UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Urban mobility data analysis in Montevideo, Uruguay

Renzo Massobrio

Programa de Posgrado en Informática–PEDECIBA
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Diciembre de 2018

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Urban mobility data analysis in Montevideo, Uruguay

Renzo Massobrio

Tesis de Maestría presentada al Programa de Posgrado en Informática–PEDECIBA, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magíster en Informática–PEDECIBA.

Director:
Sergio Nesmachnow

Director académico:
Sergio Nesmachnow

Montevideo – Uruguay
Diciembre de 2018

Massobrio, Renzo

Urban mobility data analysis in Montevideo, Uruguay / Renzo Massobrio. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2018.

XII, 93 p.: il.; 29, 7cm.

Director:

Sergio Nesmachnow

Director académico:

Sergio Nesmachnow

Tesis de Maestría – Universidad de la República, Programa en Informática–PEDECIBA, 2018.

Referencias bibliográficas: p. 84 – 93.

1. movilidad urbana, 2. ciudades inteligentes, 3. sistemas inteligentes de transporte, 4. análisis de datos urbanos, 5. matriz origen-destino. I. Nesmachnow, Sergio, . II. Universidad de la República, Programa de Posgrado en Informática–PEDECIBA. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Prof. Antonio Mauttone

Prof. Esteban Meneses

Prof. Ricardo Giesen

Montevideo – Uruguay
Diciembre de 2018

ABSTRACT

Transportation systems play a major role in modern urban contexts, where citizens are expected to travel in order to engage in social and economic activities. Understanding the interaction between citizens and transportation systems is crucial for policy-makers that aim to improve mobility in a city. Within the novel paradigm of smart cities, modern urban transportation systems incorporate technologies that generate huge volumes of data in real time, which can be processed to extract valuable information about the mobility of citizens.

This thesis studies the public transportation system of Montevideo, Uruguay, following an urban data analysis approach. A thorough analysis of the transportation system and its usage is outlined, which combines several sources of urban data. The analyzed data includes the location of each bus of the transportation system as well as every ticket sold using smart cards during 2015, accounting for over 150 GB of raw data. Furthermore, origin-destination matrices, which describe mobility patterns in the city, are generated by processing geolocalized bus ticket sales data. For this purpose, a destination estimation algorithm is implemented following methodologies from the related literature. The computed results are compared to the findings of a recent mobility survey, where the proposed approach arises as a viable alternative to obtain up-to-date mobility information. Finally, a visualization web application is presented, which allows conveying the aggregated information in an intuitive way to stakeholders.

Keywords:
urban mobility, smart cities, intelligent transportation systems, urban data analysis, origin-destination matrix.

List of Figures

List of Tables

Acronyms

AFC Automated Fare Collection 8, 14, 16, 18, 19, 20, 22, 35, 38, 44, 60

APC Automated Passenger Counter 8, 15, 17, 22, 24

AVL Automatic Vehicle Location 8, 15, 18, 19, 22, 35, 38, 60, 65, 72, 82, 83

CDR Call Detail Record 15, 22

CRS Coordinate Reference System 35

CSV Comma Separated Values 68

EDA Exploratory Data Analysis xi, 9, 10, 11, 33, 36, 38

FING Facultad de Ingeniería 34, 35

GIS Geographic Information Systems 11

IM Intendencia de Montevideo 3, 26, 34, 35, 48

INE Instituto Nacional de Estadística 25

INS Inertial Navigation Systems 8

ITS Intelligent Transportation Systems vii, xi, 1, 2, 3, 5, 6, 7, 8, 9, 12, 13, 15, 16, 22, 24, 47, 60, 63, 66, 67, 69, 70, 72, 79, 80, 81, 82, 83

IoT Internet of Things 7

OD Origin-Destination vii, viii, xi, xii, 2, 3, 6, 12, 15, 16, 17, 19, 20, 21, 22, 24, 30, 34, 35, 37, 39, 42, 45, 60, 61, 66, 67, 68, 69, 70, 72, 73, 74, 75, 76, 78, 80, 81, 82, 83

QoS Quality of Service 3, 8, 16, 21, 22, 32, 47, 48, 52, 75, 82

RFID Radio-frequency identification 8

SPMD Single Program Multiple Data 66

STM Sistema de Transporte Metropolitano vi, vii, viii, xi, 2, 3, 25, 26, 29, 30, 29, 30, 35, 36, 38, 39, 40, 41, 43, 44, 45, 47, 49, 51, 52, 55, 58

SVR Support Vector Regression 13

SaaS Software as a Service 67

UBN Unsatisfied Basic Needs vi, 26, 77, 78

kNN k nearest neighbors 13, 14

Contents

Chapter 1

Introduction

Since 1950 populations have been steadily shifting from rural to urban residencies, in a worldwide process known as urbanization (Camagni et al., 2002). Among the multiple challenges that emerge due to this intense and on-going urban expansion process, mobility of citizens constitutes a central issue in modern cities (Cardozo and Rey, 2007). The geographical organization of urban scenarios demands citizens to travel in order to engage in social and economic activities. In this context, public transportation systems play a major role in urban mobility, as they represent the most efficient, sustainable, and socially fair mode of transportation (Grava, 2000). Therefore, understanding the interaction between citizens and public transportation systems is paramount in order to design and implement policies that aim at improving mobility in a city.

Urbanization has taken place along with an increasing incorporation of information and communication technologies in the infrastructure of cities. Modern *smart cities* take advantage of technology to improve urban services (Deakin and Waer, 2011). Urban traffic and transportation systems are generally addressed under the paradigm of smart cities, in what is referred to as *smart mobility* (Benevolo et al., 2016). Related to this concept are Intelligent Transportation Systems (ITS), which make use of technology to develop and enhance transportation. In addition to improving mobility in cities, ITS allow collecting large volumes of urban data (Figueiredo et al., 2001). Large repositories of data offer a unique opportunity to gain valuable insights into the mobility of citizens. In this context, urban data analysis arises as a tool to extract meaningful information from raw urban data to help decision-making

processes in cities.

Understanding the dynamics of mobility within a city is crucial to improve transportation systems. Mobility is described through *Origin-Destination (OD) matrices*, which indicate the number of passengers moving between relevant locations in a city. Traditionally, OD matrices are generated based on surveys or manual passenger counts. However, these methods are very expensive to be carried out regularly, so they offer a partial and outdated view of the mobility patterns in a city (Ortúzar et al., 2011). ITS usually incorporate technology to locate vehicles and to simplify the process of paying for tickets in public transportation systems. Thus, data from these sources can be processed to generate OD matrices. As will be outlined throughout this thesis, the main challenge of this approach resides in accurately estimating the destination of trips, since most ITS do not require any action from passengers alighting the bus.

In 2010 an urban mobility plan was implemented in Montevideo, Uruguay, with the goal of restructuring and modernizing public transportation (Abreu and Vespa, 2010). Under this plan, public transportation in the city was integrated into a unified system named Sistema de Transporte Metropolitano (STM), which incorporates many of the characteristics common to ITS. Buses in STM were equipped with on-board GPS units and ticket selling machines operated with smart cards. These devices represent new sources of urban data, which have a huge potential to help authorities understand mobility in Montevideo.

The main goal of the research reported in this thesis is to take advantage of ITS data in order to characterize mobility patterns of citizens in Montevideo, Uruguay, following an urban data analysis approach. The main contributions of this work are:

1. A thorough review of the related works regarding urban mobility, specifically, on OD matrix generation using ITS data.

2. An urban data analysis of the use of the public transportation system of Montevideo, Uruguay.

3. An algorithm that estimates destinations of trips and generates OD matrices using ticket sales transactions and bus location data.

4. Estimated OD matrices for the public transportation system of Montevideo and their validation against a household mobility survey.

5. A visualization tool to interactively present the computed OD matrices to stakeholders in an intuitive fashion.

The work reported in this thesis resulted in several publications, which address topics included in this manuscript and other related lines of work. A list of these publications, along with a brief description, is presented next.

- Massobrio et al. (2016) presented a framework to assess the Quality of Service (QoS) offered to citizens by a transportation system. A data analysis approach was implemented to compute interesting QoS metrics, such as punctuality of buses, and the case study of Montevideo was addressed.
- Massobrio and Nesmachnow (2016) was a divulgation article presented at a conference held at Intendencia de Montevideo (IM). Partial results from the data analysis process reported in this thesis were included in this article and presented to the transportation authorities of IM during this event.
- Nesmachnow et al. (2017) presented a data analysis that combined mobility data with socioeconomic data. The case study of Montevideo was presented and QoS metrics were compared between areas of the city with different socioeconomic characteristics.
- Fabbiani et al. (2017) introduced a master-slave parallel model to compute OD matrices using ITS data. The model was evaluated using data from STM and the distributed approach allowed significantly reducing the execution times.
- Massobrio et al. (2018) proposed a cloud framework for processing large volumes of urban data. A MapReduce model implemented using the Hadoop framework was proposed and evaluated over a set of case studies based on data from STM.

Additionally, during the research phase of this thesis, a collaboration with a research group at Centro de Investigación Científica y de Educación Superior de Ensenada in Mexico was established. The results from the data analysis process reported in this thesis were used to address bus fleet scheduling and timetable synchronization problems. This collaboration led to a series of co-authored articles on the topic (Peña et al., 2016, 2017a,b, 2018).

The remainder of this thesis is structured as follows. Chapter 2 reviews the main concepts related to urban mobility, smart cities, and ITS; which comprise the theoretical framework of the research. Additionally, urban data analysis is presented as an approach to gain insight from data available in the context of smart cities. Then, Chapter 3 reviews the literature related to urban data analysis and to the problem of generating OD matrices that describe mobility using ITS data. The trip chaining methodology is thoroughly reviewed, since it provides the foundation for the destination estimation algorithm used in this thesis. Afterwards, Chapter 4 presents a study of the transportation system of Montevideo, Uruguay, following an urban data analysis approach. Firstly, the city and the STM transportation system are described. Then, the specific details of the urban data analysis process are outlined. Finally, the results of the analysis of urban data from the STM system in Montevideo are presented through several visualizations and discussed in detail. Later, Chapter 5 outlines the OD matrix estimation process using sales and vehicle location data from STM. The implementation details of the destination estimation algorithm and the OD matrix generation process are presented, and the obtained results are outlined and compared against a household mobility survey. Additionally, an interactive visualization tool is presented, which allows conveying mobility in the city in an intuitive way to stakeholders. Finally, Chapter 6 states the conclusions of the research reported in this thesis and presents the main lines of future work.

Chapter 2

Urban mobility in smart cities

This chapter reviews the main concepts related to the subject of this thesis. Section 2.1 introduces mobility as one of the key issues in urbanized geographic areas. Then, Section 2.2 presents the paradigm of smart cities and ITS as means to address mobility issues in modern cities. Finally, Section 2.3 introduces urban data analysis as an approach to gain valuable information from data available in smart cities.

2.1 Urban mobility

According to United Nations (2018), nowadays more people live in urban rather than rural areas. In 1950, an estimated 30% of the world population lived in towns or cities. Current indicators state that 55% of the population is urbanized, and projections suggest that this trend will continue, reaching an estimated 68% by 2050. The effect of rapid urban expansion has been thoroughly studied (Camagni et al., 2002). In this intense urbanization process, successful management of urban growth becomes crucial to ensure sustainable development at economic, social, and environmental levels.

The geographical organization of urban scenarios, along with the complexity of the activities developed in modern cities, impose serious challenges to the mobility of citizens (Cardozo and Rey, 2007). Facing these challenges requires user-centered policies and regulations that strive for social and spatial justice of citizens (Harvey, 1992). In this context, urban transportation systems play a major role in modern cities (Grava, 2000). The main mobility problems in urban scenarios are related to the inability of transportation systems to satisfy

the needs of a growing number of users.

Understanding the dynamics of mobility within a city is crucial to identify problematic situations related to the daily operation and long-term planning of transportation systems. Mobility is usually described using *OD matrices*, which indicate the amount of trips between relevant locations in a city. A trip is defined as a movement from a location of origin to a location of destination (Ortúzar and Willumsen, 2011). Each trip can have multiple segments, if a passenger makes intermediate stops and transfers between vehicles in order to get to their final destination. Thus, when building OD matrices, the destination of a trip is considered as the final destination of the sequence of segments, where a passenger is assumed to go to perform an activity. Different divisions for the city can be used to analyze mobility at a finer (e.g., specific locations) or coarser grain (e.g., zones). Additionally, OD matrices can be built for specific periods of time to characterize mobility in different days (e.g., working days vs. weekends) or times of the day (e.g., peak vs. non-peak hours).

Traditionally, OD matrices are generated using information from mobility surveys. Ortúzar et al. (2011) reviewed survey methodologies implemented in different cities and emphasized on the importance of collecting mobility data on a continuous basis. Unless performed regularly, surveys offer a partial and outdated view of the mobility patterns of citizens. Additionally, in large cities, where mobility analysis requires detailed zonification and time disaggregation, surveys demand very large sample sizes to compute results with statistical significance. As a consequence, surveys are usually a very expensive mean to characterize urban mobility. Therefore, mobility data repositories in many cities, especially in under-developed countries, are scarce and outdated, due to the lack of human and economic resources to perform large surveys. Taking into account the aforementioned considerations, incorporating technology to help administrations to understand the dynamics of mobility in a city is mandatory, in order to implement solutions that improve the quality of life of citizens.

2.2 Smart cities and ITS

The rapid trend of worldwide urbanization has occurred along with the emergence of information and communication technologies. Consequently, a new

term was coined, which describes the interaction between these two phenomena: *smart cities*. The concept of smart cities embraces several definitions and there is still no definitive consensus in the related literature (Cocchia, 2014; Albino et al., 2015). Among the multiple definitions, the following proposed by Barrionuevo et al. (2012) stands as the most comprehensive one:

> "Being a smart city means using all available technology and resources in an intelligent and coordinated manner to develop urban centers that are at once integrated, habitable and sustainable."

In essence, the paradigm of smart cities proposes taking advantage of information and communication technologies to improve the quality and efficiency of urban services (Deakin and Waer, 2011). Several fields, including public administration, education, health services, public safety, housing, energy, transportation, and logistics, can be improved, interconnected and become more efficient under this paradigm (Washburn and Sindhu, 2009). Modern cities are increasingly becoming sensed and instrumented. The embedding of smart devices into traditional physical systems deployed on cities, together with the emergence of citizen sensors such as mobile phones or Internet of Things (IoT) enabled domestic appliances, currently generate vast volumes of data that present unprecedented opportunities as well as challenges. Extracting insights from the gathered datasets is crucial to improve decision-making processes in cities as well as to achieve quality improvements and increase the efficiency of public services. One of the most addressed areas within the smart cities paradigm is urban mobility. In fact, initiatives aiming at improving transportation and mobility in smart cities are encompassed in the concept of smart mobility (Benevolo et al., 2016).

Related to the topic of smart mobility are ITS. ITS integrate synergistic technologies, computational intelligence, and engineering concepts to develop and improve transportation. ITS are aimed at providing innovative services for transportation and traffic management, with the main goals of improving transportation safety and mobility, while also enhancing productivity (Sussman, 2008). ITS allow gathering large volumes of data by taking advantage of different sensors and devices present in modern vehicles and infrastructure (e.g., passenger counters, GPS devices, video cameras, ticket vending machines) (Figueiredo et al., 2001). Some of the more widely available technologies in urban transportation systems, which are the source of the data that

drives the analysis presented in Chapter 4, are described next.

Automatic Vehicle Location (AVL) systems are a mean for automatically determining and communicating the geographic location of a moving vehicle (Zhao, 1997). The transmitted locations of a fleet of vehicles can be collected at a central server to overview and control the group of vehicles. Due to its widespread availability, low cost, and precision, the most common technology to determine the location of vehicles in AVL is GPS. However, for areas where GPS coverage is poor, this technology is often applied in combination with other methods such as Inertial Navigation Systems (INS) and active Radio-frequency identification (RFID). INS are systems that combine motion and rotation sensors (e.g., accelerometers, gyroscopes) to compute location using *dead reckoning*. Dead reckoning is the process of estimating the current position of a given object based on a known previous location and its estimated velocity and direction. AVL technology is frequently incorporated in ITS and constitutes a rich source of data, as it can help to monitor and control the QoS provided by the transportation system to users.

Most ITS incorporate technology to simplify the process of paying for tickets or fares. Automated Fare Collection (AFC) systems are the stack of components that automate the ticketing system of a public transportation network (Blythe, 2004). With some variations, most AFC systems are comprised of fare media, devices to read/write on to these media, communication technologies, and back office systems. Regarding fare media, contactless smart cards have become the de facto technology in AFC systems. Pelletier et al. (2011) provided a thorough literature review on the use of smart cards in public transportation systems. The review covered the most used technologies, privacy and legal concerns related to these systems, and several applications that use smart card data from public transportation systems. AFC systems generate highly valuable data, which can be processed to extract useful metrics for both day-to-day operation and long-term planning of the transportation system.

Technology has also been integrated in ITS to measure the use of vehicles within the transportation system. Automated Passenger Counters (APCs) are electronic devices that can be incorporated to moving vehicles to record boarding and alighting data (Boyle, 2008). This technology is a major improvement over traditional manual passenger counts or surveys. Several implementations of this concept have been proposed. The most frequently used ones are: i)

incorporating a set of infrared lights in the doorways of vehicles in such way that the order in which the beams are interrupted by a person moving through the door allows determining whether they are entering or exiting the vehicle; and ii) using CCTV cameras in combination with computer vision software to automatically identify and count people. The data generated by these systems allow identifying use patterns by linking boarding and alighting data with stop or station location (Furth et al., 2006).

The development of smart tools that use data gathered by ITS infrastructure and vehicles has risen in the past years. These tools rely on efficient and accurate data processing (even in real-time), which poses an interesting challenge from the technological perspective. Furthermore, ITS data can be combined with more traditional data sources, such as sociodemographic data, that are regularly and systematically collected by government agencies. The methodology for analyzing these sources of urban data in order to gain valuable insights to describe and improve the life of citizens is described next.

2.3 Urban data analysis

Data analysis is the process of collecting and processing raw data to extract meaningful information that provides supporting evidence for conclusions and helps decision-making processes. Multiple definitions and workflows have been proposed to describe the process of data analysis, and techniques under a variety of names have emerged in different fields of knowledge at both academia and industry. Figure 2.1 outlines the data science workflow proposed in Schutt and O'Neil (2013).

The data analysis process has as both, starting and ending points, the current reality. In urban contexts, the analysis starts with collecting raw data from a given city and ends with communicating findings that can potentially help stakeholders to shape the reality of that city to improve the quality of life of its citizens. In between, the data analysis process is comprised of several phases. Firstly, raw collected data must be processed. This phase may include several tasks such as placing data into structures (e.g., tables), inspecting datasets, and cleansing data to detect missing or inaccurate records. After data processing, Exploratory Data Analysis (EDA), which is described next, is performed. This phase may lead to detecting further inaccuracies in the data and potentially requiring further cleansing. After EDA, statistical mod-

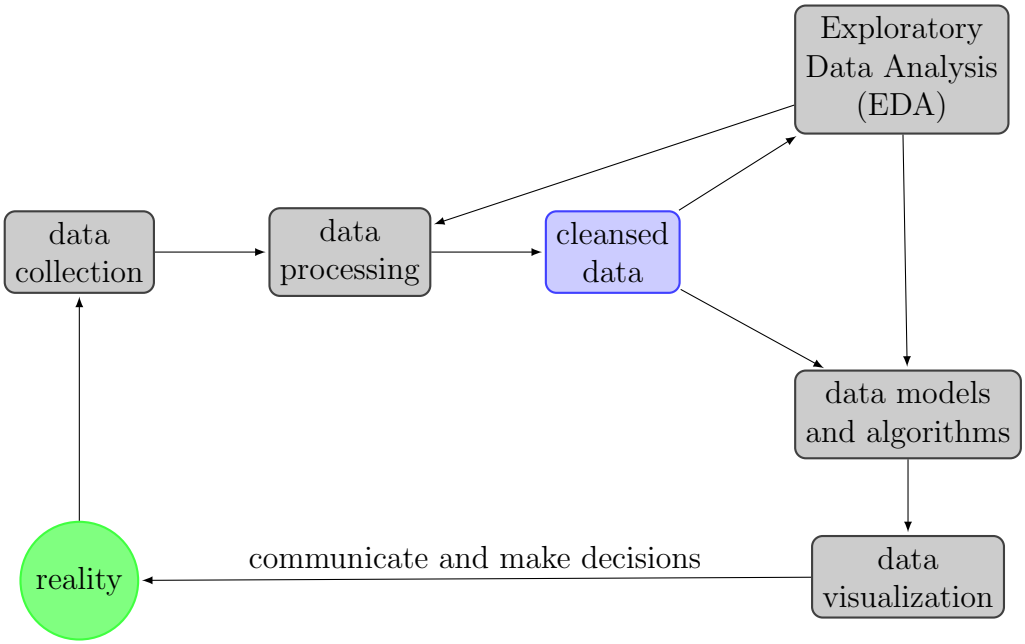


Figure 2.1: Data analysis workflow, based on Schutt and O'Neil (2013)

els and algorithms (e.g., machine learning techniques) are applied to identify relationships between the studied data (Judd et al., 2011). Finally, results are interpreted and communicated, mostly using visualization techniques. When dealing with urban data, effectively communicating results is crucial, thus, the visualization phase is described in more depth in the following paragraphs.

In 1977, Tukey argued that the field of statistics placed too much emphasis in hypothesis testing instead of using data to suggest which hypotheses to test first. As an alternative Tukey introduced EDA, which constituted a major development in statistical theory. In contrast to confirmatory data analysis, where the goal is to build a model to test a defined hypothesis, EDA aims at describing what data can tell beyond the formal modeling and hypothesis testing phase. In this regard, Tukey (1977) stated the following description of EDA:

> "Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there."

There are multiple benefits of doing proper EDA early in the data analysis process, including: gaining intuition about the data, making comparisons between distributions and datasets, performing sanity checks to datasets to

find missing and inaccurate data, and summarizing large sets of data. Since urban data tends to come from a variety of diverse and dynamic sources, EDA becomes mandatory for urban data analysis.

EDA makes an intensive use of data visualization. The main goal of data visualization is to efficiently display measured quantities through graphics. As defined by Tufte (1986), in one of the seminal books on visualization, "At their best, graphics are instruments for reasoning about quantitative information". Traditionally, data visualization techniques were mainly dominated by charts and diagrams comprised of numerical data. However, areas such as urban data analysis, which demand combining quantitative and qualitative data, require more advanced means of visualizing results for effective communication. Since urban data usually has a prevalence of geographic components, urban data visualization combines classic statistical graphics with Geographic Information Systems (GIS).

As a final remark, when performing urban data analysis, several sources of data may be considered. Public entities are among the largest collectors of data (Janssen, 2011). Thus, the willingness of public administrators to open up government data is crucial. Open data is defined as non-confidential data which is made available without restrictions of use or distribution. Several benefits associated with open data have been identified, including a higher return of investment from publicly-funded data, economic growth through stimulation of innovation, and a greater involvement of citizens with their communities (Janssen et al., 2012). Additionally, open data allows citizens to audit public organizations in a more transparent and democratic way, improving their trust in public authorities. In spite of all these benefits, convincing stakeholders to open data is far from trivial. Huijboom and Van den Broek (2011) presented a comparison of open data initiatives in multiple countries and identify the main barriers and challenges for effective open data policy. Overall, the success of open data initiatives heavily relies on citizens properly using available data to generate innovative products that benefit communities, in order to convince authorities to continue opening data. In this regard, the work presented in this thesis intends to contribute with a small step in that direction.

Chapter 3

Related Work

This chapter presents the review of the related literature. Section 3.1 reviews works related to data analysis applied to understand and improve urban mobility. Then, Section 3.2 specifically reviews works that address the problem of generating OD matrices. Later, Section 3.3 outlines the household mobility survey carried out in Montevideo in 2016, which is used to compare and contrast the results arising from the urban data analysis process and from the OD matrix estimation algorithm. Finally, Section 3.4 summarizes the literature review and presents some conclusions of the analysis of related works.

3.1 Urban mobility data analysis

The advantages of using data analysis for social transportation have been studied in a thorough manner in the general review of the field developed by Zheng et al. (2016). The authors discussed the use of several sources of information, including vehicle mobility (e.g., GPS coordinates, speed data), pedestrian mobility (e.g., GPS and WiFi signals from mobile devices), incident reports, social networking (e.g., textual posts, user location), and web logs (e.g., user identification, comments). In the review, the advantages and limitations of using each source of data were discussed. Several other novel ideas to improve public transportation and implement the ITS paradigm were also reviewed, including applying *crowdsourcing* techniques for collecting and analyzing real-time or near real-time traffic information, and using *data-based agents* for driver assistance and human behavior analysis. Finally, a data-driven social transportation system that integrates all the previous concepts and improves traffic

safety and efficiency was proposed.

Several proposals that combine computational intelligence techniques with urban data analysis have been recently applied to process ITS data in order to help decision-making processes in smart cities.

Chen et al. (2014) proposed a model that aims at efficiently predicting traffic speed on a given location using historical data from various sources including ITS data, weather conditions, and special events taking place in a city. To obtain accurate results, the prediction model needs to be re-trained frequently in order to incorporate up-to-date data. The proposed prediction model combines the k nearest neighbors (kNN) algorithm with a Gaussian Process Regression. Additionally, the results are computed using a Map-Reduce model, implemented under the Hadoop framework. The experimental evaluation was performed over a real scenario using data from Research Data Exchange, a platform for ITS data sharing (`www.its-rde.net`). The data used corresponded to the Interstate 5 Highway in San Diego, California, United States, and included speed, flow, and occupancy data measured using loop-detectors on the road, as well as visibility data taken from weather stations nearby. Experimental results showed that the proposed method was able to accurately predict traffic speed with an average forecasting error smaller than 2 miles per hour. Additionally, a 69% improvement on the execution time was achieved by using the Hadoop framework executing in a cluster infrastructure when compared with a sequential algorithm executing in a single machine.

Shi and Abdel-Aty (2015) applied the random forest data mining technique and Bayesian inference to process large volumes of data from a microwave vehicle detection system, with the main goal of identifying in real-time the contributing factors to crashes. Rear-end crashes were studied because they have a straightforward relation with congestion. The experimental evaluation of the proposed computational intelligence approach was performed considering traffic data from State Routes 408, 417, and 528 in Central Florida, United States. A reliability model was also included in the analysis. The main results allowed the authors to conclude that peak hour, higher volume and lower speed at upstream locations, and high congestion indices at downstream detection points significantly increased the probability of crashes.

Ahn et al. (2016) applied Support Vector Regression (SVR) and a Bayesian classifier for building a real-time traffic flow prediction system. Data preparation and noise filtering were applied to raw data, and a traffic flow model

was proposed using a Bayesian framework. Regression techniques were used to model the time-space dependencies and relationships between roads. The performance of the proposed method was studied on traffic data from Gyeongbu, the Seoul-Busan corridor in South Korea. The experimental results showed that the approach using SVR-based estimation outperformed traditional linear regression methods in terms of accuracy.

Xia et al. (2016) studied the real-time short-term traffic flow forecasting problem. To solve the problem, the kNN algorithm was applied in a distributed environment, following the Map-Reduce model using the Hadoop framework. The proposed solution considered the spatial-temporal correlation in traffic flow, i.e., current traffic at a certain road segment depends on past traffic (time dimension) and on the traffic situation at nearby road segments (spatial dimension). These two factors can be controlled using weights in the proposed algorithm. The experimental analysis was performed using data of trajectories obtained from more than 12000 GPS-equipped taxis in the city of Beijing, China, during a period of 15 days in November 2012. The first 14 days of data were used as the training set and the last day was used for evaluating the computed results. The proposed algorithm allowed reducing the mean absolute percentage error by up to 11.5% over three existing techniques based on the kNN algorithm. Additionally, the proposed solution achieved a computational efficiency of 0.84 in the best case.

Effectively analyzing urban datasets is complex since data usually contain both geographical and temporal components, in addition to multiple variables associated to the specific source. Ferreira et al. (2013) developed a software tool for visual exploration of large datasets of urban data. The developed tool allows data scientists to dynamically explore data variation in space and time, analyze specific events at a given time, and identify patterns across different regions of a city. In order to enhance usability, the software allows specifying the parameters of the data query visually, by setting spatial and temporal constraints interactively over a graphic user interface with a map of the studied area. The capabilities of the developed solution were demonstrated by performing different analysis over a large set of taxi trajectories from New York.

More related to the data analysis research included in this thesis, many works have studied urban mobility using smart card data from AFC in public transportation systems.

Bagchi and White (2005) discussed the role of smart card data for travel behavior analysis. Two datasets from the transportation systems of Southport, Merseyside and Bradford in England were used, which accounted for nearly 3500 cardholders. The authors performed a simple study focused on the average number of trips and transfers made by passengers. Additionally, the turnover rates were analyzed, to identify the number of active users in the system. The research concluded that smart card data allow obtaining much larger samples than surveys to characterize transportation systems. However, certain information (e.g., purpose of traveling) cannot be inferred from these data. Thus, the authors conclude that smart card transactions are not an alternative to traditional data collection methods, but a useful complementary source of data.

Utsunomiya et al. (2006) studied access and usage patterns of passengers in the transportation system of Chicago, US. Firstly, the authors discussed the analysis that can be performed using smart card sign-ups and transactions data, identifying the major issues encountered as well as general recommendations. The potential uses for smart card data were classified in several categories, including service planning, demand forecasting, pricing and fare policy definition, and market research. A dataset corresponding to seven days of recorded transactions was studied to analyze walking access distances, frequency of daily travel patterns, and passenger behavior by residential area. During the data analysis, the more frequent errors were due to missing transactions and incorrect bus route identification. In order to deal with these inconsistencies, the authors proposed combining smart card data with passenger counts and vehicle location from APC and AVL systems.

3.2 OD matrices generation

The estimation of OD matrices is a well-known problem in the field of public transportation. This problem has had a renewed interest with the increasing availability of large volumes of data from modern ITS systems and other sources. Several works have proposed generating OD matrices for urban transportation systems using a variety of data sources. Some authors have used APC systems to estimate OD flows from detailed boarding and alighting counts (Furth et al., 2006; Lu, 2008). However, since entering and exiting data cannot be assigned to individual passengers, most of the proposed mod-

els require some previously-computed OD matrix as a baseline, which is then expanded using the passenger count data. Other approaches use Call Detail Record (CDR) data from mobile phones. This is an extended method applied to building OD matrices for general road transit analysis (i.e., considering all modes of transportation). However, in order to limit the analysis to public transportation systems, it is necessary to either infer the transportation mode (Wang et al., 2010; Doyle et al., 2011) or combine CDR with data from ITS (Anda et al., 2017). Other works have proposed using Bluetooth antennas to detect when mobile devices enter and exit vehicles (Kostakos et al., 2010). This information, combined with data from AVL systems, can be used to estimate OD matrices. However, antennas must be installed on vehicles to detect on-board devices, requring an initial investment to deploy the necessary infrastructure. Furthermore, noise from mobile devices outside the vehicles must be filtered for accurate passenger sensing. Filtering external signals is not a simple task and small false-positive passenger counts can rapidly accumulate, impacting on the overall accuracy of the estimated OD matrices.

Despite the variety of sources that have been used to estimate OD matrices in urban transportation systems, the majority of the literature focuses on methods that involve using smart card data arising from AFC systems. Since this is the approach followed in this thesis, a review of the main related works regarding OD matrix generation using smart card data is presented next.

An analysis of the literature about using smart cards in ITS was presented by Pelletier et al. (2011). The review covered all the details about hardware and software needed for deploying smart card payment solutions in urban transportation systems. In addition, privacy and legal issues that arise when dealing with smart card data were also reviewed. Several examples of using smart card data to improve transportation systems were described. The studied use cases were grouped in three categories: strategic level, tactical level, and operational level. The strategic level refers to long-term planning of the transportation networks. The reviewed examples in this category focus on understanding user behavior through data analysis, as an alternative to traditional methods. On the tactical level, most of the revised work focus on taking advantage of the estimated mobility patterns to optimize vehicle schedules to improve the QoS offered to passengers. Finally, on the operational level, most works focus on auditing the transportation system by checking for timetable adherence, fare evasion, and employee mistakes.

Most AFC systems require that passengers validate their smart cards when boarding but not when alighting the bus. Thus, the origin of a trip can be accurately determined but the destination must be inferred. Li et al. (2018) presented an up-to-date survey of the literature related to destination estimation techniques for OD matrix generation using smart card data. The survey reviewed 20 articles after an initial selection of 984 published papers accessed through different databases. Three models for estimating destination based on smart card data were identified in the reviewed works: the trip chaining model, the probability model, and the deep learning model. The trip chaining model, which is the one applied in this thesis, is discussed in depth in the remainder of this section. The probability model computes the alighting probability based on the traveled distance and the number of passengers on board. The main disadvantage of this model is that it infers the total number of passengers boarding and alighting at each bus stop, as opposed to the trip chaining method which can analyze boarding and alighting of each specific passenger. The third model, based on deep learning, requires both boarding and alighting data for training, which makes it more suitable to railway or subway transportation systems where passengers are required to validate their smart cards both to enter and exit stations. The most relevant works on OD estimation based on the trip chaining method are reviewed next.

The trip chaining model for destination estimation was originally proposed by Barry et al. (2002). The model proposes inferring destinations by looking at the history of trips of each cardholder. Two hypotheses are considered: *i*) the origin of a new trip is the destination of the previous one; and *ii*) at the end of the day, users return to the origin of their first trip of the day. The authors considered data from a travel survey to backup the validity of both assumptions. The proposed model was applied to the subway system of New York, where nearly 80% of riders use smart cards. The computed OD matrix was validated using station exit counts at different times of the day and using peak load passenger volume data and a trip assignment model. The authors estimated that 90% of destinations can be accurately inferred for a 78% share of the total number of subway users.

Trépanier et al. (2007) proposed using the trip chaining model for estimating the destination for passengers boarding buses with smart cards, following a database programming approach. Based on the same two assumptions than Barry et al. (2002), the proposed approach follows the chain of trips of

each user in the system. Those trips for which chaining is not possible (e.g., only one trip in the day exists for a particular user) are compared with all other trips of the month for the same user, in order to find similar trips with known destination. The experimental evaluation was conducted using real information from the transit authority in Gatineau, Quebec. Two datasets were used, with 378,260 trips from July 2003 and 771,239 trips from October 2003. Results showed that the destination estimation was accurate for 66% of the trips. Most of the trips for which destination could not be estimated with the proposed approach took place during off-peak hours, where more atypical and non-regular trips are performed. Considering only peak hours, the percentage of trips with their destination estimated improved to 80%. However, the real estimation accuracy could not be assessed due to the lack of a second source of data (e.g., surveys, APC) for comparison.

Farzin (2008) applied the trip chaining method to data from the AVL and AFC infrastructures of the transportation system in São Paulo, Brazil. Farzin faced an additional challenge: the AVL and AFC systems studied were independent, thus, the location of each transaction was not directly recorded. Consequently, an additional step in the trip chaining algorithm was required to search for the most recent record in the AVL for each transaction, in order to find the corresponding bus stop of each ticket sale. The studied dataset accounted for only 8% of the total bus trips in the city, since it was mostly concentrated in one particular area. The computed results were compared to the findings of a household survey and the proposed approach arose as a viable alternative to understand mobility in the city. However, the household survey was performed 11 years before the comparison, so the conclusions are, at least, questionable.

Later, Wang et al. (2011) proposed using the trip chaining method to infer bus passenger origin-destination from smart card transactions and AVL data from London, United Kingdom. Origins were accurately determined by searching for the timestamp of each smart card transaction in the AVL records to assign each transaction to a bus stop. To estimate destinations, the authors used a similar methodology to the one presented by Trépanier et al. (2007), chaining trips when possible to infer destinations. Results were compared against the passenger intercept survey of Transport for London, which is performed every five to seven years for each bus route and includes the number of people boarding and alighting at each bus stop (Transport for London, 2018).

The analysis showed that destinations could be estimated for nearly 57% of all trips. When compared to the survey, the difference on the estimated destinations were below 4% on the worst case. Finally, two practical applications of the results were presented. The first one consisted of studying the daily load/flow variation in order to identify locations along each bus route where passenger load is particularly high, as well as underutilized route segments. The second application consisted of a transfer time analysis, evaluating the average time that users need to wait for transferring between buses, based on the alighting stop and the AVL data.

More recently, Munizaga and Palma (2012) presented a similar approach to the one applied by Wang et al. (2011) for estimating OD matrices in the multimodal transportation system of Santiago, Chile. The scenario considered by Munizaga et al. is more general than other previous works, since passengers can use their smart cards to pay for tickets at metros, buses, and bus stations. The proposed approach was evaluated using smart card datasets corresponding to two different weeks, with over 35 million transactions each. The origin of the trip was accurately determined for nearly every transaction and the destination and time of alighting was estimated for over 80% of the transactions. After extrapolating and post-processing, an estimated OD matrix was presented to visualize the computed results at any given time-space disaggregation. Later, the authors extended their work by validating the main assumptions of the model (Munizaga et al., 2014). The estimated OD matrices were validated using an endogenous validation (i.e., using the same data used to build the OD matrices), comparing to a detailed OD survey with a sample size of 300,000 users, and by performing personal interviews to a small sample of passengers. The authors concluded that the proposed model is highly reliable, accurately estimating 84.2% of the inferred destinations.

Some of the reviewed works deal with an additional problem: origin estimation. This issue arises in transportation systems having AFC but not AVL infrastructures. Most surveyed works that deal with this additional challenge are related to transportation systems present in China (Li et al., 2011; Ma et al., 2012). In this problem variant, boarding location must be inferred based on the timestamp of the transaction, the bus line identifier, and using timetables. The generated OD matrices in these scenarios are less reliable, since more assumptions need to be made (e.g., all buses are assumed to have a synchronous clock in order to compare timestamps of transactions, drivers are

assumed to strictly adhere to timetables). The case study addressed in this thesis does not suffer from this problem, as it uses data from the transportation system of Montevideo, Uruguay, which has both AVL and AFC infrastructure.

Alsger et al. (2015) proposed an interesting study to validate the key assumptions of the trip chaining model. The authors used smart card data from the transportation system of South East Queensland, Australia, which consists of bus, train, and ferry networks that share the same AFC infrastructure. The dataset accounted for one week of smart card transactions of 260,803 cardholders, totaling 628,479 transactions. The peculiarity of this dataset is that it contains both origin and destination records, since passengers are required to validate their smart cards when boarding and alighting. Therefore, the authors were able to study different variants of the trip chaining method and compare the resulting OD matrices against the real data from AFC records. The study focused on validating the following aspects of the trip chaining model: i) transfer time thresholds; ii) transfer walking distances; and iii) the hypothesis that passengers return to their first origin at the end of the day. For the first aspect, different allowed transfer time thresholds proposed in the literature were studied. Results showed that the best estimations were achieved when considering that consecutive trips within a 60 minutes timeframe correspond to a transfer, which matches with the regulated allowed transfer time of the transportation system. For the second aspect, several walking distances were considered and the best results were computed when using 800 m as the maximum distance a passenger is expected to walk between consecutive legs of a transfer. Both aspects needed to be addressed since the AFC infrastructure of the studied dataset did not record transfer activity directly. This is not the case for the dataset used in the research reported in this thesis, which accurately indicates whether a given transaction corresponds to a transfer or to a direct trip. Regarding the last aspect, results showed that for nearly 88% of the passengers the last destination of the day was within a walkable distance of their first origin, thus validating one of the key assumptions of the trip chaining model.

Nassir et al. (2015) used smart card data from the same transportation system than Alsger et al. (2015) for activity detection. The authors argued that a common assumption of the trip chaining model is that trips including bus transfers are always considered as multi-legged trips when, in reality, some correspond to separate trips done for specific purposes. Thus, the authors propose an activity detection method to distinguish between "real transfers"

(i.e., when a transfer is actually needed to reach the final destination) from short/hidden activities (e.g., quick shopping, picking up kids from school). The activity detection method uses vehicle schedules and optimal path calculations to separate activities from transfers. The proposed model was evaluated using data from a household travel survey. Results validated the proposed model and emphasized the importance of activity detection when building OD matrices using smart card data.

After OD matrices are estimated, several interesting metrics can be computed to characterize the transportation system and the behavior of users. Regarding transportation systems, Trépanier et al. (2009) proposed using smart card data to compute several metrics to assess the QoS offered to citizens. Statistics regarding the performance of the network (e.g., operating speed, distance per vehicle) and regarding passenger service (e.g., traveled distances, traveled times, average trip length) were computed at multiple spatial and temporal resolutions. Regarding behavior of users, Ma et al. (2013) proposed using clustering algorithms to identify travel patterns regularities. By identifying periodic travel patterns among users, authorities might be able to evaluate the impact of potential changes to the transportation network, measure transit performance, and target marketing campaigns more accurately.

Finally, some works in the literature addressed the problem of visualizing OD matrices (Boyandin et al., 2011; Guo and Zhu, 2014). This is a problem that arises not only when analyzing passenger flows in transportation systems, but in many other areas which model flows of people, goods, animals, network packets, etc., among different locations. The major challenges faced are visualizing large amounts of data effectively and including temporal analysis within the OD visualizations. To address these challenges, density estimation, normalization, and smoothing techniques are applied to OD flows in order to generate visually legible maps.

3.3 Mobility survey in Montevideo

In 2016, a metropolitan household survey was conducted in Montevideo, Uruguay, to update the mobility information, which dated back to 2009. Mauttone and Hernández (2017) outlined the methodology used to carry out the survey as well as the main findings. The survey aimed at characterizing the mobility in the city, considering all modes of transportation and also com-

prising the metropolitan area, which includes towns and villages outside of Montevideo. Face-to-face interviews were carried out during working days from August to October 2016 in 2230 households to 5946 individuals. Regarding mobility, the survey encompassed every trip done by each interviewed individual between 4.00 a.m. on the previous day of the interview to 4.00 a.m. on the same day of the interview. For each trip, mode of transportation, time and place of origin and destination, and information on each leg of the trip was recorded. Additionally, general questions about mobility habits and perceptions on the QoS offered by the public transportation service were inquired. Besides mobility, the survey included several socioeconomic indicators, e.g., education level, employment status, income, building quality of the household. Thanks to the design of the sample in the survey, the results were later extrapolated to represent each considered zone by applying a series of expansion factors and taking into account the population of each defined zone. The main findings of the urban data analysis described in Chapter 4 and from the OD matrix estimation process described in Chapter 5 are compared to those obtained from the mobility survey.

3.4 Summary

Table 3.1 summarizes the related works included in the literature review, including a brief comment on each reviewed work.

The analysis of related works allows identifying several proposals for using data analysis in the context of ITS to understand and improve urban mobility. Urban data analysis combined with computational intelligence and learning methods are often used to identify traffic patterns using a variety of data sources and to provide useful information for planning.

Regarding OD matrices estimation, several works have addressed the problem of estimating the destination by chaining consecutive trips under certain assumptions. This thesis expands the original trip chaining method proposed by Barry et al. (2002) by also considering transfers between bus lines, which are specifically recorded in the smart card dataset from Montevideo, Uruguay, used for the evaluation. Additionally, many works that do consider transfers, only take into account those made within the same bus stop. The transfer analysis methodology reported in this thesis extends that idea, since the transportation system in Montevideo allows transfers between any lines at any

Table 3.1: Summary of the related works included in the literature review

Urban mobility data analysis	
reference	*comment*
Zheng et al. (2016)	Reviewed the advantages of using data analysis for social transportation.
Chen et al. (2014)	Predicted traffic speed using historical data from various sources.
Shi and Abdel-Aty (2015)	Processed large volumes of data from a vehicle detection system to identify the contributing factors to crashes in real-time.
Ahn et al. (2016)	Built a real-time traffic flow prediction system.
Xia et al. (2016)	Studied the real-time short-term traffic flow forecasting problem.
Ferreira et al. (2013)	Developed a software tool to visualize large volumes of urban data.
Bagchi and White (2005)	Discussed the role of smart card data for travel behavior analysis.
Utsunomiya et al. (2006)	Studied access and usage patterns of passengers in the public transportation system of Chicago, US.
OD matrix estimation	
reference	*comment*
Furth et al. (2006) Lu (2008)	Estimated OD matrices using passenger counts from APC systems.
Wang et al. (2010) Doyle et al. (2011)	Proposed using CDR data from mobile phones to build OD matrices.
Anda et al. (2017)	Argued that several data sources must be combined to accurately model mobility in urban scenarios.
Kostakos et al. (2010)	Proposed using Bluetooth antennas to detect on-board mobile devices.
Pelletier et al. (2011)	Reviewed the literature on the use of smart cards in ITS.
Li et al. (2018)	Reviewed the literature on using smart card data for OD matrix estimation.
Barry et al. (2002)	Proposed the trip chaining method for destination estimation and applied it to data from the New York City subway.
Trépanier et al. (2007)	Applied trip chaining to data from Gatineau, Quebec. No comparison made due to the unavailability of a second source of mobility data.
Farzin (2008)	Applied trip chaining in São Paulo, Brazil and compared the results to a household mobility survey.
Wang et al. (2011)	Applied trip chaining to data from London and compared the results to a passenger intercept mobility survey.
Munizaga and Palma (2012) Munizaga et al. (2014)	Applied trip chaining to a multimodal transportation system in Santiago, Chile comprised of buses and metros and validated the key assumptions of the model.
Li et al. (2011) Ma et al. (2012)	Addressed the origin estimation problem in transportation systems with AFC but not AVL.
Alsger et al. (2015)	Validated key assumptions of the trip chaining model using data from an AFC system in Australia with both boarding and alighting information.
Nassir et al. (2015)	Proposed a model for identifying short/hidden activities masked as transfers in the trip chaining method.
Trépanier et al. (2009)	Used smart card data to compute statistics of the QoS offered to passengers.
Ma et al. (2013)	Identified travel patterns regularities using clustering algorithms.
Boyandin et al. (2011) Guo and Zhu (2014)	Addressed effective OD visualization in flow maps.
Mauttone and Hernández (2017)	Outlined the methodology and main findings of the 2016 mobility survey in Montevideo.

bus stop. Thus, the OD matrix estimation algorithm considers trips that may include several intermediate transfers as well as walks between bus stops to do those transfers.

In order to compare the OD matrices generated using smart card data, most authors use existing mobility surveys (passenger intercept and household surveys) or passengers counts from APC systems. In this thesis, the estimated OD matrices generated using smart card data are compared against the results of the household mobility survey performed in 2016 (Mauttone and Hernández, 2017).

The main contribution of this thesis is to apply the existing knowledge in the literature regarding urban data analysis and OD matrix generation to the transportation system of Montevideo, Uruguay. In this regard, no previous works using ITS data to understand and improve urban mobility in Montevideo, Uruguay, were found in the analysis of the related literature. Therefore, the research reported in this thesis contributes with a novel proposal to assess the transportation system and understand mobility patterns in Montevideo, Uruguay.

Chapter 4

Urban data analysis in Montevideo, Uruguay

This chapter presents a study of the transportation system of Montevideo, Uruguay, following an urban data analysis approach. Section 4.1 introduces the case study, describing Montevideo and its public transportation system. Then, the urban data analysis process and implementation details are described in Section 4.2. Finally, Section 4.3 outlines and discusses the main findings of the analysis related to describing the use of the transportation system and presents practical use cases for the information obtained through the data analysis process.

4.1 Overview of the case study

This section presents an overview of the case study considered in the urban data analysis process. Section 4.1.1 describes Montevideo, Uruguay, including geographic, demographic, administrative, and socioeconomic information. Then, Section 4.1.2 introduces STM, the public transportation system in Montevideo.

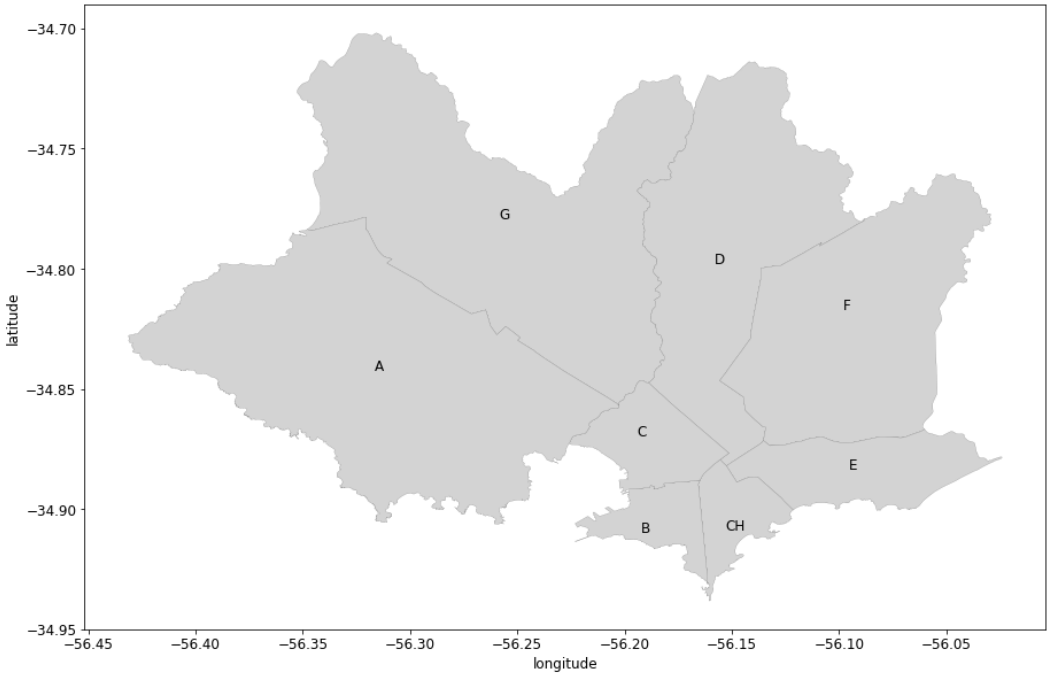
4.1.1 Montevideo, Uruguay

Montevideo is one of the nineteen departments in Uruguay. The capital city of the department, also named Montevideo, is the capital city for the country as well. Due to sharing the same name, the city of Montevideo is often confused with the department of Montevideo. For the remainder of the docu-

ment, except when explicitly stated, Montevideo refers to the department and not to the city, since the transportation system covers the whole department. Located in the southernmost part of the country, Montevideo extends to an area of only 530 km^2. From an administrative point of view, Montevideo is comprised of eight municipalities. Figure 4.1a shows a map of the municipalities that comprise Montevideo, generated using data available at Servicio de Geomática - Intendencia de Montevideo (2011). A finer-grain division, mostly used in census and surveys, is defined by Instituto Nacional de Estadística (INE). This division separates Montevideo into 1063 zones named *census segments*. Figure 4.1b shows a map of Montevideo and its division into census segments based on data from Servicio de Geomática - Intendencia de Montevideo (2014b). Both administrative divisions are referenced throughout the remainder of the document, as they constitute different units of analysis for the urban data studied.

In spite of accounting for only 0.3% of the total surface of Uruguay, Montevideo has an estimated population of 1.319.108, which represents nearly 40% of the total population of the country (Instituto Nacional de Estadística, Uruguay, 2012). The population of Montevideo is unevenly distributed over its small area, with high population densities near the coastline bordering the Río de la Plata estuary. Figure 4.2 shows a choropleth map of the population density in Montevideo, according to data provided by Servicio de Geomática - Intendencia de Montevideo (2014b).

Describing the population of Montevideo from a socioeconomic point of view is not a simple task and is out of the scope of this thesis. However, a broad picture of the social reality can be obtained by studying Unsatisfied Basic Needs (UBN). The UBN methodology aims at identifying the lack of goods or services (or critical problems accessing them) which prevent citizens from exercising their social rights. Figure 4.3 shows a choropleth map of Montevideo indicating the percentage of households with UBN, as defined by Calvo (2012). The map was generated using data available at Servicio de Geomática - Intendencia de Montevideo (2014a). It is clear that the most vulnerable citizens are located farther away from the coast and the city center, in sparsely populated areas.



(a) municipalities

(b) census segments

Figure 4.1: Administrative divisions of Montevideo, Uruguay

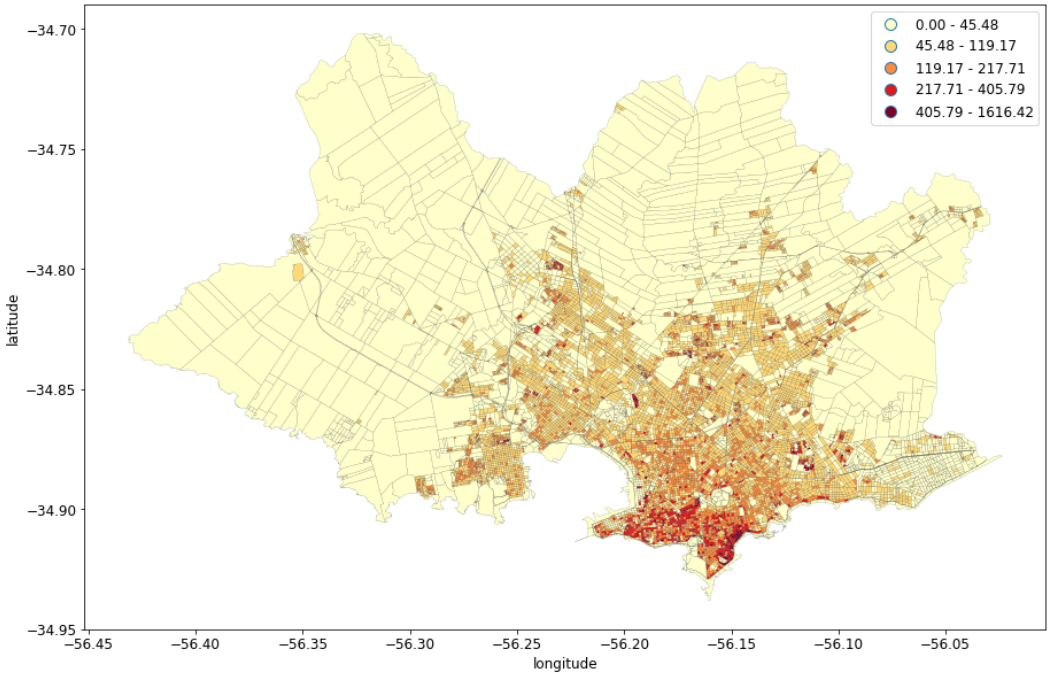


Figure 4.2: Population density in Montevideo, Uruguay (inhabitants per ha)

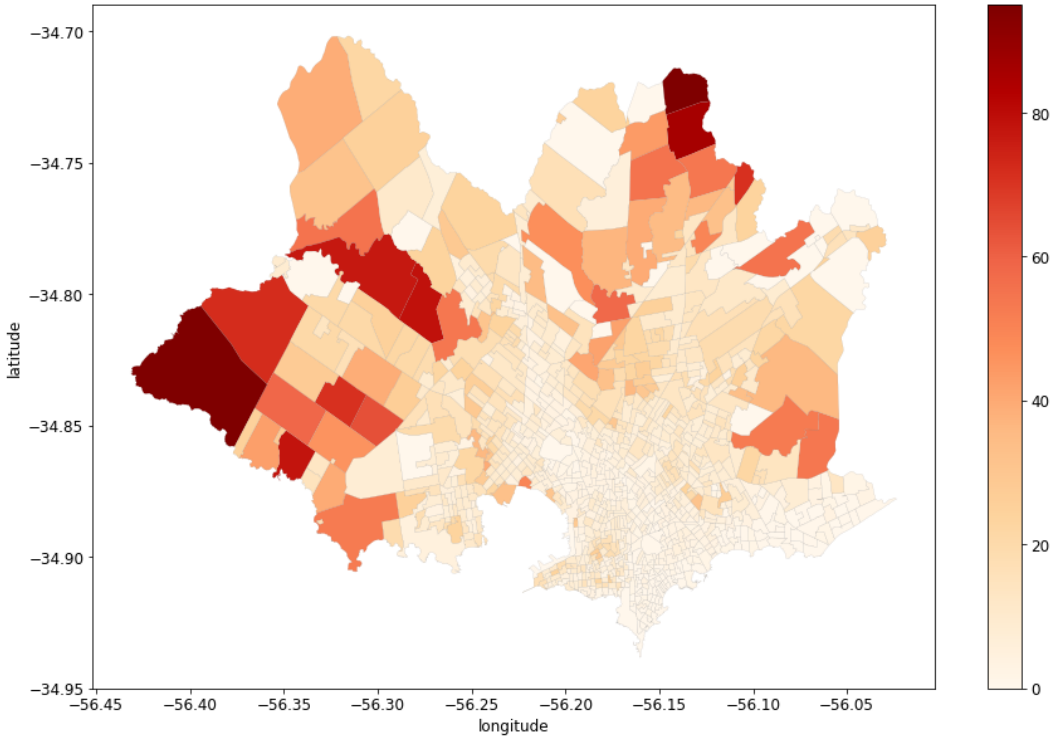


Figure 4.3: Percentage of households with UBN in Montevideo, Uruguay

4.1.2 The public transportation system

The authorities at IM proposed in 2010 an urban mobility plan with the goal of restructuring and modernizing public transportation (Abreu and Vespa, 2010). Within this plan, public transportation in Montevideo was integrated into a unified system named STM, which is comprised of 1528 buses operated by four private companies.

The bus network consists of 145 bus lines. However, each bus line usually has different variants, accounting for outward and return trips, as well as shorter versions of the same line. The total amount of bus lines when considering each variant individually is 1383. This amount seems to be remarkably large, especially when compared to the total number of buses available in STM. Figure 4.4 shows the bus lines that comprise STM, on top of a road map, according to data provided by Servicio de Geomática - Intendencia de Montevideo (1996, 2012a). It is clearly noticeable that the city center acts as a centrality in the bus network, with most lines converging to that area. Additionally, the large length of certain bus lines with respect to the area of Montevideo is also noteworthy. The average bus line length is 16.7 km (standard deviation 7.1) and the median length is 16.4 km, with the longest line spreading over 39.6 km. Intuitively, these figures strike as remarkably large, considering that the total area of Montevideo is 530 km^2 and can be circumscribed to a rectangle of 26×37 km. Furthermore, if only the two upper quartiles of census segments according to their population density are considered (i.e., census segments with a population density larger than the median), it can be stated that the most urbanized area extends only to 75.2 km^2, representing 14% of the total area of Montevideo.

The bus network is comprised of 4718 bus stops. Figure 4.5a shows a map with the location of bus stops, while Figure 4.5b shows a detailed view of the bus stops located in the city center. Bus stop location data correspond to data available from Servicio de Geomática - Intendencia de Montevideo (2012b). The density of bus stops in the city center is noteworthy, with more than one bus stop per block in some of the main avenues. This fact is consistent with the previous observation about the central role that this area of the city plays within the bus network.

With the creation of STM, fares were redefined to provide passengers with more flexibility when traveling. Firstly, smart cards were introduced to allow

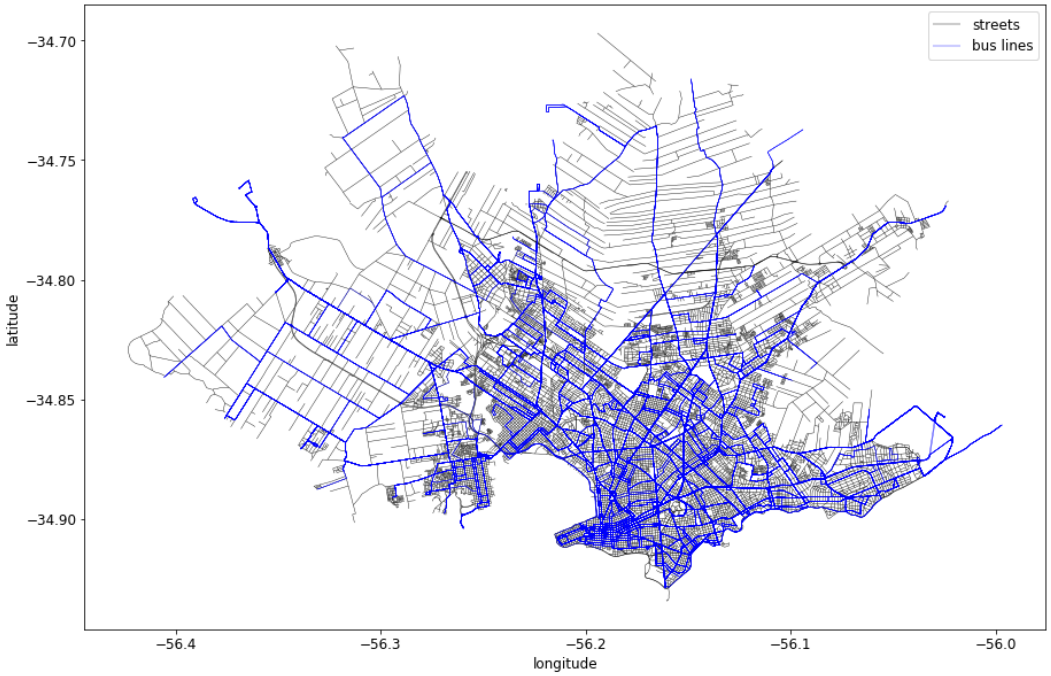
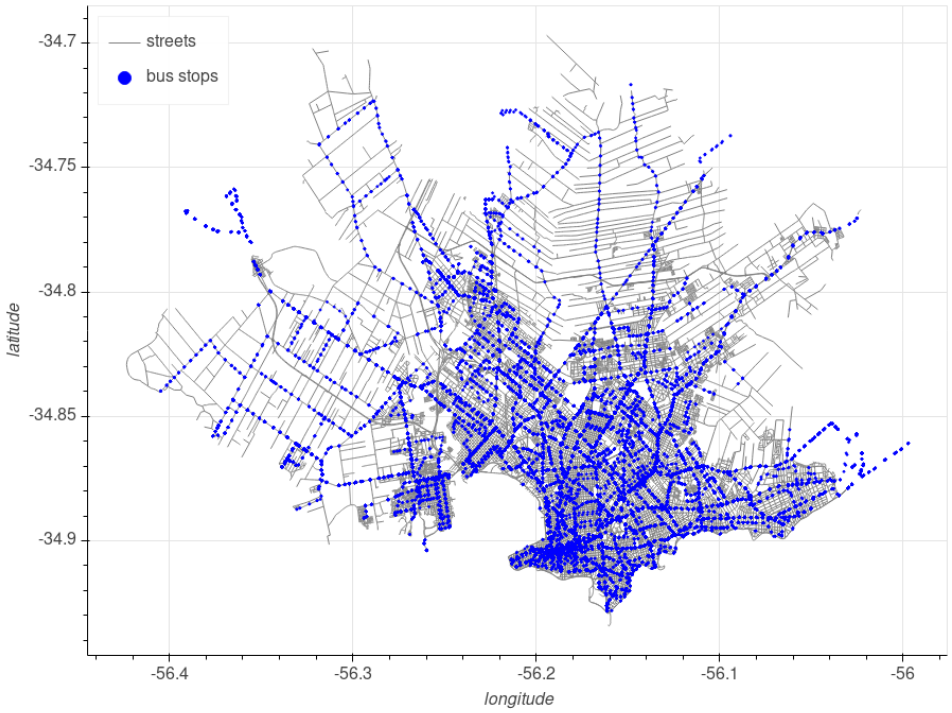


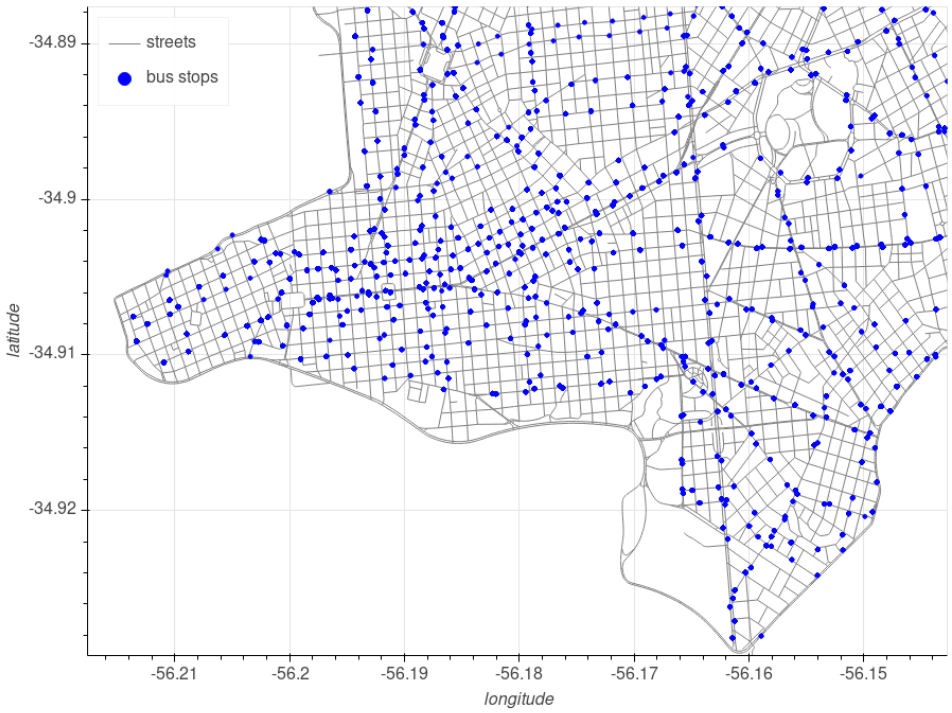
Figure 4.4: Bus lines in STM

passengers to pay for tickets without using physical money. STM smart cards are contact-less top-up cards which are linked to the identity of the owner (a valid government ID or passport is required to get one). Passengers traveling with STM cards can buy different types of tickets: regular tickets, preferential tickets (for certain bus lines with better vehicles and faster routes), local tickets (to travel within certain areas of the city), and a city-center ticket (to travel within the city center). Two different types of bus tickets exist which allow bus transfers, named *one-hour* and *two-hours* tickets. One-hour tickets allow boarding up to two buses within an hour, while two-hours tickets grant unlimited bus transfers within a period of two hours. This fare scheme supports transfers between any bus line at any bus stop. In practice, this means that a passenger can even make an outward and return trip in the same line, as long as the boarding time of the second bus is within the validity period of the ticket.

Using the STM card is straightforward: passengers indicate the type of ticket desired to the driver and approach their smart cards to the terminal, which prints out the corresponding ticket. For consecutive trips included in the valid period of time of the ticket, the user simply approaches the STM card to the terminal, which signals the validity of the ticket without printing



(a) Montevideo



(b) Montevideo city center

Figure 4.5: Bus stops location in STM

a new proof of purchase. Passengers do not validate their STM cards when alighting a bus. While this is practical for passengers, it constitutes one of the main challenges for building OD matrices, as discussed in Chapter 5. Cash payments are also allowed for users without STM cards, however only single trips can be purchased (i.e., no transfers are permitted). Authorities have recently taken measures to encourage citizens to pay using STM cards, e.g., including a price surcharge for cash payments.

4.2 Urban data analysis process

This section describes the urban data analysis process performed with the goal of characterizing how citizens of Montevideo use the public transportation system. The analysis followed the workflow outlined in Section 2.3. The data analysis reported in this section focused on the interaction of passengers and the transportation system. Additionally, some related lines of work were explored during the research process of this thesis. The QoS offered to citizens by the transportation system, according to the punctuality of buses, was studied using bus location data. The proposed solution and the main findings of this study are reported in Massobrio et al. (2016). Furthermore, a solution for processing large volumes of urban data in the cloud was implemented and is described in Massobrio et al. (2018). Finally, Nesmachnow et al. (2017) presents an urban data analysis of the public transportation system combined with socioeconomic data.

4.2.1 Computing infrastructure

This section describes the computing infrastructure used during the urban data analysis process. A description of the specific software packages used for the implementation is presented, along with the main characteristics of the hardware platform where the analysis was performed.

Software infrastructure

The majority of the urban data analysis reported in this thesis was performed using Python. Python is an interpreted high-level programming language that has been gaining sustained popularity in many fields in recent years (TIOBE,

2018). Due to the emergence of a plethora of Python libraries for data processing, machine learning, statistical analysis, and visualization; Python has been widely adopted by the scientific computing community (Oliphant, 2007). A brief description of the main libraries used for data analysis in the context of this thesis is presented next:

- *NumPy*, a Python library that efficiently manages numerical data (Oliphant, 2006). NumPy is the foundational library for scientific computing on top of which most data analysis libraries are built.
- *pandas*, a Python library that offers high-performance data structures and data analysis tools (McKinney, 2010). The library provides a data structure, named *dataframe*, which consists of a two-dimensional tabular with both row and column indices. This data structure combines the high-performance capabilities offered by NumPy with a flexible data manipulation functionality similar to that offered by spreadsheets and relational databases.
- *GeoPandas*, an extension of pandas that allows working with geospatial information (GeoPandas developers, 2013).
- *Matplotlib*, the most popular plotting library in Python to generate two-dimensional visualizations (Hunter, 2007).
- *Bokeh*, an interactive visualization library for data applications (Bokeh Development Team, 2018). Bokeh incorporates an optional component, named *Bokeh Server*, which allows deploying interactive visualizations as standalone applications.
- *Datashader*, a library for plotting large volumes of data (Bednar et al., 2016). Datashader is designed to project entire datasets on to a two-dimensional rectangular grid, by aggregating data into bins (e.g., pixels). This approach allows creating effective visualizations of large datasets without the need to sub-sample, making it ideal for EDA.

These libraries were combined into an integrated ecosystem for data analysis by using the Jupyter Notebook (Kluyver et al., 2016). The Jupyter Notebook system provides a web-based application for interactive computing. The system offers a web interface to create *notebooks*, which are documents that combine text annotations, executable code, and outputs from computations. In essence, notebooks contain the inputs and outputs of an interactive computing session. Additionally, notebooks may incorporate accompanying text, thus interleaving executable code, rich representations of computed results,

and documentation. Notebooks are internally represented as JSON files, so are suited to collaborative editing and can be version-controlled with software such as Git. Furthermore, notebooks can be exported to several static formats such as HTML, LaTeX, and PDF. All these features make Jupyter Notebook a very interesting tool for *reproducible research* (Peng, 2011).

Hardware infrastructure

Due to the large volumes of data included in the analysis, special hardware infrastructure was required, particularly for the OD matrix generation described in Chapter 5. Data was processed over the cloud infrastructure at *ClusterUY*, the national center of supercomputing in Uruguay (Nesmachnow, 2010). The main goal of ClusterUY is to provide support for solving complex problems that require large computing power. Specifically, the computations described in Chapter 5 were performed using a server comprised of 40 Intel Xeon Gold 6138 (2.00GHz) cores and 128 GB of RAM.

Additionally, following a reproducible research methodology, Jupyter Notebooks corresponding to the data analysis process were hosted at the GitLab server provided by Facultad de Ingeniería (FING). Finally, interactive Bokeh and Datashader plots generated during the data analysis process were published as standalone applications and hosted at Amazon Web Services. All these resources are available at the thesis website (`www.fing.edu.uy/~renzom/msc`).

4.2.2 Data collection and processing

On August 2010, a presidential decree was published which regulated public access to state-owned information (Presidencia de la República, 2010). Following its publication, several initiatives have been taken to strive to open up data to the public at all levels of the public administration. A web portal that acts as a hub for open data at the state level was created and is available at `www.catalogodatos.gub.uy`. Additionally, many state agencies and local governments have web interfaces for publishing open data. In the context of this thesis, the most useful web interface was the geographic information site at IM (`www.sig.montevideo.gub.uy`), which holds geographic data of Montevideo including base maps, socioeconomic indicators, and transportation network data. Open data from these sources was key for the performed analysis and

are cited frequently throughout the document.

Besides using open data publicly available, the analysis included data regarding STM accessed through a collaboration between FING and IM. The sources of these data are the AVL and AFC systems integrated in buses of the STM. The data corresponding to the full set of records of GPS bus location and bus ticket sales payed with STM cards during 2015 was released for research purposes. These large datasets comprise over 150 GB of raw data.

The bus location dataset contains information about the position of each bus in STM, sampled every 10 to 30 seconds. Each location record holds the following information:

- a unique bus line identifier.
- a unique trip identifier to differentiate trips of the same bus line.
- GPS coordinates.
- instant speed of the vehicle.
- time stamp when the GPS measure was taken.

Ticket sales data contain records related to each STM transaction made, including the following fields:

- trip identifier for the sale, which allows linking to the bus location dataset.
- GPS coordinates at the moment of the STM card validation.
- bus stop identifier.
- time stamp at the moment of the STM card validation.
- unique STM card identifier, hashed for privacy purposes.
- number of passengers traveling with the same STM card.
- leg number, for multi-leg trips that include transfers.

The fact that each ticket sale is associated to a univocally identifiable key is a fundamental feature for building OD matrices based on historic trip information of users, as outlined in Chapter 5.

The data collection process was straightforward in the case of already opened datasets. The main efforts on this phase were related to data provided by IM. Several meetings with authorities at IM were celebrated, until an agreement was signed granting access and use to the data for research purposes.

With regards to the processing phase, the studied data was structured in pandas dataframes. Among the many transformations performed to the datasets, the most significant one was related to the Coordinate Reference System (CRS). Since several sources of data were considered during the anal-

ysis, geospatial data appeared in a variety of CRS. In order to be able to combine different datasets, all geospatial data was transformed to the WGS 84 (EPSG:4326) coordinate system which is the standard used by GPS.

For the sake of clarity in the visualizations, the reported results of the analysis correspond to tickets sold during the month of May 2015. Pre-hoc analysis of the complete dataset showed that this month is representative of the trends in the full dataset. The source code for the analysis (Available at: www.fing.edu.uy/~renzom/msc) can be easily configured to process any subset of the complete dataset.

4.2.3 Exploratory Data Analysis (EDA)

An initial EDA was performed to characterize the dataset of sales with STM cards. Visualizing data early in the analysis process is crucial, specially when dealing with urban data (Tukey, 1977). Figure 4.6 shows an aggregated visualization of the geolocation of 20.4 million sales corresponding to May 2015. The visualization was generated using Datashader, as described in Section 4.2.1. The location of each STM transaction was projected on to a grid of bins of size equal to one pixel of the 900×750 image. Then, transactions on the same bin were aggregated and a color mapping was applied to generate the final image, where brighter (white) areas indicate high concentration of ticket sales whereas darker (red) areas indicate low STM transaction activity. An interactive version of this visualization is also available at the thesis web site (www.fing.edu.uy/~renzom/msc).

The initial visualization of aggregated sales location data allows uncovering several interesting facts of the underlying dataset. Firstly, the city center is clearly different from other zones, with a significant higher number of STM transactions. Additionally, the main avenues can be clearly identified due to the higher number of ticket sales. It is worth noting that the visualization only considers ticket sales data and does not include any information related to the bus lines or the city streets. Furthermore, some sales activity is registered outside of the limits of Montevideo, for instance, in the sea. This is an important insight that guided the data cleansing process described in the following section.

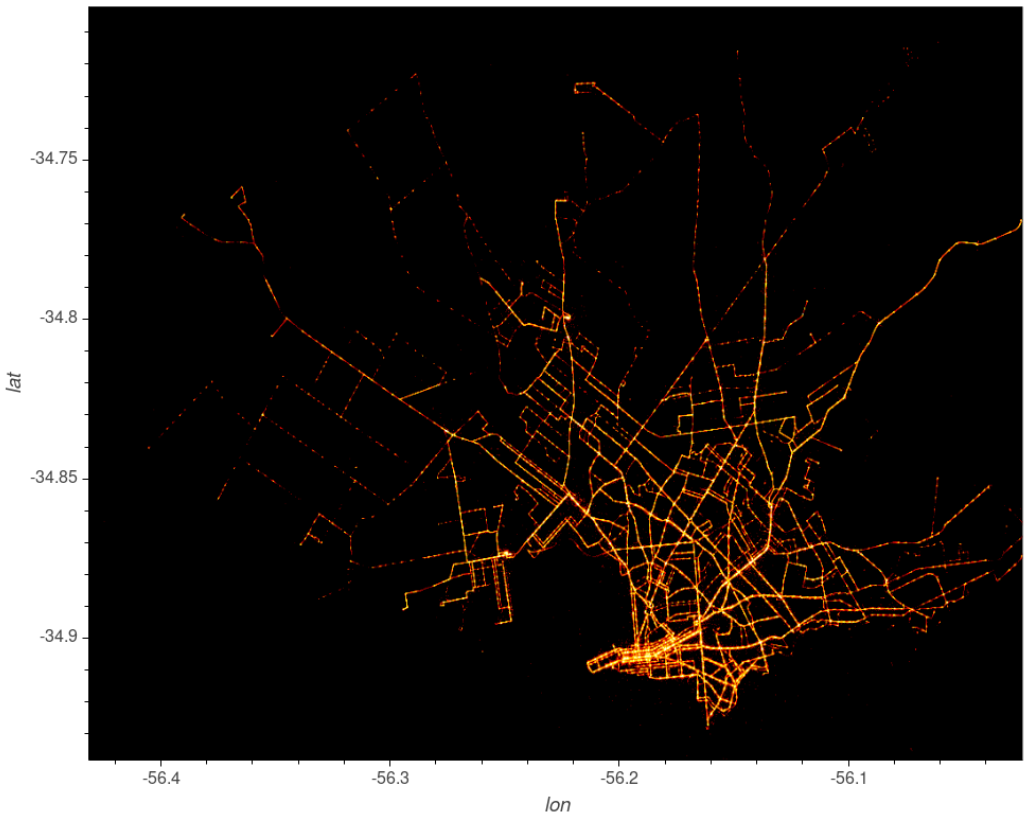


Figure 4.6: Aggregated sales with STM cards in May 2015

4.2.4 Data cleansing

Data cleansing is a mandatory step in data analysis that strives to detect and correct corrupt or inaccurate records (Rahm and Do, 2000). EDA allowed identifying several types of inconsistencies within the studied datasets. Due to the lack of a backup source of information the chosen strategy was to delete records that appeared to be corrupted. This section describes the errors encountered and the actions taken to filter the inaccurate records from the dataset.

Vehicle location using GPS is prone to errors from a variety of sources, so several methodologies have been proposed to cope with this phenomena (Jagadeesh et al., 2004). As described in Section 4.2.2, the studied dataset holds the GPS measure at the time of the transaction. The most frequent error regarding geolocation was that a large number of records had a fixed value for both latitude and longitude. The fixed location pinpoints to the middle of the Atlantic Ocean. Most likely, this was caused by an error message of the GPS unit (e.g., when no satellites are visible) being misinterpreted as a valid coordinate during data recording. Considering the sales data of May 2015, 932.176

records suffered from this issue, accounting for nearly 4.6% of the total dataset. Additionally, 29.432 records corresponded to locations outside of the bounding box of Montevideo. However, these records were not discarded, since the dataset also holds the identifier of the boarding bus stop of each transaction, which is defined using a series of measures from the on-board GPS unit. Thus, even though the GPS measure at the moment of the transaction may fail, the boarding bus stop can be accurately determined from previous measures. Consequently, the bus stop identifier is more reliable than the raw GPS measure when defining the starting point of each trip. As a result, the OD matrix estimation described in Chapter 5 was designed using bus stops identifiers as starting and ending points for trips.

Regarding time stamps of transactions, the sales corresponding to May 1^{st} were filtered, since they correspond to Labour Day, a public holiday in which the transportation system is mostly inoperative. The transactions on this day represent a clear outlier from the remainder of the dataset. Thus, the transactions occurring on this date were filtered from the dataset, accounting for 74 records. Similarly, only one transaction occurring on May 31^{st} was present in the dataset. There is no clear explanation for this issue and, therefore, the record was filtered. As a consequence, during the data analysis process, the month of May will represent STM transactions occurring between May 2^{nd} 00:00:00 to May 30^{th} 25:59:59 of 2015.

Other filters were applied to account for situations identified during the EDA. Some transactions had trip identifiers which were not present in the GPS records. In other words, some transactions appeared in the AFC system but not in the AVL system. Since these records cannot be linked to their corresponding bus line, they were discarded. This approach allowed filtering 1634 additional records.

Similarly, transactions made with the same STM card during the same trip were detected in the original dataset. In some cases, transactions occurred within few seconds of each other. This might be caused by users validating their STM card twice when boarding the bus. In other cases, the repeated records occurred after several minutes. This might be explained by a synchronization problem between the bus and the centralized server where transactions are recorded. Since no fail-proof criteria can be adopted to decide which of the repeated records corresponds to the legitimate transaction, all repeated records were discarded, accounting for 22 transactions.

Since the dataset corresponds to sales from 2015, some transactions refer to bus lines that were modified or no longer exist. In this case, the transaction cannot be linked to a bus line nor to a bus stop according to current data. These transactions were also filtered from the dataset, accounting for an additional 36.030 records. Finally, a considerable amount of transactions had identifiers of bus stops which were not part of the bus line route corresponding to the sale. Due to this issue 274.011 additional records were filtered.

In summary, the complete data cleansing process consisted in filtering 311.772 out of a total of 20.359.835 records, accounting for 1.53% of the original dataset.

4.3 Results and discussion

This section outlines the main results of the urban data analysis process aimed at characterizing the use of the public transportation system in Montevideo, Uruguay. A description of the use patterns of STM cards is presented, as well as a spatial and temporal analysis of the use of the transportation system. Additionally, some practical use cases for the information resulting from the data analysis process are presented.

4.3.1 Characterizing the use of STM

This section presents the results from the data analysis process that help to describe the use of the transportation system in Montevideo from several perspectives.

Cardholders

The sales dataset holds transactions made with 654.228 different STM cards. As explained in Section 4.2.2, the STM system allows several passengers to travel together using the same STM card. Table 4.1 shows the number of passengers traveling with the same STM card. The vast majority of passengers use their personal STM card, with over 97% of transactions corresponding to individual ticket sales. Therefore, STM cards can be confidently assumed to represent a single passenger. This is a key assumption used in the OD matrix estimation presented in Chapter 5, where all passengers under the same STM card are assumed to travel from origin to destination without splitting. Thus,

the fact that few group trips are performed using the same STM card provides a certain level of robustness to the OD matrix estimation model.

Table 4.1: Number of passengers traveling with the same STM card

# passengers	*total*	*percentage*
1	19494451	97.24%
2	510043	2.54%
3	36454	0.18%
4	5468	0.03%
5+	1647	0.01%

Another interesting aspect that can be studied through data analysis is the frequency of use of the transportation system. Table 4.2 shows descriptive statistics of daily and monthly transactions per STM card. The *mean* number of transactions is reported, along with the standard deviation (*std*). Additionally, the minimum (*min*) and maximum (*max*) values are presented, along with the 25^{th} (*Q1*), 50^{th} (*Q2*), and 75^{th} (*Q3*) percentiles. The 50^{th} percentile corresponds to the median of the distribution of transactions per STM card. Monthly statistics consider all transactions done by each cardholder during May 2015. Daily transaction statistics only consider days for which at least one transaction was made.

Table 4.2: Descriptive statistics of daily and monthly use of STM cards

	STM transactions	
	daily	*monthly*
mean	2.78	30.65
std	1.53	28.14
min	1	1
Q1 (25%)	2	8
Q2 (50%)	2	22
Q3 (75%)	4	47
max	54	528

Several interesting facts arise from use data of STM cards. When looking at monthly figures, cardholders perform over 30 transactions on average, nearly one transaction per day. However, the standard deviation is large, indicating a significant difference between regular and sporadic users of the public transportation system. The median of the monthly transactions is 22, nearly one

transaction per working day in the month. Regarding daily use, the average cardholder performs 2.78 STM transactions each day that uses the transportation system. Figure 4.7 presents an histogram of daily transactions per STM card, considering only cards that made up to 10 transactions within the same day in order to remove outliers. Most cardholders perform two transactions per day, which probably correspond to direct trips used for commuting. It is interesting to observe that more cardholders perform four rather than three transactions. This might be explained by passengers commuting to work using a trip involving a transfer, thus, two transactions correspond to the outward trip and the remaining two transactions to the return trip.

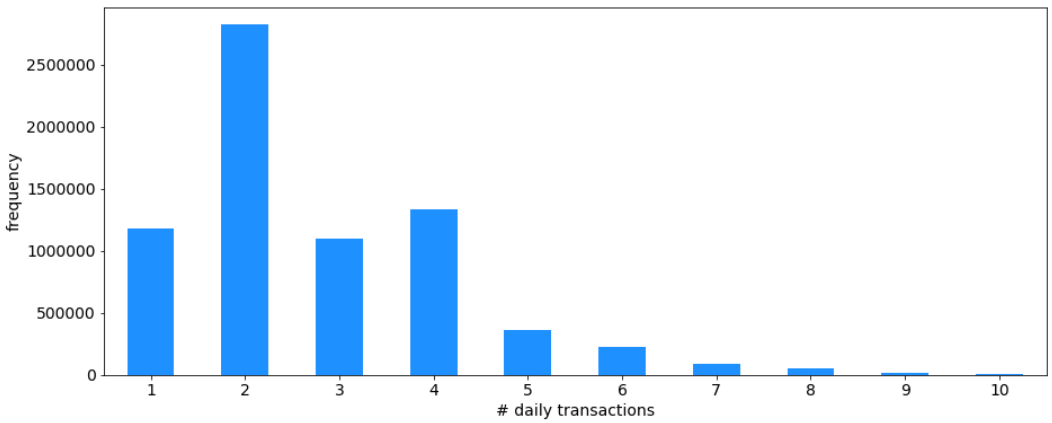


Figure 4.7: Histogram of daily transactions per STM card during May 2015

A few interesting applications arise when looking at outliers within the STM use statistics. On the one hand, cardholders with very low activity can be identified by their card ID. For instance, in the studied dataset 15.440 cardholders performed only a single trip during the whole month of May 2015. Targeted marketing campaigns could be designed to encourage disengaged citizens to use the public transportation system more frequently. On the other hand, cardholders with large number of transactions can also be identified. In the studied dataset a single card was found to perform 54 transactions within the same day. Through data analysis, authorities may further investigate these situations in order to identify possible abuses to the rules of the transportation system.

Transfers

As introduced in Section 4.1.2, STM tickets allow transfers between any bus line at any bus stop. Thus, a trip can be comprised of several legs, with bus transfers between each leg. Figure 4.8 details the percentage of trips involving different number of legs, where a trip with one leg corresponds to a direct trip. Results show that 55.99% of all transactions involve a single direct trip. Next, 40.26% of STM transactions correspond to a trip comprised of two legs and involving one transfer. The number of transactions involving more than two bus transfers are less than 4% of the total dataset. The average number of legs for the studied dataset is 1.37. According to the household mobility survey, presented in Section 3.3, the average number of legs when travelling by bus is 1.5 (Mauttone and Hernández, 2017). The slight difference between both estimations might be explained due to the fact that the mobility survey considers the walks from/to the bus stop as separate legs (if they are longer than 500 m). Since the cardholders identity is not included in the study dataset for privacy issues, personal information (e.g., home address) cannot be used to infer the walked distance to/from the bus stop from the studied dataset. Thus, direct trips requiring the passenger to walk more than 500 m to reach the bus stop are counted as two-legged trips in the mobility survey and as one-legged trips in the urban data analysis approach.

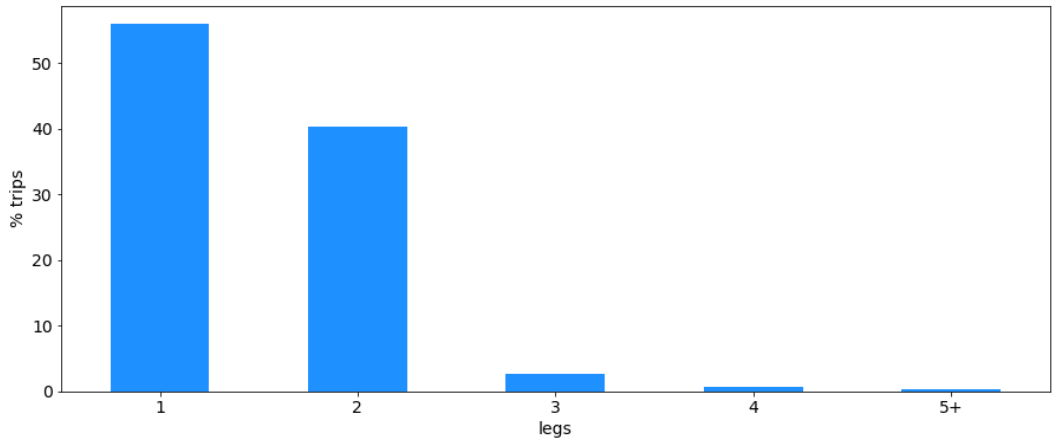


Figure 4.8: Percentage of legs per trip during May 2015

It is worth noting that consecutive transactions within the validity of a ticket are assumed to be legs of a larger trip. In reality, since the purpose of travel is unknown, a passenger may actually perform two independent short

trips using the same ticket. This problem was discussed in the analysis of the related literature and is addressed in Chapter 5 when estimating OD matrices.

Transactions per bus trip

Grouping STM transactions by their corresponding trip identifier provides a rough estimate for the number of boardings on each trip. Figure 4.9 presents a histogram of the number of transactions per trip. On average, 39.70 transactions are made in each bus trip (std: 28.16). The largest value encountered was a single trip with 249 transactions. It is worth noting that passengers might also board without using a STM card, so these figures represent a lower bound on the total number of boardings for each trip. Taking into account the capacity of the buses operating in Montevideo, some of the largest values point to one of the issues discussed in Section 4.1.2 regarding the length of some of the bus lines of the transportation system.

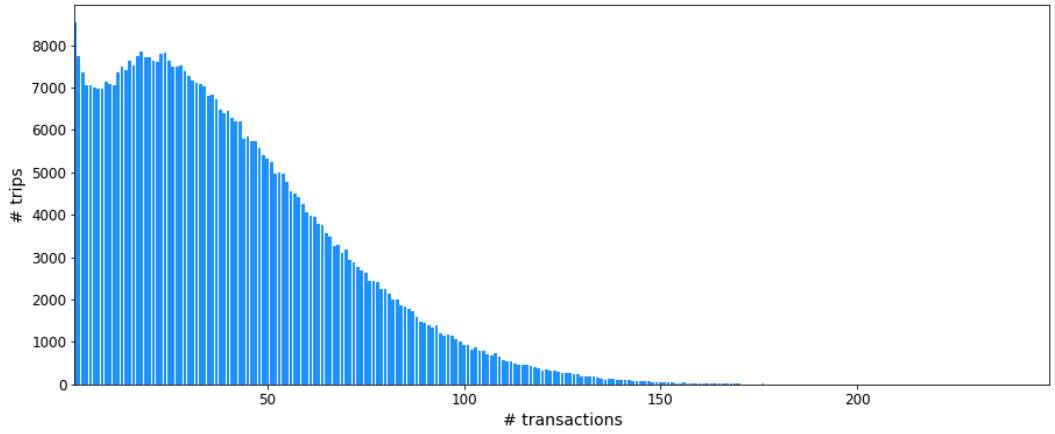


Figure 4.9: Histogram of transactions per bus trip during May 2015

Most used bus lines

Data analysis over the transaction data can be used to identify the most popular as well as the most underused bus lines. Figure 4.10 shows the ten most used bus lines. Some of the lines overlap since they correspond to different variants of the same line (e.g., outward and return lines). For each line the regular name (i.e., the name appearing in the front of the bus) is indicated in the map, along with its variant code indicated in parenthesis. The most used bus line is 183, closely followed by 181. Both lines connect the neighborhood

of La Teja, located in the west side of Montevideo, with Pocitos, located in the south by the coastline. It is interesting to notice that none of the ten most used bus lines go into the city center.

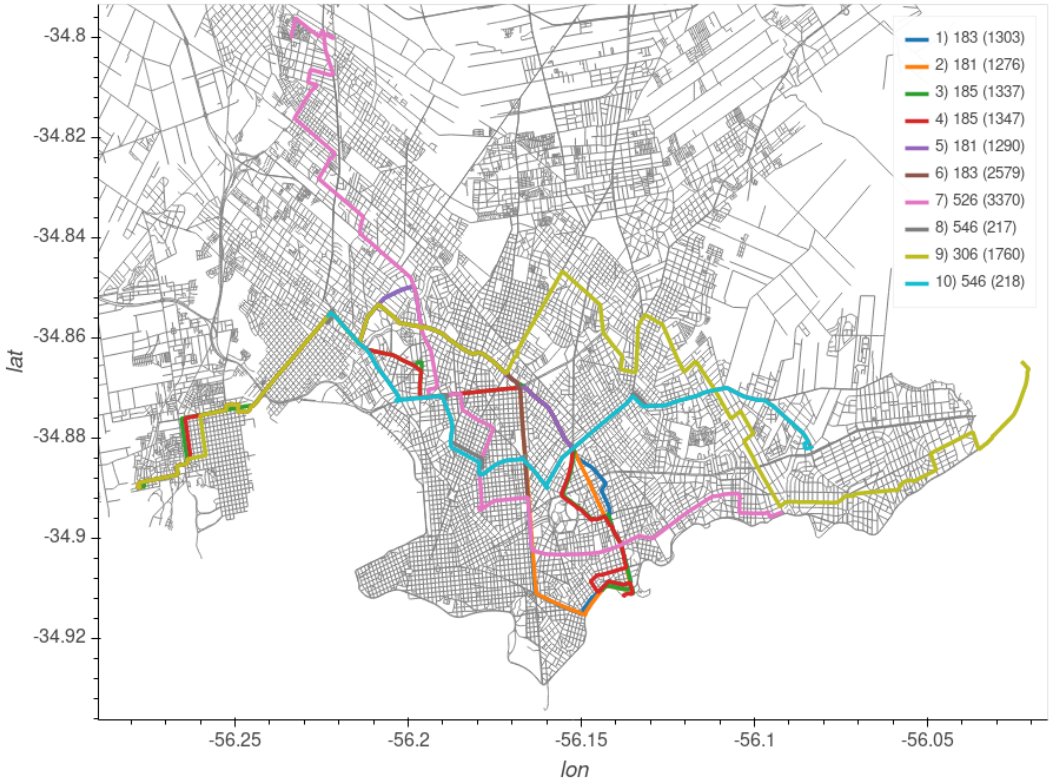


Figure 4.10: Top 10 bus lines with most STM card transactions during May 2015

Most used bus stops

Figure 4.11 shows the ten most used bus stops in the studied dataset. The first and second most used bus stops are located in the intersection of Agraciada and Freire avenues, in Paso Molino neighborhood. The third most used bus stop is located in Portones, a bus terminal within a shopping mall. It can be observed that the most frequently used bus stops are those corresponding to bus terminals or which are located in the intersection of busy avenues.

Temporal analysis of transactions

The AFC system in STM records the date and time of each transaction. These data allows analyzing the distribution of transactions across time.

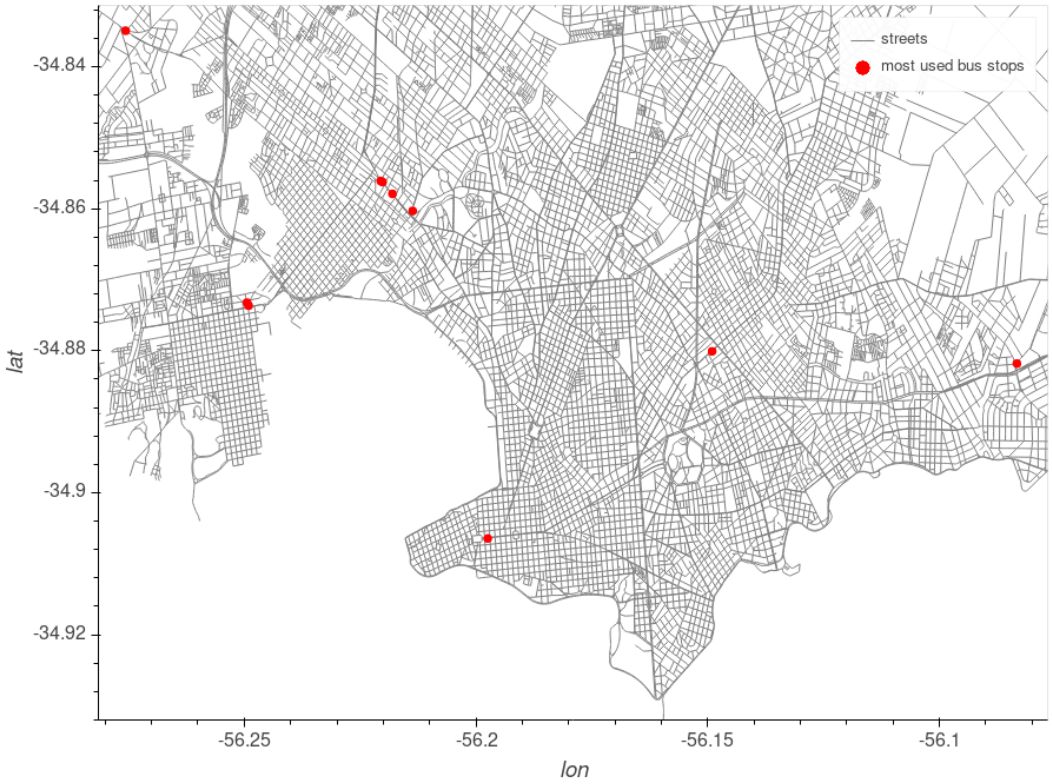


Figure 4.11: Top 10 bus stops with most STM card transactions during May 2015

Figure 4.12 outlines the number of transactions occurring each day of the week in the studied dataset. As expected, working days show the largest concentration of transactions with an average of ∼33.15M of transactions and a median of ∼34.41M. A slight decrease is noticed on Mondays, which might be explained due to the 18^{th} of May, a calendar holiday which was a Monday in 2015. In contrast, transactions during weekends drop significantly, with a clear difference between Saturdays (∼2.19M transactions) and Sundays (∼1.28M transactions).

A finer-grain analysis can be done to study the distribution of transactions across time. Figure 4.13 shows an histogram with the number of STM transactions at each hour of the day during May 2015.

As expected, two clear peaks of STM transaction activity can be noticed during the morning (7.00–8.00) and the afternoon (16.00–18.00), probably due to commuting. The morning peak is preceded by an increasing trend of sales starting at 3.00 a.m. while the afternoon peak gradually decays as the night approaches. However, an interesting observation is that another peak occurs

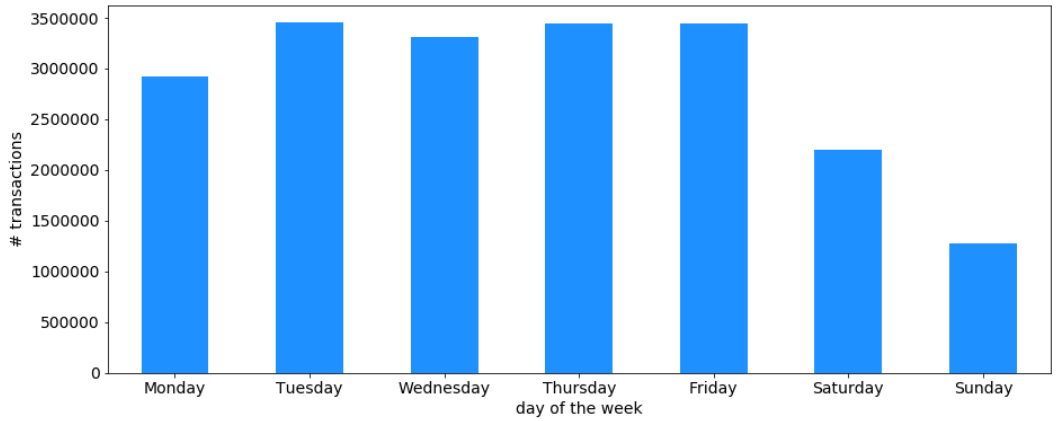


Figure 4.12: Histogram of sales with STM cards at each day of the week during May 2015

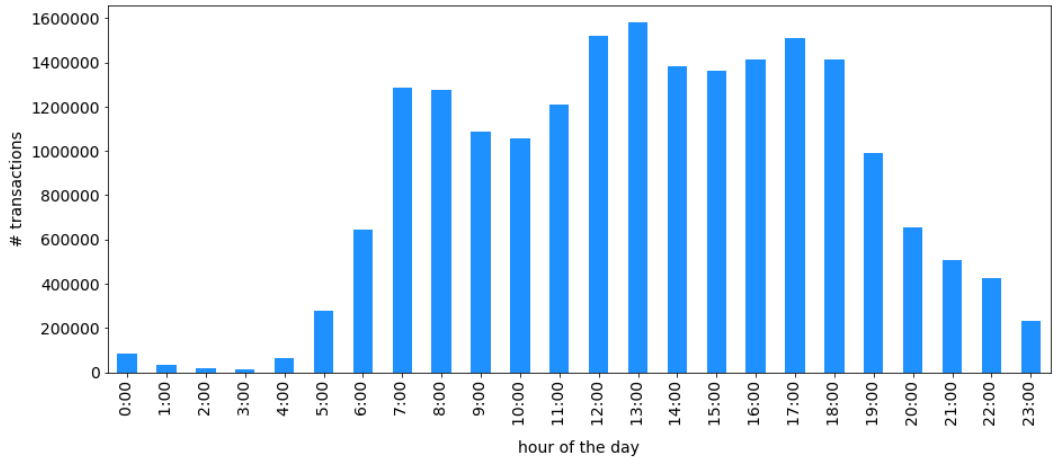


Figure 4.13: Histogram of sales with STM cards at different times of the day during May 2015

at midday (12.00–13.00) which might not be foreseen prior to the analysis. In fact, the overall largest amount of transactions occur at 13.00. Finally, it is worth noting that the lowest number of STM sales happen at 3.00 a.m. This finding is used for the OD matrix estimation algorithm presented in Chapter 5, which considers each new day as starting at 3.00 a.m., when fewer sales are made.

A similar temporal analysis was made during the 2016 household mobility survey, introduced in Section 3.3. Figure 4.14 shows the histogram of starting time of trips according to the urban mobility survey (Mauttone and Hernández, 2017). Although the survey covered trips in many modes of transportation, the histogram corresponds only to trips done by bus. The previous observations

regarding peak hours and the time of the day with fewest sales hold. According to the results from the survey, three peak hours can be identified (i.e., morning, midday, afternoon), and 2.00 a.m. is the time of the day when fewer sales occur. Consequently, the results of the temporal analysis following an urban data approach are highly consistent with those arising from the household mobility survey.

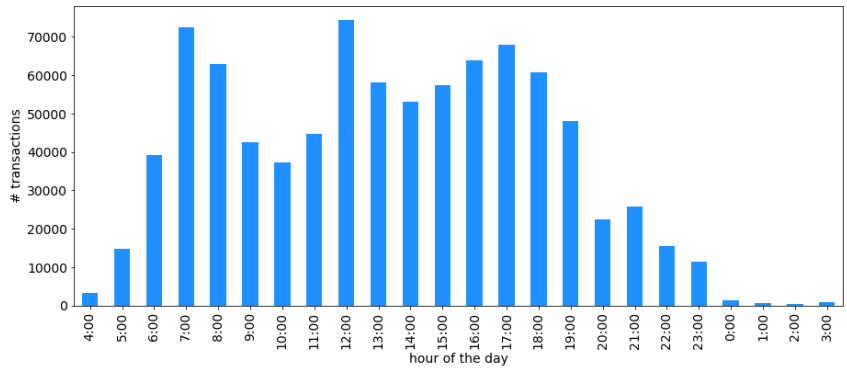


Figure 4.14: Histogram of starting times of trips in public transportation according to the Urban Mobility Survey Mauttone and Hernández (2017). Aggregated data provided by the authors. Raw data are available at https://catalogodatos.gub.uy/dataset/encuesta-origen-destino-montevideo

A different picture is obtained when studying weekends independently. Figure 4.15 outlines the number of transactions occurring at each time of the day considering only Saturdays and Sundays of May 2015. It can be seen that the distribution of transactions differs significantly from the one presented in Figure 4.13. The morning and afternoon peaks entirely disappear. Instead, the number of transactions steadily increases from the lowest value at 3.00 a.m. to the highest value at 12.00 p.m. Then, transactions gradually decrease, with a valley between 4.00 p.m. and 6.00 p.m. Unfortunately, the household mobility survey only characterizes trips done during working days. Therefore, it is not possible to assess whether or not these observations are consistent with the surveyed reality.

In this regard, it is interesting to highlight how the survey approach and the data analysis approach are not exclusive but, in fact, can complement each other. Urban data analysis can extract meaning from large volumes of data generated from sources such as ITS. This type of massive data collection would be infeasible to perform through surveys. However, ITS usually generate data as a by-product, since their main goal is not collecting data but

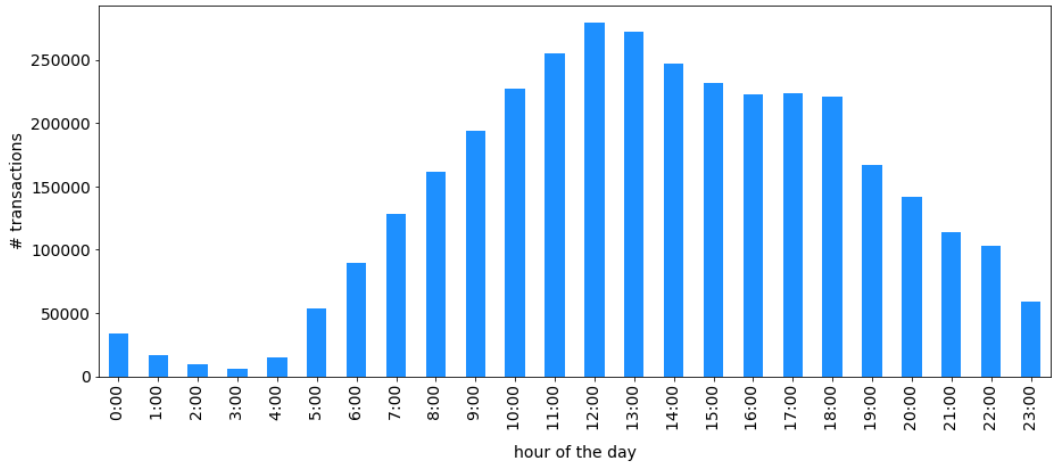


Figure 4.15: Histogram of sales with STM cards at different times of the day during weekends of May 2015

providing users with better QoS. In contrast, surveys are specifically designed to characterize the studied reality and provide answers to a series of questions. Thus, some of the information collected through surveys is hard to estimate through data analysis approaches. For instance, the household mobility survey holds information about the purpose of trips. Results show that the midday peak of trips disappears when considering only trips associated to commuting. This information cannot be easily inferred from the studied dataset, since no personal information (e.g., household or work location) is associated to each cardholder.

Spatial analysis of transactions

Since each sale record holds the geolocation of the bus at the moment the ticket was sold, interesting analysis can be performed to characterize sales activity in the spatial dimension. This type of analysis provides valuable insights to understand mobility and can help authorities in the decision-making processes aimed at improving the QoS offered to citizens. For instance, the city center of Montevideo is widely known to be one of the most troublesome areas in terms of mobility. These issues are related to the transportation network design, with many bus lines converging to the city center, as outlined in Section 4.1.2. This design leads to major congestion at peak hours in 18 de Julio, the main avenue in the city center.

In our article Massobrio and Nesmachnow (2016) the average speed of buses

running through 18 de Julio was studied. For this purpose the avenue was divided into three sections, as outlined in Figure 4.16. The study considered bus lines that run throughout the entire avenue in the East-West direction. The analysis used GPS traces of buses corresponding to the working days of the first week of September 2014. Table 4.3 shows the average speed of buses in each section of 18 de Julio at different times of the day. In the worst case, average speeds of 9.39 km/h were identified during the afternoon peak hour (17.00–19.00). Authorities at IM are concerned with the mobility issues in the city center and have proposed a plan to significantly alter the infrastructure of 18 de Julio (Intendencia de Montevideo, 2017). Decisions to address this kind of issues could be supported with evidence resulting from urban data analysis processes, as described next.

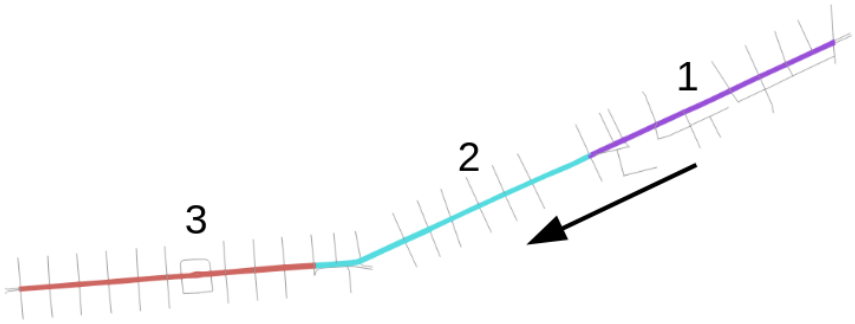


Figure 4.16: 18 de Julio avenue: sections considered for bus speed study

Table 4.3: Average speed of buses in 18 de Julio (in km/h)

	section		
	1	*2*	*3*
07:00–09:00	13.57	15.85	10.15
13:00–15:00	13.25	13.05	9.82
17:00–19:00	12.78	13.77	9.39
21:00–23:00	15.56	18.53	11.92

Figure 4.17 shows a heatmap of sales transactions in the city center during the month of May 2015. An interactive visualization for the whole area of Montevideo is available at the thesis website (`www.fing.edu.uy/~renzom/msc`). Bright (white) pixels in the heatmap indicate high concentration of ticket sales while dark (red) areas indicate low STM transaction activity.

The largest concentration of sales can be observed along 18 de Julio avenue. However, most of the streets running parallel to 18 de Julio also show a signif-

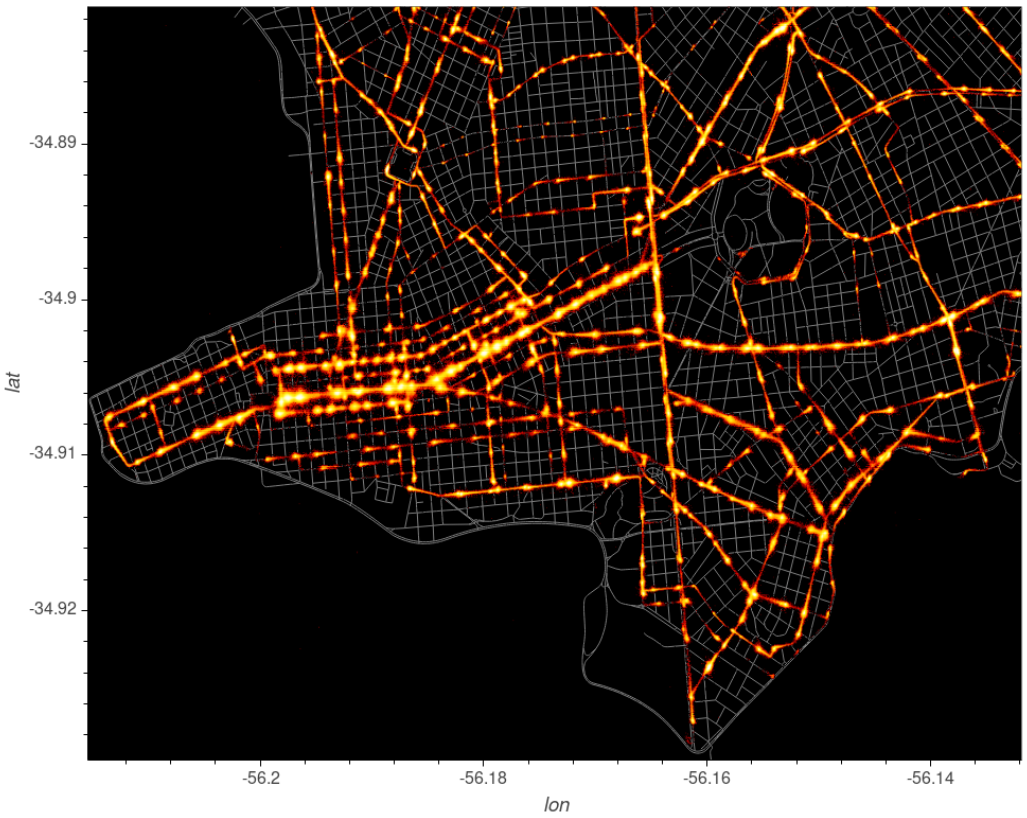


Figure 4.17: Aggregated sales with STM cards in the city center during May 2015

icant intensity of transactions. Thus, a plan that only targets the main avenue might not be successful in solving the mobility problems in the city center as a whole. Additionally, a considerable amount of sales activity is present in the old town, where streets are significantly narrower, thus aggravating the mobility issues in this area of the city.

Spatiotemporal analysis of transactions

The spatial and temporal dimensions of sales data can be combined, in order to gain insights that might not be evident when studying each dimension independently. Figure 4.18 shows an aggregated visualization of the spatiotemporal distribution of sales in Montevideo during May 2015. In this visualization the hours of the day are used as categories. Each transaction occurring at a given pixel in the image is categorized according to its time stamp. Then, the color of the pixel is set considering the amount of transactions on each category. The color mapping, which is detailed in the visualization, corresponds roughly to: red (12 a.m.), yellow (4 a.m.), green (8 a.m.), cyan (12 p.m.), blue (4

p.m.), purple (8 p.m.), and back to red, since hours and colors are both cyclic. An interactive version of the visualization of the spatiotemporal analysis is available at the thesis web site (www.fing.edu.uy/~renzom/msc).

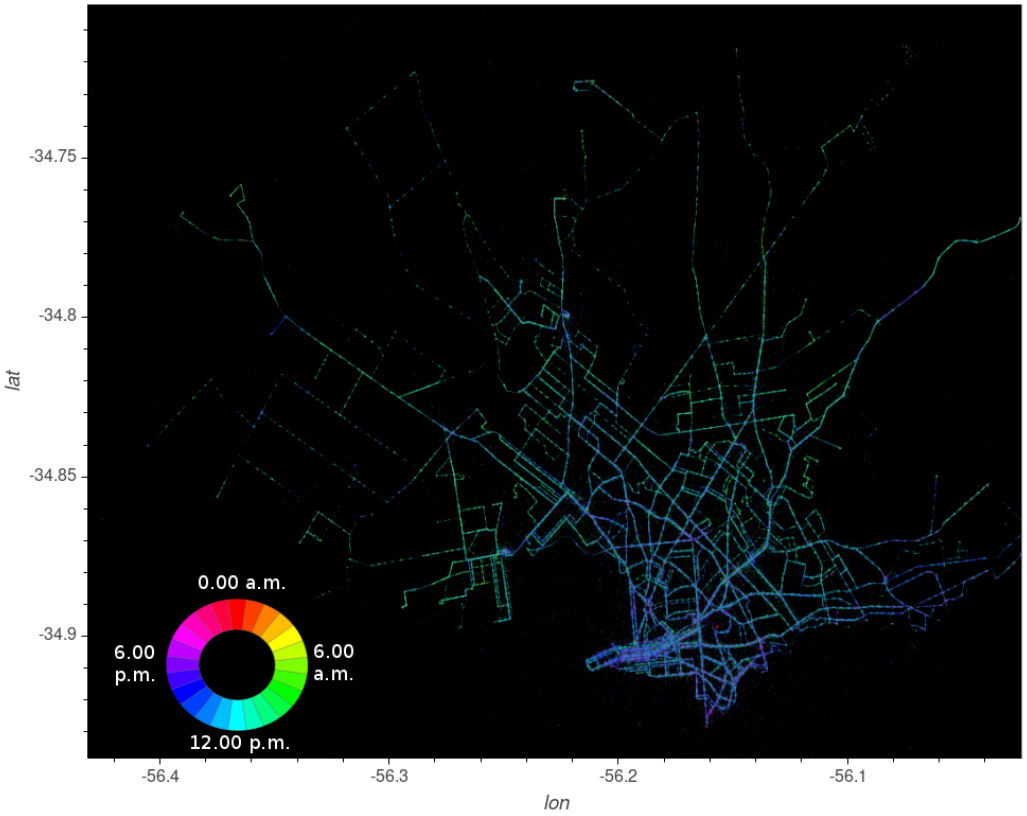


Figure 4.18: Spatiotemporal distribution of trips in Montevideo during May 2015

Firstly, it is observed that the city center has a prevalent blueish tone in the visualization. This corresponds to most transactions taking place between noon and the afternoon. This is consistent with the fact that many offices and public entities are located in this area of the city, thus, most transactions correspond to people commuting from the city center back to their homes by the end of the office-hours.

Another interesting fact arising from the spatiotemporal analysis of STM transactions is the clear difference between areas near the coast and areas farther away. It can be clearly observed that areas away from the coastline appear with more yellow and greener tones whereas areas closer to the coast have predominantly blue tones. This means that the majority of STM transactions in areas farther away from the coast occur earlier in the day than those near the coast. This can be explained by people commuting early in the day from these

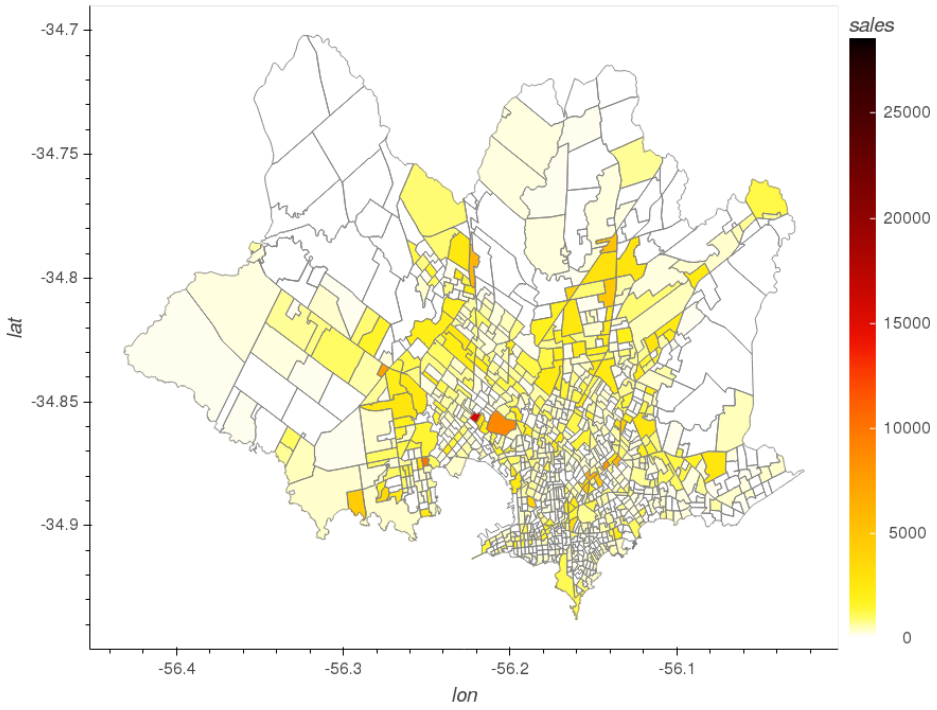
areas to workplaces located closer to the city center.

A more detailed analysis can be done by mapping transactions at different times of the day. Figures 4.19 and 4.20 show choropleth maps of the number of transactions occurring in each census segment in the morning and evening, respectively.

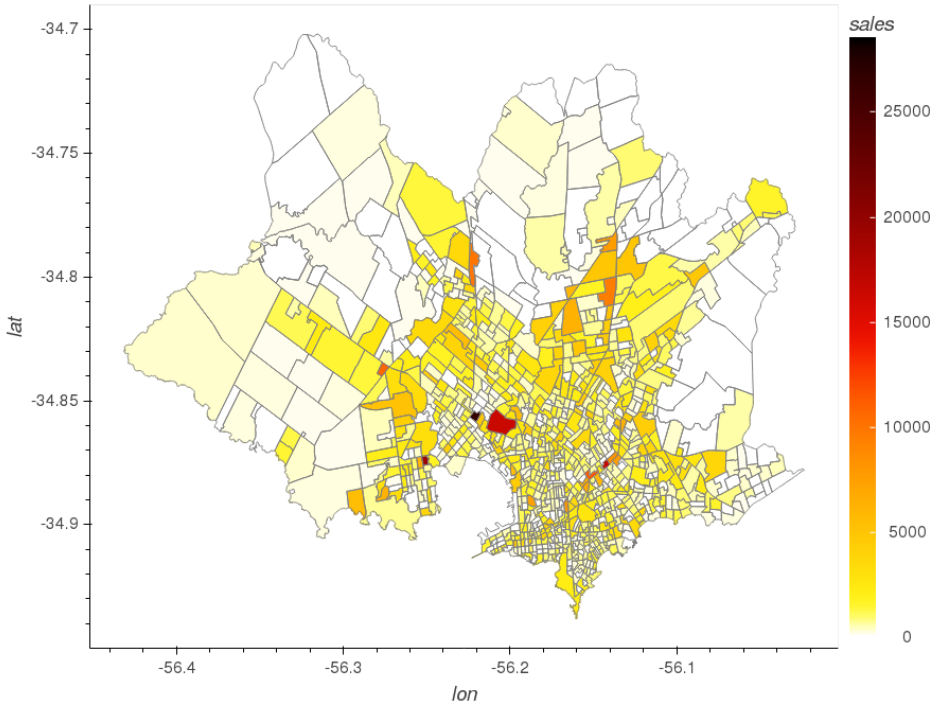
Regarding STM transactions occurring in the morning, Figure 4.19a clearly shows that those areas farther away from the city center and the coastline have higher STM transaction activity early in the morning (6.00 a.m.–7.00 a.m.) than those near the coast. Transaction activity in the city center and near the coastline intensifies an hour later, as can be seen in Figure 4.19b. Between 7.00 a.m. and 8.00 a.m. large amounts of transactions occur in most areas of Montevideo. A few census segments show a specially large number of transactions. These areas correspond to the location of bus terminals, where several bus lines converge and many transfers between bus lines occur.

Considering STM transactions occurring in the evening, Figure 4.20a shows a large number of transactions located in the city center area. This may be explained by the large amount of people returning to their homes from workplaces located in this area of the city at the end of office hours (6.00 p.m.–7.00 p.m.). When looking at transactions occurring later at night, Figure 4.20b shows that between 9.00 p.m. and 10.00 p.m. the amount of sales in the whole territory significantly drops. The areas with some remaining transaction activity are, once again, those located farther away from the city center and the coastline. This might be explained by people living in poorly connected areas taking longer to commute back to their homes by the end of the working day or also due to citizens working during night shifts and commuting to their workplace.

It is interesting to combine this analysis with the population density and socioeconomic description outlined in Section 4.1.1. Areas with transactions occurring early in the day and later at night are also the more vulnerable from a socioeconomic point of view. Our journal article (Nesmachnow et al., 2017) studied the differences in the QoS offered to citizens by the public transportation system according to their socioeconomic situation. Bus lines were characterized using the median household income of the areas they cover. A similar analysis could be performed to understand how mobility patterns vary across citizens with different socioeconomic characteristics.

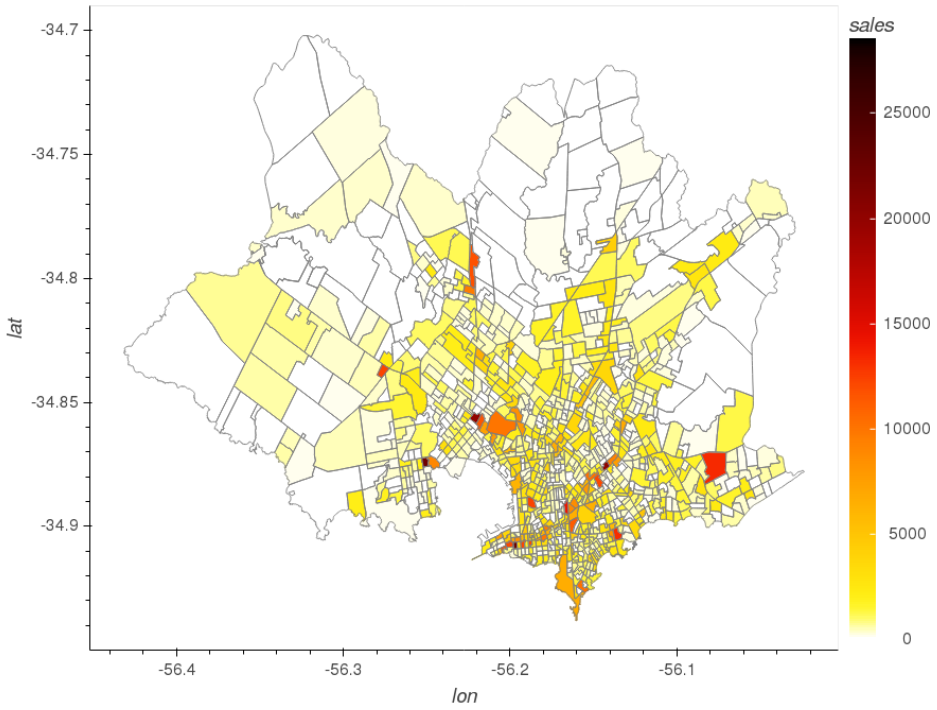


(a) 6.00 a.m.–7.00 a.m.

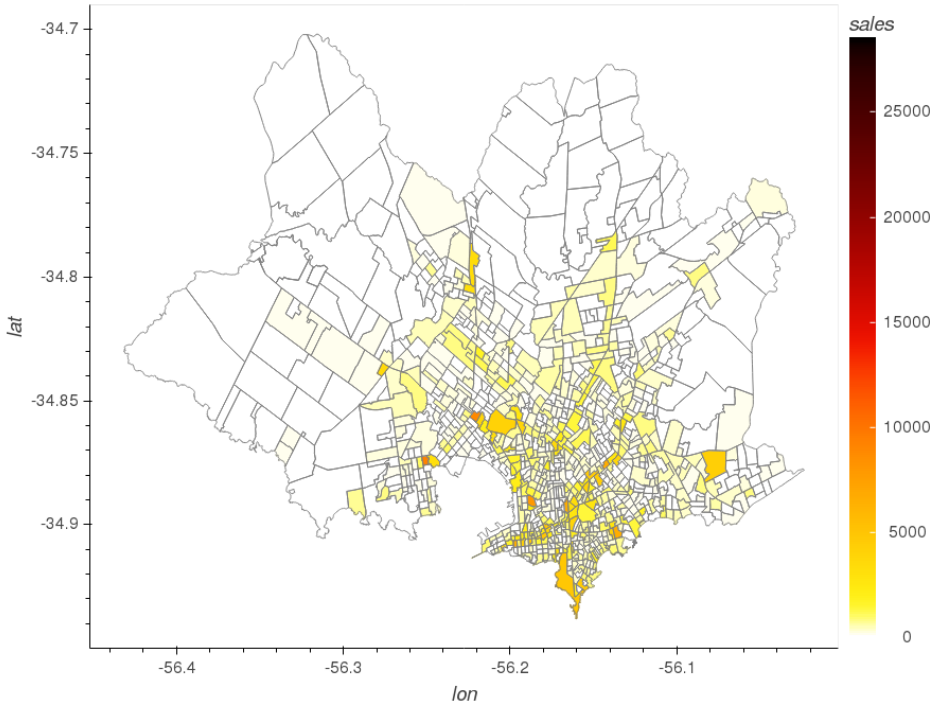


(b) 7.00 a.m.–8.00 a.m.

Figure 4.19: Choropleth map of STM transactions in the morning during May 2015



(a) 6.00p.m.–7.00p.m.



(b) 9.00p.m.–10.00p.m.

Figure 4.20: Choropleth map of STM transactions in the evening during May 2015

4.3.2 Practical use cases

Besides a purely descriptive use, urban data analysis can help authorities of the public transportation system in several ways. This section presents use cases where data analysis can be incorporated to the auditing, control, and policy enforcement workflows of public transportation authorities.

Anomaly detection in the spatial dimension

Geolocation data of sales transactions can be used to detect abnormal situations in the transportation system. As an example, Figure 4.21 shows a heatmap of transactions, along with the streets (in gray) and the bus lines (in blue). Two clusters of sales records (labeled A and B) appear in a street where no bus routes run. This represents a detour of one or more bus lines from their predefined routes. This may be due to an exceptional circumstance (e.g., road works) or due to a periodic event occurring certain days of the week (e.g., a flee market). Authorities can take advantage of this type of analysis to identify anomalies and make appropriate changes to bus routes and schedules.

Anomaly detection in the time dimension

By applying a similar methodology to the one used in the previous analysis, the time stamp of sales can be used to identify abnormal use patterns in the transportation system. Figure 4.22 shows an aggregated visualization of combined spatial and temporal information regarding STM transactions data. A small cluster of pixels in red can be observed in the map (indicated with a circle), which correspond to a group of sales occurring approximately at midnight. This pattern significantly differs from the remainder of the dataset. Given the location of these records, near an outdoor venue named *Velódromo Municipal*, the transactions probably correspond to a special event (e.g., a concert) taking place at night in this venue. In these occasions, bus companies usually assign buses to allow citizens to return to their homes at the end of the event. Authorities can use urban data analysis to identify special events taking place in the city and implement strategies that improve the mobility of those attending these events.

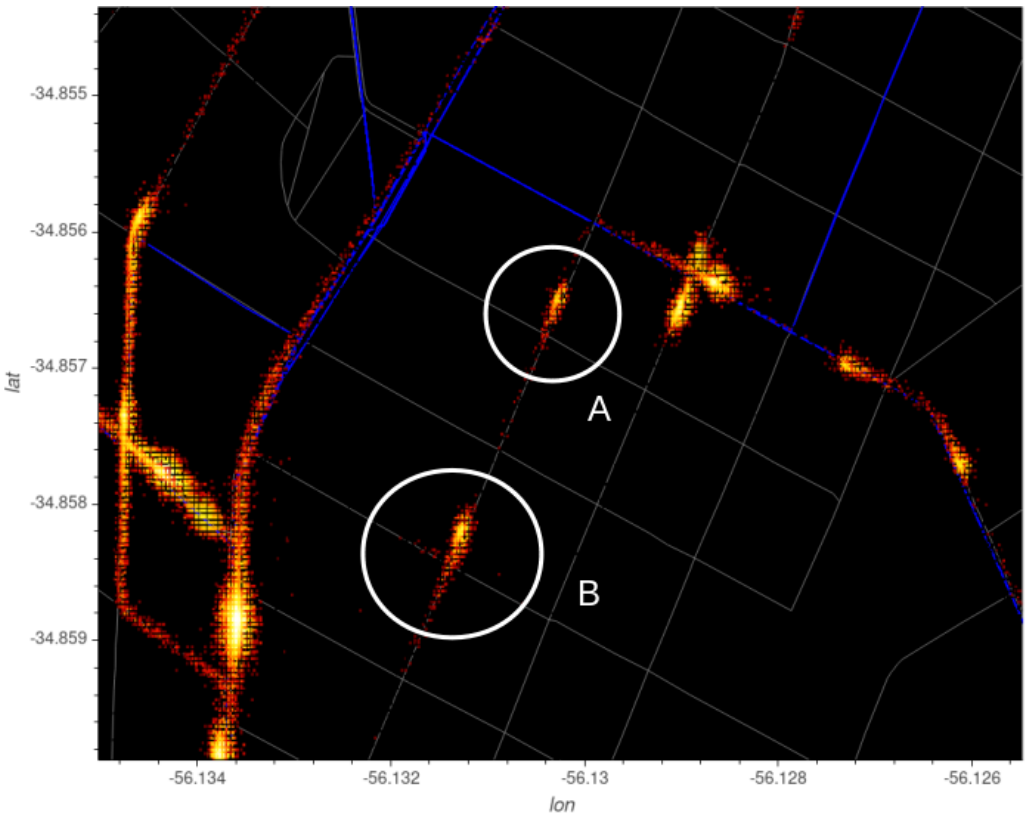


Figure 4.21: Anomaly detection: example of detour. The blue lines represent bus routes. A and B are two clusters of transactions which occurred outside of the bus network.

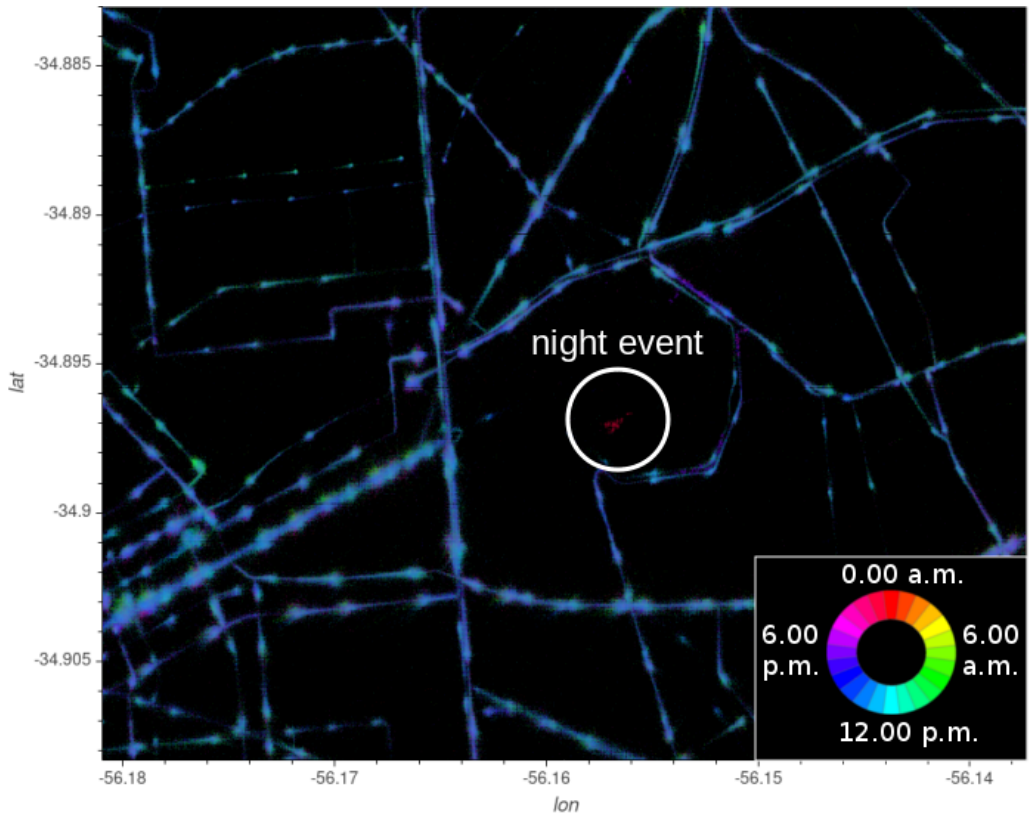


Figure 4.22: Anomaly detection: example of event taking place at midnight near an outdoor venue.

Driving behavior and safety

Another interesting use for information of sales records is to analyze the spatial distribution of sales with regards to bus stops. Figure 4.23 shows a heatmap of transactions occurring in one-way streets. Arrows indicate the direction of each street and bus stops are represented using blue circles. This visualization shows that the spatial distribution of sales is skewed with respect to the location of the bus stops. More transactions occur after the location of the bus stop than before. This uneven distribution is probably caused by drivers moving the bus before all the boarding passengers validate their smart cards. This might represent a safety issue, since passengers are standing while validating their cards. In fact, this might represent an even more serious issue, when drivers are also in charge of operating the smart card terminal. Driving and selling tickets at the same time is a risky behavior that can be seen frequently among bus drivers in Montevideo. It is interesting to see how data gives evidence that support these observations.

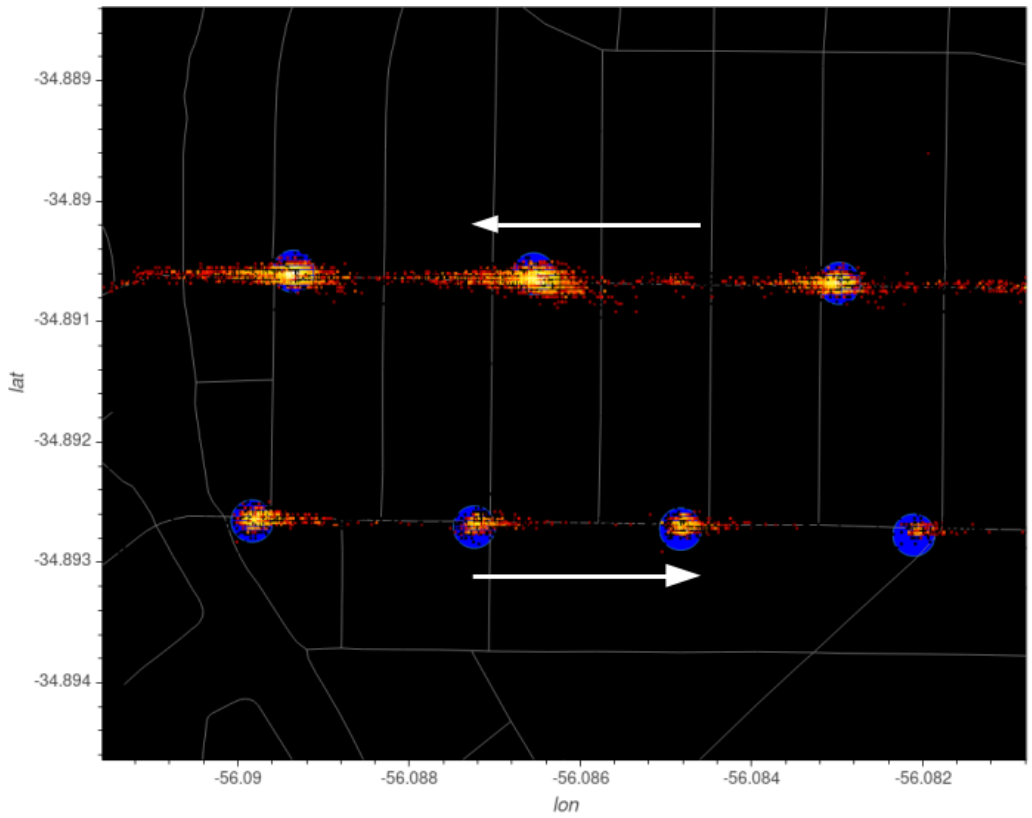


Figure 4.23: Spatial distribution of transactions with regards to stop location: one-way streets

Similarly, Figure 4.24 shows a heatmap of transactions and bus stops in the surroundings of a roundabout. It can be noticed that a large amount of transactions take place within the roundabout. This means that passengers are standing and validating their smart cards while the bus is moving. Additionally, for buses without an assistant, the driver is actually driving through the roundabout while operating the STM card terminal. Authorities can use this type of data analysis to audit driving behavior, improving the safety of passengers and drivers of the transportation system.

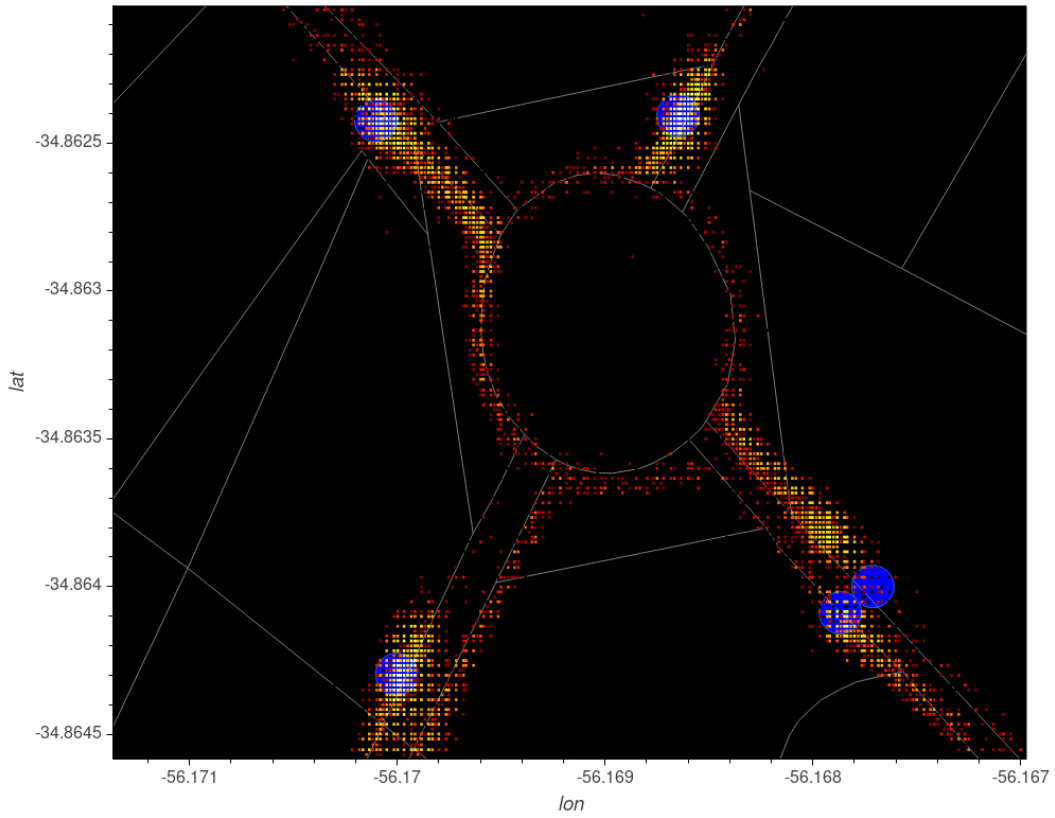


Figure 4.24: Spatial distribution of transactions with regards to stop location: roundabout

Chapter 5

OD matrices generation

This chapter describes the methodology and discusses the results of computing OD matrices using data from the ITS in Montevideo, Uruguay. Section 5.1 outlines the implemented solution for estimating OD pairs and presents a cloud computing framework that could be used to generate OD matrices. Then, Section 5.2 presents the OD matrix computed using ITS data from May 2015 and compares the result against the household mobility survey of 2016. Finally, two practical use cases of the information derived from the computed OD matrices are presented in Section 5.4.

5.1 Implemented solution

This section presents the implementation details of the proposed approach for building OD matrices using ITS data. Firstly, the details of the proposed approach for estimating destination of trips based on the trip-chaining method. Then, a cloud computing framework is presented, which could offer citizens and authorities near real-time mobility information.

5.1.1 Destination estimation algorithm

This section presents the destination estimation algorithm using trip chaining and gives the specific details used to adapt the algorithm to the case study of the ITS in Montevideo.

General overview

By combining data from the AFC and AVL systems it is possible to identify the origin of trips precisely, since the location of the bus is recorded whenever a passenger pays for a ticket using a smart card. However, since passengers are only required to validate their smart cards when boarding and not when alighting the bus, the destination of each trip is unknown and must be estimated in order to generate OD matrices. For this purpose, a destination estimation algorithm was developed based on the trip chaining method proposed by Barry et al. (2002) and later applied by other researchers, as outlined in Section 3.2.

The trip chaining method proposes estimating destinations of trips for a given passenger using information of the previous trips done by the same passenger earlier on the day. The method is based on the following two assumptions: *i*) the origin of a new trip is near the destination of the previous one; and *ii*) at the end of the day, users return to the origin of their first trip of the day. Figure 5.1 shows an example of the use of the trip chaining method to estimate destinations. In the example, the passenger performs three smart card transactions throughout the day. The boarding bus stops associated to each transaction are marked in green, and the estimated destinations of trips and trip legs are marked in orange.

In the example, the first transaction of the day occurs at 07:30, when the passenger boards bus line A at bus stop A_{19}. Later, at 08:15, the passenger boards bus line B at bus stop B_9 without paying for a new ticket. Since the boarding occurred within the validity of the previous ticket, the trip is assumed to be a transfer between buses. The closest stop from line A to bus stop B_9 is A_{23}, which is assumed to be the destination of the leg trip starting at 07:30. The last transaction of the day occurs at 17:20, when the passenger boards line C at bus stop C_4 and pays for a new ticket. Bus stop B_{12} is identified as the destination of the leg trip starting at 08:15, since it is the closest stop from line B to bus stop C_4. Since a new ticket was payed for, no further transfers are considered. Thus, an OD pair is identified between bus stops A_{19} and B_{12}. Finally, the destination of the last trip of the day is assumed to be bus stop C_8, since it is the closest bus stop of line C to the origin of the first transaction of the day (A_{19}). As a result, two OD pairs are identified, one consisting of two leg trips with a bus transfer and the other being a direct trip.

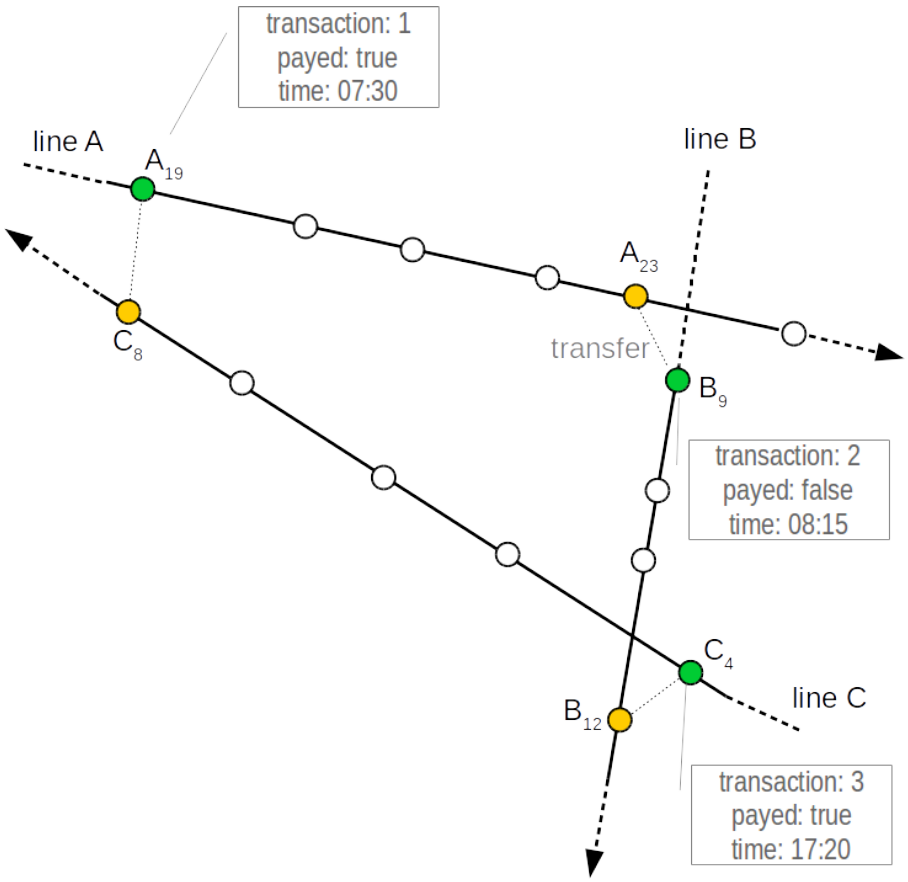


Figure 5.1: Example of the trip chaining algorithm to estimate destinations

Implementation details

The destination estimation algorithm was implemented using the Python programming language. A Python pseudo-code to describe the algorithm is presented in Listing 1. The algorithm receives as input a set of transactions sorted by their timestamp corresponding to a 24-hour period for a given cardholder. The first transaction of the day is processed independently (lines 2-7), since it is used to close the chain of trips with the last transaction of the day. Transactions are processed iteratively (lines 8-31). For each new transaction, the destination of the previous one is estimated (line 9). If the destination cannot be estimated (i.e., no stops are found within a given range), the trip chain is considered broken, so no further transactions are processed (lines 11-12). If a destination was estimated, the information regarding the identified trip is saved (lines 15-16). If the new transaction was payed for (i.e., it is not a bus transfer), then the estimated destination corresponds to the final end of the previous trip. Thus, the OD pair is recorded (lines 17-21), as it will be used

later to generate OD matrices. On the contrary, if the transaction was not payed for, then it corresponds to a bus transfer, which is recorded appropriately (lines 22-26). Then, the chain of trips is updated (lines 27-31) and the process continues with the next transaction of the day. Finally, the destination of the last transaction of the day is estimated independently, using the information of the origin of the first transaction of the day (lines 32-41). The algorithm returns three files with information corresponding to the identified trips, OD pairs, and bus transfers corresponding to the transactions received as input.

The destination estimation algorithm used during trip chaining is straightforward and is presented in Listing 2. The algorithm receives as input the identifier of the current stop, along with the identifiers of the previous line and boarding bus stop. All stops of the previous line occurring after the boarding bus stop are considered as potential destinations (lines 2-4). The distance from the current stop to all candidate stops is computed, and the closest bus stop is obtained (lines 6-8). Finally, if the distance to the closest bus stop is less than a given threshold, the algorithm returns the stop identifier as the estimated destination. Otherwise, an empty value is returned, which will break the chain of trips for that passenger.

Configuration for the ITS in Montevideo

The destination estimation algorithm processes sales data grouped in chunks corresponding to 24 hour periods. Records are split at the time of the day when the lowest sales activity is observed, as recommended by Munizaga et al. (2014). In the studied scenario, with data from the ITS in Montevideo corresponding to May 2015, the lowest amount of sales occurs at 3.00 a.m., as outlined in Section 4.3.1. A similar methodology was used by the urban mobility survey of 2016, which inquired about trips done in a 24-hour period starting at 4.00 a.m. (Mauttone and Hernández, 2017).

The destination estimation algorithm limits the search of a possible destination bus stop to a configurable radius (MAX_DISTANCE parameter on Listing 2). The search is sensitive to this parameter: large values may incorrectly identify destinations when other transport modes are used within the chain of bus trips, while a small radius might miss to identify destinations for trips that involve large walks from the bus stop to the destination. In the reviewed

```
def trip_chaining(transactions):
    #First transaction of the day
    t=transactions[0]
    n_passengers=get_number_of_passengers(t)
    first_origin_of_day=origin_for_od=previous_origin=get_origin(t)
    timestamp_for_od=previous_timestamp=get_timestamp(t)
    previous_line=get_bus_line(t)
    for t in transactions[1:]:  #Process chain of trips
        destination=estimate_destination(get_origin(t), previous_origin,
                                         previous_line)
        if not destination:
            exit() # Trip chain broken. Exit.
        else:
            #Save the identified trip
            save_trip(previous_origin, destination, passengers,
                    previous_timestamp, previous_line)
            if is_payed(t): # Final destination, record OD pair.
                save_od_pair(origin_for_od, destination,
                             passengers, timestamp_for_od)
                origin_for_od=get_origin(t)
                timestamp_for_od=get_timestamp(t)
            else: #Save info regarding the bus transfer
                save_transfer(previous_line, get_line(t),
                              destination, get_origin(t),
                              get_number_of_passengers(t),
                              get_timestamp(t))
            #Update the chain of trips
            passengers=get_number_of_passengers(t)
            previous_origin=get_origin(t)
            previous_timestamp=get_timestamp(t)
            previous_line=get_line(t)
    #Last destination of the day
    destination=estimate_destination(first_origin_day, previous_origin,
                                     previous_line)
    if not destination:
        exit() # Trip chain broken. Exit.
    else:
        save_od_pair(origin_for_od, destination,
                     passengers, timestamp_for_od)
        save_trip(previous_origin, destination, passengers,
                  previous_timestamp, previous_line)
```

Listing 1: Trip chaining algorithm to process daily transactions of a given cardholder

```
def estimate_destination(current_stop, previous_stop, previous_line):
      #Consider alighting stops visited before boarding stop
      previous_stop_index=get_line_stops(previous_line).index(previous_stop)
      candidate_stops=get_line_stops(previous_line)[previous_stop_index:]
      #Find the closest stop from within the candidates
      closest_stop = min (candidate_stops,
                      key= lambda s: compute_distance(current_stop,s))
      distance=compute_distance(current_stop,closest_stop)
      #Check distance threshold
      if distance<=MAX_DISTANCE:
          return closest_stop
      else:
          return None
```

Listing 2: Destination estimation algorithm

works of the related literature, several values were found for this parameter: 800 m (Alsger et al., 2015), 1000 m (Wang et al., 2011; Munizaga and Palma, 2012), and 2000 m (Trépanier et al., 2007). In this work the maximum distance to search for a destination bus stop was set to 1000 m, which is the median of the values found in the related literature. Additionally, 1000 m is also the maximum distance used to classify a walk as "short" according to the urban mobility survey (Mauttone and Hernández, 2017).

The proposed approach for destination estimation could be further improved. The trip chaining methodology may provide inaccurate results when mixed modes of transportation are used, since the sequence of trips using buses is broken. For instance, the trips of a passenger commuting to work by bus and returning home in a private vehicle (e.g., carpooling) would not be identified. Several alternatives could be implemented as a fallback method when trip chaining is not possible. Machine learning and clustering methods could be used to identify frequent bus stops visited by a passenger. By looking at historical data (e.g., monthly or yearly transactions) instead of relying only on the transactions occurring on the same day, frequently visited areas could be identified and assigned to trips for which the destination cannot be estimated using trip chaining.

Another aspect of the proposed approach that could be improved involves transfers. The destination estimation algorithm assumes that trips done within the validity of a ticket correspond to legs of a larger trip. In reality, passengers

may use a single ticket for independent trips in order to perform several short activities. To mitigate this issue AVL data should be used to determine the time of alighting from the first bus. Then, a simple time-based criteria could be used to decide whether a transfer corresponds to a second leg of a larger trip or to a short activity. For example, Munizaga and Palma (2012) proposed using a threshold of 30 minutes for the transfer. If the time between alighting from the first bus and boarding the second bus is larger than 30 minutes, the passenger is assumed to have engaged in a short activity and the trips are recorded separately. A more complex method could be devised, using AVL data to assess whether the passenger boarded the first arriving bus or if several buses passed by the bus stop before the passenger boarded, which may be an indicator of a short activity taking place.

5.1.2 Cloud computing framework

In our journal article, Massobrio et al. (2018), we described a framework for processing ITS data in the cloud in order to improve public transportation systems. The proposed framework is suitable to process ITS data to generate OD matrices. The framework decomposes the problem at hand into two sub-problems: i) a pre-processing stage that prepares the input data for the next phase, and ii) the parallel/distributed processing of ITS data. A master-slave model is used to define and organize the control hierarchy and processing. Figure 5.2 outlines the conceptual diagram of the proposed framework.

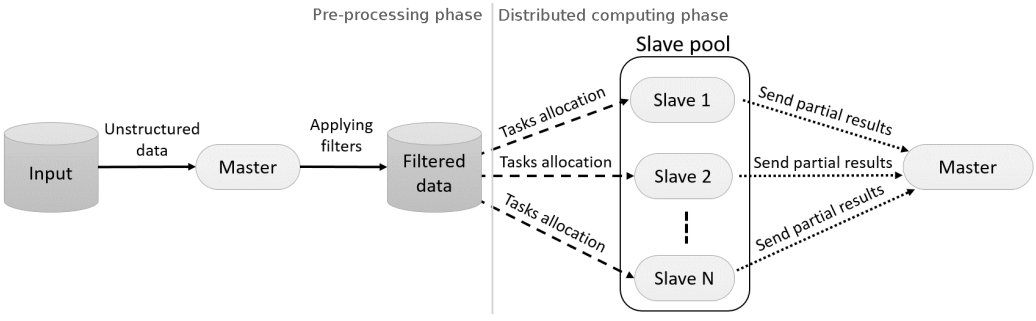


Figure 5.2: Framework for processing ITS data in the cloud

During the pre-processing phase, the master process prepares the data, filtering inaccurate or corrupt records. The filtering stage is dependent on the specific problem being solved. In the case of the OD estimation process, this phase corresponds to the data cleansing process described in Section 4.2.4.

The framework applies a data-parallel domain decomposition strategy for parallelization. After the pre-processing stage, the filtered data are split into chunks and distributed to several processing elements. The master process partitions the data and assigns each chunk to a slave for processing. The slave collaborate in the data processing, following a Single Program Multiple Data (SPMD) model.

The proposed implementation can be integrated under a Software as a Service (SaaS) paradigm in a real cloud computing environment, providing a useful service for both citizens and authorities. A diagram of the proposed system is presented in Figure 5.3. Buses send their current geographic location to a server in the cloud. The server performs the distributed processing of the collected location and sales data from buses in real time. The results from this processing are published to be consumed by intelligent ubiquitous mobile applications and websites for end-users and by monitoring applications for city authorities, following the traditional SaaS model for cloud computing. From the point of view of public transport users, information from real-time ITS data can help with mobility decisions (e.g., prefer a certain bus line over others, decide to transfer between buses). From the point of view of the city authorities, mobility data are useful for planning long-term modifications to bus routes, timetables, bus stop locations, as well as to identify specific bottleneck situations.

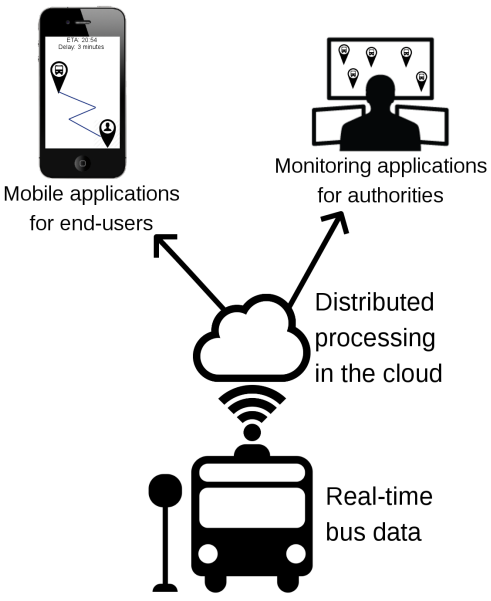


Figure 5.3: Cloud computing framework for real-time ITS data processing

5.2 Experimental results

This section reports the experimental results from the OD matrix estimation process. Firstly, the results achieved by the implemented algorithm are presented and discussed. Then, the computed results are compared and contrasted against the findings from the household mobility survey of 2016.

5.2.1 Numerical results

After the cleansing process described in Section 4.2.4, 311.772 records were discarded from the dataset corresponding to May 2015, leading to a cleansed dataset comprised of 20.048.063 records. For the destination estimation process, this dataset was split into chunks, where each chunk held the information for an entire day starting and ending at 3 a.m. Due to this splitting strategy, six hours worth of data were discarded, i.e., the first three hours of the first day and the last three hours of the last day of the dataset. Additionally, since the destination estimation algorithm requires at least two transactions to perform trip-chaining, the records associated to cardholders that only performed one transaction within a given day were filtered from the dataset. As a result, the destination estimation algorithm was applied to a set of 18.885.711 records. Out of these records, the implemented algorithm was able to assign a destination to 15.414.230 trips, achieving a success rate of 81.62%. This is a highly competitive result, considering the success rates achieved by other works in the related literature, e.g., 57% (Wang et al., 2011), 66% (Trépanier et al., 2007), 80% (Munizaga and Palma, 2012). Each identified trip holds the following information: boarding bus stop, time stamp at boarding, bus line identifier, and alighting bus stop.

OD matrices were built considering the first origin and final destination of each trip, without considering intermediate stops due to transfers. As a result, the number of OD pairs is lower than the number of identified trips, since more than 40% of trips involve at least one transfer, as shown in Section 4.3.1. Computed results allowed identifying 9.485.904 OD pairs, which were used to generate OD matrices. At the finest grain, OD matrices were generated considering each pair of bus stops. At a more coarse grain, the computed results were aggregated for each census segment. Both OD matrices are available at the thesis website (`www.fing.edu.uy/~renzom/msc`) in Comma Separated Values (CSV) files with their corresponding metadata. For the sake of visual-

ization, results are discussed at a coarser grain in this document, aggregating the computed OD pairs by municipality. Table 5.1 outlines the estimated OD matrix corresponding to the studied dataset of May 2015. Each municipality is represented by its identifying code, as described in Section 4.1.1.

Table 5.1: Estimated OD matrix by municipalities for May 2015

		destination								
		A	B	C	CH	D	E	F	G	*total*
origin	A	626388	199196	184905	98087	30108	40370	21875	73390	1274319
	B	154358	662993	224578	366865	108640	173898	119306	108469	1919107
	C	174040	260526	320368	111113	102244	64691	62188	101337	1196507
	CH	100348	334040	131089	362377	101433	156685	115310	66461	1367743
	D	48502	222110	148581	130733	321610	71018	93969	64253	1100776
	E	27463	138400	46288	110868	86344	287243	133179	28827	858612
	F	21038	127429	51570	108017	155355	82811	315573	20427	882220
	G	74482	141380	120539	57388	41670	29779	21068	379724	866030
	total	1226619	2086074	1227918	1345448	947404	906495	882468	842888	

Several conclusions arise from the computed OD matrix. Firstly, the largest values are located in the diagonal of the matrix. Values located in the diagonal represent trips starting and ending within the same municipality. This observation holds for every municipality with the only exception of trips ending at *CH*, which are mostly originated in *B* rather than *CH* by a small margin. Secondly, municipality *B* stands out as both the largest generator and attractor of trips when considering the total number of OD pairs (highlighted in gray in the table). This is consistent with the fact that the city center and other surrounding areas are within municipality *B*, where multiple workplaces, public offices, and services are located.

5.2.2 Comparison to the 2016 mobility survey

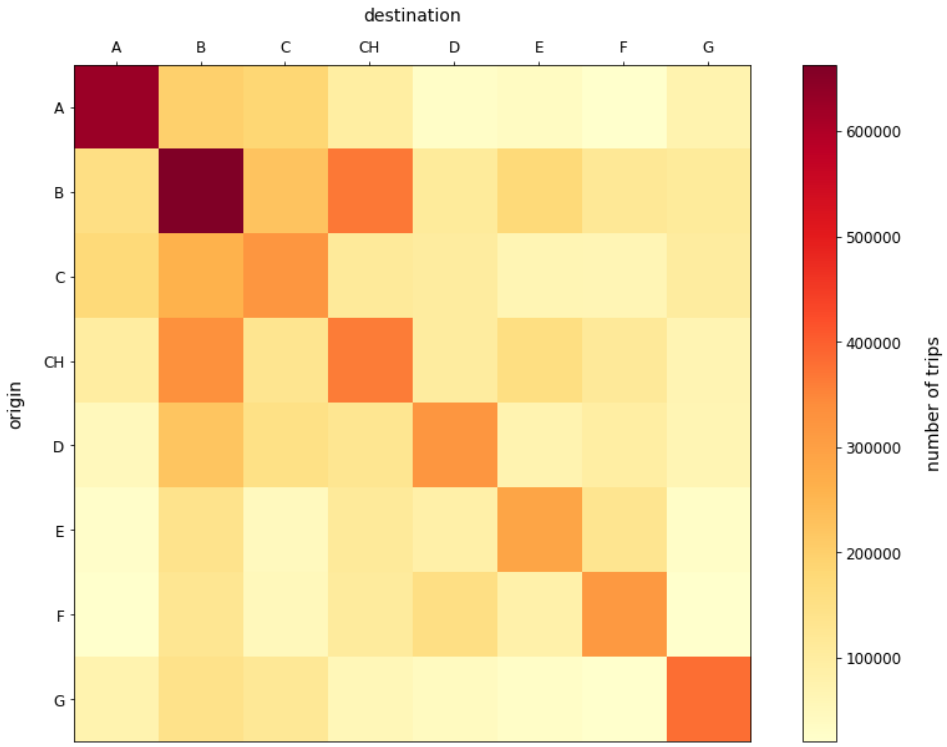
According to the best practices reviewed in the related literature, it is desirable to compare the results of the OD matrix estimated using ITS data against an alternative source of information. To this end, the results from the household urban mobility survey carried out in 2016 (presented in Section 3.3) were used. Table 5.2 shows the OD matrix of trips done between each municipality using public transportation, according to the results of the 2016 urban mobility survey (Mauttone and Hernández, 2017). Reported figures were computed by expanding the results of the survey taking into account the population of each municipality.

Table 5.2: OD matrix in public transportation by municipalities according to the 2016 household mobility survey (Mauttone and Hernández, 2017). Aggregated values by municipalities were provided by the authors. Raw data are available at https://catalogodatos.gub.uy/dataset/encuesta-origen-destino-montevideo

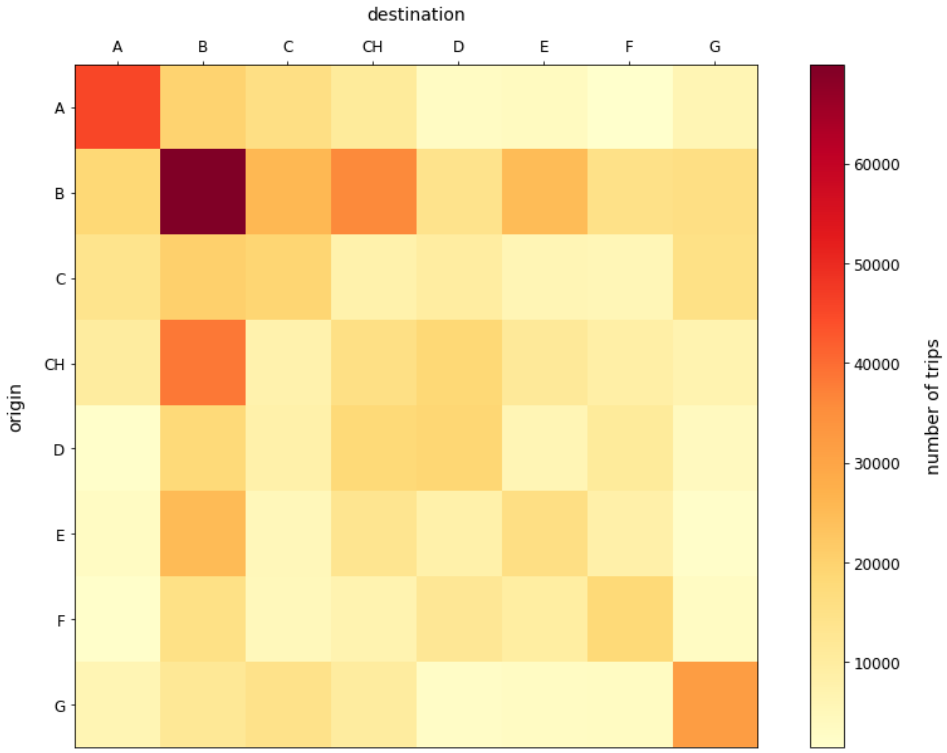
		destination								
		A	B	C	CH	D	E	F	G	*total*
origin	A	45576	19667	15733	10863	3103	3783	1436	6136	106297
	B	18301	69964	25705	36210	14245	24805	15071	15779	220080
	C	13772	20702	19314	7781	9823	5893	5622	15243	98150
	CH	10244	38628	7402	15614	18543	11360	8876	6935	117602
	D	1796	17904	7999	18275	18881	5835	10896	3954	85540
	E	3094	25004	4942	13397	8023	15801	8369	2213	80843
	F	1858	15296	4858	6957	12192	9515	18183	3304	72163
	G	6163	11984	14427	10026	2737	3061	3563	31865	83826
	total	100804	219149	100380	119123	87547	80053	72016	85429	

Results from the OD survey have many similarities with those estimated from ITS data. The previous observation of a large number of trips taking place within each municipality also applies to the results from the survey. Additionally, the survey OD also identifies municipality B as the largest generator and attractor of trips. These remarks can also be assessed in Figure 5.4, where a visual comparison between the OD matrices derived from ITS data and from the mobility survey is presented. Each OD matrix is represented as a two-dimensional grid with colors mapped according to the number of transactions occurring in each OD pair. Results derived from the ITS data are presented in Figure 5.4a whereas those derived from the mobility survey are presented in Figure 5.4b.

The visual representation of OD matrices as heatmaps on two-dimensional grids allows identifying further similarities between the results computed with ITS data and those from the mobility survey. Trips within municipalities A and B are the most dominant OD pairs according to both estimations, followed by trips within municipality G. Both figures show that trips from B to CH and vice versa are also highly dominant with regards to other OD pairs. The diagonal of the grid is mapped to more intense colors in Figure 5.4a than in Figure 5.4b. This might be a consequence of the larger number of trips considered in the OD matrix generated from ITS data. Despite this observation, an outstanding number of similar color patterns are found when comparing the grids both row-wise and column-wise.



(a) estimation for May 2015 using ITS data



(b) results from the 2016 mobility survey

Figure 5.4: Comparison of OD matrices.

Results are very promising, showing that OD matrices generated from ITS data are a valid alternative to understand mobility in a city. Several advantages can be highlighted from the proposed approach for building OD matrices. Firstly, due to the large volume of data generated by ITS compared to the number of individuals that participate in a survey, a finer-grain OD matrix can be obtained. With the approach proposed in this thesis, OD matrices at the bus stop and census segment levels were obtained, whereas the mobility survey results only apply to municipalities. Secondly, thanks to data analysis, different OD matrices can be computed applying different criteria regarding, e.g., days of the week, hours of the day. As an example to showcase this feature, Figure 5.5 shows a heatmap corresponding to the OD matrix derived from ITS data considering only weekends of May 2015. It can be seen that the role of municipality B as the largest generator and attractor of trips is significantly smoothed when considering only weekends. As stated before, several offices and workplaces are located within municipality B, which are mostly only opened during working days. The information from the mobility survey refers to trips done during working days only. Thus, in order to gain insight on the mobility of citizens during weekends a new survey ought to be carried out, with the associated costs and delays.

Regarding costs, the proposed approach for OD matrix estimation provides an attractive alternative for public administrations aiming at characterizing mobility in a city. If the ITS infrastructure is already deployed, deriving mobility information is nearly inexpensive, since value is produced from already existent data. This is clearly the case of Montevideo, where the ITS infrastructure has been deployed for nearly a decade. Besides economic considerations, it is worth noting that the proposed approach can be easily applied whenever new data becomes available. In fact, following the architecture described in Section 5.1.2, OD estimation techniques could be applied in a streamline fashion in order to obtain near real-time OD matrices. This represents a clear advantage in comparison to surveys, which demand large amounts of time to plan, carry out the survey, and process the results. As a consequence, the proposed approach allows easily obtaining an up-to-date view on the mobility of a city while surveys offer a partial and mostly outdated picture.

The previous observations are not aimed at questioning the importance and convenience of carrying out mobility surveys. On the contrary, surveys are essential to understand mobility in a city and authorities should invest in

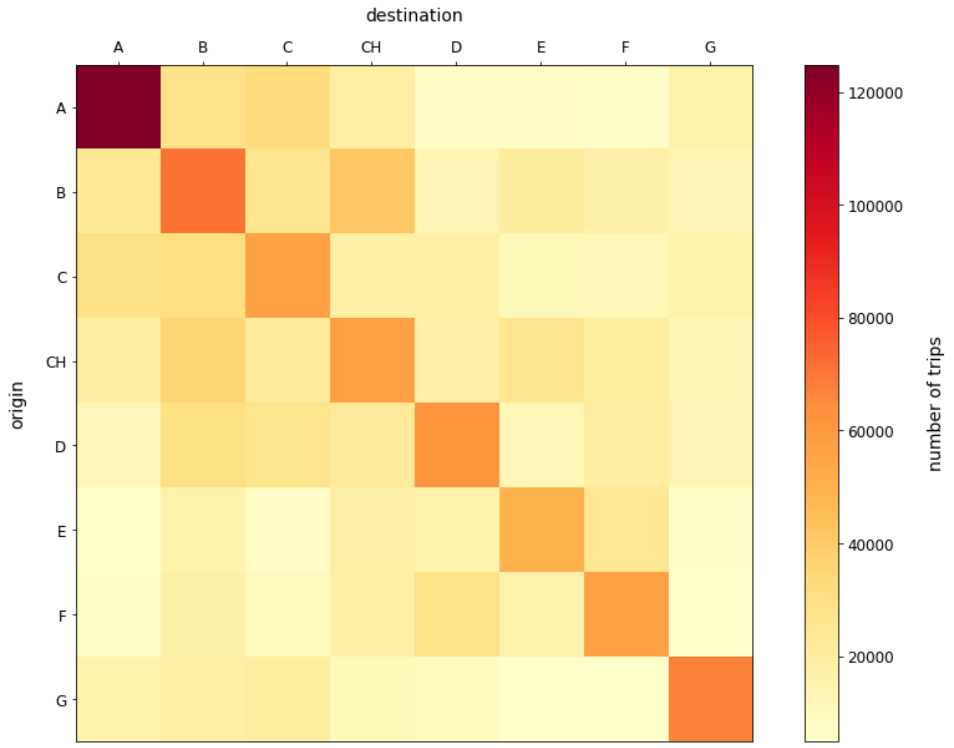


Figure 5.5: OD matrix heatmap by municipalities for weekends in May 2015

conducting them periodically. Firstly, because they serve as a ground-truth for other methodologies, such as the one proposed in this thesis. Secondly, surveys can be used to gain insights into aspects that cannot be easily derived from raw ITS data. For instance, the purpose of travel is a standard question in most mobility surveys and is not easy to state using merely data analysis (Nassir et al., 2015). Another example is leg identification in trips involving bus transfers. In the approach proposed in this thesis, all trips done within the validity of a ticket were considered as legs of a larger trip. In reality, passengers may use the same ticket to perform several short activities. Thus, surveys are less prone to errors in this aspect, since they inquire about each leg of each trip separately. Some procedures could be incorporated to the proposed approach to mitigate these errors. For example, bus location from the AVL system can be used in order to check whether a given passenger boarded the first arriving bus while doing a transfer. If several buses went through the bus stop and were not boarded by the passenger, this might be an indicator that the passenger was performing a short activity and not actually doing a bus transfer to get to a final destination.

5.3 OD matrix visualization tool

As outlined in Section 2.3, the last step of the urban data analysis workflow involves presenting results visually to communicate the main findings and to help stakeholders make decisions that can shape the studied reality (Schutt and O'Neil, 2013). For this purpose, an interactive web application was developed to show the computed OD matrices in an intuitive and friendly manner. The OD visualization tool allows users to select an area in the map and creates a heatmap indicating the number of passengers traveling from the selected area to all other areas in the map. The tool was developed in Python using Pandas for data processing, Geopandas to display the map of the city and the administrative divisions, and Bokeh to provide interactivity to the visualization. The web application is freely available at `www.fing.edu.uy/~renzom/msc`. Figure 5.6 shows the user interface of the developed tool and its main components are described next.

The OD visualization tool offers several tools for users to filter data using different criteria prior to plotting. Firstly, the canvas of the plot supports multiple tabs. These tabs are used to select the level of aggregation for the OD data. Users can select between a coarse-grain visualization consisting of municipalities or a finer-grain aggregation consisting of census segments. When a tab is selected the map is updated accordingly, to show the city division selected by the user. The map area has pan and zoom capabilities, which can be toggled on or off using the buttons located on the bottom right of the canvas. Secondly, users can select ranges of dates as well as ranges of hours in the day to consider in the visualization. These selections are done in a straightforward fashion, using range sliders to indicate the exact time frame to be plotted. Additionally, users can select the type of day to consider for the visualization among three pre-defined types, namely, *all days*, *working days*, or *weekends*.

After indicating the desired options the user can select an area (i.e., a municipality or a census segment) by clicking on the map. Then, the selected area is shown in a different color for the user to confirm the selection. Once confirmed, the application updates the color of all the areas in the map according to the amount of trips done from the selected area, considering the date, time, and type of day preferences indicated before. A color bar is presented on the right to quantify the information visually displayed. Additionally, the applica-

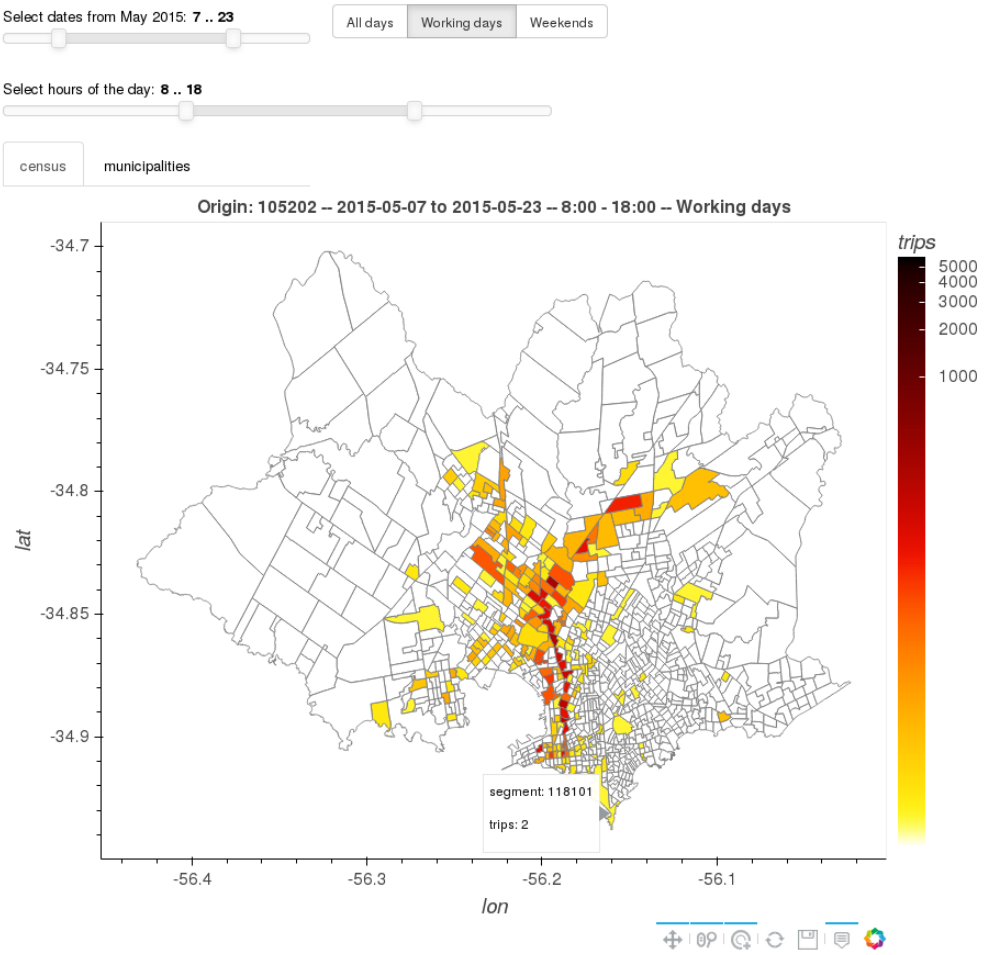


Figure 5.6: User interface of the OD matrix visualization tool

tion offers a hover tool, which displays information when the mouse cursor is over a given area. The displayed information includes the area identifier (name of the municipality or id of the census segment) as well as the exact number of trips with that destination. Finally, at every step of the visualization the user is able to export the displayed map as an image using the save button in the bottom right panel of the map.

5.4 Practical use cases

OD matrices are a key input for any optimization problem involving public transportation (e.g., bus network redesign, stop location, timetable scheduling) and to provide evidence that supports decisions aiming at improving the QoS offered to citizens. In this section, two straightforward examples of the

potential use of OD matrices are presented.

5.4.1 Bus line load profile

The set of identified trips using the destination estimation algorithm can be filtered by the bus line identifier. By doing so, all OD pairs for a given bus line can be obtained. This subset of the results holds the information of every boarding and alighting at each bus stop of a given line. A measure of the load of the line at each bus stop can be obtained when subtracting the number of alightings to the number of boardings. Figure 5.7 shows the load profile of line 183 (code 1303), the most frequently used line as shown in Section 4.3.1, considering all the identified trips of May 2015. The bus stops are represented in the X-axis in the order that are visited by the bus line, with bus stop 0 being the origin of the line. The Y-axis represents the difference between the number of passengers boarding and alighting at each bus stop.

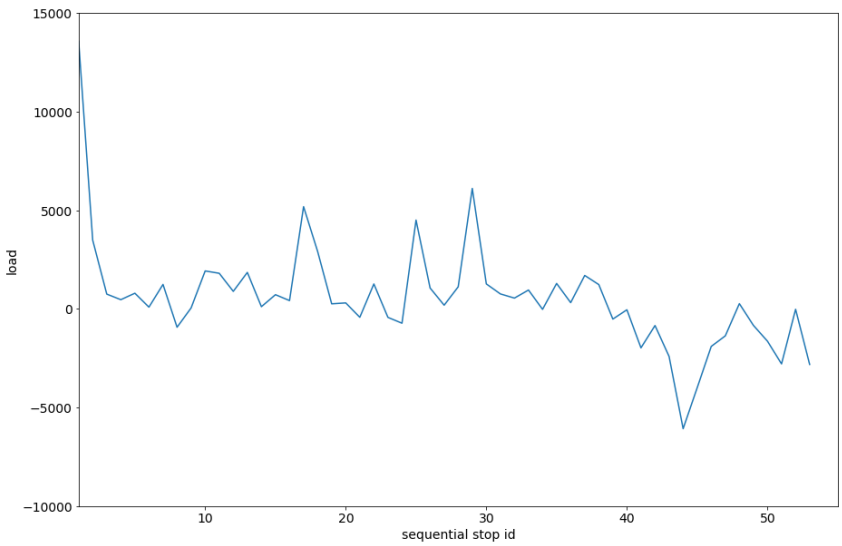


Figure 5.7: Load profile of line 183 (1303) in May 2015

Several interesting remarks arise when looking at the load profile of the bus line shown as an example. As expected, the first stop of the bus line has a significant higher number of boardings than alightings. Along the route of the bus line, three peaks can be clearly identified which have more boardings than alightings. This bus stops may correspond to areas used for bus transfers or that are located near attractions that generate large number of trips (e.g., shopping malls). Special actions can be taken to reduce the time of boarding

for passengers at this identified stops, for instance, by implementing closed bus stops and allowing passengers to pay for their tickets in advance, prior to boarding the bus. Finally, it can be noted that the bus stop where more alightings occur compared to boardings is not the last stop of the bus line, but nearly 10 stops before the end of the route. This information could be used when planning a redesign of the bus network, for example, to shorten the bus line.

5.4.2 Transfers

When identifying trips using the destination estimation algorithm, bus transfers were recorded separately, including the pair of bus lines involved in the transfer as well as the alighting and boarding stops used by the passenger. This information is highly valuable to identify pair of lines and locations where citizens transfer the most. Studying frequent transfer points might bring to light bad designs in the bus network topology that may force passengers to transfer excessively due to the absence of direct lines. Additionally, bus timetables can be synchronized in highly demanded point in order to reduce the waiting times of passengers during transfers. Table 5.3 shows the ten most frequent transfers according to data of May 2015. The identifiers of both lines involved in the transfer are presented, along with the bus stop identifier of the alighting stop of the first bus and the boarding stop of the second bus. For each identified transfer, the number of occurrences in the studied dataset is outlined.

Table 5.3: Top ten most frequent bus transfers in May 2015

lines		*stops*		*# transfers*
first	*second*	*alighting*	*boarding*	
1096	2579	2427	2426	8055
1759	1667	1885	4468	4641
1122	650	1108	4775	4192
1759	987	1942	4399	3823
170	1290	4212	5709	3806
1276	1092	2447	2437	3778
170	2579	4212	5709	3754
1122	1347	4843	4930	3613
1096	1290	2427	2426	3342
2050	1418	2392	2295	2927

The most popular bus transfer outnumbers the occurrences of the second most frequent transfer in 42.4%. This bus transfer corresponds to lines 158 (id: 1096) and 183 (id: 2579). The transfer takes place in the intersection of Burgues and Luis Alberto de Herrera, two main avenues of the city. The first bus line, 158, starts at Gruta de Lourdes in a lower-income area of the city, whereas line 183 ends in Pocitos, a highly-populated upper-class neighborhood with many services. Figure 5.8 shows both bus lines and transfer bus stops on top a choropleth map indicating the percentage of UBN of each census segment along the route of the bus lines.

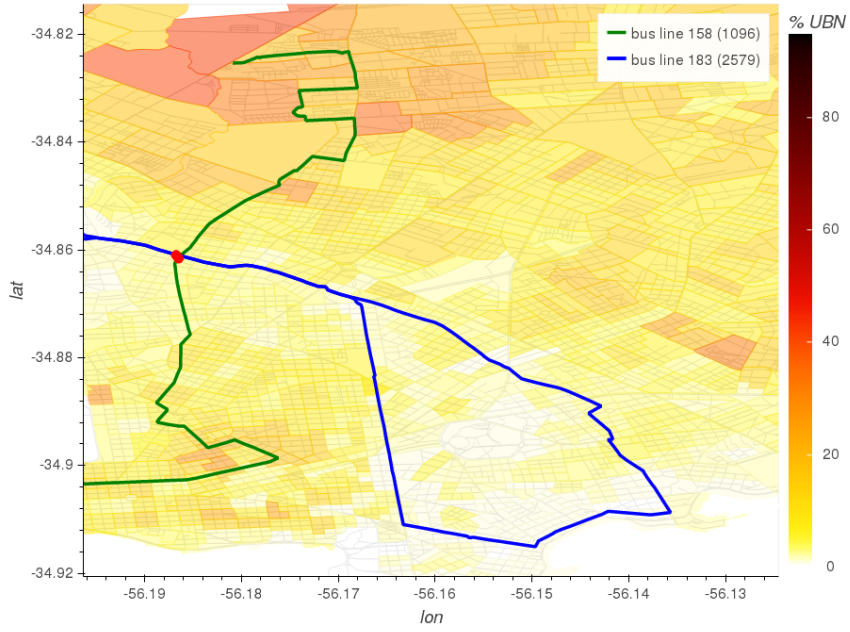


Figure 5.8: Most frequent bus transfer in May 2015. Passengers travel from areas with high percentage of UBN using line 158 (in green) and transfer to line 183 (in blue) at the bus stops marked in red to reach areas with better UBN indicators.

It can be clearly seen that the trip involving the bus transfer allows citizens from areas of the city with higher percentage of UBN to reach areas near the coastline with better socioeconomic levels. The large number of transfers might be due to people commuting to this area of the city where more job opportunities are located. This example shows how authorities can take advantage of information arising from estimated OD matrices to improve the quality of the transportation system. For instance, a direct line could be implemented to link these areas of the city or a better transfer option could be implemented that does not imply such a large detour.

Chapter 6

Conclusions and future work

This closing chapter outlines the main findings and conclusions resulting from the research reported in this thesis, along with the main lines of future work.

6.1 Conclusions

This thesis studied mobility in Montevideo, Uruguay, following an urban data analysis approach.

An intense urban expansion process has been taking place worldwide since 1950, with populations shifting from rural to urban residencies. In urban scenarios, public transportation systems play a major role in mobility, since they constitute the most sustainable and socially fair mode of transportation. Under the paradigm of smart cities, ITS have emerged to take advantage of information and communication technologies to improve public transportation. ITS allow collecting massive amounts of urban data, which can be used to extract meaningful information to help decision making in cities. This thesis studied data from the ITS in Montevideo, Uruguay, in order to characterize mobility in the city.

The studied dataset from the ITS in Montevideo consisted of GPS bus location data and smart card ticket sales data corresponding to the full year of 2015, accounting for over 150 GB. The results reported in this thesis correspond to the dataset of May 2015, accounting for more than 20.4 million bus tickets sold using smart cards. During a data cleansing process, 1.53% of the records were filtered due to inconsistencies. Several insights were obtained through data analysis of the studied dataset, including: number of passengers traveling

with the same smart card, frequency of use of the smart cards, number of bus transfers, number of transactions per bus trip, and most used bus lines and stops. A temporal analysis of ticket sales was performed, identifying three peak hours during working days, namely, morning, midday, and afternoon. These peak hours were also identified in the 2016 urban mobility survey (Mauttone and Hernández, 2017). Additionally, a spatial analysis of ticket sales was performed focusing on the city center, a troublesome area with regards to mobility. Then, both dimensions were combined into a spatiotemporal analysis which revealed that citizens from areas farther away from the coastline start trips earlier than those near the coast. Finally, some practical examples on the use of data analysis on ITS data were presented, including: anomaly detection in space (to identify bus detours), anomaly detection in time (to identify events in the city), and a characterization of driving behaviors and potential safety hazards due to reckless driving.

Besides using ITS data in a purely descriptive fashion to characterize the public transportation system, a methodology for building OD matrices was proposed and implemented. The main challenge when building OD matrices using ticket sales data from ITS is that passengers validate their cards when boarding but not when alighting the bus. Thus, while the origin is known, the destination of the trip must be estimated. For this purpose, a trip chaining algorithm was developed based on previous works on the related literature. The algorithm receives as input a historical set of sales records and aims to link together trips done by the same cardholder based on the following two assumptions: i) the origin of a trip is near the destination of the previous one; ii) the destination of the last trip in the day is near the origin of the first trip of the day. The implemented algorithm was able to estimate the destination for 81.62% of trips in the studied dataset, which is a highly competitive result when compared to other figures reported in the related literature. Grouping the trips identified by the algorithm, OD matrices were built at different levels of granularity, i.e., the bus stop, census segment, and municipality levels.

The computed OD matrix was compared against the one resulting from the 2016 urban mobility survey. Results showed many similarities between both approaches, suggesting that OD matrices built using ITS data are a valid alternative to understand mobility in a city. Both methods identified municipality B as the largest generator and attractor of trips in the city, as well as a large number of intra-municipality trips. Several advantages can be high-

lighted from the proposed approach. Firstly, taking advantage of ITS data allows studying mobility at a finer grain, obtaining OD matrices between pairs of bus stops (4718×4718) and census segments (1063×1063), whereas the OD matrix built using data from the mobility survey applies to the level of municipalities (8×8). Secondly, the proposed approach allows computing OD matrices considering different criteria, e.g., building OD matrices for specific dates, times of the day, or group of bus lines. Moreover, generating OD matrices using ticket sales data is inexpensive if the ITS infrastructure is already present. This was the case of Montevideo, where the ITS infrastructure has been deployed for nearly a decade. Finally, computing OD matrices using ITS data allows obtaining up-to-date mobility information, since new matrices can be built whenever new data is present or even in a near real-time fashion with streaming sources of data.

In order to communicate the main findings of the proposed approach, an interactive web application was developed to visually display the computed OD matrices in an intuitive way to citizens and authorities alike. The visualization tool allows users to select an area in the map of Montevideo and displays a heatmap indicating the number of passengers traveling from the selected area to all other parts of the city. The application supports working at the census segment and municipality level of aggregation for OD matrices and offers several tools to filter data, including: range of dates, range of hours in the day, and type of day (i.e., all days, working days only, or weekends only). Besides displaying the information visually, users can inspect each area to retrieve the exact number of trips with that destination. Finally, at each step of the visualization the displayed information can be exported as an image and downloaded.

In order to demonstrate the use of ITS data to understand mobility in a city two practical examples were presented. The first example considered all trips done on the most frequently used bus line in Montevideo. For this bus line a load profile was generated, indicating the number of boardings and alightings along the route of that line. The second example involved studying the most frequent bus transfers done by passengers. This example allowed identifying one bus transfer frequently made by citizens from a socioeconomically vulnerable part of the city in order to reach a high-income neighborhood.

Summarizing, the main contributions of the research reported in this thesis are:

- A review of related works regarding urban mobility, specifically, on OD matrix generation using ITS data.
- An urban data analysis of mobility in Montevideo, Uruguay, using ITS data from the public transportation system.
- An algorithm to estimate trip destinations using ticket sales transactions and bus location data.
- Estimated OD matrices for the public transportation system of Montevideo and their validation against a household mobility survey.
- An interactive web application to visually display the computed OD matrices.

The work reported in this thesis resulted in several publications including two journal articles (Nesmachnow et al., 2017; Massobrio et al., 2018) and three conference papers (Massobrio et al., 2016; Massobrio and Nesmachnow, 2016; Fabbiani et al., 2017), which address topics included in this manuscript and other related lines of work. Additionally, a collaboration with a research group at Centro de Investigación Científica y de Educación Superior de Ensenada in Mexico was established. The results from the data analysis process reported in this thesis were used to address bus fleet scheduling and timetable synchronization problems, leading to a series of co-authored articles on the topic (Peña et al., 2016, 2017a,b, 2018).

6.2 Future work

The work reported in this thesis constitutes one of the first steps towards using data from the ITS in Montevideo to understand mobility in the city. As such, many lines of research remain to be explored in order to extract more and richer information that can be used to improve the public transportation system.

The data analysis reported in this thesis mainly focused in understanding the interaction between passengers and the transportation system. However, the available data sources offer the potential to study other very interesting aspects of mobility in the city. For instance, location data from AVL systems could be used to further study the QoS offered to citizens by the transportation system in terms of punctuality, frequency of lines, and load of passengers with

regards to the bus capacity. Additionally, speed information of buses could be used to characterize the streets of the city and identify bottlenecks. This information could be used as input when designing new lines or re-designing existing ones.

Regarding the estimation of OD matrices, several improvements could be made to the proposed approach. Data from tickets sold without smart cards should be used to expand the results from the estimated OD matrices to account for all passengers of the transportation system. This would make the comparison with the mobility survey fairer, since the survey accounted for passengers traveling with and withous smart cards. Additionally, the destination estimation algorithm could be further refined. Historical passenger data could be used when trip chaining fails to estimate the destination of a trip. For instance, frequent destinations could be inferred on a per-passenger basis using clustering or other machine learning techniques. This would allow to predict the destination of a trip based on past information of the same passenger. Another aspect that could be improved from the current approach involves transfers. In the proposed method, bus transfers are always considered as part of a trip with multiple legs. AVL data could be used to help identifying short individual trips done using a single ticket. Furthermore, other related optimization problems could be tackled using the OD matrices computed in this work, e.g., synchronizing bus schedules for transfers, incorporating demand data to optimize the fleet size and vehicle schedule of buses in the system, modifying the location of bus stops, and redesigning the bus line network.

Finally, it is worth noting that this work used ITS data from 2015. Since that year, the use of smart cards to pay for tickets has risen significantly. Consequently, OD matrices estimated using up-to-date data might be more representative of the universe of passengers using the public transportation system. The proposed approach should be applied to recent ITS data when it becomes available publicly. In this regard, this work intends to contribute with a small step towards authorities opening up more data.

Bibliography

Abreu, P. and Vespa, J. F. (2010). Plan de Movilidad. Technical report, Intendencia de Montevideo.

Ahn, J., Ko, E., and Kim, E. Y. (2016). Highway traffic flow prediction using support vector regression and bayesian classifier. In *International Conference on Big Data and Smart Computing*, pages 239–244.

Albino, V., Berardi, U., and Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, 22(1):3–21.

Alsger, A. A., Mesbah, M., Ferreira, L., and Safi, H. (2015). Use of smart card fare data to estimate public transport origin–destination matrix. *Transportation Research Record: Journal of the Transportation Research Board*, 2535:88–96.

Anda, C., Erath, A., and Fourie, P. J. (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences*, 21(1):19–42.

Bagchi, M. and White, P. (2005). The potential of public transport smart card data. *Transport Policy*, 12(5):464–474.

Barrionuevo, J. M., Berrone, P., and Ricart, J. E. (2012). Smart cities, sustainable progress. *IESE Insight*, 14(14):50–57.

Barry, J., Newhouser, R., Rahbee, A., and Sayeda, S. (2002). Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, 1817:183–187.

Bednar, J. A., Crist, J., Cottam, J., and Wang, P. (2016). Datashader: Revealing the Structure of Genuinely Big Data. Presentation, 15th Python in Science Conference.

Benevolo, C., Dameri, R. P., and D'Auria, B. (2016). Smart mobility in smart city. In Torre, T., Braccini, A. M., and Spinelli, R., editors, *Empowering Organizations*, pages 13–28. Springer International Publishing.

Blythe, P. T. (2004). Improving public transport ticketing through smart cards. In *Proceedings of the Institution of Civil Engineers-Municipal Engineer*, volume 157, pages 47–54.

Bokeh Development Team (2018). Bokeh: Python library for interactive visualization. Retrieved from: http://www.bokeh.pydata.org (Last accessed: 2018-08-30).

Boyandin, I., Bertini, E., Bak, P., and Lalanne, D. (2011). Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum*, 30(3):971–980.

Boyle, D. K. (2008). *Passenger counting systems*. Transportation Research Board.

Calvo, J. J. (2012). Atlas sociodemográfico y de la desigualdad del Uruguay. Technical report, Unidad Multidisciplinaria, Facultad de Ciencias Sociales.

Camagni, R., Gibelli, M. C., and Rigamonti, P. (2002). Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion. *Ecological economics*, 40(2):199–216.

Cardozo, O. D. and Rey, C. E. (2007). La vulnerabilidad en la movilidad urbana: aportes teóricos y metodológicos. In Foschiatti, A., editor, *Aportes conceptuales y empíricos de la vulnerabilidad global*, pages 398–423. Editorial Universitaria de la Universidad Nacional del Nordeste.

Chen, X., Pao, H., and Lee, Y. (2014). Efficient traffic speed forecasting based on massive heterogenous historical data. In *IEEE International Conference on Big Data*, pages 10–17.

Cocchia, A. (2014). Smart and digital city: A systematic literature review. In Dameri, R. P. and Rosenthal-Sabroux, C., editors, *Smart City: How to*

Create Public and Economic Value with High Technology in Urban Space, pages 13–43. Springer International Publishing.

Deakin, M. and Waer, H. A. (2011). From intelligent to smart cities. *Intelligent Buildings International*, 3(3):140–152.

Doyle, J., Hung, P., Kelly, D., McLoone, S. F., and Farrell, R. (2011). Utilising mobile phone billing records for travel mode discovery. In *22nd IET Irish Signals and Systems Conference*.

Fabbiani, E., Vidal, P., Massobrio, R., and Nesmachnow, S. (2017). Distributed big data analysis for mobility estimation in intelligent transportation systems. In Barrios Hernández, C. J., Gitler, I., and Klapp, J., editors, *High Performance Computing: Third Latin American Conference*, pages 146–160. Springer International Publishing.

Farzin, J. M. (2008). Constructing an automated bus origin–destination matrix using farecard and global positioning system data in São Paulo, Brazil. *Transportation Research Record*, 2072(1):30–37.

Ferreira, N., Poco, J., Vo, H. T., Freire, J., and Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158.

Figueiredo, L., Jesus, I., Machado, J. A. T., Ferreira, J. R., and de Carvalho, J. L. M. (2001). Towards the development of intelligent transportation systems. In *IEEE Intelligent Transportation Systems*, pages 1206–1211.

Furth, P., Hemily, B., Muller, T., and Strathman, J. (2006). Using archived AVL-APC data to improve transit performance and management. Technical Report 113, Transit Cooperative Research program-Transportation Research Board.

GeoPandas developers (2013). GeoPandas 0.4.0. Retrieved from: http://geopandas.org/ (Last accessed: 2018-08-30).

Grava, S. (2000). *Urban Transportation Systems*. McGraw-Hill Professional Publishing.

Guo, D. and Zhu, X. (2014). Origin-destination flow data smoothing and mapping. *IEEE transactions on visualization and computer graphics*, 20(12):2043–2052.

Harvey, D. (1992). Social justice, postmodernism and the city. *International Journal of Urban and Regional Research*, 16(4):588–601.

Huijboom, N. and Van den Broek, T. (2011). Open data: an international comparison of strategies. *European journal of ePractice*, 12(1):4–16.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *IEEE Computing in science & engineering*, 9(3):90–95.

Instituto Nacional de Estadística, Uruguay (2012). Resultados del censo de población 2011: población, crecimiento y estructura por sexo y edad. Retrieved from: http://www.ine.gub.uy/c/document_library/get_file?uuid=12d80f63-afe4-4b2c-bf5b-bff6666c0c80&groupId=10181 (Last accessed: 2018-08-30).

Intendencia de Montevideo (2017). El corazón de montevideo se renueva. Retrieved from: http://www.montevideo.gub.uy/institucional/noticias/el-corazon-de-montevideo-se-renueva (Last accessed: 2018-08-30).

Jagadeesh, G. R., Srikanthan, T., and Zhang, X. D. (2004). A map matching method for GPS based real-time vehicle location. *The Journal of Navigation*, 57(3):429–440.

Janssen, K. (2011). The influence of the psi directive on open government data: An overview of recent developments. *Government Information Quarterly*, 28(4):446–456.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 29(4):258–268.

Judd, C. M., McClelland, G. H., and Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P.,

Avila, D., Abdalla, S., Willing, C., and et al. (2016). Jupyter notebooks - a publishing format for reproducible computational workflows. In *20th International Conference on Electronic Publishing*, pages 87–90.

Kostakos, V., Camacho, T., and Mantero, C. (2010). Wireless detection of end-to-end passenger trips on public transport buses. In *IEEE Conference on Intelligent Transportation Systems*, pages 1795 – 1800.

Li, D., Lin, Y., Zhao, X., Song, H., and Zou, N. (2011). Estimating a transit passenger trip origin-destination matrix using automatic fare collection system. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications*, pages 502–513.

Li, T., Sun, D., Jing, P., and Yang, K. (2018). Smart card data mining of public transport destination: A literature review. *Information*, 9(1):18.

Lu, D. (2008). *Route level bus transit passenger origin-destination flow estimation using apc data: Numerical and empirical investigations*. PhD thesis, The Ohio State University.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12.

Ma, X.-l., Wang, Y.-h., Chen, F., and Liu, J.-f. (2012). Transit smart card data mining for passenger origin information extraction. *Journal of Zhejiang University SCIENCE C*, 13(10):750–760.

Massobrio, R. and Nesmachnow, S. (2016). Análisis de datos de movilidad del transporte público de montevideo. In *XIX Congreso Latinoamericano de Transporte Público y Urbano*, pages 1–11.

Massobrio, R., Nesmachnow, S., Tchernykh, A., Avetisyan, A., and Radchenko, G. (2018). Towards a cloud computing paradigm for big data analysis in smart cities. *Programming and Computer Software*, 44(3):181–189.

Massobrio, R., Pías, A., Vázquez, N., and Nesmachnow, S. (2016). Map-reduce for processing GPS data from public transport in Montevideo, Uruguay. In *45º Jornadas Argentinas de Informática*, pages 41–54.

Mauttone, A. and Hernández, D. (2017). Encuesta de movilidad del área metropolitana de Montevideo. Principales resultados e indicadores. Technical report, CAF, Intendencia de Montevideo, Intendencia de Canelones, Intendencia de San José, Ministerio de Transporte y Obras Públicas, Universidad de la República, PNUD Uruguay. Retrieved from: http://scioteca.caf.com/handle/123456789/1078 (Last accessed: 2018-08-30).

McKinney, W. (2010). Data structures for statistical computing inPython. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56.

Munizaga, M., Devillaine, F., Navarrete, C., and Silva, D. (2014). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, 44:70–79.

Munizaga, M. and Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18.

Nassir, N., Hickman, M., and Ma, Z.-L. (2015). Activity detection and transfer identification for public transit fare card data. *Transportation*, 42(4):683–705.

Nesmachnow, S. (2010). Computación científica de alto desempeño en la Facultad de Ingeniería, Universidad de la República. *Revista de la Asociación de Ingenieros del Uruguay*, 61(1):12–15.

Nesmachnow, S., Baña, S., and Massobrio, R. (2017). A distributed platform for big data analysis in smart cities: combining Intelligent Transportation Systems and socioeconomic data for Montevideo, Uruguay. *EAI Endorsed Transactions on Smart Cities*, 2(5):1–18.

Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20.

Ortúzar, J. D. D., Armoogum, J., Madre, J., and Potier, F. (2011). Continuous mobility surveys: The state of practice. *Transport Reviews*, 31(3):293–312.

Ortúzar, J. d. D. and Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons.

Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Peña, D., Tchernykh, A., Nesmachnow, S., Massobrio, R., Drozdov, A. Y., and Garichev, S. N. (2016). Multiobjective Vehicle Type and Size Scheduling Problem in Urban Public Transport Using MOCell. In *International Conference on Engineering and Telecommunication*, pages 110–113.

Peña, D., Tchernykh, A., Nesmachnow, S., Massobrio, R., Drozdov, A. Y., and Garichev, S. N. (2017a). Multiobjective Optimization of Urban Public Transport Using MOCell. In 8^{th} *International Supercomputing Conference in Mexico*, pages 1–3.

Peña, D., Tchernykh, A., Nesmachnow, S., Massobrio, R., Feoktistov, A., and Bychkov, I. (2017b). Multiobjective vehicle-type scheduling in urban public transport. In *IEEE International Parallel and Distributed Processing Symposium Workshops*, pages 482–491.

Peña, D., Tchernykh, A., Nesmachnow, S., Massobrio, R., Feoktistov, A., Bychkov, I., Radchenko, G., Drozdov, A. Y., and Garichev, S. N. (2018). Operating cost and quality of service optimization for multi-vehicle-type timetabling for urban bus systems. *Journal of Parallel and Distributed Computing*. (In press).

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Presidencia de la República (2010). Decreto 232/010: Reglamentacion de la ley sobre el derecho de acceso a la informacion publica. Registro Nacional de Leyes y Decretos, 1(2):394.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23:3–13.

Schutt, R. and O'Neil, C. (2013). *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, Inc.

Servicio de Geomática - Intendencia de Montevideo (1996). Ejes de Calles. [Data file] Retrieved from: http://geoweb.montevideo.gub.uy/geonetwork/srv/es/metadata.show?uuid=828dd68a-2cd5-4ac1-b754-37ecde6f4cf1 (Last accessed: 2018-08-30).

Servicio de Geomática - Intendencia de Montevideo (2011). Municipios de montevideo. [Data file] Retrieved from: http://geoweb.montevideo.gub.uy/geonetwork/srv/es/metadata.show?uuid=b0a2cf85-af7a-4aac-998f-da124ac7d073 (Last accessed: 2018-08-30).

Servicio de Geomática - Intendencia de Montevideo (2012a). Líneas de Transporte. [Data file] Retrieved from: http://geoweb.montevideo.gub.uy/geonetwork/srv/es/metadata.show?uuid=307ffef2-7ba3-4935-815b-caa7057226ce (Last accessed: 2018-08-30).

Servicio de Geomática - Intendencia de Montevideo (2012b). Paradas de Ómnibus. [Data file] Retrieved from: http://geoweb.montevideo.gub.uy/geonetwork/srv/es/metadata.show?uuid=c6ea0476-9804-424a-9fae-2ac8ce2eee31 (Last accessed: 2018-08-30).

Servicio de Geomática - Intendencia de Montevideo (2014a). Hogares con NBI por segmento 2011. [Data file] Retrieved from: http://geoweb.montevideo.gub.uy/geonetwork/srv/es/metadata.show?uuid=e5f03b3f-7106-4d5f-9443-63b0cdc3a81b (Last accessed: 2018-08-30).

Servicio de Geomática - Intendencia de Montevideo (2014b). Personas por zona 2011. [Data file] Retrieved from: http://geoweb.montevideo.gub.uy/geonetwork/srv/es/metadata.show?uuid=e3140ca2-21f0-4a9d-9be5-4da416c3ab23 (Last accessed: 2018-08-30).

Shi, Q. and Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 58:380–394.

Sussman, J. S. (2008). *Perspectives on intelligent transportation systems (ITS)*. Springer Science & Business Media.

TIOBE (2018). TIOBE Index for August 2018. Retrieved from: https://www.tiobe.com/tiobe-index/ (Last accessed: 2018-08-30).

Transport for London (2018). London travel demand survey. Retrieved from: https://tfl.gov.uk/corporate/about-tfl/how-we-work/planning-for-the-future/consultations-and-surveys/london-travel-demand-survey (Last accessed: 2018-08-30).

Trépanier, M., Morency, C., and Agard, B. (2009). Calculation of transit performance measures using smartcard data. *Journal of Public Transportation*, 12(1):5.

Trépanier, M., Tranchant, N., and Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14.

Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

United Nations (2018). World urbanization prospects: The 2018 revision. Retrieved from: https://esa.un.org/unpd/wup/ (Last accessed: 2018-08-30).

Utsunomiya, M., Attanucci, J., and Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board*, 1971:119–126.

Wang, H., Calabrese, F., Lorenzo, G. D., and Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 318–323.

Wang, W., Attanucci, J., and Wilson, N. (2011). Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, 14(4):131–150.

Washburn, D. and Sindhu, U. (2009). Helping CIOs understand "smart city" initiatives. *Growth*, 17(2):1–17.

Xia, D., Wang, B., Li, H., Li, Y., and Zhang, Z. (2016). A distributed spatial-temporal weighted model on mapreduce for short-term traffic flow forecasting. *Neurocomputing*, 179:246–263.

Zhao, Y. (1997). *Vehicle location and navigation systems*. Artech House Publishers.

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., and Yang, L. (2016). Big data for social transportation. *IEEE Transactions on Intelligent Transportation Systems*, 17(3):620–630.