# Mobility, Data Mining and Privacy: The GeoPKDD Paradigm

GeoPKDD White paper,
KDD Lab

November 23, 2009

**Abstract**

The technologies of mobile communications and ubiquitous computing pervade our society, and wireless networks sense the movement of people and vehicles, generating large volumes of mobility data. Miniaturization, wearability, pervasiveness are producing traces of our mobile activity, with increasing positioning accuracy and semantic richness: Location data from mobile phones (GSM cell positions), GPS tracks from mobile devices receiving geo-positions from satellites, etc. The objective of the GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery) project is to discover useful knowledge about human movement behaviour from mobility data, while preserving the privacy of the people under observation. Pursuing this ambitious objective, the GeoPKDD project has started a new exciting multidisciplinary research area, at the crossroads of mobility, data mining, and privacy. This paper gives a short overview of the envisaged research challenges and the project achievements.

## 1 Introduction

Research on moving-object data analysis 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. These have made available massive repositories of spatio-temporal data recording human mobile activities, that call for suitable analytical methods, capable of enabling the development of innovative, location-aware applications. This is a scenario of great opportunities and risks: on one side, mining this data can produce useful knowledge, supporting sustainable mobility and intelligent transportation systems; on the other side, individual privacy is at risk, as the mobility data contain sensitive personal information. The GeoPKDD project [1], since 2005, investigates how to discover useful knowledge about human movement behaviour from mobility data, while preserving the privacy of the people under observation. GeoPKDD aims at improving decision-making in many mobility-related tasks, especially in metropolitan areas:

- Monitoring and planning traffic and public transportation systems
- Localizing new facilities and public services
- Forecasting/simulating traffic-related phenomena
- Geo-marketing and location-based advertising
- Innovative info-mobility services
- Detecting changes in collective movement behaviour.

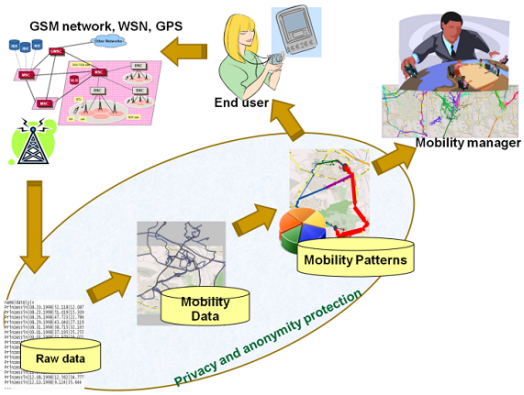


Figure 1: The GeoPKDD process

The very initial questions emerged by the above scenario are the following. How to reconstruct a trajectory from raw logs, how to store and query trajectory data? How to classify trajectories according to means of transportation (pedestrian, private vehicle, public transportation vehicle, )? Which spatio-temporal pattern and models are useful abstractions of mobility data? How to compute such patterns and models efficiently? Privacy protection and anonymity - how to make such concepts formally precise and measurable? How to find an optimal trade-off between privacy protection and quality of the analysis? To answer these questions, the basic assumption of GeoPKDD was that movement data have to be at the centre of an integrated knowledge discovery process capable to support the managing, the querying, the analysis and the interpretation of this form of data and patterns [12].

## 2 Mobility Data Management and Warehousing

A trajectory, the basic form of mobility data, is a sequence of time-stamped locations, sampled from the itinerary of a moving object. A database management system and a warehouse have been designed around this simple form of data which is then enriched by a conceptual model which supports trajectory semantics, based on the concept of stops and moves [2]. The design of the trajectory database has been influenced by the current fluerishent research on Moving Object Databases extend the traditional database technology in to order to handle modeling, indexing and query processing issues for trajectories. In such proposals the spatial and temporal dimensions are considered as first-class citizens and both past and current (as well as anticipated future) positions of moving objects are of interest [3, 4, 5]. The adopted trajectory data base, named Hermes [6] provides efficient means for reconstructing trajectories from raw location data, as well as storing and querying massive trajectory data. The trajectory reconstruction technique transforms sequences of raw sample points into meaningful trajectories according different filters (temporal gaps, spatial gaps, maximum speed, tolerance distance). In Hermes current position of a moving

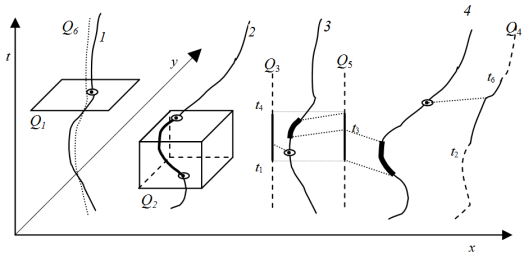


Figure 2: Some spatio-temporal primitives provided by Hermes

object is modelled as a function in time over the starting location, so that arc movement is a basic property. It is implemented on the top of a relational object oriented DBMS (ORACLE Spatial Cartridge) . Hermes provides specila indexing mechanism that support efficent queries on trajectory data such as:

- Spatial (range or NN) search : "Find all trajectories that were inside area A at time instant t (or time interval I)" or "Find the trajectory that was closest to point B at time instant t (or time interval I)"
- Topological / directional search: "Find all trajectories that entered (crossed, left, bypassed, etc.) or were located west (south, etc.) of an area" or "Find all trajectories that crossed (met, etc.) or were located left of (right of, in front of, etc.) a query trajectory TQ
- Most-similar-trajectories: Given a query trajectory TQ, show me the k- most similar trajectories to TQ (perhaps, constrained is space and/or time)

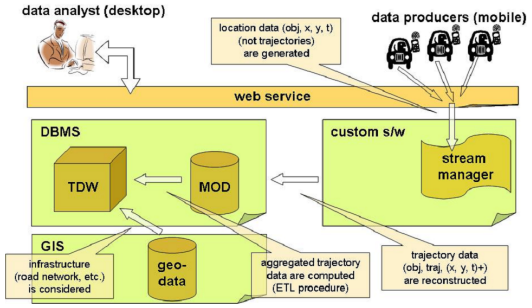


Figure 3: The trajectory database management system and warehouse

The basic analytic framework is also provided: the trajectory warehouse named T-Warehouse. It is a spatio-temporal data cube representing various aggregated measures of the moving objects, such as presence and speed. The T-OLAP engine proposed in GeoPKDD supports a variety of dimensione (temporal, spatial,thematic) and meausres (about space, time end their derivatives) enabling exploratory analysis, drilling up and down the space and time dimensions. Further investigations addressed the definition of adequate aggregate functions for OLAP operations; a new method to compute holistic functions has been developed, such as the aggregate presence measure (number of distinct trajectories in a cell) [7, 8, 9].

Globally, this first reserach lines of GeopKDD delivered a comprehensive trajectory database management system and warehouse, capable of acquiring, transforming, storing, querying and analyzing massive trajectory datasets.

## 3 Mobility Data Mining

While the OLAP analytical tools concentrate mostly on presence of moving objects, the objective, and the novelty of mobility mining research was on inventing methods capable to extract out of trajectory data movement behaviours, namely models of movement within a certain area. Given this assuntion, a method for mobility data mining has two different tasks: first to define the forms of spatio-temporal local patterns and global models to be mined from trajectory data, second, to design and implement efficient algorithms for extracting such patterns. Within GeoPKDD a wide repertoire of patterns, models and privacy-preserving techniques, have been defined and implemented, this section provides a small selection that covers the different mining task such as clustering,frequent patterns and classification.

### 3.1 Trajectory Patterns.

Frequent patterns over trajectories take the form of sequences of locations or movements. We indroduced a novel form of spatio-temporal pattern, which formalizes the idea of aggregate movement behaviour. The trajectory pattern in [11], represents a set of individual trajectories that share the property of visiting the same sequence of places with similar travel times. Therefore, two notions are central: (i) the regions of interest in the given space, and (ii) the typical travel time of moving objects from region to region. In this

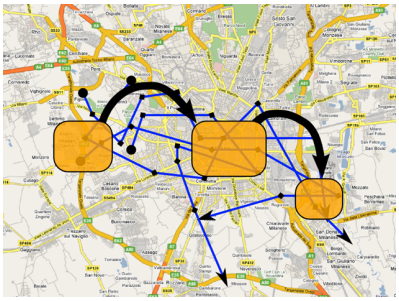

Figure 4: An Example of Trajectory Pattern

approach a trajectory pattern is a sequence of spatial regions that, on the basis of the source trajectory data, emerge as frequently visited in the order specified by the sequence; in addition, the transition between two consecutive regions in such a sequence is annotated with a typical travel time that, again, emerges from the input trajectories. For instance, consider the following two trajectory patterns over regions of interest in the centre of a town:

$$\text{Railway Station} \overset{(15min)}{\rightarrow} \text{Castle Square} \overset{(2h15min)}{\rightarrow} \text{Museum (a)}$$
$$\text{Railway Station} \overset{(10min)}{\rightarrow} \text{Middle Bridge} \overset{(10min)}{\rightarrow} \text{Campus (b)}$$

Here, pattern (a) may be interpreted as a typical behaviour of tourists that rapidly reach a major attraction from the railway station and spend there about two hours before getting to the adjacent museum. Pattern (b), may highlight the pedestrian flow of students that reach the university campus from the station: for them, the central bridge over the river is a compulsory passage. It should be observed that a trajectory pattern does not specify any particular route among two consecutive regions: instead, a instead, typical travel time is specified, which approximates the (similar) travel time of each individual trajectory represented by the pattern. More formally:

DEFINITION 3.1. *A T-pattern, is a pair* $(S, A)$*:* $S =< R_0, \ldots, R_n >$ *is a sequence of locations, and* $A = \alpha_1, \ldots \alpha_n \in R_+^k$ *are the transition times (annotations). A T-pattern is also represented as:*

$$R_0 \xrightarrow{\alpha_1} R_1 \xrightarrow{\alpha_2} \ldots \xrightarrow{\alpha_n} R_n.$$

Essentially, the algorithm computes the frequent movements and then computes their typical travel Time. The compexity of the algorithms stays into the need to dinamically apply discretization strategies in order to find similar positions in space and in time. So the prolem is formulated as density estimation problem driven by the two parameters time and space tolerance .

### 3.2 Trajectory Clustering.

In the case of clustering, the standard approach adopted consists in adapting classical distance-based algorithm by defined ad hoc distances for trajectory data, that take into account the spatial locations visited, their order, and, in some cases, the time. In our case a T-cluster is a set of similar trajectories, according to a repertoire of trajectory similarity functions; thus, a T-cluster reveals a group of objects sharing a common movement behaviour, e.g., home-work-home commuting. In particular, we adopted here the definition introduced in [13] where the concept of density-based clustering is defined over moving objects. Here, a cluster of mov-

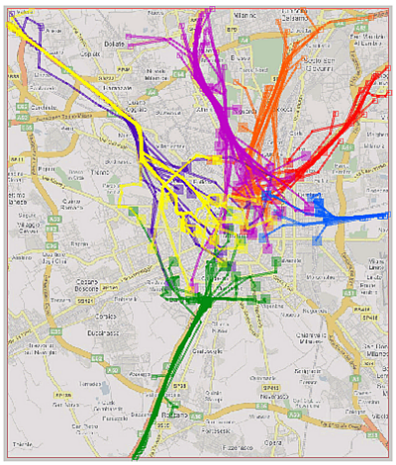

Figure 5: An Example of Clustering

ing objects contains all elements that are density-reachable respect to a tolerance value, given a minimum number of objects which must belong to a cluster. The algorithm presented in [13] computes an augmented cluster-ordering of the database objects. Using this order it builds a reachability distance plot and identifies the clusters using a user distance threshold and a minimum number of points. The algorithms is the basic brick of an interactive tool wich allow the user to progressively refine the search by using different similarity notions.

### 3.3 Trajectory Classification and Location Prediction.

Predictive models for trajectory data include a classification method for inferring the category of a trajectory, (e.g., the transportation means associated to a trajectory (private car, public transportation, pedestrian, etc.), and a predictor of the next location of a moving object given its past trajectory. Next location prediction is an hot topic in the field of

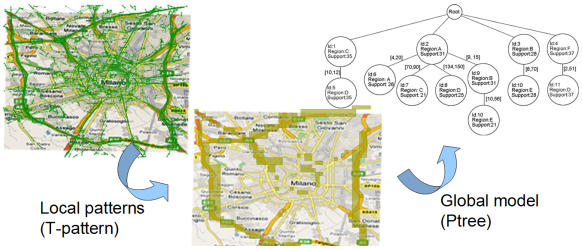


Figure 6: From local to global model as prediction tree

LBSs where learning methods are base on the individual history. The proposed method, named WhereNext[15] pretends to predict the future location of a moving object on the base of previously extracted T-patterns, coherently with the idea that global models can be built out of a collection of local patterns. Using Trajectory Patterns as predictive rules has the following implications: (I) the learning depends on the movement of all available objects in a certain area instead on the individual history of an object; (II) the prediction tree intrinsically contains the spatio-temporal properties emerged from the data and this allows to define matching methods strongly depending on such movements properties. Moreover, a set of different measures, aimed at evaluating a priori the predictive power of a set of Trajectory Patterns, has been proposed and tuned on a real life case study.

### 3.4 Trajectory Anonymity.

The standard approaches developed for protecting privacy on tabular data do not work for spatio-temporal data sets. For example, randomization techniques, which modify

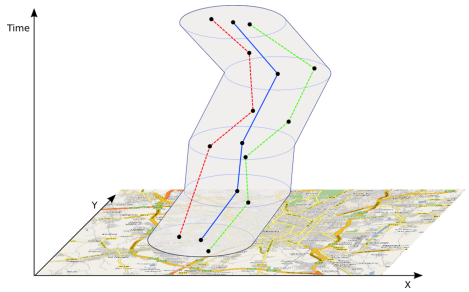


Figure 7: A step of Trajectory anonymity process

a dataset to guarantee respondents' privacy while preserving data utility for analyses, are not applicable on spatio-temporal data, due to their particular nature. Therefore, alternative solutions have been suggested: some of them belong to the category of confusion-based algorithm others belong to the category of approaches of k-anonymity for location position collection. All these techniques try to guarantee location privacy for trajectories.The approaches developed in GeoPKDD belong to the first category and provide confusion/obfuscation algorithm to prevent an attacker from tracking a complete user trajectory. The main idea is to modify true trajectories or generate fake trajectories in order to confuse the attacker. The method named Never Walk Alone [10], proposes a novel concept of k-anonymity based on co-localization that exploits the inherent uncertainty of the moving objects whereabouts. Never Walk Alone, is based on trajectory clustering and spatial translation.. It provides privacy protection by: (1) first enforcing k-anonymity, meaning every released information refers to at least k users/trajectories, (2) then reconstructing randomly a representation of the original dataset from the anonymization. A k-anonymous trajectory dataset is one where the itinerary of each person is indistinguishable from that of other k-1 persons - anonymity viewed as hiding in the crowd. Our T-anonymity methods transform a trajectory dataset into a new, k-anonymous dataset, such that the key analytical properties are preserved, together with users' privacy.

## 4 Mastering the GeoPKDD Process

In order to support the interactive, iterative, combined usage of the various tools to the purpose of discovering mobility knowledge, GeoPKDD developed two prototype platforms: a semantic-based query & reasoning systems, and a visual analytic environment.

### 4.1 Semantic-based query & reasoning system.

This system allows the user to describe the entire knowledge discovery process using a set of primitives [16, 18], based onto a **Data Mining Query Language**. The spatio-temporal query primitives support selection and preprocessing of trajectory data w.r.t. geographic background knowledge, as well as anonymization. The trajectory mining primitives allow extracting and validating mobility patterns and models. A reasoning component allows to specify domain-driven ontologies, inferring types of trajectories and patterns.

### 4.2 Visual Analytics.

The aim of this system is to help the analyst navigate through mobility data and patterns and visually drive the

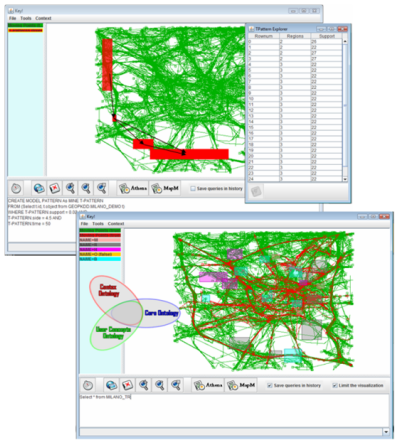

Figure 8: The Semantic-based query & reasoning system

analytical process [17]. Some key features: Visualization of T-patterns to support the navigation of the extracted patterns in the spatial and temporal dimensions. Progressive refinement of T-clusters: A user-driven exploration and evaluation of the discovered T-clusters [14], based on a step-wise iterative method. Visual exploration of the T-Warehouse [8] to browse aggregated measures of moving objects, such as density and speed.

Figure 9: The Visual Analytics System

## 5 CONCLUSION

The analysis capabilities of our system have been applied onto a massive real life GPS dataset, obtained from 17,000 vehicles with on-board GPS receivers under a specific car insurance contract, tracked during one week of ordinary mobile activity in the urban area of the city of Milan; the dataset contains more than 2 million observations, yielding more than 200,000 trajectories. By applying our mobility data mining methods to this dataset, we developed a set of novel analytical services for mobility analysis and traffic management, designed and validated in collaboration with Milan Mobility Agency. We demonstrated how the various methods and systems developed in the project support the creation of novel analytical services for mobility management, such as: i) the automated construction of origin/destination matrices from mobility data in a timely, reliable and objective manner, in order to analyze users' flows between geographical areas, overcoming the limitations of the current survey-based approach; ii) the detailed analysis and discovery of systematic movement behaviours, i.e., the movements that repeat periodically during the week, with particular emphasis to cases of home-to-work and work-to-home commuting patterns.

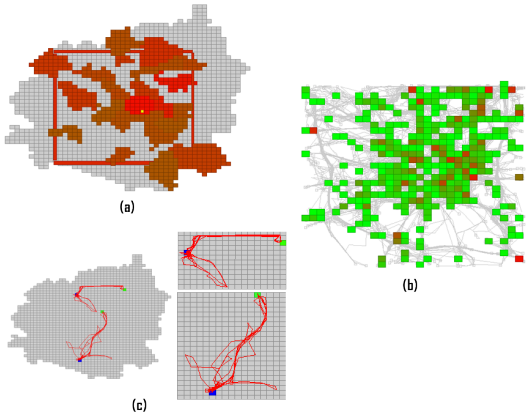


Figure 10: Some steps of the analysis: a view of the O/D matrix (a), the population distribution (b) and two home-to-work commuting patterns (c)

## References

[1] Geographic Privacy-aware Knowledge Discovery and Delivery Project. GeoPKDD, http://www.geopkdd.eu/

[2] S. Spaccapietra et al. A conceptual view on trajectories. Data Knowl. Eng., 65(1):126-146, 2008.

[3] Ralf Hartmut Guting, , et. al. SECONDO: An Extensible DBMS Platform for Research Prototyping and Teaching. In Proceeding of the International Conference on Data Engineering, ICDE, pages 1115-1116, Tokyo, Japan, April 2005.

[4] Mohamed F. Mokbel, et al. PLACE: A Query Processor for Handling Real-time Spatio-temporal Data Streams (Demo). In Proceeding of the International Conference on Very Large Data Bases, VLDB, pages 1377-1380, Toronto, Canada, August 2004.

[5] Ouri Wolfson, et al. (2002) Management of Dynamic Location Information in DOMINO (Demo). In Proceeding of the International Conference on Extending Database Technology, EDBT, pages: 769-771

[6] Pelekis, N. et al. (2006) Hermes - A Framework for Location-Based Data Management. Proceedings of EDBT. Pelekis, N. et al. (2008) HERMES: aggregative LBS via a trajectory DB engine. Proceedings of ACM SIGMOD.

[7] Marketos, et al. (2008) Building Real World Trajectory Warehouses. Proceedings of MobiDE.

[8] Orlando, S. et al. (2007) Spatio-Temporal Aggregations in Trajectory Data Warehouses. Proceedings of DaWaK.

[9] Pelekis, N. et al. (2008) Towards Trajectory Data Warehouses. Chapter in Mobility, Data Mining and Privacy: Geographic Knowledge Discovery. Springer-Verlag. 2008.

[10] O. Abul, F. Bonchi, and M. Nanni. Never Walk Alone: Uncertainty for anonymity in moving objects databases. In Proc. of the 24nd IEEE Int. Conf. on Data Engineering (ICDE'08)

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

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

[13] M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. J. Intell. Inf. Syst., 27(3):267-289, 2006.

[14] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko. Visually-driven analysis of movement data by progressive clustering. Information Visualization, 7((3/4)):225-239, 2008.

[15] Anna Monreale, Fabio Pinelli, Roberto Trasarti, Fosca Giannotti: WhereNext: a Location Predictor on Trajectory Pattern Mining.KDD 2009. 15th ACM SIGKDD Conference on Knoledge Discovery and Data Mining

[16] Riccardo Ortale, E. Ritacco, Nikos Pelekis, Roberto Trasarti, Gianni Costa, Fosca Giannotti, Giuseppe Manco, Chiara Renso, Yannis Theodoridis: The DAEDALUS Framework: Progressive Querying and Mining of Movement Data. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008).

[17] Gennady L. Andrienko, Natalia V. Andrienko: A Visual Analytics Approach to Exploration of Large Amounts of Movement Data. VISUAL 2008: 1-4

[18] Miriam Baglioni, Jos Antnio Fernandes de Macdo, Chiara Renso, Roberto Trasarti, Monica Wachowicz: Towards Semantic Interpretation of Movement Behavior. AGILE Conf. 2009: 271-288