Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN DATA SCIENCE AND COMPUTATION

Ciclo XXXIII

Settore Concorsuale: 01/A4

Settore Scientifico Disciplinare: MAT/07

Big data analytics and modeling for Human Mobility

Presentata da: Chiara Mizzi

Coordinatore Dottorato:
Prof. Andrea Cavalli

Supervisore:
Prof. Armano Bazzani

Esame finale anno 2022

Abstract

The fast development of Information Communication Technologies (ICT) offers new opportunities to realize future smart cities. To understand, manage and forecast the city's behavior, it is necessary the analysis of different kinds of data from the most varied dataset acquisition systems. The aim of this research activity in the framework of Data Science and Complex Systems Physics is to provide stakeholders with new knowledge tools to improve the sustainability of mobility demand in future cities. Under this perspective, the governance of mobility demand generated by large tourist flows is becoming a vital issue for the quality of life in Italian cities' historical centers, which will worsen in the next future due to the continuous globalization process. Another critical theme is sustainable mobility, which aims to reduce private transportation means in the cities and improve multimodal mobility. We analyze the statistical properties of urban mobility of Venice, Rimini, and Bologna by using different datasets provided by companies and local authorities. We develop algorithms and tools for cartography extraction, trips reconstruction, multimodality classification, and mobility simulation. We show the existence of characteristic mobility paths and statistical properties depending on transport means and user's kinds. Finally, we use our results to model and simulate the overall behavior of the cars moving in the Emilia Romagna Region and the pedestrians moving in Venice with software able to replicate in silico the demand for mobility and its dynamic.

Contents

Introduction **6**

1 Datasets and Cities of interest **13**
1.1 Cities like case study 13
1.2 Datasets description 17
1.2.1 The Olivetti dataset (OD) 18
1.2.2 The MTS dataset by Emilia-Romagna Region 19
1.2.3 The Bella Mossa dataset 20
1.2.4 The Octo telematics dataset 21
1.2.5 The MDT dataset by Telecom 21
1.2.6 The Venice dataset 23

2 Implemented Algorithms and Methods **24**
2.1 Algorithms for cartography extraction 24
2.2 Algorithms and techniques of analysis 30
2.2.1 Collection, filtering, and down-sampling 30
2.2.2 Georeferencing and trajectory reconstruction 33
2.2.3 Transport Means Classification 39
2.3 Simulation algorithms and techniques 44

3 Results **52**
3.1 Case study: Venezia 52
3.2 Case study: Rimini 72
3.3 Case study: Bologna 89
3.4 Simulation of car mobility: Emilia Romagna 109

A Additional information on Section 3.1 **122**

Bibliography **132**

Introduction

The fast development of Information Communication Technologies (ICT) offers new opportunities for the realization of future smart cities [1, 2, 68–70]. The ubiquitous diffusion of the technological instruments allowed the collection of large datasets, whose processing and modeling require innovative analytical tools in the framework of big data analytics, which is one of the future challenges of Complexity Science. Various authors have taken advantage of the large georeferenced datasets that have been collected on individuals mobility to study the statistical laws at the base of human movements [3, 6, 7, 10, 11, 54] and the dynamical properties of human mobility in urban contexts [8, 9].

The aim of this research activity in the framework of Data Science and Complex Systems Physics is to provide to stakeholders new knowledge tools to improve the sustainability of mobility demand in future cities [5, 12, 13]. Under this perspective, the governance of mobility demand generated by large tourist flows is becoming a key issue for the quality of life in the historic centers of Italian cities, that will become even worse in the next future due to the continuous globalization process [14–16]. On the one hand, the frailty of historical centers puts at risk the cultural heritage in the presence of big tourist flows; on the other hand, the presence of tourists heavily conditions the daily life of residents. However, the relevance of the tourism economy advises against restriction policies that would limit a priori the tourist presences. The historical center of Venice is a paradigmatic case to study the tourist flows both for the predominantly pedestrian character of the Venetian mobility and the unique features of its monuments that attract large crowds all the year.

The perspective of future smart cities should consider another much-discussed aspect: sustainable mobility. This theme is closely connected with the reduction of the overall impact of emissions on the environment. At the same time,

it seems related to the improvement of life quality of citizens. One of the critical issues for the development of effective policies towards sustainable mobility is to reduce the use of private transportation means in the cities and to improve multimodal mobility [37]. A possible approach requires understanding how the individuals realize the mobility demand in a city when different mobility networks and strategies are available [54–56]. In particular, we refer to the possibility of using various transport means (multimodal mobility) when they are present, both decision mechanisms for the choice of these and the physical interactions in the mobility networks during realization of the mobility paths [38,39].
The decision to use different transport means could depend on the length of the planned trip, on the expected duration, and the perceived convenience of choice, but also on the different individual attitudes [40]. For example, in the case of cycling or pedestrian mobility, the trip length is directly related to the fatigue necessary to perform the trip. In contrast, in the case of car mobility, one has to consider the stress related to driving in traffic conditions, the accessibility of the city to private cars, and the availability of parking places. The planning of individual daily mobility can also influence the choice of the transport means due to the necessity of performing several activities during the same day. The ICTs could provide data on individual mobility on a representative population sample, but it also opens relevant problems for privacy. However, a Statistical Mechanics approach could point out some universal features of the empirical distribution functions that allow us to suggest suitable observables able to explain how individuals use different transport means in a given city [41]. The planning of governance policies could use a complex systems approach that correlates the macroscopic statistical laws of human mobility with some features of individual behavior.
The cities of Rimini and Bologna are optimal cases to study the statistical properties of different mobility modalities and their interactions. Indeed, in a complex system like the road network, some parts will be mainly used by a single modality, and others will be traveled by more than one. It is possible to define different subnetworks and to think of the city as an overlay of them with some contact points, that is, a multiplex network [48–50,67]. Furthermore, Rimini and Bologna do not have a predominant tourist vocation as Venice, and their structure allows the usage of sustainable means of transport like buses or bikes.
If, on the one hand, the study and analysis of different datasets can add information about our knowledge of statistical properties and human behavior, on the other hand, we can use this knowledge to model and simulate the overall behavior of people living in the city. Having this in mind, we developed a software able to replicate in silico the demand for mobility and

its dynamic during the research activity [60]. This software gives the previous analysis the possibility of predicting the near-future activities, and it gives stakeholders an instrument for describing and managing the city. This instrument improves its performances and predictive capabilities if fed by a stream of real-time data, whose implementation needs inevitable cooperation among local administrations, researchers, and companies suppliers of devices and infrastructures. In this context, two important projects are born. The first one is *SCR (Smart Control Room)* involved the University of Bologna, the City of Venice, and multiple companies like Telecom [30], Fabbricadigitale [32], and Venis [31]. The second one is *S.LI. DES (Smart strategies for sustainable tourism in LIvely cultural DEStinations)* [35] that involves different partners, including the University of Bologna and five Italian and Croatian municipalities: Bari, Venice, Ferrara, Dubrovnik, and Sibenik.

This thesis work will focus on studying three cities: Venice, Rimini, and Bologna. They represent three different cases of study, each with its available datasets, analysis techniques, and results.
As regards Venice, we cope with the problem of understanding how the pedestrian flows moved on the road network of the Venice historical center during two different periods: the Carnival of Venice 2017 (from 23-02-2017 up to 02-03-2017) and during the *Festa del Redentore* (from 14-07-2017 up to 16-07-2017). We used two datasets provided by the Italian mobile phone company TIM [30] containing GPS (Global Positioning System) data about a representative sample of mobile devices, with an ID number that changes every 24 hours. The collection of GPS data is possible employing the new technologies currently being developed by NOKIA (Geosynthesis system). The data provide anonymized GPS positions of a device each time certain types of network activities are on. We have introduced restrictive conditions to identify an individual mobility path to reduce errors due to a lack of GPS data when the mobile device is in an idle state. We have considered the problem of the representativeness of our data sample performing a direct measure of the pedestrian flow on a bridge. The choice of studying the mobility during significant tourist events was made for two reasons: on the one hand, we take advantage of the presence of many people to increase the penetration of the mobile device sample to reconstruct pedestrian mobility; on the other hand, there is a specific request to study the Venetian mobility during such events, since the municipality has proposed to limit the tourist presence. Moreover, at the moment, the development of counting systems for measuring the tourist flows in Venice is under discussion, and the actual numbers are estimated using average data from transportation means. We are aware

that an exhaustive understanding of pedestrian mobility in Venice certainly requires further studies, which consider a long period of data collection but in this research activity, we limit ourselves to face the problem of how to extract relevant information on pedestrian mobility from the GPS datasets. Both the chosen events attract big crowds of tourists still they present different features. The Carnival takes place in winter and *Festa del Redentore* in summer, during the evening of 15 July. Carnival is a typical tourist festival with several scheduled events distributed throughout the city (even if the main attractions are in San Marco square), whereas *Festa del Redentore* is a religious festivity very important to the Venetians which attracts many people arriving from the Venice district and attending to the fireworks along the Giudecca Canal. For these reasons, we expect differences in the observed mobility in the two case studies to be pointed out by our analysis. The distribution of devices detected by the phone cells network has been used to measure the spatial activity patterns [17–19,65] or to estimate the evolution of crowding into different areas of a city [21–23]. After a preliminary data analysis to select the devices that have provided a suitable amount of GPS data, our approach is based on algorithms able to associate a daily mobility path to each device. The main difficulties are the occasional character of the mobile device activities that prevent the data collection at a fixed spatial scale and the signal losses mainly due to the narrow roads in Venice.
To avoid the possible introduction of biases in our analysis, we prefer to follow a big data approach reducing the numerousness of device samples drastically and only reconstructing the mobility paths that satisfy well-defined reliability criteria. Consequently, our method cannot detect critical crowding situations localized on the road network. Still, it succeeds in highlighting the dynamic features of pedestrian mobility during the considered events. In particular, the presented results refer to days of 26 Feb 2017 (Carnival Sunday) and 15 Jul 2017 (Redentore day), during which the presence of tourists was particularly relevant. We have then checked the penetration of the sample by comparing the estimated pedestrian flows aggregated at each hour, with a direct measure performed by volunteers on the Redentore bridge, a large amount of people crosses that due to the presence of fireworks show on the Giudecca Canal.
The main results of this work are about the emergence of a diffusion-like relation between the covered distance and the elapsed time $s \propto t^{\alpha}$ with $\alpha \simeq 1/2$ and the existence of preferred mobility connected subnetworks of the whole road network able to take into account the majority of the observed mobility [24]. In the first case, we suggest the existence of a travel time budget [26,27,61] for pedestrian mobility in Venice and, we introduce the concept of rest times during individual mobility, which could play an

important role in the construction of dynamic models for tourist flows. In the second case, our results highlight that mobility subnetworks can simplify the monitoring and controlling the problem of the tourist flows and help the definition of models. Thanks to the information in the datasets, we can also distinguish between Italian and foreign visitors and point out the existence of different mobility paths for the two categories.

Regarding the city of Rimini, we used data stored from 07-08-2020 to 17-08-2020, the high period of the summer season, which coincides with the Italian's favorite week for the vacation. Furthermore, due to the current pandemic situation, in 2020, there was an increased regional and domestic tourism. With this dataset, the main challenge was implementing a classifier to distinguish trips performed with a particular mean of transport starting from statistical properties [63–66]. Indeed, in contrast to the Bologna dataset, we do not have a classification made by the users. Still, the data provided by Telecom are extracts of activities potentially performed on any means of transport. We downsampled the dataset using a spatial clustering procedure with a fixed radius, then georeferenced the remained records and connected them through the best path algorithm [59]. Finally, we used more filters to discard the trips with non-significant properties, like trips too short or too long in length and time. The firsts one is due to the data acquisition system (the dataset contains just tracks of an entire real trip). Instead, the seconds one may be due to professional drivers or people who use devices constantly connected in a single activity. Furthermore, after the path reconstruction, it is possible to delete the trips with an average speed physically impossible (indicative of low accuracy or errors in GPS acquisition). We evaluate the penetration of the Telecom sample, equal to $\simeq 1.51\%$, through a cross-reference with MTS data provided by Emilia-Romagna Region for the entire week.
The analysis of global statistical properties (before classification) shows the coexistence of different kinds of mobility: the power-law trend of length and time distribution and the presence of more peaks in the average speed profile prove the multimodality presence. So we implemented the Fuzzy C-Means clustering algorithm (FCM) [57, 58] to disaggregate trips according to four features: the sinuosity, the average, the maximum, and the minimum speed. We classified four different classes representing: pedestrian mobility, slow urban mobility, vehicular mobility, and highway mobility. Excluding the faster class, we can observe that the time and length trends follow the exponential law, and they collapse on the same curve if we normalize the distributions on their average values.

Furthermore, we identified the fully connected subnetworks mainly used by each class. They involve different roads of the entire network, highlighting the interpretation of the city like a multilayer. Finally, we observed the relationship between speed and density in a more minor part of the network, following the approach of Macroscopic Fundamental Diagram (MFD) [47].

In work involving the metropolitan area of Bologna, we take advantage of the availability of two datasets to study the statistical properties of multimodal mobility in an urban context. In particular, the *Bella Mossa* [44] dataset contains a sample of the trajectories traveled by bike or by walking. These are collected using a specific app. The *Octo Telematics* [45] dataset contains a sample of private car trajectories collected for insurance reasons. These data have GPS quality and provide accurate information on the path length, the travel time, and the average velocity of each recorded path on a population sample. We perform a detailed analysis of the statistical features of these observables for the different transport to show the consistency with similar results in the literature [25, 42] using other datasets and the possibility of defining mobility energy that explains the statistical distributions according to a Maximal Entropy Principle. Our analysis suggests that we can consider the travel time as universal mobility energy consistent with the concept of travel time budget proposed in the literature [43]. The main goal is to propose a simple model that explains the main features of the travel time distribution for the different transport means and that can be related to a decision mechanism at an individual level for their choice. The economists introduced the concept of cost function and logit models to model decision mechanisms that inspired our approach. Making a correspondence between the mobility cost function and the energy concept in statistical physics related to the Maxwell-Boltzmann distribution, we define a survival model. It can describe the empirical travel time distributions in the pedestrian, bike, and private car mobility cases using three parameters whose physical meaning is time cost, convenience, and typical trip duration. The three parameters explain the differences in the travel time distributions and the observed collapse of the distributions for the bike and car mobility when one normalized the travel times with the average value. The three-time scales of the survival models could also be used to compare the features of multimodal mobility in different cities where there are present specific transport networks or specific policies that have been realized to reduce private traffic. The survival model could also be useful for the development of urban mobility models that introduce a decision mechanism at an individual level to simulate the urban mobility using different transport means, in agreement with the empirical

statistical distributions.

The thesis work is organized as follows: In the first chapter, we describe the three cities taken into account, highlighting their peculiarities and requirements. We then describe each available dataset specifically. The second chapter is dedicated to explaining the implemented algorithms, grouped in three categories: algorithms for cartography extraction and management, algorithms for analysis, and algorithms for simulation. In the third chapter, we reported the obtained results for each case of study. In the last section of each of these, the conclusive remarks are reported.
Some of the results of this thesis have been collected in two papers and have been published during the PhD:

- Mizzi, C; Fabbri, A; Rambaldi, S et al. *Unraveling pedestrian mobility on a road network using ICTs data during great tourist events* EPJ Data Sci. **7** 44 (2018).
- Mizzi, C; Fabbri, A; Colombini, G; Bertini, F; Bazzani, A *Universal properties of multimodal human mobility: a statistical physics point of view* arXiv:2107.10546 (2021)

The implemented codes are available on https://github.com/physycom.

Chapter 1

Datasets and Cities of interest

1.1 Cities like case study

The adjective *smart* referred to as city includes the ability to focus and use different kinds of data from the most varied dataset acquisition systems to understand, manage and forecast the behavior of the city. This thing is possible only by combining the most useful information extracted from many available raw data. The term *behavior* of the city, in this context, indicates the dynamics related to the mobility of users and citizens but also socio-economic and environmental factors.
Even though cities have some standard features, they are often characterized by different requirements: for instance, the commercial, industrial or tourist vocation of the city, the geographical factors, the structure of the dimension of the road network. These differences point to diversifying the analysis of cities: we can focus on some problems rather than others depending on the main features of the city of interest.

For this reason, this thesis work will focus on data, techniques, and results obtained by the analysis of three Italian cities with unique characteristics very different from each other.

1. **Venice**

   Venice is a city situated in the North-East side of Italy. It includes a modern and industrialized part accessible with any means of transportation and the historical center, which appears as an island connected to the Italian peninsula through a strip of land.
   The latter is the most interesting part of our work because it represents a paradigmatic case study. Indeed, the historical center was built

Figure 1.1: Road network of Venice (Italy).

on a shallow swamp by planting trunks of trees, one close to the other, deep enough to reach solid ground. The poles driven into the mud have become so strong and cemented together that they have remained in excellent condition for centuries. For this reason, the Island of Venice is shaped by the presence of canals and bridges that confer to this city the predominantly pedestrian character of Venetian mobility. Furthermore, the island can be reached by train and car by the stretch of land or with a water bus: few points of access exist.
For its architectural peculiarity and historical buildings, Venice is one of the most visited cities in Italy, attracting large crowds of visitors all the year. The annual report of tourism drawn up by the Department of Tourism reported almost 9 million tourists for the historic center of Venice in 2019, 88% of which are foreigners.
The historical city has a surface of $6.7 km^2$ for approximately 55,000 inhabitants. This value can grow up to double during significant tourist events. The presence of large tourist flows threatens the preservation of cultural heritage.
It is therefore evident that a city with these particular characteristics has specific problems related to the governance of mobility. Indeed, the frailty of the cultural heritage is incompatible with the presence of large tourist flows, which heavily conditions residents' daily lives.
However, the relevance of the tourism economy advises against restric-

Figure 1.2: Road network of Rimini (Italy).

tion policies that would limit a priori the tourist's number. For these reasons, for the analysis of Venice, we focused on understanding and estimating the mobility pattern both global and disaggregated for kind of road-network users: tourists, citizens, or commuters. We are also interested in overcrowding studies due to significant and rare events with the possibility to predict extremes.

2. **Rimini**

The second city which we consider is Rimini. Rimini is the capital of the homonymous province, one of the eight provinces in Emilia Romagna. Specifically, Rimini is situated on the east coast of the region, and with its almost 149,000 inhabitants (ISTAT data 2019) and, $134 km^2$ represents a summer resort of international importance. Indeed, Rimini is located on *Riviera Romagnola* where it extends for about 15 km along the Adriatic coast with a series of bathing establishments.
For these reasons, Rimini is a dynamic city with a completely different behavior during the summer season and the rest of the year. Indeed, during most days of the year, there is the classical circadian rhythm, particularly for the vehicular traffic outgoing and incoming from the city center, with two peaks corresponding with the beginning and the end of the working day. In this period, mobility from the center to the coast is scarce. Otherwise, significant fluxes towards the coast and intra coastal municipalities surrounding Rimini exist during the summer.

Figure 1.3: Road network of Bologna (Italy).

We will focus in particular on this last period since it allows us to point out the differences between various kinds of mobility performed on different means of transportation. The goal is to highlight the paths most commonly used by a particular type of city user or by a specific means of transport to provide information able to make the viability longer complies with the demand for mobility.

3. **Bologna**

The latest city taken into account is Bologna, one of the biggest cities in Central Italy. It is crucial because it is a major hub for rail and road networks connecting important economic and tourist cities like Milan, Venice, Florence, and Rome.
Bologna, with its 140.9 km^2 of dimension and a resident population of almost 395,000 inhabitants (ISTAT data 2019), represents a city with a good balance between tourists and citizens which highlights the dual nature of city: tourist and economic one. In addition, the annual report on tourist mobility of Emilia Romagna declares approximately 1.5 million presences in 2019. Then the ratio between tourists and inhabitants is about 3.8, a minimum value compared with the one for the city of Venice, which is 163.6.
Bologna is also essential for the role of the University, the oldest one in the world, situated mainly in the historical center. However, in recent years some departments were moved to more suburban areas. For this

reason, this city attracts many students from Italy and abroad who need to live and move within the city center, in most cases without a private car but rather by foot, bike or bus.
Furthermore, the province of Bologna is full of companies, and for this reason, there exist people who decide to live close to the city and travel outside daily. All these aspects generate critical circadian vehicular mobility.
For all these reasons, Bologna represents a perfect city for the multimodality study, which consists of the disaggregation, analysis, and modeling of trips with different means of transportation. The goal is to understand the actual use of the road network to provide new knowledge tools to improve the sustainability of the mobility demand in future cities. With this singular spirit, in recent years, many projects have been born to promote smart and sustainable mobility.

1.2 Datasets description

We can use different kinds of data to describe, analyze and model the behavior of the city. Generally, there exist various acquisition technologies or devices themselves. These lead to the possibility of achieving data that best highlight a particular aspect of mobility, such as the origin and destination of the trip, the daily fluxes on the roads, or the diversified use of roads from different means of transportation. The data scientist's ability consists of extracting useful information from each set of data and combining them to obtain not just a sum but an augmentation of information.
For this reason, the analysis of different data sources and the comparison and mixing of the latter is widespread, particularly when we try to understand the mobility in the city. In this case, we can use data from telephone operators or apps installed on phones and data offered by the municipal services. Many cities are equipped with coils or cameras installed for monitoring and counting the crossing of vehicles or people on the main roads of the city network.
We are now introducing the different kinds of data that we used to analyze Venice, Bologna, and Rimini. In this description, we can divide data into two sets. The first one contains a data type already known and used in literature to study other cities. The second one has innovative data.

Known datasets

1.2.1 The Olivetti dataset (OD)

The dataset provided by Olivetti [62] covers a period starting the *01-01-2020* and ending the *31-03-2020* and involves the entire Emilia-Romagna region. This data are collected thanks to a system that recognizes to which cellphone tower the device is connected. The devices considered are only those subscribed to the Italian phone company TIM [30], approximately 1/3 of the population having a device. In this way, we can know how many and which movements exist from one ACE to another. This dataset has a precision of 1 hour. It is necessary to introduce this concept since it represents the spatial precision of this dataset: ACE stands for Census Area, and it was defined by the Italian National Institute of Statistics (ISTAT). Essentially, the ACEs are areas with different shapes and dimensions specified by municipal boundaries or eventual geographical barriers like rivers, canals, or ditches. The characteristic is that these areas also consider social and demographic data to obtain inhabitants ranging from 13 to 18 thousand, subjects to exceptions. For this reason, in the center of cities, the ACE density is greater, but the spatial size is smaller than in the peripheral areas.
Now we can understand the fields present in the original dataset. We have:

- DATE: the date in which the data was recorded.
- HOUR: this is in the format "hh:mm" and indicates the reference hour respect of which there was a movement during the hour before.
- ORIGIN_ACE: an alphanumerical code defined a particular area from which the movement originated. From this code, it is possible to identify the geographical shape of ACE.
- DESTINATION_ACE: it is an alphanumerical code that defines the ACE of destination.
- OCCURRENCES: number of people that move from ACE of origin to ACE of destination.

Therefore, if we have "01-01-2020;12:00;A;B;5", it means that 5 people located in B at 12:00, were in A between the 11:00 and 11:59.

The peculiarity of this kind of data is that they allow the construction of a dynamic Origin-Destination (OD) matrix with an hourly resolution during the analysis period. These dataset results are very useful for the input of simulation described in 2.3 because they allow the reconstruction of average oriented demand for mobility, as we will see in the last part of chapter 3.

1.2.2 The MTS dataset by Emilia-Romagna Region

Thanks to a collaboration with the Emilia-Romagna region, we had the opportunity to analyze the dataset coming from sensors situated below the road surface and made of a coil sensitive to the passage of any vehicle. It is about a monitoring system, called MTS, installed in the entire regional area on the main roads to control the road network and the congestion situation on a specific street. This technology consists of about 300 coils able to determine the traffic in both directions and recognize the kind of vehicle: car, medium or a big truck, lorry and bus. There exist different formats for this dataset, but the main information is always the same:

- DATE_TIME: the day and hour in which the count is stored. The time resolution is at 15 minutes, so the count represents an aggregation of all vehicles passed in the previous quarter of an hour.
- DIRECTION: the travel direction of counted vehicles. Generally, each coil has two opposite directions identified with ID "0" and "1", but there may be rare cases of a street with more than one lane for each direction identified with numbers "2" and "3". An attached dataset provides the localization of each coil. These have geographical coordinates, and each ID's direction is expressed like "from Bologna to Modena." One of the first problems is related to matching the 0 or 1 direction with Tail-Front or Front-Tail direction of poly (see section 2.1 for more details). This operation was made by hand with the aid of some visual tools.
- COIL_ID: This number identifies the coil, and it is the same present in the dataset of location. The localization of the coil allows you to associate a street of cartography to each data stream.
- COUNT: number of vehicles from a specific type counted in the quarter of hour considered.
- VEHICLE_TYPE: this is an identification number defining a kind of vehicle.

The first available dataset covers the period from 01-02-2020 to 31-03-2020, but later we had access to other periods enabling in this way the observation of seasonal changes in the mobility demand and its distribution.

The most interesting aspect is that it is possible to reconstruct a signal of occurrences with a resolution time of 15 minutes with this kind of data.

Since there exists a period of overlapping between OD and Coil data, we can use the first as input and the second as validation in the simulation done with algorithms described in 2.3.

1.2.3 The Bella Mossa dataset

In the city of Bologna we take advantage on the availability of datasets on urban traffic that contains a sample of anonymized trajectories with GPS quality: the Bella Mossa dataset. It is recorded on a period of 6 months during 2017 (from April to September) and it contains information on the pedestrian and cycling mobility in the historical center and in the periphery of Bologna.
The Bella Mossa dataset has been made available thanks to collaboration with the company *SRM reti e mobilità srl* that offered the use of an app to the citizens with the aim to improve the sustainable mobility: this is not a open dataset but it can be accessible on request to company. When activated, the app allowed to collect GPS data each 2 seconds on the trajectory performed by an individual that has also the possibility of declaring the transportation mean associated to each trip. The data recording started when the app is switched on and terminated when the app is switched off, then each trajectory is associated to an anonymized *id* so that there is no possibility of tracking individuals. The Bella Mossa initiative has involved $\simeq 10^4$ citizens in Bologna.
The dataset consists of the following fields:

- ActivityId: an identification number for the activity. This is assigned whenever a user switches on the app.
- ActivityType: the kind of transport used for the trip declared by the user itself. The majority of these are "Cicle" and "Walk" type but "Bus" and "Car sharing" are also present.
- Time: the instant of data recording.
- Latitude and Longitude: GPS coordinates of activity record.
- Accuracy: measure in meters for the position accuracy.
- Speed: velocity in m/s measured in real-time trough the GPS system of device.

The advantage of this dataset is that the means of transport is self-declared by users, therefore it is possible to disaggregate the kind on mobility and to observe both the statistical difference between cycle and pedestrian trip and the different chosen for the two mobility strategies. Finally, it is necessary to highlight that since the Bella Mossa campaign aimed to promote the sustainable mobility through a dynamic of reward with score, it is not excluded that some false declarations about the kind of used transport could exist. Furthermore, it is necessary to consider the possible bias since the statistical sample analysed is made up of sporty and probably younger people.

1.2.4 The Octo telematics dataset

The Octo telematics [45] dataset contains information on the vehicle mobility in Bologna's historical center and peripheral area (the area is comparable with that of the Bella Mossa dataset). These data are recorded for insurance reasons on a sample of private vehicles. Previous studies have pointed out that the sample penetration can be estimated as $\simeq 5\%$ of the whole daily vehicle population moving in Bologna by comparing the expected traffic flow along the main roads with the recorded traffic flow data by magnetic coils. The dataset contains a sampling of the vehicle trajectories at a spatial scale of 2 km or at a time scale of 30 seconds with the quality of a GPS datum; the data refer to September 2016 and contains $\simeq 6 \times 10^5$ trajectories. This dataset is not open due to privacy problems, and it is possible to share only aggregated information. The reconstruction of the actual trajectories is complex due to the complexity of the urban road network. We have information about the location of each journey's starting and ending points: corresponding to the power on and off of the engine. We also know the duration and the path length. The latter is automatically computed by the devices installed on the vehicle using GPS data. Therefore, we have good quality data on the path lengths, the duration, and the average velocity of personal urban mobility with the possibility of distinguishing rush hours from normal traffic conditions.

Innovative datasets

1.2.5 The MDT dataset by Telecom

Among innovative datasets, the MDT must be mentioned. Indeed, regarding the Mobile operators, the standard ETSI 3GPP (third Generation Partner Projects) service has been introduced in the provision of 3G and 4G mobile networks in recent years. In this way, an improvement of data could be

obtained by matching the GPS position with the radio measurements information available in mobile terminals as the most recent smartphones. The data thus obtained are referred to MDT, an acronym that stands for Minimization of Drive Test. The collection of data presented in our studies was carried out through Nokia's Geosynthesis system.
Therefore, the characteristic of these data is related to the possibility of collecting GPS coordinates with great accuracy for each user "activity". This latter term means any cell phone activity as calls, sending messages, app or navigator usage, etc.
MDT data have been provided for all the three cities studied in different periods and, except for some minor features that we neglect in our analysis, the sets always have the same skeleton composed of the following fields:

- CALL_ID: identification number unique for activity. It changes whenever the user ends an activity and starts another one.
- TIME: it indicates the instant of time in which the datum was recorded. It could have different formats from the more readable *datetime* as "2020-02-10 14:00:00", to the number of seconds from an initial date.
- LATITUDE: latitude in the WGS84 geographic coordinates system.
- LONGITUDE: longitude in the WGS84 geographic coordinates system.
- UNCERT_SEM_MJR: it is expressed in meters, and it describes the distance from the major axes of the GPS uncertainty ellipse (not necessarily present).
- ALTITUDE: it is the elevation above sea levels. It is expressed in meters (not necessarily present).
- SPEED_KMH: it is the speed recorded. It is expressed in kilometers per hour (not necessarily present).

The available MDT data cover different periods according to the analyzed city. In particular, the most recent data are those of Rimini, ranging from 07-08-2020 to 17-08-2020, and those of Bologna, ranging from 18-12-2020 to 23-12-2020. Instead, the MDT data for the city of Venice are collected in two different periods of the year 2017: the first one, which extends from 14-07-2017 to 17-07-2017, corresponding to *Festa del Redentore*, the second slot of data extends from 23-02-2017 to 02-03-2017 is in correspondence of Carnival celebration.
Note that there was an upgrade regarding the mobility network in recent

years that switched from 3G to 4G technology. Therefore the Venice data are recorded with 3G; instead, for Rimini and Bologna, we used the 4G.

The MDT dataset certainly plays a fundamental role in understanding the city. They stand out compared to many available data sources like cameras, coils, or sensors. Indeed, the latter provide just local information by collecting data in a precise location. On the contrary, MDT data can supply much more distributed information, tracking the path of a single anonymized user and aggregating different paths to obtain cumulative results.

1.2.6 The Venice dataset

The last dataset that we mentioned in this Ph.D. work is the one provided within the project "Venice Smart Control Room" (SCR), in collaboration with the City of Venice, *Telecom* and *Fabbrica Digitale* company [30–32]. This project aims to create a sort of urban intelligence platform monitoring the entire lagoon city. It consists of a series of data and flows from different subsystems and sensors placed on the territory.
Although the SCR collects data coming from different structures, including ACTV/AVM, Centro Maree, and Municipality, we will focus on data provided by TIM (described above) and the network of cameras installed in the critical points of the Venetian mobility. Using Artificial Intelligence and Deep Neural Network algorithms, it is possible to detect people crossing the area covered by cameras and then provide the aggregated count in a specific time interval in real-time. This kind of data enables us to improve our simulation (described in section 2.3, comparing the two temporal fluxes: the simulated one and observed one in both directions. We will show some results in the section 3.1 dedicated to Venice.

Chapter 2

Implemented Algorithms and Methods

In this chapter, we will focus on all the algorithms implemented and used to analyze the data described in 1.2 and to simulate the mobility of a city.

2.1 Algorithms for cartography extraction

The first step for any data analysis involving the city is to reconstruct the road network, which means identifying roads and crossroads with lines and points. The best approach is to simplify the complexity of urban space with an essential graph. In this way, any street that connects two crossroads is a *link* and any road intersection is called *node*.
This allows us to use some concepts from the *graph theory* like the *best path* or the *degree* of nodes and, at the same time, to synthetically keep the information about how users can move on the city graph.
The cartography used for Venice, Bologna and Rimini are shown in figure 1.1, 1.3 and 1.2 respectively. The source of cartography information has been Open Street Map (OSM), which is a collaborative project to create free content world maps. Therefore, OSM is considered *open data* and any user can download maps for whatever purpose.
Starting from OSM, we develop different algorithms to simplify the detailed map (including private walkways or excessive detail on squares or roundabouts). This simplification is necessary to eliminate unnecessary details for the dynamics of agents on the network and speed up the execution of algorithms.
OSM provides the exporting of the map in XML format with geometry and properties of each point and way that form the shape of cartography. Therefore, initially, we transform this XML format into an internal format needed

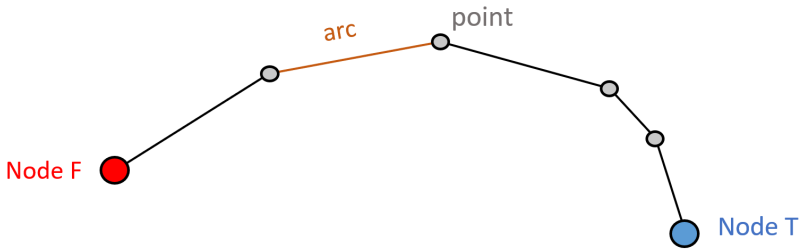


Figure 2.1: Example of poly in internal format. The first point in red is called *Node Front* or F, the last point in blue is called *Node Tail* or T, and the segment joining two consecutive points in orange is called *arc*.

to load the cartography in the software implemented by us. This latter format consists of 2 text files:

1. File with *property* information about links: in particular, we can find the identification number of link (which we will call *id_poly*), the id of first and last node (*front* and *tail* node, respectively) and other information like length of poly, the direction of travel or maximum speed if available.

2. File with *geometry* of polys. For each poly, the latitude and longitude of all the points shaping it are reported. The first and last points of the poly exactly correspond with the front and tail nodes (which we will call F and T nodes). Finally, the segment joining two consecutive points is called *arc*.

Figure 2.1 clarifies the nomenclature that we use in our internal format.

Let us now describe some of the algorithms used to manipulate the mentioned objects.

- **Removal of degree 2**

  In our cartography, generally, we want that each crossroad is a node, and each street connecting two crossroads is a poly. The raw cartography extracted by OSM does not respect this rule, so in this step, we measure the degree of each node (the number of links incoming and outgoing of it), and we act on those with degree equals to 2. Considering the crossroad as an intersection of at least three streets, it is evident

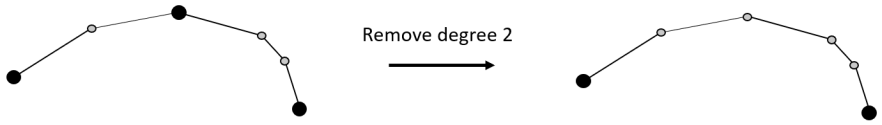


Figure 2.2: Removal of nodes with degree 2. The two poly are merged, and the node of contact is converted in point.

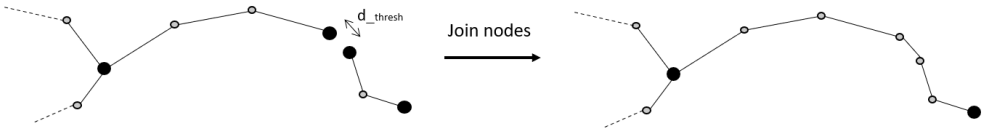


Figure 2.3: Example of joining of the pending node which is less than *d_thresh* from another pending node. Consequent removal of the obtained node with degree 2.

that we can convert each node with degree 2 in a simple point that joins the two adjacent links. Essentially, the transformation of poly with this algorithm is shown in figure 2.2. Since there may be more nodes with degree 2 one after the other, the process could be repeated until the nodes with degree 2 are eliminated.

- **Joining of nodes**

  Another essential requirement of our cartography is that it must be fully connected and without gaps: each node must be connected to a link. The OSM map does not ensure this; therefore, an algorithm makes sure to connect all pending nodes (with degree 1) that are less than a threshold distance from each other. This *d_thresh* is of the order of 12 meters, and it is a parameter to set.
  When we join a pending node with another one, we obtain a node with degree 2. Therefore, after this operation, typically, it is necessary to submit the *Removal of degree 2* algorithm again.
  The figure 2.3 shows this operation.

- **Removal of short poly**

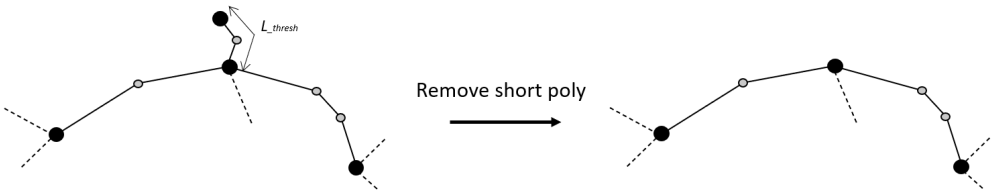


Figure 2.4: Removal of poly shorter than L_thresh meters.

Usually, we can find short polys in the raw map. Since they increase the complexity of the network and then they can slow down the execution of the analysis and simulation algorithms without adding any information, we decide to delete these pending short branches. The length for this cut is once again a settable parameter L_thresh of the order of 10 meters. The operation is illustrated in figure 2.4.

- **Merge of sub-graphs**

  It is possible to find more sub-graphs in the same cartography even after the previous operations. Indeed, the algorithm for joining the nodes can fail, for example, if the sub-graph has no pending node or if the distance of these is greater than d_thresh. In this situation, an algorithm ad hoc identifies the largest sub-graph in terms of the number of nodes and finds if a node from another sub-graph is closer than D_thresh from one arc of the main sub-graph. Generally, this distance of tolerance D_thresh is longer than d_thresh; in our case, it is of the order of 60 meters, but we can set this value according to cartography. Later, the closest node is moved to coincide with the nearest node (between F and T) of the poly to which the selected arc belongs, as shown in figure 2.5.
  This operation is done for all the sub-graphs present. The sub-graphs that cannot be merged are evaluated individually. Indeed, they could be a small network representing islands or defects originating from the bounding box that we used to select the region of interest from the OSM map. In this case, they are completely removed. However, they could be a more oversized island that we want to keep in our cartography (see Giudecca Island in Venice). In such cases, we can accept to have not a single network fully connected. Anyway, it is necessary to

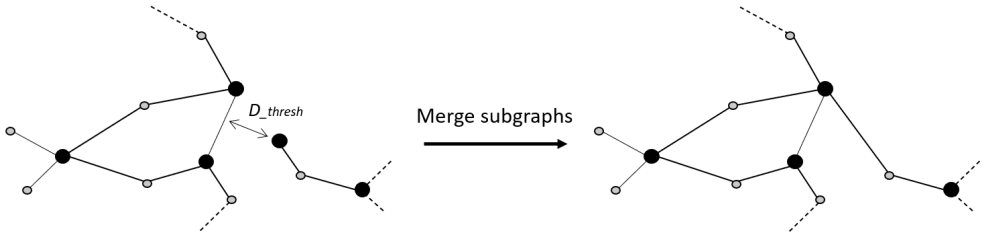


Figure 2.5: Merge of a minor sub-graph with the main one if it has a node closest than D_thresh from an arc of the main sub-graph.

point out that these are sporadic cases.

- **Merge cartographies**

  Often, having to compare the results and to receive specific information on streets from the municipal departments, it is necessary to match the provided cartography with our cartography. In particular, we need to know which poly of the other cartography corresponds with our one. We can not expect a one-to-one correspondence; therefore, we developed an algorithm that finds the poly of an external cartography E that has at least the 80% of points that are closer than 50 meters to a poly of internal cartography I.
  Generally, this operation is made because the cartography E has some information missing in the I one, like, for example, the maximum speed limit or the one-way direction. Therefore, we are just interested in matching the links to take the properties of streets without taking into account the geometry of E. Properly, since more polys of E can match with a single poly of I, usually, we kept a combination of the E properties that depends on the type of feature. For example, if the feature is the maximum speed limit, an efficient approximation takes the maximum among values of matched polys or the average of them. In figure 2.6 is shown an example of a typical situation observed comparing two cartographies.

Those shown are only some of the implemented algorithms, particularly those used to build the cartography of the three cities object of study, starting from OSM maps. However, it is necessary to specify that cartography simplification without loss of structural information is a delicate procedure and must be done to fit the specific situation. There is no fixed pipeline to

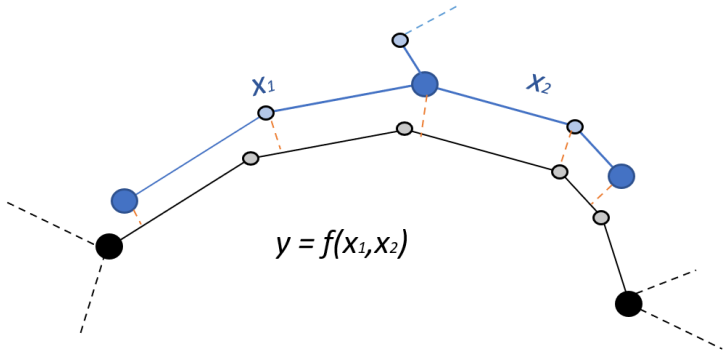


Figure 2.6: Example of a typical situation observed comparing two cartographies. In blue the external cartography E that we want to match with the black internal cartography I.

follow, and some algorithms can be more beneficial than others. Nevertheless, the common goal is to achieve a graph fully connected (in most cases) and without nodes with degree 2.

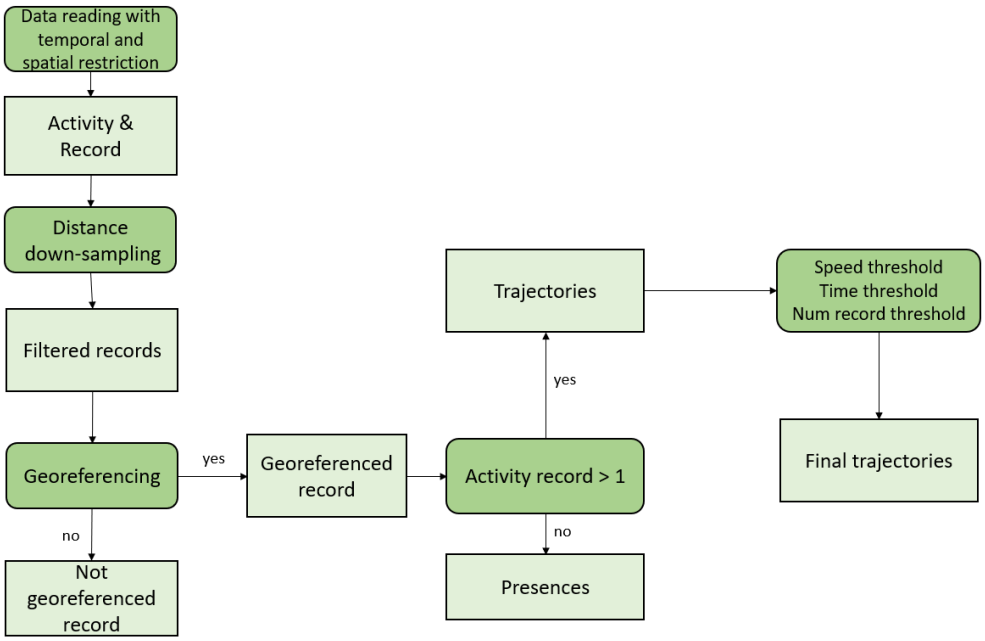


Figure 2.7: Diagram of filtering and analysis performed on GPS data. In dark green are shown the algorithms, in light green the results.

2.2 Algorithms and techniques of analysis

Once we have cartography representing a synthetic version of the city road network, we need some algorithms to interpret the interactions of data with it. All the available data include spatial and temporal information; therefore, it will be necessary to collect data and georeference them on cartography. Furthermore, we note that some kinds of data carry local information (for example, those provided by cameras or coils), while other types of them consist of a collection of GPS tracks. In the first case, we need to identify the street or the area where the datum is acquired. Then we reorganize the information to distribute it on all the cartography. Otherwise, the second kind of data needs algorithms to downsample, filter, and reconstruct the trips. We see in detail these algorithms that compose our toolchain analyzer, shown in figure 2.7.

2.2.1 Collection, filtering, and down-sampling

The algorithms of this section are dedicated to GPS track data in which each *id* corresponds to a collection of records with coordinates and timestamp, as for example those described in 1.2.3, 1.2.4 or 1.2.5. In this case, data

are being read and collected in a *key-value* map; in this manner, each key (generally the id of user or activity) corresponds to an object containing all the records acquired with that id. Just in this phase, the first spatial and temporal filters are performed. Indeed in a configuration file, it is possible to specify the maximum and minimum value of latitude, longitude, and time interval. Then we obtain a set of activities, each having an id and a vector of records.
Usually, GPS data can be acquired with fixed time intervals (like for the Bella Mossa dataset) or with an internal protocol, causing repeated data in different time instants but in the same places. Therefore, we need a downsampling algorithm. We remember that we are interested in mobility information, so we have to reduce the records that do not represent movement as much as possible. Furthermore, quite often, the records are acquired with a very short distance between consecutive points: in this way, the information results redundant, and it threatens to heavily slow down the subsequent algorithms of reconstruction and flow estimation.
Therefore, after a time sorting, we developed an algorithm that performs clustering of close spatial records in a single point that we will call stop point SP. This algorithm allows us to reduce the number of points describing the trajectory and identify and fix the spatial spike of GPS coordinates. The downsampling algorithm runs on each activity to pass from a collection of records to an array of stop points. This algorithm can be described with the Figure 2.8 and with the following steps:

1. Generate first cluster $C = \emptyset$ and push into it the record 0 (the first in time). We use the assumption that the centroid is the first record pushed into the cluster.

2. For each record i, with $1 \leq i < n_{rec}$, measure euclidean distance d_i between the record i and the centroid of cluster C. Check if this distance exceeds or not a threshold distance d_{min}.

   if $d_i < d_{min}$:
   push the record r_i into the set of points of cluster C and update the duration of the cluster, defined as the time interval between the first and last added record.

   else:

   if $SP = \emptyset$:
   $inst\_speed_C = 0$;
   promote C to stop point by adding it to SP;
   generate new $C = \emptyset$ and add the record r_i

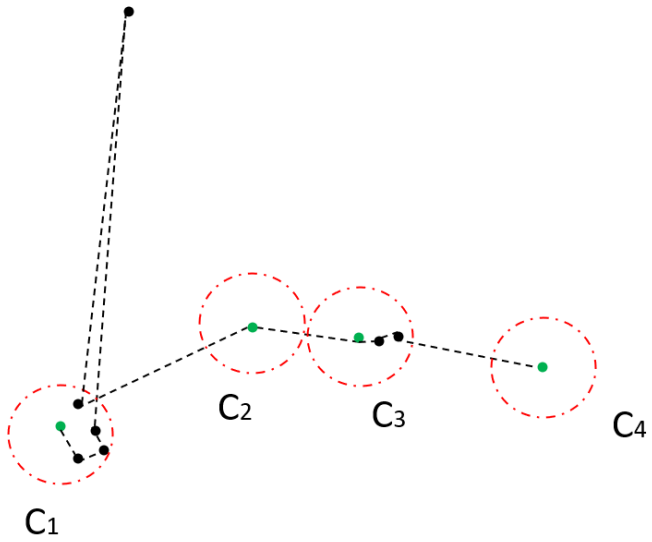


Figure 2.8: Example of spatial downsampling algorithm. Through a loop on records sorted by time, the first is elected as the centroid of the cluster (green point), and the cluster (red circle) continues to accumulate records as long as a record exceeds the distance threshold. At this moment, a new cluster is generated. In this simple example, we pass from 11 records to 4 clusters with a green centroid. These centroids constitute the successive concept of trajectory.

else:
for $0 \leq j < n_{SP}$ measure the euclidean distance d_{ij} between the record i and the stop point j;
if $d_{ij} \leq d_{min}$:
measure $inst\_speed_i = d(i, i-1)/(t_i - t_{i-1})$
if $inst\_speed_i < inst\_speed_{max}$:
push r_i into SP_j;
$SP.visited = True$;
$find\_corr = True$;
$inst\_speed_C =$ instantaneous speed of first added record;
if $inst\_speed_C < inst\_speed_{max}$ *and* $C.visited = False$:
push C into SP;
$C = SP_j$;
break;
else:
$find\_corr = True$;
break;
if $find\_corr = False$:
$inst\_speed_C =$ instantaneous speed of first added record;
if $inst\_speed_C < inst\_speed_{max}$ *and* $C.visited = False$:
push C into SP;
generate new $C = \emptyset$ and add the record r_i;

3. measure $last\_inst\_speed =$ instantaneous speed of first record added to C. Just *if* $last\_inst\_speed < inst\_speed_{max}$ *and* $C.visited = False$, push C into SP.

In the chapter dedicated to results, we will see this algorithm usually allows a strong reduction of the number of points. Generally, the rule is that the more frequently the records, the smaller the percentage of selected stop points. Finally, the previous d_{min} and $inst\_speed_{max}$ are settable parameters through a configuration file that could change depending on city; the names of this parameters are $min\_data\_distance$ and max $\_inst\_speed$ respectively.

2.2.2 Georeferencing and trajectory reconstruction

After reducing points through the downsampling algorithm, it will be necessary to georeference each point on the cartography, reconstruct the covered path between consecutive points and count for each poly the number of crossings during the analysis period. We illustrate in detail these algorithms:

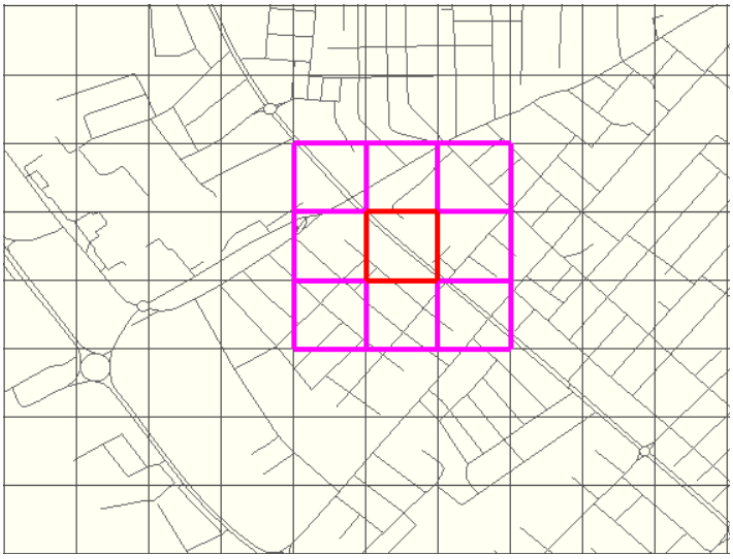


Figure 2.9: Example of cartography mapping with a $map\_resolution = 60.0m$. In red, the cell (i, j), in pink its nearest neighbors. All the nodes and arches in highlighted cells belong to the map $A[i][j]$.

Georeferencing This term refers to the attribution of a point described by spatial coordinates to an arc, and therefore to a poly, of the cartography (see figure 2.1). A stop point will be georeferenced on the arc, which maximizes a value that we will call *affinity*, and that is based on the geometrical distance.
Before describing the georeferencing algorithm, we need to map the cartography in a sort of grid. This operation will make much quicker all the future passages in which we will need to find the nodes and arches close to a point. Without treating this part in detail, just know that, depending on a settable parameter that we will call $map\_resolution$, we can obtain $N \times M$ cells with $0 \leq i < N$ and $0 \leq j < M$ defined like reported in equation 2.1.

$$\begin{aligned} i &= (lon - lon_{min})/(map\_resolution/ds_{lon}) \\ j &= (lat - lat_{min})/(map\_resolution/ds_{lat}) \end{aligned} \tag{2.1}$$

where $ds_{lat} = 111053.8$ is the conversion value from degree to meters and $ds_{lon} = ds_{lat} \cos(\frac{(lat_{max}+lat_{min})}{2} \cdot \frac{\pi}{180})$.
So we generate a map A that associates to each couple (i, j) an object pointing to arches and nodes present in the $cell_{ij}$ and in the nearest neighbors cells, such as the those red and pink respectively in the figure 2.9.

After introducing the map A, we describe the algorithm used for georef-

erencing:

1. given the stop point R that we want to associate to an arc, firstly, we measure the couple (i, j) as in 2.1, in which *lat* and *lon* are the coordinates of record.

2. *if* $A[i][j] = \emptyset$:
   return $False$ (which means that there is no arc near to the stop point, according to $map\_resolution$).

   else:
   $\forall\, arc_k \in A[i][j]$:

   measure the x and y components of- the vector from the middle point of arc k to the stop point R in the reference system of arc, as shown in 2.10.

   if $y > map\_resolution$: return $False$;

   defined $l = L_{arc}/2$ and s_0 = distance from the F node of poly (to which arc belongs) to the front point of arc, there are 3 possible cases:

   if $x < -l$:
   $s_k = s_0$
   $d_k = \sqrt{y^2 + (x + l)^2}$

   if $x > l$:
   $s_k = s_0 + 2l$
   $d_k = \sqrt{y^2 + (x - l)^2}$

   else:
   $s_k = s_0 + l + x$
   $d_k = y$

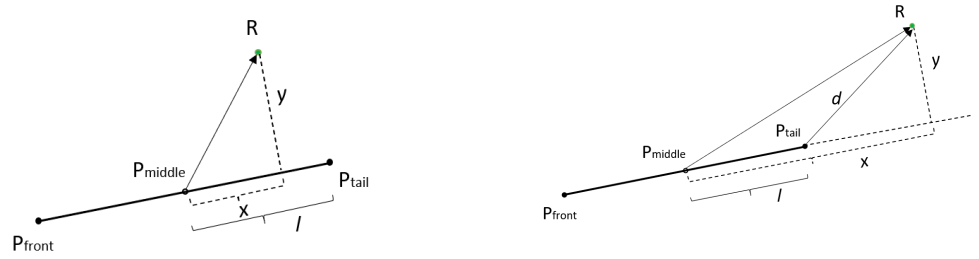
   if $d_k < map\_resolution$: return $False$;

   $x = x/l_{gauss}$, $y = y/l_{gauss}$, $l = l/l_{gauss}$;

   measure affinity $a_k = e^{-y^2}(erf(l - x) + erf(l + x))/2$

3. at the end of this loop, we have a collection of *arc*, each with the triplet (s_k, d_k, a_k): s is the distance between the start of poly and the intersection of point projection on the arc; d is the distance between point and arc measured as shown above; a is the affinity. Later, we sort all the *"affine"* arches by *affinity* in a decreasing scale and return the first arc of the list, such as the one with the greatest *affinity*.

4. *if* $d_{k\_best} < min\_poly\_distance$ the stop point is georeferenced, else it is discarded.



(a) (b)

Figure 2.10: Geometry for *affinity* evaluation of R point with the arc defined by front point P_{front}, tail point P_{tail} and middle point P_{middle}. In the case (a) the intersection with y component lies on the vector defined by $\overrightarrow{P_{middle}P_{tail}}$ and the distance d corresponds with y. While in the case in which the intersection lies over the vector (b), the distance d is the module of vector $\overrightarrow{P_{tail}R}$.

So, at the end of the algorithm execution, certain stop points are discarded; as we will see later in the results chapter, the percentage of valid data depends on the cartography and the analyzed city.

Presence vs Trajectory After the downsampling and the georeferencing procedures, there is a split between activities with a single valid point and those with more than one point, which we will call respectively *Presences* and *Trajectories*. The presences do not give any information about mobility, but we can use them to reveal the distribution of activity spots in the city through a heatmap. Furthermore, they can highlight particular accumulation points due to some events or places of interest: university centers or trade fairs. Otherwise, the trajectories can give information about how people move in the city and which roads they prefer for moving from one point to another one.
Subsequently, an additional filtering on trajectory collection can be enabled by configuration file. Once the variables *threshold_ v* (v_{thresh}), *threshold_ t* (t_{thresh}) and *threshold_ n* (n_{thresh}) are set, we can choose to discard all the trajectories with a average velocity greater than v_{thresh}, usually set with non-physical value, in order to discard possible errors survived previous algorithms. The filter also discards trajectories with a total time of activity greater than t_{thresh}. Generally, this is a large value. It excludes trips not representing the average properties of the system (for example, a bus driver moves in the city for a much longer time than the average). Finally, the last filter allows us to keep just trajectories with more than n_{thresh} points, and this

ensures the exclusion of tiny stretches of trips not valid for the representation of statistical mobility properties.

Trajectory reconstruction Regarding the trajectories, we need an algorithm able to reconstruct the path at this point of the toolchain. We want to find the list of consecutive crossed polys to get from one stop point to the next one. To do this, we implemented a version of *Dijkstra* algorithm consistent with our requirements.
The *Dijkstra* algorithm is often used to determine the *best path* between 2 nodes of a network by the minimization of cumulative weight, given by the sum of single poly weights. The weight of poly can be, for example, its length, but if available, we can also use some additional information like travel time or a ranking per the importance of the poly. Below, we explain in detail the *best path* algorithm with a generic weight.
The following commands are performed on each trajectory:

$\forall\, k$, with $0 \leq k < N_{SP} - 1$:

consider the SP_k and SP_{k-1} and find the best path connected these 2 points:

- $p1$ and $p2$ are the id of *"affine"* polys found with georeferencing algorithm, $p1l$ and $p2l$ the lengths of each of these and $s1$ and $s2$ the distance along poly between the stop point and the front node of the poly.

- *if* $p1 = p2$:
  go to the next couple of stop points

- *else*:
  defined $N1_F$ and $N1_T$ the front and tail node of $p1$, push both in a *Heap*. In computer science, this is a data structure tree-based; therefore, a defined Heap property establishes a parent-child hierarchy. Each element inserted in the Heap is determined by a key and an associated value. The hierarchical order will be performed on this value. Then, whenever a new element is pushed into the Heap, there is immediately the repositioning of every item according to Heap's property. Our property is implemented so that the first element of Heap is those with a smaller weight. The weights for the $N1_F$ and $N1_T$ are defined by equations 2.2

$$
\begin{aligned}
W_{NF} &= s1 + d_{eu} \\
W_{NT} &= p1l - s1 + d_{eu} \\
d_{eu} &= a_{eu}\sqrt{(ds_{lon}(x - lon_{SP_{k+1}}))^2 + (ds_{lat}(y - lat_{SP_{k+1}}))^2}
\end{aligned}
\tag{2.2}
$$

where x and y are the longitude and latitude respectively of the considered node. The choice to introduce this euclidean distance d_{eu} combined with the distance along the poly ($s1$ or $p1l - s1$) is a trick that, under the same distance, privileges the nodes which are in the same direction of travel. This effect can be eliminated or increased by changing the value of a_{eu} from 0 to 1.

- *while* $Heap \neq \emptyset$:

  1. take first node nw from Heap.
  2. *if* nw has already been visited from the algorithm and the current distance is greater than the stored distance, skip the successive steps and return to step 1.
  3. *if* $nw = N2_F$ or $nw = N2_T$, respectively the front and the tail node of poly $p2$, the goal is reached, so abort the *while* cycle and go to the next couple of stop points.
  4. *else*:
     for each nearest node nn of nw measure the weight $W_{nn} = dist_{previous} + d_{eu}$ and push it into the Heap, unless the node has never been visited or the current distance is smaller than the stored.

- When the while cycle was interrupted, the $SP_{N_{SP}}$ has the information about the consecutive crossing polys to get from the first stop point to the F or T node of the last stop point. Then, the last stretch of road, that is $s2$ if we have reached the front node or $p2l - s2$ if we have reached the tail node, is added to the path.

At the end of this algorithm, each trajectory brings information about which polys are crossed, in which sequence, and about the direction of travel. A visual example of the path reconstruction is shown in the figure 2.11.

Figure 2.11: Best path reconstructed with Dijkstra algorithm to go from A to B point.

Flux counter This last step allows us to collect the information about the number of activities passed through every poly with the possibility to disaggregate the passage according to the direction FT from front to tail and vice versa TF. In this way, we can observe the most widely used path in cartography. In chapter 3 we will observe some examples through the implemented visualization tools.

2.2.3 Transport Means Classification

We have already mentioned how the available data on urban mobility can be very different. In particular, there are data with punctual information, such as those provided by sensors (cameras, coils, or sniffers) installed along the streets or data coming from devices in motion; in this context, smartphones are the most widespread. These are very interesting because they can provide more distributed information and at the same time, they can track the behavior of users using different kinds of means of transport. That is why, after the algorithms which deal with filtering, georeferencing, and reconstructing the path of each activity, we need an algorithm able to classify the mean of transport used by the urban agents, starting from the features of the reconstructed trips.
Due to the nature of MDT data, we need a classifier that is:

1. *Unsupervised*
   That is a learning system able to classify a series of inputs based on common characteristics. In contrast to supervised learning, in this case,

will be provided just unmarked data, considering that the classes are unknown a priori.
Generally, the set of algorithms that allow classification of this type is note as *clustering*. The latter involves assigning samples to clusters through measures of similarity: the samples in the same cluster must be as similar as possible, and samples belonging to different clusters must be as dissimilar as possible.

2. *Soft*
This term indicates a specific kind of cluster analysis in which the classification process is not discrete (each sample can only belong to exactly one cluster). However, on the contrary, it can potentially belong to more than one cluster.

Therefore we chose to develop a classifier based on the Fuzzy C-Means clustering (FCM) algorithm. This kind of clustering is the best choice for what we need, and it is frequently used in pattern recognition.
Dunn and Bezdeck developed this method with the aim to minimize the following *target function*:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \|x_i - c_j\|^2, \qquad 1 \leq m < \infty \tag{2.3}$$

where:

- m is a real number greater than 1
- N is the number of data points (the input size)
- C is the number of clusters
- u_{ij} is the membership degree of x_i in the cluster j
- x_i is the ith sample and each sample is d-dimensional (d is the number of features selected)
- c_j is the center of the cluster j, again d-dimensional.
- $\| \star \|$ is a norm expressing the similarity between the center and the measured data.

Fuzzy partitioning is done with an iterative minimization of the target function shown above, through the update of membership u_{ij} by:

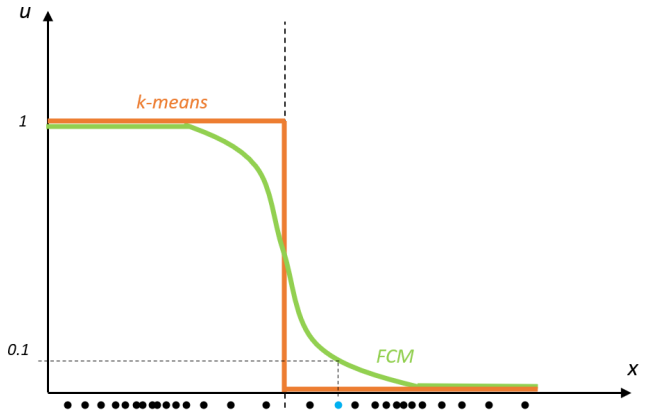


Figure 2.12: Mono-dimensional classification example with 2 clusters. In orange, the k-means membership function associates each datum to a specific centroid; in green, the FCM membership function follows a smoother line. Therefore any datum may belong to both classes.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^2} \tag{2.4}$$

and the update of cluster centres c_j by:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{2.5}$$

The iterative algorithm can be described with the following four steps:

1. Initialize the matrix U_{ij} before the first iteration which we will write as $U^{(0)}$. U_{ij} is a matrix $N \times C$, where N is the total number of data points, and C is the total number of classes of numbers between 0 and 1 that represent the degree of membership of a data point to a class.

2. For each k step, measure the vector of centres $C^{(k)} = [c_j]$ where each c_j is calculated with the equation 2.5 and using the matrix $U^{(k)}$.

3. Update $U^{(k)}$ in $U^{(k+1)}$ with the equation 2.4.

4. If $\|U^{(k+1)} - U^{(k)}\| < \epsilon$ then stop, otherwise return to step 2.

The figure 2.12 illustrates the difference between a classical hard clustering done with *k-means* and a classification done with the soft *FCM* algorithm

for a simple mono-dimensional example. Assuming that we have two possible classes, in orange, the k-means membership function associates each datum to a specific centroid; therefore, the matrix has the form of 2.6. On the contrary, the green FCM membership function follows a smoother line, indicating that any datum (as the blue one, for example) may belong to both classes with different membership degrees. In this case, obviously, the matrix U is expressed as 2.7.

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ \dots & \dots \\ 0 & 1 \end{bmatrix} \tag{2.6}$$

$$U = \begin{bmatrix} 0.2 & 0.8 \\ 0.95 & 0.05 \\ 0.7 & 0.3 \\ \dots & \dots \\ 0.1 & 0.9 \end{bmatrix} \tag{2.7}$$

This kind of approach is beneficial for MDT data since we have to consider also the hypothesis that after the analysis, there will be some trips with a single modality (the activity is fully performed with the same mean of transport). However, on the other hand, there will undoubtedly be some trips done in multi-modality (in other words, people that perform part of trip by foot and another part by bus for example). In this way, we can better understand the classification and exclude points that seem to belong predominantly to one class. In addition, we can observe between which means of transport there exists a doubt of classification.
The downside of this clustering approach is related to the necessity to specify the number of clusters. We will observe in the results that this value is set according to the city and to the number of available means of transport.
Finally, we should note that part of the analysis work in this context is related to the choice of the value of some functional parameters like the ϵ to stop the algorithm or the fuzzifier m value. The latter is connected to the level of cluster fuzziness: in the limit of m equal to 1, the degrees of membership converge to 0 and 1. While, the larger m, the smaller membership values. Without any knowledge about the domain m is typically set to 2.

In order to evaluate the goodness of clustering, we introduce the quantity

called as ***Dunn Index***, that is the ratio between the minimum distance inter clusters and the maximum distance intracluster as defined in 2.8.

$$DI = \frac{min_{1 \leq i \leq j \leq C}\, \delta(c_i, c_j)}{max_{1 \leq k \leq C}\, \Delta_k} \tag{2.8}$$

This distance can be evaluated in different ways. We choose to use the euclidean distance in the features space. Since a good clustering implies clusters well separated from other cluster but at the same time compact, can be deduced that the grater the Dunn index, the better the clustering.

2.3 Simulation algorithms and techniques

In this section, we will focus on the software implemented in order to simulate mobility in a city. Indeed the statistic analysis and the emergence of preferential paths have allowed us to implement and model a *digital twin* which takes into account the behavior of pedestrians and vehicles on the road network.
Here we are introducing some aspects we need to understand how the simulation software works:

Cartography The cartography is always at the base of any studies about the city. We use the cartography OSM rearranged with algorithms described in 2.1. This cartography is the same that we used in the data analysis done as described above in 2.2; in this way, we can use the same language (about poly and node id) to pass information acquired in the analysis to the model. For example, we can use information about the number of poly crossings to give it more or less importance in choosing the path or information about poly's actual speed and direction. Then the first step of the simulation consists of loading cartography files and storing poly and node information.

Attractions For each simulation, it is possible to define some *attractions*, that is to say, some accumulation points of the city. The presence of these points catalyzes the mobility strategies of certain types of agents, who will follow paths that pass through them. In order to model the behavior of citizens and tourists, especially for tourist cities like Venice, we have to consider the effect of historic buildings, squares, exhibitions, and museums acting as points of interest. Finally, the attractions are implemented in such a way that, by a configuration file, we can define:

- location of points expressed in geographical coordinates
- weight, seen as the relative importance among all the attractions. This weight can be a single value valid for the entire simulation or a vector that allows changing the relative importance
- maximum capacity
- visiting time if available
- time table defining the opening and closing hours

Presuming that a tourist visit more than one attraction, just as long as it finishes its travel time budget, we need to consider all the combinations of attractions. So, once defined a number max (k_{max}) and min (k_{min}) of possible consecutive attractions, we compute all the combinations, that is all the *k-subsets* of N_{attr} with k distinct elements and k ranging from k_{min} to k_{max}. For each of these, we measure the length of the trip given by the sum of best paths from one attraction to the consecutive. In addition, we also compute the cumulative weight obtained as the product of all the weights of involved attractions. When the weights change, we recalculate the weight of the specific combination.

Sources If the attractions have a role in defining and influencing the destinations of agent trips, the sources are the objects that we use to input the agents on the network. Generally, they are located in correspondence of access points to the city like, for example, train stations, ports, or bus stations. The implementation of sources is done so that all the agents generated from that source have the same characteristics defined by some variables. Now we introduce these variables that we will explain later in more detail:

- the location of points expressed in geographical coordinates. With the term *source* we mean any object that pushes agents into the simulation. Therefore, depending on the presence of some variables, we set the creation of different dynamics:
  - **SRC_DEST**: if *source_location* and *dest_location* are both present, the agents are generated in the source point and must achieve the destination point (both expressed in geographical coordinates).
  - **SRC_RND**: if just *source_location* is present and there is a tag *dest = -1*, the agents are generated in the source point and they reach a random destination.
  - **RND_DEST**: if just *dest_location* is present and there is a tag *start_node_lid = -1*, the agents are generated in a random point and they have to reach the destination point.
  - **RND_RND**: If neither of those locations is present and both *dest* and *start_node_lid* are equal to -1, the agents will be generated in a random point and they have to reach another point extracted randomly.
- creation rate: how often the source place agents on the network from its location.

- creation table: vector that defines the number of agents per hour that will be created.
- α_{we}: a value ranging from 0 to 1. In the following sections, we will see in detail how it works. For now, suffice to say that it rules the importance of poly length or poly crossing in the best path algorithm.
- α_{speed}: it ranges from 0 to 1, again useful in the best path algorithm to manage the importance of poly speed or poly crossing.
- β_{BPmiss}: it is a probability ranging from 0 to 1 that the agents have to make mistakes and not follow the best path.
- v_{min}_mps and v_{max}_mps: they are the minimum and maximum speed of a distribution from which the speed of agents are picked up.
- alternatively, it is possible to define the exact value $speed_{mps}$ expressed in meter per second.
- $cherry\_pick$: it is a tag that can be present when we have a random origin or destination, and it has a value expressed in kilometers. In this case, the random node is picked up in a circle centered in the origin or the destination point with a radius length equal to cherry_pick value. This mechanism allows us to generate and manage a sort of urban mobility in specific areas.
- TTB, an acronym that stands for Travel-Time Budget, if present, allows you to edit for that source the maximum TTB of the uniform distribution from which the value of TTB is picked up. Indeed, excluding the SRC_DEST agents, all the other types have a maximum TTB with a default value of 1 hour. The travel time budget represents the amount of time that people are prepared to take in mobility, so when the agent's lifetime exceeds its TTB, it stops its trip and disappears from the network. The value of TTB can change according to means of transport, network characteristic, or type of user (citizen, tourist, or commuter, for example).

Generation of agents and travel agendas At the start of the simulation, the first step is to load and prepare the environment; therefore, after cartography set-up, agents are generated from sources. This process happens at the instant zero of simulation, but it will be repeated whenever the source must create new agents.
Each agent has the following characteristics:

- **origin node**: if the source that generates it has specific coordinates, the assigned node is the closest. While if *start_node_lid = -1*, the node will be randomly extracted among all possible nodes. If we have additional information, such as the population for each area, we can assign a score to each node, so that node with a higher score is preferred over those with a lower score.

- **destination node**: also here, the destination's position could be expressed explicitly through coordinates, and then we can use the closest node. Otherwise, we randomly extract the destination.

- all the information about α_{we},α_{speed}, β_{BPmiss} of source are passed to agents generated with that source, influencing the individual dynamic, as we will see in the next section.

- **initial speed**: it is extracted from a uniform distribution between v_{min} and v_{max} or alternatively it is equal to $speed_{mps}$ defined in source. Note that this is just the initial speed because later, we will explain how the dynamic and the speed are influenced by agent density.

- **weighted route**: we design this option to simulate particularly the trips of tourists that arrive in a city from principal entrances (like train or bus station and parking area) and visit the city's main attractions. We just introduced the attraction object and the collection of all k-subsets consecutive attractions. Every k-subset has a length and a weight, measured as explained above. When we have to generate an agent with a weighted route, we extract a bin in the distance distribution. This is a discrete distribution in which the probability of each bin is defined by equation 2.9, where l is the length of combination, $L_c = 0.5$ is the characteristic length, $\alpha = 7.5$ and $\beta = 0.84$.

$$p(l) = \frac{\beta e^{-\beta l} \frac{(1+e^{\alpha L_c})}{1+e^{-\alpha(l-L_c)}}^{\frac{\beta}{\alpha}+1}}{1+e^{\alpha L_c}} \tag{2.9}$$

  Once we extract a bin, we make a second extraction among all the k-subset coming from that bin and take the weight into account. In this way, for each agent "tourist", we defined an agenda of destinations to visit before returning to its source point.

Dynamics of movement After the initialization of simulation, it evolves from a *start_date_hour* to a *stop_date_hour* with a step of *dt*. This value

is also settable and usually is fixed to 10 seconds. During this time step, each agent moves on the network to reach their final destination, passing or not through the attractions that define its agenda. The traveled meters during dt depend on the traveling speed on the current poly where the agent is. This speed v_{dt} has a minimum at v_c and it is related to density with the formula 2.10.

$$if\ \rho \begin{cases} > \rho_c & v_{dt} = v_c \\ \leq \rho_c & v_{dt} = (1 - \frac{\rho}{\rho_c}) v_{free} \end{cases} . \tag{2.10}$$

where $\rho = count/L$ is the ratio between number of agents on a poly and its length; ρ_c and v_c are the critical values for density and speed (they change if we consider agents as pedestrians or cars); v_{free} is the free speed of that poly that can change according to the transport mean considered. Usually, for the simulations we used $\rho_c = 4.0\, m^{-1}$ and $v_c = 0.15\, m/s$ for pawns and $\rho_c = 0.71\, m^{-1}$ and $v_c = 2.78\, m/s$ for cars.

Thanks to the best path algorithm, each agent knows the path to take to reach the destination. Indeed, during the initialization of simulation, the best path between all pairs of nodes of cartography is computed using the algorithm described in the last part of section 2.2.2 and weight given by the following equation:

$$W_{poly} = L \left[\alpha_{we} + \alpha_{speed} \left(\frac{v_{MIN}}{v_{free}} \right) + (1 - \alpha_{we} - \alpha_{speed}) \left(\frac{1/cnt - 1/cnt_{MAX}}{1/cnt_{MIN} - 1/cnt_{MAX}} \right) \right] \tag{2.11}$$

where L is the length of poly; v_{free} is the free speed (it depends on the kind of mean of transport); V_{MIN} is the minimum traveling speed; cnt, cnt_{MIN} and cnt_{MAX} are respectively the number of people passing along that poly, the minimum and the maximum number of people passing along any poly. These last values are extracted from the analysis done with algorithms described in 2.2 in the case of available data.
Finally, here we understand the role of the two parameters α_{speed} and α_{we}. Indeed, they are "buttons" that allow us to consider or not each of the three terms of the equation 2.11 in the calculation of weight. The meaning of the terms are:

1. the first term is based on the length of poly; then, in this case, the best path is equivalent to the shortest path in terms of meters. It is activated from the α_{we} coefficient.

2. the second term is based on a combination of length and speed. Under the same length, the path that appears shorter is that with higher velocity. This term is activated only by α_{speed} coefficient.

3. the last term is based on a combination of length and occurrences. In this case, instead, under the same path length, the shorter is the most popular.

Note that, given the coefficients of the third term, we can have the two extreme scenarios:

- when $\alpha_{we} = 0$, with α_{speed} it is possible to role the relative importance of second and third term. In particular, with $\alpha_{speed} = 0$ the occurrences information prevails; with $\alpha_{speed} = 1$ we consider just the combination with speed.

- when $\alpha_{speed} = 0$, the second term is canceled; therefore, with $\alpha_{we} = 0$, we consider just the occurrences factor; otherwise, with $\alpha_{we} = 1$, we consider only the length.

Another aspect that influences the dynamic is the probability of making a "mistake." Indeed, each agent knows its destination (be it a random node or an attraction node). Through the best path algorithm, it knows what the poly is to take when it reaches the crossroad, so with the β_{BPmiss}, we can introduce a more or less high probability to take another poly. So, every time that an agent arrives over a crossroad, a value from 0 to 1 is extracted from a uniform distribution; if it is greater than β_{BPmiss}, the agent is not mistaken and continues with the best path. Otherwise, an alternative poly was extracted taking into account the poly weight W_{poly}, according to 2.11.

When we introduced the *attractions*, we have mentioned the possibility of setting the maximum capacity and the visiting time. Indeed, to simulate the realistic crowd generated outside the main buildings and museums (like, for example, St Mark's Basilica in Venice), we implemented a crowd mechanism. Therefore, if the attractions are complete when the agent arrives, it "takes a ticket" that allows it to know its turn. In this way, every time some agents leave the building, the next agent can come inside.
Since different phenomena depend on the number of people, like delay on a single poly or crowd for visiting the sights, we introduced the TTB travel time budget. When the lifetime of an agent exceeds its TTB, it immediately stops its trip, wherever it is.

The last aspect that we want to mention regards the possibility to add the public transport information and to set a certain number of agents that are allowed to take the public transport if a ride is available within a time of tolerance.
The information about rides, routes, and stop times are were incorporated starting from the GTFS format. This acronym, General Transit Feed Specification, was developed by Google originally: the first capital letter was for Google, subsequently turned into General, since it now represents the standard for this kind of information. Essentially, the GTFS has a structure with six mandatory tables, described below in brief:

- **agency** is a table with information about the transport company.
- **routes** this table contains information about routes.
- **trips** this table with fields *route_id* and *trip_id* allows the association of route and trip objects.
- **stop_times** it contains the timetable of transport mean at the different stops.
- **stops** it includes geographical coordinates of each stop.
- **calendar** it defines the passage recurrences and days or periods of operation.

Depending on the day that we want to make the simulation, the software creates the *stop objects* and the *transport objects*. The first one store the geographical information of stops, the second the information about trips, such as the consecutiveness of stops, the timetable or the trip, and route id. Finally, is created a third object that acts as a proxy between instant time and available trips for each couple of stops. In this way, when an agent with public transport ticket is created, an algorithm finds if a couple of stops near its origin and destination exists, and a scheduled ride connects these points. A parameter called *tolerance time* was introduced; indeed, if there are no trips available in this slice of time, the agent chooses to go on foot. The default value for tolerance time is 30 minutes, but it can be settable from the configuration file.

Monitoring and results collection It is possible to monitor the simulation in every *dt* and know how many agents are present on every poly and the entire network. Furthermore, we develop different tools for dumping and visualizing the distribution of speed, lifetime, and occurrences during

the simulation. These are very useful for debugging and for comparing the simulated results with the real expected results.
Precisely for this reason, we implemented some synthetic barriers that count how many people cross them every time interval *dump_ cam_ dt*. Once again, this latter is a settable parameter with a default value equal to 15 minutes.

Concluding remarks The more significant the real data are, the greater the veracity of synthetic results is. In particular, when we have a continuous flow of measurements, as cam or sniffer data, we can use some of these workstations as input of our model and the other as a test of the simulation behavior. We have done this for the city of Venice. However, another possibility is to use a kind of data like the OD dataset (1.2.1) to define the origin and destination of agents and with the Coil data (1.2.2) check the model behavior and try to best fit it. In the chapter 3 we will investigate this possibility in the simulation of Emilia-Romagna car mobility.
In the end, we specify that all these implemented algorithms are developed in the C++ programming language. In contrast, the algorithms regarding pre-processing and post-processing are implemented in Python.

Chapter 3

Results

3.1 Case study: Venezia

The cartography used for Venice analysis is shown in figure 1.1, and it includes a box area defined in table 3.1, involving the historical center of Venice.

	Latitude	Longitude
Min	45.418301	12.296303
Max	45.451410	12.372170

Table 3.1: The bounding box of Venice cartography expressed in minimum and maximum latitude and longitude.

With the city of Venice, there has been a long collaboration. Indeed it is the first of the three cases of studies that we considered from 2017.
It represents the main lab in which we implemented and tested a large number of algorithms described in 2.2 and 2.3, then both in terms of statistical analysis and simulation modeling.
The first work that we illustrate is about the analysis of data provided by the Italian mobile phone company TIM already discussed in 1.2.5. This data involves tens of thousands of anonymous devices that performed an activity during eight days from 23-02-2017 up to 02-03-2017 when the Venetian Carnival was going on, and from 14-07-2017 up to 16-07-2017 on the occasion of the *Festa del Redentore*. The dataset refers to a geographical region that includes an area of the Venice province so that it is possible to distinguish commuters from sedentary people and the different transportation means used to reach Venice.

Furthermore, these data were stored with the 3G protocol. For this reason, it was also possible to obtain the roaming status, which in turn allows distinguishing between Italian and foreigners.
The activity/devices are fully anonymized, and the system automatically provides not reversible identification numbers (ID) for mobile phones and calls within the trial's scope; the ID is kept for a period of 24 hours. During each activity, a sequence of GPS data is recorded with a 2 seconds sampling rate. In this way, it is possible to follow local trajectories and detect points of interest where people stop for a specific time. As matter of fact during an activity most of people reduce their mobility except if they are on a transportation mean, so that the dataset contains a lot of small trajectories that have to be joined to reconstruct the daily mobility. Both the Carnival and the *Festa del Redentore* datasets contain $\simeq 1.5 \times 10^6$ georeferenced records concerning the mobility in the historical center of Venice during each observation day. These data mobile phone ID allows studying the daily mobility of a sample of the device population of $\simeq$ 5000 devices per day with the possibility of reconstructing the main paths used in the road network and estimating the average activity rate of a device during the circadian rhythm. The presences in the historical center of Venice during the events were of the order of 10^5 individuals per day as reported by local newspapers. We estimate an overall penetration of our sample of 3-4%. Figure 3.1 shows examples of the distribution of the GPS data recorded in the Venice historical center. In the sequel, we illustrate in detail the results of our approach for the Sunday 26-02-2017 during Carnival and for Saturday 15-02-2017 during the *Festa del Redentore* that were exceptionally crowded days.

Sample penetration estimate

We have performed a filtering process on the available datasets to extract relevant information to study the mobility on the road network. We aggregate the GPS data of each device-ID to downsample the data by starting from an initial position (pivot point) and by computing the geodetic distance with the successive points associated with the same ID. When the distance overcomes a fixed threshold (we choose a threshold value of 50 m), we keep the new point and restart the procedure using the new point as a pivot. In this way, the number of valid positions is reduced respectively to $\simeq 60 \times 10^3$ per day in the Carnival dataset and to $\simeq 90 \times 10^3$ in the *Festa del Redentore* dataset.
Each selected GPS point is located in the nearest road network link within a distance of $60m$ included the ferryboat lines; we discarded the points that cannot be attributed to any link according to this criterion. The positioning

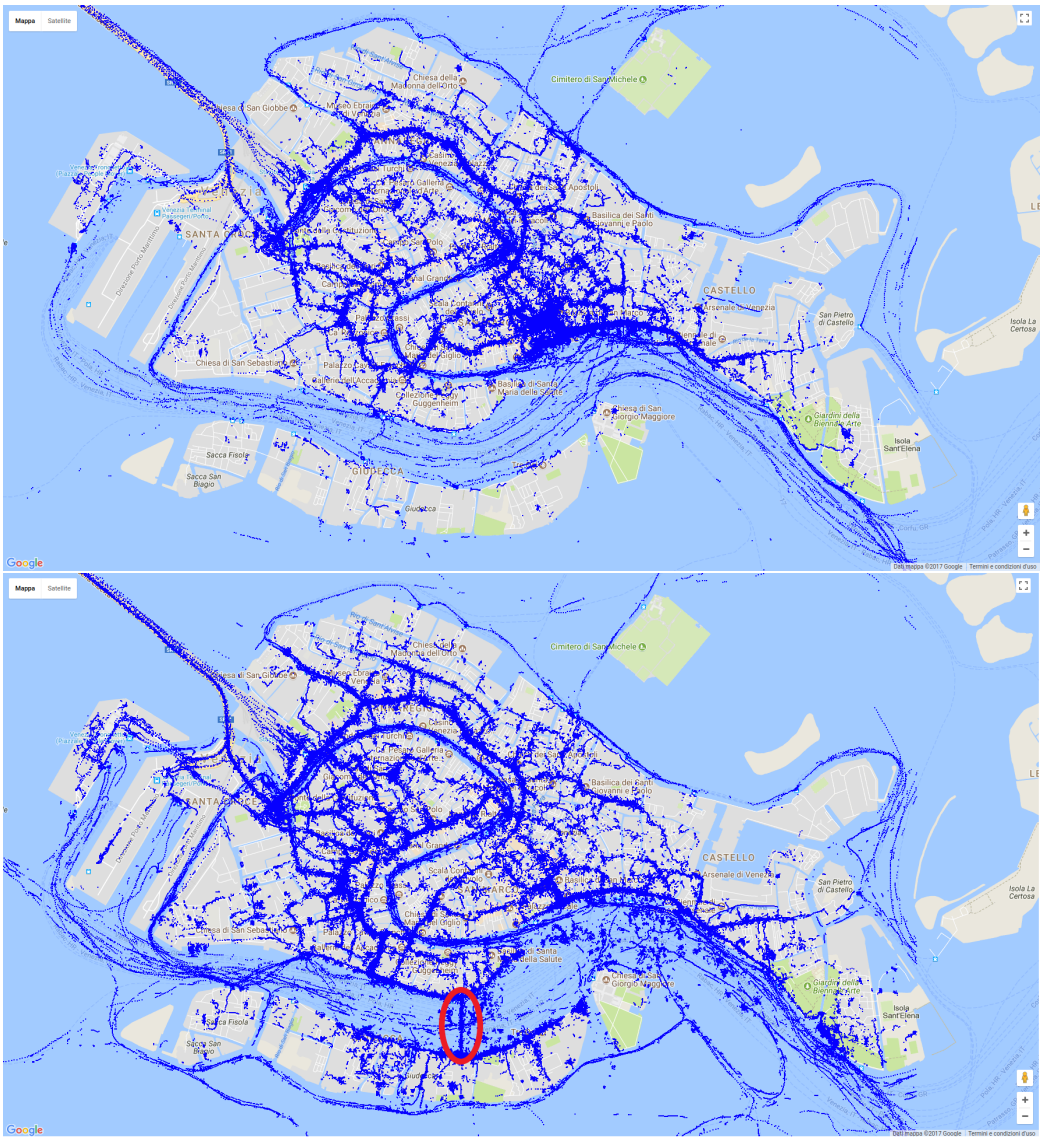


Figure 3.1: Examples of the distribution of the GPS data recorded in the Venice historical center: the top picture refers to the Carnival dataset and shows the data of 26-02-2017 from 12:00 to 14:00. The bottom picture to the *Redentore* dataset and shows the data of 15-07-2017 from 19:00 to 21:00. The red circle points out the Redentore bridge location, which is a floating bridge built for the special occasion of the *Festa del Redentore*.

procedure further reduces the valid points down to $\simeq 50 \times 10^3$ in the Carnival dataset and down to $\simeq 80 \times 10^3$ in the Festa del Redentore dataset. These positions allow dynamical information on the most visited paths on the road network and on the main points of interest where the rest time is significant.
We have performed a direct check for the representativity of the considered sample on the spatial scale of a single road. In particular, we compare the estimated flows using GPS data with the pedestrian flows directly measured by people counting on the *Redentore* bridge. CORILA [46] organized the campaign of measures, and the data were collected every 15 minutes by using people count devices. *Redentore* is a floating bridge on the *Canale della Giudecca* (see Fig. 3.1 bottom and the map in the supplementary material). The bridge has a length of $\simeq 300$ m. It was opened from 19:00 on 15-07-2017 for all the night, except during the firework show between 23:00 and 00:30.
To estimate the pedestrian flow across the bridge, we have counted the mobile devices that leave two signals at opposite sides of the bridge during the considered time interval slot to distinguish between the bridge crossing into opposite directions.
The results are reported in the figure 3.2 (top) where the estimated penetration of our sample is $\simeq 1.6\%$ according to a best fit of the measures with 20% average error (excluding the flow measured after the reopening of the bridge). The reduced sample penetration to the expected 5% is due to the small spatial scale of the bridge that requires a coincidence of two signals from the same device at the opposite sides of the bridge in a brief time interval. Then we expect that the variability of the activity rate reduces the sample penetration. We have computed the 10 minutes activity rate for the devices located in an area near the bridge from 19:00 of 15-7-2017; the results are reported in the Figure 3.2 (bottom).
Finally, we remark that the estimated flows allow reproducing the evolution of the empirical observation with good accuracy except for a single point between midnight and one o'clock a.m. when the bridge was reopened after the firework.
Indeed a significant pedestrian flow was recorded between 00:30 and 1:00, and the GPS dataset does not detect it. A possible explanation was that the activity of the mobile device in the area was dropped down during and after the fireworks. In fact, most people were mainly interested in attending the show and, afterward, crossing the bridge towards the Venice center, moving in a crowded environment. Indeed, using the direct empirical observations, we evaluate a net flow towards the Giudecca island of 8000 people from the opening of the bridge and a net flow of 14000 people after the bridge reopening (probably because some people reach the island by ferryboat).
The GPS dataset correctly estimates the incoming flow but underestimates

the outgoing flow of approximately 8000 people. This estimate could be consistent if the device activity at the bridge were reduced by a factor of 3 in the time interval from 00:30 to 1:00.

We observe as the device activity rate increases before the firework show, probably due to the people excitation before the big event of the day, and quickly drop down by a factor 2 afterward. However, the empirical evidence suggests that the selected sample recovers its representativity during the night after the first time slot. A possible explanation is that at the bridge's reopening, the high level of crowding due to the pedestrian flow incoming to Venice discourages people from using the ICT while they are walking.
As a consequence, many people move away from the area without leaving any signal in the dataset. When a normal condition is recovered, the device activity of the remaining people returns at the same level as before the fireworks. However, we have observed a lower average value since we are still considering the whole device population.
According to the previous discussion, a model for the pedestrian flows based on the ICT device activities could miss detecting localized critical situations.

Mobility paths reconstruction on the road network

The filtered GPS positions are geolocalized on the Venice road network that has to be extended to include the ferryboat lines (see figure 3.1 bottom). The procedure considers the land and water mobility separately since the two mobility networks are physically separated. It is necessary to check the transitions from one network to another carefully. We connect two successive points left by the same device using the best path algorithm on the road network to create a mobility path. We check when the path changes from land mobility to water mobility and the estimated travel speed.
To end a land path and start a water path, we require that at least two successive points in a sequence are attributed on a ferryboat line by the algorithm. In the case of a single point on a ferryboat line, we forced the localization of this point on the nearest road on land. In the second case, we discharge the paths whose velocity is inconsistent with the typical pedestrian velocity (or ferryboat velocity).
Finally, we have neglected unconventional paths which cross a very high number of roads (more than 200) or have a small number of points (less than 3). In the first case, we attribute these paths to people performing a particular activity in Venice (for example, a postman) unrelated to tourist or citizen mobility. In the second case, the associated paths are too short to study the mobility properties.
In this way, we reconstruct the daily mobility of $\simeq$ 2800 (resp. $\simeq$ 3600)

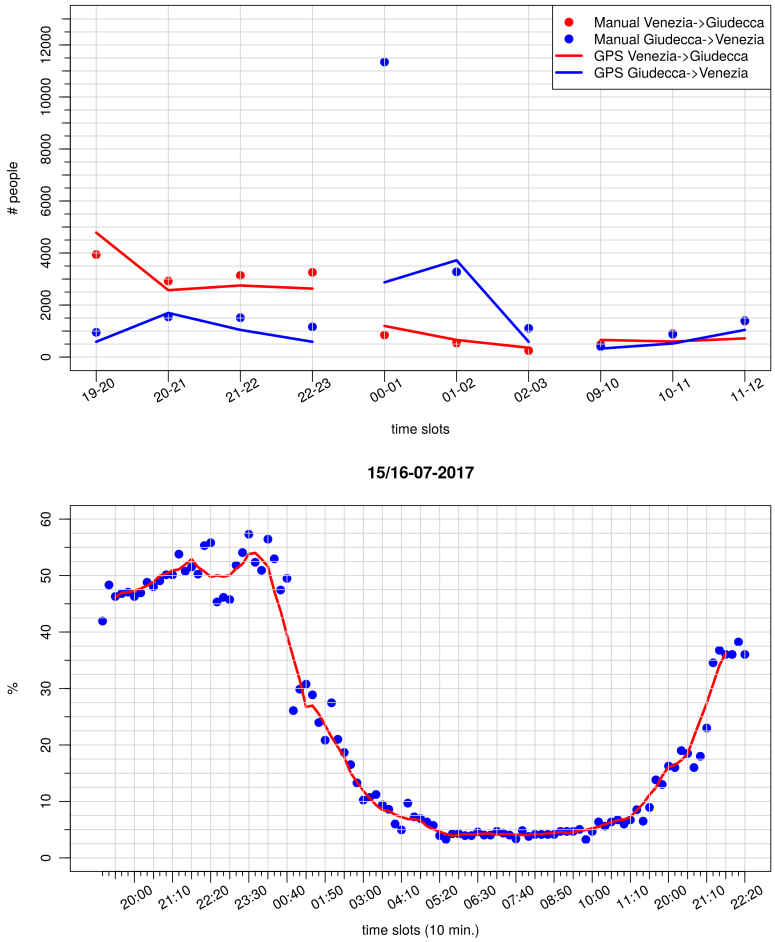


Figure 3.2: Top picture: comparison of the hourly flows on the *Redentore* bridge estimated from the GPS database (continuous curves) and the empirical measures by a direct people counting (dots). The blue data refer to the pedestrian flow from *Giudecca* island toward Venice, whereas the red data refer to the pedestrian flows in the opposite direction. The scaling factor applied to the GPS data corresponds to the penetration of 1.6% of the sample. We recall that the bridge was closed between midnight and one o'clock. Bottom picture: empirical relative frequency to get a GPS record in a time interval of 10 minutes, from a device of our sample present in the area of interest near the *Redentore* bridge; the red line is a mean average over one hour to smooth the fluctuation effect.

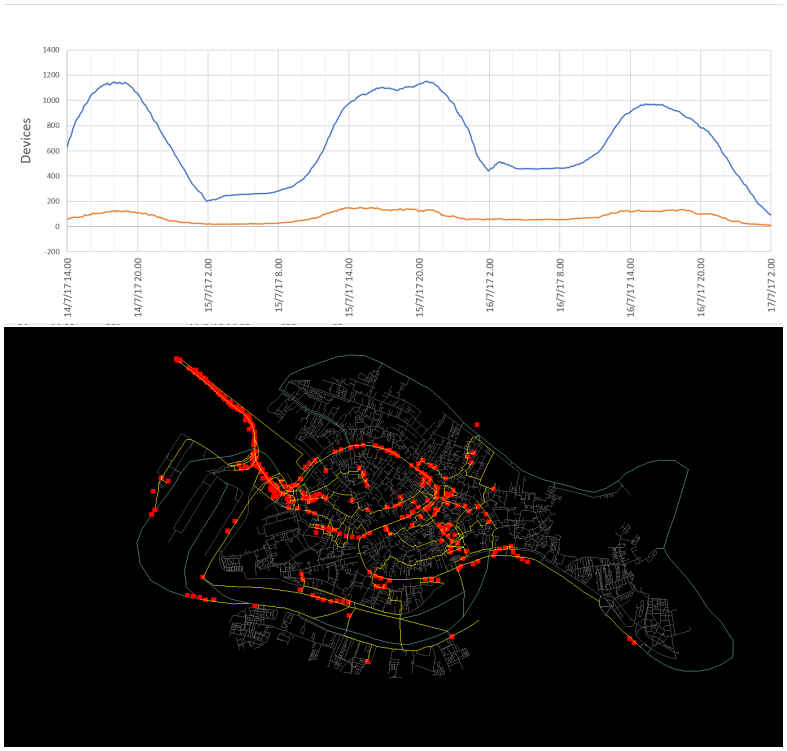


Figure 3.3: Top picture: number of selected devices present the *Redentore* dataset collected during the three days: we observe the abnormal increase of the presence during the night of 15/7/2017. Bottom picture: some examples of mobility paths reconstructed (continuous lines) on the road network of the historical center on Venice using GPS data (red dots).

different devices per day for the Carnival dataset (resp. for the *Redentore* dataset). In Figure 3.3 (top), we show the measured number of moving devices detected in the historical center of Venice. The algorithms have correctly reconstructed the mobility paths during the *Redentore* celebration: the figure refers both to the land and water mobility. Clearly, it shows the presences' circadian rhythm with a peak during the evening of 15/7/2017 on the occasion of the fireworks. According to the estimated penetration of 1.6% of our sample, the GPS data estimate a peak of $\simeq 80,000$ people during the evening of 15th July and a number of passengers of order 10^4 on the ferryboat transportation system. These numbers are consistent with estimates provided by the local newspapers. In Figure 3.3 (bottom), we report some examples of the reconstructed mobility paths on the Venice road network joined with the ferry lines on the channels.

Statistical properties of mobility paths

The mobility paths provide dynamic information on how people realize their mobility demand on the road network during the considered events. The elapsed time between two successive GPS data is used to attribute a displacement velocity that, of course, is affected by the rest times at any point of interest.

We remark that we have not a start and end point of every single trip, but only a sampling of the whole daily mobility of a device since the GPS data are recorded only in conjunction with an activity: for example, the elapsed time between two successive points may be affected by a stop for shopping.

A dynamic model to simulate the pedestrian dynamics on the Venice road network based on the individual dynamics has to include tracts covered at constant velocity and breaks due to points of interest, crowded situations, or to recover from the walking fatigue.

We consider some statistical properties of the reconstructed mobility paths to check if they are consistent with other statistical laws suggested by analyzing mobility datasets in urban contexts. In fig. 3.4 we report the daily path length distribution for both the considered datasets: the average mobility lengths are 3.1 km and 4.3 km respectively for the Carnival and the *Festa del Redentore* datasets.

The differences between the two distributions may be explained both by the effect of weather conditions (the Carnival takes place in winter whereas the *Festa del Redentore* is celebrated during summer) and by the different organization of the two events. The Venice Carnival is an ensemble of events spread on the historical center even if San Marco square is always the attractive primary location, whereas the *Festa del Redentore* is celebrated in the area near *Giudecca* Canal between the Giudecca island and the *Riva degli Schiavoni*.

Therefore one expects mobility more influenced by an origin-destination character during the *Festa del Redentore* than during the Carnival of Venice.

The path distribution in fig. 3.4 refers only to pedestrian mobility since we have excluded all the mobility paths with a tract on a ferry line. This criterion is satisfied by 2/3 of the devices in our sample, whereas the remaining 1/3 performs mixed mobility.

We propose an exponential interpolation of the path length distribution for both the datasets (cfr. dashed lines in fig. 3.4) and we observe as the exponential interpolation overestimates the short paths in the *Festa del Redentore* according to the existence of a great origin-destination component.

Assuming the existence of an average characteristic pedestrian velocity, the path length can be interpreted as a *mobility energy* distribution in agreement with a Maxwell-Boltzmann distribution. It is consistent with the concept of

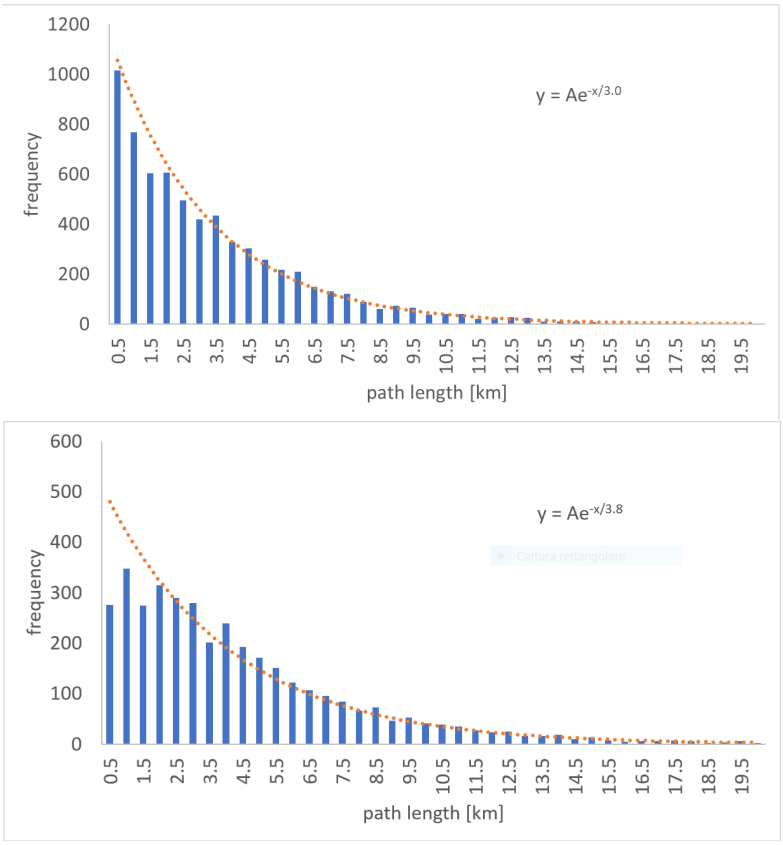


Figure 3.4: Distribution of the mobility path lengths reconstructed during the Carnival (top picture) and the *Festa del Redentore* (bottom picture) in the Venice historical center. The dashed line is an exponential interpolation of the distribution tail whose formula is reported in the pictures.

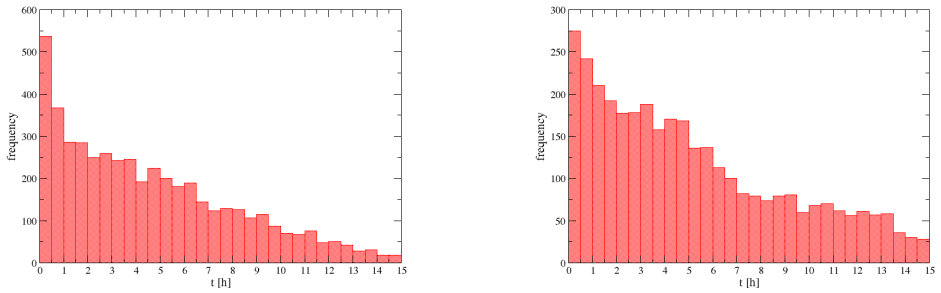


Figure 3.5: Distribution of the elapsed time associated with the daily mobility paths reconstructed during the Carnival (left picture) and the *Festa del Redentore* (right picture) in the Venice historical center.

travel time budget proposed in other studies of urban mobility. The exponential decaying defines two different characteristic lengths, 3.0 km for the Carnival dataset and 3.8 km for the *Festa del Redentore* dataset, and it suggests the propensity of the individuals to perform greater mobility in the last case.
In both cases, these distances are probably greater than the typical pedestrian mobility in a city. However, they reflect the average walking distance in the historical center of Venice, where pedestrian mobility is prevalent. The exponential distribution overestimates short mobility paths since one must cover a minimal distance to satisfy the mobility demand. The presence of short daily mobility paths could also be related to the use of the public transportation system.
To understand the statistical features of the observed mobility, we also consider the mobility time distribution associated with the mobility paths, computed as the elapsed time between the first and the last recorded GPS position of a device in the area of interest (see fig. 3.5).

The mobility time is the sum of the travel times and the rest times. The device activity during the night can affect the distribution tail not directly related to the mobility. It is a reasonable assumption that if an individual has spent more than 8 h in Venice, he has a relevant probability of spending the night in Venice. Indeed, to spend more than 8 h in Venice living outside, one has to add a commuting time between 1 and 2 h and consider the possibility of taking lunch and dinner in Venice, which could be quite expensive.
The exponential interpolation is less justified due to the increased effect of the rest times with respect to the mobility times, and we derive a dynamic model for the relation between the mobility path lengths and the mobility

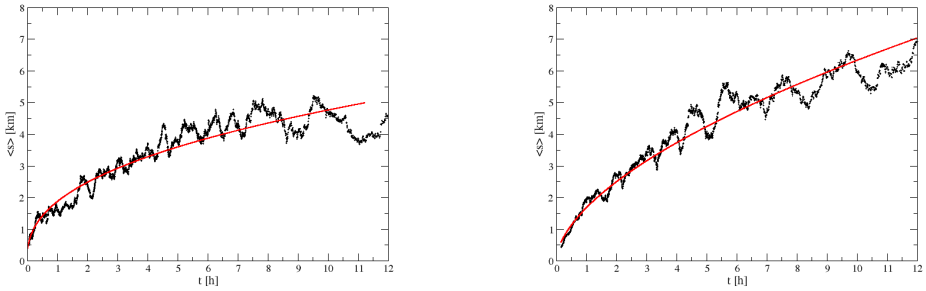


Figure 3.6: Relation between the average path lengths s and the elapsed times: the left picture refers to the Carnival dataset and the right picture to *Festa del Redentore* dataset. The plots are obtained performing a running average of length 100 on the (t, s) data. The continuous line is the result of a power-law interpolation (cfr. eq. (3.1)) with exponents $\alpha = .41$ in the first case and $\alpha = .58$ in the second one, whereas the proportionality coefficient is $\simeq 1.7$ in both cases.

time.

Dynamic properties of the mobility paths

Let us consider an ensemble of an individual moving on the road network; we define the average moving velocity $v(t)$

$$\frac{d < s >}{dt} = v(t)$$

where $< s >$ is the average path length corresponding to a mobility time t. In fig. 3.6 we show the result of an interpolation of the empirical relation between $< s >$ and t through a power law

$$< s >= ct^{\alpha} \tag{3.1}$$

where c is a suitable constant.

In normal conditions, the pedestrian dynamics is performed at a constant velocity v_0, with a stochastic variation among individuals, and a linear relation $s = v_0 t_w$ is expected where t_w is the walking time.
The statistical law $< s > \propto t^{\alpha}$ with $\alpha < 1$, where we average on the path lengths corresponding to a given mobility time t implies that the rest times, defined by the difference $t - t_w$, increase as a function of t. Therefore the relation (3.1) simulates a fatigue effect of individuals during pedestrian mobility. We remark that it is difficult to relate this effect to crowding conditions in the road network unless one could compute a fundamental diagram for the

pedestrian dynamics in the Venice road network. In our opinion, this is possible, but it requires a dataset that includes a long period of observations.
The interpolation of the empirical data gives an exponent $\alpha = .41$ in the case of the Carnival dataset and $\alpha = .58$ in the case of the *Festa del Redentore* dataset. This difference suggests less effective mobility during the Carnival than during the *Festa del Redentore*, probably due to the weather conditions in winter, but also by the many activities that could attract the attention of people.
To relate the empirical observations with a microscopic dynamic model, we propose a relation between the walking time t_w and the mobility time t of the form

$$dt_w = \frac{\alpha dt}{(1 + t/\tau)^{1-\alpha}} \tag{3.2}$$

where τ is a fatigue scale time for pedestrian mobility and $\alpha > 0$ measures the mobility efficiency: $\alpha \to 1$ is the most efficient mobility when space and time are proportional.
The relation (3.2) implies that if $t < \tau$ the mobility time practically coincides with the walking time, whereas the walking time reduces to a small fraction of the mobility time when $t \gg \tau$ as fast as $\alpha \ll 1$.
For a typical visit of 6 h in the Venice historical centre, the formula (3.2) implies that the walking time fraction is $t^\alpha \tau^{1-\alpha} \simeq 2.5$ h for a fatigue time scale $\tau \simeq 1$ h. A simple calculation gives

$$s = t^\alpha v_0 \tau^{1-\alpha} \left[\left(1 + \frac{\tau}{t}\right)^\alpha - \left(\frac{\tau}{t}\right)^\alpha \right] \simeq \bar{v}_0 \tau^{1-\alpha} \alpha t^\alpha \qquad t \gg \tau$$

so that one recovers eq. (3.1)

$$< s >= \frac{\bar{v}_0}{\alpha} \tau^{1-\alpha} t^\alpha \tag{3.3}$$

We remark that the relation (3.3) is singular when $\alpha \to 0$ (i.e. there is no mobility). Moreover, the validity of eq. (3.2) for long times t is questionable since they can be affected by the device activities at home, hotels, or restaurants. The numerical interpolation provides the value

$$\bar{v}_0 \tau^{1-\alpha} \simeq 1.7$$

so that estimating $\bar{v}_0 \simeq .5$ m/sec as a typical average pedestrian velocity, one obtains the fatigue time scale $\tau \simeq 1$ h. This approach provides an analytical formula for the mobility time distribution once the distribution of $< s >$ is known. Due to the significant individual variability in the recorded mobility, the $< s >$ distribution is no longer exponential. The approximation with a

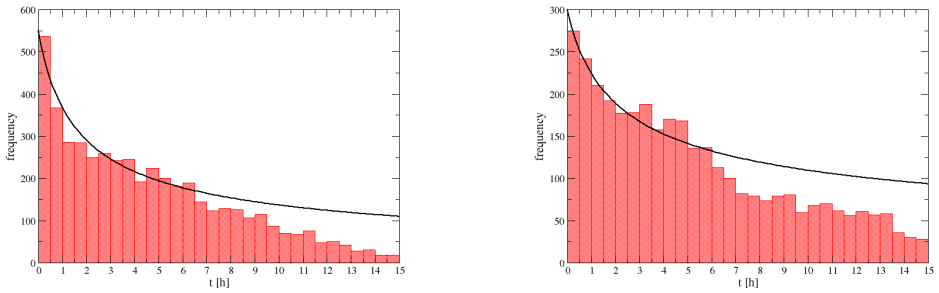


Figure 3.7: Interpolation of the empirical elapsed time distributions by using the analytical distribution (3.4) the left picture refers to the Carnival dataset, whereas the right picture to *Redentore* dataset. The continuous line is the distribution (3.4) with parameter $\alpha = .42$, $\tau = 1$ in the first case and $\alpha = .58$ and $\tau = 1$ in the second one.

constant distribution is reasonable at this stage (see supplementary material). Then one obtains a mobility time distribution of the form

$$p(t) \propto (1 + t/\tau)^{-(1-\alpha)} \tag{3.4}$$

We remark that this distribution is not summable, and we expect validity for a limited time interval.
In fig. 3.7 we show the comparison between the empirical mobility time distribution and the analytical distribution (3.4). The parameters used in the interpolation are consistent with the interpolation shown in fig. 3.7 with $\tau = 1$. We remark as the analytical law provides a quite good interpolation of the mobility time distributions with $t \in [0 : 6]$ h, whereas the distribution tail is still exponential.

Pedestrian mobility network

The reconstruction of the mobility paths also allows studying how people perform their mobility on the road network. We consider the problem of determining the most used subnetwork of the Venice road network. The existence of mobility subnetworks could be the consequence of the peculiarity of the Venice road network, where it is quite easy to get lost if you do not have a map. Therefore people with limited knowledge of the road network move according to paths suggested by internet sites or follow the signs on the roads.
To point out a mobility subnetwork, we rank the roads of Venice according to weight proportional to the number of mobility paths passing through each

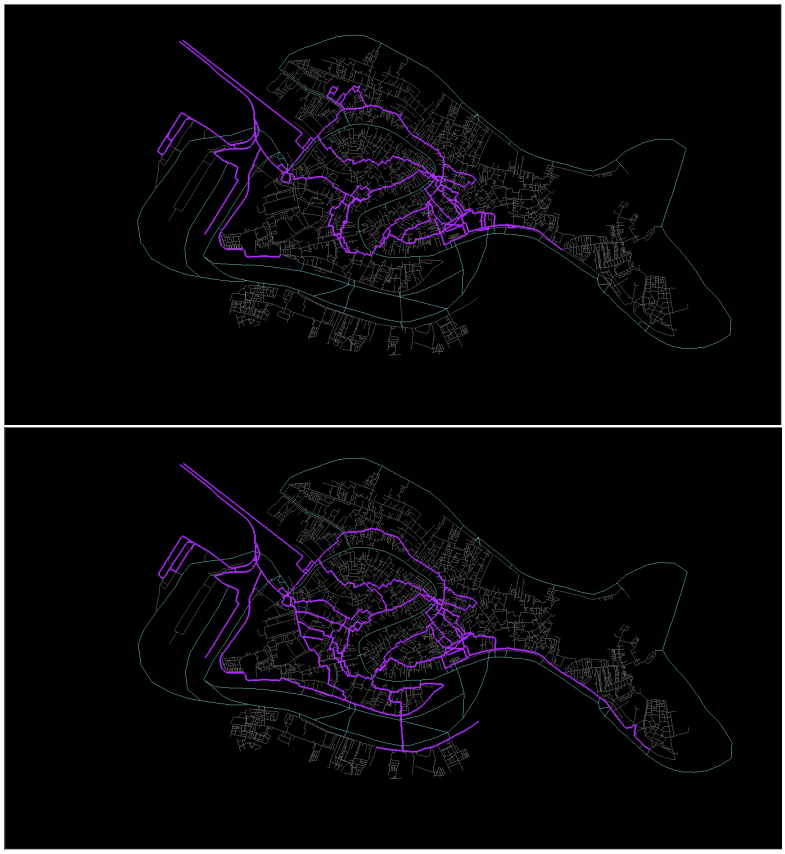


Figure 3.8: Selected subnetworks (highlighted in purple) of the total road network of the Venice historical center (in the background) that explain 64% of the recorded mobility in the datasets. The top picture refers to the Carnival mobility during 26/02/2017 and corresponds to 13% of the total length of the Venice road network. The bottom picture refers to the *Redentore* mobility during 15/07/2017 and corresponds to 15% of the total length of the Venice road network.

road. Then, we have applied an algorithm to extract a connected subnetwork, which contains the roads in the ranking able to explain a fixed percentage of the observed mobility (see supplementary material for a brief description of the main steps of the algorithm). We can extract a subnetwork that explains the 64% of the observed mobility using 13% of the total road network length for the case of the Carnival dataset and 15% of the total length in the case of the *Festa del Redentore* dataset.
The selected road subnetworks are plotted in fig. 3.8 for both the datasets. Many of the highlighted paths are also suggested by internet sites.

However, we remark some differences that can be related to the different nature of the considered events. During the Carnival of Venice, the mobility

seems to highlight three main directions connecting the railway station and the Piazzale Roma (top-left in the map), which are the main access points to the Venice historic center, with the area around San Marco square, where many activities where planned during 26/02/2017.
In the case of the *Festa del Redentore* the structure is more complex due to the appearance of several paths connecting the station and Piazzale Roma with the Dorsoduro district in front of the *Giudecca* island (see map in the supplementary materials). This geometrical structure could have a double explanation: on the one hand, the *Festa del Redentore* introduces an attractive area near the Giudecca island, where the fireworks take place in the evening; on the other hand, the *Festa del Redentore* is a festivity very much felt by the local population, that knows the Venice road network and performs alternative paths.

Foreigners versus Italians mobility

To study the possible effect on the mobility of a greater custom to visit Venice, we divide the devices in the datasets into Italian and foreign devices according to the roaming protocol. The technical details that allow this disaggregation are reported in the Supplementary Material. Of course, we have no guarantees that all the Italians are more used to visiting Venice than the foreigners, but this is a reasonable assumption since many commuter visitors come from neighboring regions during the considered events.
Then we have associated to each road two normalized weights $w_{fo,it}$ proportional to the number of mobility paths of Italians and foreigners on the road itself (the detected Italians are approximately 10 times the foreigners). In this way, we select the roads that are respectively preferred by the Italians and by the foreigners considering the distribution of the difference $w_{fo} - w_{it}$ and introducing thresholds at ± 1 rms and ± 4.5 rms. In fig. 3.9 we plot the results for the two datasets. We remark that not all the highlighted roads are present in the subnetworks in figure 3.8 since it was not possible to connect them using the high-ranked roads in our list.

It is noteworthy to observe that the majority of highlighted roads show a well-defined preference by one of the two populations (i.e., their difference $|w_{fo} - w_{it}|$ is greater than 4.5 rms).
During the Carnival, the foreigners follow a path passing through *Strada Nuova* to reach San Marco square and the Rialto bridge. In contrast, Italians prefer to go through the central part of the Venice historic center. Moreover, we have two clear attraction areas for the foreign people at the *Old Getto* (up left in the picture) and near *Palazzo Grassi* (in the center of the picture). These preferences are also observed during the *Festa del Redentore* except

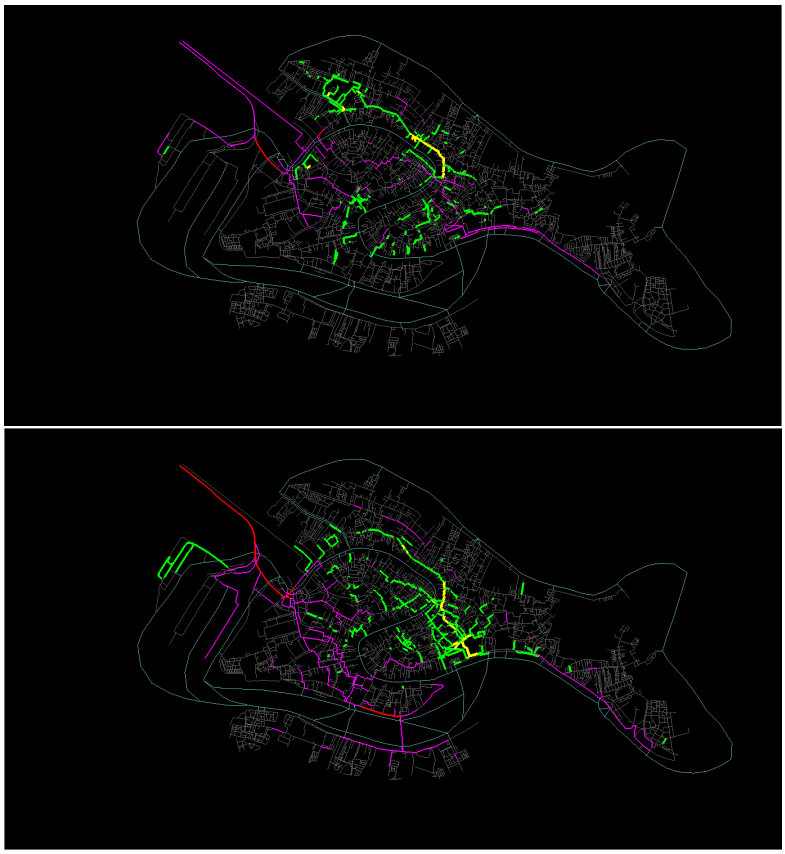


Figure 3.9: Preferred roads of foreigners and Italians in the historical center of Venice during 26/02/2017 (top) and 15/07/2017 (bottom). The foreigners have found more favorite roads highlighted in yellow and green according to the thresholds 1 and 4.5 σ in the weight difference $w_s - w_i$. In contrast, the more favorite roads by Italians are highlighted in red and purple according to the thresholds -1 and -4.5 rms.

for the *Old Getto* that was not pointed out by the algorithm. However, the attractiveness of San Marco square is increased for the foreigners with respect to the Italians that prefer to reach the area in front of *Giudecca* island.
This is consistent with the structure of the mobility subnetwork in this area that seems to be used manly by Italians (fig. 3.8 bottom).

Attractiveness of the main areas of interest

Finally, we analyze the mobility driven by the areas of greatest attractiveness like San Marco square during the Carnival and the Giudecca island during the *Festa del Redentore*.
We select the mobility paths passing through San Marco square (or the Redentore bridge) and reconstruct the mobility network defined by incoming paths. The results are plotted in fig. 3.10: for the Carnival dataset, we select $\simeq 1200$ mobility paths corresponding to the 42% of the total mobility, whereas for the *Festa del Redentore* dataset, we select $\simeq 700$ mobility paths corresponding to 19% of the total mobility.

The highlighted road networks explain the 61% (resp. 54%) of the total pedestrian mobility towards the San Marco square (resp. towards the Giudecca island) in the datasets.
In the first case, the analysis points out three main mobility pedestrian paths starting from the main entry points (the railway station and the Piazzale Roma parking area) that joins near the Rialto bridge. The observed mobility from the Rialto bridge presents a more diffusive character and does not clearly define a path. Then we have an incoming path from the *Riva degli Schiavoni* due to the ferryboat line contribution and a well-defined path between San Marco and the Accademia Bridge (see map in the supplementary materials).
We observe a single pedestrian path from the main entry points towards the Giudecca island in the second case. In contrast, we have various incoming paths along the canal banks, indicating that people arrived by ferryboat. Noteworthy, there is not a clear connection between the San Marco square and the Giudecca island, suggesting that most of the people interested in the *Festa del Redentore* in the evening have not visited San Marco before.

SCR-Smart Control Room

We introduce a short presentation about the functioning of the Smart Control Room (SCR) simulator. We will not dwell on the details of the software, well illustrated in the section 2.3, but we will give a general overview. As reported in figure 3.11, the inputs of each simulation consist of two parts, one static and the other in near real-time. The static inputs involve the knowl-

Figure 3.10: Top picture: the mobility network driven by the attractiveness of San Marco square during 26/02/2017 that takes into account the 61% of the total pedestrian mobility towards the square. Bottom picture: the mobility network driven by the attractiveness of Giudecca island during 15/07/2017 that takes into account the 54% of the total pedestrian mobility towards the bridge.

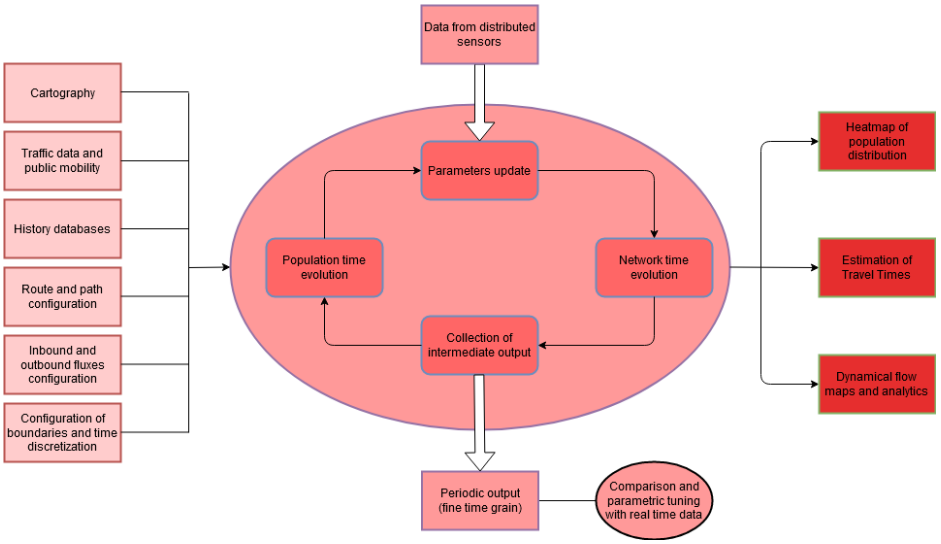


Figure 3.11: It is the flowchart of the Smart Control Room simulator. There are the static inputs to the left, at the top the real-time inputs, at the bottom the periodic output, and at the right the final outputs. In the central block, the simulator work is outlined.

edge acquired in previous research about how pedestrians move in Venice, how they prefer to walk, and what attraction points they want to reach during a city tour. Furthermore, we also have access to the GTFS data for the public transport information (ferryboat trips, stops, and timetables) and the history databases in which we collect the data acquired by the sensors in the past. In addition, there are the near real-time counters from a camera system distributed along the main paths founded in the previous study. Therefore, we can simulate both past and present pedestrian mobility, thus introducing forecasting predictions.
We can collect and dump intermediate output about the fluxes on the streets and the number of passes through some virtual barriers during the simulation. In this way, we can compare simulated with real data and then make a parametric tuning. At the end of the simulation, we can show the behavior in a heatmap of population distribution or in a dynamical map of flow. In addition, we can reproduce the statistical properties of pedestrian mobility, estimating travel times and lengths distribution.

Conclusion

The possibility of recording accurate anonymous georeferenced positions of mobile ICT devices whenever they perform an activity provides dynamic information on the people's mobility on a whole road network. Even if the

requirements to reconstruct reliable daily mobility paths strongly reduce the samples' penetration, we succeed in studying some statistical and dynamic properties of pedestrian mobility in Venice. We explicitly analyze pedestrian mobility during two significant tourist events, but our methodologies apply to any dynamic GPS dataset containing individual mobility on a road network. The historical center of Venice is an ideal experimental field to study the features of pedestrian mobility. The choice of two large tourist events (the Carnival of Venice 2017 and the *Festa del Redentore*) as case studies allows on the one hand to increase the representativeness of the sample and on the other hand to provide quantitative information to the stakeholders that are in charge of the management of tourist flows.

Our results are consistent with the existence of a 'mobility energy .' They point out the relevance of a 'fatigue effect' that reduces the average speed of a mobility path as the mobility time increases. Moreover, the distribution of the mobility paths on the Venice road network allows both to reconstruct connected subnetworks able to explain the majority of the observed mobility and give information on how people use the road network to reach the main areas of interest. These results can also be relevant for realizing a monitoring system of the pedestrian flows in Venice, suggesting where to install the people counting devices and how the local measures can be correlated to the mobility state of the road network.

The possibility of disaggregating Italians from foreigners by the roaming protocol shows some different behaviors that should be further analyzed to understand if they can be related to the different knowledge of the Venice road network. The different features of the two events (the Venice Carnival takes place for two weeks in winter, whereas the *Festa del Redentore* is a religious holiday in summer) are reflected by different dynamic properties of the observed mobility. Our results show the possibility of using the quality of the GPS data on a small sample of mobile devices to build useful tools. These can study the individual mobility at the spatial scale of the road and tune dynamic models of pedestrian flows that perform nowcasting and forecasting of the mobility state of the whole road network to avoid critical states.

We expect that in the next future, the quality and the quantity of GPS datasets provided by the ICT will continuously increase and that their study will contribute to the debate on the development of the Smart City paradigm.

3.2 Case study: Rimini

The cartography used for the analysis of Rimini city is shown in figure 1.2. The bounding box 3.2 defines it and it involves 18,894 polys, 13,842 nodes, and 51,971 arches inside a box of nearly 22×24 km^2.
The notions about node, poly and arc is discussed in section 2.1 and are displayed in figure 2.1.

	Latitude	Longitude
Min	43.96519	12.43720
Max	44.16400	12.73840

Table 3.2: The bounding box of Rimini cartography expressed in minimum and maximum latitude and longitude.

We now observe different results achieved from different kinds of data. In particular, for the city of Rimini, we have MDT data from Telecom and MTS data collected by coils installed along some roads. We have introduced all these types of data in more detail in the section 1.2.

Data analysis and path reconstruction

In this section, we will consider the results obtained with the MDT data. The available data goes from 07-08-2020 to 17-08-2020. This period involves the days just before and after the Italian festivities of *Ferragosto* (15 Aug), which coincides with the Italian's favorite week for the vacation; therefore, we can not find significant differences among the day of the week. Furthermore, due to the COVID-19 disease, in 2020, there was increased regional and domestic tourism.
In the section 2.2.1 and 2.2.2, we illustrated the algorithms used for filtering, georeferencing, and reconstructing the trajectories of each activity. These algorithms are related to specific parameters that we have to set according to the cartography or acquisition technology. For Rimini analysis, according to its cartography feature, we used the parameter reported in the table 3.3.

by referring to the diagram shown in figure 2.7, in this day, we obtain the following numerical value for the results:

- After the data reading with *timestamp* between 00:00:00 and 23:59:59 of 15 Aug 2020 and GPS coordinates inside the box described by 3.2, we collect 1,917,097 activities expressed in 16,438,255 records.

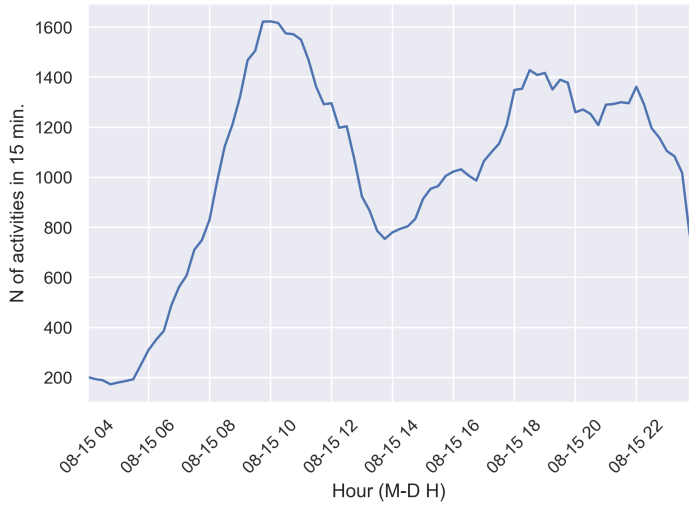
name parameter	value
map_resolution	60.0 (m)
l_gauss	10.0 (m)
max_inst_speed	50.0 (m/s)
min_node_distance	20.0 (m)
min_poly_distance	50.0 (m)
min_data_distance	50.0 (m)
threshold_v	50.0 (m/s)
threshold_t	3600 (s)
threshold_n	3
threshold_polyunique	4

Table 3.3: Values of functional parameters for MDT data analysis in Rimini.

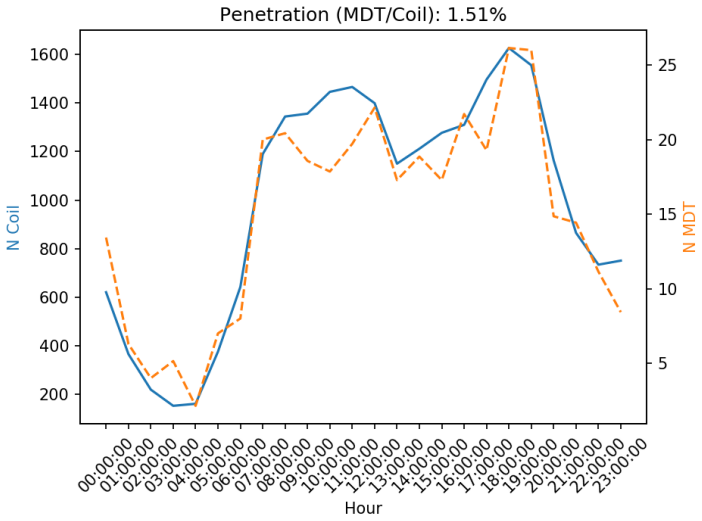
- After the record down-sampling, based on the distance between consecutive points, we keep the 16.35% of records with a total of 2,687,969 points. This substantial reduction shows an aspect of the nature of data, which is that MDT data are very close and frequent, more than is necessary to reconstruct the trips.
- After the georeferencing algorithm, 81.28% of the remaining records are georeferenced, while 18.72% are too far away from any arc of cartography.
- The 50.71% of these georeferenced data forms an activity with just one record, called *Presence* from now on. Then, we obtain 1,362,924 presences.
- From the remaining 30.58%, the 18.26% pass the threshold, and ultimately we obtain 42,880 trajectories during a single day.

Estimation of sample penetration

The plots in figure 3.12 show the statistical goodness of this kind of data after our filtering and reconstruction procedure. Indeed the plot (a) shows the number of active trajectories in 15 minutes during a single day (15 Aug 2020). The evidence of a structure with two main peaks, in the morning and evening, confirms the data goodness in the emergence of the typical circadian rhythm. In plot (b), we report a comparison done for a street in which a coil is present (see sub-section 1.2.2) to estimate the penetration of this dataset. The curves represent the average trend for the period 10-16 Aug 2020 with an hourly bin: the solid blue line is those recorded by the coil, and the dotted



(a)



(b)

Figure 3.12: (a) Number of active trajectories in 15 minutes during 15 Aug 2020. (b) The average trend for the period 10-16 Aug 2020 with an hourly bin: solid blue line refers to data recorded by the coil; dotted orange refers to MDT data. The penetration factor is equal to 1.51%.

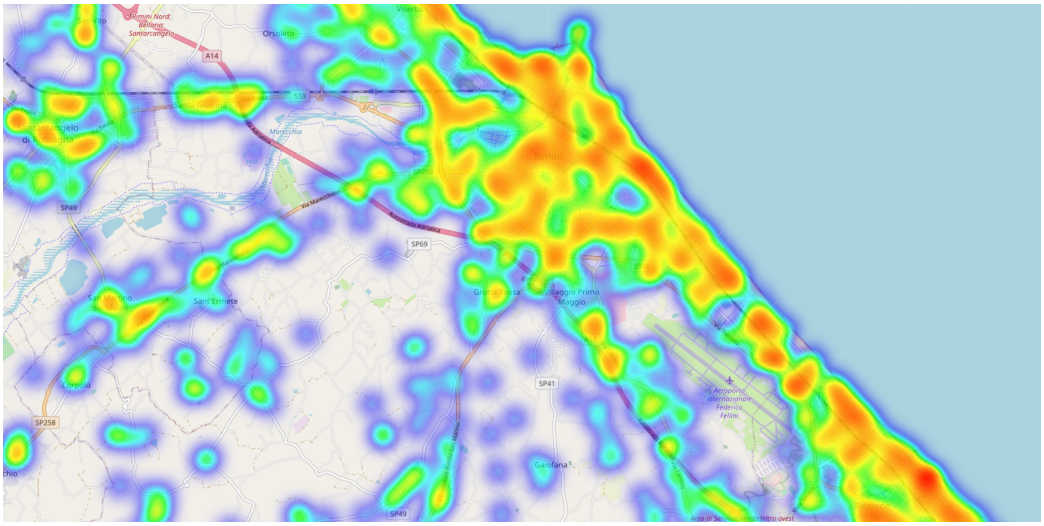


Figure 3.13: Heatmap for the center of Rimini obtained with trips that, after the analysis, collapse to a single point called presence, on 15 Aug 2020. The heatmap shows only the coordinates that has more than 20 occurrences performed by different *activity_ ids* throughout the day.

orange line is those reconstructed from MDT data. We can see that the two trends are very consistent, with a penetration factor of 1.51%.

Spatial distribution of Presences

We used just the trajectories for the study of mobility, but the presences, which have large numbers, can provide information about the spatial distribution of activities. For example, the heatmap in figure 3.13 shows the areas of increased activity.

Trajectory statistical properties

Later, the path of trajectory is reconstructed using the *Best Path algorithm* already discussed in 2.2.2. Since we know the length path for each trajectory, as the distance between its points, and time path, like the time interval between last and first record, we can measure the average speed of the trip.
In figure 3.14, we can observe the distribution of these quantities from data analysis from 10 Aug 2020 to 16 Aug 2020, after excluding the trips with a duration of fewer than 1 minute, since they do not represent the real mobility, but they are just pieces of paths. These pieces are inevitable in this kind of data since the records are stored when the user makes a phone activity. While datasets like those of Bella Mossa or Octo telematics have the advan-

tage of being closely related to the actual start and finish of travel by the user.
We note that length distribution (a), expressed in semi-log scale, reveals a power-law decay, particularly for the selected range between $1.5km$ and $10km$. As is known, this behavior is typical of a situation in which a mixture of different mobility strategies and transport means is present. As we will see below, we need to disaggregate the activity for different transport means to obtain an exponent for each kind of modality.
In the same way, the time distribution decay (b) follows a power-law function with a power factor of 1.54. The existence of different mobility strategies is confirmed by the average speed distributions, in which is evident the emergence of more than one pick. In particular, it reveals the presence of three areas: the first main pick at a low speed comparable with the slow-mobility of pedestrians and bicycles, a sort of bump in the central part in which different urban means of transport overlap, and finally a pick at hight speed representing the faster mobility of car trips on extra-urban streets and highways.

The reconstruction of activity paths allows measuring the number of passes along any roads and highlighting the parts of the network more widely used globally and by a specific class (attributable to one or more transport).

In figure 3.15 (a) we can show the roads most crossed during the 15 Aug 2020 with a threshold scale that defines the color code. In particular, we measure the quantity:

$$\sigma = \sqrt{\frac{\sum_{i=0}^{N_{poly}} {cnt_i}^2}{N_{poly}}} \tag{3.5}$$

where cnt_i represents the number of people that cross the poly during the analysis in both directions. Then, we obtain that:

- Yellow polys are those with $cnt_i > 4.5\sigma$; the most widely used.
- Red polys are those with $2.5\sigma < cnt_i \leq 4.5\sigma$.
- Green polys are those with $1.5\sigma < cnt_i \leq 2.5\sigma$.
- Blue polys are those with $1.0\sigma < cnt_i \leq 1.5\sigma$.

The plot (b) represents the distribution of crossings number per poly during the same date. As expected, the trend seems to follow a power-law decay; therefore, there are few polys with many crossings, and the vast majority of them have a small number of passages. The vertical lines correspond to the threshold limit value: from blue to yellow, respectively, σ, 1.5 σ, 2.5 σ and

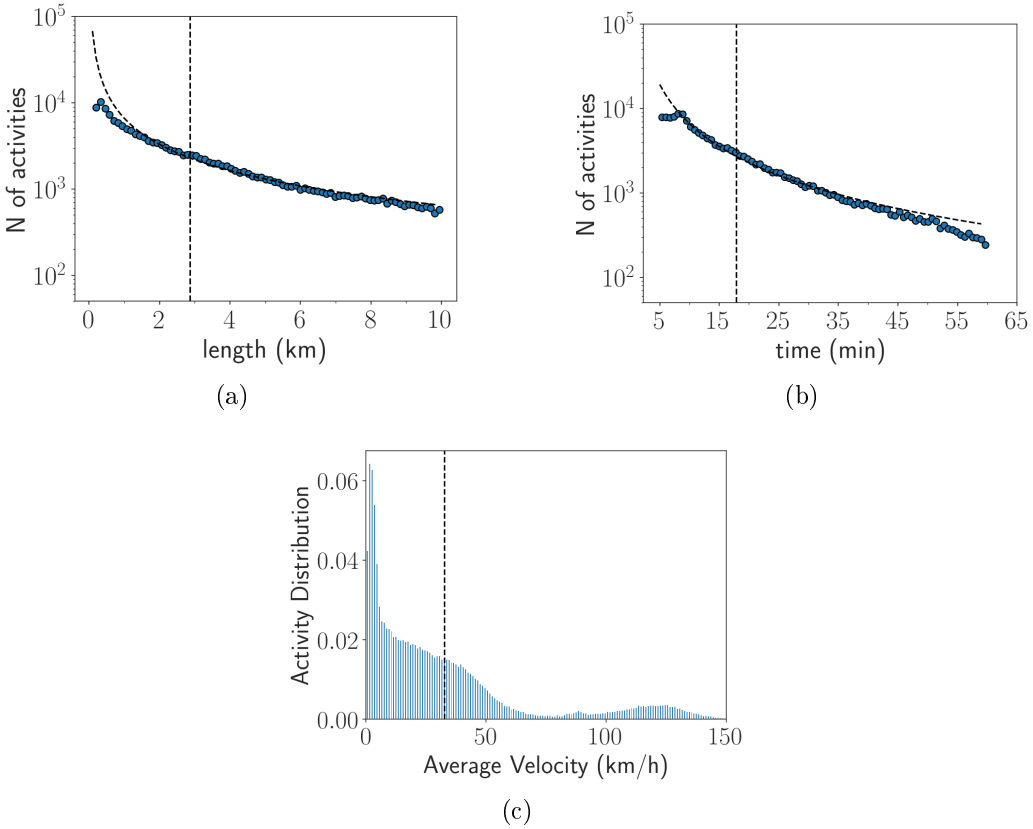
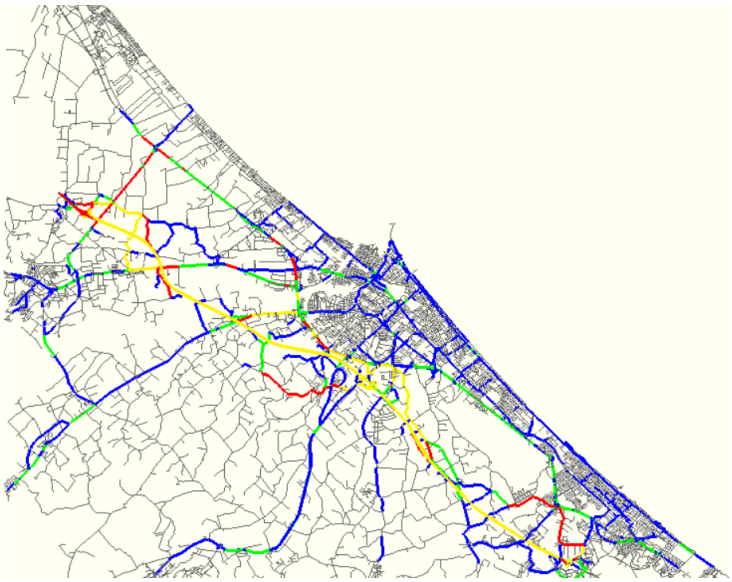
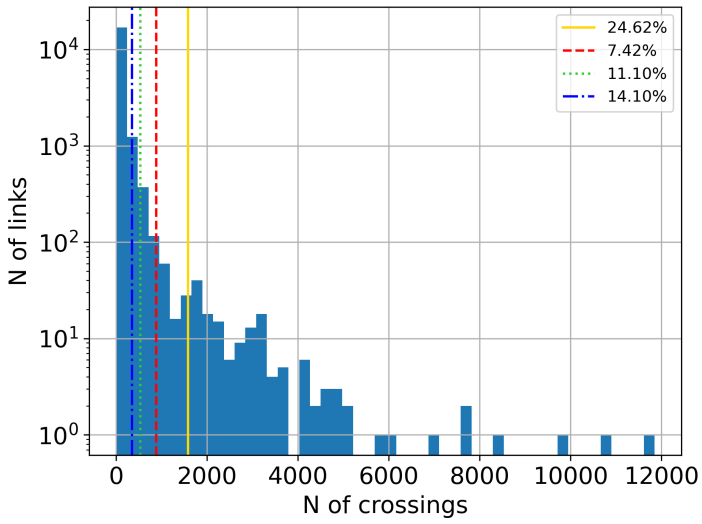


Figure 3.14: Analysis of Rimini MDT data from 10 Aug 2020 to 16 Aug 2020. (a) Trip length distribution with average $\simeq$ 2.83 km. Power-law fit parameters: $a = 6.63 \times 10^3$, $b = 1.0$ with $R^2 = 0.986$. (b) Trip time distribution with average $\simeq$ 17.8 minutes. Power-law fit parameters: $a = 22.9 \times 10^4$, $b = 1.54$ with $R^2 = 0.992$. (c) Trip average speed distribution with average $\simeq$ 32.5 km/h.



(a)



(b)

Figure 3.15: (a) Map of Rimini road-network on 15 Aug 2020 coloured according to the number of crossings per poly (cnt_i). Yellow: $cnt_i > 4.5\sigma$ (most widely used). Red: $2.5\sigma < cnt_i \leq 4.5\sigma$. Green: $1.5\sigma < cnt_i \leq 2.5\sigma$. Blue: $1.0\sigma < cnt_i \leq 1.5\sigma$. (b) Distribution of the number of crossings per poly on 15 Aug 2020. The vertical lines correspond to the threshold limit value. In the legend, the percentage of mobility is explained by polys that fall in the range of thresholds above.

4.5 σ. In the legend, the percentage mobility is explained by polys that fall in the range of thresholds above. Furthermore, we calculated the percentage of kilometers involved in each threshold range, and we found from the greater to smaller sigma, respectively the values: 3.10%, 1.27%, 2.60% and 3.53%. This means that almost the 57% of mobility is explained by just the 10.5% of network length, showing again how unbalanced the usage of roads in a city is.

Multimodality classification

Considering the previous observations about global speed distribution, we used the classification approach described in section 2.2.3 to disaggregate the trips according to the mobility strategy to reveal statistical differences and a diversified road network usage.
In this case, we try to find four different clusters (2 for slow mobility and 2 for the faster), and we use the following four features for each trip:

1. *Average speed of trip*: the ratio of trip length, measured as the sum of the distance between consecutive points and trip time, which corresponds to the difference between the last and first record of the trip.

2. *Maximum speed of trip*: the maximum average of instantaneous speeds measured among consecutive data with a settable window's size. In this case, this window has dimension 2.

3. *Minimum speed of trip*: the minimum average of instantaneous speeds measured among consecutive data with a settable window's size. And even then, the dimension of windows is 2.

4. *Sinuosity*: the ratio of the euclidean distance between the first and last record of trajectory and the length measured as the sum of all record distances. This value can be a maximum of 1.0 and indicates how curvy the path is: the smaller the sinuosity, the more tortuous the chosen path.

We noted that the usage of soft clustering allows us to obtain, for any data, the degree of membership to each cluster. We choose to impose a threshold for this degree after the classification. In particular, we assign to a trip the class which corresponds to the maximum degree, as long as it is

greater than 0.5. Otherwise, the trip is considered unclassifiable.
In this case, we obtain over 42,875, 41,277 result classifiable (the 9.36% of the available trajectories) with a *Dunn index* of 1.19×10^{-3}.

class number	average speed (km/h)	maximum speed (km/h)	minimum speed (km/h)	sinuosity
class 0	4.28	11.26	3.69	0.74
class 1	18.87	40.53	15.24	0.78
class 2	43.39	74.46	33.50	0.88
class 3	112.54	141.30	91.88	0.97

class number	occurrences
class 0	7826
class 1	7849
class 2	17850
class 3	7752

Table 3.4: Values of centers of clusters and the number of trips members of each class for MDT data in Rimini on 15 Aug 2020.

Table 3.4 reports some information about each class. In the first row, there are the values of the cluster's center; in the second one, there is the number of trips members of that class.
As anticipated from the global speed analysis, we found different speed regimes compatible with pedestrians, bicycles, and cars. In particular, the slowest class (class 0) has velocity values comparable with a mix of pedestrian and bike trips. In contrast, the other three classes have velocities attributable to car trips with different regimes. It is also noted that the greater the velocity, the greater the sinuosity. We can explain this relationship with the road network structure: the streets and then the trips performed at high velocity are located in the city's periphery, where the number of curves and deviations is minor than possible. On the contrary, polys are short in the heart of the city, and the turning points are very frequent.

In order to highlight the different usage of the road network, according to the kind of class, we used the algorithm described in the supplementary material of section 3.1 for subnetwork creation. Indeed, knowing the number of passages through each poly for each class, it is possible to define a ranking of the streets based on their usage. The algorithm allows us to extract a subnetwork fully connected and including a certain percentage of these sorted polys. Therefore, with a loop on ranked poly, we found the last set of them for which the ratio of the quantity $cnt_{ij} \cdot length_i$ where $i = 0...N_{poly}$, $j =$

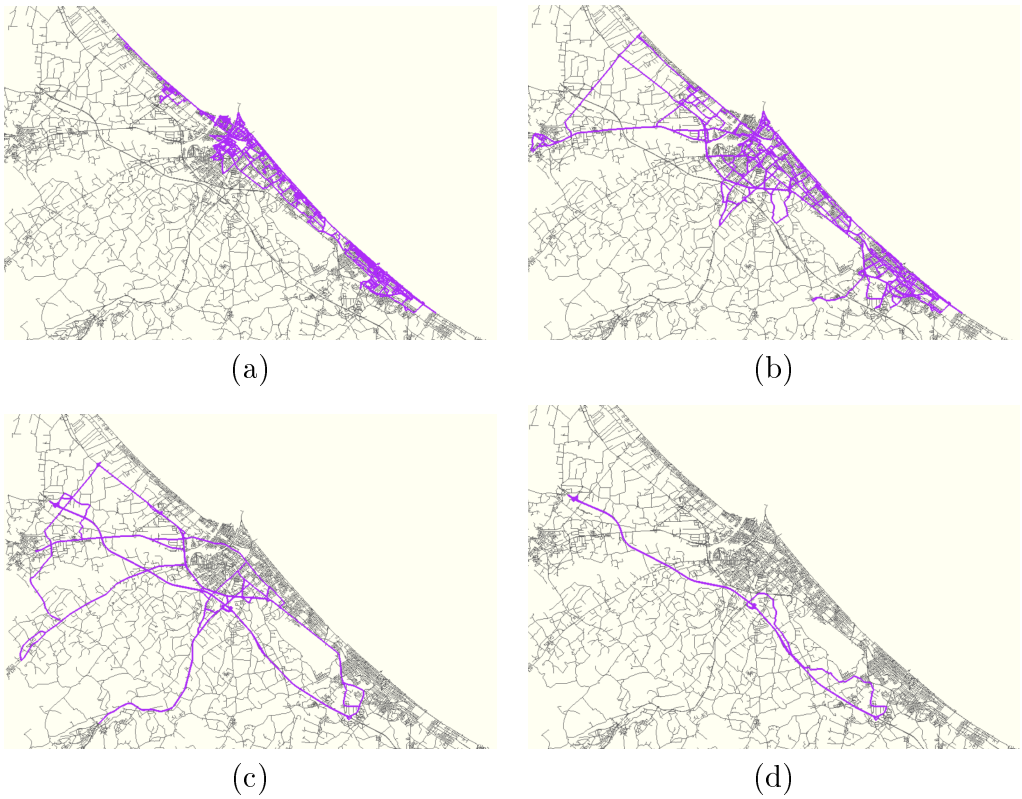


Figure 3.16: Subnetworks for each class obtained with data of 15 Aug 2020: (a) class 0 (67% of explained mobility), (b) class 1 (64% of explained mobility), (c) class 2 (68% of explained mobility), (d) class 3 (86% of explained mobility).

number class	L_m (km)	T_m (min.)	V_m (km/h)
class 0	1.17	19.21	4.91
class 1	2.85	13.74	16.54
class 2	6.64	13.29	36.00

Table 3.5: Average length (L_m) , time (T_m) and speed (V_m) for each class in the period 10-16 Aug 2021.

$0...N_{classes}$ between the selected set and the whole network is less than a value P. The value of this latter is set to 0.70 for the first three classes and 0.80 for the last one, and it represents approximately the mobility explained by each subnetwork.
In 3.16 we observe the resulting subnetworks arranged from the slower to the faster class (0-(a), 1-(b), 2-(c), 3-(d)). The plot shows how the polys involved moving from the city center to the periphery as speed increases. In particular, the last subnetwork includes exactly the motorway network, as expected from speed distribution.

Figure 3.17 reports the distributions of normalized length L/L_m, normalized time T/T_m, and average velocity (km/h) for disaggregated trips according to the classification. We excluded the fastest class (class 3) from this analysis. Indeed, it is performed on the highway, so the cut of the bounding box biases its length and time distributions. Furthermore, this mobility takes place on roads used exclusively: the interaction effects due to the simultaneous coexistence of more regimes and modalities are not present.
We observe that normalizing the L and T distributions on their average values, reported in table 3.5, the three distributions tend to collapse in the same exponential decaying: $y = Ae^{-Bx}$. The value of B is almost always close to 1; therefore, the average value of L and T corresponds to the characteristic length and time. The disaggregation of speed distributions shows three distinct peaks, and we can note that those relating to class 2 have a wider and less pronounced peak. This shape is probably due to this class's hybrid nature, which selects both faster bikes and slower cars.

Macroscopic Fundamental Diagram

Inspired by the MFD (Macroscopic Fundamental Diagram) analysis, we want to show if phenomena due to increased network load are present. Therefore we select a smaller bounding box (fig. 3.18) in the urban part of the city where we expect the more visible traffic effect.
The idea is to observe the relationship between an average macro speed and

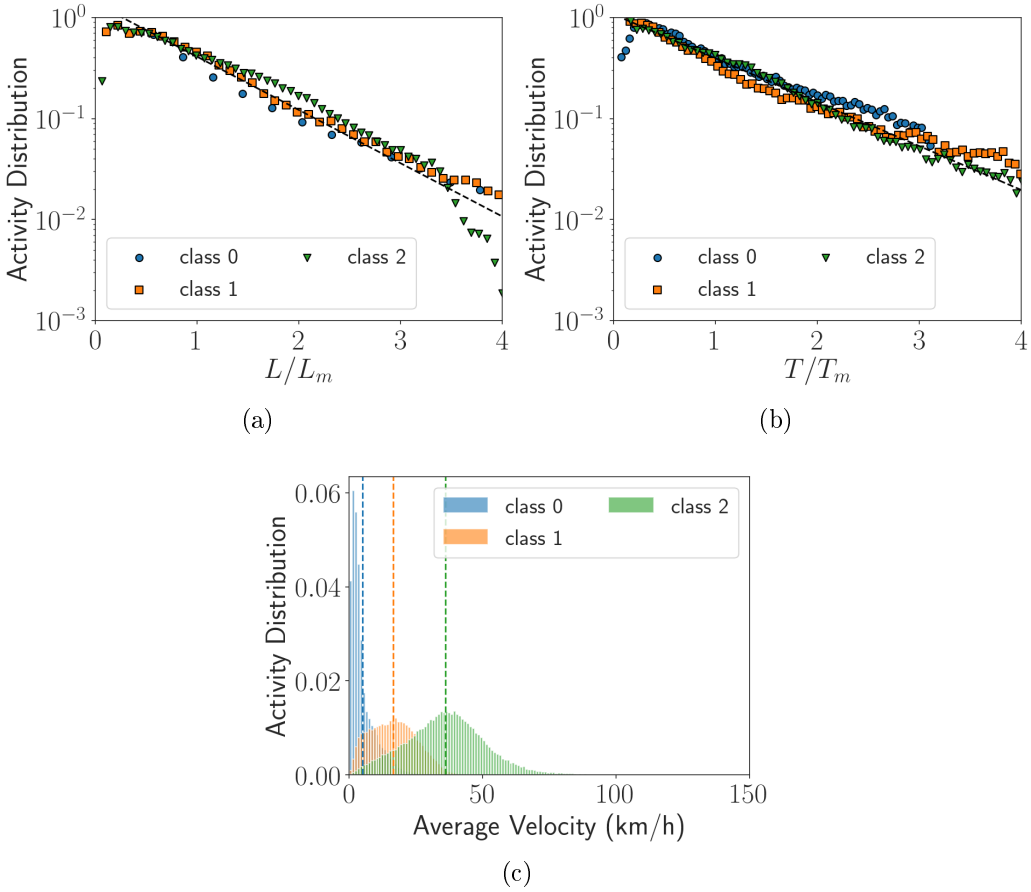


Figure 3.17: Distributions of normalized length L/L_m (a), normalized time T/T_m (b) and average velocity (km/h) (c) for disaggregated trips according to the classification with cluster's centres reported in 3.4. The average value are shown in 3.5. Fit parameters: $A_{L0} = 1.49$, $B_{L0} = 1.44$, $A_{L1} = 1.43$, $B_{L1} = 1.19$, $A_{L2} = 1.20$, $B_{L3} = 1.01$, $A_{T0} = 1.11$, $B_{T0} = 0.95$, $A_{T1} = 1.03$, $B_{T1} = 1.06$, $A_{T2} = 1.07$, $B_{T2} = 0.99$.

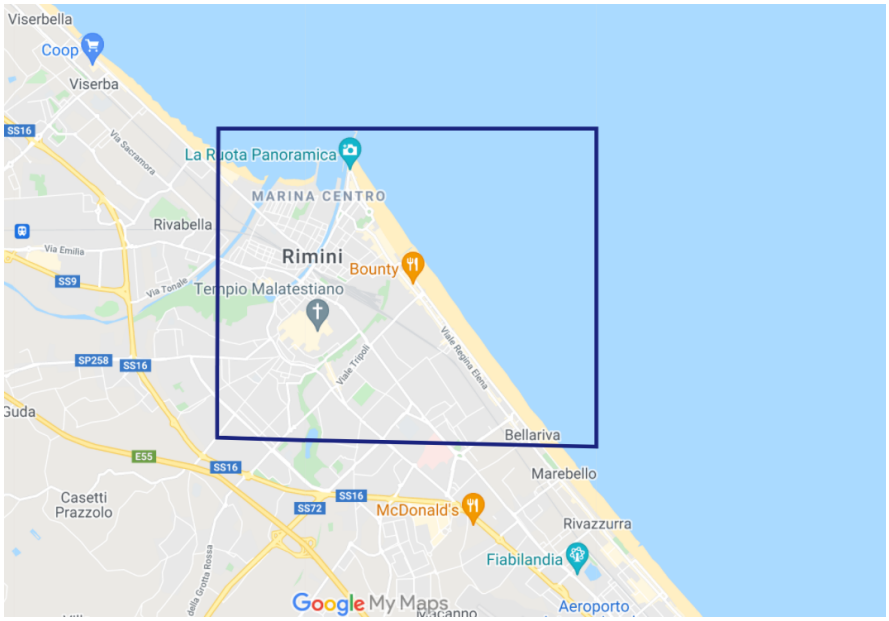
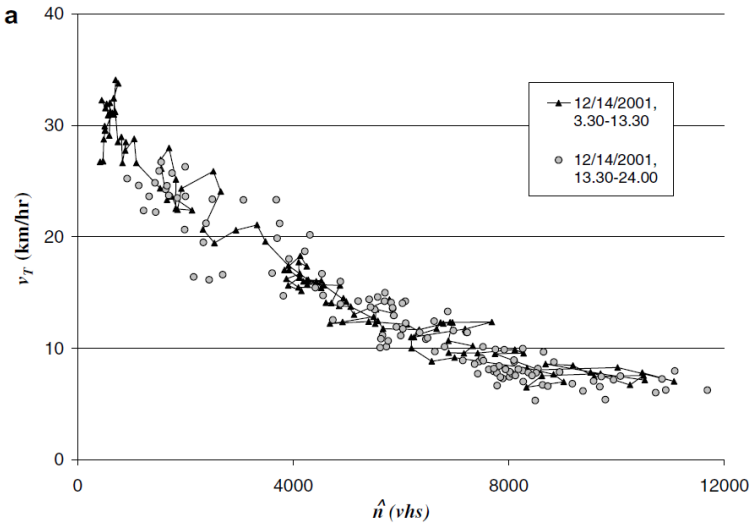


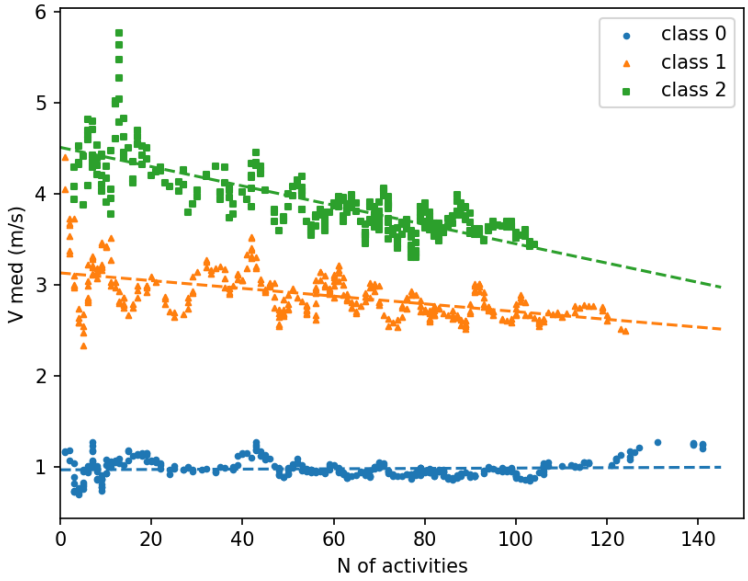
Figure 3.18: The smaller bounding box for MFD analysis of Rimini. $Lat_{min} = 44.04698$, $Lat_{max} = 44.08189$, $Lon_{min} = 12.55489$, $Lon_{max} = 12.61257$.

the number of users involved in the mobility at a specific time bin. From literature, we know that this speed decreases as the density increases following a hysteresis loop and shows traffic congestion. An example is shown in figure 3.19 (a) taken from [47], in which the authors analyzed a dataset of cabs journeys in a particular commercial area of Tokyo called Yokohama on 14 Dec 2001.

Then, for the period from 10 Aug 2020 to 16 Aug 2020, for each bin with $\Delta t = 30\ min$, we measure the total trip length and the total time of travel at that bin, according to each class. In this way, we obtain a sort of macro speed, given by the previous L and T ratio, every half hour. In addition, knowing the number of users involved in the activity of each bin, we can get the relation between speed and activity number shown in figure 3.19 (b). The increasing number of activities does not affect the macro speed for the first two classes related to pedestrian and bike mobility. Instead, when we observe the car class, the slope of the fitting is greater in absolute value indicating a stronger correlation between the number of cars and macro speed. We cannot observe the saturation due to traffic congestion, having just partial tracks of an entire trip and a minimal penetration factor. However, the different behavior among classes is further confirmation of the classification goodness.



(a)



(b)

Figure 3.19: (a) Image taken from [47]. Yokohama's estimated MFD: Scatter plot of v_T vs N. (b) Rimini's estimated MFD: Scatter plot of V_{med} vs N disaggregated for class. Blue circles correspond to pedestrians, orange triangles to hybrid slow mobility, and green squares to urban vehicles. Data have been subjected to moving average with windows size equals to 10. Interpolation parameters: $A_0 = 0.88$, $B_0 = -0.15 \times 10^{-2}$, $A_1 = 2.68$, $B_1 = -0.43 \times 10^{-2}$, $A_2 = 4.14$, $B_2 = -1.12 \times 10^{-2}$

Conclusion

The opportunity to use a dataset like those provided by Telecom and to have a tool able to disaggregate the different mobility strategies could represent a fundamental instrument for acquiring and sharing knowledge with stakeholders to improve the sustainability of mobility demand in future smart cities. Indeed, we have seen how also starting from a dataset full of redundant records, it is possible to extract just helpful information about multimodality dynamics through a toolchain of algorithms and procedures capable of downsampling, georeferencing, and filtering just the significant trajectories. Suffice it to say that we started from approximately 2 million activities per 16.5 million records. In the end, we obtained about 42 thousand trajectories for a single day: this reduction of the amount of data corresponds with the emergence of helpful information. Different expected results confirmed by observations prove the goodness of the analysis. First of all, as we saw in fig. 3.12 (a), the time evolution of the number of activities during the 15 Aug 2020 reveals the typical circadian rhythm with two peaks: the first one in the morning and the second one in the late afternoon. Previous results have shown that this shape of the curve is characteristic of industrialized cities [27, 51, 52]. In particular, we observe the vehicular transition from a specific poly in which a coil of regional MTS system is placed. Since this coil detects each passage from the street, the comparison between the flux measured by our analysis and the value stored by the coil enables us to know the penetration factor equal to 1.51%. It is necessary to specify that this is a quantity measured locally, and there is no guarantee that this percentage is consistent on the entire network. Nevertheless, the shapes of the curves seem to reproduce the same daily evolution.

We analyzed the statistical properties for all the trajectories before multimodality disaggregation. We found the typical power-law decay, a sign of the coexistence of more mobility strategies. In particular, the speed distribution reveals the presence of more than one pick, each of these related to a specific modality. The first visualization of most crossed polys (fig. 3.15) shows that there exists a small percentage ($\simeq 10.5\%$) of polys describing a significant percentage of mobility ($\simeq 57\%$), underling the existence of recurrent paths. Later, the classification of modalities through the FCM algorithm has allowed us to disaggregate different kinds of mobilities that we can interpret with different means of transport. As reported in table 3.4, class 0 is the slowest, and it can be attributed to pedestrian mobility. Class 1 represents slow mobility, including different kinds of urban modalities like runners, buses, or bikes. This interpretation is evident if we observe fig. 3.17 (c) where the class speed distributions are reported: we note that class 1 has a large bell shape due to the presence of various means of transport. The next step could be to

search for new features able to disaggregate the element of this class. Class 2 represents the typical vehicular mobility in and around the city, while class 3 represents vehicular mobility at higher speed. Accordingly, we expect that it takes place on highways and does not affect the other class mobility. We see that the classes are distributed almost equally, except for class 2 that has more than double the occurrences compared to the other classes. It could be because people use navigation apps (keeping a constant connection) frequently when they drive. Furthermore, since there exists a filter based on distance traveled in our software, a rapid connection to the email or social networks will be discarded if performed by walking rather than on a car.
The disaggregation enables us to identify the fully connected subnetworks most commonly used from each class shown in fig. 3.16. We note that the subnetwork (d) of the last class interests just the motorway section present in our bounding box and some entry lanes. Therefore we choose to discard this class from the statistical properties analysis, having this characteristic very different. In general, we note that these subnetworks explain a very high percentage of representing mobility. In particular, the networks of classes 0, 1, and 2 have some links that belong only to the class and others in common with the other classes.
The idea to have a helpful subnetwork to control and manage the behavior of a city is fascinating, and often it is a goal that especially interests the local authorities. In this perspective, the next step could be to identify some control points on this subnetwork and analyze how the correspondence changes are correlated. Furthermore, it might be interesting to find some features that characterize the subnetwork according to representing mobility.
In fig. 3.17 (a) and (b) are reported the distributions of normalized length L/L_m and time T/T_m respectively. We note that they decay following the exponential law: this is a well-known finding in the literature [27, 53]. The simulation software, implemented during the Ph.D. work and described in 2.3, models and simulates the behavior of people moving with a specific means of transport. Studying different statistical properties per transportation allows us to change the setting parameters under these findings and then simulate various mobility strategies. The next step will be to move toward a multiplex simulation in which each mobility strategy moves on a specific subnetwork, and the interaction between them are possible; for example, we can simulate the travel of a user that reaches a bus stop on foot and then continues his journey with the bus.
The last analysis regards the relationship between average speed and density did following the MFD proposed in [47]. We selected a minor area of Rimini in which there was higher activity of the first three classes, shown in fig. 3.18. The fig. 3.19 shows the results: we can note that the first class, attributable

to the pedestrian mobility, is not affected by congestion phenomena. However, instead, class 2, interpreted with the vehicular mobility, has a slope value equal to -1.12×10^{-2}. This result reveals a greater effect due to the traffic conditions.

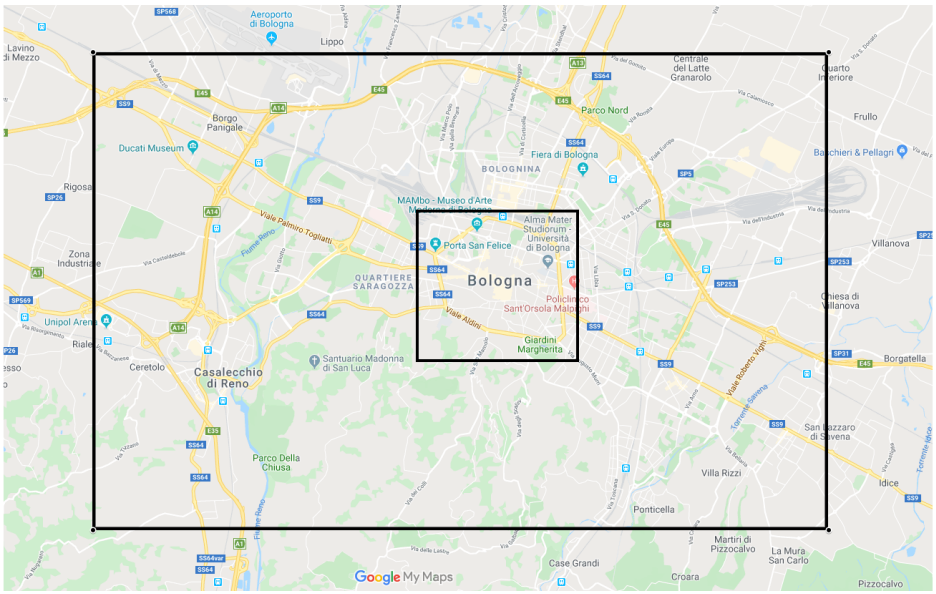
3.3 Case study: Bologna

The cartography used for the analysis of Bologna city is that shown in figure 1.3. It is defined by the bounding box 3.6 and it involves *13257* polys, *10064* nodes and *32677* arches inside a box of nearly $9 \times 14\, km^2$.
The definition of the notions of node, poly, and arc is discussed in section 2.1, and it is displayed in figure 2.1.

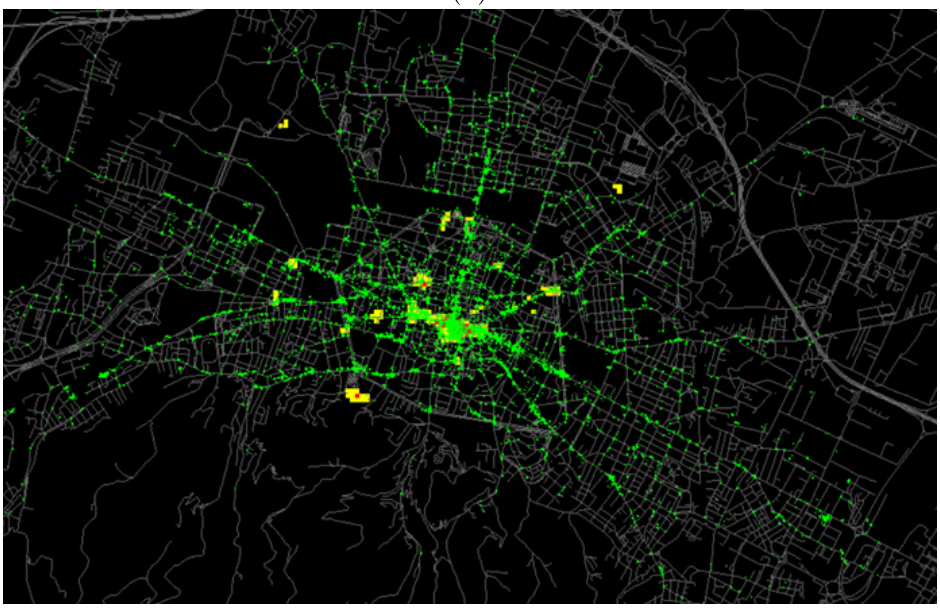
	Latitude	Longitude
Min	44.4522	11.2449
Max	44.5340	11.4213

Table 3.6: The bounding box of Bologna cartography expressed in minimum and maximum latitude and longitude.

An analysis of multimodal mobility requires considering different data sets available in the same urban context. In the city of Bologna (Emilia-Romagna region, in North Italy), we take advantage of the availability of datasets on urban traffic that contains a sample of anonymized trajectories with GPS quality. On the one hand, we had access to the Bella Mossa dataset recorded on a period of 6 months during 2017 (from April to September) that contains information on the pedestrian and cycling mobility (see description at 1.2.3) and, from the other hand, a dataset on vehicle trajectories recorded by insurance reasons during September 2016 (see description at 1.2.4).
Both datasets consider the mobility in the area shown in the figure 3.20 (a) where we have the external box defined by table 3.6 and an internal that contains the historical center of Bologna.
To show the quality of the dataset, we plot in the figure 3.20 (b) the distribution of the initial points of the trajectories whose endpoint is in the historical center: the point distribution highlights the attractiveness of the historical center for slow mobility.
Using the algorithm shown in section 2.2 it was possible to extract the trajectories from the GPS records for both datasets. We observed that we obtained an average number of $\simeq 4000$ trajectories recorded each day for the Bella Mossa dataset and a total of 136000 bike trajectories and 310000 pedestrian trajectories after a filtering procedure on the very short trips. Even if we do not control the statistical sample, the number of recorded trajectories is sufficiently significant to highlight the statistical properties of the bike and pedestrian mobility in Bologna.
In the section 2.2.1 and 2.2.2, we illustrated the algorithms used for filtering,



(a)



(b)

Figure 3.20: Picture (a): map of the Bologna metropolitan area; the larger rectangle encloses the area that has been considered by the mobility analysis, whereas the smaller one shows the historical center of Bologna ($lat\_max = 44.50591$, $lat\_min = 44.48460$, $lon\_max = 11.36208$, $lon\_min = 11.32483$). Picture (b): distribution of the GPS points of the Bella Mossa dataset which belong to trajectories with endpoints in the Bologna historical center.

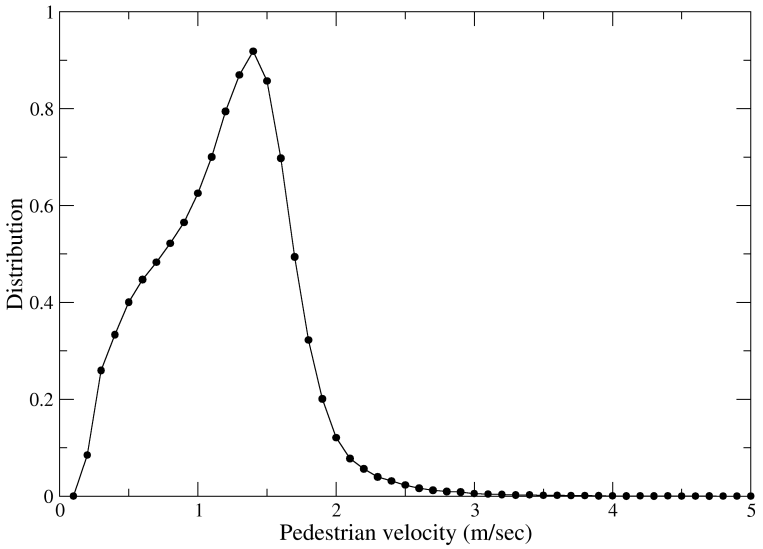
georeferencing, and reconstructing the trajectories of each activity. These algorithms are related to specific parameters that we must set according to the cartography or acquisition technology. For both datasets on Bologna cartography, we used the parameter reported in the table 3.7.

name parameter	value
map_resolution	60.0 (m)
l_gauss	10.0 (m)
max_inst_speed	50.0 (m/s)
min_node_distance	20.0 (m)
min_poly_distance	50.0 (m)
min_data_distance	50.0 (m)
threshold_v	50.0 (m/s)
threshold_t	86400 (s)
threshold_n	3
threshold_polyunique	2

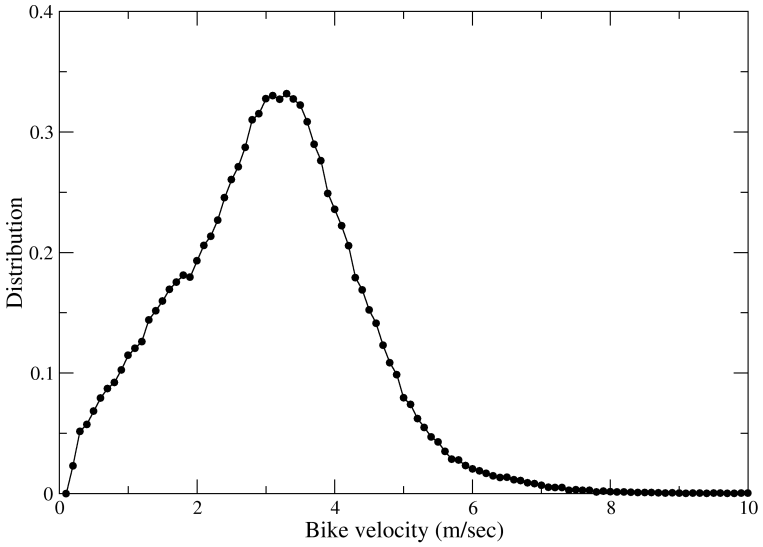
Table 3.7: Values of functional parameters for MDT data analysis in Bologna.

In this way, we can dynamically reconstruct the single trajectories. We have checked the information on the mean transportation by comparing the average velocity distribution associated with each trip with the expected typical velocities of a pedestrian or a bicycle.

In the Figure 3.21 we plot the velocity distributions for the pedestrian and bike trips: we observe that both the distributions are peaked at values consistent with an average pedestrian velocity ($\simeq$ 1.2 m/sec) and an average bike velocity ($\simeq$ 3.1 m/sec). Moreover, the pedestrian velocities are abruptly decreasing over 2 m/sec as is expected, whereas the bike velocities distribution has a larger variance due to the perturbations of the road traffic. Indeed, the very low average bike velocity could be the consequence of the traffic rules or short stops during the trip.
The velocity distributions indicate that the information in the dataset is, in general, correct, and we have studied the statistical properties of the recorded mobility according to this assumption. We remark the different features of the distribution. In the first case, we have an average velocity of 1.19 m/sec (with a peak $\simeq$ 1.4 m/sec) and a sharp distribution decay for velocities higher than 2 m/sec due to the physical limit of the pedestrian velocity. In contrast, the variability of low velocities is due to the many stops that may occur during the trip. In the second case, we have an average velocity of 3.05



(a)

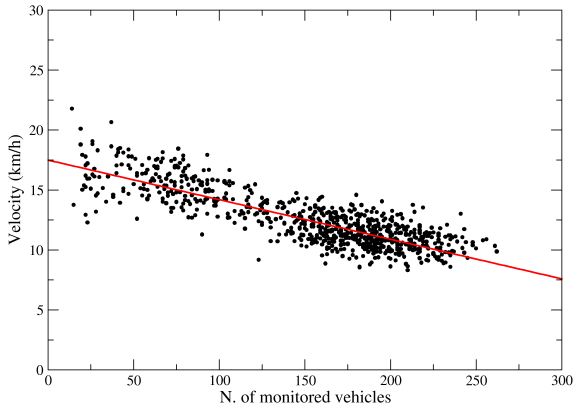


(b)

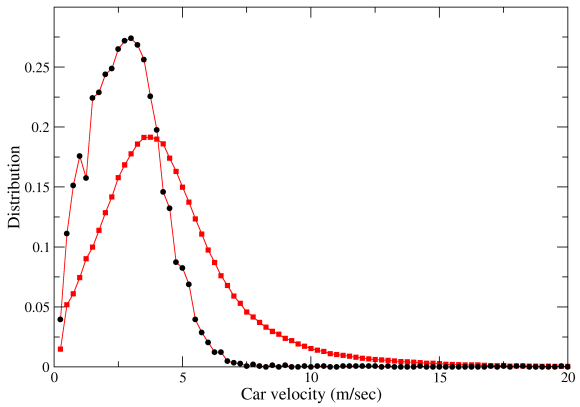
Figure 3.21: Average velocity distribution of the pedestrian trips (a) and of the bike trips (b) computed using the Bella Mossa dataset.

m/sec, which is also the distribution mode. However, the dispersion of the distribution is larger with an exponential decaying for greater velocities.
For what concerns the Octo telematics dataset (OT), we use the same parameters reported above. However, in addition, to avoid an over counting of the short trips (one could switch off the engine during short term stops), we apply an algorithm that glues the trips when the stopping time is less than 1 minute, and there is a continuity in the direction of the successive trip.
To check the quality of the OT dataset, we computed the fundamental diagram for the car trips recorded in the historical center (HC) area of Bologna (see figure 3.20) to point out the effect of traffic load on the average velocity. We have restricted the analysis to the trips inside the HC during the working days to consider the traffic dynamics on a homogeneous road network. We assume that the traffic load in Bologna is directly proportional to the number of monitored vehicles present in the considered time interval.
Each point is computed by dividing the total traveled length by the total travel time of the moving vehicles during the considered time interval of 30 minutes. Therefore it gives a congestion measure for the whole road network, and it cannot be related to the average velocity of individual trips.
The fundamental diagram (see figure 3.22) highlights the existence of different traffic regimes in the urban road network according to the traffic load that a different average velocity can distinguish: these results are consistent with the existence of a fundamental diagram for an urban road network [47].

In the fundamental diagram, we observe the points cluster along a straight line with a negative slope representing the expected decrease of the average velocity when the traffic load increases in a road network. The straight regression line suggests a free traffic average velocity $\simeq$ 17.5 km/h for the Bologna road network that reduces down to 10 km/h in case of traffic.
Moreover, the points corresponding to the high traffic regime of the city do not follow the regression line and suggest the existence of a limit velocity for the traffic regime $\simeq$ 10 km/h. Indeed, the velocity distribution computed using only the car trips in the HC during a rush hour (7:30-8:30 of a fixed working day with a limited statistics of 5600 trajectories) has an average value of 2.84 m/sec ($\simeq$ 10 km/h). The distribution is sharply decreasing when the velocity increases, suggesting a strong interaction among vehicles. In contrast, the velocity distribution in the whole metropolitan area (MA) is characterized by an average velocity of 4.53 m/sec with an exponential decaying (see fig. 3.21) when the velocity increases.
This distribution can be interpreted as the result of the coexistence of different type of road network in the same area: the urban road network near the HC whose the frequent crossing points mainly dominate traffic dynamics;



(a)



(b)

Figure 3.22: Picture (a): fundamental diagram average-velocity versus traffic load (number of monitored vehicles present) computed using the car trips recorded in the Bologna historical center area with the OT dataset during September 2016; each point corresponds to a time interval of 30 minutes and the right line is the result of a linear regression. Picture (b): distribution of the average velocity of the car trips recorded with the OT dataset in the whole metropolitan area of Bologna: the circles refer to the car trips in the historical center during rush hours, whereas the squares refer to the car trips in the whole metropolitan area of Bologna.

the country road network with larger streets that connect the periphery and the center and that tend to be congested during rush hours and the highway (the Bologna ring road) where the stop and go regime is observed.
Each road network seems to be characterized by a different velocity distribution, even if some features are universal. To highlight the features of each mobility type, we report in the table 3.8 the average values of the velocity, path length, and travel time distributions related to the different transportation means on considered areas. We understood that the statistical error was on the last digit.

Transport means	Velocity V_m	Path length L_m	Travel time T_m
pedestrians	1.19 m/sec	1.81 km	31.6 min.
bikes	3.05 m/sec	3.49 km	24.5 min.
cars HC	2.84 m/sec	2.01 km	11.3 min.
cars MA	4.53 m/sec	3.91 km	13.4 min.

Table 3.8: Average values for the different mobility distributions

The average velocity is a characteristic of pedestrian and bike mobility, whereas it depends on the traffic conditions and the road network for cars. In the considered cases (car mobility in the HC and MA of Bologna), the average velocities, 10.2 km/h and 16.3 km/h, respectively, are much smaller than the limit velocity for the urban road network 50 km/h. In particular, the average velocity in the HC during rush hours is also smaller than the bike velocity that is comparable to the average car velocity in the MA.
We remark that the average path length for MA cars is 3.91 km, which is near the average bike path length 3.49 km, whereas the average path length for HC cars 2.01 km is near the pedestrian path length 1.81 km. The citizen sample in the Bella Mossa dataset probably has a bias toward people with a propensity for pedestrian and bike mobility. Indeed, in the dataset are present long pedestrian and bike paths that are probably due to trained individual and contribute to increasing the average path length values.
Moreover, if one computes the expected travel times from the path length and the velocity, the corresponding values ($L_m/V_m = 25.3$ minutes for pedestrians and $L_m/V_m = 19.0$ minutes for bikes) are shorter than the average travel time T_m computed using the single trips that could be understood since some people may have forgotten to switch off the Bella Mossa App at the end of a trip, so we use these values as T_m in the sequel. However, the data suggest that cars are considered personal mobility, even for short trips where pedestrian mobility is inconvenient or too tiring. In contrast, the bike is an

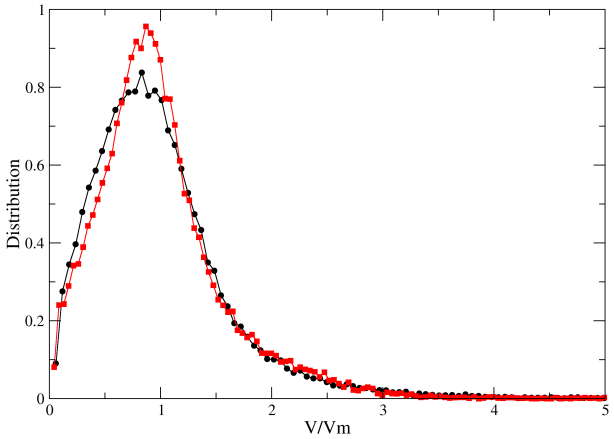
alternative transport mean to the car since it realizes similar mobility. Nevertheless, bikes' disutility probably has to be related to the energy required by riding a bicycle or the road network's discomfort. The use of the car in the HC is perceived as convenient also for short trips for the possibility of performing complex mobility with different activities for which public transportation would require much more time.
To understand the universal features of the mobility performed using the different transport means, we study the behavior of the distribution functions of average velocity, path lengths, and travel time.
Suppose one disaggregates the data according to the traffic load in the whole area. In that case, the corresponding velocity distributions are characterized by different average values: $V_m = 4.20$ m/sec during rush hours and $V_m = 5.76$ m/sec in the evening.
However, the distributions for the normalized velocity V/V_m are very similar in both cases (see fig. 3.23), suggesting that the average velocity is directly related to the traffic load. The possible congestion effects increase the distribution value slightly, corresponding to the low normalized velocities. In contrast, in the case of low traffic, the distribution is more peaked around the average value $V/V_m = 1$.
It is also interesting to observe the exponential decaying for large values of the normalized velocities that is the same in both cases: by definition, the average velocity for a given path length L (not too small) is $V = L/T$ where T is the total travel time that depends on the traffic condition on the road network. The average value T_m for the travel time is related to the traffic load on the road network, and it can be used to measure the congestion degree. Assuming that the decreasing of the travel time by a quantity ΔT with respect to an expected value is modeled by a Poisson random process (this model could also take the heterogeneity of individual behavior), one has the probability distribution

$$P(\Delta T) \propto p^{-\Delta T/T_m} \qquad \Delta T \geq 0 \tag{3.6}$$

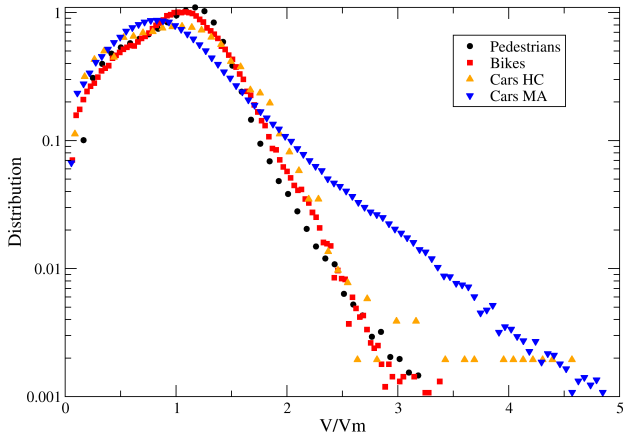
where $p \in [0, 1]$ is a suitable constant that defines the probability of the relative fluctuations. Then the corresponding relative velocity increasing is

$$\frac{\Delta V}{V_m} \simeq \frac{\Delta T}{T_m}$$

where $V_m = L/T_m$, so that if the travel time fluctuations do not depend on L, one gets an exponential decaying in the probability distribution of the normalized velocity V/V_m independently from the traffic condition, which essentially modifies V_m according to figure 3.22 (b).



(a)



(b)

Figure 3.23: Picture (a): distribution of the normalized average velocity V/V_m where V_m is the mean value for the vehicle trips: the circles refer to rush hours 7:30-8:30 and the squares to a low traffic regime 22:30-23:30 in the whole Bologna metropolitan area. Picture (b): comparison of the normalized average velocity distribution for all the considered mobility types. Right picture: distributions of the normalized average velocity for all the considered datasets using a semilog scale: in the inset, we give the correspondence of the different symbols. We remark the collapse of the distributions for the pedestrian, bike, and car HC mobility, whereas the distribution for the car mobility in the Bologna MA shows a different behavior.

However, if we compare the normalized average velocity distributions for the mobility types reported in table 3.8 (see fig. 3.23), we observe a collapse for the pedestrian, bike, and car HC distributions, whereas the car distribution in the MA shows a different decaying for high values V/V_m.
This behavior suggests that the probability p in eq. (3.6) could depend on the features of the mobility network (i.e., the presence of several crossings) as well as on the individual behavior (i.e., the heterogeneity of individuals implies different walking velocities). Indeed, pedestrian and bike mobility share the same road network with car mobility in the HC in most cases. The effect of frequent stops introduces heterogeneity in the car dynamics similar to the individual heterogeneity in walking and riding a bicycle. On the contrary, the country roads network and the highway contribute mainly to car mobility in the Bologna MA and increase the probability p in the MA. Here the road network can be viewed as a multilayer road network with different travel velocities for the different layers on which the individuals perform their mobility using strategies to change the road layer.
Under this point of view, individuals can improve mobility efficiency by reducing the travel time (therefore increasing p) for long trips for which the multilayer structure is relevant (see fig. 3.26 and [cit.]). The path lengths distribution does not enter in the previous arguments but plays a fundamental role in understanding which mobility is realized by the different transport means. Previous results have pointed out as urban mobility is mainly characterized by an exponential decreasing in the path length distribution since it reflects the habit of people to perform local mobility, and it can be justified using a Maximum Entropy Principle and introducing the concept of *mobility energy*.
The path length distributions for different mobilities are characterized by the means values reported in table 3.8, and we show in fig. 3.24 (a) the normalized path length distributions to highlight the universal features and to interpret the differences. All the normalized path lengths are characterized by an exponential decreasing, but it is possible to observe some structures. Indeed, the pedestrian and the bike distributions show an initial growth corresponding to the short paths followed by a fast exponential decrease. However, the pedestrian distribution has a slope change at $L/L_m \simeq 2.5$ that could denote the presence of long paths due to sports activities. In contrast, the bike distribution shows a bimodal structure that could be related to the presence of two types of mobility demand that could distinguish the bike mobility in the historical center from the bike mobility in the periphery. The car mobility provides two distributions where the effect of short paths is not evident, and the exponential decaying is slower. These distributions suggest that private cars could be used to perform complex mobility that realizes many activities

(i.e., not a simple origin-destination mobility) so that a Maximal Entropy Principle can be applied.
Finally, we computed the normalized travel time distributions since mobility time can be considered a universal cost for all the transport means. The results plotted in fig. 3.24 (b) highlight a collapse of the distributions for the bike and car mobility up to values $T/T_m \simeq 2.5$, whereas the pedestrian distribution has a sharp peak at $T/T_m = 0.5$. However, all the distributions have similar features with an initial fast increase for short trips and an exponential decaying, so that we consider the possibility that a single model could explain all the distributions pointing out the relevant parameters that characterize the use of different transportation means in the MA of Bologna.

A survival model for human mobility

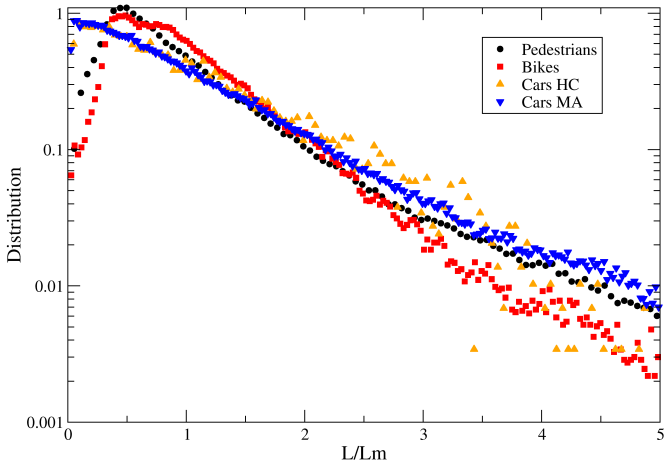
Our point of view is that the statistical properties of human mobility can be described by a simple model using the assumption of the existence of a mobility cost function based on the travel time and that the perceived utility of using a transport means is related to the travel time distribution for short trips.
Then we define $P(T)$ as the probability that a trip has a duration greater than T and we propose a simple survival model to describe the empirical trip duration distributions based on the assumption that a regular Markov process can describe the underlying microscopic dynamics

$$P(T + \Delta T) \simeq (1 - \pi(T)\Delta T)P(T) \tag{3.7}$$

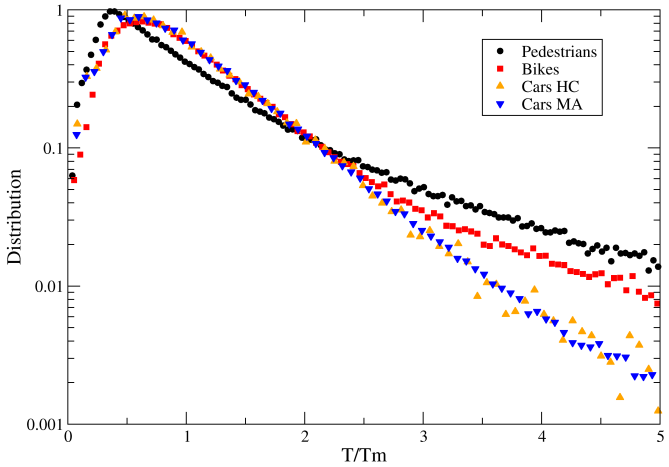
where $\pi(T)$ is the "stop" transition rate: i.e., the probability that a trip stops per unit time after a duration T.
By definition, the travel time distribution $p(T)$ is defined by

$$p(T) = -\frac{dP}{dT} \tag{3.8}$$

If $\lim_{T\to\infty} \pi(T) \to \beta$ we have an exponential decaying for $P(T)$ and we recover the exponential decaying of the empirical distribution for $p(T)$ and β^{-1}. However, to describe the distribution behavior, since the urban mobility is a consequence of the will to perform certain activities, we model the transition rate $\pi(T)$ using the concept of *utility function*. For a given transport means, let $U_c(T)$ and $U_s(T)$ a measure of the perceived utility to continue or to stop a trip of duration T in the framework of a logit model: $U_c(T)$ could be both related to expected advantage get to perform the activity that motivated the trip and to availability of a mobility network for the chosen means, whereas $U_s(T)$ is the disadvantage to continue the trip after a duration T.



(a)



(b)

Figure 3.24: Picture (a): normalized average path length distribution for all the considered datasets: pedestrians (circles), bikes (squares), cars HC (triangles up), and cars MA (triangles down) in semilog scale. Picture (b): normalized average travel time distribution for all the considered datasets using the same convention for the symbols. It is noteworthy to observe as the distributions for bike and car mobility tend to collapse.

Then the simplest assumption is that $U_c(T) = U_c^0$ is constant to measure the initial utility to perform a trip using a specific transport means, whereas $U_s(T) = \alpha T$ is an increasing function of T. Then we define $\pi(T)$ using the logistic function

$$\pi(T) \propto \frac{e^{U_s(T)}}{e^{U_c(T)} + e^{U_s(T)}}$$

so that introducing the fatigue time scale β^{-1} to define the proportionality factor, one gets

$$\pi(T) = \frac{\beta}{1 + \exp(U_c^0 - \alpha T)} = \frac{\beta}{1 + \exp(-\alpha(T - T_c))} \qquad T_c = \frac{U_c^0}{\alpha} \tag{3.9}$$

where the duration T_c can be interpreted as the characteristic duration associated to the convenience of using a certain transport means (see fig. 3.24 for examples). In the continuous limit $\Delta T \to 0$ in eq. (3.7) we get the survival model

$$\frac{dP}{dT} = -\pi(T)P(T) \tag{3.10}$$

where $\pi(T)$ plays the role of the hazard function.
The analytical solution of eq. (3.10) is explicitly written in the form

$$P(T) = e^{-\beta T}\left[\frac{1 + \exp(\alpha T_c)}{1 + \exp(-\alpha(T - T_c))}\right]^{\beta/\alpha} \tag{3.11}$$

so that we recover an exponential decay for $T \gg T_c$ and an inflection point when $\alpha T_c \gg 1$. Under this point of view, the three parameters of the model (3.10) can be associated with physical observables for understanding the statistical of human mobility. Simple algebraic manipulations provide the travel time distribution

$$p(T) \propto \frac{\exp(-\beta T)}{(1 + \exp(-\alpha(T - T_c))^{\beta/\alpha + 1}} \tag{3.12}$$

so that when $T \gg T_c$ we have the exponential decaying with a characteristic time scale β^{-1}, whereas when $T \ll T_c$ if α increases $p(T) \simeq \exp(-\alpha T_c)$ and the time T_c is related to the mode of the distribution T^* by

$$T^* = T_c - \frac{1}{\alpha}\ln\frac{\beta}{\alpha}$$

The characteristic time scale β^{-1} can be interpreted as a *time cost* for the chosen transport means (walking or bike-riding are energy-demanding activities really, but also using the car in traffic conditions is a cause of fatigue).

To perform trips with a duration $\gg \beta^{-1}$ is very improbable.
Similarly, if $\alpha T_c \gg 1$ is also improbable to observe travel times $\ll \alpha^{-1}$ so that this time scale can be interpreted as a *convenience time* scale to use a means of transport. Then we have the requirement $\beta/\alpha \ll 1$ since the convenience time scale should be shorter than the fatigue time scale.
This implies that $T_c \simeq T^*$ defines the *typical travel time* for the chosen transport means. Therefore our aim is not only to consider the interpolation of the empirical travel time distribution for pedestrian, bike, and private car mobility using the model (3.11), but also to relate the parameter values to peculiar features of each transport means in order to understand the role of travel time as universal mobility energy. In particular, the comparison of the empirical parameters for the different transport means allows a better understanding of the relation between the statistical properties of urban mobility and possible decision mechanisms at an individual level to build microscopic models for multimodal mobility.

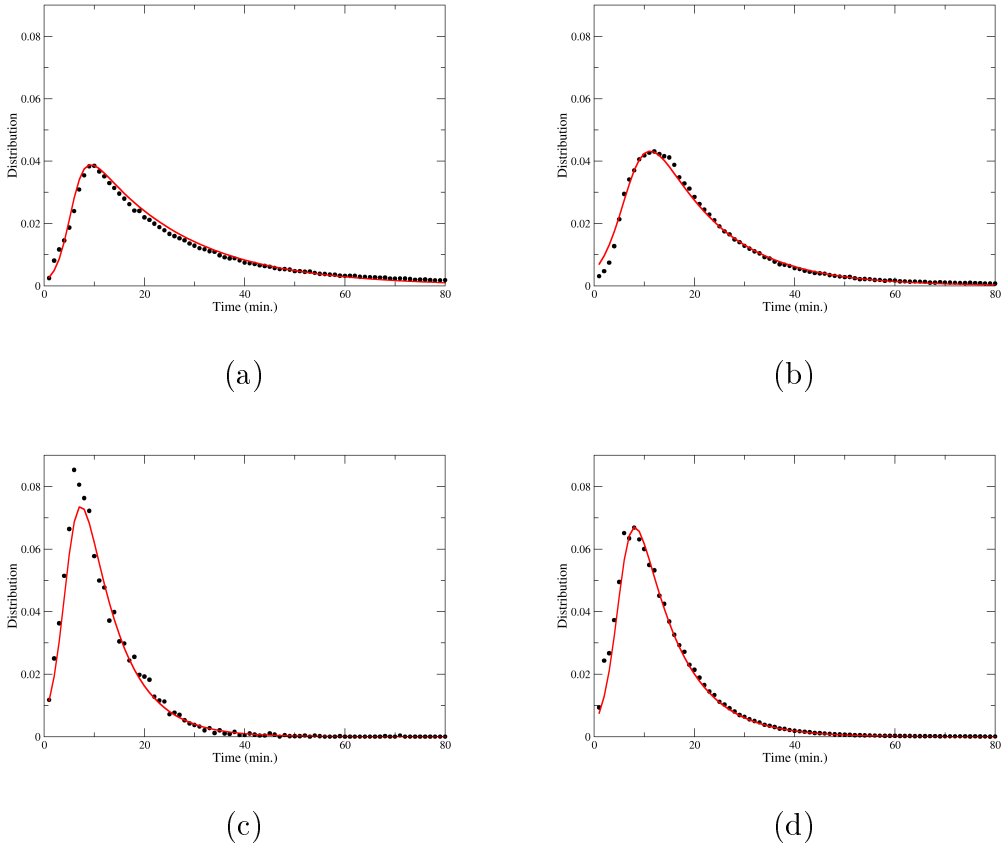
Results of the data analysis

The interpolation procedure based on the model (3.10) is performed first estimating the parameter β from the exponential decaying and then computing the parameters of the hazard function by fixing α and T_c. Our aim is not to provide an optimal interpolation but to show as the survival model (3.10) can explain the main statistical features of the travel time distributions for different transport means.
In the table 3.9 we report the parameter values used for interpolating the empirical distributions and the corresponding figures are shown by fig. 3.25.

transport means	time cost β^{-1}	convenience time α^{-1}	typical time T_c
pedestrians	18.9 min.	1.5 min.	5.5 min.
bikes	13.3 min.	2.6 min.	7. min.
cars HC	7.1 min.	1.7 min.	5.0 min.
cars MA	8.3 min.	1.7 min.	5.5 min.

Table 3.9: Interpolation parameters for the different travel time distributions

We remark that in all the considered case the survival model provides a pretty good interpolation of the empirical distributions: the main difference in the parameter values are in the time cost which varies from 7.1 min. for cars in HC up to 18.9 min. for pedestrians, whereas the other parameters



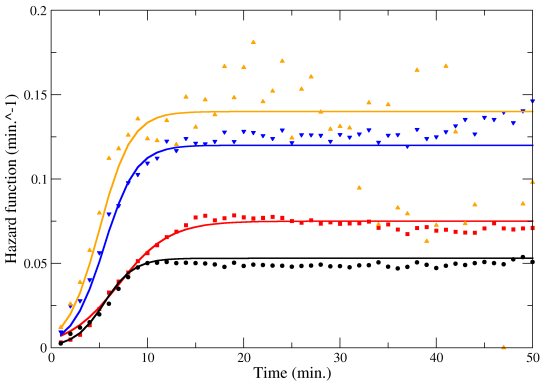
(a) (b)

(c) (d)

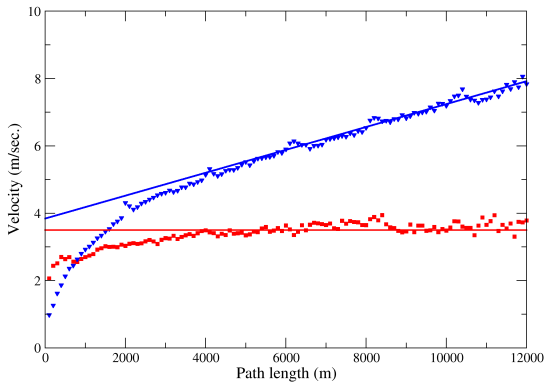
Figure 3.25: From (a) to (d), we show the empirical travel time distributions for the pedestrian, bike, car HC, and car MA mobility with the proposed interpolation using the survival model (3.10) with the parameter values reported in table 3.9.

have similar values.
More precisely, the time cost for car mobility is small, $7-8$ minutes in both cases, so the private means are perceived as convenient since they have a smaller cost with respect to bike and pedestrian mobility. It is noteworthy that convenience time scale $\alpha^{-1} = 1.7$ min. is very short for the cars (in both cases), and the distribution mode is $T^* = 9.5$ min. for cars HC and $T^* = 11$ min. for cars MA (slightly greater than T_c in both cases) suggesting the use of cars for very short trips (this is also in agreement with the path length distribution in fig. 3.24, where we do not obverse a clear drop-down of the distribution at short path lengths for cars).
This can be the consequence of complex mobility performed by people for which the car allows to fulfill many daily activities which individually would require short trips.
We finally remark the similarity of the parameter values in the case of cars HC and cars MA even if the average velocities are very different in the two cases (see table 3.8). This could be an indication of the universality of the travel time cost for urban mobility.
We have also compared the empirical hazard function computed as defined as the ratio $p(T)/P(T)$ between the distribution function $p(T)$ and the stopping probability $P(T)$, with the theoretical hazard function $\pi(T)$ defined by eq. (3.9). We remark as the hazard function (3.9) describes the behavior of empirical curves correctly for short travel times in all the considered cases.
The almost exponential distribution of the car path lengths (see fig. 3.24) can be explained since the convenience time scale is very short and the average velocity is much smaller for short trips than for long trips, as shown in fig. 3.23. This apparent acceleration for long trips is the consequence of the non-homogeneity of the underlying road network that allows individuals to change their mobility strategy according to the trip length.
We remark that this effect is not present for the bike mobility since the average velocity saturates at $V_m = 3.5$ m/sec. (see 3.26) and the road network is homogeneous for the bikes point of view.

The travel time distribution for bikes has a longer convenience time scale and fatigue time scale so that the short trips are depressed in the travel time distribution. This fact could be interpreted since the use of a bike is chosen when trips are not too short (otherwise, one chooses to walk), but then people ride the bicycle for a relatively long time (the distribution mode is $\simeq 17$ min.). The rescaling of the travel times by the means value T_m (see fig. 3.24) highlights as the distributions for pedestrians and private cars tend to collapse, suggesting that bikes and cars are used to perform mobility with different time costs.
Finally, walking mobility is characterized by a short convenience time scale



(a)



(b)

Figure 3.26: Picture (a): comparison between the empirical hazard functions $p(T)/P(T)$ (see eq. (3.10)) for the different transport means (the circles correspond to pedestrian mobility, the squares to bikes, the triangle up to the cars HC and the triangle down to cars MA) with the theoretical hazard function (3.9) using the parameters reported in table 3.9. The fluctuations observed in the car HC points are due to the poor statistics when the travel time is long. Picture (b): average velocity as a function of the trip lengths: the squares refer to the bike mobility whereas the triangle down to the car MA mobility. The straight lines are the result of the regression and show an asymptotic value of 3.5 m/sec. for the bikes and a positive slope 3.4×10^{-4} sec.$^{-1}$ for the cars. The little bumps of any 2 km are due to the spatial discretization of the long trajectories in the dataset and the limited dimension of the considered area so that some trajectories are truncated.

but a long time cost (once one has decided to walk, the time required to perform the mobility cannot be too little). The mode of the distribution is at 9.3 min. less than the average value (see table 3.8) since $T_c = 5.5$ min. is small. The peculiarity of the pedestrian travel time distribution is the long time cost with short convenience time and typical time, whose values are similar to that of the car mobility, denoting that people could choose to walk to perform with a wide range of travel time also for health implications of this activity.
As a matter of fact, the distribution mode values are very similar for all the distributions. This could indicate the existence of a travel time budget related to people's mobility in a city regardless of the specific transport means.

Conclusion

Multimodality mobility is a crucial issue for the realization of future smart cities with low environmental impact. The main goal of this research is to study the existence of universal statistical properties of multimodal mobility taking advantage of the availability of GPS data for single mobility paths and using different transport means thanks to the information and communications technologies. This study was possible in the metropolitan area of Bologna, where the app Bella Mossa collected data for six months during 2017 on bike and pedestrian mobility and where GPS data on private car trajectories were collected for insurance reasons (Octo Telematics dataset). Both the datasets may be affected by some bias since we do not control the statistical sample. However, they are sufficiently big ($\simeq 10^4$ citizens downloaded the Bella Mossa app and the OT dataset gives information on $\simeq 5\%$ of the private car population in Bologna). We have analyzed the distributions for the average velocities, path lengths, and times for the pedestrian, bike, and cars mobility (in the last case, we distinguish between an HC traffic, where many restrictions to the circulation of private cars are applied, and the MA where different types of the road network are present).
The average velocity distributions allowed to confirm the consistency of the BM dataset where the users gave the classification between pedestrian and bike mobility. Furthermore, it pointed out as the congestion effects for traffic that are present in the dataset (see fig. 3.22) affect the average value but not the shape of the velocity distribution. This result suggests that the average value for the velocity distribution is the only observable related to the traffic load at a macroscopic level and that the shape of distribution should be related to the structure of the underlying road network. This is also confirmed by comparing the distributions for the normalized velocities that highlight as the mobility in the MA of Bologna gives a different distribution

with respect to the pedestrian, bike, and car HC mobility that seem to have a similar shape. A possible explanation is that the average velocity distributions reflect the heterogeneity of individual behaviors and the small-scale structure of the road network in the HC. In contrast, in the MA, the distribution is affected by the multilayer structure of the road network that allows the realization of more complex mobility strategies. This is also confirmed by the dependence of the average velocity with respect to the path length in the MA (see fig. 3.26), and it is in accordance with previous results [9]. The path length distribution is dominated by an exponential decaying for all the transport means that is consistent with a Maximal Entropy Principle for urban mobility and the existence of finite mobility energy that defined the average path length [6, 41]. In particular, the car mobility has very short path lengths (in the car HC, we estimate $L_m \simeq 2km$), suggesting that the private cars are used to perform complex mobility that cannot be reduced to origin-destination mobility, and that could be too much time consuming if one would use public means.
Therefore we focus on the travel time distributions whose shape shows universal features for the different transport means (see fig. 3.24) that could reflect both the existence of mobility energy (i.e., the travel time budget concept [43]) and the individual behavior in the choice of the transport means. We addressed the problem of studying universal statistical properties of multimodal mobility by proposing a dynamical survival model based on three observables associated with three-time scales of the model 3.10: the time cost β^{-1}, the convenience time α^{-1} and the typical time T_c. If the time cost is directly related to mobility energy, the convenience time and the typical time could be the effect of the individual decision process underlying the use of one transport means. Our results show that the survival model can reproduce the statistical behavior of the travel time distributions sufficiently and give an explanation for the collapse of the distributions for the bike mobility and the car mobility (both in the HC and MA) when we normalized the travel time by the average value T_m. Indeed the bike and car trips realize the same mobility demand in Bologna, with different time scales but a very similar spatial scale.
The model shows that private cars are used for very short trips so that they are convenient when many activities have to be performed. In contrast, the bike has higher time and convenience costs and probably reflects origin-destination mobility. Pedestrian mobility is an alternative mobility with a significant time cost and essential health consequences for the quality of life in the city. The energy required by the pedestrian mobility implies that the normalized travel time distribution is more peaked near the mode value, that is $\simeq 1/2$ of the time cost, whereas in the other cases, the mode values are

near the time cost.
The proposed model could be helpful from different points of view. The measure of the characteristic time scales for the different transport means in a city could help plan the future smart cities [1] quantifying some effects to be considered in the realization of sustainable mobility and suggesting the best practices. The microscopic urban mobility models could integrate the survival model to introduce a decision mechanism to mimic the citizen's choice of transport means. Finally, specific governance policies (i.e., encouraging the use of electric bikes) could be developed to change the parameter values of the survival model as the bike time cost or the convenience time scale.

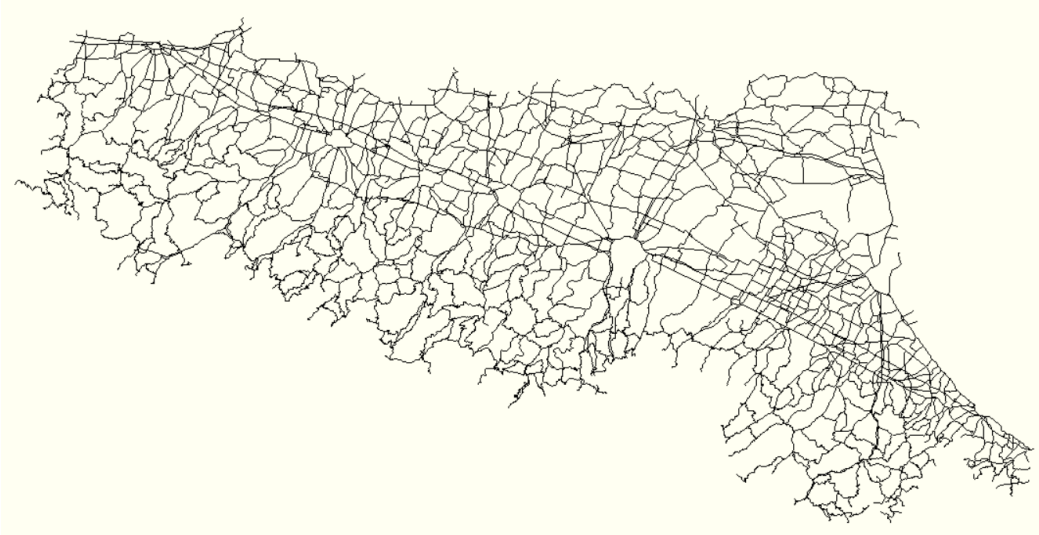


Figure 3.27: Cartography of Emilia-Romagna Region

3.4 Simulation of car mobility: Emilia Romagna

During the last two years of the Ph.D. period, we started a collaboration with the Emilia Romagna Region to develop a monitoring system for car mobility on a regional scale. In this case, we are interested in simulating fluxes among the cities. We used shared cartography with regional authorities, in which the road network of the city center is neglected. In contrast, we kept the highways, country provincial, and ring roads. The result is the cartography shown in figure 3.27.

Simulation inputs

In order to simulate the Emilia-Romagna car mobility, we used the software described in section 2.3, using the Olivetti dataset (1.2.1) to modulate the temporal demand of mobility between the cities and the MTS dataset (1.2.2) for checking and optimization of model.
Analyzing the OD data, it is possible to obtain the average occurrences curve for each Origin-Destination during a day on an hourly frequency. In order to keep the global mobility behavior and, at the same time, reduce the complexity of the model, we choose to consider the first N municipalities more *"actives"* on average. We are selecting N=50 municipalities more or less equally distributed among the 8 Emilia-Romagna provinces. The number of municipalities is a parameter that could be changed and tuned in the validation phase.

We used the 100% of the Olivetti information since in each simulation, we obtain the temporal curves for the creation of 4 kinds of agents:

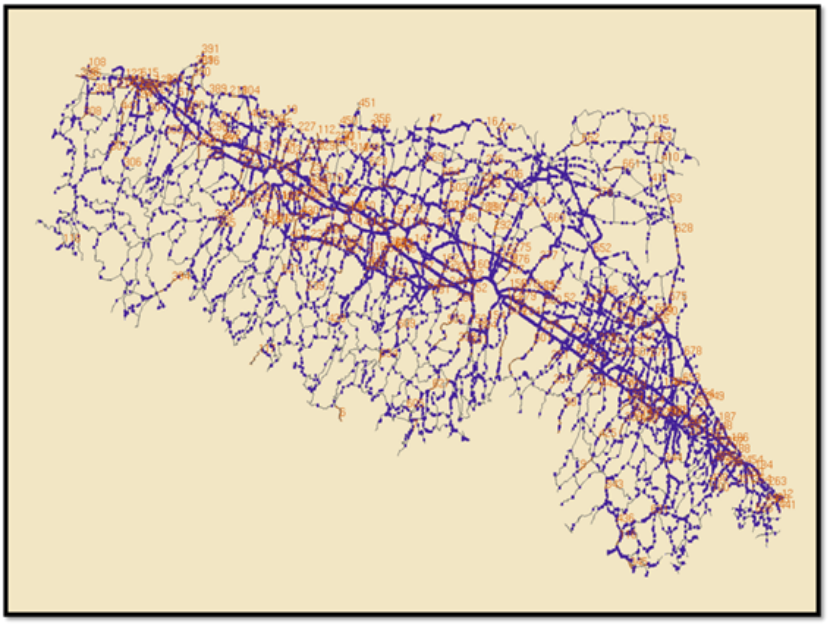
1. Between the most N active municipalities (≈ 49% of the global mobility)
2. From the most N active municipalities to random destination(≈ 17% of the global mobility)
3. From random origin to the most N active municipalities (≈ 9% of the global mobility)
4. Random Noise (≈ 25% of the global mobility): we obtained the temporal behavior from the aggregation of the counting for those municipalities that fall outside the most N active municipalities. Since we extracted the origin and destination randomly, taking into account the ISTAT population of the area, the more the population, the higher the probability.

After choosing the date and duration of the simulation, it is possible to obtain a simulation visible in our interface (figure 3.28 (a) is a screenshot of the dynamical simulation), in which each blue point is an agent that is moving on the network with the behavior already discussed, intending to reach the destination node assigned to him.
In the figure 3.28 (b), we can observe a screenshot of a dynamical heatmap of the same simulation: there is greater use of the road network in the principal axes connecting Rimini to Piacenza at the expense of the roads lateral arteries. This spatial distribution is precisely the behavior that we expected from regional car mobility.

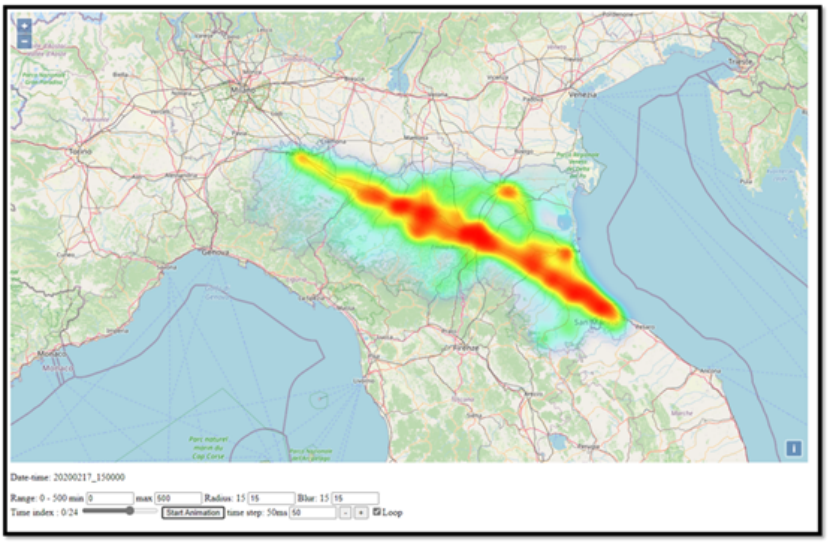
The toolchain

To simulate the mobility behavior as realistically as possible, we followed the toolchain shown below:

1. Having available the OD data for a short period of time (from *01-01-20* to *31-03-20*) we extracted the average OD matrix. This matrix is an hourly timetable representing the average behavior, for each day of the week, between couples of municipalities and couples of provinces (in particular from provinces of neighboring regions to Emilia-Romagna provinces to consider and simulate also the edge mobility).
   In this way, we can have an average behavior with a generation of agents for any simulation required, as shown above.



(a)



(b)

Figure 3.28: (a) Screenshot of an example of simulation on the entire regional road network. Each blue point represents an agent, the orange number and associated poly represent the position of a coil. (b) Heatmap of the same simulation. The red areas are those busiest; the light blue ones are those less busy.

Figure 3.29: Spatial distribution of coils on the region: red ones are the selected validation coils, blue ones are the test coils.

2. In order to improve our simulation, we select some coils that will be the validation barriers; the other will be used just for the test. In figure 3.29 it is possible to observe the spatial distribution of coils: in blue are represented all the test coils, in red the validation ones.

   We made this selection considering the activity of coils according to an equal distribution among the provinces.
   Then, we run the simulation using average OD behavior as input, and we obtain the simulated curve in correspondence of the poly associated to validation coils as shown in figure 3.30.
   Before any fitting, the trends of curves, although the simulation underestimates the observations, are very consistent, and the only usage of OD mobility demand as input reproduces the typical circadian rhythm of daily mobility. In order to measure the distance between simulation and empirical curves, we measured the ratio between the two areas.

   At this point, we can:

   - Improve the shape of the curve changing the global parameter of simulation like how much the users can make mistakes, how important is the length of the path or the duration of this in the best path selection, the value of the critical density and speed, or even the distribution and the average of initial agent's speed. For example, in the case of 3.30 the simulated pick cannot be high as

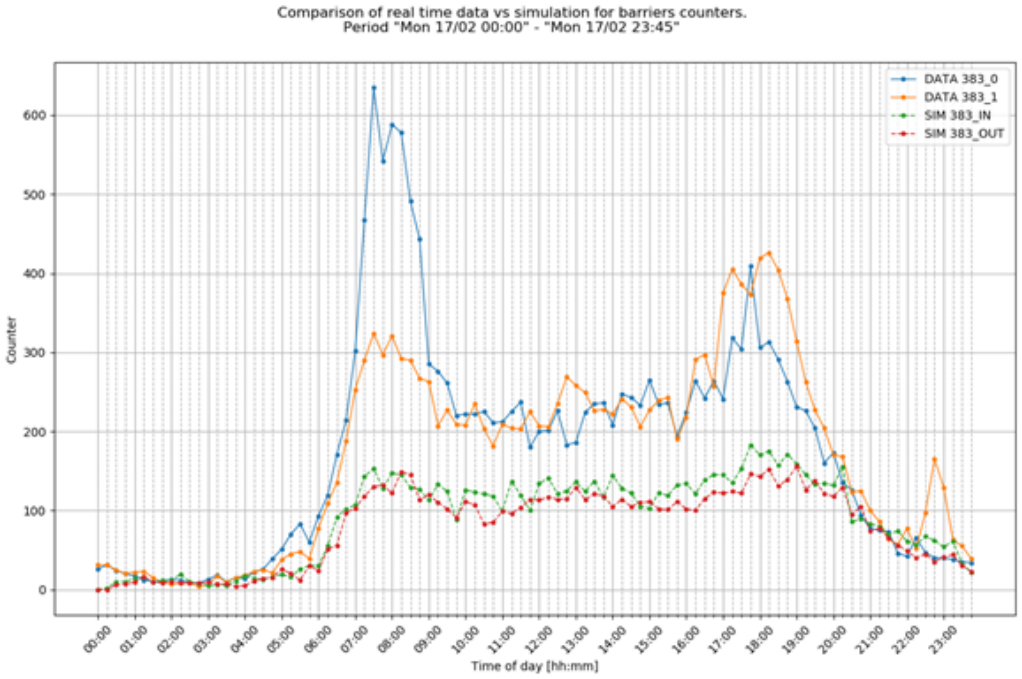


Figure 3.30: Comparison of simulated and real people crossing the coil 383. The crossings are disaggregated by direction; in particular, 0 corresponds with IN and 1 with OUT. Orange and blue are observed data; red and green are simulated ones.

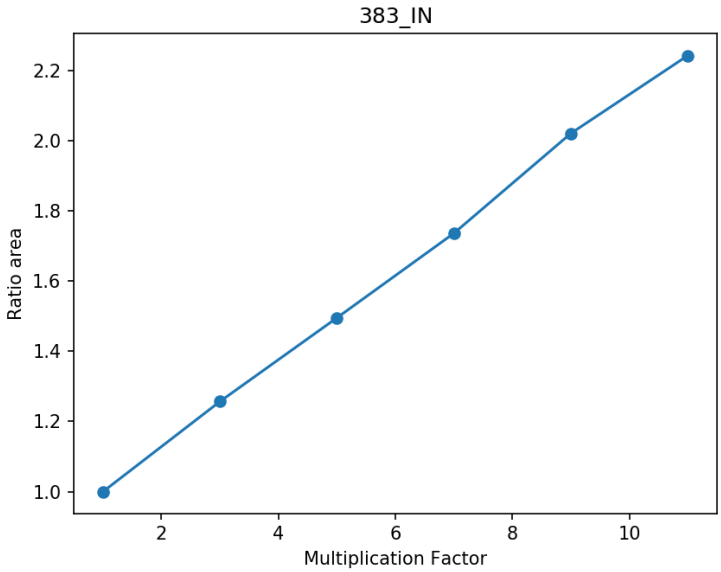
that observed. We ran a scan of parameters and discovered that it is possible to modulate the high of the pick of occurrences in general with a different value of critical density. In particular, using the average difference between real and synthetic curves recorded on the coils as a cost function to minimize, we found a critical density of 0.714.

- Improve the high of these curves by increasing or decreasing the mobility demand for OD municipalities that most influence a particular coil. Then we first performed a pre-analysis in which we identified the couple whose increasing mobility implies the increase in the simulated coils. For each OD couple we generated a simulation with the vehicle flow increased by a factor of 10 exclusively for the couple considered. We observed that the variation of the recorded flow mainly affected single coils. Then we verify if the correlation between the input flow augmentation and those recorded on synthetic coil is linear. We report this relation for a single coil, for example, in figure 3.31. We observe that, except for the small fluctuations due to the random noise, the ratio of areas between the augmented state and the initial state linearly increases with the multiplication factor. The last is the value used to multiply the entire timetable for the specific couple source-destination that influences the coil 383.

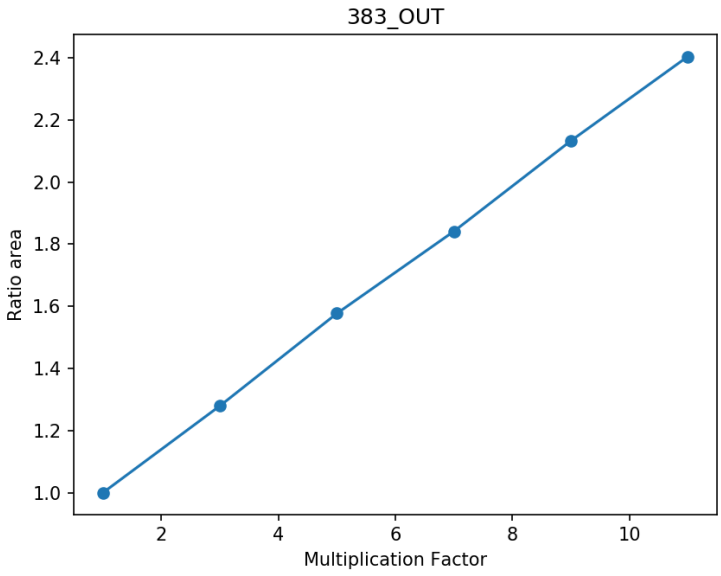
  In the end, it is possible to interpolate the curve for both directions of each coil and store this information. This step is done just one time, and it is included in the preprocessing phase, followed by the request of any simulation. The results of the augmentation step are well summarized in the figure 3.32, in which each point represents the synthetic versus the real area measured in a day by a coil of validation. For the same date, we can observe the top figure about the initial situation (without any correction) and the bottom one after the augmentation. The majority of coils reveal an excellent match between simulated and observed data.

3. Then, we set the best parameters at this point, and we know the coefficients useful to modulate the high of the curve.
   When we request a specific simulation (with a fixed date and duration), the software acts in 2 steps:

   (a) **Generation of the configuration file** in which are present all the needed information about sources, attractions, global and user's parameters. So we know the average behavior in corre-

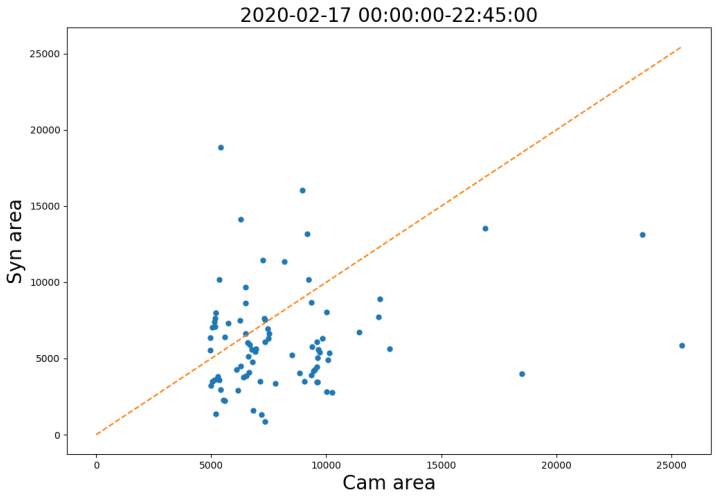


(a)

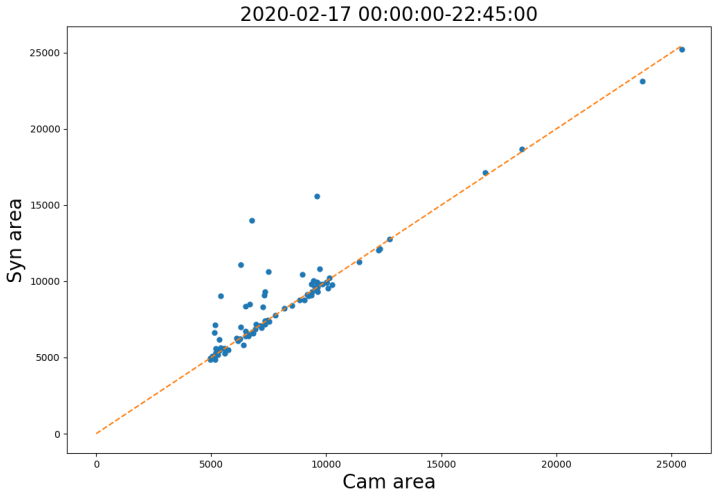


(b)

Figure 3.31: The ratio of areas between the augmented state and the initial state as a function of a multiplication factor with which the entire agent timetable was multiplied. One can observe a linear relation. IN and OUT tags refer to the two directions of travel.



(a)



(b)

Figure 3.32: Synthetic versus real area measured in the 17-02-2020 by each validation coil. (a) The simulation before any process of correction. We can note a spread distribution of points; therefore, there exist coils that overestimate and coils that underestimate. (b) The same simulation with the pre-process of augmentation: most coils are distributed along the line with slope 1, revealing an excellent match between simulated and observed data.

spondence of the validation barriers (that means the number of crossings versus time) and the same behavior recorded by the MTS system for the specific simulation date. Then, using the interpolation parameters previously fitted, it is possible to immediately calculate the multiplication factor for each sensitive OD pair and, therefore, multiply that source by the obtained value.
During this phase, the opening hours of attractions for the specific simulation time are checked, and the tourist agendas were extracted considering the total weights of available attractions and the required length distribution of paths.

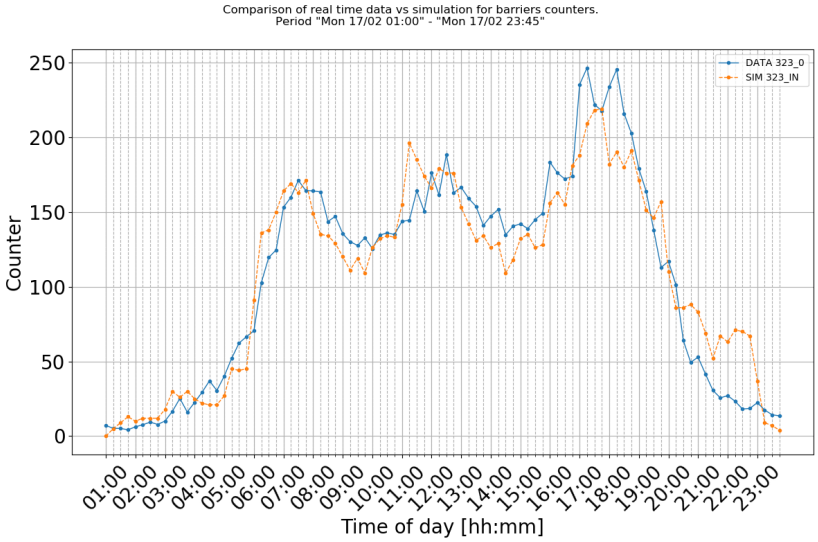
(b) **Run of simulation** with the configuration file produced in the previous step. During the simulation, all the dynamics introduced in 2.3 act. It is possible to dump some information like the timed crossings of each poly or barrier or some global statistical information about the agent's dynamic. In this way, we can obtain some comparisons like those we saw above. We report in figure 3.33 the evolution of crossings over time across two different streets in a specific direction. We can observe like the real data (the solid blue line) and simulated data (the dotted orange one) have a good match. This match happens for the majority of validation coils, as we can see in 3.32 in a single glance.

Conclusion and next steps

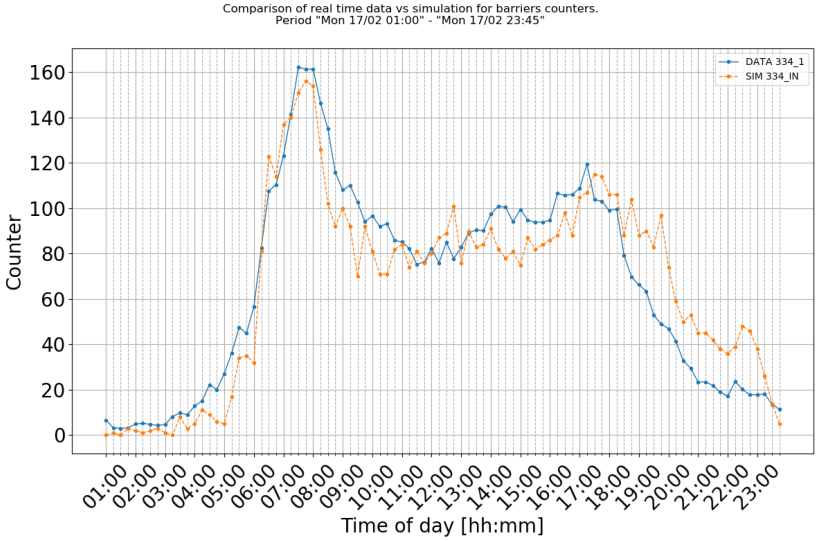
The state-of-the-art allows us to obtain a simulation for any given date and any time interval. If MTS data are available for that date, the model behavior fits the stored data; otherwise, it fits the average MTS data for that specific day of the week. Once we know the number of the crossings during the simulation, it is possible to show the entire network highlighting the discrepancies in the use of the street. In figure 3.34 we can observe these results for a simulation of 17-02-2020 from 00:00 to 23:59 of the same date.

The color code depends on the range of quantile in which the counter of crossing is. Once again, we can note that the main axes connecting Rimini to Piacenza and the access roads to the main towns are widely used at the expanse of the road lateral arteries.

As we observed in the figure 3.32, although the majority of validation coils match the simulated data, there exists some of these that continue to be distant from the bisector line (indicating a perfect match between simulated and real areas). In addition, we have a signal collected by other coils, called



(a)



(b)

Figure 3.33: (a) Comparison of real and simulated evolution of crossing through the coil 323 during the 17-02-2020. (b) Comparison of real and simulated evolution of crossing through the coil 334 during the 17-02-2020. The solid blue line represents the real data; instead, the dotted orange line is the simulated one.

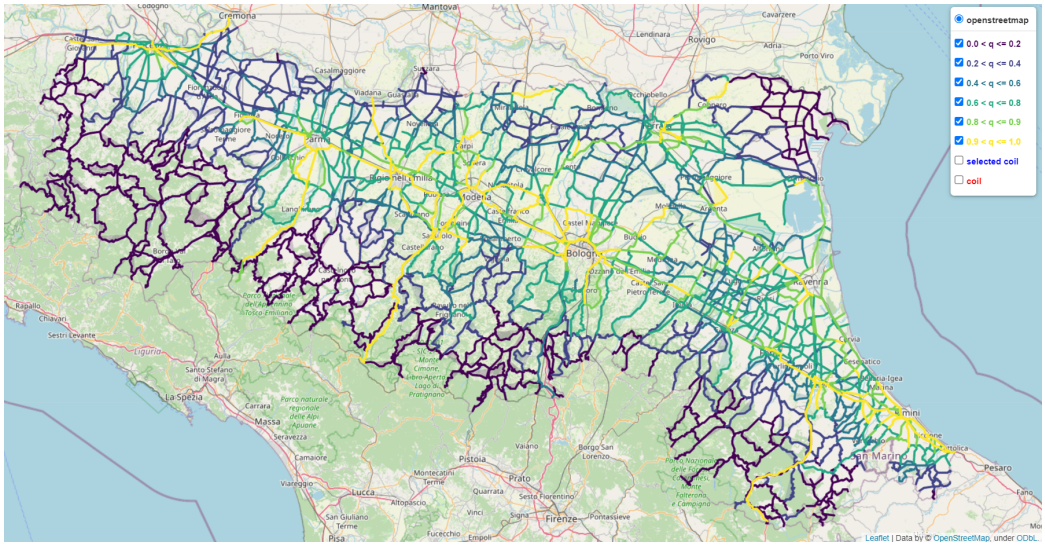


Figure 3.34: Map of Emilia Romagna with network color-coded according to the number of the crossings during a simulation (00:00 - 23:59 of the 17-02-2020). The color code depends on the range of quantile, as it is shown in the legend.

coils of the test. Observing the comparison in these coils, we will introduce a post-process algorithm performing a local correction with a consequent error propagation along the neighborhood streets. We can propagate the error by increasing and reducing the number of agents in the connected street according to their weight. The latter could be determined like the ratio between the poly crossings and the sum of crossings along all the possible adjacent streets.

The interaction and comparison of the simulated results with other kinds of data can be beneficial to analyze and correct the distribution of fluxes on the streets not monitored with a coil. As already mentioned, data coming from alarm tracking systems can add information about the ranking of streets' or path's importance and then change the path chosen by the agents. In addition, the merge between the available Regional network and the OSM one can be complicated, especially in terms of directions of traveling and forbidden turns: the interaction with GPS tracks analysis could correct this lack of information and give more realism to the simulations.

Finally, to make the interaction easier with the regional bodies concerned, infrastructure has been developed. These exhibit the model through WebAPI

with the possibility of real-time data usage coming from a database for now-casting and/or forecasting.

Appendices

Chapter A

Additional information on Section 3.1

The map in fig. A.1 highlights the main areas of interest in the Venice historical center(red circles). These areas are correlated with the mobility demand of tourist flows, and they have been considered to study pedestrian mobility in Venice.
We also recall that the mobility paths are strongly conditioned by the location of the main bridges (*Ponte di Rialto, Ponte dell'Accademia, Ponte della Costituzione*) that allow crossing the channels. The *Ponte del Redentore* is

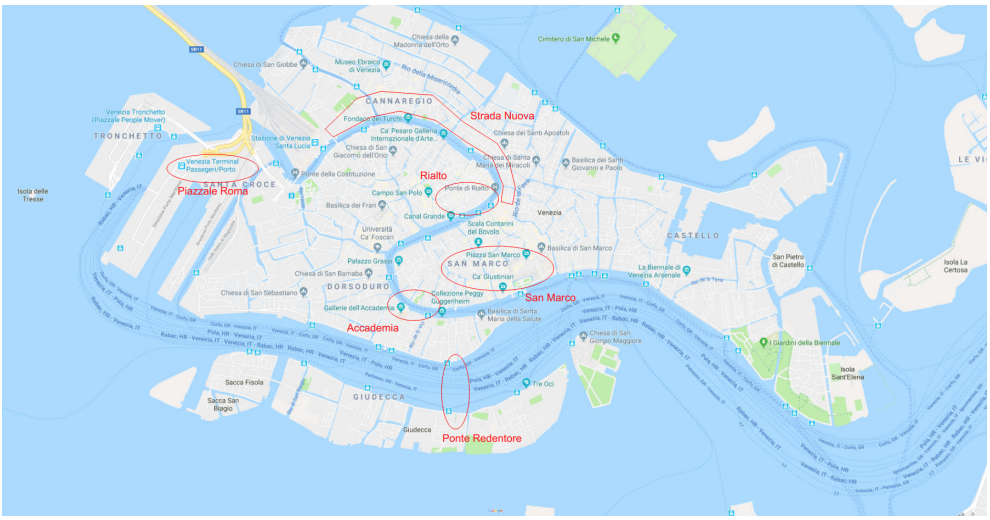


Figure A.1: Map of the historical center of Venice: the red circles enclose the main areas of interest whose name is highlighted and that have been considered in the mobility analysis discussed. We have also pointed out the location of Redentore bridge, which is a floating bridge connecting the Giudecca island with the opposite river during the *Festa del Redentore*.

a floating bridge that is placed during July 15 (we show its location on the map), and it is crossed by a large number of people interested in the firework show. We have taken advantage of this large pedestrian flow to estimate the monitored devices' sample penetration by directly counting the people moving in the opposite directions. The people counting was performed by volunteers coordinated by CORILA (www.corila.it).
Finally, we remark that the main entry points in the historical center are the train station and the parking area at the end of the highway (*Piazzale Roma*), shown in the top-left part of the map A.1.
The GPS dataset on the mobile devices is proprietary to the Italian mobile phone company TIM. According to the privacy law, the GPS data are recorded anonymously (see the section below for more details). Access to data is possible on request to the TIM company. The initial database refers to $\simeq 5\%$ of the entire population of mobile devices, and the ID of each device changes every 24 hours. Using the GPS datasets for the selected events, we succeeded in extracting the daily mobility paths of $\simeq 3000$ of devices that have been used to perform the statistical analysis discussed in the third paragraph.
The average length $< s >$ of the daily paths as a function of the daily mobility time is computed by a moving average over 100 paths ordered according to the mobility time. Due to the variability of the individual mobility, the distribution of $< s >$ differs substantially from the exponential distribution suggested in fig. 3.6. The empirical distribution of $< s >$ for the two datasets is plotted in the Figure A.2.

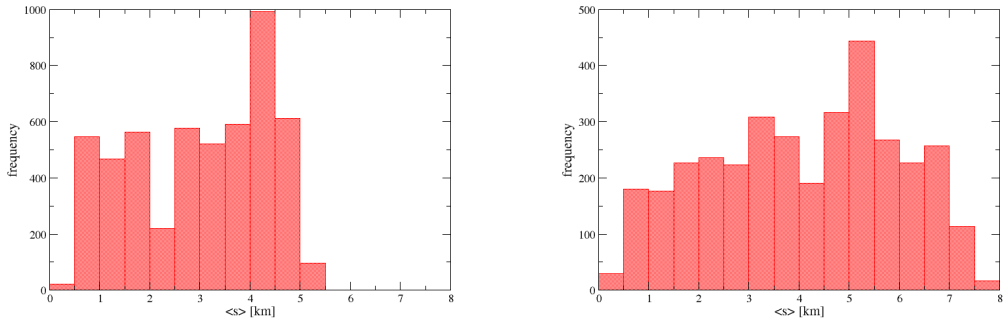


Figure A.2: Distribution of the average mobility $\bar{s}$ for the Carnival dataset (left) and the *Festa del Redentore* dataset (right).

We observe as the approximation with a constant distribution in a finite space interval is consistent with the empirical distribution of $< s >$, justify-

ing the derivation of eq.3.4 to model the rest time effect on mobility. This result could be explained by the heterogeneity of the observed population in which people with very different mobility propensities (e.g., young and older people) are present. Nevertheless, the empirical law $< s > \propto t^{\alpha}$ with $\alpha \simeq 1/2$ emerges as an average behavior of pedestrians moving in Venice.
Once we have established a robust criterion to assign to each road its time-dependent oriented pedestrian flow, we apply an algorithm to select the relevant mobility subnetworks from the whole road network of Venice. We define a relevant subnetwork as a connected subnetwork that explains a considerable fraction of the observed mobility.
We briefly sketch the core algorithm properties, whose properties will be discussed in detail in future works. Starting from the previously evaluated daily flows for each road, we order the roads according to the observed flows in a decreasing way. The algorithm scrolls down the list, adding the road to a temporary list. At every step, a 'pruning process' starts on the selected roads cutting the isolated roads in order to get a connected subnetwork. Therefore, the number of nodes of the subnetwork increases discontinuously: when we add a new road in the list, connecting several previously selected roads.
This behavior is illustrated in the figure A.3. The procedure is carried out

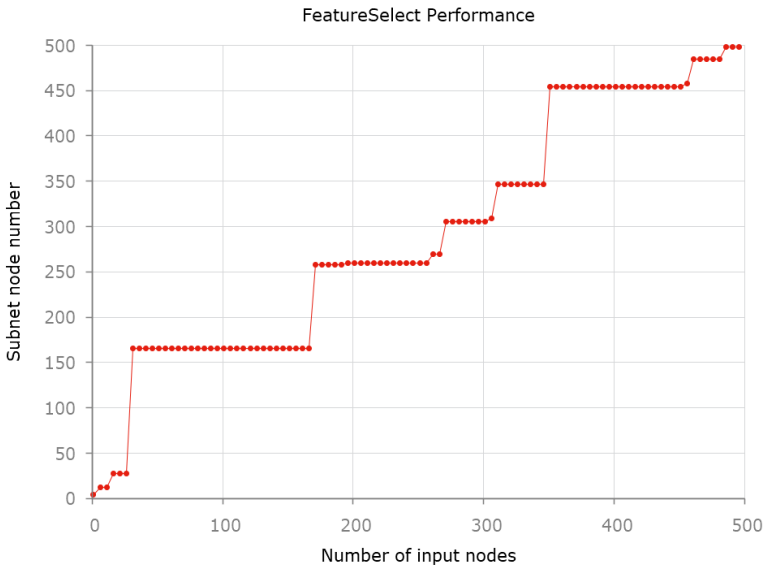


Figure A.3: Number of nodes of the selected subnetwork by our algorithm (vertical axis) as a function of the number of input nodes of the original network (horizontal axis).

iteratively, and the escape condition is obtained once a certain threshold of

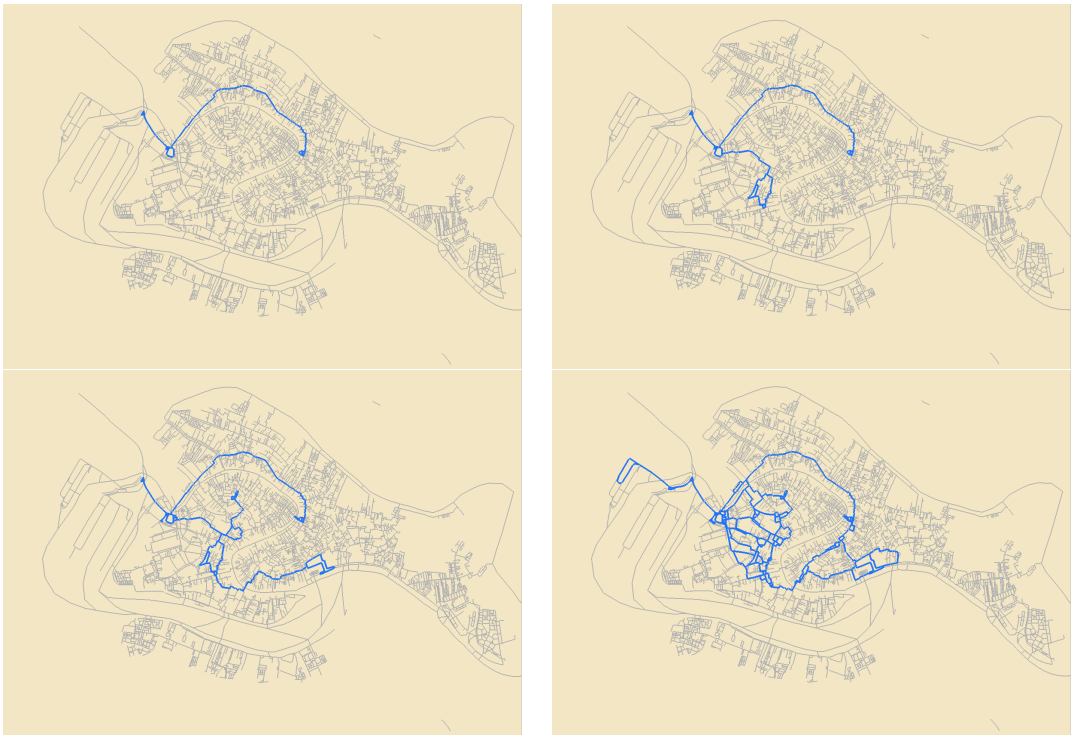


Figure A.4: From top-left to right-bottom, we plot four mobility subnetworks with increasing number of roads, selected by the our algorithm using the Carnival dataset.

the subnetwork node number is reached. After several parametric scans, we found that choosing about 10% of the nodes in the whole Venice road network is the best result for our purposes. In the figures A.4 we show four consecutive selected subnetworks in the case of the Carnival dataset to illustrate how the algorithm operates.

Dataset detailed description

The MDT function ensures that the GPS coordinates are also shown together with the radio measurements typical of the radio access network. The latter may be valid/invalid and accurate depending on the availability of the GPS satellites by the UE.
The MDT data reported from UEs and the RAN may be used to monitor and detect coverage problems in the network. The data collected by the UE are sent during the reporting phase (RRC) and do not affect the pricing to the customer in any way. The data processed in this way are aggregated, in an anonymized form, to provide information on the quality of the network with a geographical indication of the areas with the greatest problems and which

must be optimized. Therefore, the data are purely statistical to achieve the desired improvements, protecting the individual's information.
There are two ways to collect MDT data:

1) Immediate MDT: MDT functionality involving measurements performed by the UE in a CONNECTED state and reporting the measurements to eNB/RNC available at the time of reporting condition and measurements by the network for MDT purposes.

2) Logged MDT: MDT functionality involving measurement logging by UE in IDLE mode, CELL_PCH, URA_PCH states, and CELL_FACH state when second DRX cycle is used (when UE is in UTRA) for reporting to eNB/RNC at a later point in time, and logging of MBSFN measurements by E-UTRA UE in IDLE and CONNECTED modes.
For Immediate MDT, the UE provides detailed location information (e.g., GNSS location information) if available. The UE also provides available neighbor cell measurement information that may be used to determine the UE location (RF fingerprint). ECGI or Cell-Id of the serving cell when the measurement was taken is always assumed known in E-UTRAN or UTRAN, respectively. The reported location information in consists of:

1. Latitude, longitude (mandatory)
2. Altitude (conditional on availability)
3. Velocity (conditional on availability)
4. Uncertainty (conditional on availability)
5. Confidence (conditional on availability)
6. Direction (conditional on availability).

Detailed location information (e.g., GNSS location information) is included if available in the UE when the measurement was taken. If detailed location information is available, the reporting shall consist of latitude and longitude. Depending on availability, altitude, uncertainty, and confidence may also be additionally included.

Abbreviation

DRX Discontinuous Reception
ECGI E-UTRAN Cell Global Identifier
eNB Evolved NodeB
E-UTRA Evolved UTRA
E-UTRAN Evolved UTRAN

FACH Forward Access CHannel
GNSS Global Navigation Satellite System
GSMA GSM (Groupe Spéciale Mobile) Association
IP Internet Protocol
LTE Long Term Evolution
MDT Minimization of Drive-Tests
PCH Paging Channel
RAN radio access network
RNC Radio Network Controller
RRC Radio Resource Control
URA UTRAN Registration Area
UTRA Universal Terrestrial Radio Access
UTRAN Universal Terrestrial Radio Access Network

Acknowledgements

At the end of this Ph.D. cycle, I would like to acknowledge and give my warmest thanks to all those who have accompanied and supported me, during these years, with their advice, teachings, and their affection.
I would first like to thank my supervisor, Prof. Armando Bazzani, for allowing me to know and deepen many research topics through countless discussions. His continuous support during these years has been, for me, a reason for professional and human growth.
A warm thanks to Prof. Sandro Rambaldi for his precious teachings and the stimulating discussions during our meetings.
I would like to thank Telecom and Eng. Davide Micheli for the availability and clarifications on the characteristics of Telecom's MDT data.
I would also like to thank *Servizio Mobilità Regione ER* and in particular Arch. Stefano Grandi for the access to data from the Region's MTS system.
I thank all the people who have alternated in the research group during these four years; each gave me food for scientific and personal reflection.
In particular, I want to thank my colleague Alessandro Fabbri, who took me under his wing and cared for my professional growth. During this journey, he has become a reference point for me and, above all, a friend.
In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me.
Finally, I want to thank all my friends around Italy with whom I shared thoughts, smiles, doubts, and unforgettable evenings. You, too, have been an essential part of this beautiful journey, and I am immensely grateful to you.

Chiara Mizzi

Bibliography

[1] Batty, M; Axhausen, KW; Giannotti, F; Pozdnoukhov, A; Bazzani, A; Wachowicz, M; Ouzounis, G *Smart cities of the future*, The European Physical Journal Special Topics ,**214 (1)**, 481-518, (2012).

[2] Vespignani, A *Modelling dynamical processes in complex socio-technical systems*, Nature Physics **8**, 32-39, (2012).

[3] Brockmann, D: Hufnagel, L; Geisel, T *The scaling laws of human travel*, Nature **439**, 462-465 (2006).

[4] Gonzalez, MC; Hidalgo, CA; Barabasi, AL *Understanding individual human mobility pattern*, Nature **453**, 779-782 (2008).

[5] Song, C; Koren, T; Wang, P; Barabasi, AL *Modelling the scaling properties of human mobility* Nature Physics **6**(10), 818-823, (2010).

[6] Gallotti, R; Bazzani, A; Rambaldi, S *Towards a statistical analysis of human mobility*, Int. J. Mod. Phys. C **23**, 1250061 (2012).

[7] Yan, XY; Han, XP; Wang, BH; Zhou, T *Diversity of individual mobility patterns and emergence of aggregated scaling laws* Scientific reports **3**, 2678, (2013).

[8] Zhao, K; Musolesi, M; Hui, P; Rao, W; Tarkoma, S *Explaining the power-law distribution of human mobility through transportation modality decomposition* Scientific Reports **5**, 9136,. (2015).

[9] Gallotti, R; Bazzani, A; Rambaldi, S; Barthelemy, M; *A stochastic model of randomly accelerated walkers for human mobility* Nature Communications **7**, Article number: 12600 (2016).

[10] Song, C; Qu, Z; Blumm, N; Barabasi, AL *Limits of predictability in human mobility*, Science **327** (5968), 1018-21 (2010).

[11] Gallotti, R; Bazzani,A; Degli Esposti, M; Rambaldi, R; *Entropic measures of individual mobility patterns*, Journal of Statistical Mechanics: Theory and Experiment, **2013**, (2013).

[12] Lin, M; Hsu, WJ; Lee, ZQ *Predictability of individuals' mobility with high-resolution positioning data* In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, New York, 381-390, (2012).

[13] Cuttone, A; Lehmann, S; Gonzalez, MC; *Understanding predictability and exploration in human mobility*,EPJ Data Science **7**,2, (2018).

[14] Batty, M; Desyllas, J; Duxbury, E *Safety in Numbers? Modelling Crowds and Designing Control for the Notting Hill Carnival* Urban Studies, **4** (8), 1573-1590, (2003).

[15] Moussaid, M; Perozo, N; Garnier, S; Helbing, D; Theraulaz, G *The Walking Behaviour of Pedestrian Social Groups and Its Impact on Crowd Dynamics*, PLoS ONE, **5** (4), e10047, (2010).

[16] Omodei, E; Bazzani, A; Rambaldi , S; Michieletto, P; Giorgini, B *The physics of the city: pedestrians dynamics and crowding panic equation in Venezia* Quality & Quantity, **48** (1), 347-373, (2014).

[17] Candia, J; Gonzalez, MC; Wang, P; Schoenharl, T; Madey, G; Barabasi, AL *Uncovering individual and collective human dynamics from mobile phone records* Journal of Physics A: Mathematical and Theoretical *41* (22), 224015, (2008).

[18] Becker, R; Caceres, R; Hanson, K; Isaacman, S; Loh, JM; Martonosi, M; Rowland, J; Urbanek, S; Varshavsky, A; Volinsky. C; *Human mobility characterization from cellular network data* Communications of the ACM **56** (1), 74-82, (2013).

[19] Csáji, BC; Browet, A; Traag, VA; Delvenne JC;, Huens E;, Van Dooren, P; Smoreda, Z; Blondel. VD *Exploring the mobility of mobile phone user*, Physica A: Statistical Mechanics and its Applications 392 (6): 1459-73, (2013).

[20] Xu, Y; Shaw, SL; Zhao, Z; Yin, L; Lu, F; Chen, J; Fang, Z; Li, Q; *Another Tale of Two Cities: Understanding Human Activity Space Using Actively Tracked Cellphone Location Data* Annals of the American Association of Geographers **106** (2), 489-502, (2016).

[21] Ratti, C; Frenchman, D; Pulselli, RM; S. Williams, S *Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis*, Environment and Planning B: Planning and Design 33 (5), 727-748, (2006).

[22] Calabrese, F; Di Lorenzo, GD; Liu, L; Ratti, C *Estimating origin-destination flows using mobile phone location data*, IEEE Pervasive Computing **10** (4), 36-44, (2011).

[23] Xu, Y; Shaw, SL; Zhao, Z;Yin, L; Fang, Z; Li, Q *Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach*, Transportation **42** (4), 625-46, (2015).

[24] Toole, JL; Colak, S; Sturt, B; Alexander, LP; Evsukoff, A; Gonzalez, MC *The path most traveled: Travel demand estimation using big data resources*, Transportation Research Part C: Emerging Technologies **58**, Part B, 162-177, (2015).

[25] Bazzani, A; Giorgini, B; Rambaldi, S; Gallotti, R; Giovannini, L *Statistical laws in urban mobility from microscopic GPS data in the area of Florence* Journal of Statistical Mechanics: Theory and Experiment, **2010**, (2010).

[26] Mokhtariana, PL; Chenb, C; *TTB or not TTB, that is the question: a review and analysis of the empirical literature on travel time (and money) budgets* Transportation Research Part A: Policy and Practice, **38**, 9-10, 643-675, (2004).

[27] Gallotti, R; Bazzani, A; Rambaldi, S *Understanding the variability of daily travel-time expenditures using GPS trajectory data* EPJ Data Science **4**, 18 (2015).

[28] Bonabeau, E; Dorigo, M; Theraulaz, G; *Swarm Intelligence: From Natural to Artificial Systems* Journal of Artificial Societies and Social Simulation, **4**, 320, (1999).

[29] http://www.camminandoavenezia.com/itinerari/

[30] https://www.tim.it/

[31] https://www.venis.it/it

[32] https://www.fabbricadigitale.com/

[33] Barbosa-Filho, H; Barthelemy, M; Ghoshal,G: James, CR; Lenormand, M; Louail, T; Menezes, R; Ramasco, JJ; Simini, F; Tomasini, M *Human Mobility: Models and Applications* https://arxiv.org/abs/1710.00004

[34] Böcker, L; Martin Dijst, M; Prillwitz, J *Impact of everyday weather on individual daily travel behaviours in perspective: a literature review.* Transp Rev 33(1):71 (2013)

[35] https://www.italy-croatia.eu/web/slides

[36] Liu ,X-d; Song, W-g; Lv, W *Empirical data for pedestrian counterflow through bottlenecks in the channel* Transp. Res. Proc. 2 34-42 (2014)

[37] Tran, M; Draeger, C *A data-driven complex network approach for planning sustainable and inclusive urban mobility hubs and services* Environment and Planning B: Urban Analytics and City Science (2021) doi:10.1177/2399808320987093

[38] Barthelemy, M; Flammini, A *Optimal traffic networks* Journal of Statistical Mechanics:Theory and Experiment **2006-07** L07002 (2006)

[39] Ding, R; Ujang, N; Bin Hamid, H et al. *Detecting the urban traffic network structure dynamics through the growth and analysis of multilayer networks* Physica A: Statistical Mechanics and its Applications **503** 800-817 (2018)

[40] Ben-Akiva, M; Bierlaire, M *Discrete Choice Methods and Their Applications to Short Term Travel Decisions* International Series in Operations Research & Management Science **23** (1999)

[41] Kolbl, R; Helbing, D *Energy laws in human travel behaviour* New Journal of Physics **5** 48 (2003)

[42] Liang, X; Zhao, J; Dong, L et al. Unraveling the origin of exponential law in intra-urban human mobility Scientific Report **3** 2983 (2013)

[43] Marchetti, C *Anthropological invariants in travel behavior* Technological Forecasting and Social Change textbf47(1) 75-88 (1994)

[44] https://www.bellamossa.it/

[45] https://www.octotelematics.com/it/home-it/

[46] http://www.corila.it/

[47] Geroliminis, N; Daganzo, F *Existence of urban-scale macroscopic fundamental diagrams: Some experimental* Transportation Research Part B: Methodological **42**, 9 (2008).

[48] Aleta, A; Meloni, S; Moreno, Y *A Multilayer perspective for the analysis of urban transportation systems* Scientific Reports **7**, 44359 (2017).

[49] Gallotti, R; Porter, M. A; Barthelemy, M *Lost in transportation: Information measures and cognitive limits in multilayer navigation* Science Advances **2** (2016)

[50] Kivelä, M et al *Multilayer networks* Journal of Complex Networks textbf2, 203-271 (2014).

[51] Reades, J; Calabrese, F; Ratti, C *Eigenplaces: Analysing Cities Using the Space-Time Structure of the Mobile Phone Network* Environment and Planning B: Planning and Design. **36(5)** 824-836 (2009)

[52] Cheng, Z; Caverlee, J; Lee, K; Sui, D Z *Exploring Millions of Footprints in Location Sharing Services* ICWSM 201181-88. (2011)

[53] Liang, X; Zheng, X; Lv, W; Zhu, T; Xu, K *The scaling of human mobility by taxis is exponential* Physica A **391** 2135-2144 (2012)

[54] Gonzalez, M C; Hidalgo, C A; Barabasi A L *Understanding individual human mobility patterns* Nature **453** 779-782 (2008)

[55] Schläpfer, M; Dong, L;, O'Keeffe, K et al *The universal visitation law of human mobility* Nature **593** 522-527 (2021)

[56] Bettencourt, L M A *The origins of scaling in cities* Science **340** 1438-1441 (2013).

[57] Yazdizadeh, A; Patterson, Z; Farooq, B *An automated approach from GPS traces to complete trip information* International Journal of Transportation Science and Technology **8:1** 82-100 (2019)

[58] Etemad, M; Soares Junior, A; Matwin, S *Predicting Transportation Modes of GPS Trajectories using Feature Engineering and Noise Removal* Canadian AI 2018: Advances in Artificial Intelligence 259-264

[59] Fu, Z; Tian, Z; Xu, Y; Qiao, C *A Two-Step Clustering Approach to Extract Locations from Individual GPS Trajectory Data* International Journal of Geo-Information 5(10):166 (2016)

[60] Hessel, M; Ortalli, F; Borgatelli, F *Machine Learning for Parameter Screening in Computer Simulations* Conference: International Workshop on Modelling and Simulation for Autonomous Systems (2014)

[61] Vespignani, A *Predicting the behavior of techno-social systems.* Science **325** 425-428 (2009).

[62] https://www.olivetti.com/it

[63] Marra, A D; Becker, H; Axhausen, K W.; Corman, F *Multimodal passive tracking of passengers to analyse public transport use* 18th Swiss Transport Research Conference (2018)

[64] Bohte, W; Maat, K *Deriving and validating trip purposes and travel modes for multiday GPS-based travel surveys: A large-scale application in the Netherlands* Transportation Research Part C: Emerging Technologies **17**(3) 285-297. (2009)

[65] Yang, F; Yao, Z; Cheng, Y; Ran, B *Multimode trip information detection using personal trajectory data* Journal of Intelligent Transportation Systems **20:5** 449-460 (2016)

[66] Gabrielli, L; Deutschmann, E; Natale, F; Recchi, E; Vespe, M *Dissecting global air traffic data to discern different types and trends of transnational human mobility* EPJ Data Science **8**(1) (2019)

[67] Gallotti, R; Barthelemy, M *Anatomy and efficiency of urban multimodal mobility* Scientific Reports **4**(1):6911 (2014)

[68] Yin, C; Xiong, Z; Chen, H et al. *A literature survey on smart cities* Sci. China Inf. Sci. (58) 1-18 (2015)

[69] Al Nuaimi, E; Al Neyadi, H; Mohamed, N et al. *Applications of big data to smart cities* J Internet Serv Appl **6** 25 (2015)

[70] Allam, Z; Dhunny, Z A *On big data, artificial intelligence and smart cities* Cities **89** 80-91 (2019)