PhD-FSTM-2023-013
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 10/03/2023 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES DE L'INGÉNIEUR

by

Piergiorgio VITELLO
Born on 9 April 1992 in Rome, (Italy)

CROWDSOURCED DATA FOR MOBILITY ANALYSIS

Dissertation defence committee

Prof. Dr. Francesco Viti, dissertation supervisor
Professor, Université du Luxembourg

Dr. Richard D. Connors, member
Research Scientist, Université du Luxembourg

Prof. Dr. Raphaël Frank, chairman
Professor, Université du Luxembourg

Dr. Jun Pang, vice-chairman
Research Scientist , Université du Luxembourg

Prof. Dr. Jan-Dirk Schmöcker, member
Professor, Kyoto University

Abstract

The importance of data in transportation research has been widely recognized since it plays a crucial role in understanding and analyzing the movement of people, identifying inefficiencies in transportation systems, and developing strategies to improve mobility services. This use of data, known as mobility analysis, involves collecting and analyzing data on transport infrastructure and services, traffic flows, demand, and travel behavior. However, traditional data sources have limitations.

The widespread use of mobile devices, such as smartphones, has enabled the use of Information and Communications Technology (ICT) to improve data sources for mobility analysis. Mobile crowdsensing (MCS) is a paradigm that uses data from smart devices to provide researchers with more detailed and real-time insights into mobility patterns and behaviors. However, this new data also poses challenges, such as the need to fuse it with other types of information to obtain mobility insights. In this thesis, the primary source of data that is being examined and leveraged is the popularity index of local businesses and points of interest from Google Popular Times (GPT) data. This data has significant potential for mobility analysis as it overcomes limitations of traditional mobility data, such as data availability and lack of reflection of demand for secondary activities.

The main objective of this thesis is to investigate how crowdsourced data can contribute to reduce the limitations of traditional mobility datasets. This is achieved by developing new tools and methodologies to utilize crowdsourced data in mobility analysis.

The thesis first examines the potential of GPT as a source to provide information on the attractiveness of secondary activities. A data-driven approach is used to identify features that impact the popularity of local businesses and classify their attractiveness based on these features. Secondly, the thesis evaluates the possible use of GPT as a source to estimate mobility patterns. A tool is created to use the crowdness of a station to estimate transit demand information and map the precise volume and temporal dynamics of entrances and exits at the station level. Thirdly, the thesis investigates the possibility of leveraging the popularity of activities around stations to estimate flows in and out of stations. A method is proposed to profile stations based on the dynamic information of activities in catchment areas. Through

this data, machine learning techniques are used to estimate transit flows at the station level. Finally, this study concludes by exploring the possibility of exploiting crowdsourced data not only for extracting mobility insights under normal conditions but also to extract mobility trends during anomalous events. To this end, we focused on analyzing the recovery of mobility during the first outbreak of COVID-19 for different cities in Europe.

Acknowledgements

I am deeply grateful to all those who have supported me throughout the process of completing this PhD thesis. Their contributions, both to my work and my personal life, have been invaluable. I feel honored to have had the opportunity to work with such inspiring individuals who have helped shape my growth and learning over the past four years.

I would like to give a special thanks to Professor Viti for his guidance and mentorship as my supervisor over the past four years. His constant support, along with the many opportunities he provided, made a significant impact on my journey. I am also grateful for the enthusiasm, that he transmitted me and his vision, especially during the most challenging moments. His insights and guidance have been crucial in shaping the direction of my research.

I would also like to express my gratitude to Dr. Connors for his valuable supervision during my second half of PhD studies. He has provided me with new perspectives and ideas helping me to approach my research from different angles. I extend my appreciation to the members of my CET committee, Dr. Pang and Professor Frank, for their dedicated time and insightful contributions throughout my PhD. Their valuable feedback and guidance have helped me to improve my work and achieve a higher standard of research. A big thank you also goes to Professor Schmoecker for providing our group with the opportunity to work on an exciting and stimulating project that has allowed us to share our ideas and collaborate with many other individuals working on similar studies.

I am also grateful to Dr. Fiandrino for starting as my supervisor during my master's thesis and for helping me throughout my PhD journey. He has been a constant source of support and encouragement, and his guidance has been invaluable. I would like to thank Dr. Capponi for introducing me to Luxembourg and starting my research career. His support and example have been instrumental in my growth as a researcher. I am grateful to the Luxembourg National Research Fund for funding the DRIVEN project(PRIDE17/12252781/DRIVEN) which supported my research.

I would also like to acknowledge my colleagues from Mobilab, both past and present, for making the office a wonderful place to work and for the coffee breaks we shared. Finally, I would like to express my deepest appreciation to Claudia, my family, and my friends for their continuous support throughout these years. Their love has been a constant source of motivation, and without them, this journey would not have been possible.

Contents

List of Figures

List of Tables

Acronyms

CDF cumulative distribution function.

CDR call detail records.

FCD Floating car data.

GPs Gaussian Processes.

GPT Google Popular Times.

ICT Information and Communications Technology.

JSD Jensen-Shannon divergence.

LB local business.

LBSN location-based social networks.

MAE mean absolute error.

MCS Mobile CrowdSensing.

ML Machine Learning.

MLP MultiLayer Perceptron.

MTA Metropolitan Transportation Authority.

NNs neural networks.

NYC New York City.

OSM OpenStreetMaps.

POI Point of interest.

PT Public Transport.

R2 coefficient of determination.

RMSE root mean squared error.

RQ Research Question.

SARS-COV-2 Severe Acute Respiratory Syndrome Coronavirus 2.

SHAP SHapley Additive exPlanation.

SVM Support Vector Machine.

wMAPE weighted Mean Absolute Percentage Error.

WMATA Washington Metropolitan Area Transit Authority.

Part I

Introduction and State of the Art

Chapter 1

Introduction

1.1 Context and motivation

Mobility in Europe is undergoing significant changes [1]. The European Commission predicts that by 2030 automated mobility will be deployed at large scale, and that traffic on high-speed rail will double [2]. These transformations are driven by a variety of factors, including advances in technology, changing societal attitudes, and a growing focus on sustainability [3]. One of the key drivers of change in mobility is the proliferation of new technologies [4], such as electric and autonomous vehicles, as well as the development of new transportation services and platforms, such as ride-sharing and micromobility options. These technologies and services have the potential to revolutionize the way people move around and access transportation, offering more convenient, efficient, and sustainable options [2][3]. The success of digital mobility services is heavily reliant on the availability and utilization of data. Digitalization, the process of converting information and data into digital form and using digital technologies to transform traditional processes and systems, has played a vital role in enabling the development of these new mobility services. Through digitalization, it has become possible to create digital systems and platforms that enhance the movement of people in a more efficient and convenient manner.

Why data is important for mobility?

Data plays a crucial role in the mobility of modern cities. It helps to understand and analyze the movement of people, identify bottlenecks and inefficiencies in the transportation system, and develop strategies to improve individual mobility choices as well as the offered mobility services [5]. Examples of data for mobility are the routes that people are taking, the modes chosen to they reach their destination, the time of day that travel is occurring, and the level of congestion on the roads.

Data enables mobility analysis, which is a fundamental tool in transportation. This process involves the collection and analysis of data pertaining to transport infrastructure and services, traffic flows, demand, and travel behavior. By utilizing this data, mobility analysis aims to enhance the efficiency of transportation systems. Through mobility analysis, transportation authorities can identify patterns and trends in travel behavior, and use this information to improve the efficiency of the transportation system [6]. An example of the use of mobility analysis can be found in public transportation (PT), where data can be used in the short term to optimize routes and schedules, reducing waiting times and increasing the system's reliability [7]. While in the long term, it can be exploited to identify areas of the city with high demand for transportation services, and to plan for the deployment of additional infrastructure or services to meet that demand [8].

More in general, data plays a vital role in the mobility of modern cities, helping to optimize the transportation system and promote sustainable mobility [9][10]. As the amount of data available to cities continues to grow, it is becoming increasingly important for cities to develop the capacity to effectively analyze and utilize this data to improve the mobility of their citizens.

Traditional mobility data

However, the traditional data sources exploited in mobility analysis such as traffic counts and public transit ridership can have several limitations [11]. One limitation is data availability [12]. These data sources may not be available for all regions or may be limited in scope, making it difficult to get a comprehensive understanding of the transportation system state in every location. Additionally, traditional data sources can suffer the problem of granularity, i.e. the data may not be updated frequently, which can make it difficult to get an updated picture. Another limitation of traditional data sources is data accuracy [13]. These data sources may not be completely accurate due to measurement errors or data quality issues [14]. This can lead to incorrect or misleading conclusions. Furthermore, traditional data sources frequently do not adequately reflect the demand for secondary activities. These activities, such as leisure or recreational, may be overlooked in data collection efforts but can still be a significant factor in terms of mode choice. In recent years, advances in technology and the increasing availability of data from new sources have helped to address some of these limitations [15]. The use of these new data sources can help to improve the accuracy and relevance of transportation data and make it more widely available, enabling transportation planners and researchers to make more informed and effective decisions about transportation systems.

Crowdsourced data

Recently, new technologies have been introduced and deployed providing multiple sources of data that can be utilized for mobility analysis [16]. The widespread of mobile devices enables the use of Information and Communications Technology (Information and Communications Technology (ICT)) by unleashing the potential to improve the quality of mobility. The enormous number of smart devices provides a potential source of data according to the mobile crowdsensing (MCS) paradigm [17]. Mobile CrowdSensing (MCS) leverages the collective intelligence and sensing capabilities of a large number of mobile devices to gather, process, and disseminate data about various aspects of the physical world [18]. These devices, also known as sensors or crowd sensors, are typically carried by individuals and can be equipped with a variety of sensors such as cameras, microphones, GPS, accelerometers, and more. The information obtained from MCSs is called Crowsourced/Crowdsensed data. Crowdsourced data has the potential to revolutionize the way we collect, analyze, and use data about the world around us. It allows us to gather information from a wide range of sources and locations, and to do so in a timely and cost-effective manner. Crowdsourced data has been applied to a variety of fields, including environmental monitoring [19], urban planning [20], and disaster response [21].

The main motivation behind this thesis is to explore the power of this data for mobility analysis. However, also crowdsourced data come with its own set of challenges; One of the main limitations of crowdsourced data is that it is not collected specifically for mobility analysis, making it less suitable for such purposes. To address specific research questions for mobility, multiple data sources may need to be combined, which can lead to additional challenges in merging the different sources. Also, crowdsourced data is not made directly available by the providers, but often it is processed and offered in an aggregated way to preserve privacy and anonymity.

Among other crowdsourced data currently being available, in this thesis the main source we aim to study and leverage is the popularity index of local businesses and points of interest via the Google Popular Times(GPT). Since 2015, Google has made available a new feature called GPT, which consists of anonymized crowdsourced data passively collected from Google users. This data provides the temporal profile of the number of people visiting a place, such as a retail store, restaurant, or public venue. This type of data has significant potential for mobility analysis due to its unique advantages over traditional mobility data.

One advantage is its availability. Transportation system data is often difficult to obtain, as it requires new data collection efforts and may not be readily available in certain areas. GPT data, on the other hand, is already provided by Google and is globally accessible, making it possible to obtain mobility analysis in areas where traditional transportation data is not typically

available. Another advantage is the information on places and local business activities, which can provide valuable insights into secondary activities that are difficult to obtain from traditional mobility data.

However, there are some drawbacks to using GPT for mobility analysis. One limitation is the lack of transparency in the data pre-processing. While Google provides the processed data in an aggregated and normalized form, the exact methods and algorithms used for this processing are not disclosed. This lack of transparency can make it difficult to fully understand and trust the results obtained from GPT data. Another limitation is the fact that GPT data does not directly provide information on mobility. To gather mobility information from GPT data, it must be carefully processed and analyzed. Despite these limitations, GPT data can still be a valuable source of information for mobility analysis if properly used and interpreted. This thesis seeks to investigate the potential of GPT data as a source of information for mobility analysis, taking into account its limitations.

1.2 Objectives and Contributions

This dissertation aims to explore how crowdsourced data can contribute to mobility by formulating a list of objectives translated into research questions (RQ). The main RQ of this thesis can be formulated as follows:

Main RQ

How can we leverage novel crowdsourced data for mobility analysis, and overcome the limitations of traditional data sources?

Following this broad Research Question (RQ), this thesis aims to contribute to the knowledge on how crowdsourced data can be leveraged for mobility analysis and to provide insights that may be useful for researchers, policymakers, and other stakeholders interested in mobility. To investigate this potential, we identified several specific sub-questions related to the main RQ. One of the key contributions of this thesis is to improve the understanding the potential of crowdsourced data for addressing the RQs, and which may require additional data sources or approaches.

Each chapter in this thesis focuses on addressing a specific sub-question, generating a corresponding contribution. Fig 1.1 shows all research questions and the chapters where they are addressed together with the corresponding contribution.

As stated in our main research question, the objective of this dissertation is to understand the strengths and limitations of using crowdsourced data for mobility analysis, and to identify specific areas where it can be effectively utilized. To achieve this, we start by investigating the unique insights and

contributions that crowdsourced data brings to the field of mobility. To this end, the first RQ is formulated as follows:

RQ1

What are the key differences between crowdsourced data and traditional transport data in terms of their potential for mobility analysis?

We argue that crowdsourced mobility data is different from traditional mobility data in different ways, and can add value to traditional data sources by providing additional insights and information. Answering this first RQ will be crucial for the rest of the dissertation, as it will provide the foundation for our exploration of the potential of crowdsourced data for mobility research. This RQ aims to identify the key ways in which the potential of crowdsourced data can be useful for mobility research, as well as the main differences between crowdsourced and traditional approaches to collecting and analyzing mobility data.

In order to provide a solution to RQ1, Chapter 2 presents a comparison between crowdsourced and traditional mobility data. This comparison includes a detailed analysis of the different limitations of both data types, as well as an examination of the state of the art of studies that have used these data sources for mobility analysis. The chapter also explores the potential contributions that each data type can offer for understanding and improving mobility in various contexts.

To aid in the evaluation of the different data types, the chapter introduces a scoring scheme that is designed to assess the suitability of each data source for a given research objective or application. This scoring scheme takes into account various aspects such as accessibility, temporal dynamics, and disaggregation. By using this scoring scheme, researchers can more easily identify the strengths and weaknesses of each data source and select the most appropriate for their particular needs. Overall, the contribution of this chapter is to provide a comprehensive comparison of the two types of data and a framework for evaluating their suitability for mobility analysis. The main challenge of this contribution is to determine a standardized approach for evaluating the different aspects of different sources of data.

Among all the different types of crowdsourced data, one stands out as particularly useful for mobility analysis: GPT. This dataset is notable for its availability and granularity, which make it well-suited for analyzing mobility. Specifically, from Chapter 2 emerge that the impact of GPT on mobility analysis is twofold. First, the ability of GPT to provide dynamic information about points of interest allows for the analysis of secondary activities, which are often overlooked in mobility research. In addition, the characteristic of wide availability allows mobility planners and researchers to gain insights into regions where traditional mobility data is not available. These two main

Research Questions	Chapter	Main Challenge	Contribution
RQ1 What are the key differences between crowdsourced data and traditional transport data in terms of their potential for mobility analysis?	**Chapter 2** Background	Develop a standardized approach for evaluating the different sources of data	Comparison between Traditional mobility data and Crowdsourced. Including a detailed analysis of the different limitations of both data types.
RQ2 Can GPT be used to classify local businesses to understand dynamic demand profiles?	**Chapter 3** GPT as factor of local businesses attractiveness	GPT is normalized, which can make it difficult to distinguish between different types of Local Businesses with varying capacities	An analysis where we use GPT to examine LBs with the goal of identifying urban metrics that influence their popularity and using machine learning techniques to classify the category and attractiveness of LBs based on these factors
RQ3 Can GPT be used to estimate mobility patterns such as transit demand information?	**Chapter 4** Transitcrowd, the transit estimation tool	Estimating two flows(Entrances and Exits) using a single value(GPT)	The development of an estimation tool that uses GPT data to estimate the transit flows at the station
RQ4 How can we convert GPT data into transit demand information automatically?	**Chapter 5** GPT of catchment areas to estimate transit flows	Different cities can have different structures, which can make it difficult to compare the areas around stations in different locations	The development of a model that is able to estimate the transit flows at a station without the need for training data from transit
RQ5 Can Crowdsourced data be used to analyze mobility during anomalous events?	**Chapter 6** Mobility recover from Covid19	Analyze mobility of different cities merging multiple types of data	An analysis of multiple cities using crowdsourced information available from datasets, to understand the changes in mobility patterns during the outbreak and recovery of the COVID19 pandemic

Figure 1.1: List of RQs with link with the chapter where the RQ is addressed, the main challenge and the contribution

benefits of GPT for mobility prompted us to further investigate these aspects. To this end, we formulated two sub-research questions, one for each potential of GPT for mobility research. First, in order to more deeply understand the potential of GPT for secondary activities analysis we formulated the following RQ:

RQ2

Can GPT be used to classify local businesses to understand dynamic demand profiles?

GPT data can be a valuable resource for understanding the popularity and attractiveness of local businesses (local business (LB)s). By analyzing the GPT at LBs, it is possible to identify patterns and trends of attractivity that may be influenced by a variety of factors, such as the type of business, location, and centrality. The objective of this RQ is to investigate the relationship between the GPT of LBs and their attractiveness. This is important for mobility analysis because it provides dynamic information on trends in secondary activities in a city. Traditionally, mobility research on secondary activities has relied on static data sources, such as OpenStreetMap data, or has required a lot of effort to collect information through travel surveys. GPT can provide a more direct and updated understanding of secondary activities, improving the analysis of mobility. In Chapter 3 we address RQ2, by proposing an analysis where we use GPT to examine LBs with the goal of identifying factors that influence their popularity and using machine learning techniques to classify the category and attractiveness of LBs based on these factors. Our approach has two main contributions: to identify the features that can impact the popularity of LBs, and to classify the category and attractiveness of LBs based on these features. The main challenge in this RQ is the normalization of the data, which can make it difficult to distinguish between different types of LBs with varying capacities.

Having established the opportunity of GPT for secondary activities we can then focus on the second contribution of this novel data for mobility research. This second potential involves exploiting GPT to obtain information on transportation systems where data may not be available or accessible. The next research question will specifically investigate this possibility using the example of a specific transportation system, the transit system. The RQ is formulated as follows:

RQ3

Can GPT be used to estimate mobility patterns such as transit demand information?

GPT do not provide in fact direct information about mobility flows. However, the goal of this research question is to determine if it is possible to develop a methodology to process GPT in order to extract the mobility flows of transit stations. This objective could result crucial because understanding transit flows can help evaluate the efficiency and effectiveness of the transit system, and such data is not always available everywhere or may only be

available for limited periods of time.

In Chapter 4, RQ3 is analyzed. We focus on examining the relationship between the GPT of a transit station and the actual transit flows at the station. We begin proposing a preliminary analysis of the correlation between these two types of data. Then the main contribution of this chapter is the development of an estimation tool that uses GPT data to estimate the transit flows at the station. This tool is designed to exploit the GPT data for the station itself in order to provide an estimate of the transit information (in- and outflows) at the station. The main challenge of this approach is to estimate two flows using a single value. Specifically, we are employing one single dataset, the GPT of the station, to estimate the in- and outflow of transit users at transit stations.

The estimation tool presented in Chapter 4 requires real transit data for training purposes. The intended use of this tool is for researchers or mobility planners to input a limited or general amount of transit data, such as a yearly report, and for the tool to utilize the GPT of stations to provide a precise and current estimation of long-term transit data. This contribution addresses RQ3 by demonstrating the ability to extract transit flows from GPT. However, it does not exhaustively explore the potential of GPT to provide transit information in situations where such data is completely absent. As a result, in order to fully investigate this aspect, the focus of the following problem is on adapting the proposed estimation tool to obtain transit information without requiring any transit data. The problem is formulated on the following RQ:

RQ4

How can we convert GPT data into transit demand information automatically?

This research question focuses on the potential of GPT to extract transit information, to be used alone without combining it with transit data. Specifically, the main question is whether it is possible to obtain transit flows from the estimation tool developed, without having to rerun the training procedure that uses transit information. In order to address this RQ, it is necessary to consider the previous contributions obtained from RQ2 and RQ3. The former demonstrates the potential of GPT to offer valuable insights into the dynamics of secondary activities, while the latter presents an estimation tool for extracting transit information. For RQ4, the aim is to merge these findings and attempt to utilize the popularity of secondary activities as a substitute for the initial data required in the estimation tool. To summarize, the objective behind this RQ is to leverage the popularity of activities around stations as a determinant for estimating flows in and out of stations.

In Chapter 5 we address Research Question 4 by developing a model that is able to estimate the transit flows at a station without the need for training

data. Specifically, we use the GPT for the activities around the station as a substitute for the transit flows data of the station. The goal of this model is to enable the estimation of transit flows at a station even in cases where transit information is not available. This can be crucial for transportation planners and researchers who need transit data in areas where such information is not accessible or does not exist. By using the model to estimate transit flows based on the GPT for the surrounding area, it may be possible to provide valuable insights into the transit flows at a station even in the absence of traditional transit data. The challenge connected with this RQ is that different cities can have different structures, which can make it difficult to compare the areas around stations in different locations. This requires adapting the model to take in consideration for the unique characteristics of each city, such as the type and density of local businesses, sociodemographic information, and other factors that may influence transit flows of stations.

The previous sub-research questions have primarily focused on the utility of crowdsourced data, such as Google popular times (gpt), for analyzing mobility trends and flows under ordinary conditions. However, it is also important to consider how such data can be used to understand mobility during anomalous events, which are occurrences that disrupt the normal functioning of transportation systems and significantly impact mobility patterns. These events may include natural disasters, public health crises, or other unexpected situations. The last research question aims to investigate the potential of using crowdsourced data to analyze mobility during anomalous events and gain insights into how these events impact mobility patterns. The RQ is formulated as follow:

RQ5

Can Crowdsourced data be used to analyze mobility during anomalous events?

Anomalous events, such as natural disasters or public health emergencies, can have a significant impact on mobility patterns. These events can disrupt transportation networks, change the way people move within a given area, and affect the demand for different modes of transportation. This research question aims to understand how crowdsourced data can be used to analyze mobility patterns during such events, and how these patterns may differ from the patterns observed in standard conditions. Chapter 6 addresses the research question of how anomalous events, such as the COVID-19 pandemic has influenced mobility patterns. Our contribution is an analysis of multiple cities using crowdsourced information available from datasets, to understand the changes in mobility patterns during the outbreak and recovery of the pandemic. We analyze data for multiple modes of transportation, including driving, walking, and transit, in order to identify patterns of

similarity between major European cities. The challenge of this RQ was the need to analyze the mobility of different cities using multiple types of data, including crowdsourced data and COVID-19 cases information. This required merging and integrating different datasets in order to understand how mobility patterns were influenced by the pandemic in each city. This process can be complex and may require adapting data analysis techniques to account for the unique characteristics of each dataset and the way it reflects mobility patterns.

1.3 Thesis Structure

This thesis is composed of four parts. Part I introduces the thesis (with context, motivation, objectives of the thesis and the addressed challenges), background and the state of the art. Then, Part II proposes solutions to address the research questions regarding the use of GPT as an indicator of mobility in standard conditions. It follows Part III which deals with the research question on how to exploit crowdsourced data during special events. Finally, Part IV concludes the thesis. The overall structure of the thesis it is presented in Figure 1.2.

The manuscript is organized as follows:

- **Chapter 2** presents essential notions required to read the dissertation. The notions include an overview of the limitations of traditional data collections for mobility and the possible solutions that can arise from crowdsourced data, with an emphasis on Google Popular Times.
- **Chapter 3** proposes an analysis of how GPT are correlated with zones and Local businesses activities.
- **Chapter 4** presents a tool that leverages as input GPT, and it is able to estimate precisely the passenger flows of individual subway stations. Our methodology is applied in 185 stations from two different cities: New York and Washington D.C. The results are validated using two months of transit count data from the stations of the two cities.
- **Chapter 5** introduce the development of a model that is able to estimate the transit flows at a station, exploiting the GPT of the surrounding area around the station, the advantage of this method is that it does not require training data from the transit and can be applied to cities where this data is not available.
- **Chapter 6** presents an analysis for multiple cities through crowdsourced information, to shed light on the changes undergone during both the outbreak and the recovery of SARS-COVID-19 pandemic.
- **Chapter 7** concludes the work and outlines future research directions .

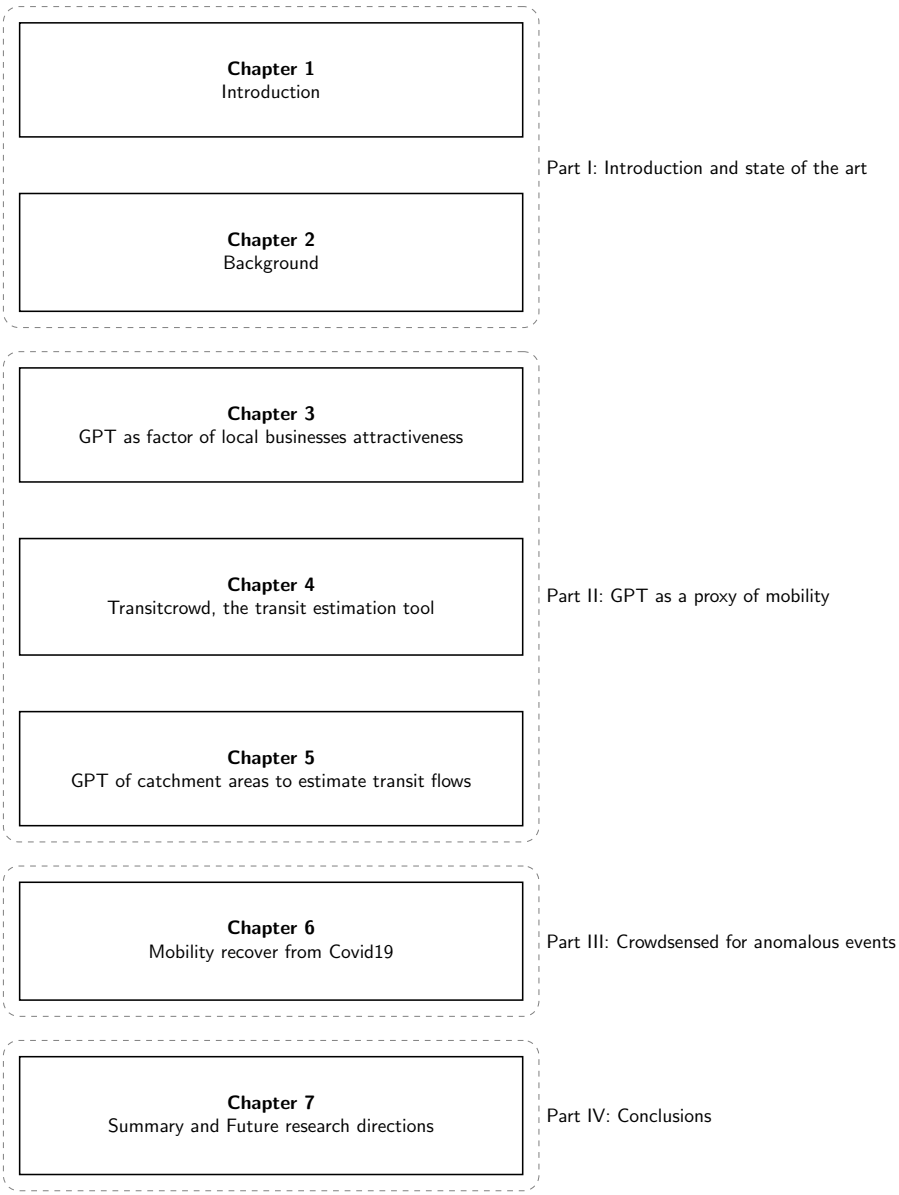


Figure 1.2: Dissertation Structure

Chapter 2

Background

*I*n this chapter, we address RQ1, " How can we leverage novel crowdsourced data for mobility analysis, and to overcome the limitations of traditional data sources? ".

The purpose of this chapter is to provide a technical background and overview of the current state of the art in datasets used for mobility analysis.

This chapter aims to provide a comprehensive overview of the various types of datasets that are used in mobility analysis and to highlight the strengths and weaknesses of each type. By understanding the characteristics and limitations of different datasets, researchers and analysts can make more informed decisions about which data sources are most appropriate for their specific needs and research questions. The chapter is divided into three main sections. In the first section, we introduce a scoring scheme that we can use to evaluate the different sources of data. In the following section, we focus on traditional mobility datasets, which are commonly used in mobility analysis. These datasets are typically collected from various sources such as travel surveys, traffic counters, smart card data, etc. They provide a wealth of information on the movement patterns of individuals, but they also have certain limitations. For example, traditional mobility datasets may not capture all forms of movement, and they may be impossible to access in certain regions. Additionally, the accuracy and granularity of the data may vary depending on the source. In the last section of the chapter, we introduce novel crowdsourced data, which refers to data that is collected and contributed by a large number of individuals through smartphones or smart devices. Crowdsourced data has the potential to offer a more comprehensive and detailed view of mobility patterns, as it can capture a wider sample of individuals and can be collected from a more diverse set of sources. However, crowdsourced data also has its own set of challenges and limitations, such as potential biases in the data collection process and the need for effective data cleaning and aggregation techniques.

2.1 Scoring scheme

In this section, we present a scoring scheme that we propose in order to evaluate the various sources of data that are used for mobility analysis. This scoring scheme is meant to compare the different datasets and assess their strengths and limitations in terms of the specific needs of mobility analysis. To do this, the scoring scheme entails several different aspects characterising the data; To identify key features to assess in the dataset, we exploited the classic big data evaluation method known as the 5Vs [22]: volume, value, variety, velocity, and veracity. To tailor this method to the needs of mobility analysis, we expanded these features as follows.

- To understand the Volume of the dataset, we took into account three key factors: **ease of collection**, **availability**, and **duration**. By evaluating the difficulty of obtaining the data, the level of accessibility of the data and the duration for which the data can be collected, we can get an overall insight into the scale and size of the dataset.

- The feature Value is directly related to the insights and knowledge

that can be gained from the data, We translated such factor in **direct measure** of mobility, which tells us which data directly measure the movement patterns of individuals or groups.

- Variety looks at the different structures of data included in the dataset and is formulated in terms of **disaggregation** level. This means evaluating the dataset based on the level of detail it provides for different aspects of mobility.

- Velocity covers the speed at which data is generated and is translated into **time dynamicity**. This relates to the frequency at which the data is updated, whether it is in real-time, daily, or weekly.

- Veracity addresses the quality and accuracy of the data and is translated into **sample size**. This includes evaluating factors such as the representativeness of the data, the potential for bias, and the completeness of the data.

For each aspect of the dataset, the scoring scheme assigns a score based on how beneficial that aspect is for mobility research. The score is defined on a scale from 1 to 3, where 1 indicates that the aspect is not useful for mobility research, and 3 indicates that the aspect is extremely beneficial. By applying this scoring scheme, we are able to identify and differentiate the most appropriate datasets for a mobility study and understand how different datasets may complement or substitute each other in order to provide a comprehensive and accurate analysis of mobility patterns. In the remaining part of the section, we will discuss the different aspects of the datasets that we will analyze using the scoring scheme.

Ease to collect

Ease of collection is an important aspect of mobility datasets, as it refers to the effort and resources required to collect the data. In the context of mobility analysis, a dataset that is easy to collect is often preferred, as it allows researchers to obtain the data more efficiently in terms of economic, time, and human resources. For example, a dataset that can be collected using automated or digital methods may be easier to collect than a dataset that requires manual data entry or field collection.

On the other hand, a dataset that is difficult to collect may be less useful for mobility analysis because it may require significant time and resources to obtain. If the dataset requires more extensive resources to collect, it may take longer to obtain, which may also influence the duration of the data collection. This may be a concern for researchers with limited time or resources. When evaluating the suitability of a dataset for a given mobility project, it is important to consider the ease of collection of the data. Datasets

that are easy to collect may be more efficient and cost-effective to obtain, making them more suitable for mobility analysis.

Availability

Availability is an important aspect of mobility datasets, as it refers to the ease with which the data can be accessed and used. A dataset that is widely available is more likely to be used by researchers and analysts, as it is easier to obtain and work with. On the other hand, a dataset that is not widely available or that has strict access restrictions may be less useful for mobility analysis, as it may be difficult or impossible for researchers to obtain and work with the data. There are several factors that can impact the availability of a mobility dataset. For example, the data may be made available through a public portal or API, alternatively, the data may be proprietary and only available to certain organizations or individuals, in which case access may be more limited. In some cases, the data may be subject to legal or regulatory restrictions, which can further impact its availability.

Disaggregation

The disaggregation aspect indicates the level of detail or granularity at which the data is collected and presented. Disaggregation level is important for mobillity analysis because it allows researchers to understand and analyze the movement of individuals at a more detailed level. In case a dataset is highly disaggregated, it may be possible to analyze the mobility patterns of specific groups or categories of individuals, such as commuters or tourists, or to examine the movement patterns within a specific geographic area. A dataset that is not highly disaggregated may be less useful for mobility analysis, as it may not provide sufficient detail to accurately understand and analyze the movement patterns of specific groups or areas. For example, a dataset that only provides aggregate data at the level of a city or region may not be sufficient to understand the mobility patterns of specific neighborhoods or groups within that city or region.

Time Dynamicity

The ability of a mobility dataset to track changes in movement patterns over time is referred to as time dynamicity. This feature allows researchers to observe the evolution of mobility patterns in different locations. A dataset with high time dynamicity, which captures changes in mobility patterns within a specific area over a longer period of time, can be useful for transportation planning and infrastructure development. It can also provide a larger sample size and be more robust for statistical tests and analyses. However, a dataset with low time dynamicity, such as one that only provides snapshot data at a

single point in time, may not have sufficient detail to accurately understand the evolution of mobility patterns over time in different locations.

In summary, it is important to consider the level of time dynamicity on mobility dataset and how it may impact the accuracy and reliability of the analysis. Datasets with high time dynamicity offer more detailed and granular data that can be used to better understand and analyze the evolution of mobility patterns over time in different locations.

Direct Measure

Mobility datasets that directly measure the movement patterns of individuals or groups, such as those using GPS traces or origin/destination matrices, provide more accurate and reliable information about mobility flows. These datasets are considered a direct measure of mobility and are useful for understanding and analyzing mobility patterns. In case of no direct measure of mobility a dataset may rely on indirect or inferred measures of mobility, this could require an elaboration of the dataset that could lead to a loss of accuracy.

Duration

The duration of a mobility dataset refers to the length of time over which the data is collected and analyzed. In the field of mobility analysis, datasets with longer durations are often preferred because they allow researchers to examine longer-term trends and patterns. For instance, a dataset with a duration of several years may provide insights into how mobility patterns have changed over time and how they may continue to evolve in the future. Consequently, datasets with shorter durations may be less useful for mobility analysis because they may not contain enough data to accurately understand and analyze mobility over different time periods. In general, it is important to consider the duration of the data when evaluating the potential of a dataset for mobility analysis. Datasets with longer durations provide more data that can be used to better understand and analyze long-term trends in mobility.

Sample Size

The number of individuals or groups whose movement patterns are captured in the data is known as the sample size, and it is a crucial aspect of mobility datasets. When conducting mobility analysis, a dataset with a larger sample size is generally preferred because it allows researchers to analyze the data with greater statistical power and accuracy. For example, a dataset with a large sample size may be able to provide insights into the movement patterns of a diverse and representative group of people, making it more representative of the overall population. Conversely, a dataset with a smaller

sample size may not provide sufficient data to accurately understand and analyze trends and patterns in mobility. Another property of small sample size data is that are subjected to sampling bias, in which the data is not representative of the overall population and may not accurately reflect the true mobility of individuals or groups.

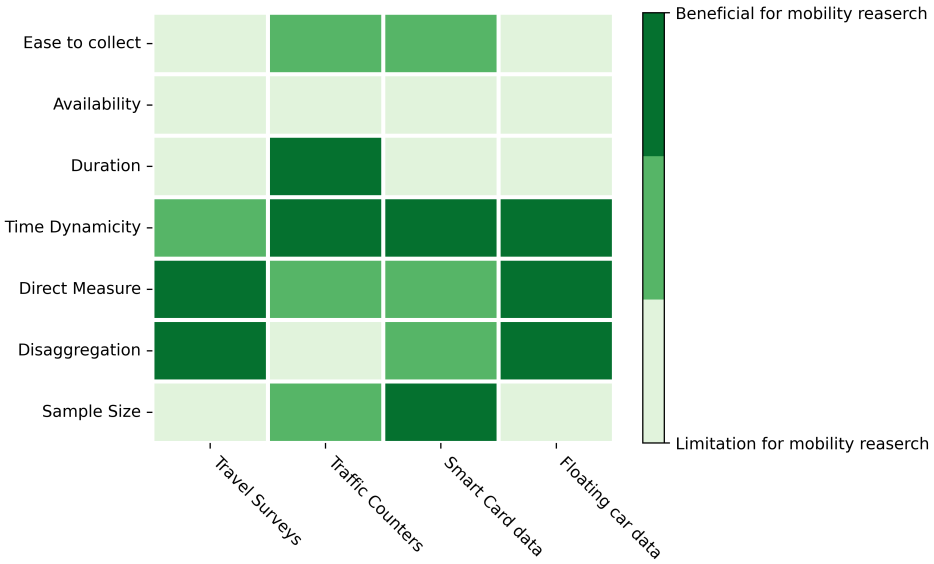


Figure 2.1: Scores traditional mobility data

2.2 Traditional mobility data

In this section, we will delve into the importance of understanding which are the datasets in the field of mobility analysis. Understanding the strengths and limitations of each datasets is fundamental to investigate the potential of data for mobility analysis [23]. Therefore, it is crucial for researchers to understand how to effectively analyze and utilize these datasets. Different types of transport and mobility data can be gathered using a wide range of techniques. To categorize the dataset that we include in our evaluation analysis, we divide them according to the way they are collected. The categories are as follows:

- Surveys: Data obtained from sources such as interviews, and administrative records fall under this category. An example of this is data from travel surveys.
- Location-based Collection: Data that is collected from specific locations, such as intersections, roads, and public transport stops, is classified under this category. An example of this is data from traffic counters.

- Paired Location Collection: Data collected from pairs of locations, such as an origin and a destination, or two points along a route, fall under this category. Examples of this include data from smart cards.
- In-Motion Collection: Data collected while in motion, using sources such as GPS trackers and onboard sensors, is classified under this category. An example of this is floating car data.

By analyzing this data, researchers can gain insights into mobility, and identify potential areas for improvement in transportation systems. While these traditional approaches have been successful in providing valuable insights into mobility, it is important to recognize their limitations as well. Figure 2.1 shows the scoring we associate to the different traditional mobility data.

In this section we will provide a detailed overview of traditional approaches for collecting mobility data including the strengths and limitations of these approaches, highlighting the unique insights they can provide as well as their potential limitations.

Surveys: Travel Surveys

Travel surveys are a common method used to gather data on mobility patterns and behaviors. These surveys are typically conducted through a variety of methods, such as online questionnaires, phone interviews, or in-person interviews, and are designed to collect information on how, when, and why individuals travel. Travel surveys can provide valuable insights into transportation patterns, such as mode choice, trip purposes, and travel behavior [24]. As shown in Fig. 2.1, a low level of benefit for mobility research is associated with the aspect of ease of collecting, this is due to the fact that travel surveys are time-consuming and resource-intensive to organize, especially if they are conducted over a long period of time. This can be a challenge for researchers and organizations who are responsible for collecting the data [25]. Consequently, travel surveys have a low score for the availability aspect because the data is not easy to extract, making it unavailable in many places. Additionally, even when the data is collected for a specific area, it is often not made available by the collector. Travel surveys tend to score highly on characteristics such as disaggregation, direct measure, and time dynamicity because they provide detailed information at the individual level with a very low level of disaggregation, direct information about respondents' trips, and information about the time when a particular person made a trip. An interesting study that reviews the practice of capturing and representing multimodal trips in travel surveys is [24], the authors analyze the implications of common practices and make recommendations to improve data collection. On the other hand, the lowest scores of this dataset are associated with the aspects of duration and sample size. Gathering

responses from a larger number of people may be more expensive and time-consuming, but it can also result in more reliable results. A smaller sample size may be easier to obtain, but it may also be less accurate. The same trade-off can be seen with the duration of a travel survey: while a longer duration may provide a more comprehensive and representative picture of travel patterns and behaviors, it may also be more difficult to organize. These two limitations, along with the possibility of incomplete reporting of travel information by respondents, can all impact the reliability and accuracy of this traditional mobility data [26]. Using GPS devices can enhance the precision of travel surveys, as outlined in [27]. The authors present a thorough examination of various techniques to exploit GPS Travel surveys for identifying trips, determining modes of transportation, and determining the purpose of travel. The article highlights the techniques utilized by researchers in the field and evaluates their advantages and disadvantages.

Location-based Collection: Traffic Counters

Traffic counting devices are used to measure the number of vehicles that pass a specific point on a road, and they are an important tool for traffic engineers and planners. These devices come in many forms, with loop detectors being the most widely used. Other types include pneumatic tubes, radar, infrared, cameras, and acoustic sensors. ALl these types of devices are commonly used in transportation planning and mobility analysis to gather data on traffic volume, speed, and other characteristics of vehicle traffic. Traffic counters are a commonly used tool in mobility research, in this section, we will examine a few examples of mobility topics for which traffic counts can be exploited. First, traffic counts can be used for transportation planning, this data can be used to analyze traffic flow patterns, predict future traffic demand, and develop transportation plans [28]. Another example of the use of traffic counts is in the analysis of environmental impacts, traffic count data can be used to understand the environmental impacts of transportation, such as the contribution of transportation to greenhouse gas emissions [29]. Moreover, traffic counts have a high impact on land use, where it is used to understand the relationship between land use and transportation, including the impact of land use patterns on traffic flow and the accessibility of different areas [30]. Overall, these examples demonstrate the utility of traffic counters data in mobility research and transportation planning. By providing data on traffic flow patterns and usage of transportation systems, traffic counters can help researchers and planners understand the characteristics of mobility and develop strategies to improve transportation systems and accessibility. The scores of fig. 2.1 highlight how the strengths of this data are in the aspects of duration and time dynamicity. Traffic counters, once installed, can continuously collect information with a high time granularity without any effort. Other positive scores are associated with ease of collection,

direct measure, and sample size. Traffic counters require an initial cost for installation, but once they are set up, the data collection process is usually smooth. These devices provide information on the flow of vehicles passing a specific location, which can be a valuable insight into mobility patterns. However, traffic counters do not typically provide information on the origin and destination of the vehicles. Additionally, while the sample size of traffic data collected by these devices is often large, it is typically limited to main roads, so it may not include data on traffic on secondary roads. However, traffic counters are not without their limitations. The lower scores are linked with the availability and disaggregation category. This data is not widely available for all locations, as not all roads and highways are equipped with traffic counters. This data is typically collected and analyzed by government agencies and transportation departments, in some cases, this data may be made available to the public but often may be available only for purchase or through a special request. Regarding the disaggregation aspect, traffic counters data is usually collected in an aggregated form, meaning that it does not include detailed information about individual vehicles. However, it is possible to obtain some level of disaggregation by dividing the data by vehicle type.

Paired Location Collection: Smart Card data

Typical approaches infer mobility of public transport users' from smartcard data. These approaches look at a variety of topics e.g., to infer bus passengers origin-destination [31], to extract information about passengers routines to predict transportation usage [32], or to measure the impact of individual characteristics on Public Transport (PT) accessibility [33]. The availability score of this data is very low. One of the major challenges with this data is that it is primarily controlled by public transportation authorities and only a limited number make their datasets publicly accessible. As a result, research utilizing smartcard data tends to be localized, with studies focusing on a particular city where the researchers have been able to secure permission to access and utilize the data. Smart card data has several strengths that make it a valuable resource for researchers and analysts. One of the main strengths is its time dynamicity, or the ability to track changes over time. By collecting data on an hourly basis, researchers can see how usage patterns and other factors change on a short-term basis. Another strength of smart card data is its large sample size. Because it includes data from all users of public transportation, it provides a comprehensive view of how the system is being used. Other important characteristics include the ease of collection, as the smart card system is already set up for the payment of tickets, so there is no need for new infrastructure. The aspects of disaggregation and direct measurement also score highly, as can be seen from Figure 2.1. This is because smart card data often provides information at the user level and

defines the entire trip of the user from the tap-in to the tap-out of the public transportation system

In-Motion Collection: Floating car data

Floating car data (FCD) involves the collection of data from GPS devices installed in vehicles as they move around on the roads. This data can be used to analyze various aspects of mobility, such as traffic patterns, road conditions, and travel times. Floating car data (FCD) has been used to improve transportation planning and operations by providing more accurate and up-to-date information about traffic conditions. For example, transportation planners can use FCD to identify bottlenecks or congestion on certain roads and make adjustments to the transportation network to improve efficiency [34]. Similarly, FCD can be used by transportation operators to optimize routes and schedules, reducing fuel consumption and emissions [35].

In addition, FCD has been used to inform the development of new transportation technologies and services, such as intelligent transportation systems or ride-sharing platforms, and determine the demand that may be able to be redirected from cars [36]. One of the main benefits of FCD is its ability to provide real-time data about the movement of vehicles on the roads. This is in contrast to traffic counters, which rely on fixed sensors or manual observations and may not be as up-to-date. FCD can also provide more detailed and specific information about individual vehicles and their trips, as it is able to collect data from individual vehicles. This allows FCD to provide not only a snapshot of current traffic conditions but also direct information about mobility patterns. The limitations of FCD are primarily related to the data collection process. FCD is often collected using GPS sensors installed on individual vehicles, which can be difficult to install on a large portion of the vehicle population. This leads to a small sample size and can make the data difficult to collect and potentially unavailable due to privacy concerns.

2.3 Crowdsourced data for mobility analysis

The pervasive adoption of mobile phones worldwide, with 91 percent of people owning a mobile phone, 86 percent owning a smartphone [1], together with the technological evolution of smartphones has provided unprecedented opportunities for collecting data in motion out of different sensors, such as GPS, accelerometers, sound recording, cameras, etc.

Mobile devices have become a concrete alternative to traditional mobility datasets; the novel Mobile CrowdSensing (MCS) paradigm allows to collect crowdsourced data from users, e.g. identifying their usual habits and inferring special events [37]. As already specified in Chapter 1.3, crowdsourced

[1] mobspread

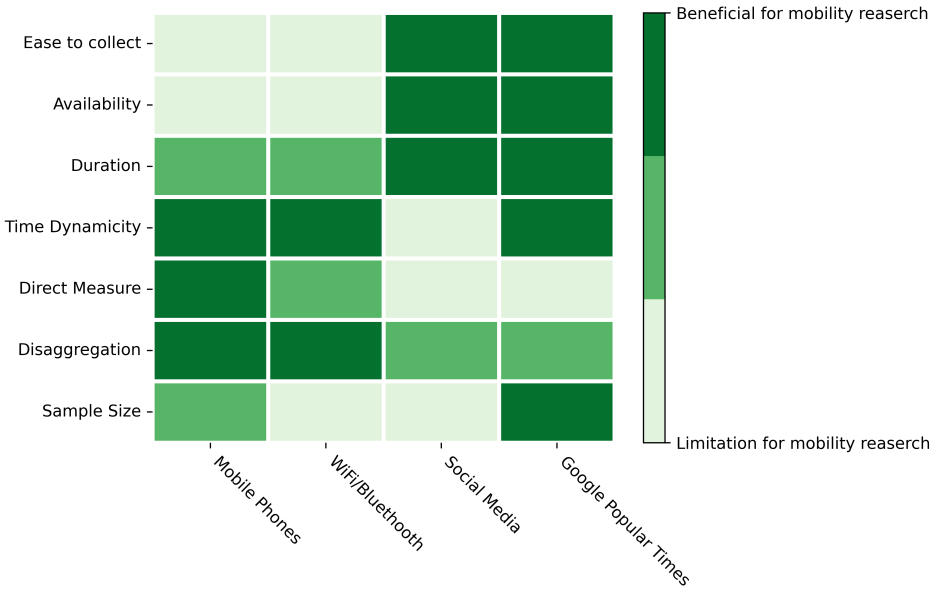


Figure 2.2: Scores Crowdsourced data

data refers to data that is collected and aggregated from a large number of people, typically through the use of mobile devices. Crowdsourced data is collected through the process of MCS, which involves the use of mobile devices, sensors, or other types of technology to gather data from a large number of people. The data is then transmitted back to a central provider, where it is aggregated, processed, and then often offered back to the public as services or information.

Crowdsourced data can contribute to addressing the drawbacks of the traditional mobility dataset detailed in section 2.2, One of the main advantages of crowdsourced data is that it can be collected in real-time, which means that it is more up-to-date and accurate than traditional datasets. Additionally, crowdsourced data can be collected over a longer period of time, and the information collected is characterized by a large sample size. The use of crowdsourced data has become a win-win solution in different domains of transportation, such as monitoring traffic dynamics and demand analysis on special events [38]. Crowdsourced-based approaches can be applied to better tackle transit demand and understand citizens' mobility. For example, crowdsourced data from the web can help to detect origin and destination of passengers in public transport [39]. Crowdsourced data approaches can be applied to better tackle urbanization issues and understand citizens mobility [40]. At the same time it allows to directly gather data from users and infer their mobility patterns with high accuracy, e.g., classify residents and visitors and identify special events [41]. Crowdsourced data have been introduced in transportation through different applications

to identify special events and disruptions or to monitor travel behavior and provide complementary information [42].

In this section, we will provide a detailed description of the main crowd-sourced datasets used for mobility analysis. We will analyze the scores shown in Figure 2.2 and describe the studies that have already exploited this novel type of data for mobility analysis.

Mobile Phones

During the past few years, several researches have exploited cellular network usage (i.e., LTE) for mobility analysis. Particularly data from loads at cell towers or call detail records (call detail records (CDR)), can be used to track the movements of individuals over time. This type of data can be useful for mobility analysis because it can provide insights into how people move within and between different areas, how long they stay in a particular location, and how their movements may vary over time. There are multiple scopes that can be addressed with such data. The authors of [43] provide a comprehensive examination of research and projects that utilize mobile phone network data for determining individuals' locations and travel patterns.

One crucial objective of mobile phone data research for mobility is determining the mode of transportation used by individuals. The authors of [44] provide a complete overview of different methods proposed for transport mode detection that use mobile phone data. In addition, the authors in [45] present a review of existing studies that have employed mobile phone data for understanding and analyzing travel behavior. This review is important because mobile phone data provides a wealth of information about an entire population, with comprehensive temporal coverage, which can be leveraged to gain insights into travel patterns and behaviors that traditional data collection methods may not reveal. Other interesting applications of mobile phones data to mobility include [46], the authors created a new framework able to exploit cellular data to measure passenger flows in subway stations in Paris, France. Mobile and wireless network data analysis can also be applied to classify subway users, distinguishing subway residents from commuters [47]. In [48] the authors created a methodology that leverages cell phone usage as a proxy to extract passengers' travel demand. Their findings help PTAs examine their public transportation options and effectively develop new transit routes or expand current routes to meet users' requirements. Unfortunately, these approaches carry significant drawbacks due to technical constraints, such as lack of location accuracy, poor network coverage, and the unwillingness of network operators to share their datasets [49]. Consequently, the availability of such data is limited, and when it is available may be difficult to obtain due to privacy concerns or legal restrictions. The scores in Fig. 2.2 indicate that, aside from the availability and ease of collection aspects that we already discussed, all the other scores are high. The data provided by mobile phones is

indeed disaggregated at the user level and varies over time. Additionally, this data describes a real mobility pattern of the user that includes the locations visited and the duration of the stay.

Wifi/Bluetooth

Similar to mobile phone data, WiFi and Bluetooth technology have emerged in the literature to capture the mobility of users. WiFi technology can be used to track the movement of users through the use of WiFi hotspots. When a device connects to a WiFi hotspot, the hotspot can record the address of the device, which can be used to identify it. By tracking the movement of devices between different WiFi hotspots, it is possible to infer the movement of the users associated with those devices. Similarly, Bluetooth technology can also be used to track the movement of users. Bluetooth devices transmit signals that can be detected by other Bluetooth devices within range. By tracking the movement of Bluetooth devices, it is possible to detect mobility patterns. Both WiFi and Bluetooth technology have been widely used for mobility analysis. In particular, WiFi sensors have been exploited to identify trajectories of metro passengers [50], to estimate real-time passengers' peak flow in order to avoid accidents [51], and to measure bus passengers' loads [52]. Although Bluetooth connections are explored more for proximity-based studies, in [53] the authors leverage this technology to detect bus passengers' origin and destination, while in [54] the authors analyze passenger dynamics and connectivity in Beijing subway. Interesting results are obtained also integrating information from WiFi with data from cameras. Several works obtained interesting results studying the integration of video information with WiFi connection especially for monitoring crowds[55][56][57]. Although these approaches are considered accurate and obtain promising results, they require every time new data collection campaigns for each specific city. This problem raises the issue of comparing a developed methodology in different cities since it would be challenging to carry multiple data collections. Regarding the scores, WiFi and Bluetooth have similar results compared to mobile phones. However, the main limitations of this data are the availability and the complexity of setting up new data collection campaigns. The size of the sample for this data is limited when compared to mobile phones, due to the technical range of WiFi hotspots and Bluetooth devices affecting the number of users that can be identified through these technologies.

Social media

In recent years, social media platforms have emerged as valuable tools for mobility research. These platforms offer a wealth of data that can be used to study human mobility patterns and understand how individuals move through urban areas. A mapping of the use of social media in transportation

is proposed in [58], the authors examine the state of social media in transportation by reviewing key studies in the literature, categorizing popular social media platforms based on their strengths and limitations.

One way that researchers have used social media data for mobility research is by analyzing the geographic information associated with content created by users, such as tweets or Instagram posts. Researchers can use the location data embedded in tweets to track the movement patterns of Twitter users [59][60]. This can provide insight into how people move through a city, including the most popular routes and destinations.

Another important contribution of social media data for mobility research is the analysis of hashtags [61]. Researchers used these hashtags to identify specific events or activities that are happening in a particular location and understand how they impact mobility patterns.

Additionally, Social Media such as Twitter or Facebook can be used to understand the experience of travelers [62], for example by analyzing the posts of travelers about their experiences with public transportation.

Social media data is attractive for various purposes due to its easy collection, widespread accessibility, and prolonged duration of data collection. This data has also big potential in addressing the limitations of traditional mobility data such as travel surveys, in [63] the authors investigate the use of social media data as a complement to travel demand survey data, and present methods for extracting relevant travel information from social media. Despite these strengths, the limitations are not negligible. First, social media data is often incomplete, as not all users are active on social media or willing to share their location data. As a result, the sample size is limited because it includes only a portion of users and may be biased. Another limitation is that data from users is not regularly distributed over time, making it difficult to reconstruct the direct mobility pattern of a user. In order to gather the trip of an individual, we would need a post for every location visited by the user.

2.4 Google Popular Times

In the context of MCS, this thesis focuses on a specific crowdsourced dataset, the Google Popular Times (GPT), due to its characteristics of wide availability and ease to collect it. In this section we introduce this type of data, and dedicate the following chapters to analyse the opportunities offered by GPT in mobility analysis. Google Popular Times (GPT) is a service within Google Map queries that visualise the temporal profile of the number of people visiting a place (points of interests such as retail shops, restaurants, public places) as a vector of normalized per-hour weekly values in the range $[0 : 100]$ (0: closing hours, 1: lowest amount of visits per-hour in a week and 100: the highest). The use of normalized values indicates the trend of an activity during a week and inherently the factors that influence such behaviour (e.g.,

a restaurant that has more success during weekends in touristic areas or at lunchtime in business districts). However, the provision of a normalised score hides the absolute quantity of the demand, i.e. the real number of customers, hence it is a relatively qualitative indicator.

The GPT is generated from data sent anonymously by smartphones with the google history location enabled, the location of these devices is tracked in the background and sent to Google through WiFi or mobile networks. To use the popular times feature, a user simply needs to search for a business or location on Google Maps and click on the business to see more information. The popular times information is displayed in a bar chart, with the horizontal axis representing the days of the week and the vertical axis representing the hours of the day. Fig. 2.3 shows how the GPT of a specific place is displayed on Google maps. The blue bars represent the standard week profile, which describes how usually busy is the place during different times of the day and based on average popularity over the last several weeks. On the other hand, the pink bar reveals the live GPT value, this value tells how busy is the place at the time of requesting the information and it is updated every hour. Together with the standard weekly profile, GPT provide a live value for the current hour, which indicates the actual level of crowding at the place. In addition to the temporal profile, GPT also displays the average duration of the activity and an estimated process time for the current time (e.g. waiting time, or typical duration of the activity performed in that place).

Fig. 2.2 indicates how GPT is associated with high scores also for the aspects of sample size, duration, and time dynamicity. The sample size for GPT represents a large and diverse dataset of individuals, based on the location history of every Google Maps user who has enabled this feature on their device, including Apple users who have installed Google Maps. Regarding the duration, GPT is continuously updated every hour, with no time limit. This has also a positive impact on the time dynamics of the data, as it allows to have information on how the data evolves over different times of the day. The disaggregation aspect of GPT is also interesting because it is not at the user level, like mobile phones, but rather at the location level. On one hand, this characteristic does not capture the entire trip of a user. On the other hand, it allows for a more detailed analysis of how different locations are being visited and used over time. This lack of a direct link with the user brings us to the main limitation of this data. GPT does not provide direct information about mobility, it does not directly measure the movement of people to and from a location. Instead, it relies on the presence of users with location tracking enabled at a given location at a particular time. To overcome this limitation and gather more accurate information about mobility, it is necessary to process and analyze the data.

The GPT can be helpful for a variety of reasons, it is a valuable tool for both individuals and businesses. For example, a tourist that is planning a trip to a museum can use the feature to see the busiest times and try to

visit during off-peak hours to avoid long lines and crowds. By analyzing the data from GPT, businesses can get a better understanding of their busiest times and use that information to better organize their activity and improve customer service. Aside of the clear added value of this information for customers and owners, GPT can be also a powerful source for mobility research. This data can provide insights into patterns of human movement and behavior. Researchers might use GPT to study how mobility patterns change over time or how they differ between different types of locations (e.g., retail stores versus restaurants). This information can be useful for a variety of applications, such as understanding how people move around a city or region, for assessing the demand of transportation infrastructure, and analyzing the impact of events or policies on mobility patterns and transportation services. In this thesis we will showcase examples of these potential application opportunities. More specifically, the potential of GPT for mobility analysis derives from the peculiar characteristics of this data. These opportunities are the following:

Opportunity 1. Large availability: contrary to the majority of mobility dataset GPT is available worldwide, at least where Google services are allowed.

Opportunity 2. Collected continuously: Google collects the information to create GPT constantly, GPT data is accessible and does not require any new data collection campaign.

Opportunity 3. Different purposes: thanks to the variety of places where GPT is available this data can help mobility research for different objectives.

Together with these relevant aspects, there are various limitations that make it challenging to analyze or manipulate this data in the context of mobility research. In order to fully exploit the powerful features of GPT, we need to overcome the following challenges:

Challenge 1. Data are normalized: as previously mentioned Google normalizes the number of users who have shared their location history. This normalization process means that the data may not reflect the actual number of people at a location. The consequence of this process is that is challenging to accurately compare the crowdedness of different places, particularly if they have different capacities.

Challenge 2. No control on the data: Google does not provide any details on how the real data of users are transformed into GPT of a location. As a result of this issue, there is no information on how Google creates the Historical and Live GPT.

Challenge 3. No direct mobility knowledge: GPT mainly provide insights regarding location crowdedness, it is challenging to extract from such data quantitative information of mobility, such as mobility flows and demand patterns.

In this thesis, we aim at addressing these limitations of GPT and explore its potential for mobility research.

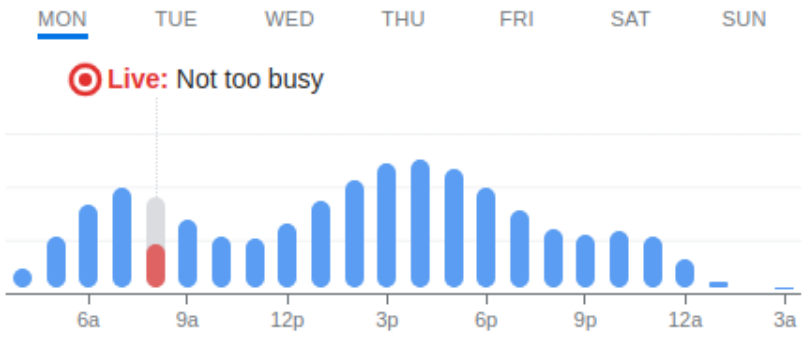


Figure 2.3: An example of GPT record

2.4.1 Related works

While there is potential for using GPT in the mobility domain, there have been relatively few studies that have explored this possibility. In this section, we describe the existing studies on GPT in the mobility domain and classify them based on the purpose of the study.

Land use and Local businesses

Since GPT provides a direct insight into the crowdedness of locations, one of the main research topics is the popularity of local businesses and their connection with land use. In [64] the authors present a machine learning-based approach for predicting venue popularity using GPT and passive sensor data. They developed a WiFi microcontroller to measure the real number of people in a place, the comparison of their data with the corresponding GPT revealed promising results. In [65], the authors use data on park use, GPT data, and land use diversity, as well as demographic and socioeconomic characteristics, to identify the factors that influence park use in the city. The results of the study indicate that land use diversity is positively correlated with park use, and that this relationship is stronger in neighborhoods with higher levels of socioeconomic disadvantage. [66] presents a method for using GPT data to model the time spent at tourism destinations. The results of the study show the value of using GPT and other online resources to analyze and predict individual behavior at tourism destinations.

Charging Stations and Parking

An interesting analysis of GPT is [67], this study analyzes the temporal variation of electric vehicle charging demand using the GPT of the activities around the charging stations. A similar approach is taken by [68], this paper investigates the demand for electric vehicle charging at popular amenities,

such as shopping and fitness centers. the findings suggest that the proposed method is a useful approach for characterizing electric vehicle charging demand at popular amenities and can be applied to various categories of businesses. GPT can be also useful for improving parking management in urban areas as demonstrated in [69]. This paper discusses the challenges of predicting parking availability and introduces a machine learning-based approach that uses data from sensors, cameras, and GPT data to predict parking availability in real-time.

Anomaly events

In the pursuit of achieving global efforts against the negative impacts of the pandemic, GPT resulted a fundamental data source during Covid19 pandemic. Google, together with Apple, decided to share data to analyse global mobility and activity trends. As result, many studies exploited the GPT dataset to analyze citizens' mobility during lockdown [70] [71], since live GPT values can be an important source to make comparisons between different time periods. In [72] the authors discuss various sources of data that can be used for estimating tourism flows with a focus on the impact of the COVID-19 pandemic. The data exploited include traditional survey data, passive sensor data, and GPT data. The paper also discusses the limitations and potential biases of these data sources and the need to carefully consider these issues when using data to estimate tourism flows.

Mobility flows

Unlike the above studies, there have been relatively few studies that have used GPT to estimate mobility flows. Notably, [73] is one of the few studies that investigated the possibility of using GPT to predict traffic volumes in a specific area. The paper discusses the challenges of using GPT for traffic management and introduces a method for using this data to predict traffic volumes in urban areas. The paper also discusses the evaluation and performance of the method and the potential applications of the results for improving traffic management and reducing traffic congestion. Overall, the paper suggests that GPT has the potential to support sustainable traffic management and can be used to improve traffic prediction and management in urban areas.

2.5 Conclusions

The aim of this chapter was to give a detailed and technical overview of the current state of mobility datasets, and to evaluate their strengths and weaknesses. This was achieved by outlining the main characteristics and

limitations of different mobility datasets, as well as by proposing a scoring system to assess various aspects of these datasets. These aspects included availability, or how easily the data can be accessed and used; temporal dynamics, or how frequently the data is updated and reflects changes in mobility patterns; and sample size, or the number of observations or data points included in the dataset. In addition to traditional mobility datasets, the use of crowdsourced data as a possible addition or alternative was also introduced and discussed. Crowdsourced data was evaluated using the same scoring system as traditional datasets, and it was explained how it could be used to overcome some of the limitations or gaps in traditional datasets. In summary, this chapter highlights that crowdsourced data has several desirable attributes, such as wide availability and prolonged periods of collection, making it a valuable resource for enhancing traditional datasets. In particular, GPT, a specific type of crowdsourced data, is particularly promising for two reasons. First, it is able to provide dynamic information about secondary activities that traditional mobility data is unable to provide. Second, due to its high availability, it can serve as a substitute for mobility data in areas where it is lacking. In the following chapters, we will explore how we leveraged these unique characteristics of GPT.

Part II

GPT as a proxy of mobility

Chapter 3

GPT as factor of local businesses attractiveness

*I*n this chapter, we address RQ2, "Can GPT be used to classify local businesses to understand dynamic demand profiles?"

We introduce an analysis of GPT of local businesses (LBs) with a twofold purpose. First, we investigate features that can influence the popularity of LBs. Second, we feed ML techniques on such dataset to classify category and attractiveness of LBs according to the considered features.

*T*his chapter is based on work that has been published in the following paper:

- Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities
  A. Capponi, P. Vitello, C. Fiandrino, G. Cantelmo, D. Kliazovich, U. Sorger, P. Bouvry
  2019 IEEE Symposium on Computers and Communications (ISCC)

Local businesses (LBs) require their owners (e.g., companies, individuals, and institutions) to take decisions for profit maximization and to offer competitive services to customers. The most crucial decisions include the typology of a LB and its location when opening, but also setting prices, the number of required employees per hour, and opening hours for proper management. Effective strategies to boost LBs require knowledge of the complex dynamics of urban environments, which depend on the spatial distribution of citizens and locations [74]. For instance, understanding real-time citizens' mobility as well as forecasting significant flows of citizens moving to a specific urban area for a special event helps municipalities to manage crowds and entrepreneurs in deciding suitable locations and required staff.

Traditional approaches to investigate LBs popularity rely either on surveys that capture users' preferences or cellular traces that infer urban mobility [75]. However, such approaches are prone to users misbehavior, technical limitations (e.g., poor network coverage), and datasets available only from network operators [49]. Crowdsensed data-driven approaches may provide novel solutions in this direction by exploiting MCS systems and services like GPT that make available accurate information on travel times and popularity of LBs.

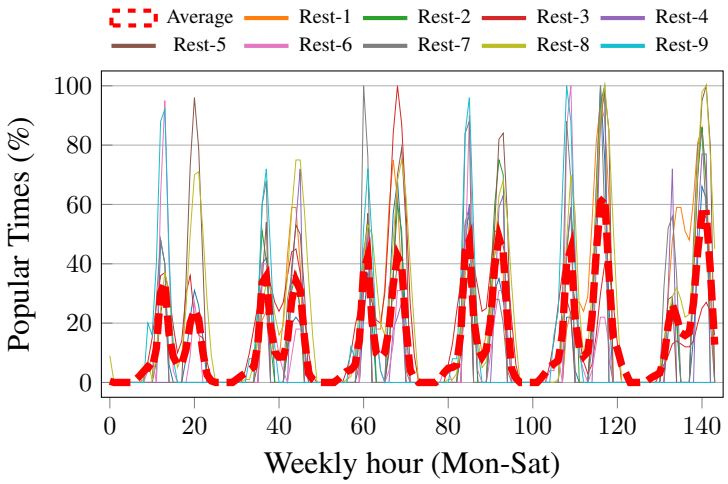
This work aims to bring one step further the research on urban computing and to boost LBs popularity by overcoming the limitations of historically experience-driven approaches. We leverage GPT data to enforce highly-accurate classification of LBs category and attractiveness with ML techniques that are powerful to handle massive data volumes and widely employed from a variety of applications, such as to infer and predict human mobility in an urban context [76]. In this work, we show that typical urban metrics (e.g., the centrality of places in street networks) fail to properly classify LBs, while combining GPT information (e.g., peak hours in LBs) with basic ML techniques supports and improves typical experience-driven approaches. To illustrate with few representative examples, restaurants and pubs usually concentrate in close areas and influence one with each other, while LBs like pharmacies are uniformly distributed over a city. Also, reachability by public transport significantly impacts on LBs popularity.

3.1 Preliminary Analysis

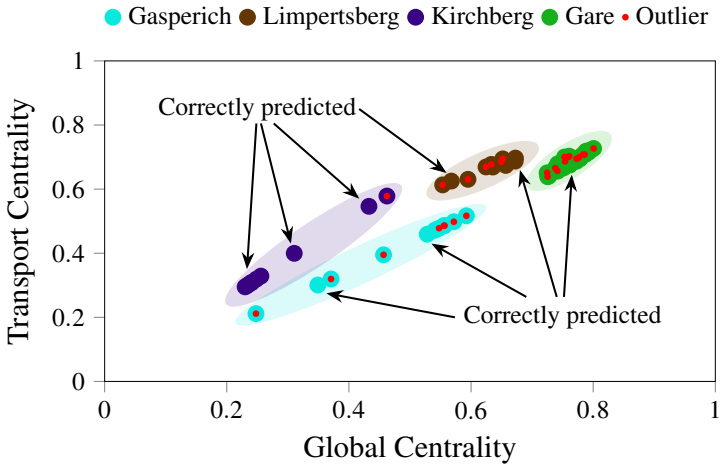
This Section grounds the roots of our work by showing why traditional urban metrics fail to classify and predict LBs attractiveness properly.

Weekly popularity:

To analyze the popularity trend over time of a LB, we use GPT data. This enables us to examine the trends of LBs throughout a week. However, because the GPT data is normalized, it hides the level of success of a single LB (such as if it has more customers than others).



(a) Average popularity of restaurants in Ville Haute



(b) Centrality and similarity

Figure 3.1: Data aggregated from different Luxembourg districts for restaurants

Fig. 3.1(a) presents Popular Times of nine restaurants and their average from Monday to Saturday in Luxembourg city (Ville Haute), a district with offices (banks, public institutions), shops, tourists spots and places for nightlife (bars, pubs). Sunday is excluded because no information was available in the dataset. The peaks of popularity approximately at 12, 20, 36, 44, etc., correspond to lunch (12 PM) and dinner (8 PM) times of each day. Analyzing the peaks in pairs, we can compare the trend of restaurants day by day and understand the lifestyle of the district. During weekdays the peaks are around lunch time or equally spread at lunch-dinner time (restaurants full of workers) while on Saturday at dinner time because most offices are closed.

Friday is the most popular day at both lunch and dinner times because both workers, tourists, and citizens populate restaurants.

Centrality and similarity: We can observe that the popularity of LBs depends on their proximity with other LBs and accessibility through public transportation. The centrality metric, which defines the importance of individual nodes in a network, can quantify popularity. Specifically, we consider the *closeness centrality*, defined as the sum of the length of the shortest paths between a node and all other nodes within the street network. We measure *global-centrality* and *transport-centrality*. The *global-centrality* defines the proximity of a LB with all other LBs:

$$C_B(k) = \frac{N_B - 1}{\sum_{i \neq k} d_{ki}}, \tag{3.1}$$

where k is the k-th node, N_B is the total number of LBs and d_{ki} is the distance between a couple of nodes. The *transport-centrality* measures the proximity of a LB with respect to transport facilities:

$$C_T(k) = \frac{N_T}{\sum_{i \neq k} d_{ki}}, \tag{3.2}$$

where N_T is the total number of transportation access points (e.g., bus stops or metro stations) and d_{ki} is the distance between the considered LB and a transport node. Considering the Earth as an oblate ellipsoid, the distance is computed with the shortest geodesic path [77]. While popularity measured with centrality identifies time-invariant characteristics of a LB, the *similarity* compares two LBs temporal profiles. The similarity aims to correlate LB weekly popularity to the average of all LBs in the same district. To measure similarity, we exploit the symmetric index of Jensen-Shannon divergence (Jensen-Shannon divergence (JSD)) that outperforms the asymmetric Kullback-Leibler divergence (KLD) [78]. The similarity of two LBs i and j is:

$$J(D_i, D_j) = H\Big(\frac{D_i + D_j}{2}\Big) - \frac{H(D_i) + H(D_j)}{2}, \tag{3.3}$$

where H is the Shannon entropy, D is the temporal profile of a LB, and J represents the divergence of two temporal profiles, The similarity can assume values in the range $[0 - 1]$. 0 represents the maximum similarity (e.g., the temporal pattern of a shop with itself) and 1 represents the maximum divergence.

Fig. 3.1(b) links centrality and similarity metrics in Luxembourg city. The clusters represent four districts, in line with global- and transport-centrality of restaurants. The red dots represent LBs whose weekly temporal demand is closer to the average of other districts (outliers). On the contrary, dots of the same dominant color have a weekly pattern more similar to their

geographical district. With the sole exception of Kirchberg district, most of the LBs are marked as outliers. Therefore, this analysis unveils that centrality and similarity are not enough to assess the popularity of LBs and their relationship with districts. The analysis correctly predicts the popularity of LBs in Kirchberg because the district is geographically separated from other districts of the city and it is home of European agencies, insurance and financial companies making the LBs in the area to share peculiar popularity trends.

The remainder of this chapter shows how to overcome this shortcoming by enforcing a ML-based analysis of the same dataset.

3.2 ML-augmented Methodology

This Section describes the methodology for applying well-established ML techniques to crowdsensed data. The GPT datasets undergo a procedure to extract features and determine the most suitable inputs to train the ML algorithms. We select only the features that augment the output accuracy after the training phase, while the others are discarded. For space reasons, we omit this preliminary selection. Next (§ 3.2.1), we introduce the ML algorithms. Then (§ 3.2.2), we discuss the considered multi-classification problems, extracted input features, and output classes. Each output is classified by exploiting a one-vs-all approach. For each LB, the element corresponding to the predicted class is set to one, all others to zero.

3.2.1 Machine Learning Techniques

This study considers Support Vector Machine (Support Vector Machine (SVM)) with a Gaussian kernel and MultiLayer Perceptron (MultiLayer Perceptron (MLP)) neural network techniques for multi-classification problems. The choice is due to the characteristics of our study, which presents a small number of features N (e.g., $1 - 1\,000$), and an intermediate number of M training samples (e.g., $1 - 50\,000$). The chosen ML approaches perfectly fit this scenario. Similar ML techniques like logistic regression or SVM without kernel (or linear kernel) have not been considered because they perform better when N is relatively large if compared to M (e.g., 10 000 and M between 1 and 1 000). In the following, we briefly analyze the considered ML techniques.

Support Vector Machines (SVMs) aim to classify input samples into output classes by dividing a hyperplane with an optimal boundary through kernel methods. To this end, it is crucial to perform fine tuning of the regularization parameter, typically named C. Furthermore, employing a kernel based on a gaussian function, it is required to set the standard deviation, indicated as γ. Paramerer C trade-offs the correct classification of training samples and the smooth decision boundary. Small values lead to simple decision

functions, which correspond to a higher tolerance to errors and smooth the classification on the training dataset. On the contrary, high values correspond to a classification with minimal error and a hyperplane with a small margin. Intuitively, γ defines how a single training sample influences other points according to its distance from the boundary.

Multilayer Perceptron (MLP) is a feedforward artificial neural network that takes a vector as input and maps it into another vector as output. It is based on different hidden layers that connect inputs to outputs. Each layer includes a certain number of nodes and nodes of different layers are connected by links with different weights. The output of a node at each layer is given by the weighted sum of all inputs. Each node in the hidden layers is connected to all nodes of next and previous layers for a fully connected topology.

3.2.2 Predicting LBs Category and Attractiveness

We formulate two multi-classification problems to predict LBs category and attractiveness by feeding ML techniques with input features extracted from crowdsensed data.

Extracted features: We select as input features from the large available datasets those that performed better and we categorize them as *intrinsic* and *extrinsic*. Intrinsic features are given by geo-location characteristics and owners' decisions, which do not present a high variability over time (e.g., opening hours and type of service offered). These properties are already widely exploited in traditional approaches for urban analytics. Extrinsic features depend on the temporal interactions of citizens with LBs, such as waiting time and average time of staying. They change more rapidly than the intrinsic ones and depend on several factors, e.g., special events, time of day, day of the week, etc. The intrinsic features we consider are 1) *global-centrality*, 2) *transport-centrality*, 3) *opening hours*, and 4) *category*. The parameters that define centrality have already been discussed in Sec. 3.1. *Opening hours* consists of an array of 144 binary values (Mon-Sat) that shows when a LB opens. For each hour in a weekday, the value 0 indicates closing time and 1 opening time. The category depends on the service offered by LBs. The extrinsic features are *popular times*, *average time of visit*, and *average waiting time*. Popular times were discussed in Sec. 3.1. The average remaining time defines in minutes the duration of customers' visits. The average waiting time indicates the minutes while waiting to access the service.

Output classes: LB categories depend on the service offered by a LB. Output classes are *public*, *store*, *health*, *restaurant*, and *bar*. The class *public* indicates generic services and offices for the community, such as institutions, post, financial and insurance companies. *Store* includes each kind of shop or seller for any goods, such as supermarkets, clothing, bakeries, etc. *Health*

comprises public and private places related to healthcare, e.g., hospitals, medical centers, dentists, and specialists. *Restaurant* includes all LBs that prepare meals with seating places. *Bar* consists of LBs selling mainly drinks, but can also include meals, e.g., pubs.

LB attractiveness is classified into *working*, *nightlife*, *weekend*, *business hours* (Bus. H.), and *shopping hours* (Shop. H.). *Working* indicates LBs with peak hours during break times of working areas, such as weekdays at mid-morning and lunchtime. It comprises typically shopping malls, bars, fast foods and some types of restaurants. The class *nightlife* shows peak hours at dinner times during all week and overall on weekends, including restaurants, pubs, and clubs. *Weekend* describes low popularity on weekdays and peak hours at weekends, which is typical of shopping malls located far from working areas and touristic places. *Business hours* indicates typical opening times and consequent popular hours of public offices, from early morning to mid-afternoon including lunch breaks. *Shopping hours* include the typical popularity hours of shops for different goods, presenting a uniform distribution on both weekdays and weekends during daytime.

3.3 Data-driven Evaluation

This Section first presents simulation set-up and performance metrics, then the obtained results.

3.3.1 Setting

To conduct the evaluation, we employ publicly available Popular Times of LBs for Luxembourg city and the city of Munich downloaded between July 21st and July 30th, 2018. These two cities present different characteristics in terms of morphology, size, street topology, and lifestyles of residents and visitors. This permits to conduct an effective analysis and discussion of the obtained performance. The datasets include 1 084 and 3 784 LBs for Luxembourg city and Munich respectively and are proportionally divided in 80%, 10%, and 10% for training, cross-validation, and test phases respectively. The performance evaluation exploits Scikit-learn, which is a Python-based open-source library.

To predict the LBs category, the input features are: *average opening hours*, *time spent*, *global-*, and *transport-centrality*. In this case, we restrict the datasets to the LBs for which information on time spent is available (800 and 1 600 LBs for Luxembourg city and Munich respectively). The hyperparameters in the SVM approach are set to $C = 2^8$ and $\gamma = 2^{-12}$. We will further discuss the rationale about the selection of parameters (see discussion Fig. 3.3a). In the MLP approach, an exhaustive search with a grid-search algorithm leads to the choice of one hidden layer with 13 nodes.

Table 3.1: Statistics for LB category and attractiveness prediction

		Precision				Recall				F1 score				Accuracy			
		SVM		MLP		SVM		MLP		SVM		MLP		SVM		MLP	
		Lux	Mun	Lux	Mun	Lux	Mun	Lux	Mun	Lux	Mun	Lux	Mun	Lux	Mun	Lux	Mun
Category	Public	0.67	0.60	1.00	0.67	0.40	0.75	0.40	0.50	0.50	0.67	0.57	0.57	0.84	0.81	0.81	0.76
	Store	0.82	0.88	0.81	0.87	0.95	0.92	0.89	0.91	0.88	0.90	0.85	0.89				
	Health	1.00	0.75	0.75	0.67	0.40	0.92	0.60	0.92	0.57	0.83	0.67	0.77				
	Restaurant	0.93	0.79	0.90	0.77	0.93	0.87	0.88	0.79	0.93	0.83	0.89	0.78				
	Bar	0.60	0.80	0.45	0.60	0.75	0.53	0.62	0.47	0.67	0.63	0.53	0.53				
	Average	0.85	0.81	0.83	0.75	0.84	0.81	0.81	0.76	0.84	0.80	0.81	0.75				
Attractiveness	Working	0.85	0.60	1.00	0.57	0.58	0.60	0.40	0.40	0.69	0.60	0.57	0.47	0.84	0.87	0.80	0.88
	Nightlife	0.60	0.88	0.81	0.86	0.75	0.73	0.89	0.77	0.67	0.80	0.85	0.81				
	Weekend	0.80	0.71	0.75	0.79	1.00	0.67	0.60	0.73	0.89	0.69	0.67	0.76				
	Business hours	0.93	0.96	0.90	0.96	0.96	0.97	0.88	0.97	0.94	0.97	0.89	0.97				
	Shopping hours	0.80	0.81	0.45	0.81	0.80	0.91	0.62	0.92	0.80	0.86	0.53	0.86				
	Average	0.85	0.87	0.83	0.87	0.84	0.87	0.81	0.87	0.84	0.87	0.81	0.87				

For LBs attractiveness, the considered features are *opening hours*, *category*, *district*, *popular times*, *global-*, and *transport-centrality*. In this case, the entire datasets were employed. The methodology followed to set the hyperparameters is as for the LBs category. For SVM, the parameters are $C = 2^6$ and $\gamma = 2^{-10}$ (likewise above, the rational is discussed in Fig. 3.3b), MLP consists of 8 nodes per layer with 2 hidden layers.

3.3.2 Performance Metrics

We consider precision, recall, F1 score, and accuracy indexes. While precision, recall, and F1 score are per-class measures, the accuracy averages the measures of all the classes. For completeness of the analysis, we consider i) *true positive* (tp) and *true negative* (tn) values to indicate respectively a *correct* prediction of positive or negative class; (ii) *false positive* (fp) and *false negative* (fn) values to denote an *incorrect* prediction. In this context, a positive observation indicates the class under analysis, while a negative observation indicates all the other classes, according to the one-vs-all approach.

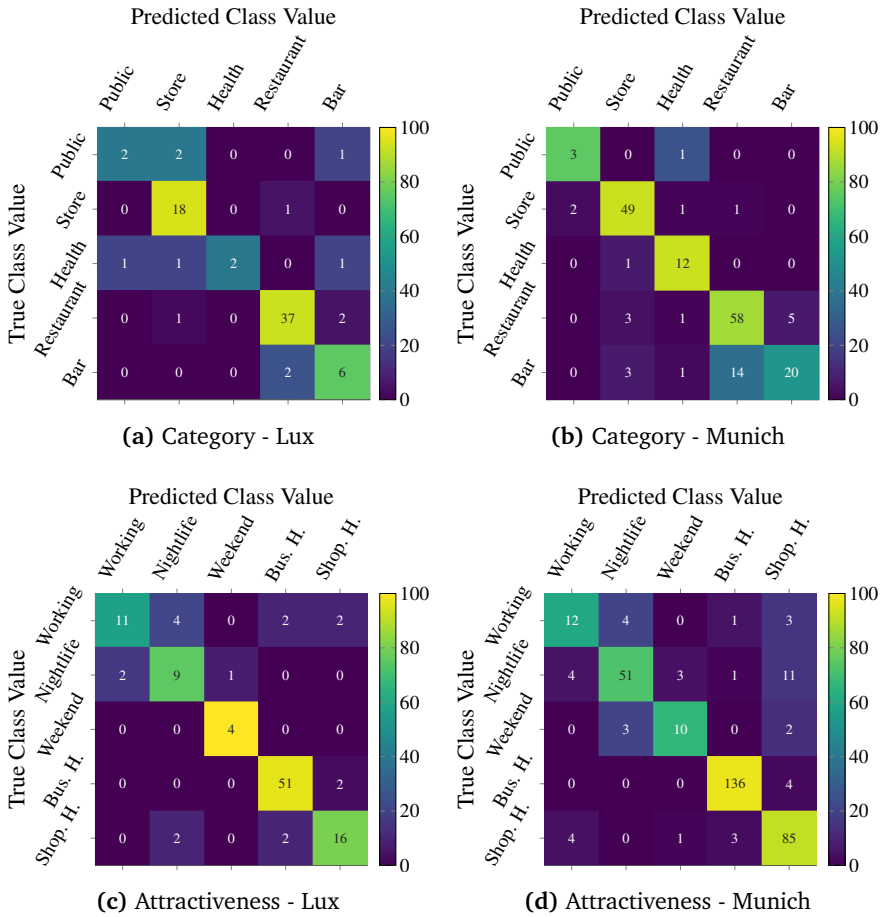
The *precision* indicates the ratio of correct positive predictions over the total predicted positive occurrences ($tp/(tp + fp)$). In other words, it indicates the capacity of the model to *not* predict another true class as the actual class. The *recall* is the ratio of correct predictions on positive observations to all the occurrences in class under analysis ($tp/(tp + fn)$). It indicates the capability of the model to catch all the samples of a class. The *F1 score* is the weighted average of precision and recall indexes and analyzes incorrect predictions. Typically, the F1 score is very useful to unveil insights from results when false positives and false negatives have different costs. The *accuracy* is computed as the ratio of correct predictions over the total occurrences and defines the performance of a classifier. Specifically, accuracy is the optimal performance indicator when the classes are symmetric, i.e., incorrect predictions have the same weights.

3.3.3 Results

Table 3.1 presents detailed results on precision, recall, F1 score, and accuracy for the predicted categories in both cities with MLP and SVM approaches. The prediction on LB categories presents higher accuracy for Luxembourg city with both ML techniques. Viceversa, Munich shows higher accuracy in predicting LBs attractiveness. Regarding the ML techniques, SVM presents an overall accuracy higher than MLP. The Table clearly shows that precision achieves high values for categories of restaurant, health, and store, while it is low for bar and public because these categories share common characteristics with other categories. The LBs prediction precision varies in the two cities because it depends on specific characteristics of each city, mainly type of visitors (e.g., tourists, workers, or residents) and city lifestyle (e.g., commercial, touristic, or working areas). For example, restaurants

present higher values of precision in Luxembourg city because the opening hours are not as international as in a larger city like Munich, while bars are predicted with higher precision in Munich. In Luxembourg city, bars and restaurants share opening hours while in Munich pubs and clubs open until late night, unlike restaurants. Regarding the attractiveness, on the one hand, the class working is predicted with much higher precision in Luxembourg. The reason is as follows: LBs with peak visits during job breaks are typically not popular, i.e., receive lower visits during other moments of the day or with another type of customers (e.g., tourists at the weekend). On the other hand, the class working is not well predicted in Munich because LBs are popular at different times during the day with no distinctive working areas. Business and shopping hours present higher values in Munich because of its urban plan characterized by LBs concentrated in specific districts with easily recognizable peak hours (e.g., the city center and shopping malls). For similar reasons, note that the model catches most samples of class (recall index) for restaurants and stores with both cities and both techniques when predicting the category, and business and shopping hours when predicting the attractiveness. F1 score analyzes the incorrect predictions by presenting a weighted average of recall and precision and the results are in line with previous considerations.

To gain additional insight, Fig. 3.2 depicts confusion matrices to highlight single occurrences for each true and predicted class and summarizes the prediction results. Each cell contains a value that indicates the number of occurrences of a predicted class when testing true inputs. The colors in legend bars represent the percentage of correct predicted occurrences over the total of true class values, which corresponds to the recall index between 0 and 1. The columns show predicted class values. The sum of all values in each row indicates the total occurrences for such class. The occurrences of correct predictions for each class are in the diagonal. The accuracy is the sum of all elements on the diagonal on all elements of the matrix. The analysis on the confusion matrices allows to i) discuss and compare behaviors of different LBs and ii) extend the discussion in point i) to different cities. As expected and already pointed out in Table 3.1, categories with distinctive features present a better prediction. The results in the table, however, do not show the wrong occurrences as confusion matrices allow. The categories *restaurant* and *store* achieve higher recall for both ML techniques in both cities because LBs in these categories share distinctive characteristics like *opening hours*. On the opposite, *public* and *bar* have a lower recall, and their wrong predictions occur respectively in *store* and *restaurant*. These LBs offer services with similar daily patterns, e.g., stores - public offices, and bars - restaurants. Fig. 3.2(a) and Fig. 3.2(b) clearly highlight these considerations because in Luxembourg city 2 bars over 8 are predicted as restaurants whereas for Munich this occurs for 14 LBs over 34. Note that the *health* category achieves significantly different results in the two cities. The



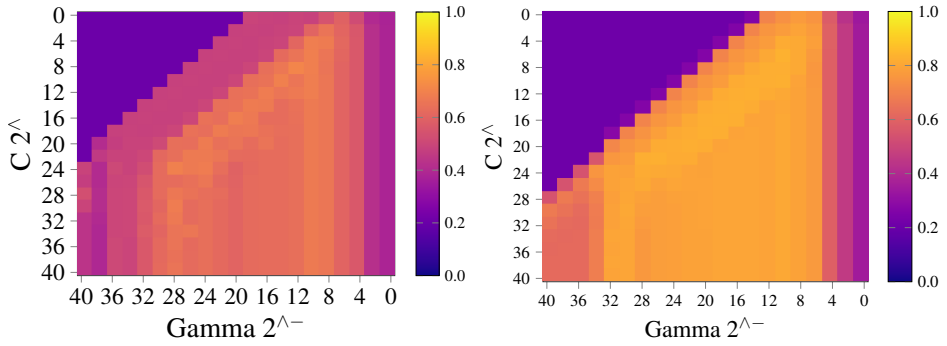
(a) Category - Lux **(b)** Category - Munich

(c) Attractiveness - Lux **(d)** Attractiveness - Munich

Figure 3.2: Confusion matrices for LBs category and attractiveness prediction with SVM technique. The rows show true class values and the columns show predicted class values.

motivation is the different number of LBs in the available datasets. In this case, higher precision is attributed to a larger dataset.

When analyzing the attractiveness, Fig. 3.2(c) and Fig. 3.2(d) unveil that the highest number of prediction errors occur for *working* and *nightlife* classes. As previously discussed, the motivation is that restaurants and bars exhibit a high popularity at lunch and dinner times, which are typical characteristics shared between *working* and *nightlife* classes. For instance, Fig. 3.2(c) and Fig. 3.2(d) respectively show that in 4 occurrences over 15 and in 4 over 16 working class true values are predicted as nightlife. The highest number of correct predictions occur for *business hours* (51 over 53 in Luxembourg city, 156 over 160 in Munich), as the popularity is uniform during all weekdays.



(a) Category **(b)** Attractiveness

Figure 3.3: Analysis of F1 score to optimize the SVM parameter selection for Munich. The values range between 0 and 1.

By comparing the two cities, Fig. 3.2(c) and Fig. 3.2(d) show that it is easier to predict the *weekend* class in Luxembourg city (4 over 4) than in Munich (10 over 15). While Luxembourg city is a destination popular for business and not for tourism, the amount of visits in LBs varies consistently between weekdays and weekends. On the opposite, in Munich it varies only a little.

Fig. 3.3 shows an analysis on the dataset of Munich for choosing the best parameters fitting the SVM technique in predicting LB category and attractiveness. Results are obtained by considering the F1 score to seek a balance between Precision and Recall. Specifically, SVM optimization parameters are $C = 2^8$ and $\gamma = 2^{-12}$ for LB category prediction, while they are $C = 2^6$ and $\gamma = 2^{-10}$ for LB attractiveness prediction.

3.4 Conclusion

This chapter aimed to explore the potential of GPT as a valuable source for dynamic information on trends in secondary activities in a city. Such information is lacking in traditional mobility data. To this end, we applied ML techniques on GPT to perform accurate predictions of LB category and attractiveness. Specifically, the work shows that ML-driven analysis outperforms historical urban computing metrics. After a preliminary analysis, the LB category and attractiveness are predicted using two different subsets of features extracted from crowdsourced data. The conducted evaluation shows that data-driven approaches outperform traditional urban metrics. The results unveil that classes exhibiting similar behaviors present higher errors when predicting their occurrences. For instance, the attractiveness of nightlife and working in a large-scale city like Munich can be miscategorized because they both include many restaurants and bars. The findings of this study can result valuable to analyze the trends of activities in a city. This is

because By quantifying and differentiating the level of attractiveness of LBs in a zone, we can also extrapolate this information to the zonal level. The attractiveness of zones of destination is a key determinant of the demand for travel to those areas. This source can be a valuable input for mobility models, for example on the generation and distribution of steps in the conventional 4-step model[79]. Attractiveness can be used to inform demand modeling in several ways. It can be used to predict the likelihood of individuals choosing to travel to a particular area for a specific activity at a specific time. It can also be exploited to predict the likelihood of individuals choosing to live or work in an area based on the presence of desirable LBs. Additionally, it can be used to inform the development and management of transportation infrastructure in a given area, such as the placement of bus stops or the development of bike-sharing stations. In this chapter, we examined how GPT can overcome the limitations of traditional mobility data and identify trends in secondary activities. In the following chapters, we will investigate how GPT can enhance the availability of mobility data in cases where it is lacking. This is the other key advantage of GPT as a source to tackle the limitations of traditional mobility data.

Chapter 4

Transitcrowd, the transit estimation tool

In this chapter, we address RQ3, " Can GPT be used to estimate mobility patterns such as transit demand information? ".

This chapter aims to investigate the possibility of using Google Popular Times (GPT), to estimate the passenger flows of individual subway stations. Since GPT only provide popularity trends of the stations in terms of crowding, we provide a tool that leverages as input GPT, and it is able to estimate precisely both entrances and exits profiles. Our methodology is applied in 185 stations from two different cities: New York and Washington D.C. The results are validated using two months of transit count data from the stations of the two cities.

The content of this chapter is based on a work that has been accepted for presentation in the following paper:

- Exploring the Potential of Google Popular Times for Transit Demand Estimation
  P. Vitello, C. Fiandrino, R.D. Connors, F. Viti
  Transportation Research Board 102nd Annual Meeting

4.1 Introduction

In this chapter we want to investigate the potential of GPT as a source of information for public transport demand. This dataset has the advantage of being already provided by Google and not requiring new data collections. Moreover, the worldwide availability of GPT opens up the possibility of estimating public transport (PT) demand in areas where such information is not collected. However, the main limitation of GPT is the lack of transparency in the data processing, as Google only provides the processed data in an aggregated and normalised way. This study aims at overcoming this shortcoming by combining the GPT with real public transport data in order to leverage the GPT data to estimate the in- and outflow of transit users at subway stations.

The literature on studies that explore the importance of GPT for the transportation field is thin. In this chapter, we precisely aim at filling this gap, focusing on PT. A main motivation for focusing on this transport service is the general lack of data for such systems with respect to the other main mode of transport, i.e. car transport. To achieve this goal, we need to overcome the following challenges:

Challenge 1. Unavailability of transit data. Most mobility operators do not provide any transit information. Since GPT is worldwide available, can it be exploited to estimate this data where it is not accessible?

Challenge 2. Granularity of transit data, when available, differs city by city. GPT has a value per hour, it can enrich transit data where the granularity is low.

Challenge 3. Estimate two flows using a single value. Specifically, we are employing one single dataset, the GPT, to estimate the in- and outflow of transit users at subway stations.

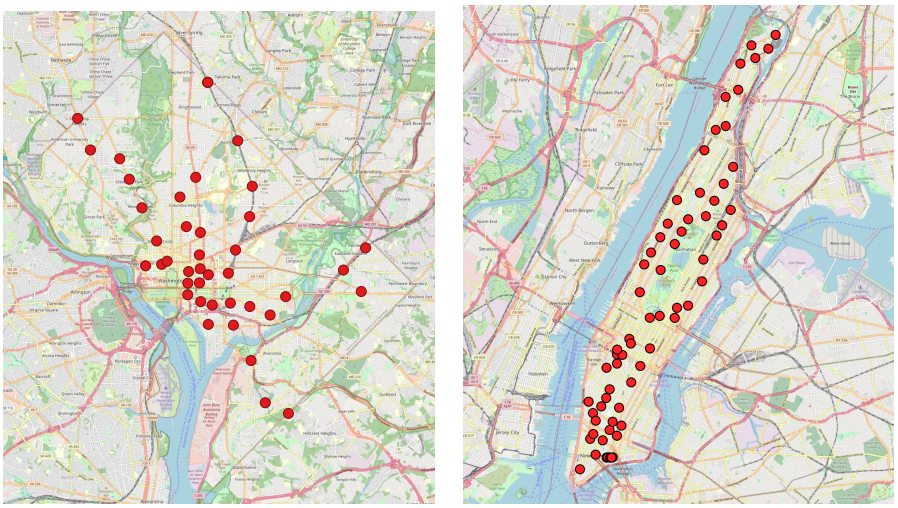
To overcome these challenges, we design TransitCrowd, a framework that is able to make live estimations of transit data exploiting only the GPT of stations. In summary, the synopsis of contributions we make with this chapter is as follows.

Contribution 1. TransitCrowd estimates live transit data regardless of the granularity of the input transit data. Our tool is composed by two different estimators.

Contribution 2. The first estimator (Reg estimator) is trained separately in every single city, it requires an initial transit dataset and focuses on obtaining the maximum accuracy in the trained area. This tool is suited for areas where a transit dataset is available with low granularity or that is limited on time.

Contribution 3. The second estimator (Sig estimator) is more flexible. It gives the possibility to transfer the methodology without requiring starting transit data but at the cost of lower accuracy. This estimator can be leveraged in case no transit data is available for the area under analysis.

In the remainder of the chapter, the next section presents the description of the data exploited in the study, followed by the methodology behind



(a) Washington DC stations

(b) Manhattan (New York City (NYC)) stations

Figure 4.1: Maps of the cities considered in our study

the framework, and the evaluation of the results. Finally, the last section concludes the work and highlights the final remarks.

4.2 Dataset and First Observations

In this section, we describe the dataset we exploit in our analysis of transit stations demand.

4.2.1 Google Popular Times

In this work, we focus on analyzing GPT of the subway stations to investigate if such information can be exploited to determine the inflow and outflow of users at the station. Our dataset includes the GPT for 105 subway stations from the Manhattan region, NYC, and 80 subway stations from Washington DC, USA.

4.2.2 Turnstile Data

For analyzing the public transport demand we considered the data shared by the Public Transit Authorities (PTAs) of the two cities in our dataset. For New York the Metropolitan Transportation Authority (Metropolitan Transportation Authority (MTA)) provides information for boarding and alighting passengers

for all the subway stations[1], while for Washington we exploited entrances and exits data provided by the Washington Metropolitan Area Transit Authority (Washington Metropolitan Area Transit Authority (WMATA))[2]. The data of New York consists of the number of turnstile entries and exits for subway stations aggregated in four hour intervals. The information we considered includes 1.135 unique turnstile positions that are associated with 732 station entrances or exits of 105 subway stations within the island of Manhattan. The data of Washington include directly the information of entrances and exits per hour for every subway station in the city, our dataset contains the entrances/exits values for 80 subway stations in Washington area. We collected two months of transit dataset for both cities.

In order to compare the transit data with GPT we needed a dataset of the same length. To this end, we exploited the first month of transit to create a typical weekly profile made by averaging the transit data of the same hours and days of the week.

4.2.3 Preliminary Analysis

In this first phase, we want to detect which information from the transit dataset of a station is the most similar to the GPT profile. The scope is to understand how the increase or decrease of the GPT percentage is correlated with the real amount of passengers entering or exiting from the stations. To analyze the transit usage data and its correlation with the GPT we use the following linear regression model:

$$G_{h,s} = \beta T_{h,s} + \epsilon, \tag{4.1}$$

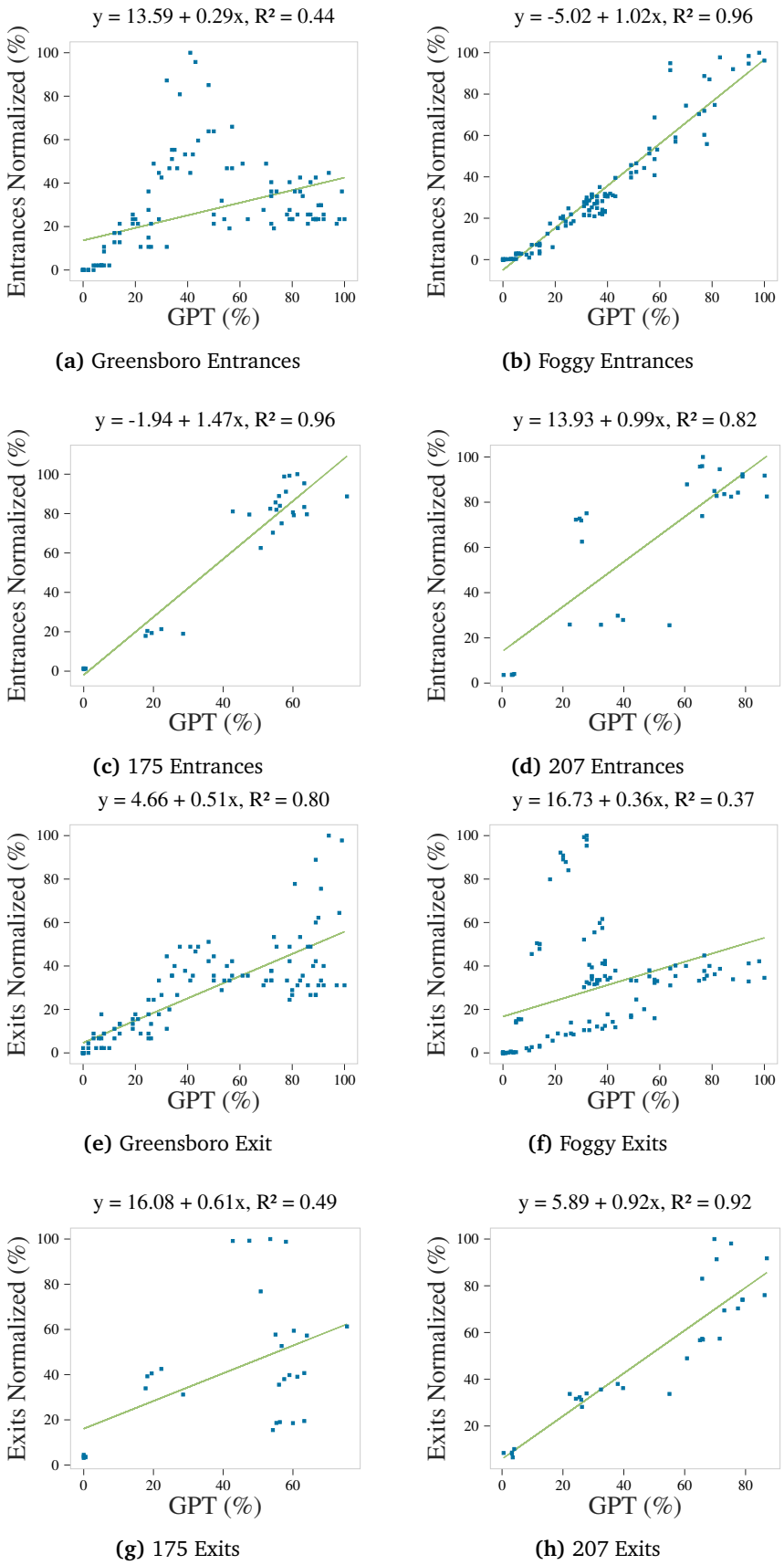
where $G_{h,s}$ is the GPT value for station s and hour of the week h, β represents the regression coefficient, ϵ is the residual error, and T is the transit data.

We tested the regression model for both transit information (entrances and exits), the sum, and the difference between the two.

The performance of the regression models are evaluated using the coefficient of determination, i.e., R^2 score, which is the proportion of variation explained by independent variables. We first start by analyzing the total results obtained in the two cities in our dataset. We applied the linear regression described in (4.1) to the standard GPT of all stations, together with the data from entrances, exits, the sum entrances+exits, and the difference entrances-exits. Fig. 4.3 shows the radial plots of the results of the regression of all stations. Specifically, it shows the average R^2 score between the GPT standard of all stations of the two cities. Looking at the R^2 it is clear that for

[1]Source: `http://web.mta.info/developers/turnstile.html`

[2]Source: `https://www.wmata.com/initiatives/ridership-portal/Rail-Data-Portal.cfm`



(a) Greensboro Entrances

(b) Foggy Entrances

(c) 175 Entrances

(d) 207 Entrances

(e) Greensboro Exit

(f) Foggy Exits

(g) 175 Exits

(h) 207 Exits

Figure 4.2: Correlation between GPT and transit data for 4 exemplifying station, 2 in New York and 2 in Washington

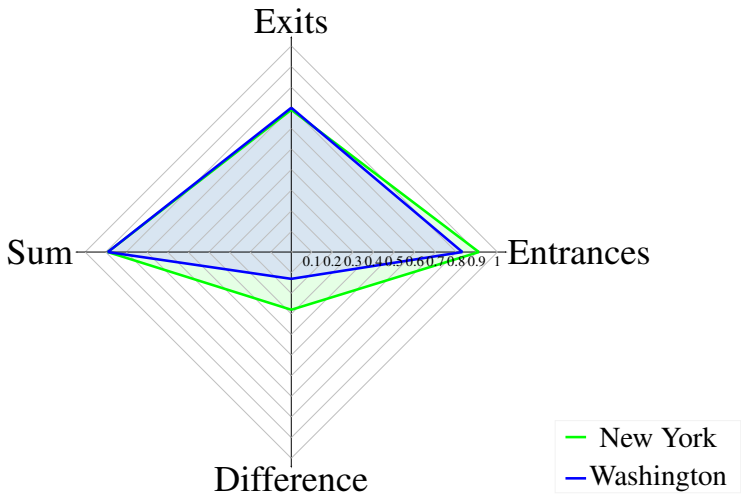


Figure 4.3: R^2 of all stations between GPT and Transit data

both cities the entrances have a better correlation (R^2 = 0.91 New York and R^2 = 0.81 Washington) than the exits (R^2 = 0.70 New York and R^2 = 0.71 Washington), the sum entrances+exits does not improve on the entrances result (R^2 = 0.89 New York and R^2 = 0.89 Washington), at the same time, the difference entrances-exits obtains the lowest scores of correlation (R^2 = 0.29 New York and R^2 = 0.16 Washington). This outcome could be explained by the fact that passengers entering a station have to wait for the subway to arrive, leaving a longer trace at the station as picked up by GPT, while the process of exiting a station is generally faster. This may explain why GPT information, which is related to presence of people in a station, is more correlated to the flow of travellers arriving at a station rather than leaving it. Despite the general trend suggesting that GPT is mainly driven by the entrances profiles, at the single station level we notice the existence of a minority of the stations where the relationship is the opposite and GPT is more correlated with the exit flows. Fig. 4.2 shows this important aspect of the GPT-Transit relationship, we selected 2 stations per city, the R^2 values and the regression lines reveal that certain stations have high correlation with entrances (fig.4.2b,4.2c) and low correlation with exits (fig.4.2b, 4.2c) and at the same time some stations reveal an opposite behavior; examples are stations Greensboro (Washington) and 207st (New York). For both stations the GPT is more correlated with exits (fig.4.2e,4.2h) than entrances (fig.4.2a,4.2d), but these remain a large minority of all analysed stations. This characteristic of peculiar similarity to the exits of some stations leads us to develop a specific profile for each station able to identify the interconnection between the GPT and the transit data for a generic week.

4.3 The TransitCrowd Estimation Framework

Fig. 4.4 shows the methodology behind the TransitCrowd framework developed by this study. The methodology aims at estimating the exit and entrance profiles of every subway station in a city for a specific week. The inputs are the standard GPT, the averaged transit data, and the live GPT. The framework is composed by two different estimation tools, the *Sig Estimator* and the *Reg Estimator*. Both tools estimate the flows of entrances and exits at subway stations, but with different characteristics. The Reg estimator is based on Machine Learning (Machine Learning (ML)) regression models, and it prioritizes the accuracy of the results, while it focuses only on a single city without allowing to transfer the methodology without a new training process. The Sig Estimator is based on simpler statistics methods, at the cost of a lower accuracy compared to the Reg estimator. The upside of Sig estimator is that thanks to the concept of the *signature*, it has the potential to be transferred to different cities without requiring a new training process involving transit data. In the following, we describe the details of the two estimation tools.

4.3.1 Reg Estimator

With the aim of estimating the entrances/exit flows from each subway station, we selected as input the corresponding standard GPT and the averaged entrances and exits to train the ML models. The whole estimation process is separated for entrances and exits.

A set of ML models were trained among the most widely and successfully used across literature dealing with regression problems [80]. The stratified k-fold cross-validation method has been implemented to validate the trained models. This method is commonly used to assess the performance of classification models performed, thanks to its capability of reducing any bias produced by the models. Moreover, each ML model has a set of hyperparameters that need to be tuned in order to improve its performance. This process, commonly known as "hyperparameter tuning", is carried out by implementing the random search method, which allows assessing the values of the hyperparameter with a larger impact on model performance.

Using R as performance parameter, we assessed that the best-trained model for our approach is the Extra trees regressor. It is a model of ensemble learning technique that aggregates the results of different de-correlated decision trees. Once the training and the choice of the model are done, we move to the real estimation step. In this phase, we replace the GPT standard used for training with GPT live of a specific week that we want to estimate.

$$\text{wMAPE} = \sum_{i=1}^{n} \frac{(\overline{y_i} - y_i)^2}{\sum_{j=1}^{n} y_i}, \tag{4.2}$$

where $\overline{y_i}$ are the estimated values, y_i the observed values, or ground truth, and n is the length of these two series.

4.3.2 Sig Estimator

The Sig estimator is composed by two interconnected phases: Signature extraction, and Live Estimation. The first phase is signature extraction, it aims at extracting the signature that characterizes the relationship between the GPT of a single station and corresponding entrances and exits profiles. We exploit the standard GPT and the averaged entrances and exits as inputs. First, we need to transform the entrances and the exits data from the transit dataset in order to replicate the GPT scale (0-100).

We apply to both entrances and exits a mix-man normalization scaling the dataset on the 0-1 interval and we then multiply by 100. The scaling procedure is the following:

$$t_{scaled} = \frac{t - \min(T)}{\max(T) - \min(T)} \cdot 100, \forall t \in T, \tag{4.3}$$

where T represents the exits or the entrances dataset for a single station, min and max are the corresponding minimum and maximum values, these two values are stored for each station and will be used in the live estimation phase. Once scaled the transit data, we compute the signature of the stations. The signature represents the scaling factor between the standard GPT and the scaled exits and entrances. For each station we compute two signatures, one for the entrances and one for the exits. The signature calculation is the following:

$$S_{en,s} = En_{scaled} - G_s, \tag{4.4}$$

$$S_{ex,s} = Ex_{scaled} - G_s, \tag{4.5}$$

where S is the signature for the station s corresponding to the transit data of entrances En or exits Ex.

In the second step, we try to estimate the real values of users exiting and entering the subway stations for a specific week by leveraging the corresponding GPT Live data. Specifically, we exploit as input the signatures $S_{ex,s}$ and $S_{en,s}$ extracted in the previous phase using past information and we combine them with the information of the current week from the Live GPT. The estimation function for the Exits profile of a week w is the following:

$$Ex_{w,s} = (S_{ex,s} + GL_{s,w}) \cdot (\max_{ex,s} - \min_{ex,s}) + \min_{ex,s}, \tag{4.6}$$

where max and min are the same stored from (4.3), $S_{ex,s}$ is the signature of exits for station s, and GL is the GPT Live data extracted for station s

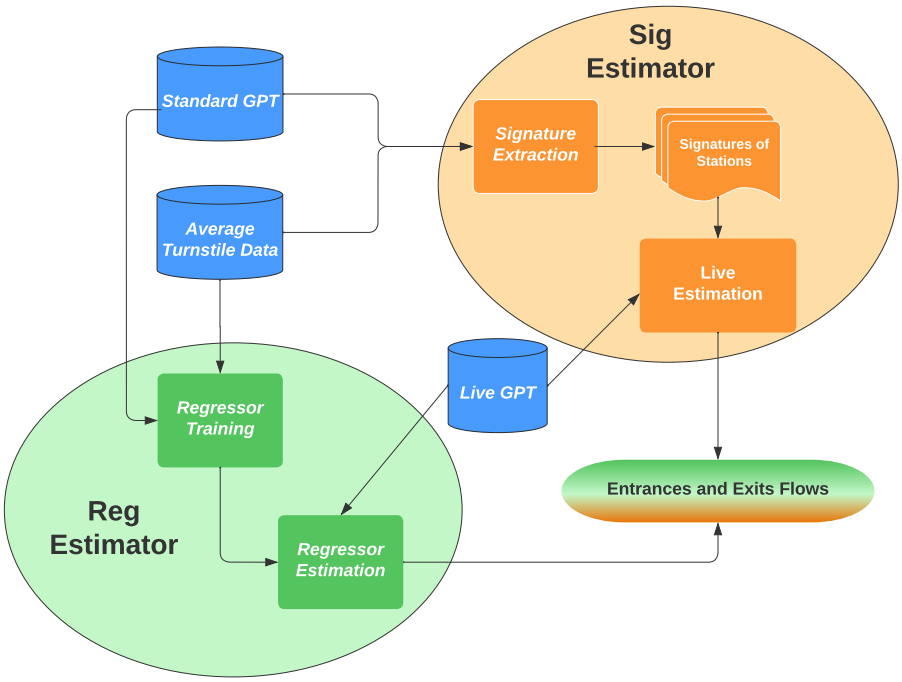


Figure 4.4: TransitCrowd framework, blue symbols represent input data, orange blocks are the Sig estimator, and green ones the Reg estimator

during week w. The same function applies also to the estimation of the entrances profile and it is repeated for every station in the dataset for 12 different weeks after the signature extraction. Similarly to the Reg estimator, the estimation error is computed using the weighted Mean Absolute Percentage Error (wMAPE) described in (4.2).

4.4 Performance Evaluation

We evaluate the performance of TransitCrowd calculating the estimation error at station level using the weighted Mean Absolute Percentage Error (wMAPE) [81]. We start analyzing the results provided by the Sig Estimator. As described in the previous section, the signature extraction is the first step of Sig estimator. Fig. 4.5 presents the signatures of entrances and exits for the subway station "50th" in New York, the first row of the plot reveals the three datasets exploited for the signature extraction: standard GPT, entrances, and exits (scaled 0-100). The second and the third row of the plot show the signature for the entrances and the one for the exits obtained by applying (4.4) and (4.5). It is interesting to note that the signatures for this station are almost always negative for the full period; above all it is clear

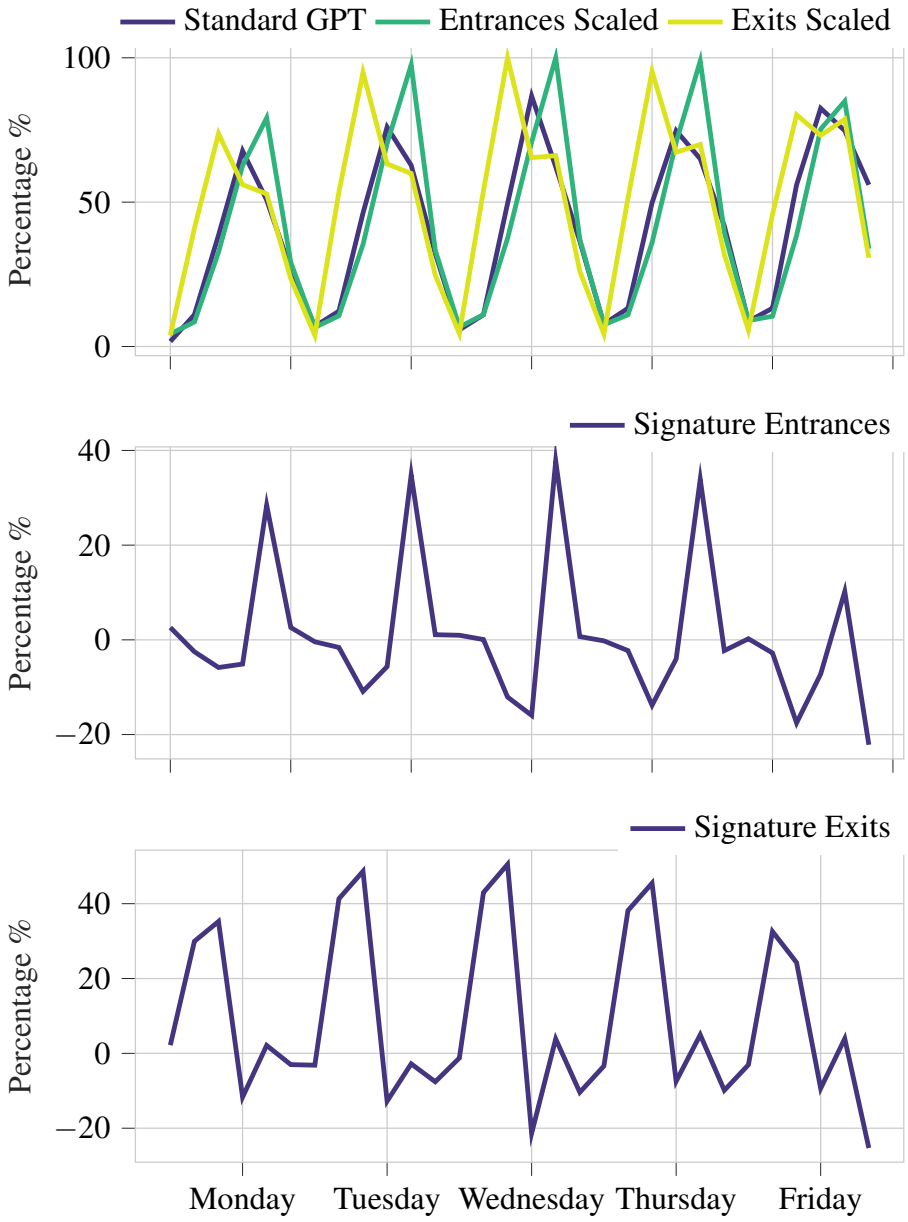


Figure 4.5: Extraction of signatures profiles for subway station named “50th”

that the biggest differences between the GPT and the transit arise during morning peaks. GPT seems not to display the same high percentages of exits and entrances during mornings, this information contained in the signatures will be crucial for the estimation process.

Once the signatures for all stations are extracted from the reference month, we are ready to leverage the GPT Live data for estimation of the

real flows of entrances and exits. Fig. 4.6 shows the result of the estimation process in a single station (Dupont Circle, Washington) for 1 week following the signature extraction month. The upper part of the figure reveals the profile of the GPT Live for the corresponding week, then the lower part presents the real estimation for entrances and exits produced by applying the matching signature. The figure depicts a good result for this single station, most of the peaks reached by the ground-truth are replicated by the estimated flows. It is interesting to notice that the estimation error for this station is stable throughout the week, this is a first signal that our prediction results are not deteriorating along different days.

We continue our analysis by looking at the results of Reg estimator. Fig. 4.7 presents the entrances estimation errors (wMAPE) of Reg for New York stations at different hours of the day. From the maps it is interesting to notice that the stations in the center of Manhattan are characterized by higher errors throughout the day. Moreover, Fig. 4.7c depicts how the errors in the evening are larger than in other day periods.

Table 4.1: Estimation error for all stations New York

Week after training	Error (wMAE)		Error (wMAE)	
	Sig	Reg	Sig	Reg
1	0.378	0.350	0.370	0.305
2	0.309	0.218	0.278	0.118
Validation set	0.309	0.218	0.278	0.118
1	0.306	0.236	0.278	0.150
2	0.308	0.263	0.271	0.178

Having illustrated the estimation results for single stations for Sig and Reg estimator, Fig. 4.8 shows the performances of our framework on entrances for every station in our dataset. The results are in form of a cumulative distribution function (CDF), every station contributes to the plot with a value of wMAPE that represents the estimation error made by the framework to estimate the entrance flow.

As expected, the Reg estimator produces lower errors; it is clear that for both plots the violet line representing Reg is always on the left of the Sig line. The Reg estimator obtains errors lower than 0.2 for the 60% of the estimations, while the errors of Sig estimator are less than 0.3 for the 60% of dataset in both cities.

The difference between the two cities is more evident in the interval $[0.6 - 1]$, here we can notice that New York CDF shows higher errors, both estimators reach values greater than 0.5 for a small portion of estimations (10%). The main outcome of the CDFs is that Reg estimator obtains better

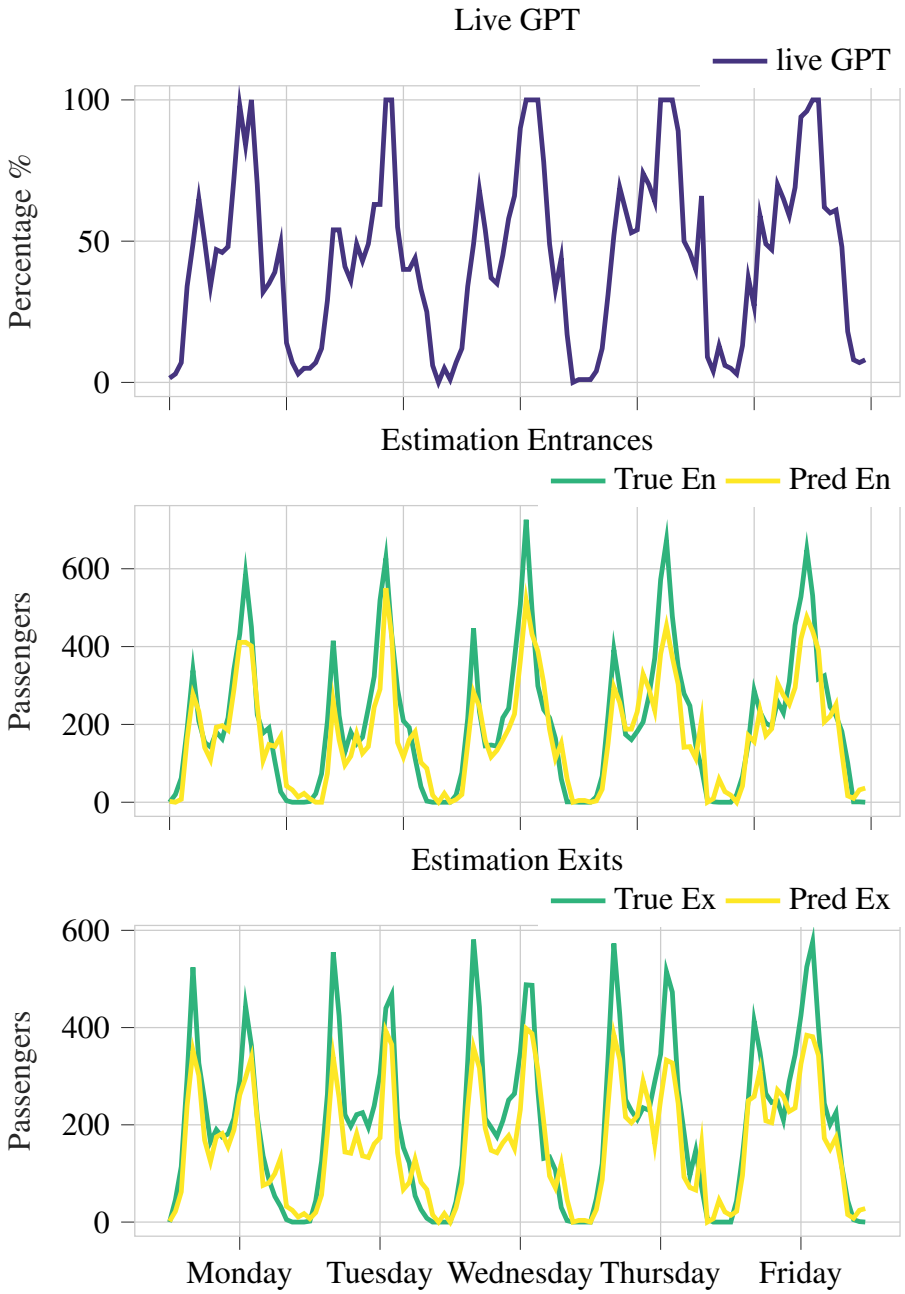
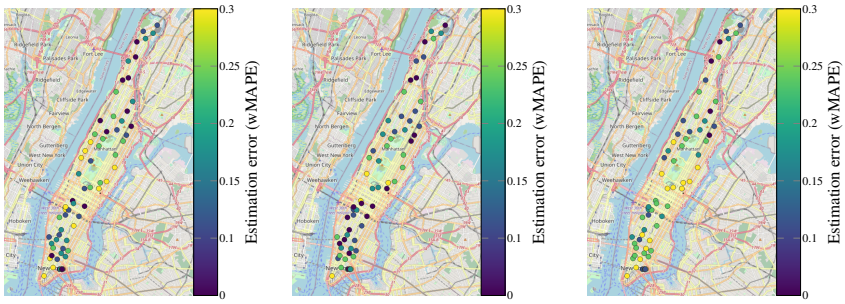


Figure 4.6: The profiles of the predicted and true values of turnstile data for week 1 after the signature extraction, for station Dupont Circle,Washington

estimation results than Sig tool. Therefore, the proposed idea of an estimator prioritizing accuracy (Reg tool) is confirmed.

Once analyzed the estimation performances on the full dataset we want



(a) H: 12:00 **(b)** H: 16:00 **(c)** H: 20:00

Figure 4.7: Estimation error for stations in New York at different hours of a working day

to analyze the evolution throughout the weeks, the scope is to recognize if our results are deteriorating along the weeks after the training suggesting that GPT trends tend to evolve in time.

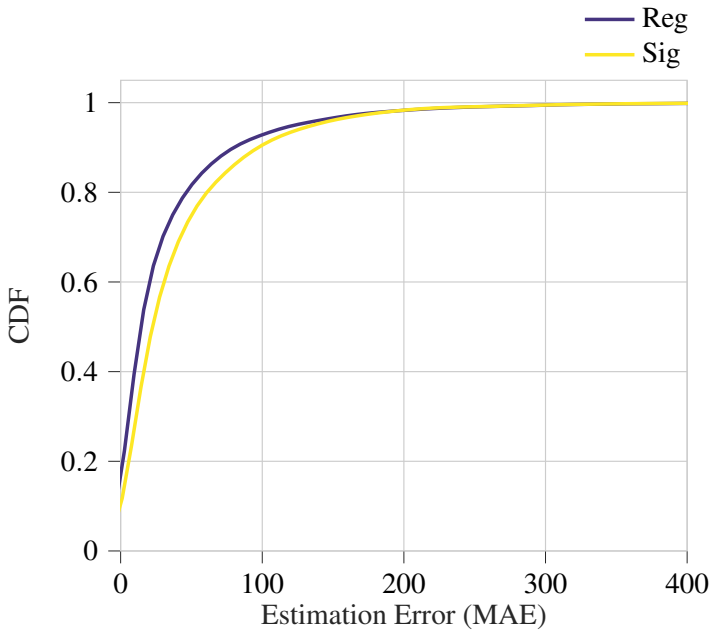
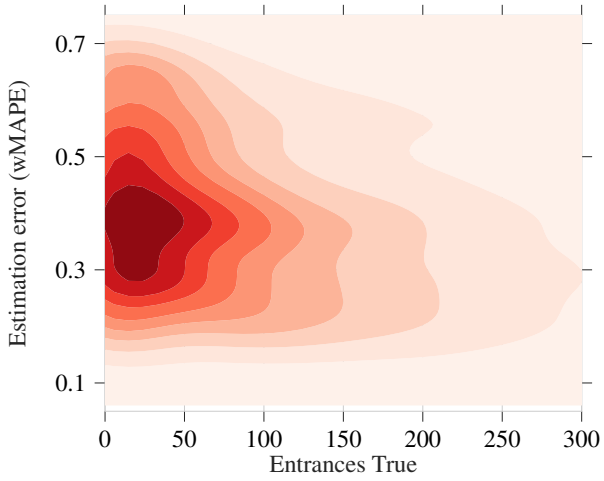
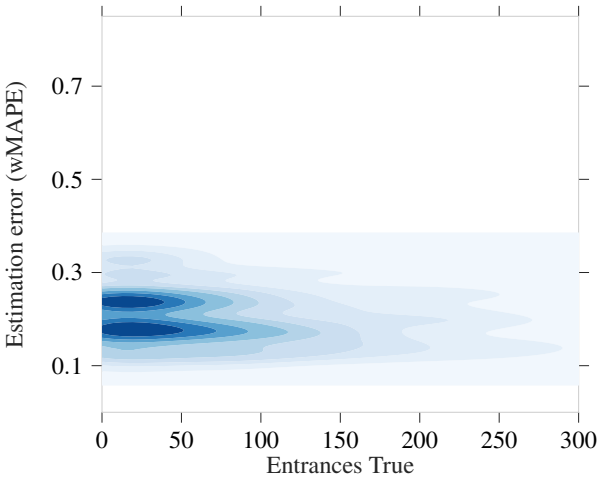


Figure 4.8: Cumulative error for all stations in Washington using both Sig and Reg estimator

Tab. 4.1 shows the performance of this estimation process on the New York dataset period for the two estimators, it includes the average wMAPE of the estimations in all stations for the entrances and the exits for all the weeks in the data collection interval. The table shows that the estimation



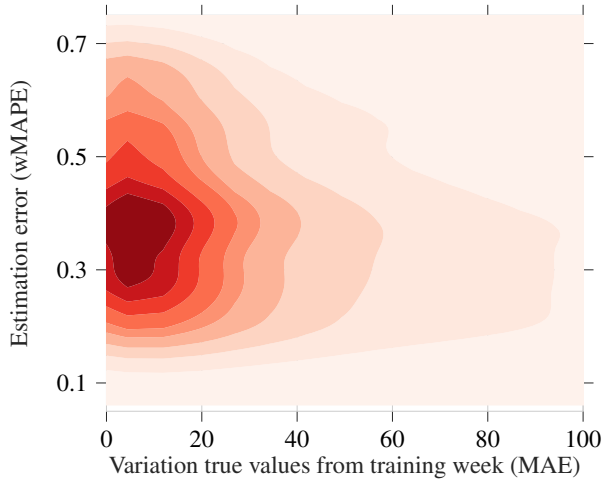
(a) Distribution estimation error Sig on test set



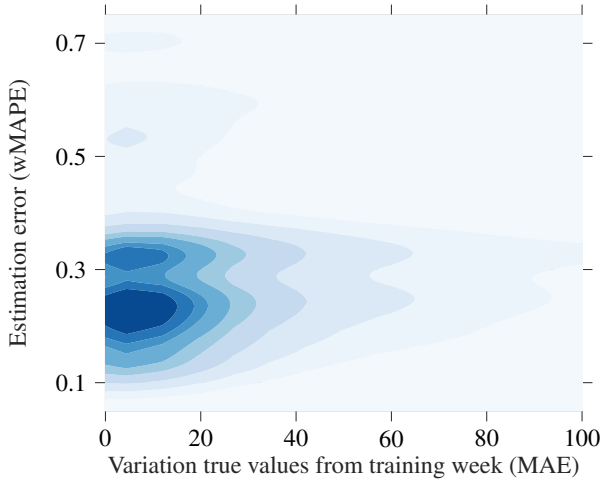
(b) Distribution estimation error Reg on test set

Figure 4.9: Distribution estimation error entrances

process is stable over the weeks, and the wMAPE is always contained in the interval [0.2 – 0.3]. it is notable that for both estimators the error does not appear to systematically increase along the different weeks, moreover the week with the lower errors is the 5th after training. This means that for Sig estimator the signatures extracted before week 1 are still valid also after the 2 months of the data collection, at the same time the Reg estimator does not require new training process after several weeks of estimations. Continuing with our analysis, We want to comprehend if the estimation errors of TransitCrowd are influenced by the amount of entrances/exits we are estimating. Fig. 4.9 presents two density plots in order to visualize the distribution of the



(a) Distribution estimation error Sig



(b) Distribution estimation error Reg

Figure 4.10: Distribution estimation error entrances

estimation errors over the values of entrances. The density plots are based on Washington results and concern both the Sig estimator(fig.4.10a) and the Reg estimator(fig.4.10b). Errors from Sig estimator are concentrated around wMAPE values of 0.35. In contrast, as for the previous results, the performances of Reg estimator are slightly better, and the errors focus on the interval [0.25 – 0.3]. The significant outcome of Fig. 4.9 is that for both estimators errors are not increasing with the rise of entrances values, it is remarkable that for entrances around 200 – 250 the errors remain the same that for entrances 0 – 50. Finally, Fig. 4.10 presents the relationship between the prediction error of entrances and the variation of true values from the

values of the reference week (i.e. the week of the signature extraction). The y-axis indicates the accuracy of the estimations (wMAPE between prediction and true values), while the x-axis depicts the deviation of the true values from the reference week (Mean absolute error between reference week and true values). These density plots reveal that the estimation error is not influenced by the variation of each week, this means that TransitCrowd is able to estimate with similar errors standard weeks and weeks different from the reference one.

4.5 Conclusion

In this chapter, we investigated the potential to leverage GPT to estimate public transport demand flows, specifically focusing on the subway. By exploring this crowdsourced data, we identified that GPT can be correlated with the entrance pattern of the majority of subway stations, while the crowdedness of a subset of stations is linked with the exits flows.

We developed TransitCrowd, a framework that exploits GPT to make live estimations of transit data at subways station level. Our framework is flexible and composed of two distinct estimator tools. The first, Reg estimator, prioritizes the accuracy of results focusing on a city level. The second, Sig estimator, extracts signatures from stations revealing the temporal profile of the correlation between GPT and entrances/exits. Through this fundamental information, it is possible to apply the presented methodology to other cities. Finally, we evaluated the performance of TransitCrowd, estimating two months of entrance/exit flows using as input the GPT Live data from each station.

The estimation process produced promising results whose accuracy appears to be stable over the different weeks. We observed that TransitCrowd is able to estimate properly weeks different from the training one, and the errors are not influenced by the high or low values of entrance/exit flows.

The next steps will focus on analyzing the signatures of different stations to identify influential factors, such as activities around the stations. Once such factors are detected, the final goal is to estimate signatures for stations in another city in order to test the transferability of our estimation process to a new environment.

Chapter 5

GPT of catchment areas to estimate transit flows

In this chapter, we address RQ4, "How to convert GPT data into transit information automatically?".

In this chapter, we investigate the feasibility of utilizing GPT data to measure the popularity of catchment areas surrounding transit stations, with the goal of proving that this information can be employed to estimate the transit demand for the station. This is an important problem because it can help transportation planners and researchers who need to analyze transit flows in cities where transit data is not available.

5.1 Introduction

In this study, we have the goal to leverage the popularity of places around stations as a determinant for estimating flows in and out of stations. . Specifically, we want to exploit the information on the catchment areas around stations to analyze the transferability of Transitcrowd tool to other cities, without having to retrain the model using transit data.

This is an important question because it has the potential to help transportation planners and researchers who need to analyze transit flows in cities where transit data is not available.

In order to apply Transitcrowd in a city without transit data we need to obtain the signatures of the stations, the concept of signature of station have been described in Chapter 4 and represent a time series that characterizes the relationship between the GPT of a single station and the corresponding entrances and exits profiles. To this end, we developed a ML framework that uses data on the activities around a station as a substitute for the training transit data. Specifically, we exploit GPT data of the activities in catchment areas around the station and exploit such data to estimate the signatures in a city without the transit information. The rest of the chapter is organized as follows: the following section presents the description of the data used in the study, followed by the methodology of the framework. Then, the results are discussed and evaluated. Finally, the last section provides the final remarks.

5.2 The Dataset

In this section, we present the types of data we used in our analysis for transferring TransitCrowd to a new city. Our framework was trained using data from New York City, which we refer to as the "training" city. The framework was then evaluated using data from Washington D.C., which we refer to as the "testing" city. For both cities, we collected the same data, which we divided into input data and target data.

Input data

The input data in our framework consists of information about the local businesses (LBs) providing GPT data around the stations. We collected this data for both the training and testing cities because it will be used during the training process and will be the primary source of information for the estimation process. We divided the inputs into two types of categories:
Static data, we first exploit the static information on how many LBs are located in the area surrounding each transit station. We extracted from Google maps the information regarding their location and their category of business. Based on the different types of LBs provided by Google we created

11 categories to divide all the different LBs types. The categories we defined in this study are: Financial, Public, Food, Transit, Stores, Attractions, Bar, Health, Gas station, and Pharmacy. Our study includes information on the type and location of $40,000 LBs$ in New York and $23,000$ LBs in Washington. **Dynamic data**, We used the dynamic information on the popularity trends of LBs in the areas around the stations. To do this, we used the popularity trends from the standard GPT, which describes the normalized weekly trend of visits at LBs based on the average of a few months. This information was collected in New York and Washington for a subset of the static dataset on LBs, as not all LBs in Google Maps provide GPT information. To divide the different types of GPT of LBS , we used the same macro categories described for the static data. Our dynamic dataset includes the GPTs for $16,000$ LBs in New York and $9,000$ in Washington.

Target data

The primary aim of this study is to assess the transferability of TransitCrowd. To do this, we have focused on the signatures of stations as our primary target. These signatures are essential for estimating the transit flows at each station using the Sig estimator method, which has been introdiced and tested in Chapter 4. The signatures represent the scaling factors between the GPT data for a station and the trends of exits and entrances of passenger flows. As a result, our targets will be divided into two types: signatures of entrances and signatures of exits, which will be analyzed independently of each other. We have employed the TransitCrowd tool to extract the signatures for both the training and testing cities. The signatures for the training city (New York) will be used to train our framework, while the real signatures extracted from transit data for the testing city (Washington) will be used to evaluate the accuracy of our predicted signatures.

5.3 Methodology

The proposed methodology involves two steps: *Data pre-processing* and *Signature estimation*. In the first step, we prepare the data from the area surrounding the station using GPT data for both the training and testing cities. The second step involves using machine learning techniques to generate accurate estimates of the signatures of stations in the testing city. This section provides the details of the two phases, which are also illustrated in Figure 5.1.

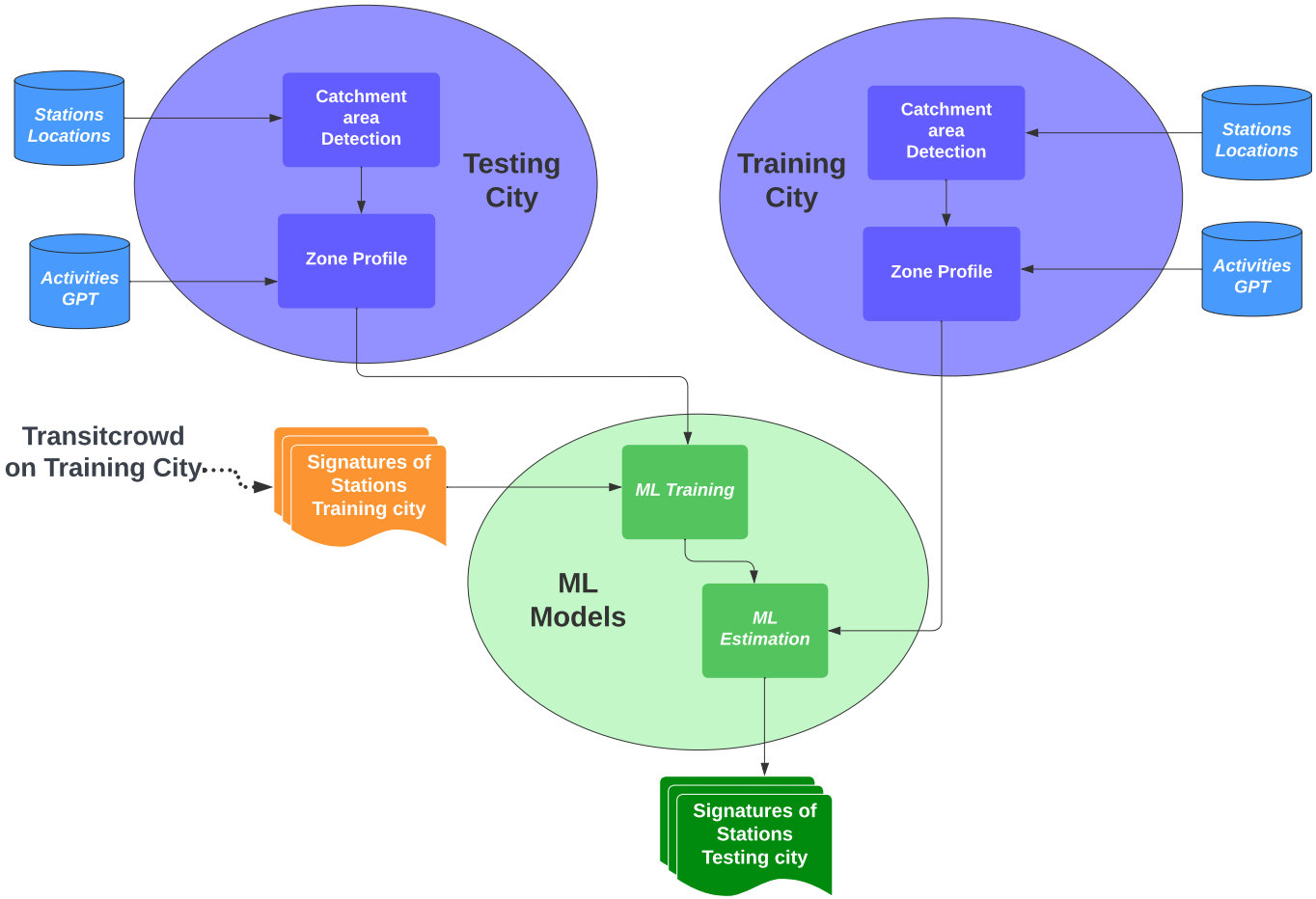


Figure 5.1: General framework of proposed methodology

5.3.1 Data pre-processing

In this phase, we focused on processing the data to create information that describes the characteristics of the surrounding area around the stations. This information will be crucial in the next step when we use it to estimate the signature of each station. In order to extract the zone profile we exploit GPT and the location of the stations.

Catchment area

To fully understand the demand profile of a station area, it is important to define its boundaries, which can be defined by the catchment area. The catchment area helps to clearly identify the geographical area that will be analyzed and ensure that all relevant activity is included in the profile. This area helps to identify the areas that are most closely tied to the station and the businesses that are most likely used in connection with it. By defining these catchment areas, we can accurately assess the characteristics of the station area for the purpose of signature estimation. In this study, we proposed two different methods to determine the station catchment areas: Voronoi and Weighted distance.

The first approach is based on **Voronoi** diagram [82], this method is a way of dividing a plane into regions based on the distances to a set of points, known as seeds. Using this method, each point within a region is closer to the seed of that region than any other seed. In the context of this study, the plane represents the city area, and the seeds represent the stations. Using the Voronoi diagram method, the city is divided into regions, with each region corresponding to a catchment area for a specific station. In this way, the catchment area of a station is defined as the region of the city that is closest to that station, as determined by the Voronoi diagram. This method is widely used in mobile network studies for visualizing and analyzing the coverage areas of base stations [83]. In the context of this study, the Voronoi diagram method was selected as our analysis has analogies with the study of coverage areas of base stations. Despite this approach well defines the catchment areas of a station, it does not consider the areas that are located between multiple stations as the approach partitions the area in sub-areas that do not overlap. However, these areas could have an influence on the behavior of different stations since the travellers performing activities in some places may be well accessing one or more stations. To address this, we developed a second approach to identify catchment areas that can split the overlapping zones between multiple stations. This method called **Weighted distance**, assigns a weight to each point in the area based on the distance of the point from each station. This way, the influence of each station on the overlapping area would be proportional to its distance from that area. This method measures the distance between two points by considering the actual street network,

rather than a straight-line distance. This can produce more accurate results, as it accounts for the geographical and network structures of the area. The weight $W_{i,s}$ assigned to a specific point i for station s is calculated using the following formula:

$$W_{i,s} = \begin{cases} \frac{d_{i,s}}{Max_w} & \text{if } d_{i,s} \leq Max_w \\ 0 & \text{if } d_{i,s} > Max_w \end{cases} \tag{5.1}$$

Where $d_{i,s}$ is the network distance of point i from station s and Max_w is the maximum walking distance. This equation calculates the weight of each point as the ratio of its distance from a particular station and the maximum walking distance if the distance is less than or equal to Max_w. If the distance is greater than Max_w, the weight is 0. This ensures that the weight is always between 0 and 1, is proportional to the distance from the station, and is 0 if the distance is larger than Max_w. In order to compare the effectiveness of both catchment identification methods, we used them in parallel, analyzing both the training and testing cities. By doing this, we were able to assess the performances of the two methods and determine which one was more accurate.

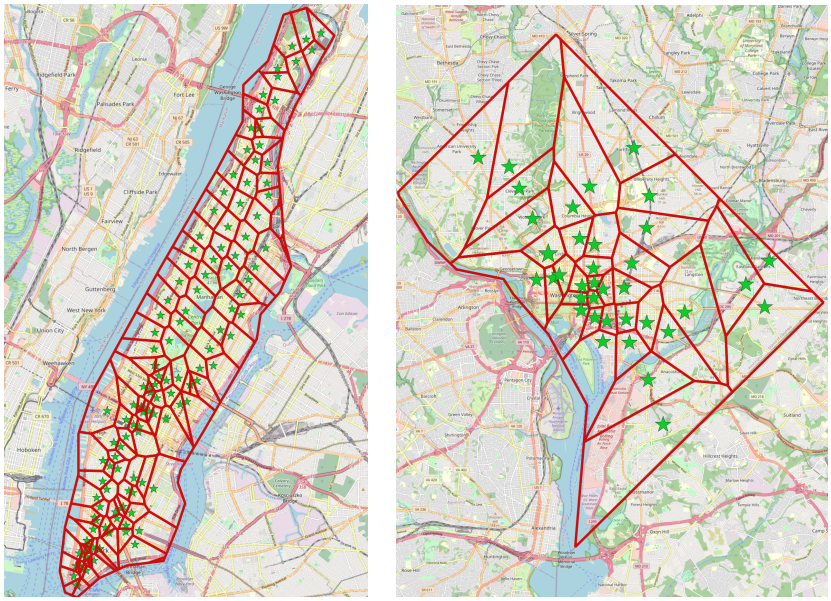
Zone Profile

Once the catchment areas around the station have been defined, we need to identify the characteristics of these areas. To this end, we exploit static and dynamic data of the area, following a similar approach to our previous work on Chapter. The dynamic data consists of the GPT of LBs, which give us an indication of how popular these businesses are at different times of the day. The static data includes the location and business category of each venue, which do not change significantly over time. We combine this static and dynamic data into a single "zone profile" for each station. The zone profile includes the distribution of business categories in the catchment area, as well as the popularity trends of those businesses over time. This information allows us to understand the types of businesses and services available in the catchment area, as well as how crowded they are. The zone profile of each station consists of a value of popularity for each hour of the week and it is repeated for each category of LBs, the categories of LBs are the ones described in Section 5.2. The zone profile of station s is represented by the following equation:

$$Z_{s,c} = \sum_{i=1}^{N_c} (G_i * W_i) \tag{5.2}$$

where N_c represents the number of local businesses in the catchment area for category c. The term $(G_i * W_i)$ represents the contribution of the LB i to the overall zone profile, with G_i being the GPT of the business for

a standard week (168 values) and W_i being the weight assigned to that business. Note that the weight depends on the strategy of the catchment area. For the Voronoi method, the weights are equal to 1 for all local businesses in the catchment area, while for the weighted distance method, the weight is described in Eq. 5.1.



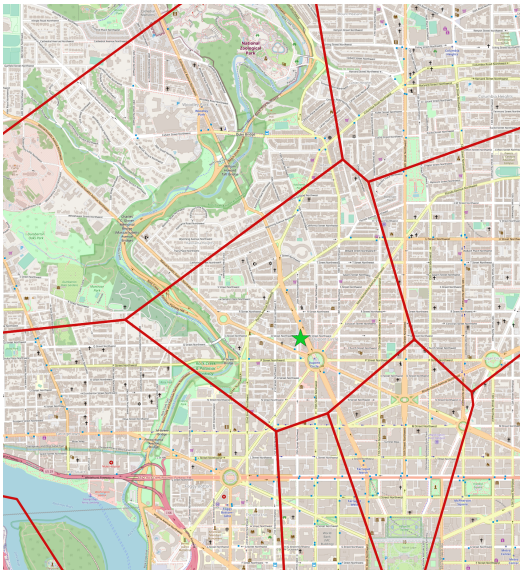
(a) New York **(b)** Washington DC

Figure 5.2: Voronoi catchment areas

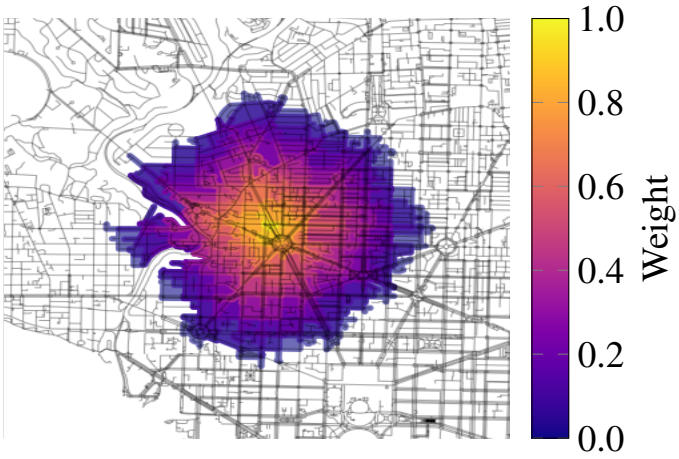
5.3.2 Signature Estimation

Once we developed a method for calculating the zone profile of a particular transit station, we used this information to estimate the signatures for the testing city, where we aimed to estimate the signature values without using traditional transit data. To achieve this, we employed machine learning regression algorithms on data from the training city. The algorithms were trained to predict the signature of the stations based on the input of the zone profiles. The signatures used as prediction targets in this phase were those generated by the Transitcrowd tool. It is worth noting that we conducted separate training for both the exits and entrances signatures, as these quantities may vary independently of one another.

The use of machine learning regression models allows us to analyze the relationship between the zone profile and the signature of the station, and to make predictions about the signature. By training the models on data from the training city, we were able to evaluate the accuracy and reliability



(a) Voronoi



(b) Weighted distance

Figure 5.3: Comparison between Voronoi and Weighted distance catchment area, for Dupont Circle Station, Washington DC

of these predictions and determine the efficacy of our method. Once trained and validated we applied the ML models to the testing city.

ML models

We trained 12 ML models to predict the signature of stations. These models were chosen because they belong to a category of models that have been widely and successfully applied to various regression problems in previous

research, indicating their suitability for use on datasets with similar sizes and compositions. This made them well-suited for our purposes. The description of the different models is shown in table5.1.

Model selection

In order to assess the performance of the different models, in this study, we used several performance metrics to evaluate the accuracy of our machine learning models for signature estimation. We also fine-tuned each model's parameters using a technique called random search, which helps identify the values that have the greatest effect on the model's performance [84]. The performance metrics exploited are commonly-used techniques for evaluating the performance of regression models. The metrics are mean absolute error (mean absolute error (MAE)), root mean squared error (root mean squared error (RMSE)), and coefficient of determination (coefficient of determination (R2)).

Mean absolute error (MAE) is a measure of the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as the average of the absolute differences between the predicted and actual values. It is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.3}$$

Root mean squared error (RMSE) is a measure of the average magnitude of the error, taking into account the direction of the error. It is calculated as the square root of the mean squared error (MSE). Can be calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.4}$$

Coefficient of determination (R2) is a measure of how well the predictions fit the actual data. It is calculated as the proportion of the variance in the dependent variable that is explained by the independent variables. It is formulated with the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5.5}$$

where $\overline{y}$ is the mean of the true values.

After analyzing the ML models using the aforementioned evaluation metrics on the training city, we selected the best model. Once we obtained the predicted signatures from the selected model for the testing city, we analyze the accuracy of these results.

Important Feature analysis

To understand the factors that are driving the predictions of our chosen model, we applied the SHapley Additive exPlanation (SHapley Additive exPlanation (SHAP)) method. This method allows us to identify the importance and impact of each feature in explaining the model's results [85]. By analyzing the SHAP values, we can see how each feature is contributing to the model's predictions, and whether certain features are having a larger or smaller impact on the final output. Using SHAP also enables us to understand the interactions between features, which can provide valuable insights into the behavior of the model. For example, we might find that certain combinations of features have a particularly strong influence on the model's predictions, or that certain features are only important in certain contexts. By analyzing these patterns, we can get a better sense of the overall behavior of the model and how it is making decisions.

5.3.3 Performance Evaluation and Discussion

As already mentioned in Section 5.2, we consider New York City as the training city and Washington D.C. as the testing city. The goal is to estimate the signature in the testing city without any dataset on transit. In order to identify the catchment area of each station in both cities, we employ the Voronoi and weighted distance methods described in Section 5.3.1. As illustrated in Figure 5.2, the Voronoi method was applied to the structures of both New York and Washington. The stations in the dataset are represented by the stars, and the red lines outline the catchment areas surrounding each station. This method ensures that every point within the catchment area is closer to the corresponding station than to any other station in the city. When comparing the catchment areas of New York and Washington, it becomes evident that the distinct structures and quantities of stations in each city have an impact on the size of the catchment areas. Specifically, the catchment areas in New York tend to be small and condensed, while the catchment areas in Washington tend to be larger in size. This difference in catchment area size can be attributed to the layout of the city and the distribution of stations. A city with a denser network of stations and a more compact structure may have smaller catchment areas, as each station serves a smaller geographic area. As outlined in Section 5.3.1, we also wanted to evaluate the performance of a second catchment area detection method, the weighted distance approach, which takes into account overlapping areas between stations. For every point in the street network surrounding a station, a specific weight is assigned according to Eq. 5.1. To determine the maximum distance parameter, we based our assumption on previous studies [86], which suggest that the maximum influence radius of a station is 1.3km. This distance is equivalent to a 15-minute walking distance assuming an average walking speed of 5

Table 5.1: ML models trained for signature estimation

Model	Description
AdaBoost Regressor	A boosting algorithm for regression that combines the predictions of multiple weak models, typically decision trees.
Bayesian Ridge	A regularized linear regression method that uses Bayesian techniques to find the optimal weights for the model.
Decision Tree Regressor	A decision tree model for regression that splits the data based on features that maximize the variance reduction.
Elastic Net	A linear regression model that combines the L1 and L2 norms of the weight vector as a regularization term in the cost function.
Extra Trees Regressor	An ensemble learning method for regression that uses multiple decision trees and outputs the mean prediction of the individual trees.
Gradient Boosting Regressor	An ensemble learning method for regression that produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
Huber Regressor	A regression method that is robust to outliers in the data by using the Huber loss function.
K Neighbors Regressor	A k-nearest neighbors regression model that makes predictions based on the mean of the k nearest neighbors.
Light Gradient Boosting Machine	A machine learning library for gradient boosting on decision trees. It is designed to be distributed and efficient with the following advantages: faster training speed and higher efficiency.
Linear Regression	A statistical method for modeling the linear relationship between a dependent variable and one or more independent variables.
Passive Aggressive Regressor	An online learning method for regression that updates the model's weights aggressively for misclassified instances and passively for correctly classified instances.
Random Forest Regressor	An ensemble learning method for regression that constructs a multitude of decision trees at training time and outputs the mean prediction of the individual trees.

km/h. In Figure 5.3, the catchment area of the Dupont Circle station in Washington is depicted using the weighted distance method. The colors in the map represent the weight assigned to each location, with lighter colors indicating locations that are closer to the station and have a higher weight, and darker colors indicating locations that are farther from the station. White zones in the map correspond to areas that are outside of the catchment area for the station and have a weight of 0.

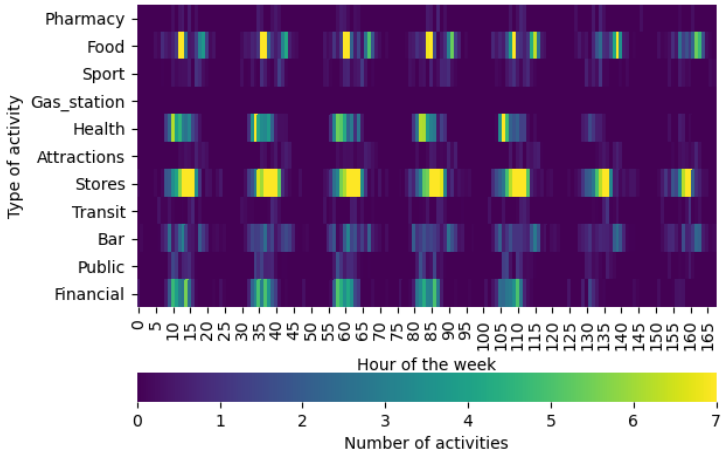


Figure 5.4: Example of zone profile for 42 St-Bryant Park Station

In the second phase, we focus on profiling the catchment areas that we just defined. The zone profile is represented by a heatmap where the x-axis indicates the hours of the week, while the y-axis expresses the different categories of business. Each cell in the heatmap depicts how many LBs are experiencing their popularity peak time along the day for the corresponding category and hour of the week as described in Eq. 5.2. Fig. 5.4 shows an example of a zone profile of the Voronoi catchment area for the subway station of 42 St-Bryant Park Station, New York. The chosen metro station is located close to Times Square and the surrounding area is characterized by a wide variety of activities. The zone profile of this station clearly shows that stores are the most popular category with a trend of uniformly distributed popularity throughout the weekdays. On weekends, however, popularity is more concentrated in the latter part of the day. This example demonstrates the characterization made using the zone profile of a specific station's catchment area. We extract the same information for all stations in our dataset for both the training and testing cities, using both the Voronoi and weighted distance methods. Once we obtained the catchment areas and the zone profiles, we can move to the final step of our work, which consists on signature estimation. The idea of this phase is to analyze if it is possible to predict the corresponding signature of a subway station given as input the

zone profile.

Table 5.2: ML model performance on signature estimation

Model	MAE	RMSE	R2
Extra Trees Regressor	5.4149	7.0969	0.7332
Light Gradient Boosting Machine	5.6416	7.3731	0.7114
Random Forest Regressor	5.7339	7.4837	0.7032
Gradient Boosting Regressor	5.7425	7.4384	0.7028
AdaBoost Regressor	6.2728	7.7249	0.6771
Linear Regression	6.0670	7.7864	0.6701
Bayesian Ridge	6.0632	7.7931	0.6699
Huber Regressor	6.0898	7.8856	0.6650
Passive Aggressive Regressor	8.2456	10.3111	0.4484
Decision Tree Regressor	7.5693	10.2300	0.4359
K Neighbors Regressor	8.9660	12.2478	0.2212
Elastic Net	9.7770	13.1909	0.1178

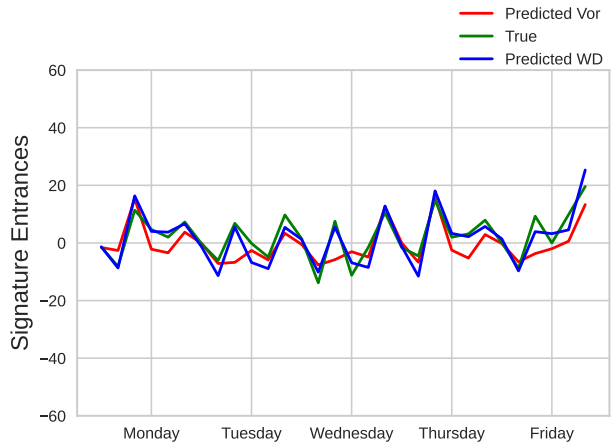
We initially trained the 12 ML models identified in Section 5.3.2 using data from the training city to determine the most suitable model for signature estimation. Table 5.2 shows the results of different ML models trained for signature estimation. The models are ranked according to their MAE, with the model with the lowest MAE being the best performer.

We can see that the Extra Trees Regressor, Light Gradient Boosting Machine, and Random Forest Regressor have the lowest MAE values, indicating that they are the top performers among the models considered in this study. These models are all learning methods that use multiple decision trees to make predictions, with the Extra Trees Regressor and Light Gradient Boosting Machine being more efficient and faster to train compared to the Random Forest Regressor. The RMSE and R2 values also follow a similar trend as the MAE values, with the top performers having the lowest RMSE and highest R2 values. Overall, we selected the Extra Trees Regressor as the best model due to its excellent performance based on the MAE, RMSE, and R2 values. It is worth noting that the results shown in Table 5.2 are based on zone profiles computed using the Voronoi method for signatures of entrances. We also trained the ML models using the weighted distance method and the signature of exits consistently founding the Extra Trees Regressor to be the best model. With the best model selected, we proceeded to test its performance on the stations in the testing city of Washington DC. We used the previously generated zone profiles as input and evaluated the accuracy of the model using both catchment area methods for entrance and exit signatures.

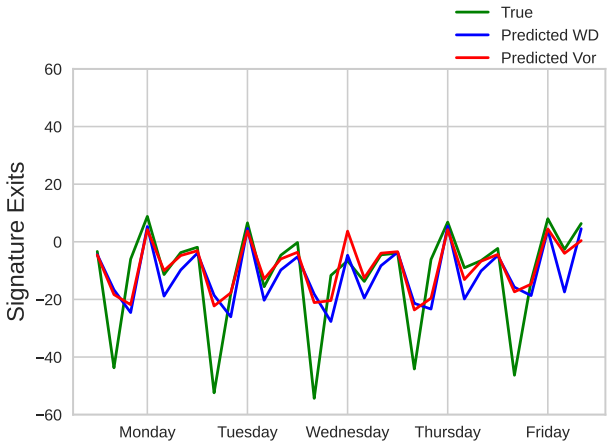
An example of signature estimation for a single station in Washington DC is shown in Figure 5.5. The green line represents the actual signature values,

while the blue and red lines show the estimates made using the weighted distance and Voronoi zone profiles, respectively. The figure illustrates that the weighted distance method (2.92 MAE) performs better than the Voronoi method (4.69 MAE) in terms of estimating signature entrances for this particular station. However, for signature exits, the error of estimation is higher for both catchment area methods, particularly for lower values of the signatures where the estimation fails to predict accurately. The MAE for signature exits is higher, with values of 7.65 for the weighted distance method and 10.09 for the Voronoi method. In Figure 5.5, we showed only the performnces on a single station. In Figure 5.6, we present the errors for all stations in the testing city (Washington DC) in the form of a cumulative distribution function (cumulative distribution function (CDF)) for the signature estimation of entrances (Figure 5.6a) and exits (Figure 5.6b). Each station is represented by a value of MAE that represents the error made by the framework in estimating the signature. As expected, the Weighted distance estimator produces lower errors than Voronoi. This is evident in both plots, where the blue line representing Weighted distance is always to the left of the Voronoi line, indicating lower errors. Specifically, for the signature of entrances, the errors obtained by Weighted distance are lower than 7 for 80% of stations, while for Voronoi they are lower than 9 for 80% of stations. The difference between the two catchment area methods is less pronounced in the signature of exits estimation. In this case, the CDF shows lower errors for the Weighted distance method, but the two lines are relatively close to each other. This suggests that both methods perform similarly for this estimation. It is worth noting that the estimation of signature exits generally produces higher errors. This is evident in the CDF, where errors are lower than 6 for 80% of stations for exits, while for entrances they are lower than 6 for 80% of stations. This suggests that it is more difficult for our model to accurately estimate the signature of exits compared to entrances. The main conclusion drawn from the CDFs is that it is more challenging for our model to accurately estimate the signature of exits compared to entrances. Additionally, the Weighted distance catchment areas method generally performs better than the Voronoi method. These findings suggest that there may be particular characteristics of the signature of exits that make them more difficult to accurately estimate. Finally, we performed an analysis of feature importance in the signature estimation process using SHAP values. The results are shown in Figure 5.7, which plots the SHAP values on the x-axis and the value of the feature on the colormap. Hour is the input feature with the highest impact, but it is surprising that static information like the number of stores, food, and transit has a greater impact than the dynamic data of GPT for the same categories. However, these three categories appear to be the most influential in signature estimation. The results of the feature importance analysis indicate that static data could be important for estimating signatures. This data could include details about

the city structure or the demographic information of the district surrounding the station. Incorporating this information may help to improve the accuracy of signature estimation.

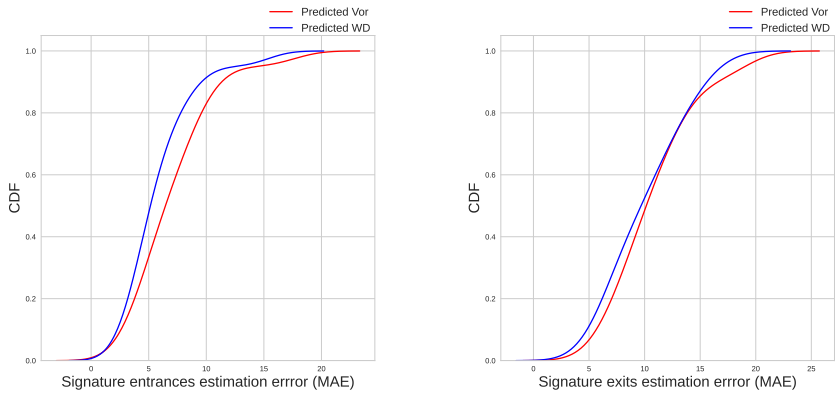


(a) Signature Entrances



(b) Signature Exits

Figure 5.5: Signature Estimation for station "Columbia Heights" in Washington DC



(a) Signature Entrances **(b)** Signature Exits

Figure 5.6: Cumulative error for all stations in Washington for signature entrances and exits

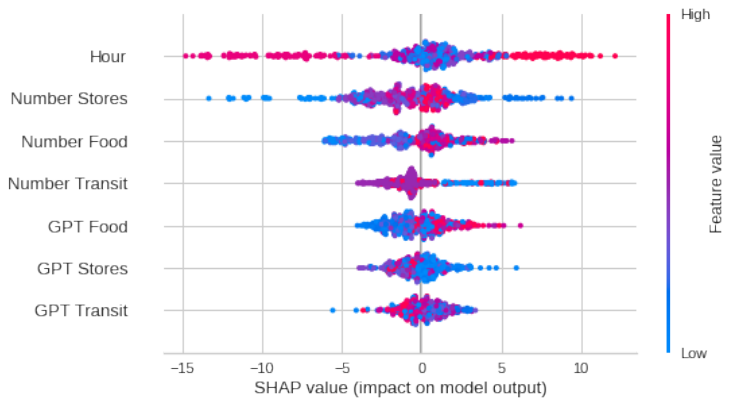


Figure 5.7: SHAP feature importance

5.4 Outlook

The purpose of this study was to investigate the feasibility of utilizing GPT data to measure the popularity of catchment areas surrounding transit stations, with the goal of proving that this information can be employed to estimate the transit demand for the station.

To achieve this, we develop a framework that estimates the signatures of a station. Signatures are the key concept that allows us to transfer the Transitcrowd estimation tool to a city without transit data. Therefore, we trained a ML model using signatures from one city together with the GPT data of the catchment areas around the stations. The results showed that

the model was effective at predicting the signature of transit stations in a different city.

The Extra Trees Regressor was found to be the most effective model for this task. We also experimented with various catchment area detection methods and found that using weighted distance provided the best results.

In future research, it would be interesting to test the performance of these models on a wider range of cities to see how well they generalize to other locations. This could help to confirm the robustness and applicability of the framework to different contexts. Additionally, incorporating more data sources, such as demographic and land use data, may further improve the accuracy of the predictions. This could provide a more comprehensive view of the factors that influence the signature of a transit station. In previous chapters and this one, we have examined how crowdsourced data, particularly using GPT, can overcome the limitations of traditional mobility data during normal circumstances. In the next chapter, we will continue to investigate the potential of crowdsourced data for mobility by analyzing the use of crowdsourced data in monitoring and responding to unexpected events in mobility.

Part III

Using crowdsourced data for anomalous events

Chapter 6

Mobility recover from Covid19

In this chapter, we address RQ5, ' Can Crowdsourced data be used to analyze mobility during anomalous events? "

We perform an analysis for multiple cities through crowdsourced information available from datasets such as Apple Maps, to shed light on the changes undergone during both the outbreak and the recovery of SARS-COVID-19 pandemic. Specifically, we exploit data characterizing many mobility modes like driving, walking, and transit. With the use of Gaussian Processes and clustering techniques, we uncover patterns of similarity between the major European cities. Further, we perform a prediction analysis that permits forecasting the trend of the recovery process and exposes the deviation of each city from the trend of the cluster.

This chapter is based on a work that has been published in the following journal paper:

- The Impact of SARS-COVID-19 Outbreak on European Cities Urban Mobility
  P. Vitello, C. Fiandrino, A. Capponi, P. Klopp, R.D. Connors, F. Viti
  Frontiers in Future Transportation

6.1 Introduction

The Severe Acute Respiratory Syndrome Coronavirus 2 (Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2)) [87] was declared as a global emergency by the World Health Organisation (WHO) as of January 30, 2020. The global outbreak of the pandemic uncovered the unpreparedness of the vast majority of healthcare systems [88] and led worldwide public institutions to take containing measures such as *social distancing, cancellation of public events*, and *closure* of businesses, education, and recreational activities. As a result, business and education systems moved to *remote* working and teaching, which stressed the limits of fixed and mobile networks [89]–[91].

The pandemic outbreak caused an unprecedented change to daily habits, including the way we move. Reducing and controlling human movement has been of the utmost importance to contain the pandemic spread and track infections. For example, by employing the DELPHI Epidemiological Model developed at M.I.T. [92] to Manila's metro transportation system, the study of [93] unveiled that the confinement measurements adopted by the authorities successfully prevented the rapid spread of infection.

In this chapter, we aim at drawing attention to two aspects. First, we aim to gain insight into how has mobility *changed* in urban environments during the first pandemic wave. Second, we study how such changes - driven by a mix of confinement policies and self-isolation measures - have impacted daily activities and, in turn, have contributed to limit the spread of the virus. Our objective is to perform an analysis encompassing several cities from different European countries and with different properties, to shed light on commonalities between the contrasting mobility reactions to the pandemic. These patterns can help cities to understand how they reacted to this first pandemic wave in terms of mobility and can be useful to detect similar cities helping to predict what will happen for future waves. The insights coming from this work are very important for the concerned stakeholders, e.g., government bodies, decision-makers, and city planners to re-think the existing urban landscape and drive more sustainable city planning. For instance, transportation authorities may monitor cities that reveal similar mobility patterns, and eventually apply policies that were demonstrated effective in those cities. For such a study, we rely on crowdsensed data that providers such as Apple make available. Specifically, we analyze the Apple Maps data that provide aggregated and anonymized information about the variation of popularity in using different transportation systems. Employing Gaussian Processes and clustering techniques, we combine the crowdsensed dataset with information about the number of daily infections. This approach allows identifying patterns of similarity between the cities considered and performing a prediction analysis to forecast the trend of the recovery process. In the remainder of the chapter, Section 6.3 illustrates the data employed in the analysis, which is described in Section 6.4. Finally, Section 6.5 concludes

the work and highlights the final remarks.

6.2 Related works

The studies that investigate the relation between Covid data and Mobility can be divided into three main categories. The first category includes the works analyzing the impacts of mobility on Covid trends, the authors of [94] investigate the importance of governmental policies and human mobility in the mitigation of the virus spread, their study draws attention to the correlation between the variations of mobility and the pandemic burden (measured in terms of deaths). The second category includes studies that given the mobility data try to forecast the pandemic evolution, this subject has been approached through different methodologies. In [95] the authors exploited graph neural networks techniques, while for [96] has been developed a partial differential equation model. Finally, the last group of studies deals with the influence of the pandemic severity on mobility, the authors of [97] created an impact analysis platform able to compute the effects of SARS-COVID-19 metrics on human mobility and social distancing.

Our analysis falls into the third category, whereas most of these works focus on determining the factors that influenced mobility, we decided to examine the similarities and the differences between citizens mobility for distinct urban areas. To perform our analysis, we exploit a crowdsourced dataset. Mobile CrowdSensing (MCS) has become a popular paradigm to perform sensing campaigns using sensors embedded in mobile devices like smartphones [17]. To combat the epidemic, many applications have been developed to monitor and establish contact tracing systems [98]–[100]. Corona-Warn-App, Immuni, and Radar COVID are examples respectively adopted by Germany, Italy, and Spain, and subscribers of the latter helped identify that loss of smell and taste could indicate the presence of the infection [101]. This approach falls in the category of participatory MCS that requires some efforts from the participants' side. With these applications, the users have to manually register and possibly declare themselves infected. Then the system takes care of controlling whether the level of exposure is high with the risk of contacting infected people. At the other extreme of the MCS landscape is the opportunistic paradigm: here, participants make no effort and the application takes care of sharing relevant information from the mobile device to the system. The crowdsourced dataset exploited for this study belongs to the opportunistic paradigm, many recent works used a similar dataset for mobility analysis. In [102], the authors combined GPS data and SARS-COVID-19 case data to understand how pandemic and restrictions affected the citizens' mobility in the USA. In [103], the authors exploited crowdsourced data from google to analyze the different impacts of the pandemic in 88 countries. Recent studies exploited the popularity of

Point of interest(Point of interest (POI)s) to quantify the mobility patterns of a city, the information on Pois can be extracted from different sources, the authors of [104] used data from Google popular times, while in [105] the dataset of Pois is taken from SafeGraph Places data. While these studies analyzed the general mobility of citizens our approach aims at investigating more in deep the different modes of transport. Other studies focused on the mobility of a specific country, the authors of [106] investigated how mobility in France changed before and during lockdown using mobile phone data, while in [107] the authors analyzed the reactions of citizens under mild policies in Sweden. Another important characteristic of our work is the focus on the European situation, in the closest to our work[108] the authors perform a socio-demographic analysis nationwide in Europe. By contrast, we work at a resolution of single cities.

6.3 Dataset

This Section explains the dataset we employ for the analysis. Specifically, we highlight the cities for which we obtain real data from different sources, i.e., Apple Maps [1] and Joint Research Centre (JRC)[2]. Besides illustrating the types of data considered for mobility and SARS-COVID-19 cases, we also delve into analyzing the morphology of the cities, population, and other metrics on the urban fabric. In such a way, the reader is provided with all the details necessary to understand the analysis of Section 6.4.

6.3.1 The Apple Maps Data

Mobile users have at their disposal several ways to share data such as location-based social networks (location-based social networks (LBSN)) (e.g., Facebook, Foursquare, and Twitter), and crowdsourced applications (e.g., OpenStreetMap, Waze) [17]. Such contributions have made available large datasets that enable an analysis of citizens' mobility, travel behaviors, and accessibility of urban areas.

Apple Maps data provides information on transportation modalities worldwide with zero privacy leakage, i.e., data is anonymized, and no information about the single users is disclosed. This is in line with what other popular crowdsensed providers like Google do (e.g., with Google Popular Times [109]). Rather, the data comes in a way that shows the aggregate requests for directions in Apple Maps for a given transportation mode, e.g., driving, walking, or site, e.g., transit, stations. Further, the data is provided as a relative increase or decrease with respect to the average past request, i.e. following pre-SARS-COVID-19 outbreak values, starting from January

[1] https://www.apple.com/covid19/mobility
[2] https://covid-statistics.jrc.ec.europa.eu/

13th, 2020. Our study analyzes the Apple dataset from February 23rd 2020, when the first lockdown measures were applied in Europe.

6.3.2 SARS-COVID-19 Cases

The JRC collects the numbers of contagious individuals and deaths at sub-national levels (admin level 1) for all the European countries. The data are imported directly from the National Authorities sources (National monitoring websites, when available). Since our analysis is at the city level, we considered the trend of the corresponding region. We extracted the evolution of the cumulative number of contagions normalized by the total number of contagions for each area.

6.3.3 The Considered Cities

After having described the type of data that will be employed for the analysis, we now describe the cities that have been selected. We began to collect data for Milan first, being one of the earliest cities hit by the SARS-COVID-19 outbreak and, for comparison, Luxembourg City that during the same period was not in the same situation. We started to pay attention to the possible dynamics of the virus diffusion, and this led to the monitoring of Valencia, where during 10/03/2020 a Champions League football match took place with an Italian team from Bergamo, Lombardia (shortly reported as one of the worst-hit areas in Italy). The study then was extended to consider multiple cities within Europe.

Table 6.1 shows the list of considered cities. For each of them, we include the population (as of 2018 from Eurostat database[3]), its morphological properties, and whether Apple data have also been recorded. Concerning the morphological properties, we take into consideration properties that define the urban network. Specifically, we resort to OpenStreetMaps (OpenStreetMaps (OSM)), which defines the street network with a graph $G_{\text{OSM}} = (V, E)$, where V is the set of vertices or nodes and E the set of edges. Each node is characterized by a unique identifier called OSMID, the latitude (y), the longitude (x). Further, each edge comes with a set of attributes: access, bridge, highway, lanes, maximum speed, name, oneway, osmid, service, tunnel, width, and the OSMIDs of the adjacent nodes of an edge.

Given that OSM is based on voluntary contributions, different cities might have a different precision level. For a fair comparison, we provide in the table the information given by the Augment-OSM Precision algorithm (AOP) [77]. Specifically, AOP augments the graph that OSM provides by adding through additional interpolation edges so that the resulting street graph contains nodes with a fixed distance, e.g., 1 m. A high density of nodes defines cities

[3] https://ec.europa.eu/eurostat/web/cities/data/database

Table 6.1: Comparison of population, number of edges, average initial edge length of each edge, and nodes for different cities

CITY	POPULATION	NODES G	EDGES G	AVG_LEN EDGES (M)
London	9,126,366	127,005	298,959	97
Berlin	3,748,148	28,073	73,187	146
Madrid	3,223,334	30,632	61,588	99
Rome	2,844,750	42,864	89,709	125
Paris	2,140,526	10,025	19,535	96
Bucharest	2,106,144	16,536	40,343	100
Vienna	1,911,191	16,083	36,105	126
Hamburg	1,899,160	21,490	51,949	145
Warsaw	1,790,658	18,823	43,370	137
Budapest	1,768,073	23,460	61,959	128
Milan	1,404,239	13,351	26,468	97
Prague	1,324,277	20,856	48,449	119
Stockholm	974,073	12,752	29,678	114
Amsterdam	873,555	11,520	26,580	98
Marseille	862,211	13,206	27,575	98
Copenhagen	794,128	6,990	17,649	102
Valencia	791,413	7,899	14,635	95
Krakow	779,115	8,925	19,859	152
Athens	664,046	8,822	18,302	64
Rotterdam	651,870	11,238	25,377	95
Helsinki	648,042	9,618	20,870	122
Glasgow	610,271	15,957	37,617	97
Dublin	554,554	11,193	26,030	95
Antwerp	525,935	7,990	17,913	120
Lisbon	506,654	9,769	20,571	91
Malmö	344,166	5,391	12,878	118
Graz	294,630	16,083	36,105	126
Brussels	176,545	1,437	2,719	87
Luxembourg City	122,326	2,146	5,000	126

with an extensive mobility network, e.g., streets and squares. Further to this information, we also include the number of edges graph and the average edge length in the city. In this case, we prune the street graph so that only two nodes define a street. As a result, we get knowledge about the degree of connectivity and regularity of the urban fabric.

6.4 Analysis

This Section presents the analysis of the dataset illustrated in Section 6.3. Specifically, in Subsection 6.4.1, we show for Luxembourg City the trends of infected individuals and fatalities, in relation to the lockdown measurements taken by the country, and the impact of such measures on the cities' mobility patterns (driving, walking, transit). In Subsection 6.4.2, we show

the methodology employed to verify the similarity trends observed in the mobility categories, group together those cities with similar patterns, and derive a prediction method to forecasts future mobility trends per category. In Subsection 6.4.3, we show the results for clustering and forecasting. Finally, in Subsection 6.4.5, we analyze the correlation between mobility and the trend of SARS-COVID-19 infected cases. Practically, we verify whether the mobility-based clusters of cities identified in Subsection 6.4.2 are still applicable when looking at the evolution of the number of SARS-COVID-19 infected cases.

6.4.1 A Primer

Fig. 6.1 shows in a comprehensive graph for Luxembourg City, from bottom to top, the evolution over time of infected cases and fatalities, the lockdown measurements taken by the government, the percentage increase or decrease of mobility modes usage, and the presence in activities. The time evolution spans from February 23rd to July 3rd, and during such a time frame we have data for both Apple Maps and SARS-COVID-19 trends.

The figure shows how the government started imposing hard lockdown measures as from March 15th, aligning with what other EU countries did despite a relatively low number of identified cases. However, in terms of mobility, it is possible to notice that the population started to reduce moving earlier, indicating an anticipation of actions following the announced restrictive measures or self-adaptation of the population to the emergency conditions. For example, were driving activities reduced already in a notable way from March 10th to March 11th while the other categories from March 11th to March 12th. The rate of decrease is comparable for all categories, revealing a pattern for all of them, reaching the lowest rate after a week from the start of the lockdown. Then during the lockdown the rate slowly increased, further increasing after the lockdown policy was stopped on May 11th. Transit experienced a substantial decrease, it started from 120% of users before the Lockdown, and after March 15th, it reached the lowest values (less than 20%) compared to Driving and Walking. Interestingly, during the first phase of the lockdown (March 15th - April 20th) Walking category experienced the highest values. This was due to good weather, which prompted people to take the opportunity to enjoy being outdoors while this was allowed. After lockdown, both walking and driving were observed to restore to normal conditions and even exceeded the expected rates, whereas transit did not. From this data there is clear evidence of a reduction in Transit ridership and an increase for driving and walking mode, the reasons behind this change could be different. A possible explanation could be a potential mode shift from public transport to walking and driving as using public transport was perceived as a more risky alternative in terms of potential contacts with infected people. On the other hand, another explanation for

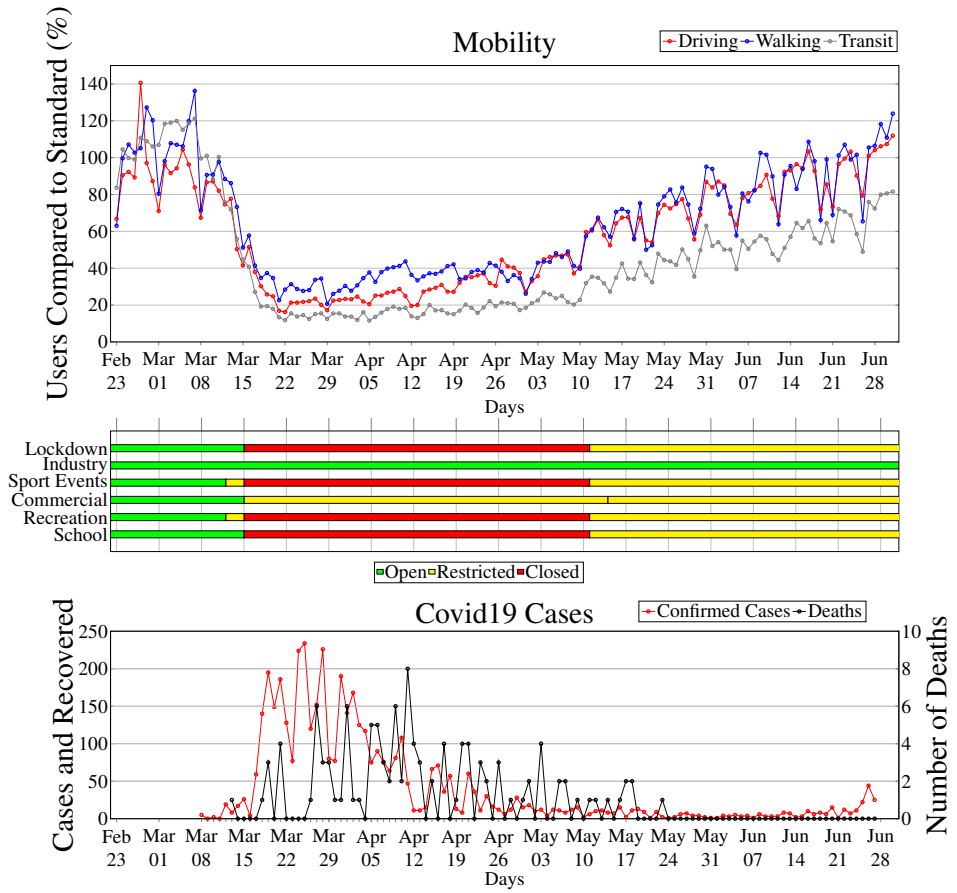


Figure 6.1: Comprehensive timeline with SARS-COVID-19 cases, lockdown measures, impact on mobility, and cities' activities

transit reduction can be found on the working from home policy that caused a drastic decrease in transit commuters. The case of Luxembourg City exemplifies the evident correlations between the three collected data sets, namely the mobility patterns via the Apple data, the lockdown policies, and the COVID-19 data. Similar correlations were observed in the collected cities. The aim of this study is to identify commonalities in how mobility trends changed across Europe as a consequence of the epidemic spread and the imposed restrictive measures. Capturing these commonalities may help in understanding how specific reductions in mobility have contributed to limit the spread of the virus, to better predict the evolution of future waves and suggest which policies may be more indicated.

6.4.2 City-level Analysis: Clustering and Forecasting Methodology

The objective of this subsection is to identify similarity trends observed in the mobility categories of Apple data for the cities considered (see Table 6.1). For this, we resort to clustering techniques. The proposed methodology consists of three interconnected components:

- regression with Gaussian processes;
- clustering with unsupervised machine learning models;
- prediction with again Gaussian processes.

We start from the raw Apple dataset at a city level, exploiting the full dataset (February 23rd - July 3rd). In this phase (regression), we want to obtain a mean function for each city that characterizes each category well (namely, Driving, Walking, Transit). The scope of this function is to find the general trend of the original data, avoiding outliers and peaks that could influence the clustering process. We noted that the interpolation could be affected by outliers due to data changes occurring in the presence of specific events (e.g., the Catholic and Orthodox Easter days). To obtain the mean function, we employ the Gaussian Processes (Gaussian Processes (GPs)) models that are one of the most commonly employed stochastic processes for application to datasets with data evolving over space and time (time series are a good example). When selecting the methodology to use, we explored both GPs and neural networks (neural networks (NNs)) like Multi-layer Perceptron and General Regression NNs. Unlike GPs, neural networks appear to be more suitable for larger and more complex datasets than the one at our disposal. Furthermore, GPs can be optimized exactly, i.e., there is no need for complex training procedures to tune the hyper-parameters. The main characteristic of GPs is that they are entirely determined by the mean and the covariance. This aspect helps the model fitting because only the first- and second-order moments should be specified. The covariance of the GPs is described by a Kernel (covariance function), in this work we use a kernel based on the combination by addition of a Matèrn component, an amplitude factor, and observation noise. The hyperparameters of the GP model are optimized by the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm (LM-BFGS) [110]. To prevent the possibility of finding a local maximum in the marginal likelihood, we run the optimization algorithm three times, using randomly-chosen starting coordinates. Once we obtained the mean functions, they are used to represent the city behavior for a specific category.

In the second phase, we first determine for each city a reference day that represents the *arrival in town* of the SARS-COVID-19 pandemic. Since the virus was observed to start spreading at different time periods in Europe, and in order to align the data seeking for comparability, we defined a reference

point as the moment in which the city (or the region containing it) reached 1% of total SARS-COVID-19 cases. Next, starting from these reference dates, we create windows of time with a fixed duration given in the number of days (e.g., 80 days). These windows are common in all cities. Once all the time windows are defined, we extract the corresponding intervals of the mean function obtained from the Apple Maps dataset for each city.

For the clustering technique, we use a hierarchical approach with a well-known distance metric:

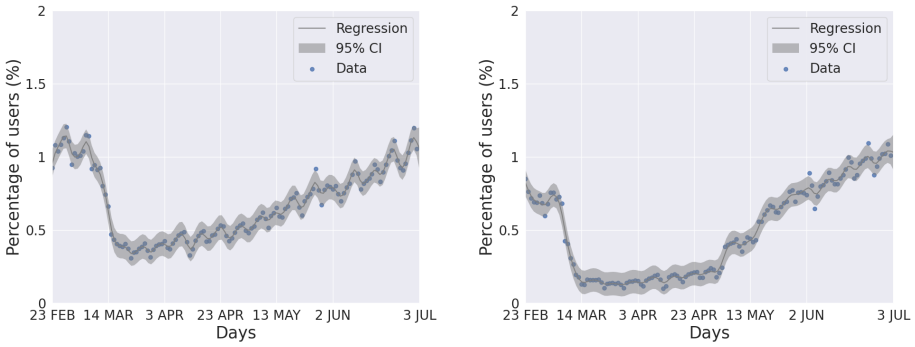
$$\text{Distance Metric} = JSD(M_{City1}, M_{City2}) \tag{6.1}$$

where M is the mean function from apple dataset and JSD is the Jensen-Shannon divergence function that measures the similarity between two distributions. We choose the JSD because it outperforms the asymmetric Kullback-Leibler divergence (KLD) [78] and it always returns a finite value. We preferred the hierarchical approach to other clustering techniques such as K-means or Dbscan, because its output is relatively easier to understand. The hierarchical algorithm produces dendrograms, which represent the similarities and the distances between the different clusters and at the same time highlight the distances between the objects in the same clusters. Such a hierarchical approach creates clusters based on both information on the mobility and evolution of SARS-COVID-19 cases. The distance between two clusters is defined according to the complete linkage or farthest neighbor method. The proximity between two clusters is the proximity between their two most distant objects.

In the third step, we re-apply the same GP model applied in the first step, but this time at the cluster level and for prediction. Indeed, GP can be employed not only for regression but also for prediction, and we are now interested in this feature. Specifically, we removed from the cluster dataset the values of the last 10 days and use them as ground truth to evaluate the prediction results. For the evaluation results, we compute the absolute error for each day within the prediction interval and then average this value with the ones obtained for the whole period. In such a way, we determine the average prediction error.

The reason why we look at predictions for the cluster is in the application of the method for early detection of future waves. Some city in the cluster could be a few days earlier in the virus spread than others, hence our analysis is of help to predict what will happen if the city in the cluster keep the same policy in terms of confinement policy or deviate and use that of another cluster if this may result to be more effective.

Fig. 6.2 shows an example of the application of the above procedure obtained for Amsterdam and Milan, to identify the trend from the data of the driving category. Note that Italy's more rigid restrictions rules reduced the variability of Milan's weekly patterns significantly. By contrast, in Amsterdam



(a) Amsterdam **(b)** Milan

Figure 6.2: Application of GP on two different cities for driving category

the recovery started earlier and the weekly patterns exhibit an increase in variability magnitude that becomes higher with time since March 15th.

6.4.3 Results

This Section analyzes the results obtained with the methodology explained above for both clustering and forecasting.

Clustering

We first start by analyzing the results obtained by the clustering operation for each of the three categories, i.e., Driving, Walking, and Transit. For each city, we extracted data concerning 20, 40, 60, and 80 days since when the 1% of total SARS-COVID-19 cases were reported. We applied the clustering approach to these different intervals to investigate the mobility evolution along the time. Fig. 6.3, 6.4, and 6.5 plot respectively the transition from the cluster obtained at an interval and the next one, i.e., Fig. 6.3 shows the difference between the clusters obtained using the first 20 days, and the period between 20 and 40 days, respectively. Clusters are rendered in the form of dendrograms that are a natural way of showing hierarchies and exposing similarities. In Fig. 6.3, 6.4, and 6.5, the dashed lines highlight the clusters while the red lines between the dendrograms indicate a change of cluster. For space reasons, we only include the plots obtained for Driving.

We resort to using only six clusters that better balance the number of cities per cluster for all the results. The x-axis represents the distance between each cluster. Note that the similarity between the two dendrograms is very high, and the only cities changing of cluster are Bucharest and Graz. Bucharest, originally a one-city-only cluster, becomes part of the blue cluster (with Paris, Luxembourg, etc.) while Graz does the opposite and transits from the purple

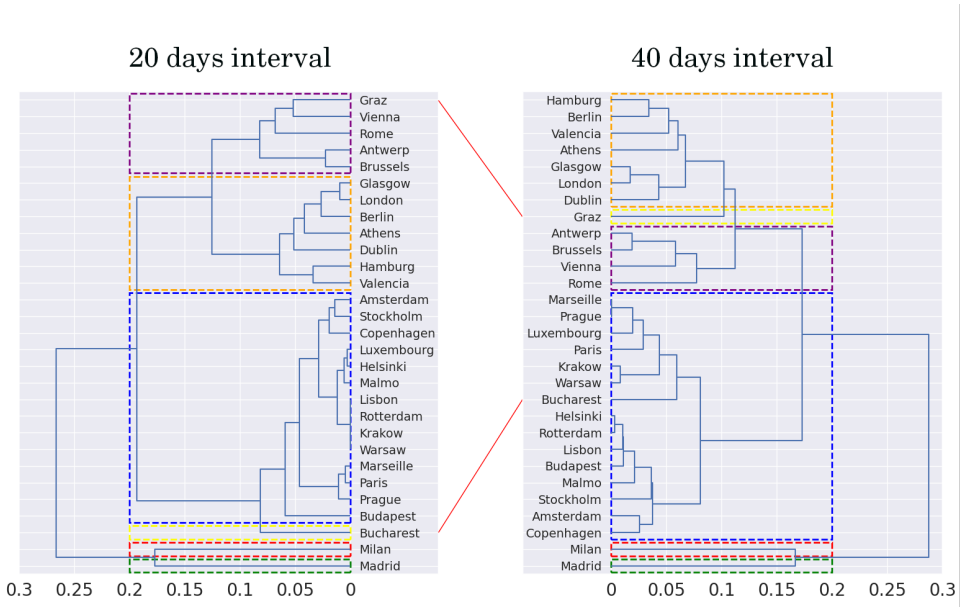


Figure 6.3: Comparison of dendrograms obtained as result from 20 and 40 days intervals for Driving Category

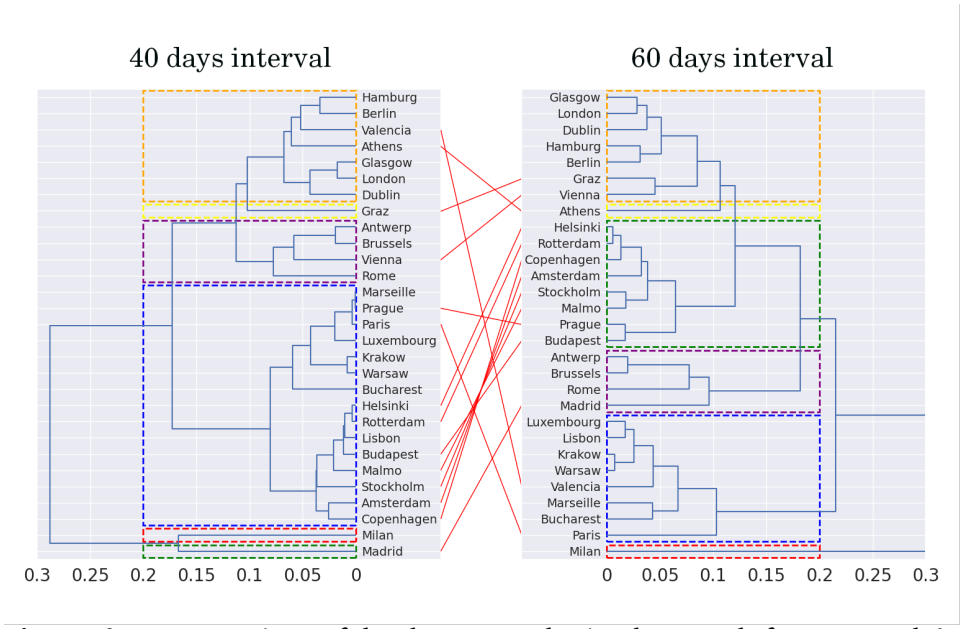


Figure 6.4: Comparison of dendrograms obtained as result from 40 and 60 days intervals for Driving Category

cluster (Vienna, Rome, etc.) shift to a one-city-only cluster. Fig. 6.4 shows the dendrograms of the clusters transiting from a window of 40 to 60 days. In such a timeframe, we witness a more extensive transformation. For example, the blue cluster splits into two smaller groups, one that comprises mostly

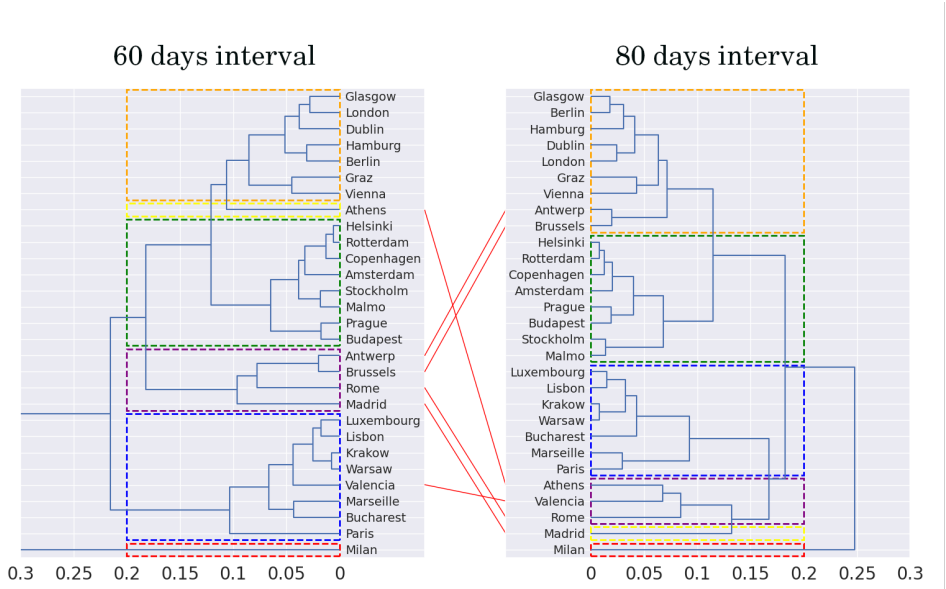


Figure 6.5: Comparison of dendrograms obtained as result from 60 and 80 days intervals for Driving Category

the Scandinavian cities, while the other is a mix of different cities, including French and Eastern European towns. The creation of a specific cluster for the Scandinavian cities highlight the consequences of the specific public health measures taken by these countries, which were notably different than other European countries.

Some of the results obtained are easy to explain, others not. Specifically, considering the 80 days dendrogram on the right of Fig. 6.5, the sets are as follows:

- Cluster 1 comprises most of the cities from the central-north European region (Berlin, Hamburg, London, Glasgow, Dublin, Vienna, Graz, Antwerp, and Brussels);
- Cluster 2 comprises Scandinavian (Stockholm, Malmö, Copenhagen, Helsinki) and Dutch (Amsterdam, Rotterdam) cities, plus Budapest and Prague;
- Cluster 3 comprises French (Paris and Marseille) and Polish (Warsaw and Krakow) cities, plus Luxembourg, Lisbon, and Bucharest;
- Cluster 4 comprises cities from the south European region (Rome, Valencia, Athens);
- Cluster 5 comprises Madrid;
- Cluster 6 comprises Milan;

Most clusters identify cities belonging to the same geographical region like cluster 2 and cluster 4 representing Scandinavian and Southern European regions. These two groups are an example of two radically different approaches to tackle the pandemic. The public institutions of cluster 4 applied very strong lockdown policies, while Scandinavian countries applied soft restrictions by encouraging citizens to follow government instructions at the same time.

Concerning clusters that include cities from different geographical regions, the explanation for being grouped is profound and has to be found in the pandemic spread in the city, the specific measures taken by authorities, and the citizens' reaction. Cluster 3 is an example of such clusters as it combines cities from eastern Europe (e.g., Bucharest, Warsaw, Krakow) with cities from western Europe (e.g., Luxembourg, Paris, Marseille).

Looking at all the clusters, it is interesting to note how there is no strong correlation between the clusters and the morphology of cities. With reference to Tab 6.1, we can see how cities with similar average edge lengths (i.e., cities with roads of similar length) like Helsinki (122) and Antwerp (120) or Milan (97) and Rotterdam (95) end up on different clusters. Another important morphology parameter is the number of edges that together with the population of the city provides a measure of urban density. We can see how the clusterization is not influenced by this parameter. An example is given by Cluster 3 which includes Paris and Bucharest that have the same population (2,14 and 2,10 million residents - accounting only for the residents in the municipality and not the neighboring areas), but while Paris has a number of edges close to 20 thousand, Bucharest has a complete different urban density with a number of edges that is double, i.e., more than 40 thousand.

It is also interesting to note how cities from the same country can belong to different clusters. For example, the Italian cities (Rome in cluster 4 and Milan in cluster 6) and Spanish cities (Valencia in cluster 4 and Madrid in cluster 5), although they share similar mobility trends, differ in the evolution of the number of SARS-COVID-19 infected people. Specifically, Madrid and Milan had the earliest outbreak of the pandemic in their respective countries and in general in the considered set of cities in this work.

Forecasting

Next, we perform a forecasting analysis per mobility category (Driving, Walking, and Transit) using as history the time horizon of 80 days after the 1% of cases and we obtain forecasts for the subsequent 10 days. Please note that the starting day from which we count the 80 days is different for each city and that the 6 resulting clusters are different for each category. For example, for the driving category (dendrogram on the right inside of Fig. 6.5), there are two one-city-only clusters for Milan and Madrid, but

this is not valid any longer for transit and walking categories. For sake of completeness we decided to show all clusters including the one-city-only, in this way we show the peculiarity of these cities that justifies being clusters of their own.

We first show the prediction results, and later we show the error made computed with the Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\sum_{1=1}^{n} \frac{(\overline{y_i} - y_i)^2}{n}}, \tag{6.2}$$

where $\overline{y_i}$ are the predicted values, y_i the observed values, or ground truth, and n is the length of these two series. Please note that we obtain one prediction per cluster.

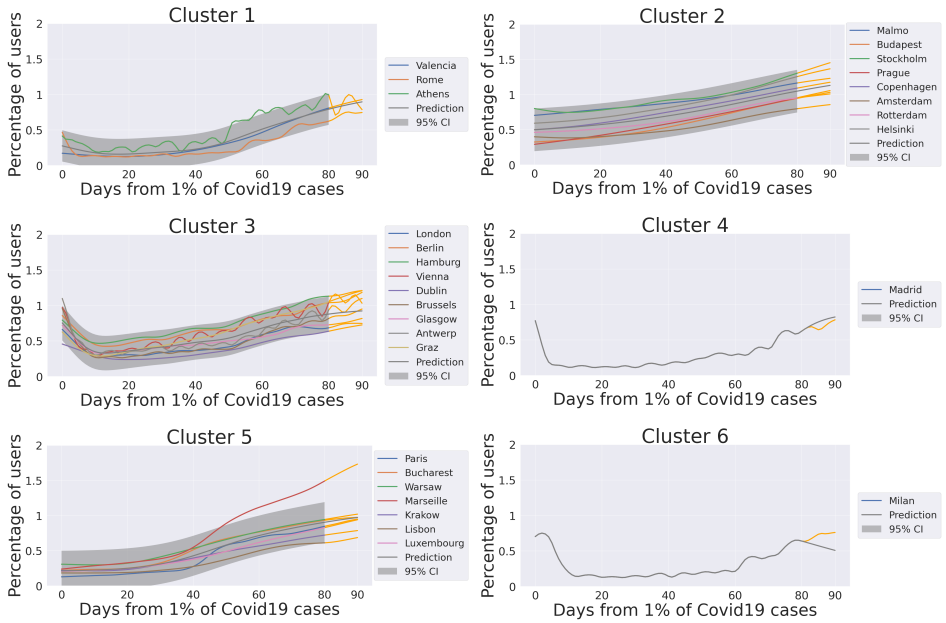
Driving. Fig. 6.6 shows the results obtained for Driving. As mentioned above, two of the six clusters are one-city-only clusters for Madrid and Milan. In Fig. 6.6(a), these cities are included in cluster 4 and 6 respectively. Their behavior is characterized by the fact that the municipalities had the earliest cases of SARS-COVID-19 in Europe. The plots show how in the first 10 days the level of driving activities follows standard trends in both cities and is followed by a rapid decrease caused by the application of the confinement policies. By contrast, cluster 1 shows cities that in their first 10 days are already at a low level of driving activities. The reason is that cluster 1 includes cities like Valencia and Rome for which the citizens learned the lesson from Madrid and Milan, and they reduced their mobility before reaching a high number of SARS-COVID-19 cases.

Fig. 6.6(b) shows that the predictions for cluster 5 are the worst of the category (average error 18%), the highest error is attributed to Marseille (68%). We observe a much earlier re-start for this city than in all the other cities in the cluster. Note that Marseille never hit the low level of driving activity possible to observe for the other cities of this cluster (i.e., a decrease of around 20%).

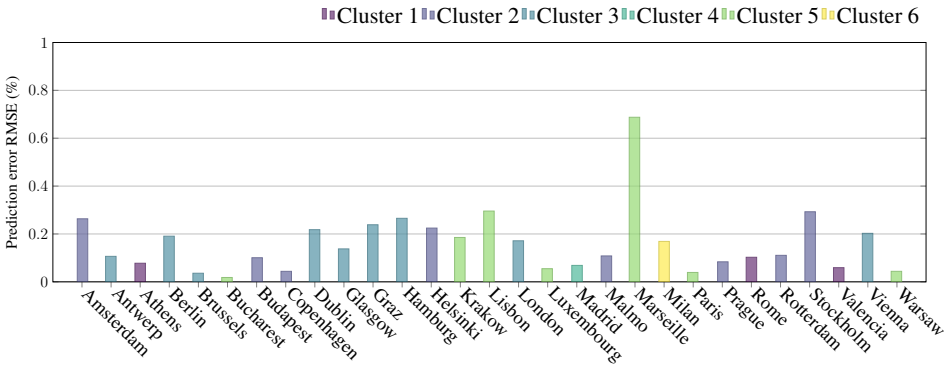
Within cluster 2, cities of Sweden like Malmö, Stockholm have benefited from mild lockdown restrictions, which explains why their driving activities profile is high. The average forecasting error for the cluster is 15%.

As for the remaining cluster, Cluster 3, the predictions are reasonably accurate (the prediction error is on average 17%). We observe the following interesting fact. Compared to the other cities, the cities of the UK and Ireland (e.g., London, Glasgow, and Dublin) did not apply strong lockdown restrictions, but their driving activities profile is the lowest of the cluster. It indicates that citizens reduced their driving activities themselves.

Walking. Fig. 6.7 shows the results obtained for Walking. First of all, please note that compared to Driving, the different groups of cities in the clusters are different. For example, the new Cluster 1 now includes some cities that in the Driving category belong to Cluster 4 (e.g., Marseille and Bucharest).



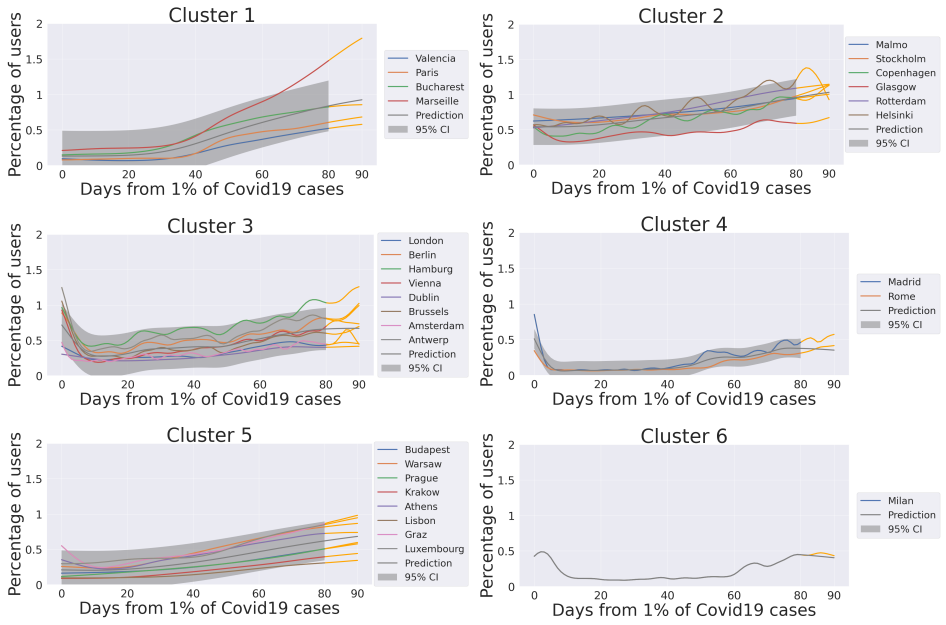
(a) Forecast



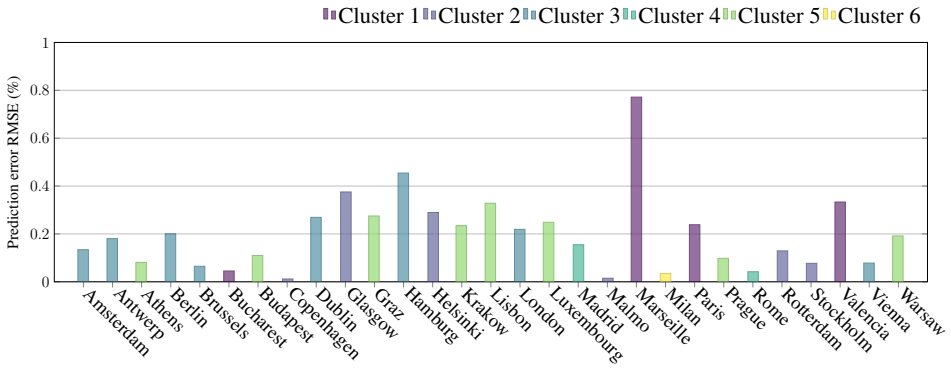
(b) RMSE

Figure 6.6: Forecasting analysis for the different clusters on Driving category

The cities in Cluster 1 have in common the following characteristic: low walking activities values during the first 40 days after having reached the 1% of SARS-COVID-19 cases and a high increase in the second 40 days. The prediction accuracy for this cluster (see Fig. 6.7(b)) is high, except for Marseille that shows the highest re-start compared to other cities (close to 200%) and the corresponding highest error (77%). This confirms that the response of Marseille in tackling the pandemic was unique as both driving and walking activities differ significantly from those of the respective comparable



(a) Forecast



(b) RMSE

Figure 6.7: Forecasting analysis for the different clusters on Walking category

cities per-category. Low values characterize the profile of Cluster 5 in the first half, likewise Cluster 1, but the recovery is slower during the second half of the observation window. The prediction error of this cluster is reasonably accurate, as it is always under 30%.

With respect to the Driving category, Cluster 2 does not contain anymore Amsterdam, Budapest, and Prague that moved to Cluster 3 and 5. This cluster now contains mainly cities from northern Europe, and the forecasting error is low (i.e., %14). The forecasting error increases for Cluster 3, mainly because of the presence of Hamburg that behaves differently from the cities

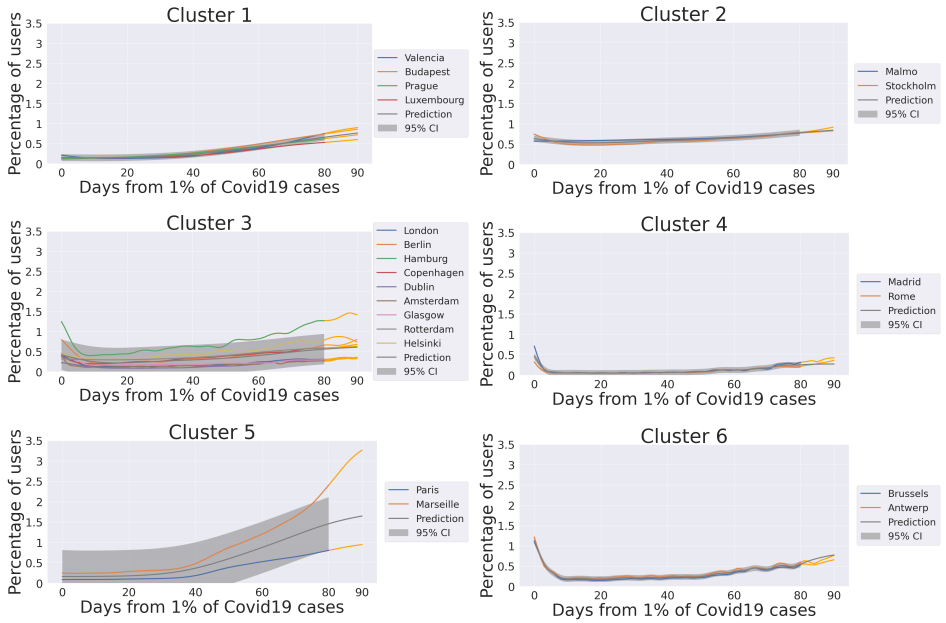
of the cluster (with a prediction error of 46%).

Regarding cluster 4, the only difference with respect to the Driving category is the addition of Rome. Madrid and Rome share a similar profile for walking activities, and the reasons for this similarity can be found in the analogous type of reactions enforced by the local authorities and the comparable size and population of the two cities. Cluster 6 still includes only Milan.

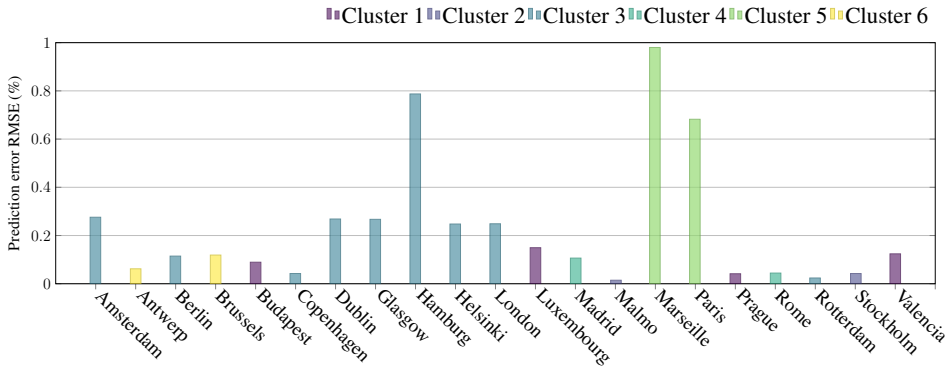
Transit. Finally, Fig. 6.8 shows the results for the Transit category for which the Apple dataset does not provide data for 8 cities (Athens, Bucharest, Graz, Krakow, Lisbon, Milan, Vienna, Warsaw). The reason is that not every city allows Apple Maps to give indications on public transport, which forces users to make use of alternative applications.

As a consequence, the clusters are very different from the ones obtained for the other categories. It is worth noticing that 3 clusters include only a pair of cities belonging to the same country, namely Malmö-Stockholm in Cluster 2, Paris-Marseille in Cluster 5, and Brussels-Antwerp in Cluster 6. The reason for this regional-based clusterization is indeed the lower dimension of the dataset due to the missing cities, but we can observe in fig. 6.8(b) that we obtained different forecasting results. While the forecasting precision for the Belgian and the Swedish clusters is high (error of 2% and 1%), the French cities differ significantly one with the other (forecasting error of 86% mainly because there are only two cities with Marseille being radically different). This confirms that the response of the citizens of Marseille has been unique in terms of mobility for all the transportation modalities. At a lower scale, Hamburg has tackled as well the pandemic differently from other cities of the clusters it belongs to. In this category, the error is 78%. It is remarkable how the values for transit have reached lower percentages than the other two categories and also the restart is slower. This can be appreciated because most of the cities never reached 100% of transit users even after 80 days from 1% of cases. It is interesting to notice that the restriction policies on public transport do not have a strong influence on the composition of clusters. Cluster 4 is a clear example, this group is composed of Madrid and Rome. We analyzed the policies of the corresponding countries through the Oxford COVID-19 Government Response Tracker dataset[4], we verified that Spain and Italy adopted a different strategy for public transport in April. While Spain only recommended citizens to not use transit services, in Italy public authorities enforced a policy of strict closure and reduction of capacity. These different strategies reveal that clusters are not lead only by governmental decisions, this is something we would like to study in deep in future researches, in order to detect which are the factors that influenced the most citizens behavior.

[4]https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker#data



(a) Forecast



(b) RMSE

Figure 6.8: Forecasting analysis for the different clusters on Transit category

6.4.4 Assessing The Impact of Mobility on Activities

This Subsection analyzes the impact of mobility on cities activities. To this end, we verify whether the clusters of cities identified in Subsection 6.4.2 are applicable to activity categories of the GPT dataset. We choose *Transit Stations* that is category directly related to mobility and *Stores*.

Fig. 6.9 shows the results obtained. The GPT dataset contains a lower subset of cities than the Apple Maps dataset, hence pick representative clusters from the driving category:

- Cluster a): Copenhagen, Stockholm and Malmö;
- Cluster b): Luxembourg;
- Cluster c): London;
- Cluster d): Milan, Madrid and Valencia.

The clusters are depicted in red in the graphs.

First of all, we observe that the overall similarity is higher for *Transit Stations* than *Stores*, which was expected. Fig. 6.9a shows how Luxembourg is an outlier in the matrix, the reason is that transit station category includes the subway stations and Luxembourg is the only city in the matrix that does not have a metro line. On the same figure we can see how inside the clusters defined from Apple driving the similarity values are stronger than outside, this means that the clusterization extracted from the Apple dataset is reflected also on the *Transit Stations* data. In the specific, we have the highest similarity values inside Cluster d while on Cluster a we have high values with the only exception of the similarity between Stockholm and Copenhagen. Looking at the *Stores* category on fig. 6.9b, please note that we are missing the city of Valencia for this dataset. For this matrix it is clear that we have misplaced Stockholm inside Cluster a, it is interesting to notice that for Stores category the city of Stockholm is more similar to Madrid and Milan than to other Scandinavian cities.

6.4.5 Assessing correlation of Mobility and SARS COVID-19

This Subsection analyzes the correlation between mobility and the evolution of the number of SARS COVID-19 cases in various cities. To this end, we verify whether the clusters of cities identified in Subsection 6.4.2 reflect the same similarities also in terms of the number of infected cases. We choose the clusters obtained from *Driving* category with a timeline of 80 days after the 1% of contagious, *Driving* is the category with the largest dataset and 80 days is the longest available timeline in our dataset. To compute the SARS COVID-19 similarity between cities, we extract the number of cases for each city for the same timelines of *Driving* clusters (80 days) and then we normalize by the total number of cases in the city. The SARS COVID-19 dataset contains a lower subset of cities than the Apple Maps dataset. Compared to the driving dataset, we excluded four cities, Bucharest, Budapest, Paris, and Marseille.

Fig. 6.10 shows the results obtained. The similarity matrix compares the different SARS COVID-19 trends while the red lines display the *Driving* clusters. The similarity is computed with the JSD metric that is explained in Subsection 6.4.2. The cities are sorted by driving cluster and the order inside the cluster is given from the similarity distance between the cities for driving data.

(a) Transit stations

(b) Stores

Figure 6.9: Similarity Matrix of popularity trend from GPT

We first observe the presence of different outliers, Malmö and Luxembourg, because have the lowest values of similarities with respect to other cities. As for Malmö, our mobility analysis highlighted that the city behaves similarly to Stockholm for all the categories while in terms of the number of infections it is interesting to note that the cities are profoundly different. A possible reason could be that many residents in Malmö commute daily to Copenhagen, Denmark, hence it has in terms of potential contacts a very distinct behavior with respect to Sweden's capital city. As for Luxembourg, the difference can be justified by *i)* the peculiarity of the city (in terms of population, number of workers commuting every day from neighboring countries) and the large scale testing applied by the government.

Next, we observe that it is possible to identify strong intra-cluster similar-

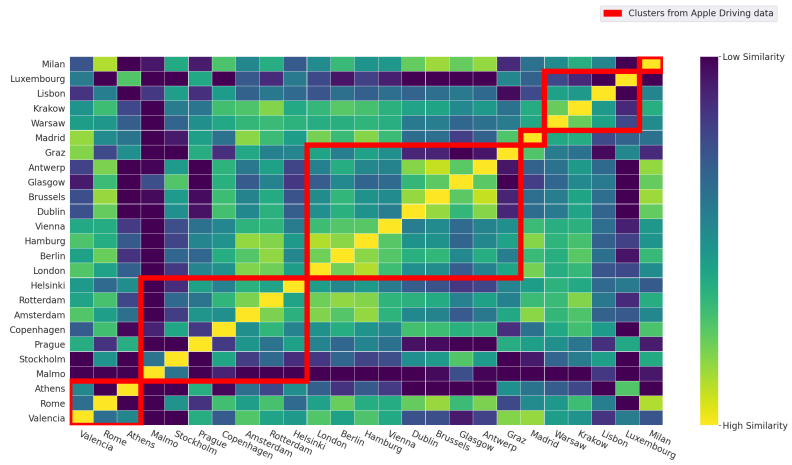


Figure 6.10: Similarity Matrix of contagious trends of SARS COVID-19

ity for Clusters 2, 3, and 5. For example, with regard to Cluster 2, we can identify a sub-cluster of Amsterdam, Rotterdam, and Helsinki. With regard to Cluster 3, we can identify two sub-clusters: London to Vienna and Dublin to Antwerp. Further, by performing a clustering analysis solely based on the number of SARS COVID-19 cases, the resulting clusters would be different. For example, Amsterdam, Rotterdam, London, Berlin, Hamburg, and Vienna would be assigned to a single cluster.

6.5 Concluding Remarks and Future Research Directions

In this paper, by using different crowdsensed datasets, we perform an analysis to uncover the impact of the SARS COVID-19 outbreak on the changes in mobility in urban environments. Specifically, we use Gaussian Processes and clustering techniques on the Apple Maps data to uncover patterns of similarity between the major European cities and perform a prediction analysis that permits forecasting the trend of the recovery process.

We identify a range of interesting behaviors. For example, the repetition of our clustering methods over different intervals highlighted an evolution of the mobility trends of many cities along the days after the outbreak. We detected a group of cities that defined a cluster only after many days after the outbreak, such as the Scandinavian cities that became a proper cluster only after 60 days from the outbreak. Apart from few changes, our methodology produced stable clusters, most of them region-wise, from which we extracted a common trend useful to understand the behaviors of different cities and improve the forecasting of the next days.

Regarding the forecasting, we exploited the 80 days after the outbreak to

predict the coming 10 days, we predicted the trend of each cluster obtaining low prediction errors, on average we obtained prediction errors of 14% for driving category, 19% for walking, and 24% for transit. We identified outlier cities like Marseille and Hamburg, i.e., cities where citizens have used transportation modes radically differently from the cities in the respective clusters.

The results of this study are useful for municipalities and local authorities to identify other towns with a similar reaction to the pandemic spread in terms of mobility. The possible application of the mobility clusters and their patterns is to help cities to perform a critical assessment of the efficacy of confinement measures enforced and whether might be more convenient to adopt a different policy used by cities in other clusters.

In our future research, We would like to exploit additional crowdsourced datasets, the Apple maps data is based on Apple users who asked for directions while using multiple sources of data could help on representing the true travel behaviors of all citizens.

Part IV

Conclusion

Chapter 7

Summary and Future research directions

The chapter concludes the thesis by presenting a summary and outlining future research directions.

7.1 Summary

Traditional data sources for mobility, such as traffic counts and public transit ridership, can be limited in availability and granularity. A potential solution to these limitations is the use of a crowdsourced dataset, which can provide more updated and granular data on mobility. In this context, this thesis aims to combine the latest data-driven methodologies with novel crowdsourced datasets. The main goal of the studies presented in this dissertation was to examine the potential of a crowdsourced dataset for mobility research. The results of these studies will be useful for understanding the usefulness of crowdsourced datasets to overcome the main limitations of traditional mobility datasets. This thesis consists of several chapters, each addressing a specific RQ and presenting a contribution and takeaway message. A flowchart summarizing the RQs, related contributions, and corresponding takeaway messages can be found in Figure 7.1. The first part of this thesis presented the context, research questions, and contributions of this work. It began by explaining the significance of data for mobility and how crowdsourced data can be beneficial in this field. The research questions (RQs) addressed in this thesis were then described, along with the various obstacles encountered in achieving the results.

In **Chapter 2**, we addressed RQ1, " What are the key differences between crowdsourced data and traditional transport data in terms of their potential for mobility analysis?". The purpose of this chapter was to provide a more in-depth and technical understanding of the current state of mobility datasets and to assess their strengths and weaknesses. This was accomplished by detailing the main features and limitations of different mobility datasets, as well as by presenting a scoring scheme to evaluate various aspects of these datasets. These aspects included availability, or how easily the data can be accessed and used; time dynamicity, or how frequently the data is updated and reflects changes in mobility patterns; and sample size, or the number of observations or data points included in the dataset. Along with these traditional mobility datasets, we also introduced and discussed the use of crowdsourced data as a potential supplement or alternative. We evaluated crowdsourced data using the same scoring scheme as for traditional datasets, and discussed how it could be used to address some of the limitations or gaps in traditional datasets.

Overall, the main message from this chapter tells us that Crowdsourced data has several attractive characteristics, including wide availability and long periods of collection, which make it a valuable resource for improving traditional datasets. It is highlighted that crowdsourced data has several advantageous attributes such as its wide availability and prolonged collection periods, making it a valuable resource to enhance traditional datasets. The chapter also focuses on the benefits of GPT, a specific type of crowdsourced data. We outline the main characteristics and limitations of GPT. This tool

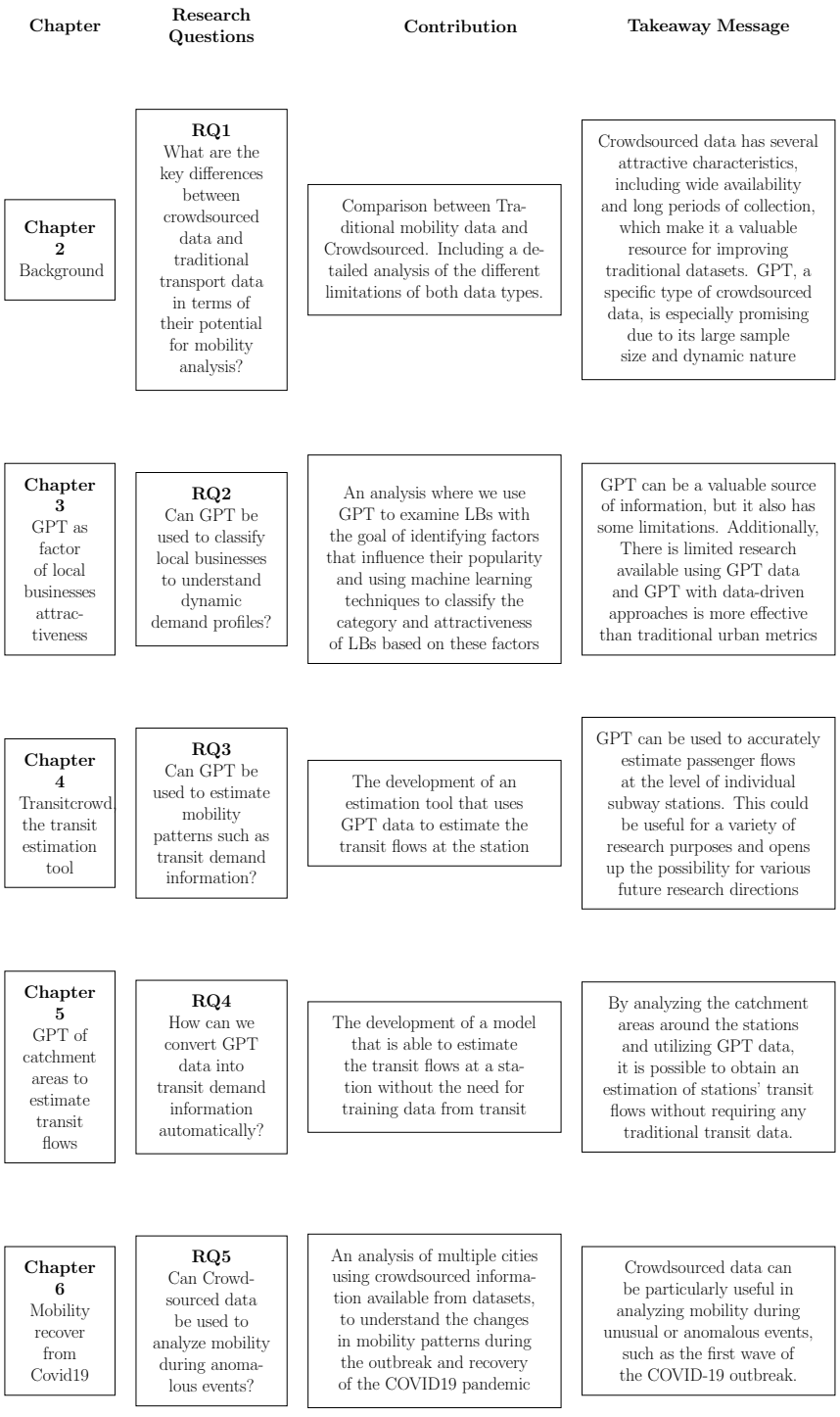


Figure 7.1: Summary of RQs, Contributions, and takeaways messages

offers users real-time information about the level of crowdedness at local businesses level, but it is important to note that there are different challenges associated with this data: first, the data is normalized, meaning that it does not accurately reflect the actual number of people at a location. Additionally, we have no control over how the data is collected or processed, and there is limited information available about how GPT data is created. Finally, it can be difficult to extract information about mobility patterns from GPT data, as it primarily provides insights about crowdedness at activity locations. From this chapter, it emerges that there are two main potential contributions of GPT for mobility analysis. The first is that GPT can provide dynamic information about secondary activities that traditional mobility data does not cover. The second is that GPT can be a substitute for mobility data in areas where it is lacking. Now that we have determined that GPT is a promising dataset for mobility research, we want to focus on investigating these two potentialities.

In **Chapter 3** we started by analyzing the potential of GPT for secondary activities, we address RQ2, "Can GPT be used to classify local businesses to understand dynamic demand profiles?". We conducted an analysis of GPT and LBs. As GPT data is collected at the level of local businesses, we wanted to examine the possibility of extracting dynamic demand information for secondary activities. Such data is difficult to obtain with traditional mobility data. Information on secondary activities is crucial for mobility models that usually have to rely only on static information, such as the number of points of interest in a certain area. To do this, we first investigated urban metrics that may influence the popularity of LBs such as the centrality of places in street networks. Then, we used machine learning techniques to classify the category and attractiveness of LBs based on the considered features. This analysis was conducted with the goal of better understanding how GPT data can be used to analyze mobility patterns. To summarize, this chapter teaches us that using GPT with data-driven approaches is more effective than traditional urban metrics at estimating the attractiveness of LBs.

Having examined the key value of GPT for secondary activities, we are now ready to move on to other potentialities and evaluate the possibility of extracting mobility flows from this data. As we already mentioned, one of the limitations of GPT is the lack of a direct connection to mobility. To address this, we preprocessed GPT and combined it with other sources in order to extract mobility insights.

This is what we did in **Chapter 4**, where we address RQ3, "Can GPT be used to estimate mobility patterns such as transit demand information?". This chapter focuses on the potential for extracting transit information from GPT. Specifically, we explored the use of GPT to estimate passenger flows at individual subway stations. Since GPT only provides crowding trends for stations, We developed a framework called TransitCrowd that uses GPT data to make real-time estimates of transit activity at the level of individual subway stations. Our framework is flexible and consists of two separate estimator

tools. The first, the Reg estimator, prioritizes the accuracy of results at the city level. The second, the Sig estimator, extracts signatures from stations to reveal the temporal patterns of the correlation between GPT data and actual entrances and exits. This information allows the presented methodology to be applied to other cities. Finally, we evaluated the performance of TransitCrowd by using it to estimate two months of entrance and exit flows at each station using GPT Live data as input. The results of this process were promising, with the accuracy of the estimates appearing to be stable across several consecutive weeks. We also found that TransitCrowd is able to accurately estimate entrance and exit flows in weeks different from the ones used for training, and that errors were not influenced by high or low values of entrance and exit activity. The main takeaway from this chapter is that GPT can be used to accurately estimate passenger flows at the level of individual subway stations. This could be useful for a variety of research purposes and opens up the possibility for various future research directions, which we will discuss in Section 7.2.

However, the contribution of this chapter does not exhaustively explore the potential of GPT to provide transit information in situations where traditional transit data is completely absent. As mentioned, the Sig estimator claims to be able to be applied to other cities than the one it was trained on, which would allow for the prediction of transit flows in areas where there is not enough traditional transit data to train our estimation tool.

In order to fully investigate this aspect, in **Chapter 5** we addressed RQ4, " How can we convert GPT data into transit demand information automatically?". We focus on the possibility of estimating signatures in a city without requiring any traditional transit data. Specifically, it uses GPT data for activities around the station as a substitute for traditional transit flows data. The goal is to enable the estimation of transit flows at a station even in cases where traditional transit information is not available, providing valuable insights for transportation planners and researchers in areas where such data is not accessible or does not exist. To achieve this, we develop a framework that estimates the signatures of a station. The results indicated that the model was effective at predicting the signature of stations in a different city, with the Extra Trees Regressor being the most effective model. We also experimented with various catchment area detection methods and found that using weighted distance provided the best results. Overall, this chapter emphasizes the importance of considering the surrounding context of transit stations as a valuable source for understanding the relationship between GPT data and transit data. By analyzing the catchment areas around the stations and utilizing GPT data, it is possible to obtain an estimation of stations' transit flows without requiring any traditional transit data.

We have previously examined the hypothesis of using crowdsourced data to estimate mobility under normal conditions, but we have not yet considered the possibility of using crowdsourced data to analyze mobility during unusual

or unexpected conditions.

In **Chapter 6** we addressed RQ5, " Can Crowdsourced data be used to analyze mobility during anomalous events?". Our focus is particularly on analyzing mobility during a specific unusual event, the first wave of the COVID-19 pandemic. We perform an analysis by using different crowdsourced datasets to uncover the impact of the SARS COVID-19 outbreak on the changes in mobility in urban environments. Specifically, we use Gaussian Processes and clustering techniques on the Apple Maps data to uncover patterns of similarity between the major European cities and perform a prediction analysis that permits forecasting the trend of the recovery process. We found several intriguing patterns. For instance, repeating our clustering techniques on various intervals revealed changes in mobility trends in many cities over the days following the outbreak. We discovered a group of cities that formed a cluster only after a significant number of days from the outbreak, such as the Scandinavian cities that became a cluster only after 60 days from the outbreak. With a few exceptions, our method consistently generated stable clusters, many of which were regional, which allowed us to extract a common trend that helped us understand the behaviors of different cities and improve forecasting for future days. Regarding the forecasting, we exploited the 80 days after the outbreak to predict the coming 10 days, we predicted the trend of each cluster obtaining low prediction errors. The results of this study are useful for municipalities and local authorities to identify other towns with similar responses to the pandemic in terms of mobility. The potential use of these mobility clusters and their patterns is to help cities evaluate the effectiveness of their confinement measures and consider alternative policies implemented by cities in other clusters. One key conclusion from this chapter is that crowdsourced data can be particularly useful in analyzing mobility during unusual or anomalous events, such as the first wave of the COVID-19 outbreak. By using crowdsourced data to analyze mobility during these types of events, researchers and policymakers can gain insights into how people are moving and potentially make informed decisions about how to respond to such events.

7.2 Future research directions

This section discusses the potential for future research in the area of using crowdsourced data for mobility analysis. This is a key finding of this thesis, as there is currently limited research on the use of GPT for mobility analysis. An important result of this work is to open new research directions that can be explored using crowdsourced data for mobility, with a focus on GPT.

Each chapter in this thesis, along with its corresponding contribution, creates new research directions that generate research questions that are not addressed in this work but can be explored in future research. This means

that while this thesis may provide answers to certain RQs, contributing to augmenting the knowledge of exploiting crowdsourced for mobility, there are still many areas that remain unexplored and could be the focus of future studies. These new research directions may be related to the specific topics addressed in each chapter or may be broader in scope. These new research directions offer the potential for further exploration and the opportunity to build on the insights gained from this thesis.

GPT as a queue process

In Chapter 3, we presented the main characteristics of GPT data and analyzed its use for local businesses. One of the main limitations highlighted in the chapter is the lack of direct information about mobility and the absence of control of the raw data. To further understand these limitations and the underlying mechanisms of GPT data, a potential future research direction is to investigate the applicability of a queuing process model for representing GPT data. This model can help to better understand the dynamics of the inflow and outflow of people at local businesses, and how the GPT's crowdedness value reflects the length of this queue. Additionally, it is important to specify how the duration of activity at the LB, provided by GPT in the form of the mean duration of users' stays at the LB, can be incorporated into this model to understand the capacity and service rate. Therefore, a more specific research question could be: "How can the queuing process model be applied to represent GPT data for local businesses?"

Transitcrowd application to other scenarios

In Chapter 4, we introduced a method for utilizing GPT data to gain insights into mobility patterns. We developed Transitcrowd specifically for the purpose of analyzing the demand flows of transit service. While we have already proposed evaluating Transitcrowd in various cities, and have conducted tests in New York and Washington, there is still room for further exploration. One potential direction for future research is to continue testing the effectiveness of Transitcrowd in cities with different characteristics and transportation systems. This will help to determine the generalizability of our approach and whether we can identify the urban environments where this approach is applicable. In addition to evaluating Transitcrowd in different cities, it would also be valuable to consider the potential for expanding our analysis to other types of mobility services. For example, we could use GPT data to analyze the flow of traffic entering and leaving a specific district, or to study the usage of car-sharing or bike-sharing services in a certain area. This would enable us to create a more comprehensive tool that utilizes GPT to understand mobility trends for different transportation services in a given area, and could lead to the development of new strategies for optimizing transportation systems.

Connection between signature, activities, and structure of the city

Chapter 5 examined the feasibility of applying the Transitcrowd methodology to other cities. Specifically, we focused on estimating the signature values of a station using GPT data on the activities in the surrounding area. While we employed data-driven approaches, there is still much to be learned about the relationship between the signature of a station and the activities in the surrounding area.

One potential direction for future research is to delve deeper into the connection between the signature of a station, the surrounding area, and the structure of the city. This could involve analyzing the ways in which the signature is influenced by factors such as the density of population and personal incomes, the availability of transportation options, and the overall layout of the city. By gaining a deeper understanding of these connections, it may be possible to improve the accuracy and refinement of methods for predicting the signature of a station and enhance the transferability of Transitcrowd to other cities.

Merging Crowdsourced data for anomaly detection

In chapter 6, we explored the use of crowdsourced data to analyze mobility during unexpected events. Specifically, we looked at how mobility patterns changed for different cities during the COVID-19 pandemic. In this research we knew of the presence of a specific event, in this case, a pandemic, a potential direction from this study can be to use crowdsourced data to detect the occurrence of anomalies and unexpected events in mobility.

Future research could use crowdsourced data from different sources to identify and monitor unusual changes in mobility patterns, which could be indicative of an unexpected event. By analyzing changes in mobility over time, it may be possible to detect the early warning signs of such events and respond accordingly. Identifying anomalies in mobility patterns can be a challenging task, as it requires distinguishing between normal variations in demand or mobility, and unusual or unexpected changes. To accurately distinguish between the two, a deep understanding of the underlying patterns and dynamics of mobility is necessary.

One powerful way to tackle this challenge is by utilizing advanced statistical techniques to analyze continuous time-series data such as GPT. These techniques, such as time series decomposition, can help in understanding recurrent fluctuations and identifying random vs correlated variations in different indicators. Additionally, machine learning methods, such as anomaly detection algorithms, can be applied to identify patterns and behaviors that deviate significantly from the norm, thereby detecting anomalies and determining which factors are influencing the most variation.

In conclusion, by detecting changes in mobility patterns, it may be possi-

ble to anticipate and prepare for unexpected events, potentially mitigating their impact.

Appendix A

List of Publications

A.1 Journals

- The Impact of SARS-COVID-19 Outbreak on European Cities Urban Mobility
  P. Vitello, C. Fiandrino, A. Capponi, P. Klopp, R.D. Connors, F. Viti
  Frontiers in Future Transportation[111]
- Mobility-driven and energy-efficient deployment of edge data centers in urban environments
  P. Vitello, A. Capponi, C. Fiandrino, G. Cantelmo, D. Kliazovich
  IEEE Transactions on Sustainable Computing[112]
- A mobility-based deployment strategy for edge data centers
  M Girolami, P Vitello, A Capponi, C Fiandrino, L Foschini, P Bellavista
  Journal of Parallel and Distributed Computing[113]

A.2 Conferences

- Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities
  A. Capponi, P. Vitello, C. Fiandrino, G. Cantelmo, D. Kliazovich, U. Sorger, P. Bouvry
  2019 IEEE Symposium on Computers and Communications (ISCC)[114]
- The CORONA business in modern cities P Vitello, A Capponi, P Klopp, RD Connors, F Viti, C Fiandrino
  Proceedings of the 18th Conference on Embedded Networked Sensor Systems[115]
- The impact of human mobility on edge data center deployment in urban environments
  P Vitello, A Capponi, C Fiandrino, G Cantelmo, D Kliazovich
  2019 IEEE Global Communications Conference (GLOBECOM)[116]

Bibliography

[1] J. Armoogum, C. Garcia, Y. Gopal, *et al.*, "Study on new mobility patterns in european cities," 2022 (cit. on p. 17).

[2] M. Sheller, "Sustainable mobility and mobility justice," *Mobilities: New perspectives on transport and society*, p. 289, 2011 (cit. on p. 17).

[3] K. Gkoumas, F. L. Marques dos Santos, M. Stepniak, and F. Pekár, "Research and innovation supporting the european sustainable and smart mobility strategy: A technology perspective from recent european union projects," *Applied Sciences*, vol. 11, no. 24, p. 11 981, 2021 (cit. on p. 17).

[4] M. Kamargianni, W. Li, M. Matyas, and A. Schäfer, "A critical review of new mobility services for urban transport," *Transportation Research Procedia*, vol. 14, pp. 3294–3303, 2016 (cit. on p. 17).

[5] T. Teoh, B. van Berne, I. Hindriks, *et al.*, *Leveraging big data for managing transport operations*, 2019 (cit. on p. 17).

[6] S.-h. An, B.-H. Lee, and D.-R. Shin, "A survey of intelligent transportation systems," in *2011 Third International Conference on Computational Intelligence, Communication Systems and Networks*, IEEE, 2011, pp. 332–337 (cit. on p. 18).

[7] K. Dziekan and K. Kottenhoff, "Dynamic at-stop real-time information displays for public transport: Effects on customers," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 6, pp. 489–501, 2007 (cit. on p. 18).

[8] S. Porru, F. E. Misso, F. E. Pani, and C. Repetto, "Smart mobility and public transport: Opportunities and challenges in rural and urban areas," *Journal of traffic and transportation engineering (English edition)*, vol. 7, no. 1, pp. 88–97, 2020 (cit. on p. 18).

[9] L. Chapman, "Transport and climate change: A review," *Journal of transport geography*, vol. 15, no. 5, pp. 354–367, 2007 (cit. on p. 18).

[10] A. Nurdden, R. Rahmat, and A. Ismail, "Effect of transportation policies on modal shift from private car to public transport in malaysia," *Journal of applied Sciences*, vol. 7, no. 7, pp. 1013–1018, 2007 (cit. on p. 18).

[11] D. Milne and D. Watling, "Big data and understanding change in the context of planning transport systems," *Journal of Transport Geography*, vol. 76, pp. 235–244, 2019, ISSN: 0966-6923 (cit. on p. 18).

[12] B. L. Smith, W. T. Scherer, and J. H. Conklin, “Exploring imputation techniques for missing data in transportation management systems,” *Transportation Research Record*, vol. 1836, no. 1, pp. 132–142, 2003 (cit. on p. 18).

[13] N.-E. El Faouzi, H. Leung, and A. Kurian, “Data fusion in intelligent transportation systems: Progress and challenges–a survey,” *Information Fusion*, vol. 12, no. 1, pp. 4–10, 2011 (cit. on p. 18).

[14] M. Clarke, M. Dix, and P. Jones, “Error and uncertainty in travel surveys,” *Transportation*, vol. 10, no. 2, pp. 105–126, 1981 (cit. on p. 18).

[15] A. I. Torre-Bastida, J. Del Ser, I. Laña, M. Ilardia, M. N. Bilbao, and S. Campos-Cordobés, “Big data for transportation and mobility: Recent advances, trends and challenges,” *IET Intelligent Transport Systems*, vol. 12, no. 8, pp. 742–755, 2018 (cit. on p. 18).

[16] Z. Tao, J. Tang, and K. Hou, “Online estimation model for passenger flow state in urban rail transit using multi-source data,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 6, pp. 762–780, (cit. on p. 19).

[17] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry, “A survey on mobile crowdsensing systems: Challenges, solutions and opportunities,” *IEEE Communications Surveys Tutorials*, pp. 1–49, 2019 (cit. on pp. 19, 97, 98).

[18] H. Ma, D. Zhao, and P. Yuan, “Opportunities in mobile crowd sensing,” *IEEE Communications Magazine*, vol. 52, no. 8, pp. 29–35, 2014 (cit. on p. 19).

[19] M. N. Kamel Boulos, B. Resch, D. N. Crowley, *et al.*, “Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, ogc standards and application examples,” *International journal of health geographics*, vol. 10, no. 1, pp. 1–29, 2011 (cit. on p. 19).

[20] C. A. Le Dantec, M. Asad, A. Misra, and K. E. Watkins, “Planning with crowdsourced data: Rhetoric and representation in transportation planning,” in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 1717–1727 (cit. on p. 19).

[21] M. Poblet, E. García-Cuesta, and P. Casanovas, “Crowdsourcing tools for disaster management: A review of platforms and methods,” in *International Workshop on AI Approaches to the Complexity of Legal Systems*, Springer, 2013, pp. 261–274 (cit. on p. 19).

[22] Y. Demchenko, C. de Laat, and P. Membrey, “Defining architecture components of the big data ecosystem,” in *2014 International Conference on Collaboration Technologies and Systems (CTS)*, 2014, pp. 104–112 (cit. on p. 29).

[23] T. F. Welch and A. Widita, “Big data in public transportation: A review of sources and methods,” *Transport Reviews*, vol. 39, no. 6, pp. 795–818, 2019 (cit. on p. 33).

[24] K. Clifton and C. Muhs, “Capturing and representing multimodal trips in travel surveys,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2285, pp. 74–83, Dec. 2012 (cit. on p. 34).

[25] P. Stopher and S. Greaves, "Household travel surveys: Where are we going?" *Transportation Research Part A: Policy and Practice*, vol. 41, pp. 367–381, Feb. 2007 (cit. on p. 34).

[26] M. Clarke, M. C. Dix, and P. Jones, "Error and uncertainty in travel surveys," *Transportation*, vol. 10, pp. 105–126, 1981 (cit. on p. 35).

[27] L. Shen and P. R. Stopher, "Review of gps travel survey and gps data-processing methods," *Transport Reviews*, vol. 34, no. 3, pp. 316–334, 2014 (cit. on p. 35).

[28] M. Jiber, I. Lamouik, Y. Ali, and M. A. Sabri, "Traffic flow prediction using neural network," in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, IEEE, 2018, pp. 1–4 (cit. on p. 35).

[29] A. Sookun, R. Boojhawon, and S. D. Rughooputh, "Assessing greenhouse gas and related air pollutant emissions from road traffic counts: A case study for mauritius," *Transportation Research Part D: Transport and Environment*, vol. 32, pp. 35–47, 2014 (cit. on p. 35).

[30] T. Wagner, "Regional traffic impacts of logistics-related land use," *Transport Policy*, vol. 17, no. 4, pp. 224–229, 2010 (cit. on p. 35).

[31] W. Wang, J. P. Attanucci, and N. H. M. Wilson, "Bus passenger origin-destination estimation and related analyses," 2011 (cit. on p. 36).

[32] S. Foell, G. Kortuem, R. Rawassizadeh, S. Phithakkitnukoon, M. Veloso, and C. Bento, "Mining temporal patterns of transport behaviour for predicting future transport usage," New York, NY, USA: Association for Computing Machinery, 2013, ISBN: 9781450322157 (cit. on p. 36).

[33] M. Dixit and A. Sivakumar, "Capturing the impact of individual characteristics on transport accessibility and equity analysis," *Transportation Research Part D: Transport and Environment*, vol. 87, p. 102 473, 2020, ISSN: 1361-9209 (cit. on p. 36).

[34] Z. Yong-chuan, Z. Xiao-qing, C. Zhen-ting, *et al.*, "Traffic congestion detection based on gps floating-car data," *Procedia Engineering*, vol. 15, pp. 5541–5546, 2011 (cit. on p. 37).

[35] A. Gühnemann, R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner, "Monitoring traffic and emissions by floating car data," 2004 (cit. on p. 37).

[36] M. Nigro, M. Castiglione, F. Maria Colasanti, *et al.*, "Exploiting floating car data to derive the shifting potential to electric micromobility," *Transportation Research Part A: Policy and Practice*, vol. 157, pp. 78–93, 2022, ISSN: 0965-8564 (cit. on p. 37).

[37] B. Pender, G. Currie, A. Delbosc, and N. Shiwakoti, "Social media use during unplanned transit network disruptions: A review of literature," *Transport Reviews*, vol. 34, no. 4, pp. 501–521, 2014 (cit. on p. 37).

[38] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Using data from the web to predict public transport arrivals under special events scenarios," *Journal of Intelligent Transportation Systems*, vol. 19, no. 3, pp. 273–288, 2015 (cit. on p. 38).

[39] T. Moyo and W. Musakwa, "Using crowdsourced data (twitter & facebook) to delineate the origin and destination of commuters of the gautrain public transit system in south africa," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-2, pp. 143–150, Jun. 2016 (cit. on p. 38).

[40] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation research part C: emerging technologies*, vol. 68, pp. 285–299, 2016 (cit. on p. 38).

[41] M. S. Kaiser, K. T. Lwin, M. Mahmud, *et al.*, "Advances in crowd analysis for urban applications through urban event detection," *IEEE Trans. on Intelligent Transportation Systems*, pp. 1–21, 2017 (cit. on p. 38).

[42] N. Nandan, A. Pursche, and X. Zhe, "Challenges in crowdsourcing real-time information for public transportation," in *2014 IEEE 15th international conference on mobile data management*, IEEE, vol. 2, 2014, pp. 67–72 (cit. on p. 39).

[43] J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities," *GeoJournal*, vol. 78, no. 2, pp. 223–243, 2013 (cit. on p. 39).

[44] H. Huang, Y. Cheng, and R. Weibel, "Transport mode detection based on mobile phone network data: A systematic review," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 297–312, 2019, ISSN: 0968-090X (cit. on p. 39).

[45] Z. Wang, S. Y. He, and Y. Leung, "Applying mobile phone data to travel behaviour research: A literature review," *Travel Behaviour and Society*, vol. 11, pp. 141–155, 2018, ISSN: 2214-367X (cit. on p. 39).

[46] V. Aguiléra, S. Allio, V. Benezech, F. Combes, and C. Milion, "Using cell phone data to measure quality of service and passenger flows of paris transit system," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 198–211, 2014, Special Issue with Selected Papers from Transport Research Arena, ISSN: 0968-090X (cit. on p. 39).

[47] X. Lou and M. Yan, "Classifying subway passengers based on mobile network data analysis," in *Proc. of IEEE/ACIS ICIS*, 2021, pp. 92–96 (cit. on p. 39).

[48] M. G. Demissie, S. Phithakkitnukoon, T. Sukhvibul, F. Antunes, R. Gomes, and C. Bento, "Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: A case study of senegal," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 9, pp. 2466–2478, 2016 (cit. on p. 39).

[49] X. Wang, Z. Zhou, Z. Yang, Y. Liu, and C. Peng, "Spatio-temporal analysis and prediction of cellular traffic in metropolis," in *IEEE ICNP*, 2017, pp. 1–10 (cit. on pp. 39, 49).

[50] J. Zhao, L. Zhang, K. Ye, *et al.*, "Gltc: A metro passenger identification method across afc data and sparse wifi data," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–15, 2022 (cit. on p. 40).

[51] "The passenger flow status identification based on image and wifi detection for urban rail transit stations," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 119–129, 2019, ISSN: 1047-3203 (cit. on p. 40).

[52] T. Oransirikul, R. Nishide, I. Piumarta, and H. Takada, "Measuring bus passenger load by monitoring Wi-Fi transmissions from mobile devices," *Procedia Technology*, vol. 18, pp. 120–125, 2014, International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014, Warsaw, Poland, ISSN: 2212-0173 (cit. on p. 40).

[53] V. Kostakos, T. Camacho, and C. Mantero, "Towards proximity-based passenger sensing on public transport buses," *Personal and Ubiquitous Computing*, vol. 17, pp. 1807–1816, Dec. 2013 (cit. on p. 40).

[54] R. Wu, Y. Cao, C. H. Liu, P. Hui, L. Li, and E. Liu, "Exploring passenger dynamics and connectivities in beijing underground via bluetooth networks," in *2012 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2012, pp. 208–213 (cit. on p. 40).

[55] X. Hu, H. Zheng, W. Wang, and X. Li, "A novel approach for crowd video monitoring of subway platforms," *Optik*, vol. 124, no. 22, pp. 5301–5306, 2013, ISSN: 0030-4026 (cit. on p. 40).

[56] J. Zhang, J. Liu, and Z. Wang, "Convolutional neural network for crowd counting on metro platforms," *Symmetry*, vol. 13, no. 4, 2021 (cit. on p. 40).

[57] G. Solmaz, P. Baranwal, and F. Cirillo, "CountMeIn: Adaptive crowd estimation with Wi-Fi in smart cities," in *Proc. of IEEE PerCom*, 2022, pp. 187–196 (cit. on p. 40).

[58] E. Chaniotakis, C. Antoniou, and F. Pereira, "Mapping social media for transportation studies," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 64–70, 2016 (cit. on p. 41).

[59] J. Pang and Y. Zhang, "Deepcity: A feature learning framework for mining location check-ins," in *Eleventh International AAAI Conference on Web and Social Media*, 2017 (cit. on p. 41).

[60] X. Zhou and L. Zhang, "Crowdsourcing functions of the living city from twitter and foursquare data," *Cartography and Geographic Information Science*, vol. 43, no. 5, pp. 393–404, 2016 (cit. on p. 41).

[61] Y. Zhang, M. Humbert, T. Rahman, C.-T. Li, J. Pang, and M. Backes, "Tagvisor: A privacy advisor for sharing hashtags," Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, ISBN: 9781450356398 (cit. on p. 41).

[62] J. T. Méndez, H. Lobel, D. Parra, and J. C. Herrera, "Using twitter to infer user satisfaction with public transport: The case of santiago, chile," *IEEE Access*, vol. 7, pp. 60 255–60 263, 2019 (cit. on p. 41).

[63] E. Chaniotakis, C. Antoniou, J. M. S. Grau, and L. Dimitriou, "Can social media data augment travel demand survey data?" In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2016, pp. 1642–1647 (cit. on p. 41).

[64] S. Timokhin, M. Sadrani, and C. Antoniou, "Predicting venue popularity using crowd-sourced and passive sensor data," *Smart Cities*, vol. 3, no. 3, pp. 818–841, 2020, ISSN: 2624-6511 (cit. on p. 44).

[65] D. Fry, J. A. Hipp, C. Alberico, J.-H. Huang, G. S. Lovasi, and M. F. Floyd, "Land use diversity and park use in new york city," *PREVENTIVE MEDICINE REPORTS*, vol. 22, 2021, ISSN: ["2211-3355"] (cit. on p. 44).

[66] A. Mahdi and D. Esztergár-Kiss, "Robust linear regression-based gis technique for modeling the processing time at tourism destinations," in *HCI in Mobility, Transport, and Automotive Systems*, H. Krömker, Ed., Cham: Springer International Publishing, 2022, pp. 557–569, ISBN: 978-3-031-04987-3 (cit. on p. 44).

[67] J. Dixon, I. Elders, and K. Bell, "Evaluating the likely temporal variation in electric vehicle charging demand at popular amenities using smartphone locational data," *IET Intelligent Transport Systems*, vol. 14, no. 6, pp. 504–510, (cit. on p. 44).

[68] J. Dixon, K. Bell, and I. Elders, "Electric vehicle destination charging characterisations at popular amenities," in *E-Mobility Power System Integration Symposium 2018*, 2018 (cit. on p. 44).

[69] V. Paidi, "Short-term prediction of parking availability in an open parking lot," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 541–554, 2022 (cit. on p. 45).

[70] V. Mahajan, G. Cantelmo, and C. Antoniou, "Explaining demand patterns during COVID-19 using opportunistic data: A case study of the city of munich," *European Transport Research Review*, vol. 13, no. 1, pp. 1–14, 2021 (cit. on p. 45).

[71] P. Vitello, A. Capponi, P. Klopp, R. D. Connors, F. Viti, and C. Fiandrino, *The corona business in modern cities: Poster abstract*, 2020 (cit. on p. 45).

[72] J.-D. Schmöcker, "Estimation of city tourism flows: Challenges, new data and covid," *Transport Reviews*, vol. 41, no. 2, pp. 137–140, 2021 (cit. on p. 45).

[73] J. M. Bandeira, P. Tafidis, E. Macedo, *et al.*, "Exploring the potential of web based information of business popularity for supporting sustainable traffic management," *Transport and Telecommunication Journal*, vol. 21, no. 1, pp. 47–60, 2020 (cit. on p. 45).

[74] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014 (cit. on p. 49).

[75] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transportation research part C: emerging technologies*, vol. 26, pp. 301–313, 2013 (cit. on p. 49).

[76] Y. Zhou, B. P. L. Lau, C. Yuen, B. Tunçer, and E. Wilhelm, "Understanding urban human mobility through crowdsensed data," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 52–59, 2018 (cit. on p. 49).

[77] P. Vitello, A. Capponi, C. Fiandrino, P. Giaccone, D. Kliazovich, and P. Bouvry, "High-precision design of pedestrian mobility for smart city simulators," in *Proc. IEEE ICC*, 2018, pp. 1–6 (cit. on pp. 51, 99).

[78] K. D'Silva, A. Noulas, M. Musolesi, C. Mascolo, and M. Sklar, "Predicting the temporal activity patterns of new venues," *EPJ Data Science*, vol. 7, no. 1, p. 13, 2018 (cit. on pp. 51, 104).

[79] M. G. McNally, "The four-step model," in *Handbook of transport modelling*, Emerald Group Publishing Limited, 2007 (cit. on p. 60).

[80] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, vol. 78, pp. 273–282, 2017, ISSN: 0957-4174 (cit. on p. 67).

[81] S. Kolassa and W. Schütz, "Advantages of the mad/mean ratio over the mape," *Foresight: The International Journal of Applied Forecasting*, pp. 40–43, 2007 (cit. on p. 69).

[82] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991 (cit. on p. 81).

[83] A.-E. Baert and D. Seme, "Voronoi mobile cellular networks: Topological properties," in *Third International Symposium on Parallel and Distributed Computing/Third International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks*, IEEE, 2004, pp. 29–35 (cit. on p. 81).

[84] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, and M. DATA, "Practical machine learning tools and techniques," in *Data Mining*, vol. 2, 2005 (cit. on p. 85).

[85] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017 (cit. on p. 86).

[86] A. Hoback, S. Anderson, and U. Dutta, "True walking distance to transit," *Transportation Planning and Technology*, vol. 31, no. 6, pp. 681–692, 2008 (cit. on p. 86).

[87] M. F. Boni, P. Lemey, X. Jiang, *et al.*, "Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic," *Nature Microbiology*, 2020, ISSN: 2058-5276 (cit. on p. 96).

[88] M. Simsek and B. Kantarci, "Artificial intelligence-empowered mobilization of assessments in COVID-19-like pandemics: A case study for early flattening of the curve," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, 2020, ISSN: 1660-4601 (cit. on p. 96).

[89] T. Favale, F. Soro, M. Trevisan, I. Drago, and M. Mellia, "Campus traffic and e-Learning during COVID-19 pandemic," *Computer Networks*, pp. 1–1, 2020, ISSN: 1389-1286 (cit. on p. 96).

[90] Ericsson Research, *Mobility report june 2020*, Jun. 2020 (cit. on p. 96).

[91] A. Feldmann, O. Gasser, F. Lichtblau, *et al.*, "The lockdown effect: Implications of the COVID-19 pandemic on internet traffic," in *Proceedings of the ACM Internet Measurement Conference (IMC)*, 2020, 65–72 (cit. on p. 96).

[92] D. Bertsimas, L. Boussioux, R. Wright, *et al.*, *COVIDanalytics project*, 2020 (cit. on p. 96).

[93] B. Egwolf and N. Austriaco, "Mobility-guided modeling of the COVID-19 pandemic in metro Manila," 2020 (cit. on p. 96).

[94] P. Bryant and A. Elofsson, "Estimating the impact of mobility patterns on covid-19 infection rates in 11 european countries," *PeerJ*, vol. 8, e9879, Sep. 2020, ISSN: 2167-8359 (cit. on p. 97).

[95] A. Kapoor, X. Ben, L. Liu, *et al.*, "Examining covid-19 forecasting using spatio-temporal graph neural networks," *arXiv preprint arXiv:2007.03113*, 2020 (cit. on p. 97).

[96] H. Wang and N. Yamamoto, "Using a partial differential equation with google mobility data to predict covid-19 in arizona," *Mathematical Biosciences and Engineering*, vol. 17, no. 5, 2020 (cit. on p. 97).

[97] L. Zhang, S. Ghader, M. L. Pack, *et al.*, "An interactive covid-19 mobility impact and social distancing analysis platform," *medRxiv*, 2020 (cit. on p. 97).

[98] M Kendall, M Parker, C Fraser, *et al.*, "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 368, no. 6491, 2020 (cit. on p. 97).

[99] S. Whitelaw, M. A. Mamas, E. Topol, and H. G. Van Spall, "Applications of digital technology in COVID-19 pandemic planning and response," *The Lancet Digital Health*, 2020 (cit. on p. 97).

[100] J. H. Reelfs, O. Hohlfeld, and I. Poese, "Corona-Warn-App: Tracing the start of the official COVID-19 exposure notification app for germany," in *Accepted as Poster in Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM)*, 2020, 1–3 (cit. on p. 97).

[101] C. Menni, A. Valdes, M. B. Freydin, *et al.*, "Loss of smell and taste in combination with other symptoms is a strong predictor of covid-19 infection," 2020 (cit. on p. 97).

[102] S. Engle, J. Stromme, and A. Zhou, "Staying at home: Mobility effects of covid-19," *Available at SSRN*, 2020 (cit. on p. 97).

[103] M. M. Rahman, J.-C. Thill, and K. C. Paul, "Covid-19 pandemic severity, lockdown regimes, and people's mobility: Early evidence from 88 countries," *Sustainability*, vol. 12, no. 21, 2020 (cit. on p. 97).

[104] V. Mahajan, G. Cantelmo, and C. Antoniou, "Explaining demand patterns during covid-19 using opportunistic data: A case study of the city of munich," *European Transport Research Review*, vol. 13, Dec. 2021 (cit. on p. 98).

[105] A. Roy and B. Kar, "Characterizing the spread of covid-19 from human mobility patterns and sociodemographic indicators," Oct. 2020 (cit. on p. 98).

[106] G. Pullano, E. Valdano, N. Scarpa, S. Rubrichi, and V. Colizza, "Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the covid-19 epidemic in france under lockdown: A population-based study," *The Lancet Digital Health*, vol. 2, e638–e649, Dec. 2020 (cit. on p. 98).

[107] M. Dahlberg, P.-A. Edin, E. Grönqvist, *et al.*, *Effects of the covid-19 pandemic on population mobility under mild policies: Causal evidence from sweden*, 2020. arXiv: 2004.09087 [econ.GN] (cit. on p. 98).

[108] "Examining the association between socio-demographic composition and covid-19 fatalities in the european region using spatial regression approach," *Sustainable Cities and Society*, vol. 62, p. 102 418, 2020, ISSN: 2210-6707 (cit. on p. 98).

[109] A. Capponi, P. Vitello, C. Fiandrino, *et al.*, "Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities," in *Proc. of IEEE Symposium on Computers and Communications*, 2019, pp. 1–6 (cit. on p. 98).

[110] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989 (cit. on p. 103).

[111] P. Vitello, C. Fiandrino, A. Capponi, P. Klopp, R. D. Connors, and F. Viti, "The impact of sars-covid-19 outbreak on european cities urban mobility," *Frontiers in Future Transportation*, vol. 2, 2021 (cit. on p. 128).

[112] P. Vitello, A. Capponi, C. Fiandrino, G. Cantelmo, and D. Kliazovich, "Mobility-driven and energy-efficient deployment of edge data centers in urban environments," *IEEE Transactions on Sustainable Computing*, vol. 7, no. 4, pp. 736–748, 2022 (cit. on p. 128).

[113] M. Girolami, P. Vitello, A. Capponi, C. Fiandrino, L. Foschini, and P. Bellavista, "A mobility-based deployment strategy for edge data centers," *Journal of Parallel and Distributed Computing*, vol. 164, pp. 133–141, 2022, ISSN: 0743-7315 (cit. on p. 128).

[114] A. Capponi, P. Vitello, C. Fiandrino, *et al.*, "Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities," in *Proc. of IEEE ISCC*, 2019, pp. 1–6 (cit. on p. 128).

[115] P. Vitello, A. Capponi, P. Klopp, R. D. Connors, F. Viti, and C. Fiandrino, "The corona business in modern cities: Poster abstract," New York, NY, USA: Association for Computing Machinery, 2020, ISBN: 9781450375900 (cit. on p. 128).

[116] P. Vitello, A. Capponi, C. Fiandrino, G. Cantelmo, and D. Kliazovich, "The impact of human mobility on edge data center deployment in urban environments," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6 (cit. on p. 128).