D2.2 Semantic-enriched data-driven theory of mobility demand and final framework for integration

DATA science for SIMulating the era of electric vehicles

www.datasim-fp7.eu

Project funded by the European Community under the Information and Communication Technologies Programme - Contract ICT-FP7-270833

Abstract

This report presents the work pursued during the second year of the project by the partners participating in the Workpackage 2. The document propose an abstract simulation pipeline to discuss how the methods and patterns investigated in the WP may be integrated in a new generation of agent-based simulation systems.
# Contents

1 **Introduction** .................................................. 4  
2 **Global Models for Mobility and Social Behavior** .......... 7  
   2.1 Radiation Model ............................................... 7  
   2.2 Heterogeneity of Human Mobility .............................. 8  
   2.3 Diffusion of Online Service Innovations ..................... 11  
3 **Semantic Enrichment of User Activities** ................. 15  
   3.1 Human Assisted Semantic Enrichment ......................... 16  
      3.1.1 Finding Significant Places ............................ 16  
      3.1.2 Annotating Trajectories with DayTag ................ 18  
      3.1.3 Annotation using web data ............................ 21  
   3.2 Automatic Semantic Enrichment ............................... 24  
      3.2.1 Annotation using background knowledge and POIs .... 24  
      3.2.2 Semantic annotation from trip features ............... 26  
      3.2.3 Semantic annotation from GSM data .................... 28  
      3.2.4 Annotation using trip and activities features ......... 29  
      3.2.5 Annotation using Individual Mobility Networks ....... 31  
   3.3 Semantic Quantification of User Profiles ................... 34  
      3.3.1 Segmenting population with GSM Call Habits ........ 34  
4 **Understanding Mobility with Complex Network Analysis** .... 38  
   4.1 Time-dependent Mobility Borders ............................ 38  
   4.2 Purchasing Behavior and Mobility ............................ 41  
   4.3 Big Data and Network Analytics for Boosting Car Pooling 42  
5 **Knowledge Transition to Simulator and Future Directions** .. 44  
   5.1 Agent-based Simulation Process ................................ 45  
      5.1.1 Big Data Annotation ..................................... 45  
      5.1.2 Generative Model Extraction ............................ 46
5.1.3 Mapping to Geography ........................................ 47

6 Directions and Future Plans ........................................ 52

6.1 Generating Synthetic Mobility ................................. 52
6.2 Impact of Electric Vehicles on Mobility ..................... 54
6.3 Diffusion of car pooling and electric cars .................... 54
6.4 Consolidation of Patterns and Models ......................... 54
Chapter 1

Introduction

One of the main objectives of the workpackage is the study of human mobility to root a new generation of simulators to the models and patterns emerging from big data. We propose a paradigm shift from the current agent-based simulation systems, by assigning a central role to the behaviors extracted from data. While traditional simulators are rooted on the availability of paper-and-pencil annotated surveys, we candidiate the patterns and model emerging from big data through automatic analysis to be embedded within the simulation process. We imagine two parallel steps to pursue this objective: a) we exploit big data to automatically produce annotated surveys to be fed to existing simulators; b) we embed specific behavioral models directly into the simulation pipeline.

A prerequisite for both the two steps is the availability of semantically rich data. Semantic rich data are rare and usually they originate from ad hoc collection campaigns. However, we can witness a broad diffusion of services and online platforms that allow users to create content that can be linked to their activities in space. The steady growth of such services suggests that in a recent future it would be possible to derive rich data opportunistically from big data. Since, at the present, the availability of this kind of data is not sufficient to support a modelling process as it is, for the short-term objectives of the project we have devised several methods to infer semantics from available data. In Chapter 3 we proposed a repertoire of methods to enrich raw movement data with user activity purpose. Beside allowing to produce rich data for our analyses, the methods developed helped to obtain deeper insights of the behaviors that drive human mobility. Some of the results may be generalized directly from this process to obtain a high level description of internal mechanisms that govern the mobility of individuals.

Step b), involving the improvement of the core simulation process, clearly presents an important scientific challenge. In this respect, we plan to plug the
models describing human mobility within the simulator pipeline directly. We have generalized the simulation pipeline into five macro steps, showed sketchily in Figure 1.1. At each step, we propose to embed the descriptive models developed within the workpackage to augment the quality of the resulting simulation. An overview of these methods is presented in Chapter 2 and Chapter 4. Here, one of the challenges we have to face is to exploit such methods as generative processes. We already started investigating some integrations within the current simulator available from WP3. In particular, in Chapter 5 we propose a blueprint of a simulator enriched with the contributed models and patterns and we give a first overview of their integration.

Figure 1.1: Abstract pipeline for an agent-based simulator.

This documents presents the activities and research works performed within Work Package 2. The workpackage is organized into four tasks:

- **Task 2.1: Person-based constrained trajectory mining based on GPS data.** The objective of this task is to demonstrate how the combination of network science with data mining within a uniform analytical framework is able to develop macro-micro models of human mobility and achieve an unprecedented explanatory and predictive power.

- **Task 2.2: Semantic enrichment of mined trajectories.** This task investigates a semantic enrichment and reasoning process to characterize movement data with domain-dependent behavior definitions. The discovered mobility models and patterns will be expressed in terms of realistic semantics of movements. The semantic annotation/reconstruction of mobility patterns will be based on the mobility and social networking data, and the socio-demographic data provided by WP1.

- **Task 2.3: Social network mining.** In this task we plan to analyze the topological and the dynamical properties of the social network of users, to the
purpose of better characterizing the mobility behaviours of sub-populations on the basis of their social relations. The analysis of networks brings into the project another semantic dimension which may contribute at better understanding the mobility behavior at society-wide scale.

- **Task 2.4: Knowledge transition to Simulator.** This task will integrate the patterns, models of the previous tasks to blend the behavioural and semantic description of individual traveler into the novel agent-based simulation system developed in WP3, which is able to evaluate policy and behaviourally reasoning on a country-wide scale.

At the end of Year 1, in Deliverable 2.1, we proposed a taxonomy of analytical tools based on the type of patterns and models extracted: *individual patterns*, based on information and patterns of each individual; *collective patterns* describing the commonalities among groups of individuals; *global models* defining general rules of mobility at society-wide scale. These three lines of investigations have been maintained during Year 2 activities. In some cases –like the methods for semantic enrichment (Chapter 3) – we used a combination of models and patterns of the three categories. In this report we will present a description of the main methods developed by the consortium during Year 2.

The deliverable is organized as follows. In Chapter 2 we present models that describe general laws of human mobility. This year the emphasis was posed on the study of the relation between human mobility and its surroundings. Chapter 3 presents an overview of the semantic enrichment methods developed by the consortium. This set of works focus mainly on individual mobility. However, one of the partners has proposed semantically rich methods to quantify and annotate groups of individuals. The process of semantic enrichment requires dependable supervised datasets on top of which we can base our study. For this reason, the consortium has developed a series of human assisted frameworks to assign or revise semantics of movement (Section 3.1). Since we are coping with huge amount of data, the process of semantic enrichment should be as much as automatic as possible. The methods of Section 3.2 represent different approaches to this problem. Chapter 4 presents a selection of studies where the methods from network analytics are exploited to analyze inter-personal relations among individuals, on the basis of their mobility. Chapter 5 provides a workplan to integrate the methods presented in a new generation of agent-based simulator. The Chapter gives a rationale of a simulation pipeline and for each step it highlights how the proposed methods may be exploited. In conclusion, an actual timeline of implementing the new simulator is provided, which will be implemented within Task 1 of Workpackage 3, during the last year of the project.
Chapter 2

Global Models for Mobility and Social Behavior

In this chapter we challenge the problem of describing social and mobility behavior at a large scale with the objective to grasp the general laws that subsume societal phenomena. These global approaches consider mobility and socio-economic context (Section 2.1 and Section 2.2) and social interactions and influences 2.3

2.1 Radiation Model

Destination selection can be intended as the trip distribution step of the 4-step model, in which a destination is assigned to each trip. The gravity model [67, 21] is the traditional framework to predict origin-destination (OD) matrices.

Recently a new family of models, called radiation model [55, 56], has been proposed. In [57], BME, IMOB and CNR has worked on the task of including the radiation model in the simulator, testing its performance in different scenarios, and developing various methods to enhance its predictive power and versatility. To this end the authors focused on the prediction of commuting trips’ OD matrix, and they used the radiation model with selection [56] to estimate home-to-work trips between the 309 Flanders’ municipalities.

To validate the predictive power of the model, the authors first approach was to use population as a proxy to estimate the number of employment opportunities in each location, $a$, assuming in first approximation a linear relationship between population and job openings. The model’s prediction depends also on the metric used to calculate the distance between locations, as the variable $a$ depends on the particular metric selected. The proposed approach uses three different definitions of distance: the geographic distance, the road distance, and the average travel time.
A second test consisted in a modification of the original assumption that the number of job openings, $a$, is proportional to the resident population in each location. Indeed, highly populated residential areas usually have fewer job openings than low populated industrial areas, suggesting that residential population might be a good proxy for job openings only at large spatial scales. The authors thus repeated the previous analysis estimating the number of job openings, $a$, using the total number of travelers arriving in each location, instead of its total population (the total number of arrivals was obtained from the data).

The first model is over-estimating small fluxes (less than 200 travelers) and under-estimating large fluxes (more than 200 travelers). Moreover, the choice of a particular distance metric has a marginal effect on the model’s outcome (see Fig.2.1). The second model’s performances are more accurate and, again, the influence of the three different metrics is marginal (see Fig.2.2).

On the one end the result of these tests indicates that the use of a particular distance metric other than the geographic distance (e.g. road distance or travel time) has a small influence on the radiation model’s estimates. On the other end, a careful choice of the method to estimate the number of job openings has an dramatic effect on the model’s performance. In order to enhance the radiation model’s prediction we plan to further study the optimal way to estimate the number of job openings in each location using, for example, more detailed land use information.

### 2.2 Heterogeneity of Human Mobility

Recently, thanks to the huge social microscope provided by Big Data, researchers discovered that traditional mobility models adapted from the observation of particles or animals (such as Brownian motion and Lévy-flights) [20, 49], and recently from the observation of dollar bills [9], are not suitable to describe people’s movements. Indeed, at a global scale humans are characterized by a huge heterogeneity in the way they move, since a Pareto-like curve was observed in the distribution of the characteristic distance traveled by users, the so called radius of gyration [29, 45]. This means that most of us live our mobility within a very small circle of few miles, moving back and forth with high regularity among several nearby locations. This highly localized majority coexists with many people who move routinely dozens of miles each day, and a fewer individuals, but still many, who travel routinely more than hundreds of miles. A considerable presence of outliers emerges from the observation of human mobility, suggesting that the distances we cover in our daily mobility are not comparable. We do not live in a world where Poisson or Gaussian distributions describe our travels, but in a planet where a great variability characterizes our movements. Despite the observed heterogeneity in
people’s movements, by observing their past mobility history the whereabouts of most individuals can be predicted with a very high accuracy, greater than 80% [18, 58]. Moreover, when it comes to the predictability of our actions, outliers disappear and power laws are replaced by Gaussians: regardless of the age, people covering hundreds of miles each day are just as predictable as people covering only few [58].

Although these discoveries have doubtless shed light on interesting and fascinating aspects about human mobility, the origin of the observed patterns still remains unclear: Why do we move so differently? What are the factors that shape our mobility? How does the observed heterogeneity in human movements relate to that found in other areas of society, like social networks and the distribution of income?

To address the above questions we exploited the access to two GSM and GPS
datasets representing mobility information of individuals. The GSM dataset was collected by a European mobile phone carrier for billing and operational purposes and contains date, time and coordinates of the phone tower routing the communication for each call and text message sent by 67,000 customers, in a period of 3 months. The GPS dataset stores information of 40,000 cars who performed travels during one month (May 2011) in an area corresponding to central Italy. For each user, we computed several individual mobility measures. The radius of gyration of a user is the characteristic traveled distance, a measure of how far she is from her center of mass. In mathematical terms, it is defined as the root mean square distance of the locations from the center of mass. The k-radius of gyration of a user in the radius of gyration computed taking into account only the k most frequent location of a user. Starting from these two ingredients, we started investigating the role of frequent locations in the value of the radius of gyration. From the correla-
tion between the radius and the k-radius two profiles of users immediately emerge: while for a fraction of abscissa users we need to observe the entire set of visited locations to determine their radius, the characteristic circle of diagonal users is mainly determined by the first k most frequent locations only. In particular, diagonal users are characterized by k well separated locations, acting as planets around which other minor satellite locations gravitate. Conversely, abscissa users show a number of planets higher than k. In order to identify those meaningful groups of locations, which we called mobility hearts, we used a DBSCAN density-based clustering algorithm, using different values of the eps and minPts parameters. After the clustering procedure, we computed for each user her cluster-radius, that is the radius computed taking into account only the most frequent locations of the clusters detected by the algorithm. The scatterplot of radius vs cluster-radius revealed an interesting finding: abscissa users disappear, meaning that we can characterize the radius of each user only by observing the dominant location in their mobility hearts. Moreover, the distribution of the radius of gyration does not show a huge heterogeneity anymore: it is more similar to an exponential distribution, having hence a characteristic value.

2.3 Diffusion of Online Service Innovations

One of the objectives of the DATASIM project is the study of advantages for society in terms of adoption of the new technology of electric cars. This exploration is pursued by means of the understanding of many social behaviors. For example, car pooling represent a case study that is studied by many partners of the consortium. In general, we can think of car pooling adoption or electric vehicles spreading as a case of diffusion of innovation. Propagation of innovations can be interpreted as a social contagion phenomenon [8, 61, 5, 43], evolving in a social network and being driven by the entanglement of individuals’ decision-making processes [48], as well as by the influence of media and social interactions between individuals. Although network effects in contagion processes have recently been shown to be crucial [7], our knowledge about the social network itself is rather limited since its structure and dynamics usually remain hidden. In this respect the online age has opened up unprecedented opportunities as social network services like Facebook and MySpace, or VoIP services such as Skype and Google Talk, record detailed information about the connections and activities of their users. These services not only partially decode the structure of the underlying social network, but also provide accurate records on the adoption behavior of users. To understand the motivations behind diffusion of innovation, in [36] BME tackle the problem of analyzing this model in the case study of online VoIP service adoption. The authors
Figure 2.3: **Sample of the aggregated Skype network of Switzerland.** A snowball sample of four steps, where product users and confirmed links between them are shown. Nodes are colored according to their adoption state: grey for future users, orange for current adopters, and purple for terminated accounts (corresponding to states $S$, $A$ and $R$ in our compartmental model). B shows the pure adoption, C the pure termination networks.

Opening a user account constitutes a decision-making process that may be triggered by the influence of media or by the social environment[64, 5]. On the other hand, a user may terminate an account for several reasons including vanishing demand or dissatisfaction, by switching to another product permanently, or by simply abandoning the service with a chance of re-adoption (e.g. due to loss of password or intention for lower monitorability). Some of these finer processes are, indeed, accessible by investigating the data. More can be understood by further decomposing this structure into adoption and termination networks. In the former, some nodes appear disconnected and indicate individuals that have adopted Skype prior their friends, an action referred to as *spontaneous* adoption (where individual factors and media influence play a role). Alternatively, at the time of adoption many nodes are already connected with existing users, a common pattern at later stages of dynamics. This scenario indicates influence of the social environment and we refer to it as *peer-pressure* adoption. In contrast, the termination network consists mostly of disconnected nodes at all times suggesting a negligible influence of social interactions on the termination process.

The spreading of the online service is determined by the competing processes of adoption and termination, as indicated by the evolution of the corresponding rates $R_a(t)$ and $R_t(t)$ (thin green and blue curves in Fig.2.4A, respectively). These
simple rate functions already disclose interesting features of the adoption dynamics, since their overall growth signals continuously accelerating adoption and termination processes. The net adoption rate \( R_n(t) = R_a(t) - R_t(t) \) characterizes the real evolution of the spread of the service (thin red curve in Fig.2.4A).

An analysis of the evolving network structure around a given individual can help us detecting whether the ego adopted/terminated the product before any of the neighbors did, or else followed the previous decisions made by a fraction of them. In this way we may label the performed action as either spontaneous or driven by peer pressure. Our measurements show that after an initial, transient period, the probability of spontaneous adoption \( p_a(t) \) (orange curve in Fig.2.4B) and the probability of overall termination \( p^-(t) = p^-_a(t) + p^-_p(t) \) (blue curve in Fig.2.4D) become constant apart from small fluctuations. This tendency also holds for the probabilities of spontaneous \( [p^-_a(t)] \) and peer-pressure \( [p^-_p(t)] \) termination separately, as inferred from the yellow and orange curves in Fig.2.4D. The time invariance of these probabilities, an evident assumption for most biological epidemics, has never before been empirically shown in the case of social contagion phenomena, despite its wide use [32, 63, 10]. Our results thus validate an important assumption used in the conventional modeling of social spreading processes, where probabilities analogous to the ones described here are treated like constants at the outset.

The analogy between epidemic spreading and social contagion has been widely used to model diffusion processes in a society [28, 17, 32, 34]. Here we take this approach to build a compartmental model based on the identified mechanisms behind Skype usage, aimed at a generic description of the large-scale adoption dynamics of technological innovations. We depict individuals as agents in one of three non-overlapping states, susceptible (S), adopter (A) and removed (R), respectively describing people who may adopt the product later, are users already, and will never use it again. Using the data and a theory based on rate equations we are able to describe the empirical observations about the spreading of the online service innovation. Fitting parameters of the model to the first 5 years, we are able to make medium term prediction about the adoption dynamics. We could establish correlations between the speed of adoption and the economics and liberty status in different countries.
Figure 2.4: **Empirical rates and probabilities.** (A) Thin curves denote empirical rates of adoption (green), termination (blue), and net adoption (red), while symbols are their corresponding binned values. Note that a binned data point between $2T$ and $3T$ was removed due to systematic bias in $R_t(t)$. A shaded (white) area indicates the training (predicted) period for the theoretical fit of our model, drawn as thick lines with the same colors as the empirical rates. (B) Probabilities of spontaneous $[p_a(t)]$ and peer-pressure $[p_p(t)]$ adoption per unit time (orange and brown lines, respectively). (C) Conditional probability of adoption $p(n)$, as a function of the fraction of adopting neighbors $n$. Error bars indicate standard deviation and the orange line is a linear fit. (D) Probabilities of overall termination $[p^-(t)]$, and of spontaneous $[p_a^-(t)]$ or peer-pressure $[p_p^-(t)]$ termination per unit time (blue, yellow and orange curves, respectively). The inset depicts a zoom from time $2T$ onwards. In panels (A, B, D) $T$, $r$ and $c$ are arbitrary linear scaling constants, with time dimensions for $T$. In panels (B, D) the black lines are fitted constants.
Chapter 3

Semantic Enrichment of User Activities

The objective of this task is to provide a set of techniques to associate semantic information to mobility data sensed from different data sources: GPS, GSM, and Social web data. Semantic enrichment may be pursued with different levels of human interaction involvement. We present in this chapter two main techniques of semantic extraction from movement data, using user-assisted visual analytics frameworks (Section 3.1) and automated methods based on machine learning and data mining tools (Section 3.2). We also introduce a third analytical problem where the focus is the quantification of a population according to their mobility habits (Section 3.3).

The problem on annotating movements consists in creating a system capable of assigning a label, i.e. a semantics, to a move on the basis of some properties of either the movement itself, or the individual, or the information from the context. The vast literature on the subject may be organized according to the type of movement observed. In fact, many works focus on the movement of individuals to recognize gestures, indoor activities (like cooking or diving) [38, 6], physical activity levels (i.e. walking, running, stairing), surveillance, outlier and intrusion detection. We are mainly interested in those works considering the movement as the physical change of position of the individual, thus leaving a geographical place to arrive to another one (thus we are not considering indoor movement tracking). We can identify two large groups of inference methods: supervised and unsupervised. Among the supervised approaches, there are methods that try to infer the mode of transportation [66, 50], the activity performed in a specific location [39, 65, 11] and the combination of the twos [41]. The methods that deals with the transportation mode try to infer if the user is moving by foot, by car, by bike or by public transportation. This annotation exploits several features of the movements, like the
speed, the acceleration and, when available, other context data like accelerometer measurements. The learning approaches are based on discriminative methods, like decision trees [50] and conditional random fields [41, 40], and generative ones, like Hidden Markov Models. In our context we are not interested in the transportation mode of the individual, since our focus is the prediction of the activity at destination. However, the learning approaches applied to this problem can also be generalized to our classification task since they are based on similar solutions. When considering the activity from movement, however, we can distinguish two main approaches: sequence learning approaches consider the activity of a user as a sequence in a fixed temporal period (usually one day) and try to predict the labels for the whole sequence [41]; episode learning approaches try to label each single movement episode independently from the others [50].

Unsupervised methods are mainly based on clustering techniques [35, 19, 25] or dimensionality reduction [35, 23]. In the paper [35] the authors analyze an activity-based travel survey conducted in the Chicago metropolitan area with the aim of exploring the daily activity structure and the clusters of behavior of people. They describe how the considered population can be clustered into eight (weekdays) or seven (weekends) representative groups according to the activities performed by the individuals.

3.1 Human Assisted Semantic Enrichment

Movement data by themselves are semantically poor. To annotate such data with semantics we can exploit visual analytics methods to tag raw mobility data. Such approach, however, is feasible only for small dataset but it provides very reliable semantic information, that can be used for automatic learning approaches as presented in Section 3.2. In this section we present a selection of approaches developed by the consortium to annotate mobility data with semantics.

3.1.1 Finding Significant Places

The process of assigning annotation to movement data involves also the extraction of relevant places visited by individuals. In [2], Fraunhofer and CNR extend the approach in [1] to implement a visual analytics procedure to assist the analyst to identify and analyze relevant places. The places of interest are selected according to a combination of spatiotemporal constraints provided by the analyst and to a two-stages cluster analysis. At the first stage, spatiotemporal clusters of mobility events are found, that is, the event are clustered according to their positions in space and time. As a result, occasional events are removed from the analysis. At
the second stage, spatial clustering is applied only to the subset of events that belong in the spatiotemporal clusters. This stage unites the spatiotemporal clusters having the same or close position in space. To provide a scalable implementation of the procedure capable of working with limited amount of RAM, the authors propose a clustering analytical process that extends traditional density-based clustering methods [22, 4]. The algorithm exploits the spatiotemporal constraints on the distance function to determine if, at a given time, one item can be disposed from main memory. When each event \(e\) is represented as a multidimensional point \(\langle c_1^e, c_2^e, \ldots, c_M^e \rangle\), the spatial threshold for neighborhood can be seen as a vector of thresholds \(\langle D_1, D_2, \ldots, D_M \rangle\). If any two events \(e_1\) and \(e_2\) have a local distance between the two components \(c_1^{e_1}\) and \(c_1^{e_2}\) greater than \(D_1\) then the overall distance between the two events is set to infinity. The selection of a neighborhood is done by selecting all the events in the dataset with distance within the given threshold. The novel proposal consists in executing a preprocessing step where for each event \(e\) it is selected the list of objects with a distance that is compatible with the distance thresholds. This is done by defining a relevant zone (RZ) of event \(e\) for each dimension \(c_i\) as the interval \([c_i^e - D_i, c_i^e + D_i]\), where \(c_i^e\) is the value of dimension \(c_i\) for the event \(e\). A Lower Relevant Zone (LRZ) denotes the lower half interval of a relevant zone, i.e. the interval \([c_i^e - D_i, c_i^e]\). By definition of the distance function, fixed a dimension \(i\), two events cannot be neighbors if one of them is not contained within the RZ of the other. To exploiting this properties, the events are sorted according to one of the dimensions available and then they are scanned in this order. When two events are ordered, i.e. \(e_1 < e_2\), then \(e_1\) and \(e_2\) may be neighbors only if \(e_1\) is contained in the LRZ of \(e_2\). This property is used to unload events from memory. At each step a new event is loaded in memory and its LRZ is evaluated. An event is disposed from main memory when none of the LRZ of the objects currently loaded contains it. The work of this algorithm is illustrated in Figure 3.1. Events are represented by points in two dimensions. The neighborhood areas of the events are represented by circles around the points. The horizontal dimension (x) is used for ordering and defining the relevant zones. The LRZ of the currently processed event is marked in each step by a gray-shaded rectangle. For events 1 and 2, the LRZs contain no other events. For event 3, the LRZ contain event 2; however, the distance between 2 and 3 exceeds the threshold, i.e., these events are not neighbors. Event 4 has events 2 and 3 in its LRZ and its distances to these events are below the threshold. Hence, the lists of neighbors of all three events 2, 3, and 4 are updated. The lists of events 2 and 3 include 4 and the list of event 4 includes 2 and 3. The approach has a complexity of \(\Theta(n)\), where \(n\) is the number of events to cluster, when the dataset is already sorted.

The effectiveness of the approach is measured by evaluating the number of instances loaded into memory during the algorithm execution. The variation of the
Figure 3.1: Illustration of event pre-clustering. The events, represented as two-dimensional points, are ordered according to the x dimension. The neighborhood areas are represented by circles. The lower relevant zone in each step is marked with a gray rectangle.

RZ sizes in the course of performing the algorithm using different event orderings is presented graphically in Figure 3.2.

3.1.2 Annotating Trajectories with DayTag

The study of the mobility behavior of people in urban areas is essential in any transportation management and planning scenario. For this reason, traditionally this information is collected by means of paper-and-pencil surveys, that are filled in by a limited number of selected volunteers. These surveys detail a typical day of a citizen moving in a city, thus reporting the main trips including the location and time of the daily activities (go to work, go shopping, etc).

Diaries manually collected are semantically very rich and can be very useful for mobility data analysis, since the user may express specific activities he/she performed at the stopped locations. However, this kind of collection lacks in spatial and temporal accuracy, relying only on the user memory. Also, people tend not to report small movements like stopping at the ATM to get cash or at a coffee shop for a coffee, just limiting the reporting to the main activities. Furthermore, these diaries usually represent one specific day and not a typical behavior of the individual over a longer period. Moreover, they miss several spatio-temporal essential information such as the georeferenced location where the user stopped, the route covered during the movement, the duration of the single trips and the single stops. On the other hand, collecting diaries by semantically annotating the traces of individuals auto-
matically gathered by GPS-enabled devices offers a cheap and easy way to collect accurate spatio-temporal information. Almost any modern smartphone can be used as a GPS tracker, thus getting a precise location of the movements. However, the downside is that the manual annotation of activities from a GPS track is a burdensome work and it can cause errors because of user mistaking the annotations.

Since the GPS track collections are becoming more and more common and useful in several application domains, we propose a tool, called DayTag, designed as a personal assistant to help an individual to annotate GPS trajectories with activities, thus reconstructing her/his daily diary. The GPS logs are analyzed to extract the visited location (the stops) together with the background geographical information. The system offers a graphical interface by automatically highlighting the relevant visited locations and the time spent during the visit. The user interacts with the software to visualize, correct, complete and annotate her/his trajectories. This results in a simple and quick way to get reliable diaries.

In DayTag we support the annotation of both stops and moves: the user can annotate each segment specifying the following start and end date and time of the stop (or the move), the purpose for the movement - or activity, the transportation mean and the weather conditions.

The annotation of this attributes using a state-of-the-art GPS data viewer is not a simple task for a user. Let us consider as an example the traces of a volunteer during a day depicted in Figure 3.3. We can see that while the spatial component is quite clear the other annotation attributes of the diary are still difficult to be grasped.

Figure 3.2: The graphs depict the variation of the RZ sizes in the course of scanning of the list of low speed events in Milan ordered according to longitude (red), latitude (blue), time (green), and direction (brown). Top: the Y-axis is linear. Bottom: the Y-axis is logarithmic.
Figure 3.3: Visualization of a raw trajectory collected by a smartphone in the area of Pisa, with no annotations. Screenshot of the annotation file to be filled in manually by the user. The link between the annotation information and the GPS track have to be done manually.

For example, how much time does it take for the user to move from one stop to the successive one? A manual annotation procedure may include the collection of the user annotations in a separate file and then manually join these annotation with the spatio-temporal data coming from the GPS device.

With the purpose of supporting the user in annotating her/his own diary from the GPS tracks, we developed DayTag offering a visual interface and trajectory mining algorithms to compute stop places. DayTag is developed in Java and it exploits available Open Gis Consortium standards to represent and store spatial data on a DBMS. Indeed, the persistence layer is based on the PostgreSQL with PostGIS extension [26]. The graphical interface is developed using Swing libraries. The tools allows the user to load the raw GPS tracks and to preprocess them by removing noise points and by automatically detect stops and moves. Once the mobility episodes are extracted, the user can annotate her activities. After the annotation process is concluded, the spatial and temporal data, along with the corresponding annotations, are stored in the DB.

DayTag graphical user interface is composed of two linked displays: the Time Display and the Map Display (see Figure 3.4). The Time Display shows the temporal evolution of the movement, while The Map Display renders the raw GPS points and the visited locations on a map. The Time Display shows the user position with respect to a conceptual reference system of visited locations. To detect such places, the system analyzes the GPS trace of the user to automatically split the sequence of points into moves and stops [59, 44]. When two stops are geographically close to each other they are represented by a unique location. This generalization enables
the user to abstract from the actual GPS coordinates and to focus only on the relevant visited places. The Time Display shows a time line oriented horizontally and, for each location, it shows a distinct axis. The position of the user along time is presented with a linestring along the same location axis, i.e. the user stays in the corresponding location, or with a line connecting two location axes.

3.1.3 Annotation using web data

For spatial planning applications it is important to know the typical movement-related behaviors and needs of the people on a given territory in order to find suitable planning options and predict their possible effects, e.g., by means of simulation. A usual approach is asking a sample of the population to provide their diaries, i.e., reports about the daily activities and travels. The diaries are then generalized to a set of typical mobility and activity profiles, which can be used to generate synthetic populations for simulation of various “what if” scenarios. However, such population surveys are expensive and the resulting data are very limited in their temporal extent (usually each person reports about a single day or a few days). Unlike personal diaries, movement data can be collected automatically by GPS trackers or by recording positions of mobile phones and, hence, are cheap and easy to acquire. However, the data consist of just geographic coordinates and time stamps and lack any semantic information. To make a reasonable alternative to diaries, movement data need to be semantically enriched by attaching meanings to the spatial positions and travels of the people. Semantic information can be derived by comparing the positions with locations of predefined places of inter-
est (POI) [11]. This approach, however, does not uncover personal POIs such as home, work, child’s school or kindergarten, and regularly visited grocery. In the area of human geography, mobile phone calls data are analyzed to find personal POIs and classify some of them as home or work places based on the frequency of the persons calls from each place, their average time of the day, and the standard deviation of the time of the day [1]. There is no attempt to classify other kinds of POIs. Interactive visualizations allow a human analyst to discover and interpret personal POIs and mobility patterns from movement data of one person [2]; however, it is not feasible to do the same for many people as analyst’s time is a very limited and costly resource. In [3], Fraunhofer shows how to extend this visual analytics approach [2] to much larger datasets reflecting movements of many people. For this purpose they define a computational procedure that extracts stops, finds spatial clusters of the stops, which correspond to individual POIs, and creates “temporal signatures” characterizing the temporal distribution of person’s presence in each POI. The signatures are then visualized to enable understanding of place meaning and, hence, classification of the places and trips between them. The approach is demonstrated empirically with three distinct dataset. The first dataset contains 609,241 time-stamped GSM positions (cell tower coordinates) of 67 persons in France for the period of 49 days. The data have been obtained by an active collection procedure: selected people were pinged every 7 minutes. The second dataset contains 81,389 GPS positions of a single person in the USA for the period of 351 days. All these data were provided by volunteers for use in research. Finally, the third dataset contains individual semantically meaningful places from a data set of twitter messages: 163,203 tweets of 2,607 local Twitter users of two-month period (August 8th to October 8th, 2011) from the greater Seattle area in Washington State, USA.

The interactive visual interface for POI exploration, interpretation, and semantic annotation includes two main components: a map and a set of time graphs. The map shows the spatial positions and extents of the POIs. One of the temporal aggregates can be represented by embedded diagrams at the POI locations. Additionally, the stop points and the trajectories can be drawn semi-transparently. The time graphs represent the computed time series.

Data reflecting movements of people, such as GPS or GSM tracks, can be a source of information about mobility behaviors and activities of people. Such information is required for various kinds of spatial planning in the public and business sectors. Movement data by themselves are semantically poor. Meaningful information can be derived by means of interactive visual analysis performed by a human expert; however, this is only possible for data about a small number of people.

Figure 3.5 gives an example of temporal signatures of personal places of one
Figure 3.5: The time graphs show temporal signatures of personal places of one Twitter user. A, B, C: time series of place visits by hours of the day for the work days (A), Saturdays (B), and Sundays (C); D: time series of place visits by days. 

person from her Twitter activity. Four time graphs show the time series of place visits by hours on the work days (A), Saturdays (B), and Sundays (C) and the time series of visits by days (D). The line coloured in blue demonstrates a typical time series for a work place.

To be able to analyze a large set of places without considering the temporal signatures of each place one by one, two approaches are possible: similarity analysis of the time series and clustering of the time series.

Figure 3.6: By means of similarity analysis, places with temporal signatures similar to selected ones have been found. Upper row: likely work places; lower row: likely home places.

For similarity analysis, the analyst selects a place with previously assigned semantic interpretation and uses the distance function to compute the distances between the time series of this place and those of all other places. Then the analyst applies interactive dynamic filtering by the distances and looks at the time graphs to select a subset of sufficiently similar time series. The places these time series belong to can be given the same interpretation as the exemplar place. An illustration is given in Figure 3.6.

Figure 3.7 demonstrates selected results of clustering. The cluster presented in the upper row (197 members) consists mostly of time series characteristic for work places. The cluster in the middle row can be interpreted as a cluster of likely home places (74 members). The cluster in the lower row, probably, includes places of
Figure 3.7: Temporal signatures of personal places have been clustered by similarity. The images show three selected clusters.

stops of public transport (134 places), where people mostly appear in early morning hours of the work days.

3.2 Automatic Semantic Enrichment

In this Section we show how the availability annotated mobility data may enable the development of automatic processes where new raw mobility data may be analyzed and enriched with purpose movement annotation.

3.2.1 Annotation using background knowledge and POIs

The work [14] of CNR is essentially based on an improvement and extension of the work in [60], which in turn is based on the pioneering work of Spaccapietra et al. in [59]. Here authors propose a conceptual model for semantic trajectories. While trajectories are defined as a time-space function that record the changing of the position of an object moving in space during a given time interval, semantic trajectories are defined as sequences of stops (where the moving object stays still during a time interval) and moves (the part of a trajectory where the position of the object changes). The basic assumption behind the notion of stop is that the place where a person stops is of some interest for her/him. Therefore, each stop is somehow associated to a POI. The association between a POI and a trajectory stop is the objective of several approaches, ranging from the simplest (associating the closest like in [44]) to more sophisticated proposals [46]. However, most of the approaches do not explicitly consider the temporal validity of the association (i.e. if the POI exists or it is accessible during the actual stop), neither the probability value associated to each stop-POI pair, nor the concept of activity and the time dimension.
The semantic enrichment process aims at annotating a raw trajectory with a list of activities that a user moving by a vehicle could perform when he stops. With the assumption that the GPS is installed into a vehicle we have to consider that a person needs to park the car (the stop place) and then he/she starts walking to reach the destination place. The enrichment process is done by gathering the environmental information around the stop place and in particular by exploiting the POIs nearby. The semantic enrichment process includes two phases: a start-up and preprocessing phase in which the POIs are collected and integrated, and a second phase where the most probable activities associated to the POIs are identified and used to annotate the stops.

In short, among all the available POIs, some of them are filtered out by using the set of spatio-temporal domain rules provided by the experts. The spatial filter aims at selecting the POIs within a certain spatial range defined by the maximum walking distance a user is willing to travel. The temporal constraints verify instead the temporal compatibility between the stop time and the opening time of the POIs. For the remaining POIs, the probability of “being visited” is computed by using a gravity based function. At the end, for each stop the most probable activity is returned.

Example 3.2.1 shows how the method works.

Example 3.2.1. Let us suppose to have the trip \( tr \) from \( s_1 \) to \( s_2 \) performed on Sunday by the user \( U \). \( U \) stops in \( s_2 \) from 11:50 am to 12:05 am. Let us suppose to have the list of the POIs in the area of interest and the mapping \( \mu \) of the POI category to the Activities (Figure 3.8 (A)). From the POIs in the area of interest, we first select the candidate POIs using the spatio-temporal rules. The spatial constraint, derived by the \( M_{wd} = 500 \) mt (depicted by the blue circle in the example), excludes the POIs too far from the stop \( s_2 \). Then the temporal rules are applied to the remaining POIs (Bank, Dentist, Church and Bar) in order to
verify the temporal compatibility. The Bank and the Dentist are excluded because they are closed on Sunday, while the Church and the Bar are selected because the duration of the stop in $s_2$ is compatible with the Sunday Holy celebration and the Bar is opened almost every days (Figure 3.8 (B)). For these two candidates, the probabilities $P(\text{Church}, s_2)$ and $P(\text{Bar}, s_2)$ of being visited are computed. Exploiting the mapping POI category - Activity, the list of the most probable activities associated to $s_2$ is returned (Figure 3.8 (C)).

This process is outlined in Figure 3.9.

![Figure 3.9: A schema of the semantic enrichment process](image)

3.2.2 Semantic annotation from trip features

The extraction of significant places allows us to reason about the destinations visited by users to reconstruct their activities. One of the method that we propose aims to infer activity types from GPS traces by developing a decision tree-based model [51]. The model only considers as exploratory variables the activity start times and activity durations revealed by raw GPS data. Based on the decision tree classification, a probability distribution and a point prediction model were constructed. The probability matrix describes the estimated probability of each activity type for each class of the input scenarios (i.e. combination of activity start time and activity duration); while the point prediction model selects the activity type
that has the highest probability. Two types of data were collected in 2006 and 2007 in Flanders, Belgium, i.e. traditional activity-travel diary data and GPS data [37]. In the diary survey, the respondents recorded trip information during the course of one week, such as the trip start and end time, trip origin and destination, activity type, transportation mode, and so on. Half of the households were given a GPS-enabled PDA, in which GPS logs such as the longitude, latitude, and the timestamp of a trip were recorded on a second-to-second basis. The information from the part of the diary survey which is not complemented with GPS data is used for model training and validation. The data from the associated GPS survey is used to test the model and to assure the method is applicable when only GPS data is available. The optimal classification tree constructed comprises 18 leaves. Consequently, 18 if-then rules were derived. An accuracy of 74% was achieved when training the tree. The accuracy of the model for the validation set, i.e. 72.5%, shows that over fitting is minimal. When applying the model to the test set, the accuracy was almost 76%. Although the method (i.e. decision trees) is fairly simple and straightforward, the technique solves important problem and is potentially very useful in the domain of annotation. For instance, the models indicate the importance of time information in the semantic enrichment process; it also contributes to future data collection in that it enables researchers to directly infer activity types from activity start time and duration information obtained from GPS data. Because no geographic information is needed, this research can be easily and readily implemented to millions of individual agents as well as transferred to other regions.

**Model input and output**  
The raw GPS string data first undergoes a trip end identification procedure as to determine the trips and the corresponding start and end times of the trips. The trip information is then converted into activity start time and end time data sets, from which the activity durations are also obtained. The activity start time before the first trip of a day and the activity end time after the last trip of a day for each respondent are assumed to be 0:00 AM and 24:00 PM, respectively. Both the activity duration and activity start time are subsequently used as exploratory variables to predict the activity type. The predicted variable is divided into six categories based on the original design of the diary survey, including home, work/school, bring/get, shopping, social visits and leisure. The home activity encapsulates all time spending at home, while the work/school refers to all work or school related activities outside home. The bring/get activity is for picking up/dropping people and the shopping for any goods shopping. Regarding the remaining two activity types, the social visit refers to all visit activities to friends, colleagues or family members and the leisure accommodates all recreational activities such as indoor or outdoor sports, eating or drinking at restaurants, and tour.
The annotation model outputs either the probability of each of these activity types, or a particular activity type which has the highest probability based on point prediction method.

### 3.2.3 Semantic annotation from GSM data

A more advanced annotation method has been developed for mobile phone call location data then the previous approach for GPS data. This approach [42] explores an integration between machine learning algorithms and the characteristics of underlying activity-travel behavior which originates the call location traces. As a result of the nature of how activity and related travel decisions are made in daily life, human activity-travel behavior exhibits a high degree of spatial and temporal regularities as well as sequential ordering. This study is to investigate to what extent the behavioral routines could reveal the activities being performed at mobile phone call locations. The annotation process consists of four steps. First, a set of comprehensive temporal variables characterizing each call location are defined. Feature selection techniques are then applied to choose the most effective variables in the second step. Next, a set of state-of-the-art machine learning algorithms including Support Vector Machines, Logistic Regression, Decision Trees and Random Forests are employed to build classification models. Alongside, an ensemble of the results of the above models is also tested. Finally, the inference performance is further enhanced by a post-processing algorithm. Using call location data collected from 80 peoples real life over more than a year, we evaluated this approach via a set of extensive experiments. Based on the ensemble of the individual classifiers, we achieved prediction accuracy of 69.7%. Furthermore, using the post processing algorithm, the performance obtained a 7% improvement. The experiment results indicate the feasibility to infer activities using information drawn from calling locations; they also demonstrate the importance of the integration between regular machine learning algorithms and the characteristics of underlying activity-travel behavior when annotating the massive movement data. The advantage of using this approach is that it does not depend on additional sensor data and geographic details, the data collection cost is low and the results are generic to be deployed to other areas. Both the methodology and data requirement needed to apply this method are fairly simple. In the world where simple phones are still prevalent which account for nearly 85-90% of total global handsets in use today especially in developing or under-developing countries, this research has undoubtedly important contributions to the semantic interpretation of the massive location data. In addition, as the big data sets on individuals movement behavior continuously expand and researches on the massive data increasingly gather pace, concerns over privacy issues have also been growing. An annotation approach, which is independent of
precisely geometric positions of an individual and a detailed map, like the method proposed in this study, would be preferable in terms of reducing privacy worries, and can thus be recommended as one of the potential solutions to addressing this issue when annotating the data.

**Model input and output**  The input for this annotation approach is a set of mobile phone call data, consisting of full mobile communication patterns of users over a certain period of time, recording the location and time when each user conducts a call activity. Specifically, for each user, the information includes the base station (cell ID) where the user is located, the day, time and duration of the call activity, the type of the call activity i.e. voice call and message, and the direction of the information flow, i.e. incoming, outgoing and missed calls for voice call and incoming and outgoing for message. The model output is activity types, which are clustered into the same categories as the previously adopted classification for GPS data annotation. However, due to the small labeled training set, the activities of bring/get and shopping are combined into one category of non-work obligatory, as these two activities are expected to subject to a similar level of spatial and temporal constraints. As a result, all the locations in the training set are classified into 5 activity types, including home, work/school, non-work obligatory, social visit and leisure. The annotation model then predicts the possible activity out of all these activity types, given a call location and its relevant information.

3.2.4 **Annotation using trip and activities features**

Neural Networks constitute a methodology that presents special interest in the solution of problems related to semantic trajectories. Lint et. al. (2005) used recurrent neural networks to predict travel time, which provide information that can reduce the traffic congestion. Hu et. al. (2004) used Kohonen neural networks to predict the trajectory class of a moving object. In (Petzold et al., 2005) the performance of the prediction of the in-door next location was investigated using Bayesian Neural networks, while also comparing with several prediction methods. In (Kumar & Venkataram, 2002), the prediction of the future movement of a mobile host depending on its movement pattern history was proposed, using MLP networks. Ju et. al. (2004) proposed a new approach for evaluating a real-time trajectory of a moving object, while the object position is obtained from the image data of a camera, using Kohonen neural networks.

Depending on the way that the neurons are interconnected we distinguish several neural network architectures, the most popular being:

- Multi-Layer Perceptron (MLP) networks
Radial Basis Function (RBF) networks.

The procedure during which the neural networks learn the relation between the input and the output variables is called training. Training or Learning is a fundamental capability of neural networks, allowing them to learn from their environment and improve their behavior. Training refers to the process of achieving a desired behavior by updating the synaptic weights. The most popular method for training MLP networks is the back propagation algorithm. However, in this work, the MLPs were trained by the Levenberg - Marquardt algorithm (Hagan & Menhaj, 1994), which combines the advantages of the back-propagation method and the Gauss-Newton technique.

The RBF networks form a special neural network architecture which is characterized by two main advantages: The simple structure and the fast training algorithms. Figure 3.10 shows the RBF architecture structure. In addition, the RBF networks training stage is fast because it is divided into two stages. In the first stage the calculation of hidden node centers locations are calculated and the most

---

Figure 3.10: RBF Network Structure

The RBF networks form a special neural network architecture which is characterized by two main advantages: The simple structure and the fast training algorithms. Figure 3.10 shows the RBF architecture structure. In addition, the RBF networks training stage is fast because it is divided into two stages. In the first stage the calculation of hidden node centers locations are calculated and the most
popular method is the k-means algorithm means (Darken & Moody, 1990, Macqueen, 1967, Moody & Darken, 1989). In the second stage the network weighted connections are calculated by using linear regression.

In [12] UPRC has adopted neural networks for predicting:

- i. the future position category
- ii. the activity type
- iii. the POI category

by using moving objects data, collected from Flanders, Belgium. These data were collected using two sources, namely travel diaries and GPS-enabled PDAs. The available data were split randomly into three different independent datasets: training dataset, validation dataset and testing dataset. The training dataset was composed of 50% of the entire dataset and the three minimum and three maximum values for each input. The validation and the testing datasets are composed of the 25% of the entire dataset. The training dataset was used for tuning the models parameters, while the validation dataset was used for selecting the best model. The testing dataset was an independent dataset that was used only to evaluate the model. In this work, the authors trained two different neural networks models for each case. The RBF networks were trained by using the Fuzzy Means algorithm and the modified Thin Plate Spline as the activation function. The MLP networks were trained by using the Levenberg-Marquardt algorithm. Also, the MLPs architecture consists of two hidden layers modified. In order to find the best model, all possible combinations of nodes were examined.

The approach has been tested on three case studies. In the first case, the future position category was predicted by using as inputs the start location category, the activity duration and the activity start time. The location categories (start locations and future/end locations) were generated by processing longitude and latitude data using the k-means algorithm. In the second case, the activity type (14 classes) was predicted by using as inputs the activity duration and the activity start time. In the third case, the POI category (4 classes) was predicted by using as inputs the activity type and the episode sequence.

3.2.5 Annotation using Individual Mobility Networks

In [] CNR tries to understand the dualism between systematic versus non-systematic mobility behaviours because of the different impact on the mobility demand of a territory. To this aim, an important aspect is the definition of a realistic data-driven model capable of capturing and summarizing concisely the main features that characterize the mobility habits of an individual: what are the locations around which
the mobility of a person gravitates? What are the routes and the locations that dominate her movements? In other words, how does she see and live her mobility space?

A novel challenge is to study the interplay between the individual behaviors and collective behaviors. The first are obtained by applying mining methods to the histories of the individuals, the latter are obtained by mining the overall collection of individual data. Both methods yield to useful analyses with rather different objectives. From this perspective new question arises: may collective behavior be useful for increasing ones self-knowledge? is there a new concept of personal data management that ethically may support such interplay? In this paper the authors show how the combination of data mining and network analytics supports a novel personal data management store tailored on the analysis of individual and collective mobility. Such representations incorporate the spatial and temporal dimensions of individual mobility patterns and habilitate high level learning and reasoning, independently from specific space and specific time.

In this context the authors propose a data mining-based methodology to obtain a high level knowledge of individual mobility out of raw GPS traces that abstracts from the physical geography and synthesizes the individual mobility habits, by summarizing the salient features of movements. The method is composed by two elements:

- the abstract representation of the individual mobility as the network of the places visited by the individual, the trips that connect such places equipped with three class of measures: centrality, predictability and hubbiness aimed at capturing the basic characteristics of the individual mobility;
- a inference process that extracts the individual mobility network by mining the history of the movements of each single individual from GPS trajectories.

**Individual Mobility Networks** (IMNs) are a first approximation of mobility diaries in their full form, providing information about how the movements and the visited places are structured, allowing also to perform a classification of places that appear to be important for the individual, such as (most likely) dwelling and work place, but also all the locations which are occasional for such individual.

The IMN synthesizes the mobility history of a user \( u \) in a compact model, whose basic bricks are a representation of the visited locations and the description of the trips between any two of these locations.

Formally, an **individual mobility network** of a user \( u \) is a weighted directed graph \( G_u = (V, E, \omega, \tau) \) where: \( V \) is the set of nodes, which identifies the locations visited by \( u \); \( E \) is the set of edges, each of whom describes a connection between two nodes; \( \omega : E \rightarrow \mathbb{N} \) is a function that given an edge returns its weight
(i.e. the number of travels performed by $u$ on that edge); $\tau : V \rightarrow \mathbb{N}$ is a function that given a location returns the time spent in it.

The information regarding the frequency of the movements are described as edge attributes through the $\omega$ function, while the time spent in each location is modeled as a node’s attribute accessible through the $\tau$ function. In particular, during the construction of the IMN we can dynamically extract the aggregated time $\tau(a)$ spent in a location $a$ (i.e. the sum of all the rest times of the individual in that location), and the aggregate number of travels $\omega(e)$ that involve an edge $e$. For example, in Figure 3.11 the edge $e = (c, a)$ with weight $\omega(e) = \omega(c, a) = 5$ indicates that the user moved five times from location $c$ to location $a$, while $\tau(a) = 8$ tell us that the user stops for a total of 8 time units in that location. As we can notice from the graph, in a mobility network some locations could be wells (no travels start from it), due to the limited period of the observation of the dataset.

The model is put at work on the analytical task of automatic activity extraction. The authors use individual mobility networks to infer information on the activity associated to trips (edges of the network). The IMN model does not provide a complete order of the trips, since it is not possible to link each location with the previous one according to the edge connecting both nodes. However, it is possible to extract structural dependencies by exploiting the topological properties of the IMN graph. The authors identify three classes of features to describe the inherent structure of the IMN components: centrality, predictability, and hubbiness. Such features are borrowed from network analytics methods. The problem of activity recognition is usually tackled into two step: first a classifier is learned from a set of annotated data items; then the classifier is applied to new data items when they arrives. In this work the authors exploit the network features extracted from the IMNs to learn a classifier from the supervised data. For the experiments the authors used a large dataset of around 150k vehicles moving in Tuscany during May.
2011. A small part of these movements were annotated by volunteers to reconstruct their activities during this period. In particular, 35 vehicles’ movements were annotated, for a total of 4099 distinct trips associated with behavioral tagging. The annotation was performed by using 13 distinct activities: going home, working, daily shopping, shopping, social activities, leisure, services, refuelling, education and training, bring and get, touring, other.

3.3 Semantic Quantification of User Profiles

3.3.1 Segmenting population with GSM Call Habits

In [27], CNR presents a strategy for mobility behavior identification based on aggregated calling profiles of mobile phone users. This compact representation of the user call profiles is the input of the mining algorithm for automatically classifying various kinds of mobility behavior.

The segmentation of the population with GSM Call Habits is a strategy for mobility behavior identification based on aggregated calling profiles of mobile phone users. This compact representation of the user call profiles is the input of the mining algorithm for automatically classifying various kinds of mobility behavior. In particular, the authors focus on the identification of Residents, Commuters, Visitors and in the discovery of new profiles that may emerge from the data. The analysis process is synthesized in Figure 3.12.

Figure 3.12: The analysis process to infer the user categories from GSM data.

To an abstract level, the process foresees two phases: a **bootstrap** (shown at the top of the figure) and a **running phase** (shown at the bottom). The bootstrap phase is aimed at finding a temporal profile of the calls for the users, and is based on the analysis of the raw GSM data. The running phase starts with the temporal...
calls profiles and performs inferences to extract the user categories. The distinction between bootstrap and running phase is particularly important since we can decouple the analysis of GSM raw data - a huge dataset of limited use by privacy and storage issues - with the calls profiles needed to infer the user categories. In other words, the bootstrap phase can be done only once to establish the the calls profiles needed to infer the user categories. Once the format of these aggregated data has been chosen and tested in the bootstrap phase, the running phase iterates the process, but starting from these temporal calling profiles instead of the GSM raw data. The importance of having established the calls temporal profiles is that in the running phase, the data provider limit the data supply to these aggregated profiles, thus avoiding the transfer to the analyst of raw big data. The running phase can therefore be started with new temporal call profiles every time we need to refresh the dataset for a continuing monitoring of the area.

During the bootstrap phase we have a first computation step called \textit{Spatio-temporal selection and aggregation}. Here, the CDRs are first analyzed and some statistics are calculated to learn about the data and get some aggregate information. These statistics are useful for the choices of parameters and thresholds in the next steps. Then, starting from the CDRs, the call behavior of each user is modeled in terms of space (where a call started and where it is terminated) and time (when the call happens and its duration) to define some user’s \textit{Calling Profiles}. Different calling profiles can be computed at this stage, each of them with different degree of complexity and useful for different kinds of users categorizations. When we have to combine several temporal levels (e.g. hour of the day and day of the week) with the spatial constraints, a complex profile has to be built by performing several temporal projections over the CDRs.

The \textit{Multi-Dimensional Call Profile} is a profile that captures the user’s presence during the whole period of observation, splitted into specific temporal slots. In particular, two operations are performed on the original CDR:

1. Aggregation of the days in weekday and weekend slots, and
2. Definition of the time bands representing the interesting time windows during the day. For example we can have:

   - $t_1 = [00:00:00 - 07:59:59]$, Early in the morning when people are usually at home;
   - $t_2 = [08:00:00 - 18:59:59]$, Mid day when people are out for work/school or other activities;
   - $t_3 = [19:00:00 - 23:59:59]$, Late in the evening and night when people are typically back home.
The Inductive step takes as input the call profiles (which can be simply Temporal, or Spatio-constrained Temporal or Multi-Dimensional (depending of the kind of user categorization we want to define) and, according to their relative similarities, group them into homogeneous groups of similar call profiles. Here we use an implementation of the Self Organizing Map (SOM). A SOM is a kind of neural network based on unsupervised learning that produces a one/two-dimensional representation of the input space using a neighborhood function to preserve the topological properties of the input space. The SOM forms a sort of semantic map where similar samples are mapped close together and dissimilar apart are mapped far away. This produces a set of nodes, where each node represents a group of users with similar profile. We call each node the General User Profile. For example, we can have groups of users whose general profile is represented by a multidimensional profile having calls in the mornings of weekdays and no calls in the week ends.

Figure 3.13 (b) shows an example of general user profile which includes 359 users. The columns are the sequence of weekdays and weekends (for 2 weeks), and the rows identify the temporal bands of the day \( t_1, t_2, t_3 \) as exemplified above.

The automatic classification of the General User Profiles in categories like Residents, Commuters and Visitors is done by using a K-Nearest neighbors algorithm where the training examples are a set of prototypes of these categories. In the classification phase, each General User Profile is assigned with the label corresponding to the most similar prototype. This phase returns a first attempt of automatic classification that can be later modified by the expert by using the system interface.

The quantification step concludes the analysis evaluating the percentage of
users for each category and computing a general statistics.
Chapter 4

Understanding Mobility with Complex Network Analysis

In this Section we address the research challenge of understanding the complex relations among the entities in a mobility scenario. Here we consider this relations in a very broad sense. In particular, we present three distinct approaches. In Section 4.1 we investigate how temporal characteristics influence mobility and its predictability, by exploiting the mobility borders method. The topic of Section 4.2 is the comprehension of mobility of customers in a huge purchase dataset. In Section 4.3 the potential social relationships among individuals are extracted by analyzing their mobility habits.

4.1 Time-dependent Mobility Borders

In the real world, different events may dramatically change how people move on the territory. Such events may be unpredictable or not frequent, like natural disasters, but most of them are not. The most natural regular and predictable event is the transition between working and non-working days. During Saturdays and Sundays, people usually abandon their working mobility routines for different paths, obeying to completely different criteria. Another example may be organized human social events, like manifestations in a particular town or sport events.

In [16], CNR systematically proves that to mine human mobility and to extract from it useful knowledge is necessary to take into account these phase transitions. A dataset of undifferentiated trajectories, without taking into account when they were performed, may lead to increased and unexpected noise effects, lowering the quality of the results and, in extreme cases, hiding interesting patterns. The authors address the temporal dimension problem by analyzing with complex network tech-
niques a set of GPS trajectories and then understand their predictive power of the movements of the observed vehicles over the time span of a month. The complex network analysis technique used is the so called community discovery. Community discovery in complex networks aims to detect a graph’s modular structure, by isolating densely connected sets of nodes called communities. For the temporal dimension, the communities observed at time $t$ are used to predict the communities observed at time $t + 1$. In the proposed framework, the authors generate sets of network with different criteria (temporal and spatial) and then apply community discovery on these networks, following their previous works [54, 15], to identify the borders of human mobility. The focus is to evaluate which temporal perspective and which grid resolution is leading to the best results. Each network results are evaluated both quantitatively, using different quality scores, and qualitatively, by looking at the resulting borders and confronting them with background knowledge on Tuscany mobility.

Figure 4.1: Some statistics for the daily network snapshots.

The author fixed a minimum temporal interval of one day and then generated daily snapshots of the movement graphs. Figure 4.1 shows some of the basic statistics of these daily networks. We can see that there are remarkable differences between weekday and weekend networks (we recall that May 8th, 15th, 22nd and 29th 2011 were Sundays). Saturdays and Sundays networks usually have less edges, somewhere between 62-75% of the edges of a weekday (Figure 4.1(a)); they have more components, i.e. the networks are more fragmented, with areas not connecting at all to each other (Figure 4.1(b)); and finally their average path length is significantly higher, May 8th presents a lower peak, but the whole preceding week was lower than the following, due to the fact of Italian national holiday of May 1st (Figure 4.1(c)).

To evaluate how much the communities discovered in a particular temporal interval are meaningful, authors check if they are preserved in different time periods, by comparing each other by means of the measures of precision and recall.
The general procedure is the following: first consider the clusters extracted in the network representing the first week and then calculate the Precision and the Recall for each of the other networks. A high score means that the target network contains similar clustered information, therefore it is predictable using the source network. The results are depicted in Figure 4.2.

To understand how to read Figure 4.2, let us consider its leftmost scatter plot: in this case the source clustering is calculated using each of the Weekday network. Each dot represent the quality results, according to Precision (x axis) and Recall (y axis), for each of the other network considered in this article. The dot color represent the kind of network to which we are applying the prediction: green for Weekday, blue for Weekend and red for Week. Since we are dealing with four weeks and three different network views for each week (Weekday, Weekend and Week) we have a total of 48 points, 4 of which scores 1 for both Precision and Recall as they are clusterings applied to themselves: since we are considering the leftmost plot, the 4 perfect scores are all green dots, each representing a Weekday clustering applied to itself.

Now we can find evidences about the lower quality of the Weekend predictions by considering all the three plots. As we can see, the central plot, the one representing the prediction results using the Weekend clusters, scores lower performances for all networks, both in Precision and Recall. Not only Weekend clusterings are not able to predict Weekday and Week clustering: they also score poorly in predicting themselves, proving that from one weekend to another the trajectories vary significantly, and therefore they cannot be predicted efficiently using the simple assumption that the same period in the week should behave in the same way across time.
4.2 Purchasing Behavior and Mobility

In the economic literature, market society is considered driven by rationality and the expression of this rationality is the price system. According to this view customers are rational beings: they try to minimize the amount of money they are spending, while at the same time maximizing the amount of goods they are purchasing [33]. Therefore, price is a generic utility function that each customer tries to minimize, and it is the same for everybody. However, customers are also driven by their own personal needs and desires [24]. Many of these needs are shared with other customers, such as the basic needs for survival, but many others are intimately bound to each individual and possibly different from the ones of everybody else. A customer is driven both by a generic utility function (cost minimization) and by a personal utility function (fulfillment of unique desires).

If we are able to quantify the personal utility function for each customer, then we can address a question with repercussions on a seller’s market strategy: which function will win the arms race in influencing the purchase behavior of a customer, the generic one or the personal one? If the generic one is stronger, then a seller is forced to compete mostly on the price; while if a customer’s needs are more important, then it is the quality of the choice that matters the most.

In [47], CNR develops an analytic framework based on mining big customer transaction data, aimed to quantify the strength of both utility functions. The authors test the customer behavior in terms of distance traveled, under the assumption that customers want to minimize their travel length. They observe that customers do not always go to the closest supermarket: there is a range effect for each product, due to the intrinsic characteristics of the product. To explain and predict the range effect it is proposed a method to compare the strength of the generic and the personal utility function in the customer’s mind. This comparison boils down to the question: given that customers travel on average \( x \) meters to buy product \( p \), are they doing that because \( p \) is expensive or because \( p \) satisfies very particular needs?

While the price is an explicit information of the product, the needs the product itself is satisfying are not. Such needs are quantified by evaluating the sophistication of each product and customer, following [31]. It comes out that the sophistication of a product is better than the price in explaining a customer’s behavior.

The authors provide empirical evidence of these claims with real world data about customer behavior, by analyzing digital traces of customer purchases in the database of a large supermarket chain in Italy. The more sophisticated is the need a product satisfies, the longer a customer will travel to purchase it on average, almost regardless of its price. Intuitively, this means that to buy bread people will just settle with the closest shop where it is available, while to buy blank DVDs, with roughly the same price and available in all the supermarkets of the chain,
a customer will travel a significantly longer distance. While the product range concept may be quite intuitive, in this paper we provide a system able to quantify it better than just assuming that it is proportional to the product price.

There are many consequences for sellers from the ability of predicting a product’s range. For instance, to know the range of all the products of a supermarket implies that the supermarket’s marketing strategies can be tailored according to the distance of a customer from the nearby points of sales. Customers far away from a point of sale need to be stimulated on more sophisticated needs, while nearby customers may be more susceptible to more basic needs.

A second application is in point of sale placement, as this methodology can be used in conjunction with the central place theory [13]. Besides the construction costs, each point in the city space is altering the minimum distance between a customer and a product. Therefore, given the range effect, each point in the city space has one optimum in its product assortment. In the paper, the authors provide the proof that this problem can be formally addressed to find a good approximate solution.

The final contribution of this paper is to show how to accurately predict how long a customer will travel (or which shop she will choose) to buy a given product, as a function of the product’s sophistication. In other words, product sophistication reveals as a powerful predictor feature for a challenging predictive task, because most people shop preferably at the closest store for most products, so it is difficult to accurately characterize for which products a customer will travel more. This last points is also very relevant for simulation systems, since their precision and quality is mainly based on accurate prediction of agent’s behavior.

4.3 Big Data and Network Analytics for Boosting Car Pooling

In [30] CNR presents a network analytics approach to unveil hidden information in mobility data: the so called ranking measures. A ranking is a relationship between a set of items such that, for any two items, the first is either “ranked higher than”, “ranked lower than” or “ranked equal to” the second. In mathematics, this is known as a total preorder of objects: it is not necessarily a total order of objects because two different objects can have the same ranking. How ranking measure can be used to discover hidden information in mobility data? If mobility data can be mapped to models such as complex networks, then link analysis can be applied to these networks in order to discover useful information. Using these information, in a car pooling scenario is it possible to say if a user is a good passenger or a good driver, which are the routes most crowded at the same time and who are the best actors for
the proposal of a car pooling service.

The authors built a potential carpooling network starting from GPS dataset of trajectories adopting the procedures explained in [62] to detect inclusive trajectories and to identify who are the users that shares some route. Then the HITS algorithm was applied on the carpooling graph to rank a user how a good passenger or a good driver. In addition, the authors performed community detection algorithms to detect group of users that are highly related in the car pooling sense and we observed how the rank scores are distributed among the communities. Some of these communities with certain characteristics about the ranking measures, could help in finding the nodes that with an high probability are good seeds for proposing a car pooling service. Finally, they propose a novel approach to assign passenger to driver based both on topology measures and on ranking measures, and we show the performances of this algorithm with respect to a random selector.
Chapter 5

Knowledge Transition to Simulator and Future Directions

The general workflow to instantiate an agent-based simulator consists in providing to the simulator a set of annotated diaries of real agents. Such diaries are then synthesized into a set of internal rules that are exploited to generate new agents during the simulation. Examples of such internal rules may be marginal distribution of movement dimensions (i.e. movement length distribution, time of departure per given activity type, speed distribution, etc.). In this context, the methods proposed in Section 3 can be exploited with two approaches. From big data we can assign to raw trajectories of individuals their semantics. When fully annotated, travel diaries can be reconstructed (i.e. estimation of the complete daily activity-travel schedule of individuals) from the data and can be fed into activity-based models (e.g. Feathers) or agent based models. On the other hand, the models and patterns based on semantic information may be used to enhance the internal rules of the simulator, for example by deriving more precise distributions according to the analysis of big data.

The main purpose of this chapter is the description of a system that integrates the proposed annotation methods and constructs activity-travel diaries based on the big data. The description of this Chapter provides an executive plan of the next steps that will be followed in the project to instantiate the new simulator. The actual implementation of this architecture will be pursued during Year 3 of the project within Task 1 of Workpackage 3.
5.1 Agent-based Simulation Process

We imagine a general process divided into two phases: a bootstrap phase where the internal rules of the simulator are learned from the big data; a generative phase where an iterative process is executed for each virtual agent to synthesize the corresponding mobility. In Figure 5.1 we report a schematic visualization of this process.

![Diagram of the agent-based simulation process](image)

Figure 5.1: Abstract pipeline for an agent-based simulator.

During the bootstrap, the available data is analyzed and, eventually, raw data is annotated with activities. From the annotated sequences, a generative model is learned that will be exploited in the successive phase. During the generation phase, for each virtual agent a diary schedule is generated as a sequence of activities to be performed. The planned activities are then mapped to the real geography by assigning position of each destination to the territory. We represent individual mobility as tours among the relevant locations of the user and trips to secondary destinations for minor activities. In the following sections we give some rationale about the actual implementation of these steps.

5.1.1 Big Data Annotation

This step exploits the methods developed in Task 2 for semantic enrichment of raw data. We believe that a combination of the semantic enrichment methods presented in Section 3.2 may provide a reliable set of annotated data to be used for the successive step. Our plan is to combine the strengths of each method to obtain a pipeline capable of characterizing raw movement with activity annotation. For instance, Individual Mobility Networks provides a basic tools to partition the virtual agents into subsets with similar characteristics observed from real data. This step may enable us to implement specific generation phases according to the profile of each user. For example, the mobility of an unemployed agent is more entropic than the
mobility of a person who has a job. According to the IMNs classifier, it is possible to reach high levels of accuracy for the other activities as well. However, such performances may be augmented if new constraints are pushed into the framework. The work presented in [14] provides, for example, a set of constraints to determine the most probable activity according to specific places in the territory, by combining several spatial and temporal constraints on the candidate activities that can be performed in a specific place.

Moreover, the semantic annotation methods may be exploited as generative processes where, given an activity type, they are capable of determining the geographic location where that particular activity may be performed. This feature will be better explained later in Section 5.1.3.

5.1.2 Generative Model Extraction

When synthesizing the mobility of a virtual agent, it is necessary to determine a sequence of activities that it will execute during the simulation period. The reader may think to this step as a word generation with a given alphabet, where the letters are the distinct activities and the resulting words are individual schedules. Given a set of letters (i.e. activities), not all combinations are valid words. Thus it is necessary to have a model capable of generating valid sequences. Clearly, these valid sequences are not necessarily ruled by a finite grammar. On the contrary, all the possible combinations of activities should be first observed and learned from data. When using paper-and-pencil diaries, the repertoire of different behaviors is limited. On the contrary, the observations of many individuals from big data, enable us to sense many different combinations. While the large availability of data produces flexible schedules, it requires new methods to aggregate and summarize such data in order to be able to produce new virtual valid sequences. An example of such a generative model may be the Hidden Markov Models (HMM). The foundation of HMM is the characterization of probability distributions in terms of activities, activity-travel times and activity locations associated with each individual, based on the observed activity-travel behavior of the individual revealed by the big data. The represented HMM is then used to simulate activity-travel patterns for the individual for a certain day.

A HMM is a probabilistic representation that can capture statistical relevant information implicit in a group of related sequences. It is based on the assumption that the probability of a letter in a sequence depends on the occurrence of previous letters. The information extracted from a sequence includes: (i) a sequence of positions, each with its own distribution overall all possible letters; (ii) the possibility for inserting extra letters between consecutive positions.

A HMM based on a daily activity-travel sequence for a working day is de-
signed as follows (Fig. 1). It contains a sequence of positions (match states), corresponding to the anchor points (i.e. home and work) of the activity-travel pattern. The anchor points represent some locations that have an higher relevance for the agent. The match states segment a daily pattern into a sequence of tours. The other activities of the day may be performed within such tours among the relevant anchor points. Each match state is complemented with an insertion state that can add an activity between this state and its next match state, representing the individual trips in the corresponding tour. A insertion state can emit an activity from all possible activities governed by a distinct emission probability distribution. At each of the match and insertion states, other activity attributes can be profiled with corresponding probability distribution, such as activity starting time and duration.

At each match state, two possible transition probabilities are assigned to reflect the likelihood of movement between this state to the next match state as well as between this state and the next insertion state. At each insertion state, however, three possible transition probabilities are anticipated: from this state to the two neighboring match states as well as to the next insertion state.

5.1.3 Mapping to Geography

Once the schedule for a virtual agent has been determined, each activity should be placed on the geography.
In the first step of the simulation pipeline, the anchor points for each activity are determined. The current approach in the literature to determine such distribution is based on the marginal distribution derived from the surveys. However, we showed in Chapter 2 and 3 different methods to derive such information from big data. For example, the Individual Mobility Network model presented in Section 3.2.5 provides a formal and concise way to observe the position of relevant locations in a population. The novel contribution of such approach is the definition of individual points of interest that concerns the individual and private sphere of each agent. In Section 2.1 we showed how the radiation model may be used to determine relative distances of home and work locations by estimating origin destination matrices for distinct activity types. The heterogeneity model (Section 2.2) provides the link between the socio-demographic information of small areas and the mobility habits of that population.

Individual mobility networks allow to model the activities performed in shared places, i.e. in locations where different individuals choose to perform the same activity. For example, a commercial mall is a place where many users go for shopping. Since IMNs model the user mobility in an abstract space, it is necessary to instantiate such location on real geography. The choice and the location of general activities should be related to the facilities available on the territory. With this objective, the constraints defined over the POIs (Section 3.2.1) may be investigate to be used as a generative model to determine the most suitable location for an agent for a given activity purpose.

CNR and IMOB are investigating the exploitation of mobility borders in a semantic context where the mobility networks are projected on specific activity types. The result yielding from mobility borders for general mobility was that people tend to move within basins of mobility that naturally emerge from their mobility. We empirically showed in Section 4.1 that such borders do strongly depend on the temporal dimension, having distinct partitions, for example, in weekdays and weekend. We are now tackling the problem of answering the question: do mobility borders depend on the specific activity type?

Figure 5.3 shows the resulting borders according to two distinct activity type mobility – going home (top) and working (bottom) – in the region of Flanders in Belgium.

Starting from a series of semantic origin destination matrices, where each link counts only the movements that corresponds to the activity type of the matrix, we studied the network characteristics of each matrix and applied the mobility border analysis. From the Figure we can notice how the home borders are mainly located around big urban centers. The eastern and western borders are maintained in both semantic matrices. This suggests that people living there has a work location near their home. In the central part of the region, around Brussells, we can notice the
Figure 5.3: Mobility Borders based on two distinct activity type movements: going home (top) and working (bottom)
differences among the two partitions. This results presents a dual situation of the previous one: working locations are concentrated in close places and people tend to live outside the main city. These results suggests that mobility borders may be used effectively also during the generation of trip destinations, by looking in different border distributions according to the activity type to generate.

In [53], CNR proposes to exploit the decision tree learned from a small dataset of IMNs to a large collection of IMNs derived from a fleet of 80k private cars. The aim of this annotation is twofold: on one hand it allows a mobility manager to reason on the collective activities performed by people; on the other hand it provides each contributed user with the possibility of comparing her mobility habits with those emerging from the collectivity.

![Distribution of predicted activities by time of day](image)

**Figure 5.4: Distribution of predicted activities by time of day**

Figure 5.4 shows the distribution of each activity on a typical day as derived after the annotation of 80k vehicles. We can notice how the most frequent activities, i.e. home and work, present a recognizable pattern with three peaks during the day. It is also interesting to note how such time-dependent activities, like shopping, present a compatible distribution in the day.

Behavioral choices of individuals are driven by their personal objectives dur-
ing the day. It is clear that during working days, the structure of the daily schedule is somewhat constrained to the working activity. Besides, working days provide many other constraints for personal mobility like, for example, the task of bringing children at school. Intuitively, the models describing and generating mobility should be diverse for the two different time periods. In [16] (Section 4.1), CNR extends the previous work [15] and shows empirically from a large proxy of human mobility, how this assumption is also supported by data. This result becomes also relevant when simulating the daily schedules for a synthetic population.

One of the basic assumption behind such different scenarios is the working habits of each individual. To the best of our knowledge, there is no contribution in the literature that provide a proof of this hypothesis and, this is a research challenge to be coped with. However, in transportation science it is largely accepted as a valid rationale to explain the different mobility behaviors. Thus, when simulating a synthetic population, it is crucial to understand also the characteristics of each individual. In other words, it should be modeled the relation between the mobility of a user and her mobility. In [27] (Section 3.3), CNR has studied a framework to automatically analyze human traces to assign a label to each individual to establish if she is resident, commuter or visitor. Even if such method is not directly applicable to our problem, it provides a starting point to investigate the relation between mobility and job status.
Chapter 6

Directions and Future Plans

Many of the activities of this workpackage enable new advances in the other workpackages of the project. We have identified two main groups of tasks. We see the contributions of the WP2 as a novel approach to synthesize human mobility and provide realistic simulations. These simulations may also be driven by specific behavioral modification to witness the consequences of short- and long-period planning. While the development of a new simulation system is still starting, we are also experimenting virtual simulation of synthetic behaviors: we use real mobility sensed by available data to simulate how human mobility is compatible with new behavioral models.

6.1 Generating Synthetic Mobility

The blueprint of the simulator presented in the previous sections has, among the others, a novel and relevant characteristic: it depends strongly on the models and patterns emerging from the big data. This is also one of the main objectives of the workpackage, i.e. providing a strong comprehension of the internal characteristics of human mobility to a new generation of simulators.

As a proof of concept of these results, we plan to create an instance of the simulator according to the contributions presented above. A basic requirement for such experiment is the availability of several data sources within the same territory. As presented in DLV 1.1, within the project consortium we have identified the territory of Tuscany as a valid candidate for this study: many of the available mobility datasets based on GPS and GSM are related to this region. The consortium, moreover, has started collecting new data to have a comprehensive vision of mobility in the region, including some relevant demographic and energy-probe data that at the present are not available. The plan of the WP2 participants for Year 3 is detailed
as follows.

**Data selection and collection**  As basic step to instantiate an agent-based simulation system it is necessary to have a minimum set of data describing the background and the constraints of the territory. In particular, as a very basic layer, we already collected road network data and demographic data. To stress the necessity of replicating the experiment in different geographic regions, we try to use, when possible, public available data. The road network was extracted from the OpenStreetMap project, whereas demographic data is publicly available from the national bureau of statistics (ISTAT) in Italy.

**Collecting Mobility Diaries**  Starting from an instance of one of the existing agent-based simulations, we need to feed it with annotated travel diaries. In the selected area there is no ready availability of survey data. Rather than a limitation, this situation provides us with the challenge of deriving semantically rich diaries from big data. A general description of this approach is outlined in Section 5.1.1. CNR is also implementing a sensing application capable of collecting movement data from volunteers. The annotation of such movements will be performed by each volunteer by means of the DayTag tool [52].

**Data-driven Mobility Models and Patterns**  Once all the data are available, the models and patterns presented in the previous sections will be instantiated and integrated within the simulator. Here the scientific challenges are twofold: on one hand we have to find a balance among the different data sources, trying to mitigate possible conflicts; on the other hand, we can explore how the internal models of the simulator may be overridden by the models derived by big data analytics. We may refer to these two scenarios as blackbox approach and whitebox approach respectively. The blackbox approach considers the simulator as an opaque toolbox to be fed with semantically enrich data: the quality of the output, thus, depends solely on the quality of the input provided. Although this may seems a simplistic approach, it gives us the possibilities of comparing different outcomes by varying the semantic information in input. The techniques and methods for the comparisons are being developed and studied in Work Package 6. The whitebox approach gives us the possibility of enhancing specific steps of the simulation pipeline. The integration of the radiation model presented in Section 2.1 is a preliminary step towards this approach.
6.2 Impact of Electric Vehicles on Mobility

One of the scenarios to be investigated within the project is the impact of electric vehicles on mobility habits of users. CNR and UPM has started to investigate an analytical method to check how much the constrained range of electric vehicle may influence individual mobility. In the current status of work, the real mobility of a large dataset of private cars is analyzed to simulate the coverage of each individual in the case each vehicle should be battery operated. The objective of this activity is twofold: on one hand we plan to obtain a deeper comprehension of human mobility and their compatibility with the new generation of electric vehicles; on the other hand we define an objective process to measure the impact of EV on human mobility. The second approach is crucial when evaluating different simulation scenarios to evaluate how initial policies for improve EV adoption are effective.

6.3 Diffusion of car pooling and electric cars

We have identified two application scenarios where the novel data-driven mobility theory may be fully exploited: adoption of electric vehicles and automatic services to boost car pooling. The success of such innovations strongly depends on the adoption they will receive along time. Exploiting the work on diffusion of innovation from BME [36], we plan to investigate the evolution of these two applications in a simulated population, to explain how the initial condition may influence the resulting success, or failure, of the two scenarios.

6.4 Consolidation of Patterns and Models

We plan to continue the investigation of mobility models and patterns into two directions. First, we will consolidate the existing methods to improve their precision and efficiency. Second, we will continue exploring large mobility data to identify new patterns that are better suited to describe specific characteristics of sub-populations.
Bibliography


