Understanding Human Mobility with Big Data

Comprendere la Mobilità Umana usando i Big data

Fosca Giannotti and Salvatore Rinzivillo

Abstract The paper illustrates basic methods of mobility data mining, designed to extract from the big mobility data the patterns of collective movement behavior, i.e., discover the subgroups of travelers characterized by a common purpose, profiles of individual movement activity, i.e., characterize the routine mobility of each traveler. We illustrate a number of concrete case studies where mobility data mining is put at work to create powerful analytical services for policy makers, businesses, public administrations, and individual citizens.

Abstract Il lavoro illustra i metodi basici del mobility data mining progettati per estrarre dai dati di mobilità comportamenti collettivi, cioè scoprire gruppi di viaggiatori caratterizzati da mobilità simile, i comportamenti routinari di ogni individuo. Illustriamo alcun casi di studio conncreti, dove il mobility data nining è la base per costruire servizi di supporto alle decisioni per politici, amministratori pubblici e cittadini.

Key words: mobility data mining, Big data analytics,

1 Introduction

The large availability of location aware services allows the collection of huge repository of movement data. These new sources of data give an unprecedented opportunity to have a social microscope of individual collective and global behaviours- Here we focus on mobility data, such as the call detail records from mobile phones and the GPS tracks from car navigation devices, which represent society-wide proxies of human mobile activities. These big mobility data help us

[1] Fosca Giannotti, ISTI CNR, email: fosca.giannotti@isti.cnr.it
Salvatore Rinzivillo; email: Salvatore.Rinzivillo@isti.cnr.it

understand human mobility, and discover the hidden patterns and profiles that characterize the trajectories we follow during our daily activity.
The paper illustrates the basic methods of mobility data mining, designed to extract from the big mobility data the patterns of collective movement behavior, i.e., discover the subgroups of travelers characterized by a common purpose, profiles of individual movement activity, typical path followed by many travellers. These methods are the basic breaks to support analytical questions such as:

- What are the most popular itineraries followed from the origin to the destination of people's travels? What routes, what timing, what volume for each such itinerary?
- How do people leave the city toward suburban areas (or vice-versa)? What is the spatio-temporal distribution of such trips?
- How to understand the accessibility to key mobility attractors, such as large facilities, railway stations or airports? How do people behave when approaching an attractor?
- How to detect an extraordinary event and understand the associated mobility behavior? How and when do people reach and leave the event's location? What is the spatio-temporal distribution of such (portion of) trips?
- What will be the areas with highest traffic volume in the next hour(s)? To what extent are our predictions accurate?
- Are there geographic borders that emerge from the way people use the territory for their daily activities? If so, how do we define such borders? Are these borders matching the administrative ones?

In this paper shortly illustrate the system Urban Mobility Atlas, that visually synthetizes the complex analytical processes in a toolset of measures for various mobility dimensions of a geographical area. We focus on the challenge of constructing novel mobility indicators from Big Data, capable of capturing the mobility vocation of a territory: what is the relationship between systematic and non systematic behavior? Is a territory amenable for adopting a new mobility behavior such as car-pooling or for massive diffusion of electric vehicles?

In the following we will consider a big dataset of GPS traces of private vehicles circulating in central Italy, in the region of Tuscany. The owners of these cars are subscribers of a *pay-as-you-drive* car insurance contract, under which the tracked trajectories of each vehicle are periodically sent (through the GSM network) to a central server for anti-fraud and anti-theft purposes. This dataset has been donated for research purposes by *Octo Telematics Italia S.r.l* (oct), the leader for this sector in Europe. The whole dataset describes about 150,000 cars tracked during a month (May, 2012) in Tuscany.

2 Mobility Data Analysis

The collection of different sources of data is the very first step of an analytical process that involves several transformations of data to gather useful and novel knowledge from it. In particular, for mobility data it is necessary, at first, to process the raw observations of each position at a given time into a *trajectory*, i.e. an higher level object representing the movement of an individual from an origin to a destination.

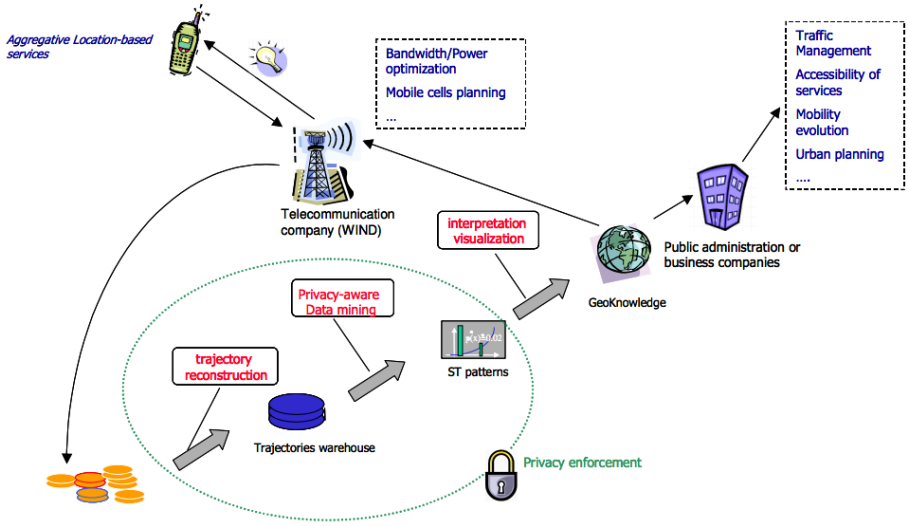

Figure 1: The GeoPKDD process [3].

This process is very complex and depends mainly on the spatio-temporal granularity of the raw data. In [2], it is settled the ground for a new research field, namely Mobility Data Mining (MDM). The majority of methods of MDM are centred around the concept of trajectory. In [2], the authors present the advantages of combining different MDM algorithm and method to derive more complex analytical processes that can be easily deployed as stand-alone services.

Origin Destination matrix model provides a very compact representation of traffic demands, by abstracting the single trajectories to flows between any two regions. The large availability of sensed tracks enables us to automatically extract OD matrices, by dividing the territory into a partition of cells and by counting the movements from each origin to each destination. The resulting model can be presented to the analyst of to the traffic manager to let her browse the interesting flows with an advanced visual interface. The combination of the OD matrix model with clustering methods, for example, may enable an analyst to understand which are the most popular itineraries followed by vehicles entering (or exiting) to a given territory. The extraction of OD matrix models strictly depends on the spatio-temporal granularity of the movement data. When using a data source with a variable sampling rate, like for example GSM data, it is necessary to plan

reconstruction strategies for missing data. In [4], the authors use data mining methods and frequency analysis to derive OD matrix from a large collection of GSM data. In this approach, the systematic movements are first identified within the raw data and the OD matrix is derived from this subset. The outcoming OD model is then fed to a traffic simulator for traffic assignment to the road network.

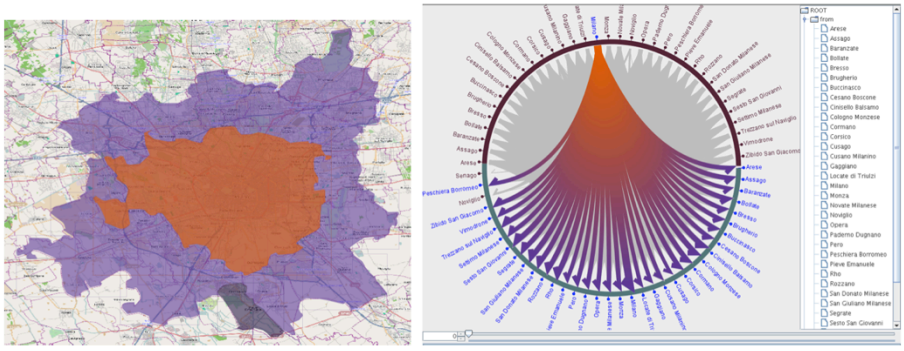

Figure 2: An example of Matrix model exploration with GPS data. (left) Spatial partition of the territory: the center region (in orange) contains the administrative borders of a city; the purple cells represent the adjacent cities. (right) The visual interface to browse the OD Matrix: each region is represented with a node, nodes are displayed in a circular layout. The arc connecting two nodes represents the flow, i.e., the number of trips from the origin to the destination node; the arc width is proportional to the flow.

Using the OD Matrix Exploration we may select relevant flows that may be further investigated. For example, we can focus on the set of trajectories living the central city and moving towards North-East (Figure 3 (left)). Despite all these trips originate in the city center and end in the NE suburbs, a broad diversity is still evident. In order to understand which are the most popular itineraries followed by the selected travels, we apply an algorithm that automatically detects significant groups of similar trips.

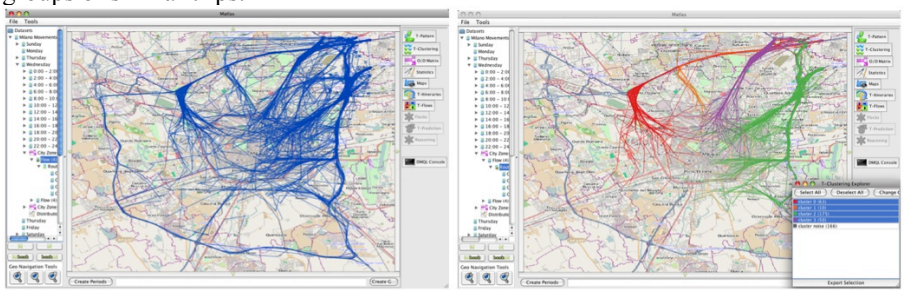

Figure 3: The result of clustering from the trajectories moving from the center to the North-East area. (left) The input dataset for the clustering algorithm: the trajectories moving from the center to the North-East area. (right) The resulting clusters using the Route Similarity distance function. The clusters are visualized using a themed color, and the analyst can select and browse them separately.

In particular we use the density-based clustering algorithm with the Spatial Route distance function. The clustering algorithm produces a set of clusters, each of which

can be visualized by means of a thematic rendering where the trajectories in the same cluster are drawn with the same color. Figure 2(right) shows how the most popular clusters highlight the main routes used by drivers to leave the center towards NE. The frequent behaviors identified by the clustering processing may be also analyzed by the temporal dimension. For example, in Figure 4, it is presented the temporal distribution of the trips in each cluster along a typical day. We can use such statistics to interpret the semantic of each group of trips.

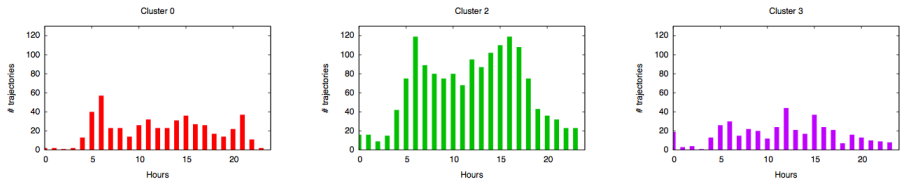

Figure 4: Temporal distribution of the trajectories in the clusters of Figure 1.6(right) on the hours of weekdays. Cluster 0 and Cluster 3 (left, right) do not exhibit significant peaks, while cluster 2 (center) has a peak in the morning and one in the afternoon. The tempo- ral profile of Cluster 2 captures two commuting behaviors: a group leaving the city in the morning (commuters going to work outside), and a larger group leaving the city in the late afternoon (commuters coming back home in the suburbs after work.)

From the analysis of collective movements we are also able to identify the actual borders that emerge from daily behaviours of people [6]. The aim of discovering borders at a meso-scale is motivated by providing decision- support tools for policy makers, capable of suggesting optimal administrative borders for the government of the territory. This analytical process is based on techniques developed for complex network analysis. Figure 5 shows the resulting border of analyzing the movements of around 40k vehicles for one month in the center-west part of Tuscany. From the figure is possible to recognize the relevant socio-cultural district of the region. For comparison, we also draw the current administrative borders of each municipality.

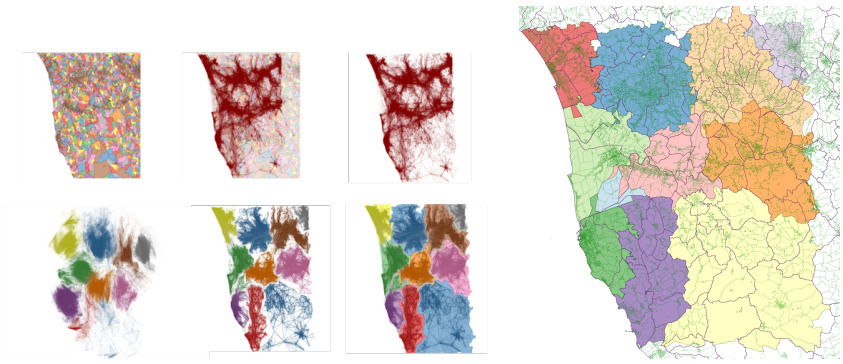

Figure 5: Visualization of the mobility borders in Tuscany. As a reference with the existing administrative borders, the perimeter of each town is drawn with a thicker line. Regions within the same cluster are themed with the same color.

Shifting from the collective behaviors to the individuals we can investigate movement from a semantic point of view, giving relevance to the daily routines of each person. 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 [7], 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 6:

1. divide the whole history of the user into trips (Figure 6(a));
2. group trips that are similar, discarding the outliers (Figure 6(b));
3. from each group, extract a set of representative trips, to be used as mobility profiles (Figure 6(c)).

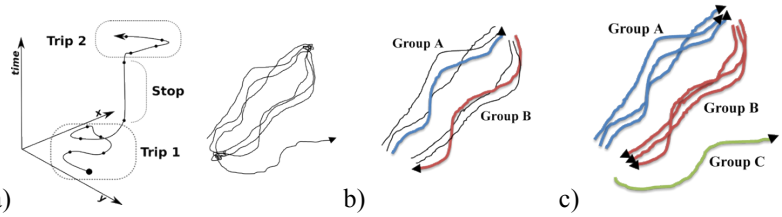

Figure 6: Mobility profile extraction process: (a) trip identification; (b) group detection/outlier removal; (c)selection of representative mobility profiles.

The concept of individual profile is also exploited in [1], where the authors reconstruct the calling habits of customers of a mobile operator to determine a class for each individual among a set of predefined ones: resident, visitors, commuters. In this case, it is exploited the relations between the individual and the collectivity.

3 Human Mobility Indicators for the Urban Mobility Atlas

Starting from the analytical processes presented in the previous section, it is interesting to derive quantitative estimations of the main characteristics that distinguish the mobility of a territory. We call such estimators *mobility indicators* and, for each territory, we compute and aggregate several measurement to have an overview of the whole mobility. As a first step, it is crucial to identify the mobility analysis that may be relevant for our objectives. In particular, we describe here two complex mobility indicators that are based on individual characteristics of the observed population: the *systematicity index* measures the tendency of a population of following frequent routines to move among places; the *radius of gyration* measures the tendency of a population of exploring places far away from their home. Both these measures are associated to an individual: we need to extend them from a single user to a group of users. We can imagine the mobility of a user as composed of several locations she has visited during her movements. Some of these locations are more relevant to the individual, for example since she spend the majority of her time in a few places. By analysing the individual history, we can determine the *Most*

Frequent Location (MFL) and we may assume this as the *home location*. Given the MFL of a user *u*, we may univocally associate *u* to the area where her MFL is located. This association can be exploited to define a derived estimator for an area *A* as an aggregation of the measures of all users whose MFL is located in *A*.

To determine the systematicity of a user, we exploit the mobility profile extraction method [7]. This method allows us to separate systematic movements from the unfrequent ones. The individual systematicity index of a user *u* is given by the ratio of the systematic trips over the non-systematic ones. This index is extended to a territory *A* by averaging the individual indexes of each individual linked to *A*.

Moreover, by means of cluster analysis, it is possible to aggregate individual profiles of users to determine common routes to access a territory. We call such routes *access* patterns. The access patterns can be exploited to reason about the shared strategies of users to access a destination. In combination with the systematicity index, it is possible to give a qualitative measure of such routes. We have designed a visual widget to represent the combination of access patterns for a specific municipality. Figure 1 shows a compact representation of access patterns arriving to Florence: linestrings represent the individual mobility profiles, similar profiles are rendered with the same color, and for each group we have a pie-chart representing the ratio of systematic mobility over the whole movements.

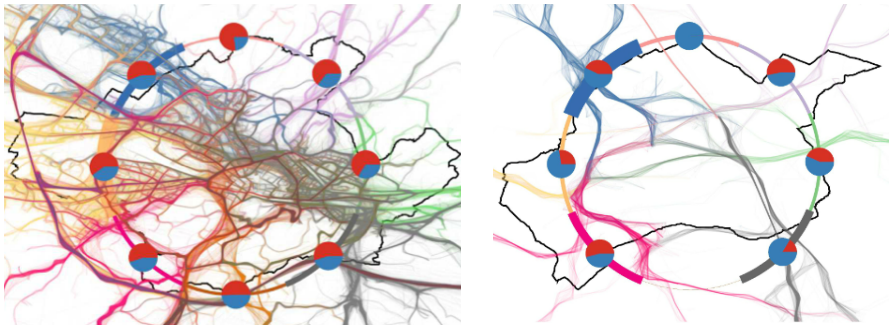

Figure 7: Incoming traffic to Florence (left) and Montepulciano (Right). Movements are aggregated by similarity of access point to the city. Each route is annotated with the Systematicity indicator: red color represents systematic movements, blue color the remaining mobility.

From the comparison of the two plots in Figure 7 we can grasp easily the differences between two different cities: on the left we have a large metropolitan area capable of attracting mobility for services and work; on the right we have a small town whose mobility is mainly influenced by the neighbouring municipalities. Besides the differences in volume, we can notice which access patterns are used for systematic movements. In Florence the ratio of systematic trips is very high for every access path. In Montepulciano, on the contrary, there differences among the different incoming routes. This result may highlight the relationship and exchange of commuters with neighbouring cities.

At a glance, it is possible to notice the main routes of systematic movements arriving from North. On the contrary, the trajectories coming from South are equally populated by systematic and non-systematic trajectories.

The concept of Radius of Gyration gives a measure of the spreadness of visited locations of a user over the geography. It is based on the concept of *center of mass r_cm* of a user, defined as a two-dimensional vector representing the weighted mean point of the locations visited by an individual. We can measure the mass associated to a location with its visitation frequency or the time spent in the location, obtaining the following definition:

$$r_{cm} = \frac{1}{N}\sum_{i \in L} n_i r_i$$

where L is the set of locations visited by the user, r_i is a two-dimensional vector describing the geographic coordinates of location i; n_i is the visitation frequency or the total time spent; and N is the sum over all the locations of n_i (i.e. total number of visits or time spent).

The radius of gyration r_g of a user characterizes how spread out the visited locations are from the center of mass. In mathematical terms, it is defined as the root mean square distance of the locations from the center of mass:

$$r_g = \sqrt{\frac{1}{N}\sum_{i \in L} n_i (r_i - r_{cm})^2}$$

where L is the locations set visited by the user; n_i is the visitation frequency or the time spent; N is the sum of all the single frequencies/times; r_i and r_{cm} are the vector of coordinates of location i and center of mass respectively.

The radius of gyration provides us with a measure of mobility volume, indicating the typical distance traveled by an individual and providing an estimation of her tendency to move.

By analysing the distribution of the values of r_g in a city, we can understand which are the characteristic movements of a city. In Figure 8(Left) we may notice a skew distribution of r_g in the observed population in Florence. The majority of people has a very limited value of radius, mainly within the city boundaries. However, there are still a consistent part of people travelling long distances.

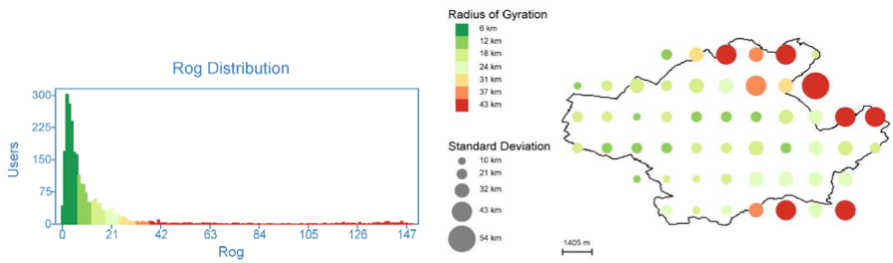

Figure 8: Visual widgets for Radius of Gyration indicator for the city of Florence. (Left) distribution of the r_g on the resident population. (Right) Spatial distribution of mean and standard deviation of r_g per cell.

Since each individual is linked to her MFL, we design also a spatial distribution of the r_g. In particular, we consider a spatial partition of the territory (we have used a regular grid in Figure 8(Right)) and for each cell we aggregate the corresponding values of r_g: the mean value is coded with a color scale (linked also to the distribution histogram); the standard deviation is represented as the circle radius. From the widget in Figure 8 (Right) we can notice how the internal areas of the city have both a low mean value of r_g and also a low standard deviation. On the boundary, instead, we can notice higher values of r_g and also a very high variance.

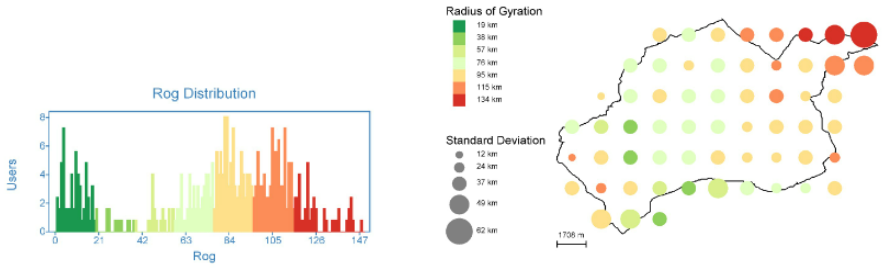

Figure 9: Visual widgets for Radius of Gyration indicator for the city of Montepulciano.

If we consider another city, like for example Montepulciano in Figure 9, we can highlight the different behaviours in mobility of the respective residents. In this case, people tend to travel longer distances, as it emerges from the r_g distribution. This may be justified with the need of finding services that are not directly accessible in a small town.

4 The Urban Mobility Atlas

In the previous section we demonstrated the advantages of having a compact visual representation of two complex mobility indicators. We extend such approach to different indicators and we try to aggregate them in an organic and comprehensible composition. The visual representation of these indicators is presented in Figure 10. The layout is composed of four sections: an header containing general information about the city (number of vehicles, period of observation, etc); a series of mobility statistics (distribution of length of trajectories, distribution of duration, etc.); distribution of individual mobility characteristics (radius of gyration, time spent per location,etc.); incoming traffic statistics; outgoing traffic statistics. The Urban Mobility Atlas can be created on a territory given the availability of a large dataset of vehicular trajectory. We created an instance of the Urban Mobility Atlas in Tuscany, exploiting the GPS dataset provided by *OCTO Telematics* covering around the 5% of the circulating vehicles. An interesting question is to assess if the dimension of such dataset, as well as the spatial coverage that it shows, are enough to support the accuracy and representativeness of the overall vehicular mobility; several studies on the literature [5] have tackled this problem by building an

extremely accurate predictor that, given GPS data observation, estimates the real traffic values as measured by road sensors (ground truth).

The user can navigate through a visual interface, where she can browse each city on a map.

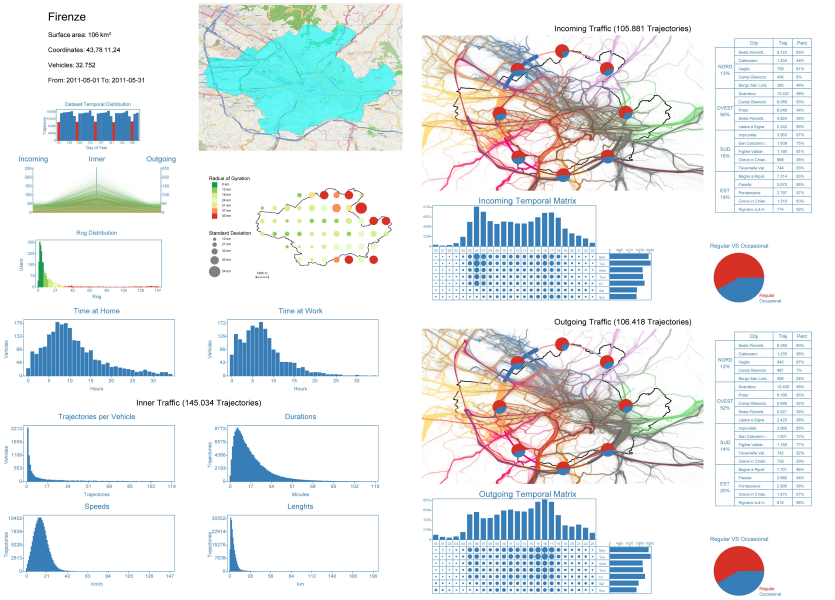

Figure 10: Urban Mobility Atlas for the city of Florence.

References

1. B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp ’12, pages 17–24, New York, NY, USA. ACM Press (2012).
2. 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. *VLDB J.*, 20(5):695–719, 2011.
3. Giannotti, F. Pedreschi, D. Mobility, Data Mining and Privacy. Springer (2008)
4. Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Peter Van Der Mede, Joost De Bruijn, Erik De Romph, Gerard Bruil. MP4-A Project: Mobility Planning For Africa. NetMob 2013, Third conference on the Analysis of Mobile Phone Datasets (2013)
5. L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti. Under- standing the patterns of car travel,. *European Physics Journal Special Topics*, 215:61–73, 2013.
6. Rinzivillo, S, Mainardi, S. Pezzoni, F. Coscia, M. Pedreschi, D. Giannotti, F. Discovering the Geographical Borders of Human Mobility. Künstliche Intelligenz 26(3): 253-260 (2012)
7. R. Trasarti, F. Pinelli, M. Nanni, and F. Giannotti. Mining mobility user profiles for car pooling. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’11, pages 1190–1198, New York, NY, USA, 2011. ACM.