

Mobility, Data Mining and Privacy: A Vision of Convergence

F. Giannotti and D. Pedreschi

The comprehension of phenomena related to movement – not only of people and vehicles but also of animals and other moving objects – has always been a key issue in many areas of scientific investigation or social analysis. The human geographer, for instance, studies the flows of migrant populations with reference to geography – places that are sources and destinations of migrations – and time. The historian, another example, studies military campaigns and related movements of armies and populations. (A famous instance is the depiction of Napoleon’s March on Moscow, published by C.J. Minard in 1861, discussed in Chap. 1 of this book (see Fig. 1.1); this figure represents with eloquence the fate of Napoleon’s army in the Russian campaign of 1812–1813, by showing the movement of the army together with its dramatically diminishing size during its advance and subsequent retreat.) The ethologist studies animal behaviour by the analysis of movement patterns, based on field observations or, sometimes, on data from tracking devices.

Today, in the extremely complex social systems of the gigantic metropolitan areas of the twenty-first century, the observation of the movement patterns and behavioural models of people is needed for the traffic engineers and city managers to reason about mobility and its sustainability and to support decision makers with trustable knowledge. The very same knowledge about people movement and behaviour is precious for the urban planner, e.g. to localise new services, to organise logistics systems and for the timely detection of changes that occur in the movement behaviour. At a finer-grained spatial scale, movement in contexts such as a shopping area or a natural park is an interesting subject of investigation, either for commercial purposes, as in geo-marketing, or for improving the quality of service.

In all the above cases, albeit so different from each other, two key problems recur:

- First, how to *collect mobility data* about extremely complex, often chaotic, social or natural systems made of large populations of moving entities.

F. Giannotti
KDD Laboratory, ISTI-CNR, Pisa, Italy, e-mail: fosca.giannotti@isti.cnr.it

- Second, how to turn this data into *mobility knowledge*, i.e. into useful models and patterns that abstract away from the individual and shed light on collective movement behaviour, pertaining to groups of individuals that it is worth putting into evidence.

In other words, by the observation of (many) individual movements – of a migrant, of one of Napoleon’s soldiers, of an animal, of a commuting worker in a city, of a tourist in a park – we aim at understanding the general movement patterns or models – a migratory flow, an army’s path, a frequently followed trajectory in the savannah, on the urban street network or in a park – that suddenly become usable knowledge, which makes the original system easier to understand by revealing some of its motion laws, hidden in the chaos. Simple and useful mobility knowledge is learned from complex systems of moving entities.

If this has been a long-time dream, never fully realised in practice, a chance to get closer to the dream is offered, today, by the convergence of two factors:

- The *mobility data* made available by the wireless and mobile communication technologies
- *Data mining* – the methods for extracting models and patterns from (large) volumes of data

1 Mobility Data

Our everyday actions, the way people live and move, leave digital traces in the information systems of the organisations that provide services through the wireless networks for mobile communication. The potential value of these traces in recording the human activities in a territory is becoming real, because of the increasing pervasiveness and positioning accuracy. The number of mobile phone users worldwide was estimated as 1.5 billion in 2005, with regions, such as Italy, where the number of mobile phones is exceeding the number of inhabitants; in other regions, especially developing countries, the numbers are still increasing at a high speed. On the other hand, the location technologies, such as GSM and UMTS, currently used by wireless phone operators are capable of providing an increasingly better estimate of a user’s location, while the integration of various positioning technologies proceeds: GPS-equipped mobile devices can transmit their trajectories to some service provider (and the European satellite positioning system Galileo may improve precision and pervasiveness in the near future), Wi-Fi and Bluetooth devices may be a source of data for indoor positioning, Wi-Max can become an alternative for outdoor positioning, and so on.

The consequence of this scenario, where communication and computing devices are ubiquitous and carried everywhere and always by people and vehicles, is that human activity in a territory may be *sensed* – not necessarily on purpose, but simply as a side effect of the ubiquitous services provided to mobile users. Thus, the wireless phone network, designed to provide mobile communication, can also be viewed

as an infrastructure to gather mobility data, if used to record the location of its users at different times. The wireless networks, whose pervasiveness and localisation precision increase while new location-based and context-based services are offered to mobile users, are becoming the *nerves* of our territory – in particular, our towns – capable of sensing and, possibly, recording our movements.

From this perspective, we have today a chance of collecting and storing mobility data of unprecedented quantity, quality and timeliness at a very low cost: in principle, a dream for traffic engineers and urban planners, compelled until yesterday to gather data of limited size and precision only through highly expensive means such as field experiments, surveys to discover travelling habits of commuting workers and ad hoc sensors placed on streets.

However, there's a long way to go from mobility data to mobility knowledge. In the words of J.H. Poincaré, 'Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.' Since databases became a mature technology and massive collection and storage of data became feasible at increasingly cheaper costs, a push emerged towards powerful methods for discovering knowledge from those data, capable of going beyond the limitations of traditional statistics, machine learning and database querying. This is what *data mining* is about.

2 Data Mining

Data mining is the process of automatically discovering useful information in large data repositories. Often, traditional data analysis tools and techniques cannot be used because of the massive volume of data gathered by automated collection tools, such as point-of-sale data, Web logs from e-commerce portals, earth observation data from satellites, genomic data. Sometimes, the non-traditional nature of the data implies that ordinary data analysis techniques are not applicable.

The three most popular data mining techniques are predictive modelling, cluster analysis and association analysis.

- In *predictive modelling*, the goal is to develop *classification models*, capable of predicting the value of a class label (or target variable) as a function of other variables (explanatory variables); the model is learnt from historical observations, where the class label of each sample is known: once constructed, a classification model is used to predict the class label of new samples whose class is unknown, as in forecasting whether a patient has a given disease based on the results of medical tests.
- In *association analysis*, also called *pattern discovery*, the goal is precisely to discover patterns that describe strong correlations among features in the data or associations among features that occur frequently in the data. Often, the discovered patterns are presented in the form of association rules: useful applications of association analysis include market basket analysis, i.e. the task of finding items

that are frequently purchased together, based on point-of-sale data collected at cash registers.

- In *cluster analysis*, the goal is to partition a data set into groups of closely related data in such a way that the observations belonging to the same group, or cluster, are similar to each other, while the observations belonging to different clusters are not. Clustering can be used, for instance, to find segments of customers with a similar purchasing behaviour or categories of documents pertaining to related topics.

Data mining is a step of *knowledge discovery in databases*, the so-called KDD process for converting raw data into useful knowledge. The KDD process consists of a series of transformation steps:

- *Data preprocessing*, which transforms the raw source data into an appropriate form for the subsequent analysis
- *Actual data mining*, which transforms the prepared data into patterns or models: classification models, clustering models, association patterns, etc.
- *Postprocessing* of data mining results, which assesses validity and usefulness of the extracted patterns and models, and presents interesting knowledge to the final users – business analysts, scientists, planners, etc. – by using appropriate visual metaphors or integrating knowledge into decision support systems

Today, data mining is both a technology that blends data analysis methods with sophisticated algorithms for processing large data sets, and an active research field that aims at developing new data analysis methods for novel forms of data. On one side, classification, clustering and pattern discovery tools are now part of mature data analysis systems and have been successfully applied to problems in various commercial and scientific domains. On the other side, the increasing heterogeneity and complexity of new forms of data – such as those arriving from medicine, biology, the Web, the Earth observation systems – call for new forms of patterns and models, together with new algorithms to discover such patterns and models efficiently. One of the frontiers of data mining research, today, is precisely represented by spatiotemporal data, i.e., observations of events that occur in a given place at a certain time, such as the mobility data arriving from wireless networks. Here, the challenge is particularly tough: which data mining tools are needed to master the complex dynamics of people in motion and construct concise and useful abstractions out of large volumes of mobility data is, by large, an unanswered question. Good news, hence, for researchers willing to engage in a highly interdisciplinary, highly risky and highly promising area, with a large potential impact on socially and economically relevant problems.

3 Mobility Data Mining

Mobility data mining is, therefore, emerging as a novel area of research, aimed at the analysis of mobility data by means of appropriate patterns and models extracted by efficient algorithms; it also aims at creating a novel knowledge discovery process

explicitly tailored to the analysis of mobility with reference to *geography*, at appropriate scales and granularity. In fact, movement always occurs in a given physical space, whose key semantic features are usually represented by geographical maps; as a consequence, the geographical background knowledge about a territory is always essential in understanding and analysing mobility in such territory. Mobility data mining, therefore, is situated in a Geographic Knowledge Discovery process – a term first introduced by Han and Miller in [2] – capable of sustaining the entire chain of production from raw mobility data up to usable knowledge capable of supporting decision making in real applications.

As a prototypical example, assume that source data are positioning logs from mobile cellular phones, reporting user's locations with reference to the cells in the GSM network; these mobility data come as streams of raw log entries recording users entering a cell – (*userID*, *time*, *cellID*, *in*) – users exiting a cell – (*userID*, *time*, *cellID*, *out*) – or, in the near future, user's position within a cell – (*userID*, *time*, *cellID*, *X*, *Y*) and, in the case of GPS/Galileo equipped devices, user's absolute position. Indeed, each time a mobile phone is used on a given network, the phone company records real-time data about it, including time and cell location. If a call is taking place, the recording data-rate may be higher. Note that if the caller is moving, the call transfers seamlessly from one cell to the next. In this context, a novel geographic knowledge discovery process may be envisaged, composed of three main steps: *trajectories reconstruction*, *knowledge extraction* and *delivery* of the information obtained, described in the following.

- (1) *Trajectory reconstruction.* In this basic phase, the stream of raw mobility data has to be processed to obtain trajectories of individual moving objects; the resulting trajectories should be stored into appropriate repositories, such as a trajectory database or data warehouse.

Reconstruction of trajectories is per se a challenging problem. The reconstruction accuracy of trajectories, as well as their level of spatiotemporal granularity, depend on the quality of the log entries, since the precision of the position may range from the granularity of a cell of varying size to the relative (approximated) position within a cell.

Indeed, each moving object trajectory is typically represented as a set of localisation points of the tracked device, called *sampling*. This representation has intrinsic imperfection mainly due to two aspects. The first source of imperfection is the measurement error of the tracking device. For example, a GPS-enabled device introduces a measurement error of a few metres, whereas the imprecision introduced in a GSM/UMTS network is the dimension of a cell, which could be from less than hundred metres in urban settings to a few kilometres in rural areas. The second source of imperfection is related to the sampling rate and involves the trajectory reconstruction process that approximates the movement of the objects between two localisation points. Although some simple approximated reconstruction techniques are sometimes applicable, more sophisticated reconstruction of trajectories from raw mobility data is to be investigated, to take into account the spatial, and possibly temporal, imperfection in the reconstruction process.

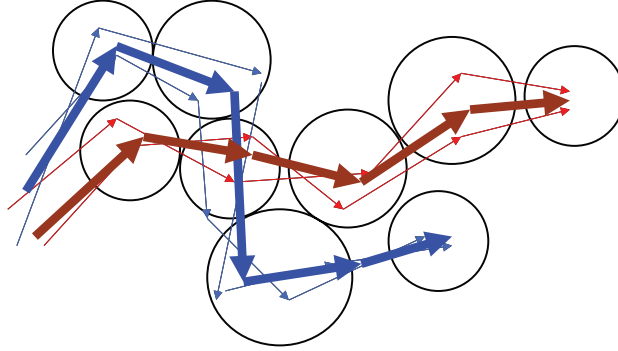


Fig. 1 Trajectory clustering

The management and querying of large volumes of mobility data and reconstructed trajectories also poses specific problems, which are only partly solved by currently available technology, such as moving object databases.

(2) *Knowledge extraction.* Spatiotemporal data mining methods are needed to extract useful patterns out of trajectories. However, spatiotemporal data mining is still in its infancy, and even the most basic questions in this field are still largely unanswered: What kinds of patterns can be extracted from trajectories? Which methods and algorithms should be applied to extract them? The following basic examples give a glimpse of the wide variety of patterns and possible applications it is expected to manage¹:

- *Clustering*, the discovery of groups of ‘similar’ trajectories, together with a summary of each group (see Fig. 1). Knowing which are the main routes (represented by clusters) followed by people or vehicles during the day can represent precious information for mobility analysis. For example, trajectory clusters may highlight the presence of important routes not adequately covered by the public transportation service.
- *Frequent patterns*, the discovery of frequently followed (sub)paths (Fig. 2). Such information can be useful in urban planning, e.g. by spotlighting frequently followed inefficient vehicle paths, which can be the result of a mistake in the road planning.
- *Classification*, the discovery of behaviour rules, aimed at explaining the behaviour of current users and predicting that of future ones (Fig. 3). Urban traffic simulations are a straightforward example of application for this kind of knowledge, since a classification model can represent a sophisticated alternative to the simple ad hoc behaviour rules, provided by domain experts, on which actual simulators are based.

¹ In the figures, circles represent cells in the wireless network.

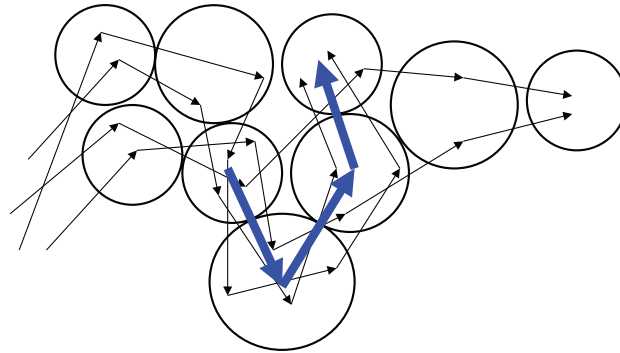


Fig. 2 Trajectory patterns

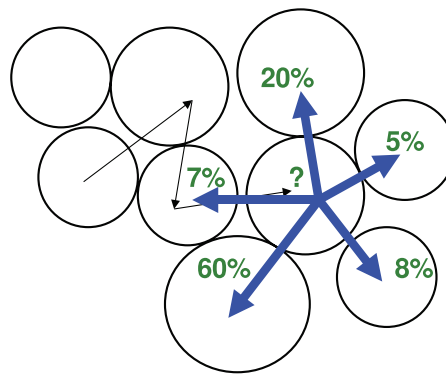


Fig. 3 Trajectory prediction

(3) *Knowledge delivery.* Extracted patterns are very seldom geographic knowledge *prêt-à-porter*: It is necessary to reason on patterns and on pertinent background knowledge, evaluate patterns' interestingness, refer them to geographic information and find out appropriate presentations and visualisations. Once suitable methods for interpreting and delivering geographic knowledge on trajectories are available, several application scenarios become possible. The paradigmatic example is sustainable mobility, namely how to support and improve decision making in mobility-related issues, such as

- Planning traffic and public mobility systems in metropolitan areas
- Planning physical communication networks, such as new roads or railways
- Localising new services in our towns
- Forecasting traffic-related phenomena
- Organising postal and logistics systems
- Timely detecting problems that emerge from the movement behaviour
- Timely detecting changes that occur in the movement behaviour

4 Privacy

Today we are faced with the concrete possibility of pursuing an *archaeology of the present*: discovering from the digital traces of our mobile activity the knowledge that makes us comprehend timely and precisely the way we live, the way we use our time and our land today.

Thus, it is becoming possible, in principle, to understand how to live better by learning from our recent history, i.e. from the traces left behind us yesterday, or a few moments ago, recorded in the information systems and analysed to produce usable, timely and reliable knowledge. In simple words, we advocate that mobility data mining, defined as the collection and extraction of knowledge from mobility data, is the opportunity to construct novel services of great societal and economic impact.

However, there is a little path from opportunities to threats: We are aware that, on the basis of this scenario, there lies a flaw of potentially dramatic impact, namely the fact that the donors of the mobility data are the citizens, and making these data publicly available for the mentioned purposes would put at risk our own privacy, our natural right to keep secret the places we visit, the places we live or work at and the people we meet – all in all, the way we live as individuals. In other words, the personal mobility data, as gathered by the wireless networks, are extremely sensitive information; their disclosure may represent a brutal violation of the privacy protection rights, established in increasingly more laws and regulations internationally.

A genuine positivist researcher, with an unlimited trust in science and progress, may observe that, for the mobility-related analytical purposes, knowing the exact identity of individuals is not needed: anonymous data are enough to reconstruct aggregate movement behaviour, pertaining to whole groups of people, not to individual persons. This line of reasoning is also coherent with existing data protection regulations, such as that of the European Union, which states that personal data, once made anonymous, are not subject any longer to the restrictions of the privacy law. Unfortunately, this is not so easy: the problem is that anonymity means making reasonably impossible the re-identification, i.e. the linkage between the personal data of an individual and the identity of the individual itself. Therefore, transforming the data in such a way to guarantee anonymity is hard: as some realistic examples show, supposedly anonymous data sets can leave unexpected doors open to malicious re-identification attacks. Chapter 4 discusses such examples in different domains such as medical patient data, Web search logs and location and trajectory data; moreover, other possible breaches for privacy violation may be left open by the publication of the mining results, even in the case that the source data are kept secret by a trusted data custodian.

The bottom-line of this discussion is that protecting privacy when disclosing mobility knowledge is a non-trivial problem that, besides socially relevant, is scientifically attractive. As often happens in science, the problem is to find an optimal trade-off between two conflicting goals: from one side, we would like to have precise, fine-grained knowledge about mobility, which is useful for the analytic

purposes; from the other side, we would like to have imprecise, coarse-grained knowledge about mobility, which puts us in repair from the attacks to our privacy. It is interesting that the same conflict – essentially between opportunities and risks – can be read either as a mathematical problem or as a social (or ethical or legal) challenge. Indeed, the privacy issues related to the ICTs can only be addressed through an alliance of technology, legal regulations and social norms. In the meanwhile, increasingly sophisticated privacy-preserving techniques are being studied. Their aim is to achieve appropriate levels of anonymity by means of controlled transformation of data and/or patterns – limited distortion that avoids the undesired side effect on privacy while preserving the possibility of discovering useful knowledge. A fascinating array of problems thus emerged, from the point of view of computer scientists and mathematicians, which already stimulated the production of important ideas and tools. Hopefully, in the near future, it will be possible to reach a win-win situation: obtaining the advantages of collective mobility knowledge without divulging inadvertently any individual mobility knowledge. These results, if achieved, may have an impact on laws and jurisprudence, as well as on the social acceptance and dissemination of ubiquitous technologies.

5 Purpose of this Book

Mobility, data mining and privacy: There is a new multi-disciplinary research frontier that is emerging at the crossroads of these three subjects, with plenty of challenging scientific problems to be solved and vast potential impact on real-life problems. This is the conviction that brought us to create a large European project called GeoPKDD – *Geographic Privacy-aware Knowledge Discovery and Delivery* [1] – that, since December 2005, is exploring this frontier of research. The same conviction is the basis of this book, produced by the community of researchers of the GeoPKDD project, which is thoroughly aimed at substantiating the vision advocated above.

The approach that we followed in undertaking this task is twofold: first, in Part I of the book, we set up the stage and make the vision more concrete, by discussing which elements of the three subjects are involved in the convergence: mobility (Which data come from the wireless networks?), data mining (in which classes of applications can be addressed with a geographic knowledge discovery process) and privacy (Which is the interplay between the privacy-preserving technologies and the data protection laws?). Second, in the subsequent parts of the book, we identify the scientific and technological ingredients that, from a computer science perspective, are needed to support a geographic knowledge discovery process; for each such ingredient we discuss the current state of the art and the roadmap of research that we expect.

More precisely, the book is organised as follows.

In Part I (*Setting the stage*), Chap. 1 introduces the basic notions related to the movement of objects and the data that describe the movement; Chap. 2 characterises

the next generation of mobility-related applications through a privacy-aware geographic knowledge discovery process; Chap. 3 discusses tracking of mobility data and trajectories from wireless networks and Chap. 4 discusses privacy protection regulations and technologies, together with related opportunities and threats.

In Part II (*Managing moving object and trajectory data*), Chap. 5 discusses data modelling for moving objects and trajectories; Chap. 6 deals with trajectory database management issues and physical aspects of trajectory database systems, such as indexing and query processing; Chap. 7 discusses the first steps towards a trajectory data warehouse providing online analytical tools for trajectory data and Chap. 8 discusses the location privacy problem in spatiotemporal and trajectory data, also taking into account security.

In Part III (*Mining spatiotemporal and trajectory data*), Chap. 9 discusses the knowledge discovery and data mining techniques applied to geographical data, i.e. data referenced to geographic information; Chap. 10 deals with spatiotemporal data mining, i.e. knowledge discovery from mobility data, where the space and time dimensions are inextricably intertwined; Chap. 11 discusses the privacy-preserving methods (and problems) in data mining, with a particular focus on the specific privacy and anonymity issues arising in spatiotemporal data mining; Chap. 12 discusses the quest towards a language framework, capable of supporting the user in specifying and refining mining objectives, combining multiple strategies and defining the quality of the extracted knowledge, in the specific context of movement data and Chap. 13 considers the use of interactive visual techniques for detection of various patterns and relationships in movement data.

This is more a book of questions, rather than a book of answers. It is clearly devoted to shape up a research area, and therefore targeted at researchers that are looking for challenging open problems in an exciting interdisciplinary subject. This is why we tried to speak, as far as possible, a language comprehensible to researchers coming from various subareas of computer science, including databases, data mining, machine learning, algorithms, data modelling, visualisation and geographic information systems. But, more ambitiously, we also tried to speak to researchers from the other disciplines that are needed to fully realise the vision: geography, statistics, social sciences, law, telecommunication engineering and transportation engineering. We believe that at least the material in Part I, and also most of the remaining chapters, can reach the attention of researchers who are interested in the inter-disciplinary dialogue, and perceive the interplay among mobility, the information and communication technologies and privacy as a potential ground for such a dialogue. Most of, if not all, open challenges of the contemporary society are intrinsically multi-disciplinary, and require solutions – hence research – that cross the boundaries of traditional disciplines: we like to think that this book is a little step in this direction.

References

1. GeoPKDD.eu – Geographic Privacy-aware Knowledge Discovery and Delivery. <http://www.geopkdd.eu/>.
2. H.J. Miller and J. Han (eds). *Geographic Data Mining and Knowledge Discovery*. Taylor & Francis, 2001.



<http://www.springer.com/978-3-540-75176-2>

Mobility, Data Mining and Privacy
Geographic Knowledge Discovery
Giannotti, F.; Pedreschi, D. (Eds.)
2008, XIV, 410 p., Hardcover
ISBN: 978-3-540-75176-2