# Individual Mobility Profiles: Methods and Application on Vehicle Sharing*

Roberto Trasarti[1], Fabio Pinelli[2], Mirco Nanni[1], and Fosca Giannotti[1]

[1] KDDLab, ISTI-CNR, Pisa, Italy: email name.surname@isti.cnr.it
[2] IBM Research Lab, Dublin, Ireland: fabiopin@ie.ibm.com

**Abstract.** In this paper we present 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 we present an experiment which is performed on a massive dataset containing GPS traces of private cars.

## 1 Introduction

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. 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. The standard approach adapts classical distance-based algorithms and defines ad hoc distances for trajectory data [7], possibly with limited ad hoc refinements [3] or ad hoc solutions include variants of model-based clustering [4], collective movements detection methods [6], and others. As opposed to existing solutions, in our proposal, already published in [1], the evaluation of similarity between individuals is not realized as a direct comparison of trajectories. Instead,we propose 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.

## 2 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,

* Extended Abstract

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

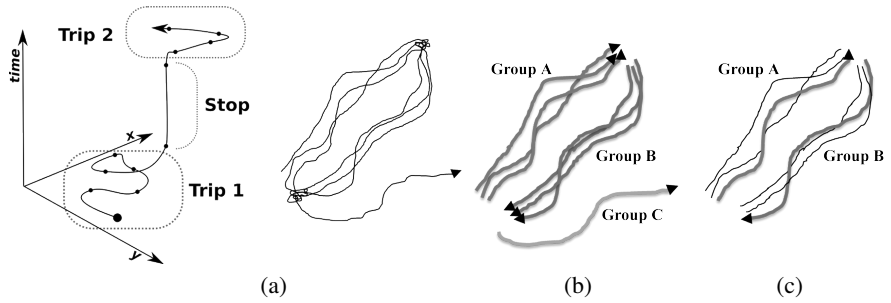



**Fig. 1.** Mobility profile extraction process: (a) trip identification; (b) group detection/outlier removal; (c) selection of representative mobility profiles.

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:

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)).

### 2.1 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 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:* $S = \langle p_m \dots p_k \rangle \,|0 < m \leq k \leq n \wedge \forall_{m \leq i \leq k} Dist(p_m, p_i) \leq th^{stop}_{spatial} \wedge Dur(p_m, p_k) \geq th^{stop}_{temporal}$.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

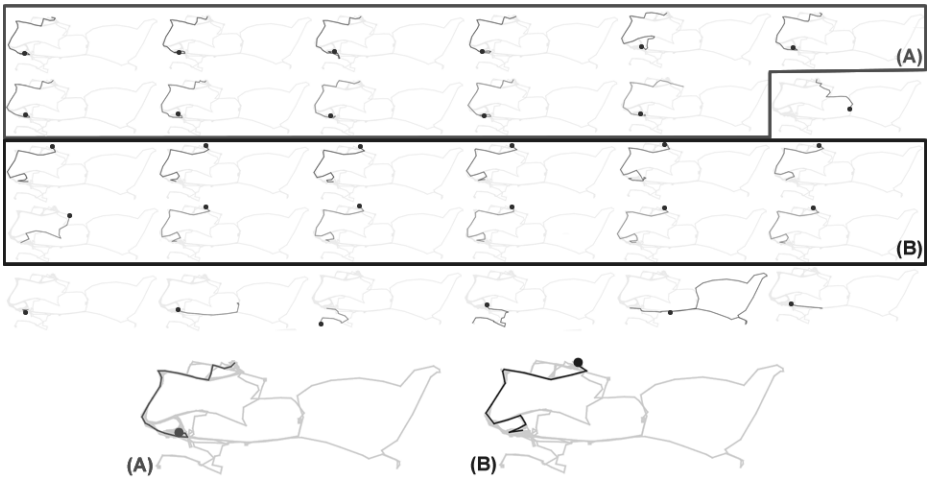



**Fig. 2.** Trajectories of a user and the corresponding groups and routines extracted (A and B). Of the 30 trips, 11 are part of group A, and 12 of group B, while the remaining 7 are noise. The two routines are spatially similar, yet move in opposite directions (points represent the end of trips), i.e.,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

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:

**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* $S_i \in ActS \wedge k = \min\{j | j > i \wedge S_j \cap S_i = \emptyset\} < \infty \quad \Rightarrow \quad S_k \in ActS$

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*
- $T = \langle p_k, \ldots, p_n \rangle \, | 0 < k \leq n \wedge \exists i (S_i = \langle \ldots, p_k \rangle)$.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

The set of extracted trips $\overline{T} = \langle T_1 \dots T_c \rangle$ 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 $\delta : \overline{T}^2 \to \mathcal{R}$, a temporal alignment constraint $\alpha : \overline{T}^2 \times \mathcal{R} \to \mathcal{B}$ 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 $\delta$ 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 $\alpha$ 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 $\delta$ used to compute them, the user's* mobility profile *is defined as his/her corresponding set of routines:*

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

**Mobility profile construction.** The definitions provided in the previous section were kept generic w.r.t. the distance function $\delta$. 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

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

scalability). Obviously, the results obtained by different instantiations can vary greatly. Our proposal is to use a clustering method to carry out this task. We choose the clustering algorithm for trajectories proposed in [3], 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.

### 2.2 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* [2]. The process is implemented using the data mining query language provided by the M-Atlas system [8]. We processed this dataset of observations using the *mobility profile construction* algorithm, with the following parameters:

$\delta$ **and** $\alpha$ : we adopted the *route similarity* function described in [3] as spatial distance function ($\delta$). 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 ($\alpha$) 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 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.

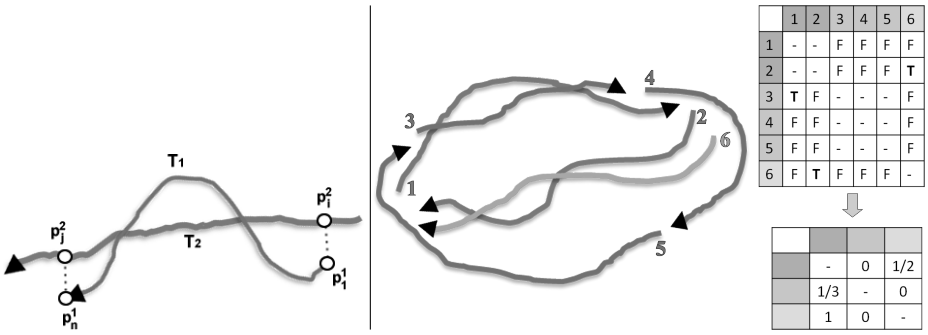
Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

Globally during the execution of the algorithm, a set of 46,163 trips is generated 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). For space reason we don't report the complete analysis presented in [1] on how the method is effected by the parameters and how much the profiles extracted remain persistent and stable in time.

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




| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | - | - | F | F | F | F |
| 2 | - | - | F | F | F | **T** |
| 3 | **T** | F | - | - | - | F |
| 4 | F | F | - | - | - | F |
| 5 | F | F | - | - | - | F |
| 6 | F | **T** | F | F | F | - |

| | | | |
|---|---|---|---|
| | - | 0 | 1/2 |
| | 1/3 | - | 0 |
| | 1 | 0 | - |

**Fig. 3.** On the left: example of routine containment test for Definition 8. On the right: there is an example of the mobility profile matching process where the routines of the same color belong to the same mobility profile: the routines matrix-containment (top) and the profile share-ability matrix (bottom).

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:

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

$th^{walking}_{distance}$ : 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.

$th^{wasting}_{time}$ : 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.3(left)). 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 $th^{walking}_{distance}$ and $th^{wasting}_{time}$, 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 $\tilde{T}_2$:*

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

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 algorithm first builds a *routine containment matrix* over single mobility routines, (ii) then the results corresponding to each pair of users are collapsed to form a mobility profile share-ability matrix, by applying the Definition 9. A visual example of the result is shown in Fig.3(right).

### 3.1 The car pooling service with GPS data

We used Algorithm presented in previous section to perform the matching process on our data with different parameter settings. The results in Figure 4(right) 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 4(left) 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

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

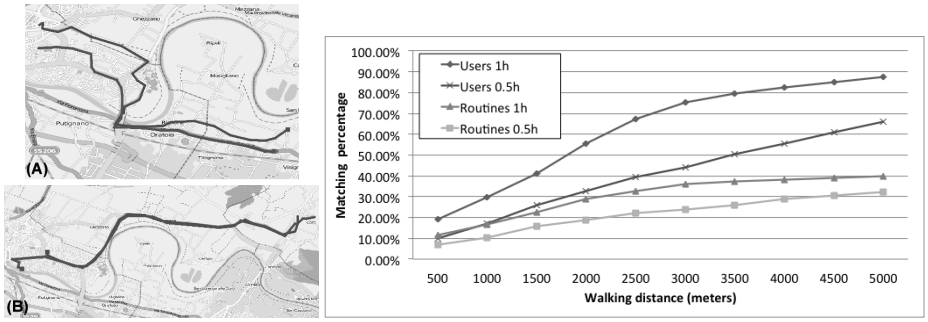



**Fig. 4.** Examples of routine containment: red routines are contained in the violet ones (left). Matching percentages of users (upper curves) and routines (lower curves) for different settings of the spatial and temporal thresholds (right).

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 2.2, 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.

## References

1. R. Trasarti, F. Pinelli, M. Nanni and F. Giannotti *Mining mobility user profiles for car pooling*. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD 2011.
2. *Octotelematics*. http://www.octotelematics.com/.
3. 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.
4. 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.
5. F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339, 2007.
6. 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.
7. 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.
8. F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *Very Large Database*, 20(5), 2011.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 