, 34(6):704–714, 2011.

[8] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[9] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[10] F. Calabrese, M. Diao, G. D. Lorenzo, J. Joseph Ferreira, and C. Ratti. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301–313, 2013.

[11] G. Colombo, M. Chorley, M. Williams, S. Allen, and R. Whitaker. You are where you eat: Foursquare checkins as indicators of human mobility and behaviour. In *Proceedings of 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, PerMoby'12, 2012.

[12] B. C. Csáji, A. Browetc, V. Traag, J.-C. Delvenne, E. Huensc, P. V. Doorenc, Z. Smoredae, and V. D. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473, 2013.

[13] M. Daoui, A. M'zoughi, M. Lalam, M. Belkadi, and R. Aoudjit. Mobility prediction based on an ant system. *Computer Communications*, 31(14):3090–3097, 2008.

[14] A. D. Domenico, E. C. Strinati, and A. Capone. Enabling Green cellular networks: A survey and outlook. *Computer Communications*, 37(0):5–24, 2014.

[15] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[16] A. Forster, K. Garg, A. H. Nguyen, and S. Giordano. On context awareness and social distance in human mobility traces. In *Proceedings of the Third ACM International Workshop on Mobile Opportunistic Networks*, MobiOpp'12, 2012.

[17] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[18] R. Hariharan and K. Toyama. Project Lachesis: Parsing and Modeling Location Histories. *Geographic Information Science*, 3234:106–124, 2004.

[19] S. Isaacman, R. Becker, R. Caceres, and S. Kobourov. Identifying Important Places in People's Lives from Cellular Network Data. In *Proceedings of the 9th International Conference on Pervasive Computing*, Pervasive'11, 2011.

[20] J. H. Kang, W. Welbourne, B. Stewart, and G. Boriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, WMASH'04. ACM, 2004.

[21] M. Kim and D. Kotz. Extracting a mobility model from real user traces. In *Proceedings of the 25th IEEE International Conference on Computer Communications*, INFOCOM'06. IEEE, 2006.

[22] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

[23] M. Lin, W.-J. Hsu, and Z. Q. Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12. ACM, 2012.

[24] E. Martin, K. H. P., S. Jorg, and X. Xiaowei. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2th International Conference on Knowledge Discovery and Data Mining*, KDD'96, 1996.

[25] E. Miluzzo, M. Papandrea, N. D. Lane, A. M. Sarroff, S. Giordano, and A. T. Campbell. Tapping into the vibe of the city using vibn, a continuous sensing application for smartphones. In *Proceedings of 1st International Symposium on From Digital Footprints to Social and Community Intelligence*, SCI'11. ACM, 2011.

[26] S. Mudda and S. Giordano. Regula: Utilizing the regularity of human mobility for location recommendation. In *International Workshop on GeoStreaming*, SIGSPATIAL '15, 2015.

[27] A. Noulas, S. Salvatore, L. Renaud, P. Massimiliano, and M. Cecilia. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE*, 7:1–10, 05 2012.

[28] M. Papandrea and S. Giordano. Enhanced localization solution. In *Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, PerMoby'12, pages 241–246, 2012.

[29] M. Papandrea and S. Giordano. Location prediction and mobility modelling for enhanced localization solution. *J. Ambient Intelligence and Humanized Computing*, 5(3):279–295, 2014.

[30] M. Papandrea, S. Giordano, S. Vanini, and P. Cremonese. Proximity marketing solution tailored to user needs. In *Proceedings of the 2010 IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*, WoWMoM'10, pages 1–3, 2010.

[31] M. Papandrea, M. Zignani, S. Gaito, S. Giordano, and G. P. Rossi. How many places do you visit a day? In *Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, PerMoby'13, 2013.

[32] F. Rebecchi, M. Dias de Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti. Data offloading techniques in cellular networks: a survey. *Communications Surveys & Tutorials, IEEE*, 17(2):580–603, 2015.

[33] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transaction on Networking*, 19(3):630–643, 2011.

[34] A. Rodriguez-Carrion, S. K.Das, C. Campo, and C. Garcia-Rubio. Impact of Location History Collection Schemes on Observed Human Mobility Features. In *Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, PerMoby'14, 2014.

[35] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010.

[36] A. S. Thakur, U. Kumar, and A. H. a. W.-J. Hsu. Analysis of Spatio-Temporal Preferences and Encounter Statistics for DTN Performance. In *Proceedings of the 7th International Wireless Communications and Mobile Computing Conference*, IWCMC'11, 2011.

[37] H. Tuncer, S. Mishra, and N. Shenoy. A survey of identity and handoff management approaches for the future internet. *Computer Communications*, 36(1):63–79, 2012.

[38] Zheng, Yu, Zhang, Lizhu, Xie, Xing, Ma, and Wei-Ying. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW'09, 2009.

[39] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW'08. ACM, 2008.

[40] C. Zhou, N. Bhatnagar, S. Shekhar, and L. Terveen. Mining Personally Important Places from GPS Tracks. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, ICDEW'07. IEEE, 2007.

[41] M. Zignani. Geo-CoMM: A geo-community based mobility model. In *Proceedings of the 2012 9th Annual Conference on Wireless On-demand Network Systems and Services*, WONS'12, pages 143–150. IEEE, 2012.

[42] M. Zignani and S. Gaito. Extracting human mobility patterns from GPS-based traces. In *Proceedings of IFIP Wireless Days*, WD'10, pages 1–5. IEEE, 2010.

[43] M. Zignani, S. Gaito, and G. P. Rossi. Extracting human mobility and social behavior from location-aware traces. *Wireless Communications and Mobile Computing*, 13(3):313–327, 2013.

[44] M. Zignani, M. Papandrea, S. Gaito, S. Giordano, and G. P. Rossi. On the key features in human mobility: Relevance, time and distance. In *Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications Workshops*, PerMoby'14, 2014.

Appendix A. Pre-processing and general statistics

In this appendix we describe the filtering process and characterize the datasets specifying their most important properties. In particular we present some methods which allows us to reduce the different mobility traces to a sole representation, i.e. a sequence of temporal annotated Points of Interest (PoIs).

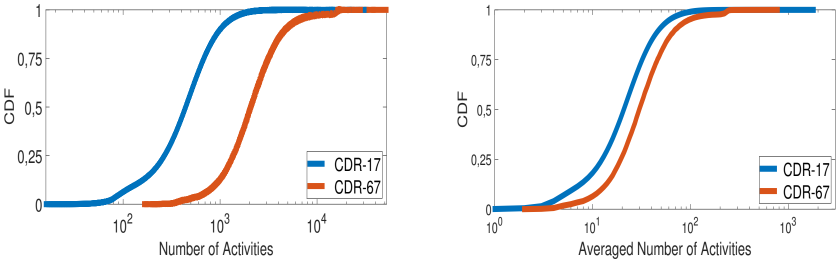
Appendix A.1. CDR datasets

To extract mobility characteristics of individuals we need to have enough CDR samples to study the movement of users. Therefore we select users with at least one activity per day in each trace and we restrict our analysis to this subset of users. Also, we combine call/SMS and Internet traffic records to get more data about users' positions. An Internet traffic record has the same format as an SMS one. Specifically, it reports the position of the user every 10 Mb of traffic and at midnight. This way, we can consider as Points of Interest for a user, the cells he/she visits, *i.e.* where he/she performs an on-the-phone activity. The number of users and the number of visited cells covered by each dataset have been indicated in Table A.7. The results indicate the portion of active users w.r.t. the total number of users by increasing the geographic area.

Figure A.19a reports the cumulative distribution function (CDF) of the aggregated number of activities (SMS or call). To fit the empirical distributions, we compare different distributions, whose parameters have been estimated by

Table A.7: The number of users and network cells in the CDR datasets. The last column reports the number of users that our analysis is based on.

Dataset	Users	Cells	Users with at least one activity per day
CDR-17	1,291,416	12,898	543,085
CDR-67	734,149	5,398	17,400



(a)

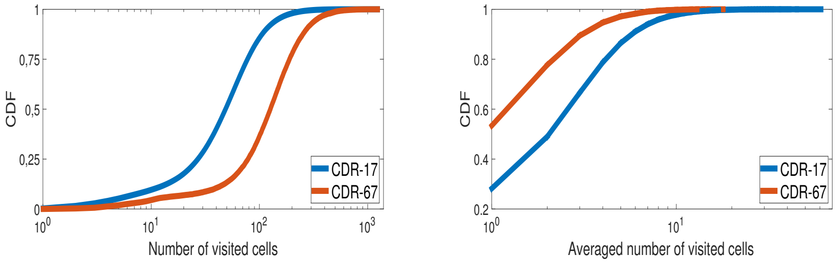
(b)

Figure A.19: (a) Distribution of the number of activities per user. (b) Distribution of the averaged number of activities per user per day.

MLE; and from those that pass the Kolmogorov-Smirnov (KS) goodness-of-fit test[2], we select the model which gets the lowest KS statistic. The evaluated distributions are Log-Logistic (3P), Log-Logistic, Pearson, Log-Pearson, Log-Normal, Log-Normal (3P), Weibull (3P), Weibull, Gamma, Log-Gamma, Exponential, Pareto, Levy, Chi-Squared. According to the above method the Log-Logistic (3P) distribution with parameters $\alpha = 2.4575$, $\beta = 1978.8$ and $\gamma = 83.932$ (p-value ≈ 0.2632) obtained the best result for CDR-67 traces. For CDR-17, none of the mentioned distributions passed the test. The average and standard deviation of the number of activities per user in CDR-17 traces are circa 532 and 412 contacts; in CDR-67 traces these values are higher, 2,722 and 2,578 respectively, as the observation period is much longer.

Figure A.19b shows the CDF of the number of activities per user, averaged over the span of a day. We observe that the distribution related to CDR-17 is located above the one related to CDR-67. We applied the average over the day in order to have comparable values: the measured average corresponds to 25 ($\sigma = 20$) in CDR-17 and 40 ($\sigma = 38$) in CDR-67. In general, by combining the information of the above distributions, the set of users captured by the CDR datasets are quite active and some of them are very active. That represents a good advantage since active users result in more mobility data.

[2] 'Data follow the distribution X ' is the null hypothesis. A p-value greater than 0.05 usually indicates that the null hypothesis has not been rejected.



(a) (b)

Figure A.20: (a) Distribution of number of distinct visited cells per user. (b) Distribution of averaged number of distinct visited cells per user per day.

In Figure A.20a we report the distributions of the number of distinct visited cells per user for each dataset. First of all, almost 90 percent of users have visited fewer than 100 and 260 distinct cells, respectively in CDR-17 and CDR-67 traces. This implies that most of the people visit a limited number of cells (places), while only a few of them visit a huge number of cells [36]. The CDF of CDR-67 lies under the 17-day CDR trace, implying that over a longer period people are more likely to discover and visit new places [17]. The best fitted distributions (from those on the already mentioned list) of the number of distinct visited cells are Log-Normal (3P) with parameters $\sigma = 0.6108$, $\mu = 4.125$,$\gamma = -14.693$ and p-value ≈ 0.646 for CDR-17, and Log-Logistic (3P) with parameters $\alpha = 3.6538$, $\beta = 183.1$ and $\gamma = -57.57$ (p-value ≈ 0.6455) for the CDR-67 dataset. In broader terms, the number of distinct visited cells follows a heavy-tailed distribution.

Figure A.20b reports the CDF of the number of distinct visited cells per day and per user. Most people visit on a daily basis a very low number of cells, median values are 1 in CDR-67 and 2 in CDR-17; but there is a long tail accounting for people who visit many cells every day. As the considered mobility area is larger in the 17-day CDR dataset, this dataset captures a higher number of locations visited per day by users.

Although our CDR traces have a higher number of users than the other two GPS datasets, we should note that CDR traces are more sporadic in the temporal dimension and coarse in the spatial one w.r.t GPS dataset. However, we are able to extract the distribution of the pause time in CDR-67 as reported in Figure A.21. We note that cells visited for periods shorter than an hour are very frequent, while locations where people spend more than 7 hours exist and are limited in number (25% of visits).

Appendix A.2. Trajectories dataset

Although GeoLife represents the most reliable dataset publicly available, it was not collected to find visited locations. So, for our purposes, we had to pre-process trajectories in order to determine the most meaningful locations. The need for a pre-processing phase is dictated by the dataset bias which favors

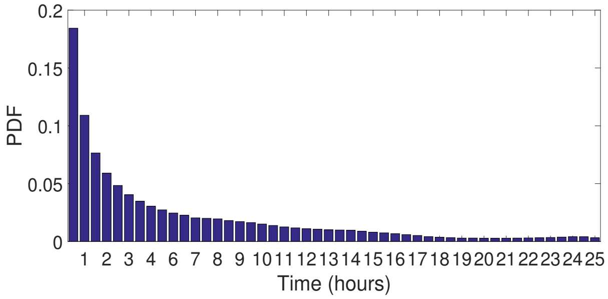


Figure A.21: Probability Distribution Function (PDF) of the pause-time in the CDR-67 dataset. Each bin is one hour size.

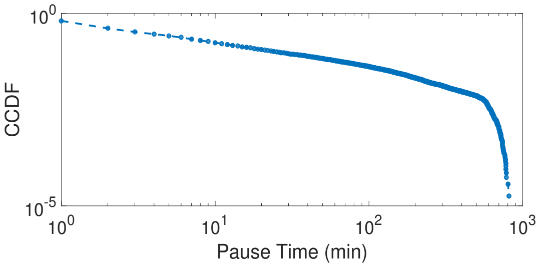


Figure A.22: CCDF of the aggregated pause times in the stay-locations.

movements, while we are interested in the activity of visiting PoIs. In particular we aim to densify trajectory points corresponding to the pause phase by a filling heuristic. Meanwhile, we remove the points belonging to users' movements.

Indoor filling. Mobility data collected by GPS devices present gaps because GPS signals are often disrupted inside buildings. This represents a big problem, especially if we are interested in detecting the PoIs of a user. In fact, in many cases most of the PoIs visited by a person during the day are buildings or other indoor locations. This situation has been depicted in Figure 2b, where a user reappears after about 20 minutes at a position close to the previous one. To overcome the problem given by missing records [23], and to avoid an underestimation of the number of PoIs, we apply the following simple rule. When the ending and beginning GPS points of a gap are within a distance of 35 meters and the gap duration is greater than 5 min, the user is taken as residing at the same location during that time. This rule also works in the situation where the individual enters a building, or where the individual turns off the GPS devices in an indoor place. Practically, we add as many GPS points equal to the entry point as the duration in sec of the gap. After the trajectory reconstruction phase, we noticed a big increment of points, anyway limited by the threshold imposed on the gap duration.

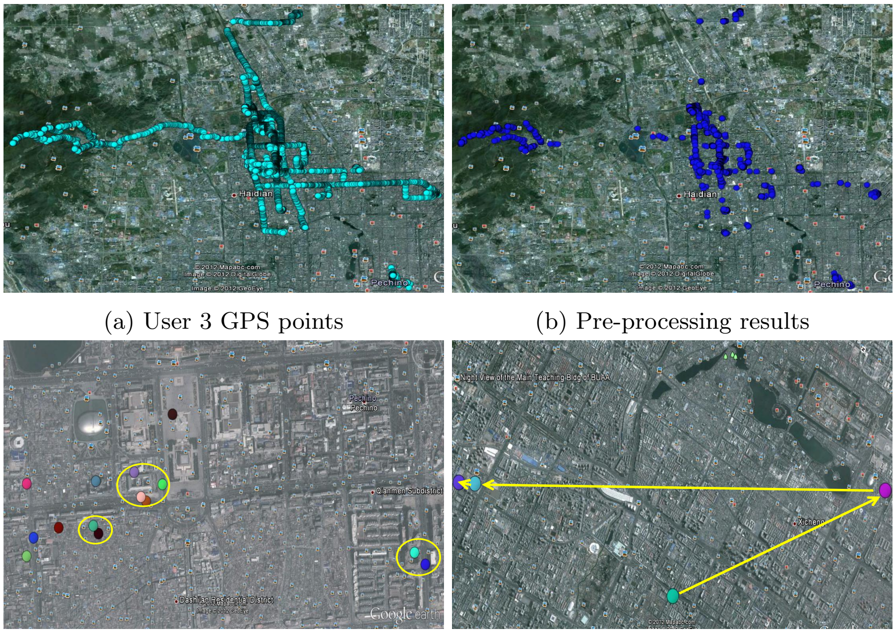
Movement phase reduction. We apply a filter with the goal of leaving out data which describe the movements among the PoIs that a user visits, thus reducing the number of points to analyze. This way we consider the periods in which a user stands still in a place, assuming that users manifest their interests by spending a certain amount of time in such places. In order to extract the pause periods and their related GPS points from the whole individual trace, we apply the heuristic proposed in [43, 42], where a similar but smaller dataset has been analyzed. If two points p_i and $p_i + 1$, with timestamps indicated by $t(p_.)$, do not satisfy

$$\frac{\|p_{i+1} - p_i\|}{t(p_{i+1}) - t(p_i)} \leq \Delta \tag{A.1}$$

then we delete p_{i+1} from the original trace, since it belongs to the movement phase. Analyzing walking mobility data, we set the threshold to the very low value of $\Delta = 1.3m/s$, according to the fact that we observe that human walking speed is about 4-5 km/h (1.1-1.4 m/s). It seems a reasonable value as generally, in a location, people do not reach the maximum speed. This way, we capture points where a person is standing still or is moving very slowly inside a small area. The result of the speed filtering process is a sequence of points that forms the trajectory $S = ((p_1, t_1), ..., (p_n, t_n))$, where t_i is a timestamp and $p_i \in \mathbb{R}^2$, on which we apply the PoIs extraction methodology proposed in Section Appendix A.4. In Figure A.23b we show the results of the movement phase reduction applied to the raw trace reported in Figure A.23a.

Users' selection. The point reduction also has effects on the number of users and the number of days, per user, from which we can extract places of interest. The reduction is mainly due to the fact that the GeoLife dataset has been built for the transportation prediction task, and, as a consequence, it favors movements.

To overcome these limitations we classify the users by considering two properties: the period (in hours) a single day trace spans and the number of days the single user traces cover. In particular, for each user, we only consider the daily traces that record more than h hours. On these tracks we count the number of users that have more than d days of data. In particular, for all the users of the dataset we filter out all the days of sampling (data collected within 24 hours, from 00:00 A.M. until 11:59 P.M.) which have $h \leq 3$ hours of sampling. All the remaining days are considered *relevant days*. After this first processing, we filter out all the users which collected fewer than 20 *relevant days* of data ($d = 20$): the resulting number of users is 21, out the total number of 178 users. The above thresholds have been chosen to optimize the trade-off among the importance of having a large number of users, the chance to generalize our analysis and the need to deal with sampled data which does not only correspond to trajectories. For example, only by increasing the threshold h by one hour we obtain a number of users insufficient for purposes of our goal (10 users). Note that the resulting dataset, even with a reduced number of users, still almost fully spans the original GeoLife as to time period.



(a) User 3 GPS points

(b) Pre-processing results

(c) Sub-PoIs issue

(d) Compact representation

Figure A.23: PoIs extraction applied to the user 3's trajectories. In (a) we plot all the recorded points (raw data). In (b) we show the points resulting from the application of the pre-processing phase. In (c) we depict the sub-PoIs that have to be grouped in the real PoI (yellow circle) while (d) is a compact representation of user 3's mobility during a single day.

Appendix A.3. Continuous Mobility Dataset

Even if the tracking service runs continuously, for privacy reasons we allowed the users to manually pause it. Thus, the collected data is not always a 24 hour continuous data flow, but may present some holes. Also from this dataset we select a subset of significant users which have collected at least 14 relevant days of data (two weeks), where a relevant day includes at least 6 hours of location sampling. The resulting number of relevant users we are considering for our study is 7. To identify the user's relevant PoIs, in this case, we only act on the algorithm tuning [29]).

With respect to the number of detected places visited by users, we observe that on average the number of distinct visited PoIs is 16, while the median amounts to 1, like the previous datasets.

Appendix A.4. From GPS traces to Points of Interest

GPS datasets, like the ones we are analyzing, present many difficulties concerning the PoIs extraction task as to the mobility data inferred from geo-coded or geo-tagged social networks [11] (e.g. Foursquare, Facebook Places,. . .). In our context we do not have any information about the interest expressed by the user, but we must rely only on the periods when a user is standing still.

If we assume a constant sampling rate, as in our case, the pause periods and the places visited by users translate into a higher concentration of recorded points. Thus, the PoIs extraction corresponds to the unsupervised task of density-based clustering. In particular, we are extending the methodology proposed in [43], adopting a two-level density-based clustering combined with a thresholding mechanism based on pause in the regions extracted by the first clustering phase.

All the points of a trajectory belong to the pause phase and are the starting points for extracting the PoIs. To reach this goal, we first find the possible regions of interest via a clustering algorithm and then we detect the real PoIs considering the pause time feature.

Formally, we capture the possible regions by introducing the concept of stay-location L.

Definition 1. Let S be a trajectory and $L = \{L_1, \ldots, L_k\}$ a partition of $\{p_1, \ldots, p_n\}$ s. t. for each $L_i \in L$, L_i is maximal w.r.t. the property that for each $p_u, p_v \in L_i$ exists a sequence $(p_u = p_w, ..., p_{w+j} = p_v)$ of points in L_i, s.t. $\|p_{w+k} - p_{w+k+1}\| \leq \delta, k = 0, ..., j-1$ for a fixed δ. A stay-location is an element of L.

Informally, a stay-location is an area where a person stops, independently of how long s/he stays there. Let us consider individual traces in order to extract stay-locations and analyze their properties. To find stay-locations we apply the density-based clustering algorithm DBSCAN [24]. As DBSCAN parameters we use $\delta = 10$ mt and $\epsilon = 2$ neighbors (δ represents the maximum distance such that two points are considered neighbors, while ϵ is the minimum number of neighbors that a node must have to be considered in a cluster).

We observe that in daily movements there are many stay-locations where an individual stays for a short amount of time. These stay-locations are meaningless as they represent small pauses in the movement towards the real destinations that we call Points of Interest.

Definition 2. Let S be a trajectory and $L_i \in L$ a stay-location. L_i is a Point of Interest (PoI) if in S there exists a subsequence $((p_i, t_i), \ldots, (p_{i+k}, t_{i+k}))$ such that $p_{i+j} \in L_i$ for $j = 0, \ldots, k$ and $t_{i+k} - t_i \geq \phi$.

In the analysis of the dataset performed in this paper, we set the threshold $\phi = 5$ min, which corresponds to the average of the pause distribution in stay-locations, shown in Figure A.22. We must underline that we do not consider the sum of the pause times in a stay-location; rather, we consider the single values. The thresholding results in the meaningful PoIs, although we observe situations, like those presented in Figure A.23c, where we have many sub-PoIs of the same general PoI. To overcome this impasse we run DBSCAN with a larger ϵ on the centroids of the sub-PoIs detecting the real points of interest. This way we obtain two important results: we drastically reduce the number of stay-locations and we can infer which are the main destinations, i.e. the PoIs.

In addition to finding PoIs, the above methodology has the ability to express human mobility as a compact trace that summarizes the transitions between PoIs and the users' pause time in them as shown in Figure A.23d.

The detection of the PoIs allows us to compare the mobility habits in terms of visited places with the CDR datasets. In fact we obtain an average number of PoIs per user comparable to CDR-67 datasets, *i.e.* 148, and the same median of the number of places visited per day.