Mining Mobility User Profiles for Car Pooling

Roberto Trasarti Fabio Pinelli Mirco Nanni

Fosca Giannotti

KDD Lab, ISTI, CNR
email: name.surname @ isti.cnr.it

ABSTRACT

In this paper we introduce a methodology for extracting mobility profiles of individuals from raw digital traces (in particular, GPS traces), and study criteria to match individuals based on profiles. We instantiate the profile matching problem to a specific application context, namely proactive car pooling services, and therefore develop a matching criterion that satisfies various basic constraints obtained from the background knowledge of the application domain. In order to evaluate the impact and robustness of the methods introduced, two experiments are reported, which were performed on a massive dataset containing GPS traces of private cars: (i) the impact of the car pooling application based on profile matching is measured, in terms of percentage shareable traffic; (ii) the approach is adapted to coarser-grained mobility data sources that are nowadays commonly available from telecom operators. In addition the ensuing loss in precision and coverage of profile matches is measured.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: User profiles and alert services; H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms

Keywords

Trajectory patterns, Spatio-temporal data mining

1. INTRODUCTION

The analysis of movement data has been recently fostered by the widespread diffusion of new techniques and systems for monitoring, collecting and storing location-aware data, generated by a wealth of technological infrastructures, such as GPS positioning and wireless networks [7]. These have made massive repositories of spatio-temporal data available that record human mobile activities, such as data location from mobile phones, and GPS tracks from mobile devices.

There are many potential opportunities, and movement data have already been recognized by private and public institutions as a valuable source of information to assess the lifestyle, habits and demands of citizens in terms of mobility.

The traditional use of mobility data, for instance in the context of urban traffic monitoring and transportation planning, mainly focuses on inferring simple measurements and aggregations, such as density of traffic, and car flows on road segments. Recent research in mobility analysis has been extended in order to identify the behaviors that people (taken as a whole group, rather than as individuals) consistently follow, such as groups of trajectories with a common route [2] or popular itineraries [6]. This can also be extended to problems that need to link different geographical areas. This is because we can now infer information in terms of origin-destination pairs of areas that exchange traffic, as well as the routes along which the exchange occurs.

Despite the great attention that this area has attracted, current work on mobility analysis largely neglects a key element that lies in between single trajectories and a whole population, i.e. the individual person, with his/her regularities and habits, that can be differed from the population. In fact, analysing individuals (rather than just large groups) provides the basis for an understanding of systematic mobility, as opposed to occasional movements, which is fundamental in some mobility planning applications, e.g. public transport. This paper proposes a framework that accommodates a "middle level element", providing a two-phase process: first an individual-centered mobility model extraction; then a population-wide analysis based on the individual models. Our framework can be seen as a new approach in the learning paradigm since it provides a local-to-global analysis.

The main contributions of this paper are: (i) we introduce the concept of the *mobility profile* of a user as the set of his/her *routine trips*, and define a general method based on trajectory clustering to extract such profiles; (ii) we show an instantiation of the method on the GPS data of vehicles with a *route similarity* function. We perform an empirical evaluation of the effects of the different parameter settings of the method; (iii) we propose a car pooling service on the basis of the GPS-based profiling method; (iv) we study the robustness of the general method w.r.t. the downgrading of the spatio-temporal richness of the data. We thus generated a synthetic GSM-like dataset out of the GPS data, and instantiated the method on these data. The final comparison

KDD'11, August 21–24, 2011, San Diego, California, USA.

confirms the goodness of the method for profile extraction, in the sense that comparable profiles are extracted for each user. In contrast, the car pooling service does not appear to be well supported by the second scenario.

The rest of the paper is organised as follows. In Sec.2 we present the background on mobility data mining, in Sec.3 we present the methodology for the mobility profile extraction in terms of GPS-equipped vehicles. In Sec.4 we define the match between a pair of profiles and its application to a car pooling service. In Sec.5 a downgrading quality precision process is presented and its impact on the mobility profile extraction and matching is investigated in detail. Section 6 highlights the most promising future lines of research.

2. BACKGROUND AND RELATED WORK

The core contribution of this paper consists in a methodology for mobility analysis, aimed in particular to match users based on their individual mobility behaviours. While mobility data mining [7] as a whole is a recent research field, a significant amount of works with similar objectives already exists.

A close body of research can be found in the mobility clustering literature, aimed at finding sets of trajectories that are similar. The standard approach adapts classical distance-based algorithms and defines ad hoc distances for trajectory data [10], possibly with limited ad hoc refinements [2]. Alternative, ad hoc solutions include variants of model-based clustering [5], collective movements detection methods [9], and others. Based on similar solutions a recommendation system is proposed in [14,15] where the user trajectories are translated into sequence of regions of interest and then compared. As opposed to existing solutions, in our proposal the evaluation of similarity between individuals is not realized as a direct comparison of trajectories. Instead, a two-phase analysis is performed, that first finds typical behaviours for each single user (and this task clearly falls within the scope of trajectory clustering), and then compares pairs of users through a comparison between their corresponding sets of behaviours. The research in analyzing the mobility data is increasing; most works try either to infer simple movement aggregates (average speed, etc.) in local areas, with an emphasis to real-time applications, or to discover macro-level laws of human mobility, such as the law governing the distribution of traveled distances [8, 11], with applications to high-level evaluations of predictability of human mobility [12]. The analysis framework proposed in this paper can adapt to different mobility data types, and indeed a comparative study is performed on several of them, which provides insights on how spatial granularity (GPS precise points vs. GSM large cells) and sampling rate (continuous vs. phone call rates) affect the process.

3. MOBILITY PROFILES EXTRACTION

The daily mobility of each user can be essentially summarized by a set of single trips that the user performs during the day. When trying to extract a *mobility profile* of users, our interest is in the trips that are part of their habits, therefore neglecting occasional variations that divert from their typical behavior. Therefore in order to identify the individual mobility profiles of users from their GPS traces, the following steps will be performed - see Figure 1:

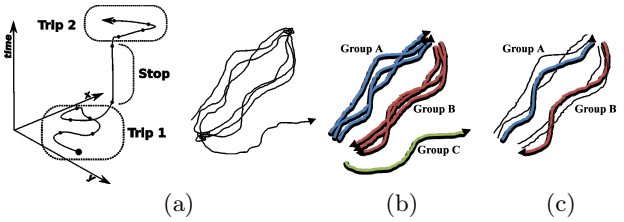

Figure 1: Mobility profile extraction process: (a) trip identification; (b) group detection/outlier removal; (c) selection of representative mobility profiles.

1. divide the whole history of the user into trips (Figure 1(a))
2. group trips that are similar, discarding the outliers (Figure 1(b))
3. from each group, extract a set of representative trips, to be used as mobility profiles (Figure 1(c)).

3.1 Mobility profile definitions

Trips. The history of a user is represented by the set of points in space and time recorded by their mobility device:

DEFINITION 1 (USER HISTORY). *The user history is defined as an ordered sequence of spatio-temporal points* $H = \langle p_1 \dots p_n \rangle$ *where* $p_i = (x, y, t)$ *and* x, y *are spatial coordinates and* t *is an absolute timepoint.*

This continuous stream of information contains different trips made by the user, therefore in order to distinguish between them we need to detect when a user stops for a while in a place. This point in the stream will correspond to the end of a trip and the beginning of the next one. In literature there are two main approaches: clustering-based [3] and heuristic-based [14]. In this paper we adopt the latter for computational efficiency reasons. Thus we look for points that change only in time; i.e. they keep the same spatial position for a certain amount of time quantified by the temporal threshold $th^{stop}_{temporal}$. Specularly, a spatial threshold $th^{stop}_{spatial}$ is used to remove both the noise introduced by the imprecision of the device and the small movements that are of no interest for a particular analysis.

DEFINITION 2 (POTENTIAL STOPS). *Given the history* H *of a user and the thresholds* $th^{stop}_{spatial}$ *and* $th^{stop}_{temporal}$, *a* potential stop *is defined as a maximal subsequence* S *of the user's history* H *where the points remain within a spatial area for a certain period of time:* $S = \langle p_m \dots p_k \rangle | 0 < m \leq k \leq n \wedge \forall_{m \leq i \leq k} Dist(p_m, p_i) \leq th^{stop}_{spatial} \wedge Dur(p_m, p_k) \geq th^{stop}_{temporal}$.

where *Dist* is the Euclidean distance function defined between the spatial coordinates of the points, and *Dur* is the difference in the temporal coordinates of the points. Potential stops can overlap with each other (yet, none of them can completely contain the other, for the maximality condition), making it difficult to use them as a basis for further analysis. In order to avoid this, a criterion of *early selection* is adopted to remove any overlaps:

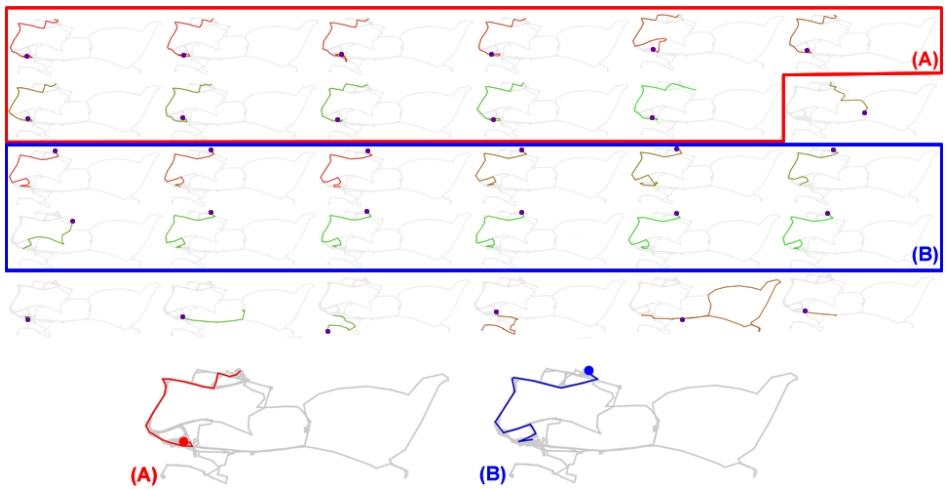

Figure 2: Trajectories of a user and the corresponding groups and routines extracted (A and B). Of the 30 trips, 11 are part of group A, and 12 of group B, while the remaining 7 are noise. The two routines are spatially similar, yet move in opposite directions (points represent the end of trips), i.e., south (A) vs. north (B).

DEFINITION 3 (ACTUAL STOPS). *Given a sequence of potential stops* $S_{set} = \langle S_1, \ldots, S_N \rangle$, *sorted by starting time (i.e.,* $S \leq S' \Leftrightarrow S = \langle (x, y, t), \ldots \rangle \wedge S' = \langle (x', y', t'), \ldots \rangle \wedge t \leq t'$*), the corresponding sequence of* actual stops *ActS is defined as the minimal sequence of potential stops such that:*

1. $S_1 \in ActS$
2. *if* $S_i \in ActS \wedge k = \min\{j | j > i \wedge S_j \cap S_i = \emptyset\} < \infty \Rightarrow S_k \in ActS$

We indicate with $\overline{S} = \langle S_1 \ldots S_t \rangle$ the set of all actual stops over H. Once we have found the stops in the users history we can identify the trips:

DEFINITION 4 (TRIP). *A trip is defined as a subsequence* T *of the user's history* H *between two consecutive actual stops in the ordered set* $\overline{S}$ *or between an actual stop and the first/last point of* H *(i.e.,* p_1 *or* p_n*):*

- $T = \langle p_m, \ldots, p_k \rangle \, | 0 < m \leq k \leq n \wedge \exists i (S_i = \langle \ldots, p_m \rangle \wedge S_{i+1} = \langle p_k, \ldots \rangle)$, *or*
- $T = \langle p_1, \ldots, p_m \rangle \, | 0 < m \leq n \wedge \exists i (S_i = \langle p_m, \ldots \rangle)$, *or*
- $T = \langle p_k, \ldots, p_n \rangle \, | 0 < k \leq n \wedge \exists i (S_i = \langle \ldots, p_k \rangle)$.

The set of extracted trips $\overline{T} = \langle T_1 \ldots T_c \rangle$ in Fig. 1(a), are the basic steps to create the user mobility profile. Notice that the thresholds $th^{stop}_{spatial}$ and $th^{stop}_{temporal}$ are the knobs for expressing specific analytical requirements.

Trip groups. Our objective is to use the set of trips of an individual user to find his/her routine behaviors. We do this by grouping together similar trips based on concepts of spatial distance and temporal alignment, with corresponding thresholds for both the spatial and temporal components of the trips. In order to be defined as *routine*, a behavior needs to be supported by a significant number of similar trips. The above ideas are formalized as follows:

DEFINITION 5 (TRIP GROUP). *Given a set of trips* $\overline{T}$, *spatial and temporal thresholds* $th^{group}_{spatial}$ *and* $th^{group}_{temporal}$, *a spatial distance function* $\delta : \overline{T}^2 \rightarrow \mathcal{R}$ *and a temporal alignment constraint* $\alpha : \overline{T}^2 \times \mathcal{R} \rightarrow \mathcal{B}$ *between pairs of trips, and a minimum support threshold* $th^{group}_{support}$, *a* trip group *for* $\overline{T}$ *is defined as a subset of trips* $g \subseteq \overline{T}$ *such that:*

1. $\forall t_1, t_2 \in g . \delta(t_1, t_2) \leq th^{group}_{spatial} \wedge \alpha(t_1, t_2, th^{group}_{temporal})$;
2. $|g| \geq th^{group}_{support}$.

Condition 1 requires that the trips in a group are approximately co-located, both in space and time, while condition 2 requires that the group is sufficiently large. Again, the thresholds are the knobs that the analyst will progressively tune the extraction process with.

Mobility Profile. Each group obtained in the previous step represents the typical mobility habit of a user, i.e., one of his/her routine movements. Here we summarize the whole group by choosing the central element of such a group:

DEFINITION 6 (ROUTINE). *Given a trip group* g *and the distance function* δ *used to compute it, its* routine *is defined as the medoid of the set, i.e.:*

$$routine(g, \delta) = \arg\min_{t \in g} \sum_{t' \in g \setminus \{t\}} \delta(t, t')$$

Notice that the temporal alignment is always satisfied over each pair of trips in a group, therefore the alignment relation α does not appear in the definition. Now we are ready to define the users mobility profile.

DEFINITION 7 (MOBILITY PROFILE). *Given a set of trip groups* G *of a user and the distance function* δ *used to compute them, the user's* mobility profile *is defined as his/her corresponding set of routines:*

$$profile(G, \delta) = \{ routine(g, \delta) \mid g \in G \}$$

Algorithm 1: *Mobility profile construction*

INPUT:

User's Observations	D
Spatial distance measure	δ
Temporal alignment relation	α
Set of Thresholds	$th^{stop}_{spatial}$, $th^{stop}_{temporal}$, $th^{group}_{spatial}$, $th^{group}_{temporal}$, $th^{group}_{support}$

OUTPUT:

Mobility profile of the user	P

```
BEGIN
H = OrderByTime(D);
T = BuildTrips(H, th^stop_spatial, th^stop_temporal);
C = SelectGroups(T, δ, α, th^group_spatial, th^group_temporal, th^group_support);
P = ∅ ;
FOR EACH c ∈ C DO
    IF size(c) > th^group_support THEN
        P = P ∪ {routine(c,δ)} ;
END;
```

Mobility profile construction. The whole mobility profile extraction – from the initial user history to the final mobility profiles – is summarized in Algorithm 1. The definitions provided in the previous section were kept generic w.r.t. the distance function δ. Different choices can satisfy different needs, possibly both conceptually (which criteria define a good group/routine assignment) and pragmatically (for instance, simpler criteria might be preferred for the sake of scalability). Obviously, the results obtained by different instantiations can vary greatly. Hence w.r.t. Algorithm 1 the crucial point is the *SelectGroup* procedure. Our proposal is to use a clustering method to carry out this task. We choose the clustering algorithm for trajectories proposed in [2], consisting of two steps. First, a density-based clustering is performed, thus removing noisy elements and producing dense – yet, possibly extensive – clusters. Secondly, each cluster is split through a bisection k-medoid procedure. Such method splits the dataset into two parts through k-medoid (a variant of k-means) with $k = 2$, then the same splitting process is recursively applied to each sub-group. Recursion stops when each resulting sub-cluster is compact enough to fit within a distance threshold of its medoid, by removing sub-clusters that are too small. The bisection k-medoid procedure guarantees that requirements 1 and 2 of Definition 5 are satisfied. The clustering method adopted is parametric w.r.t. a repertoire of similarity functions, that includes: *Ends* and *Starts* functions, comparing trajectories by considering only their last (respecively, first) points; *Route similarity*, comparing the paths followed by trajectories from a purely spatial viewpoint (time is not considered); *Synchronized route similarity*, similar to Route similarity but considering also time.

3.2 Profiling GPS-equipped vehicles

In this section we present the results of our method applied to a real dataset of GPS observations of 2,107 real car users in Tuscany in a time period of 12 days covering different kind of territories such as urban and suburban areas. This is a sample of data obtained by a private company employed specifically as a service for insurance companies and other clients called *octotelematics* [1]. The process is implemented using the data mining query language provided by the M-Atlas system [13]. We processed this dataset of observations using the *mobility profile construction* algorithm, with the following parameters:

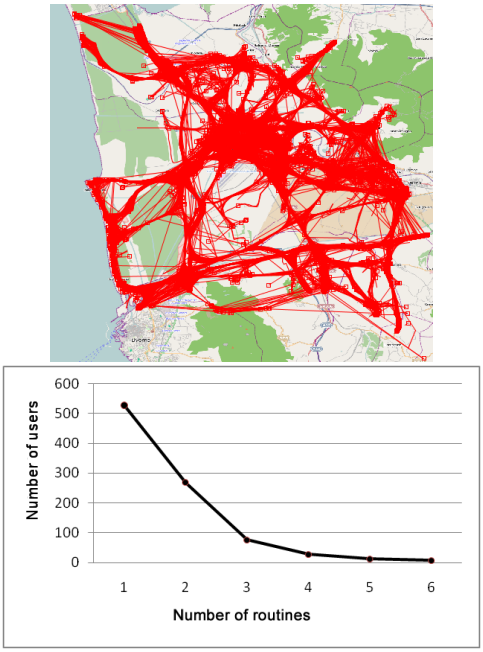

Figure 3: The set of trips extracted from the user observations (top) and the distribution of routines by the users (bottom).

δ **and** α : we adopted the *route similarity* function described in [2] as spatial distance function (δ). The route similarity function performs an alignment between points of the trajectories (trips) that are going to be compared, and then computes the sum of distances between corresponding points. In addition, we adopted a temporal alignment constraint (α) which simply computes the temporal distance between the starting points of the two trips, and compares it against the temporal threshold.

$th^{stop}_{spatial}$ **and** $th^{stop}_{temporal}$: 50 meters and 1 hour, this means that we consider a stop when a user stays with his/her car in an area of 50 m^2 for at least one hour. Single trips of a user are thus the movements between these stops.

$th^{group}_{spatial}$ **and** $th^{group}_{temporal}$: 250 meters and 1 hour, we want to group trips which are *similar* considering a maximum of 250 meters and a temporal alignment of 1 hour.

$th^{group}_{support}$: 4 trips, only the groups with at least 4 trips survive the pruning process, the others are not considered interesting enough for the mobility profiles.

An example of how the *mobility profile construction* works is shown in Fig.2. As can be seen, two main routes are frequently repeated, each time with small variations. In addition, they appear to represent symmetric trips, such as home-to-work and work-to-home routine movements. The

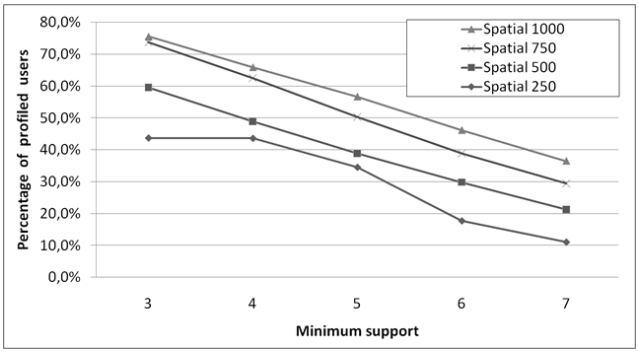

Figure 4: Different settings of parameters lead to a different mobility profile extraction.

corresponding mobility profiles are depicted at the bottom of the figure. Notice that seven user trips were occasional trips that did not fit any consistent habit, and therefore were (correctly) filtered out by our algorithm.

Globally during the execution of the algorithm, a set of 46,163 trips is generated (Fig.3(top)) and the result of the mobility profile construction is a set of 1,504 routines that form 919 mobility profiles (i.e., for 43.6% of the 2,107 users a profile was extracted). Figure 3 (bottom) shows the distribution of the number of routines per user, with almost every user having one or two routines, which usually correspond to the commute to (and from) work (not always at the same time).

To understand how the process is affected by the parameters, using different configurations we analyzed the percentage of users with a mobility profile. The results are shown in Fig.4. They confirm that by using loose constraints, a profile can be built for almost 77% of users. Such percentage decreases to 12% when a strict set of constraints is applied. When looking at the results obtained with a low spatial constraint, we must consider that the clustering method groups together fewer trips and thus pruning using the support threshold becomes more effective. Finally the temporal threshold for the *mobility profile construction* does not seem to have much influence. In fact we discovered that the results significantly change only with a very high threshold. This is simply because we mixed together trips in different periods of the day, thus for clarity's sake, we only show the threshold equal to one hour.

The purpose of the following analysis is to show to what degree the mobility profiles remain persistent and stable considering two time windows: *days 1-6* and *days 7-12*. Figure 5(top) presents the number of profiles extracted in the two time windows and the number of common profiles. It is important to note that the percentage of users profiled in both periods is around 50%, which means that their behavior is *persistent* in time. However although they are profiled in both periods, we want to understand if their routines are similar considering the same spatial threshold used for the grouping phase. In other words, if a routine extracted in the first period remains in the same hypothetical group. The results in Fig.5(bottom) show that for the same experiment we have a percentage of 74% of routines remaining stable.

The thresholds used in the rest of the paper correspond to the ones presented at the beginning of this section. There are two reasons why we chose this configuration: (i) from Fig.4

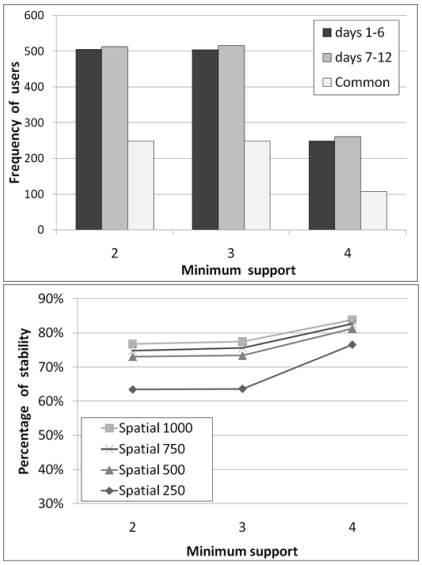

Figure 5: The process is performed on two different time windows and then compared to show how many mobility profiles are persistent (top) and stable (bottom).

it seems that the process reaches a critical point before the decay in performance, and (ii) we want to maintain a critical point of view in order to discover the real capability of the process.

4. MOBILITY PROFILE MATCHING

In this paper we focus on a *car pooling* application aimed at identifying pairs of users that could most likely share their vehicle for one or more of their routine trips. The service might be deployed as a system that provides pro-active suggestions to facilitate the matching process, without the need for the user to explicitly describe (and update) the trips of interest. The starting point of this analysis is the set of representative trips which make up the user mobility profiles. These mobility profiles represent their different typical behaviors, and by comparing them, we can understand if a user can be *served* by another user.

DEFINITION 8 (ROUTINE CONTAINMENT). *Given two mobility routines* $T_1 = \langle p_1^1 \dots p_n^1 \rangle$ *and* $T_2 = \langle p_1^2 \dots p_m^2 \rangle$, *and thresholds* $th_{distance}^{walking}$ *and* $th_{time}^{wasting}$, *we say that* T_1 is contained in T_2, *denoted contained*$(T_1, T_2, th_{distance}^{walking}, th_{time}^{wasting})$

$$\textit{iff: } contained(T_1, T_2, th_{distance}^{walking}, th_{time}^{wasting}) \equiv \exists i, j \in \mathcal{N} \mid 0 < i \leq j \leq m \wedge Dist(p_1^1, p_i^2) + Dist(p_n^1, p_j^2) \leq th_{distance}^{walking} \wedge Dur(p_1^1, p_i^2) + Dur(p_n^1, p_j^2) \leq th_{time}^{wasting}$$

Thresholds $th_{distance}^{walking}$ and $th_{time}^{wasting}$ represent the total spatial and temporal distances allowed between the two routines in space and time, in other words:

$th_{distance}^{walking}$: represents the maximum distance the user which *is served* could walk to reach the meeting point and then to reach their final destination at the end of the trip.

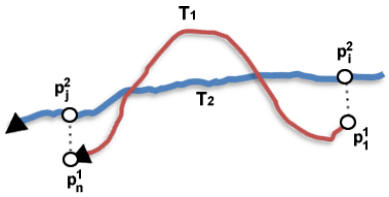

Figure 6: Example of routine containment test for Definition 8: the start point and end point of T_1 are considered and matched against their corresponding nearest points in T_2.

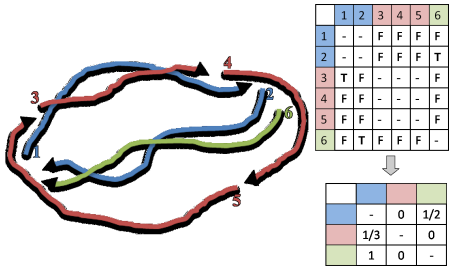

Figure 7: Example of the mobility profile matching process: the routines of the same color belong to the same mobility profile. On the right the matrix of containment between routines (top) and the matrix of the profile share-ability (bottom) with instantiated values.

$th_{time}^{wasting}$: represents the maximum delay the user which *is served* allows, considering the departure and the arrival time.

It is important to note that the *contains* relation is not reflexive because one trip can include the other but not vice versa. This is a basic requirement in the car pooling application because the destinations of the user which *serves* the other can be very far from the destination of the one who is *served* (Fig.6). Extending the definition to the mobility profiles of the users, we can compute the share-ability level of each pair of users:

Definition 9 (Mobility profile share-ability). *Given two mobility profiles $\tilde{T}_1$ and $\tilde{T}_2$, and thresholds $th_{distance}^{walking}$ and $th_{time}^{wasting}$, the* Mobility profile share-ability *measure between $\tilde{T}_1$ and $\tilde{T}_2$ is defined as the fraction of routines in $\tilde{T}_1$ which are contained in at least one routine in $\tilde{T}_2$:*

$$profileShare(\tilde{T}_1, \tilde{T}_2, th_{distance}^{walking}, th_{time}^{wasting}) = \frac{\left| \left\{ p \in \tilde{T}_1 \mid \exists q \in \tilde{T}_2 . Share(p, q, th_{distance}^{walking}, th_{time}^{wasting}) \right\} \right|}{\left| \tilde{T}_1 \right|}$$

By applying this definition to all possible pairs of users (i.e., to their corresponding profiles) we can build a matrix of share-ability, thus expressing how good the match of each pair is. The process is presented in Algorithm 2, where first a *routine containment matrix* is built over single mobility routines, then the results corresponding to each pair of users are collapsed to form a mobility profile share-ability matrix, by applying the Definition 9. In the algorithm, $\hat{P}$ represents the set of all routines of all users, while function $get\_user(p)$ returns the user that owns routine p (identical routines generated by different users are distinguished, here). A visual example of the result is shown in Fig.7.

Algorithm 2: *Matching matrices*

INPUT:
Set of Users U
Users' routines $\hat{P}$
Set of Thresholds $th_{distance}^{walking}, th_{time}^{wasting}$
OUTPUT:
Routine containment matrix $M : \hat{P}^2 \rightarrow Bool$
Mobility profile share-ability matrix $C : U^2 \rightarrow Real$

```
BEGIN
FOR EACH (p, q) ∈ P̂² DO
    IF get_user(p) ≠ get_user(q) THEN
        M(p, q) := contains(p, q, th_distance^walking, th_time^wasting) ;
FOR EACH (u, v) ∈ U² DO
    IF u ≠ v THEN
        ∀x ∈ {u, v}.P̂_x := {p ∈ P̂ | get_user(p) = x};
        C(u, v) = profileShare(P̂_u, P̂_v, th_distance^walking, th_time^wasting));
END;
```

4.1 The car pooling service with GPS data

We used Algorithm 2 to perform the matching process on our data with different parameter settings. The results in Figure 9 show how the performances are affected, in terms of percentage routines and mobility profiles that have at least one match. Note that by allowing a *walking distance* of 5 km and a *wasting time* of 1 hour, 89% of profiled users have (at least) one match, which decreases to 66% if the *wasting time* becomes half an hour. Figure 8 shows two examples of matching between two users. The red user can be served by the violet user on the basis of the routines shown. In the two examples it is interesting to see that in the first case (A), the starts and ends of the routines are quite close, therefore these users can both serve or be served by each other; in the second case (B) the relation is unidirectional, since the red routine ends much earlier than the other, and therefore the *contain* relation does not hold in the opposite direction.

Considering a hypothetical car pooling service built on top of the proposed method, using a *walking distance* of 2.5 km and a *wasting time* of 1 hour, we can calculate some statistics regarding the potential impact of the service. In fact 684 users, corresponding to 32.4% of participants, receive at least one indication of a possible host for one of their routines. This means that if everybody takes the opportunity of sharing a / their car using this system, traffic could be decreased significantly. As previously mentioned, one advantage of the system is that users do not need to manually declare their common trips (indeed, routines are automatically detected), which is a major flaw of current car pooling systems, and probably contributes substantially to their failure. As shown in section 3.2, the system can keep reasonably up-to-date routines and profiles by executing the profiling process once every two weeks (or more), using a temporal sliding window on the data.

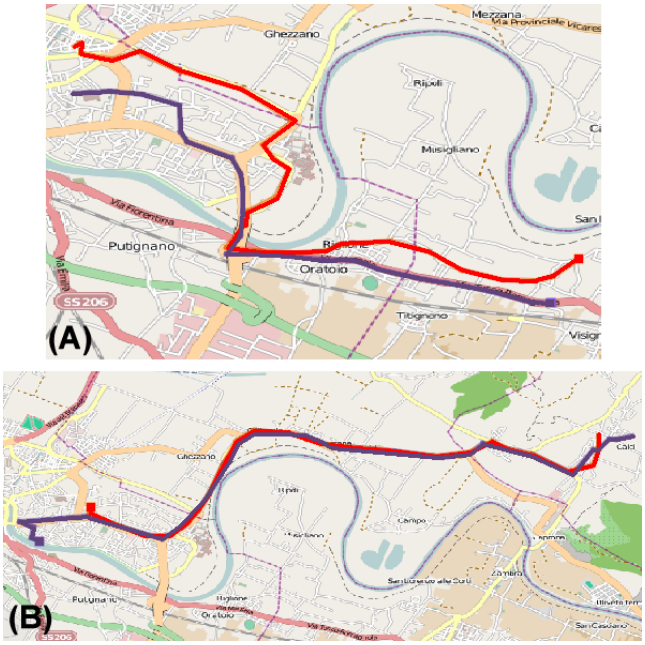

Figure 8: Examples of routine containment: red routines are contained in the violet ones.

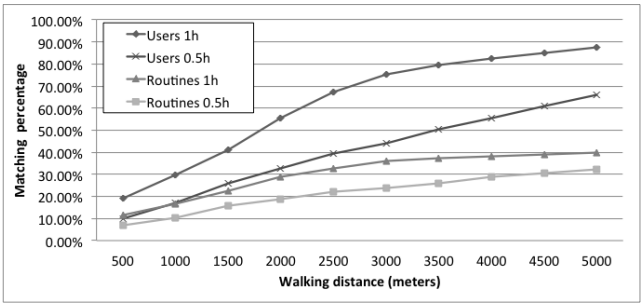

Figure 9: Matching percentages of users (upper curves) and routines (lower curves) for different settings of the spatial and temporal thresholds.

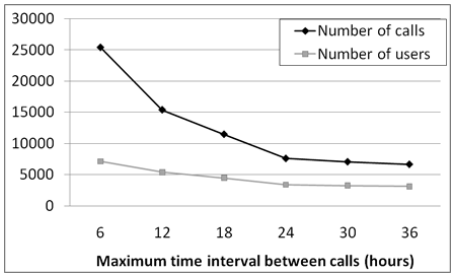

Figure 10: Number of users and number of calls in the Network-based simulation

5. DOWNGRADING DATA PRECISION

User profiles can also be extracted and compared on the basis of mobile phone network traces that are commonly (and massively) available from telecom operators. It is therefore natural to wonder whether the loss in spatial and temporal precision and completeness that characterizes this kind of data can compromise the current usability of our methodology. The general methodology introduced so far in this paper can be directly applied to this kind of data and enables us to perform exactly the same kind of analysis on both GPS and GSM data and compare the results. In this section we discuss three issues, how to: (i) set up a test to compare profile extractions over different spatio temporal granularities of the same dataset; (ii) instantiate the profile extraction method w.r.t. different spatio temporal granularities; and (iii) compare and interpret the profiles of the same user extracted at different spatio temporal granularities.

5.1 GSM data simulation

In the context of GSM technology there are two different forms of data collection: in the first, named *Handset-based*, the mobile device collects the history of its positions as the sequence of the towers that it traversed. In the second, named *Network-based*, the network collects the sequence of towers that serve a calling device.

Handset-based simulation. In order to simulate a *Handset-based* localization system we apply a transformation on GPS data named *GPS generalization* defined as follows:

DEFINITION 10 (GSM GENERALIZATION). *Given a user GPS history $H = \langle p_1, \dots p_n, \rangle$ and a set of GSM cells C, let functions $tower : C \to \mathcal{R}^2$ and $cover : C \to 2^{\mathcal{R}}$ denote, for each cell, the position of its tower and its spatial coverage, respectively. Then, the GSM generalization of H is defined as the sequence $H_{GSM} = \langle g_1, \dots g_n \rangle$ such that $\forall_{0<i\leq n}.g_i = (x^*, y^*, t) \wedge p_i = (x, y, t) \in H \wedge \exists c \in C.tower(c) = (x^*, y^*) \wedge (x, y) \in cover(c)$.*

The result is a sequence of points with timestamps, each point represents the tower (and therefore the cell it covers) that captures a corresponding point in the GPS traces of the original dataset.

Network-based simulation. The particular feature of the *C*DR data is the fact that they are collected only when the user is calling. Therefore to simulate such data starting from the GPS traces, we need to insert holes corresponding to the missing data due to the not-calling status of the user. We therefore use two different probability functions:

- Pr_{call}: To determine when a user starts a call, we use an exponential distribution specifying the maximum time interval Int in which the user performs at least one call.

- $Pr_{duration}$: In [4], the authors present a study of the distribution of durations, which determines a way to define the probability of a call duration. They define a function called TLAC which simulates over 96% of the duration of a user's call. We use this function in order to simulate when a call terminates.

With the definition of probabilities Pr_{call} and $Pr_{duration}$, we can use different values for the maximum time interval Int to obtain different scenarios. In the following, we denote each of these datasets as H^{Int}. Figure 10 shows some basic statistics of six different datasets obtained with a time interval $Int \in \{6; 12; 18; 24; 30; 36\ hours\}$.

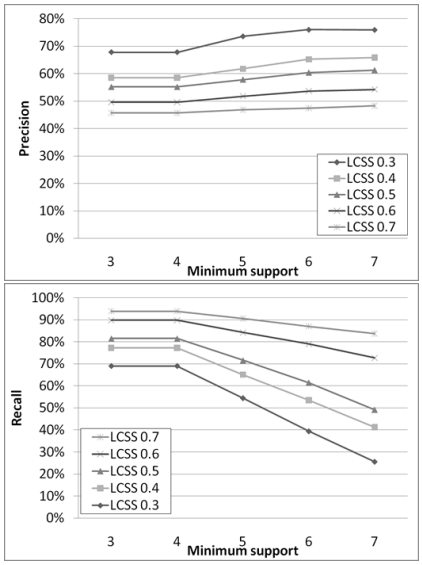

Figure 11: *Precision* and *recall* of the profile construction in GSM_H data

5.2 Profiling users with GSM data

The particular features of the GSM context require only two small changes w.r.t. the GPS in order to adapt the method: the first is related to the spatial granularity of stops to construct the trips and the second is the similarity function adopted.

- In computing stops and trips we set the spatial threshold $th^{stop}_{spatial} = 0$. This is because each location in the history of the user already represents an area, namely the coverage area of the tower serving the device.
- The route similarity adopted in the case of GPS traces is replaced with the *Longest Common Sub-Sequence* (LCSS) distance. We adopt this distance function in order to measure the longest sequence of towers crossed by two users.

To summarize the effect of spatial downgrading we used the following elements:

GPS is the set of profiles extracted with the GPS profile extraction method

GSM_H is the set of profiles extracted with the GSM-profile extraction method on the handset-based data.

GSM_{CDR} is the set of profiles extracted with the GSM-profile extraction method on the network-based data.

We use *precision* and *recall* measures to compare the quality of GPS versus GSM_H profiles and GPS versus GSM_{CDR} profiles. The idea, in both cases, is to use the GPS profiles as a ground truth against which to verify the GSM profiles.

Handset-based. Fig. 11(top) shows the *precision* of the GSM_H profiles considering different support values and varying the minimum $LCSS$ distance. Note that a greater value of the minimum support threshold corresponds to an increase in *precision*. In addition, using a smaller minimum distance leads to a more precise set of mobility profiles. This means that the mobility profiles extracted from

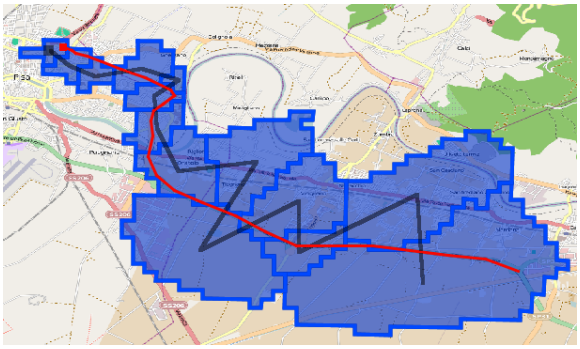

Figure 12: The red line represents the GPS routine, the blue areas connected through the black line indicate the GSM routine.

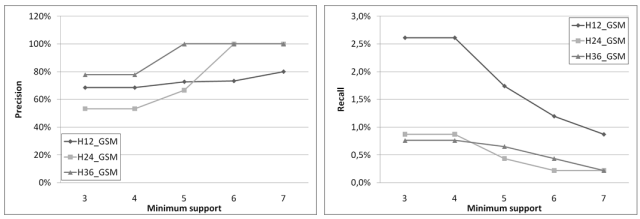

Figure 13: *Precision* and *recall* of the profile construction in GSM_{CDR} data considering a time interval $Int \in 12, 24, 36$

GSM_H data correspond, in most cases, to the profiles obtained with the GPS data. The *recall* measure is depicted in Fig. 11(bottom), where it can be seen that with a greater support threshold, the *recall* measure decreases. Furthermore, the decreasing behavior is enforced using a more restrictive distance value. It is worth highlighting that for some configurations, we are able to obtain a precision close to 60% and a *recall* greater than 70%. Figure 12 shows the results of selecting the same user in the GPS and GSM_H datasets and performing the profile construction in both cases, where the precise correspondence between the GPS routine and the GSM routine is clear.

Network-based. The effect of the loss of spatial precision and temporal completeness becomes more remarkable considering GSM_{CDR} data. In fact, on this kind of data the process is only able to build a profile for few users. This is clearly reflected in Fig.13 where we obtain high values of *precision* but a very low level of *recall* for all configurations of the input parameters. This is because the process constructs few mobility profiles on this data, which often correspond to the profiles contained on the GPS ones.

5.3 The car pooling service with GSM data

In this case the spatial distance constraint is included in the spatial approximation introduced with the GSM generalization of each point. Therefore, the matching between two users is computed by means of the same $Contains$ and $ProfileShare$ functions defined in Section 4, setting the threshold $th^{walking}_{distance} = 0$. To investigate the impact of the downgrading process on the mobility profile matching, again we study the trend of *precision* and *recall* measures w.r.t. the mobility profile share-ability matrix discussed in Sec. 4.1.

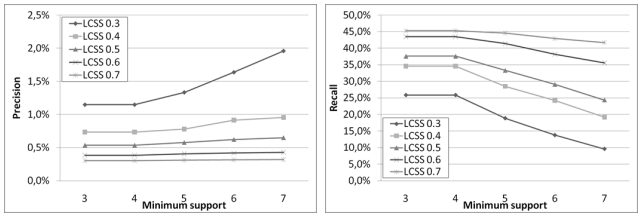

Figure 14: *Precision* and *recall* of the matching in GSM_H data

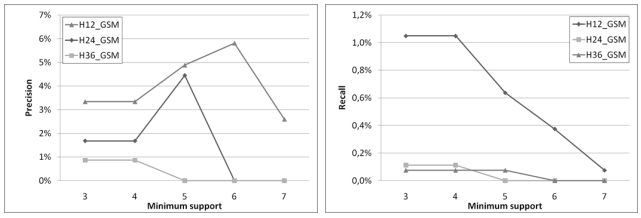

Figure 15: *Precision* and *recall* of the matching in GSM_H data considering a time interval $Int \in 12, 24, 36$

The influence of the spatial degradation introduced with the GSM generalization is substantial as shown in Fig. 14 where the *precision* reaches low values and the *recall* measure shows a decreasing trend but maintains considerable stability. In fact, the spatial generalization enables the GSM_H mobility profile matching process to verify the share relation between a pair of users in several more cases, by producing a bigger share-ability matrix w.r.t. the one obtained with the GPS data. Concerning an evaluation of the matching process on GSM_{CDR} data, shown in Fig.15, we obtain low results both for *precision* and *recall*. This is due to the fact that the generalization data does not allow any matching between a pair of mobility profile users.

6. CONCLUSIONS

In this paper we proposed an analysis aimed at discovering and matching mobility profiles, then we showed as it may foster an intelligent car pooling service. We applied the proposed analytical framework to two different scenarios: GPS vehicular data and (simulated) GSM mobile phone data.

Future developments include a wide empirical evaluation of the persistence of user profiles, as well as their tolerance to transient changes in mobility habits (caused by road works, extraordinary events, etc.) and their reaction times to steady changes. Also, alternative profile matching schema will be explored, in order to deal with the loss of performance with less rich forms of data, and a method to automatically suggest the parameters in all the steps involved in the process will be investigated to increase the applicability in real scenarios. Finally, more complex and realistic simulations of GSM and CDR data might be considered, moving from the static tower assignment used in this work, based on shortest distance, to a dynamic one that better models the non-deterministic and noisy nature of actual assignments.

7. REFERENCES

[1] *Octotelematics*. http://www.octotelematics.com/.

[2] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. *Interactive Visual Clustering of Large Collections of Trajectories*. VAST: Symposium on Visual Analytics Science and Technology, 2009.

[3] V. Bogorny, C. A. Heuser, and L. O. Alvares. A conceptual data model for trajectory data mining. In *GIScience*, pages 1–15, 2010.

[4] P. O. V. de Melo, L. Akoglu, C. Faloutsos, and A. A. Loureiro. *Surprising Patterns for the Call Duration Distribution of Mobile Phone Users*. ECML PKDD: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2010.

[5] S. Gaffney and P. Smyth. Trajectory clustering with mixture of regression models. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 63–72. ACM, 1999.

[6] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339, 2007.

[7] F. Giannotti and D. Pedreschi, editors. *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery*. Springer, 2008.

[8] M. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[9] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *Proceedings of 9th International Symposium on Spatial and Temporal Databases (SSTD'05)*, pages 364–381. Springer, 2005.

[10] N. Pelekis, I. Kopanakis, I. Ntoutsi, G. Marketos, and Y. Theodoridis. Mining trajectory databases via a suite of distance operators. In *ICDE Workshops*, pages 575–584, 2007.

[11] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 7:713–, 2010.

[12] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

[13] R. Trasarti, F. Giannotti, M. Nanni, D. Pedreschi, and C. Renso. *A Query Language for Mobility Data Mining*. IJDWM: International Journal of Data Warehousing and Mining., 2010.

[14] X. Xiao, Y. Zheng, Q. Luo, and X. Xie. Finding similar users using category-based location history. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[15] H. Yoon, Y. Zheng, X. Xie, and W. Woo. Smart itinerary recommendation based on user-generated gps trajectories. In *Proceedings of the 7th international conference on Ubiquitous intelligence and computing*, 2010.