Urban mobility data analysis in Montevideo, Uruguay

Renzo Massobrio
Universidad de la República
Uruguay
renzom@fing.edu.uy

***Abstract*—Understanding the interaction between citizens and transportation systems is crucial for policy-makers that aim to improve mobility in a city. Modern urban transportation systems incorporate technologies that generate huge volumes of data, which can be processed to characterize mobility of its users. This article summarizes the main findings of a M.Sc. thesis which studied 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. Furthermore, origin-destination matrices, which describe mobility patterns in the city, are generated using a destination estimation algorithm, 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.**

***Index Terms*—urban mobility, smart cities, intelligent transportation systems, urban data analysis, origin-destination matrix**

I. Introduction

Since 1950 populations have been steadily shifting from rural to urban residencies, in a worldwide process known as urbanization [1]. 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 [2]. 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 [3]. 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. Urban traffic and transportation systems are generally addressed under the paradigm of *smart cities*, in what is referred to as *smart mobility* [4]. 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 [5]. 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 [6]. 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.

The main goal of this research 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 remainder of this article is structured as follows. Section II reviews the literature related to urban data analysis and to the problem of generating OD matrices that describe mobility using ITS data. Afterwards, Section III presents a study of the transportation system of Montevideo, Uruguay, following an urban data analysis approach. Later, Section IV outlines the OD matrix estimation process using sales and vehicle location data from the public transportation system of Montevideo, Uruguay. Finally, Section V states the conclusions and outlines the main lines of future work.

II. Related Work

This section outlines the review of the related works in the use of data analysis applied to understand and improve urban mobility. The review also covers works that address the problem of generating OD matrices and the mobility information available in Montevideo, Uruguay.

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. [7]. 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 *crowd-sourcing* 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.

The use of smart cards to simplify the process of paying for tickets or fares is steadily rising in public transportation systems. Pelletier et al. provided a thorough literature review on the topic, including the most used technologies, privacy and legal concerns related to these systems, and several applications that use smart card data from public transportation systems [8]. Bagchi and White discussed the role of smart card data for travel behavior analysis [9]. Two datasets from the transportation systems of Southport, Merseyside and Bradford in England were used, which accounted for nearly 3500 cardholders. The authors studied the average number of trips and transfers made by passengers, as well as 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.

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 Automatic Passenger Counters (APC) systems to estimate OD flows from detailed boarding and alighting counts [10], [11]. However, since entering and exiting data cannot be assigned to individual passengers, most of the proposed models 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 [12], [13] or combine CDR with data from ITS [14]. Other works have proposed using Bluetooth antennas to detect when mobile devices enter and exit vehicles [15]. However, antennas must be installed on vehicles to detect on-board devices and noise from mobile devices outside the vehicles must be filtered for accurate passenger sensing.

Despite the variety of sources that have been used to estimate OD matrices, the majority of the literature focuses on methods that involve using smart card data from Automatic Fare Collection (AFC) systems. 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. presented an up-to-date survey of the literature related to destination estimation techniques for OD matrix generation using smart card data [16]. 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 research, 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, thus, it is not suitable to analyze boarding and alighting of individual passengers. 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. [17]. 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. proposed using the trip chaining model for estimating the destination for passengers boarding buses with smart cards, following a database programming approach [18]. 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 could be estimated for 66% of the trips (80% if considering only peak hours). However, the real estimation accuracy could not be assessed due to the lack of a second source of data (e.g., surveys) for comparison.

Later, Wang et al. proposed using the trip chaining method to infer bus passenger origin-destination from smart card transactions from London, United Kingdom [19]. Origins were accurately determined by searching for the timestamp of each smart card transaction in the bus location records and trip chaining was used 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 [20]. 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: identifying daily load/flow variations and evaluating the average time that users need to wait for transferring between buses.

More recently, Munizaga and Palma applied trip chaining to estimate OD matrices in the multimodal transportation system of Santiago, Chile [21]. 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 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 [22]. 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.

Regarding the case studied in this research, a metropolitan household survey was conducted in Montevideo, Uruguay in 2016, with the goal of updating mobility information, which dated back to 2009 [23]. The survey aimed at characterizing mobility in the city, considering all modes of transportation and also comprising 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 Quality of Service (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. The main findings of the urban data analysis described in Section III and from the OD matrix estimation process described in Section IV are compared to those obtained from the mobility survey.

The analysis of related works allows identifying several proposals for using data analysis in the context of ITS to understand and improve urban mobility. The main contribution of this article 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 article contributes with a novel proposal to assess the transportation system and understand mobility patterns in Montevideo, Uruguay.

III. Urban mobility data analysis

This section presents a study of the transportation system of Montevideo, Uruguay, following an urban data analysis approach. The case study, methodology, and implementation details are described and the main findings of the analysis related to describing the use of the transportation system are outlined.

A. *Overview of the case study*

Montevideo is one of the nineteen departments in Uruguay and includes the capital city of the country. Located in the southernmost part of Uruguay, Montevideo extends to an area of only 530 km^2. From an administrative point of view, Montevideo is comprised of eight municipalities. A finer-grain division, mostly used in census and surveys, separates Montevideo into 1063 zones named *census segments*. 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 [24]. 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. Fig. 1 shows a choropleth map of the population density in Montevideo [25].

Describing the population of Montevideo from a socioeconomic point of view is not a simple task and is out of the scope

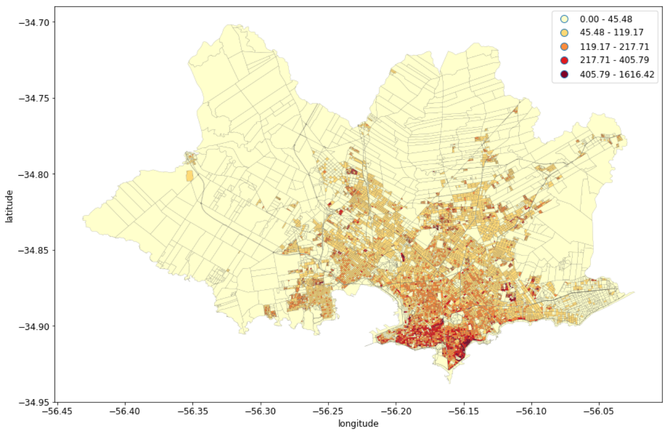


Fig. 1: Population density in Montevideo, Uruguay (inhabitants per ha)

of this research. 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 [26]. Fig. 2 shows a choropleth map of Montevideo indicating the percentage of households with UBN [27]. It is clear that the most vulnerable citizens are located farther away from the coast and the city center, in sparsely populated areas.

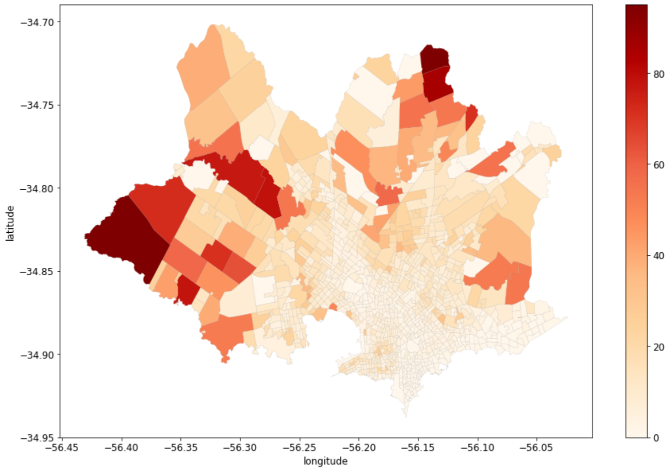


Fig. 2: Percentage of households with UBN in Montevideo, Uruguay

City authorities in Montevideo proposed in 2010 an urban mobility plan with the goal of restructuring and modernizing public transportation [28]. Within this plan, public transportation in Montevideo was integrated into a unified system named Sistema de Transporte Metropolitano (STM), which is comprised of 1528 buses operated by four private companies.

The bus network consists of 145 bus lines and 4718 bus stops. 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. Fig. 3 shows the bus lines that comprise STM [29], on top of a road map [30]. 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), 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.

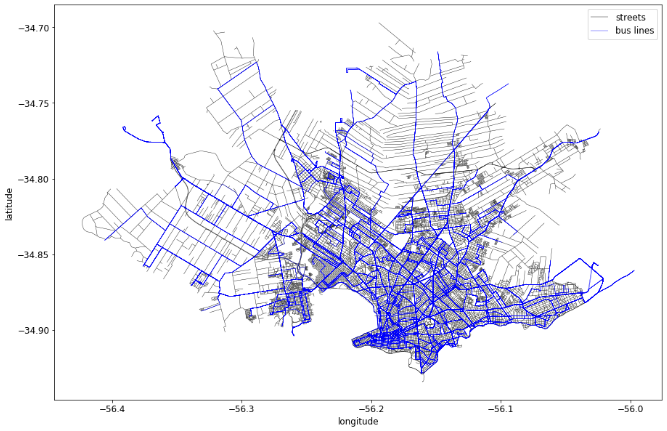


Fig. 3: Bus lines in STM

With the creation of STM, fares were redefined to provide passengers with more flexibility when traveling. Firstly, smart cards were introduced to allow passengers to pay for tickets without using physical money. STM smart cards are contactless top-up cards which are linked to the identity of the owner (a valid government ID or passport is required to get one). 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.

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 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 Section IV. 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.

B. Data analysis methodology

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. In this research the data science workflow proposed by Schutt and O'Neil was used [31].

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, placed in data into structures (e.g., tables), inspected, and cleansed to detect missing or inaccurate records. Afterwards, Exploratory Data Analysis (EDA) is performed, which aims at describing what data can tell beyond the formal modeling and hypothesis testing phase [32]. 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. After EDA, statistical models and algorithms are applied to identify relationships between the studied data [33]. Finally, results are interpreted and communicated, mostly using visualization techniques [34]. Urban data analysis, which demands combining quantitative and qualitative data, requires advanced means of visualizing results for effective communication. Additionally, since urban data usually has a prevalence of geographic components, urban data visualization combines classic statistical graphics and charts with Geographic Information Systems (GIS).

C. Computing infrastructure

The majority of the urban data analysis reported in this article was performed using Python. Several data analysis libraries were combined into an integrated ecosystem by using the Jupyter Notebook [35]. 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.

Due to the large volumes of data included in the analysis, special hardware infrastructure was required, particularly for the OD matrix generation described in Section IV. Data was processed over the cloud infrastructure at *ClusterUY*, the national center of supercomputing in Uruguay [36]. The main goal of ClusterUY is to provide support for solving complex problems that require large computing power. Specifically, the computations described in Section IV were performed using a server comprised of 40 Intel Xeon Gold 6138 (2.00GHz) cores and 128 GB of RAM.

Following a reproducible research methodology [37], Jupyter Notebooks corresponding to the data analysis process were hosted at a GitLab server. Additionally, interactive plots generated during the data analysis process were published as standalone applications. All these resources are freely available at www.fing.edu.uy/~renzom/msc.

D. Data analysis process

This section describes the urban data analysis process following the workflow outlined in Section .

1) Data collection: On August 2010, a presidential decree was published which regulated public access to state-owned information [38]. Following its publication, several initiatives have been taken to strive to open up data to the public at all levels of the public administration. In the context of this research, the most useful web interface was the geographic information site at www.sig.montevideo.gub.uy, which holds geographic data of Montevideo including base maps, socioeconomic indicators, and transportation network data.

Besides using open data publicly available, the analysis included data regarding STM accessed through a collaboration with the city authorities in Montevideo. 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.

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.

2) Exploratory Data Analysis (EDA): An initial EDA was performed to characterize the dataset of sales with STM cards. Fig. 4 shows an aggregated visualization of the geolocation of 20.4 million sales corresponding to May 2015. 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 www.fing.edu.uy/~renzom/msc.

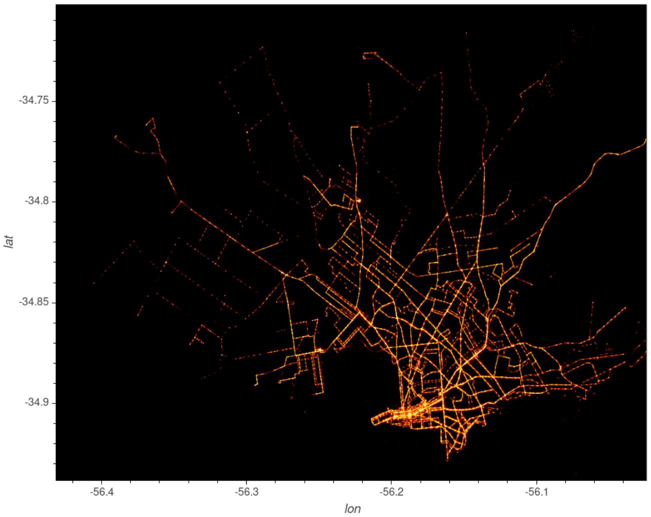


Fig. 4: Aggregated sales with STM cards in May 2015

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.

3) Data cleansing: 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.

As described in Section III-D1, the studied dataset holds the GPS measure at the time of the transaction. Vehicle location using GPS is prone to errors from a variety of sources [39]. However, 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 Section IV 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. 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.

Some transactions had trip identifiers which were not present in the GPS records. 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.

E. 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.

1) Cardholders: The sales dataset holds transactions made with 654.228 different STM cards. Despite the STM system allows several passengers to travel together using 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 Section IV, where all passengers under the same STM card are assumed to travel from origin to destination without splitting.

Another interesting aspect that can be studied through data analysis is the frequency of use of the transportation system. Table I 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. 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 I: 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. 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.

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.

2) Transfers: As introduced in Section III-A, 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. 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 amount of transactions involving more than two bus transfers is 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 II, the average number of legs when traveling by bus is 1.5 [23]. The slight difference between both estimations might be explained due to the fact that the mobility survey considers 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 studied 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.

3) Temporal analysis of transactions: The STM system records the date and time of each transaction. These data allows analyzing the distribution of transactions across time. Data show that the largest concentration of transactions corresponds to working days, with an average of ∼33.15M of transactions and a median of ∼34.41M. 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. Fig. 5 shows an histogram with the number of STM transactions at each hour of the day during May 2015.

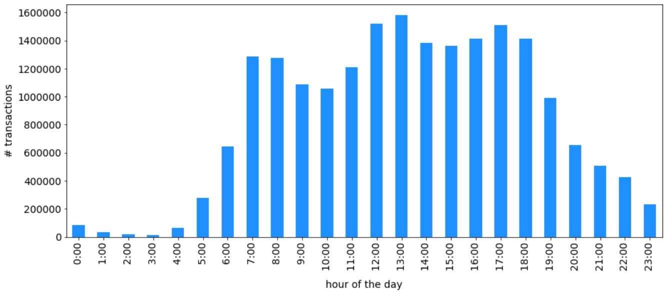


Fig. 5: Histogram of sales with STM cards at different times of the day during May 2015

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 at midday (12.00–13.00) which might not be foreseen prior to the analysis. In fact, the overall largest amount of transactions occurs 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 Section IV, 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 II. Fig. 6 shows the histogram of starting time of trips according to the urban mobility survey [23]. 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.

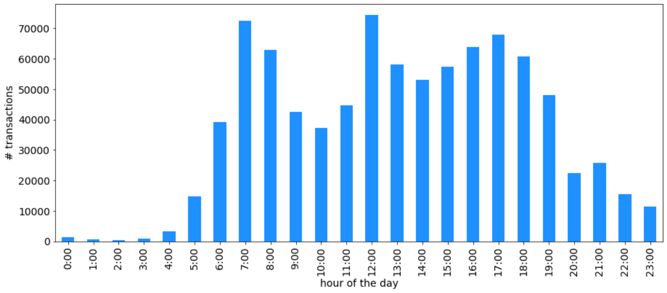


Fig. 6: Histogram of starting times of trips in public transportation according to the Urban Mobility Survey [23]. Aggregated data provided by the authors. Raw data are available at https://catalogodatos.gub.uy/dataset/encuesta-origen-destino-montevideo

A different picture is obtained when studying weekends independently. Fig. 7 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 Fig. 5. 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 unfeasible to perform through surveys. However, ITS usually generate data as a by-product, since their main goal is not collecting data but 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. 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.

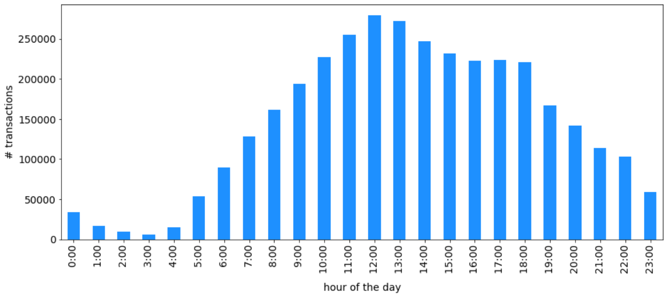


Fig. 7: Histogram of sales with STM cards at different times of the day during weekends of May 2015

4) 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 III-A. This design leads to major congestion at peak hours in 18 de Julio, the main avenue in the city center [40]. City authorities in Montevideo are concerned with the mobility issues in the city center and have proposed a plan to significantly alter the infrastructure of 18 de Julio [41]. Decisions to address this kind of issues could be supported with evidence resulting from urban data analysis processes, as described next.

Fig. 8 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 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 significant 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.

5) 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. Fig. 9 shows an aggregated

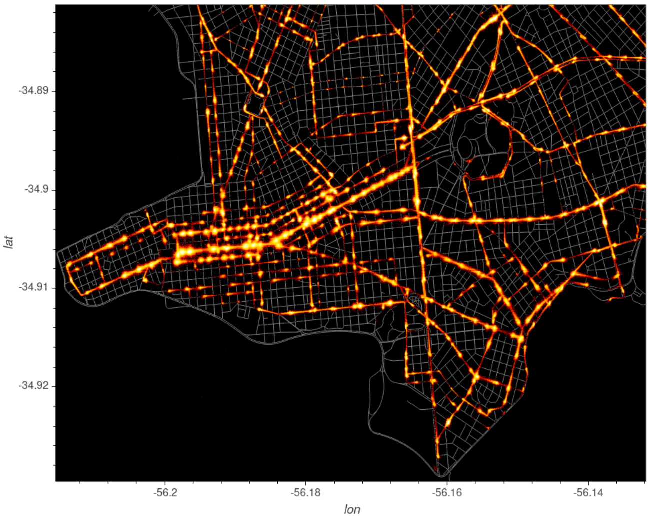


Fig. 8: Aggregated sales with STM cards in the city center during May 2015

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 www.fing.edu.uy/~renzom/msc.

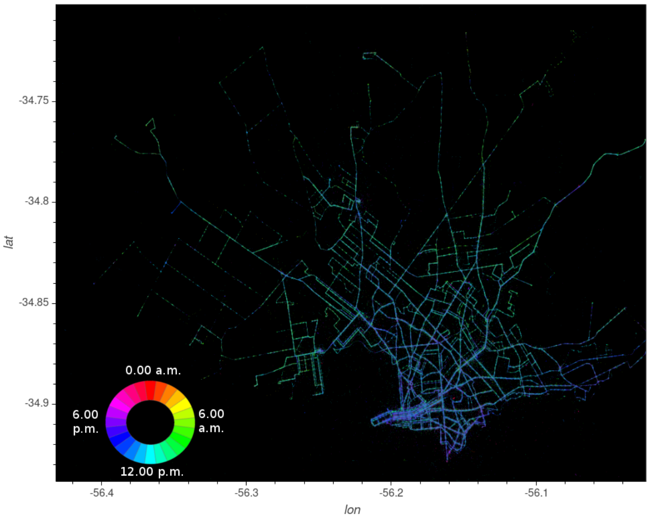


Fig. 9: Spatiotemporal distribution of trips during May 2015

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. Figs. 10 and 11 show choropleth maps of the number of transactions occurring in each census segment in the morning and evening, respectively.

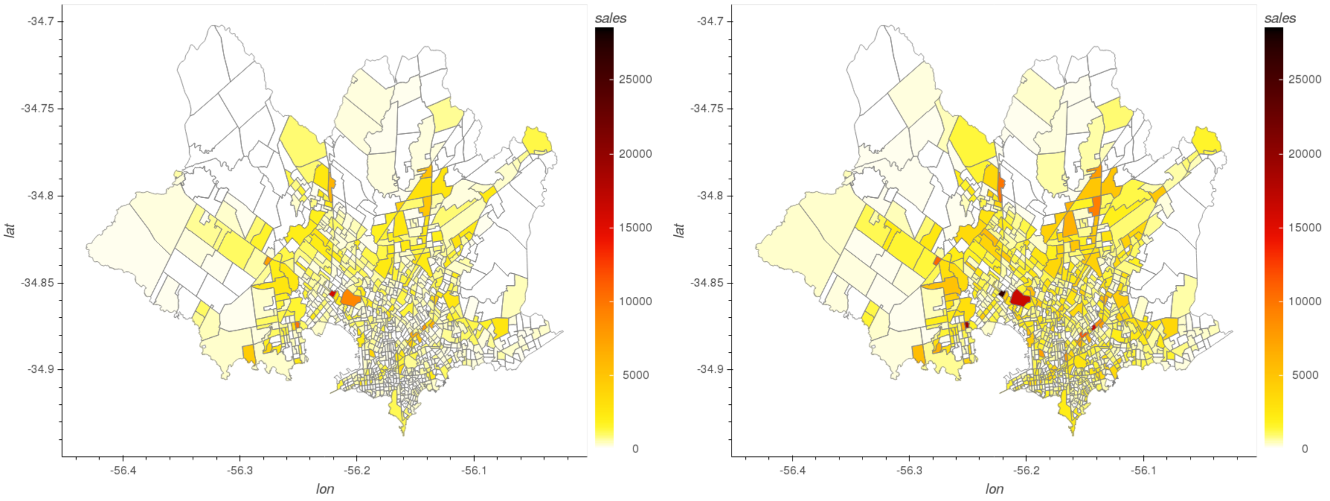
Regarding STM transactions occurring in the morning, Fig. 10a clearly shows that those areas farther away from the city center and the coastline have higher STM transaction activity early in the morning (6.00 a.m.–7.00 a.m.) than those near the coast. Transaction activity in the city center and near the coastline intensifies an hour later, as can be seen in Fig. 10b. 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, Fig. 11a 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, Fig. 11b 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 III-A. Areas with transactions occurring early in the day and later at night are also the more socioeconomically vulnerable [42]. By combining different sources of data it is possible to understand how mobility patterns vary across citizens with different socioeconomic characteristics.

F. Practical use cases

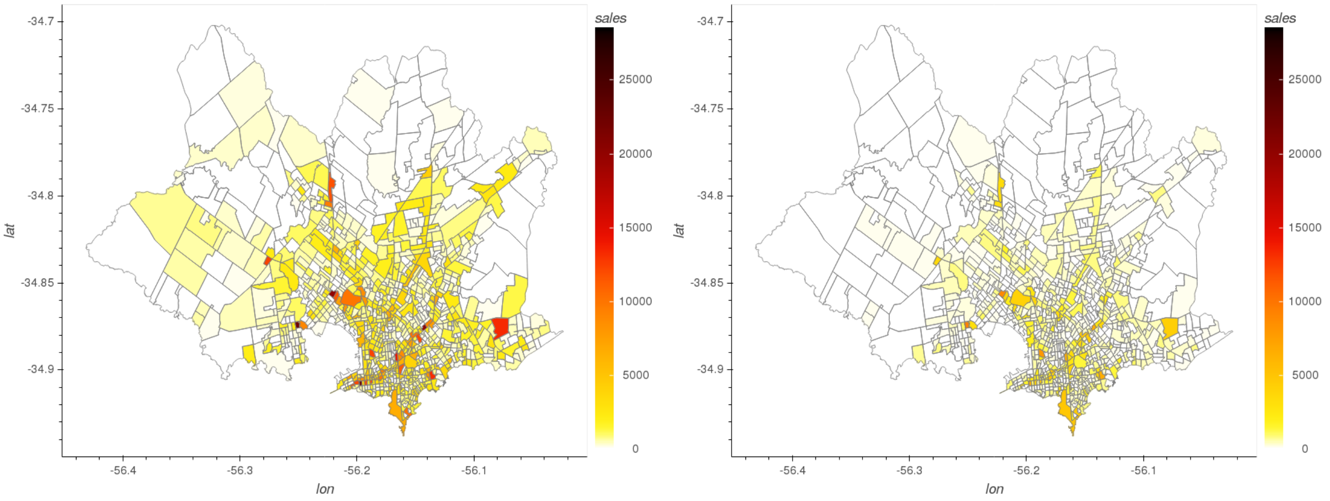
Besides a purely descriptive use, urban data analysis can help authorities of the public transportation system in several ways. This section presents use cases where data analysis can be incorporated to the auditing, control, and policy enforcement workflows of public transportation authorities.



(a) 6.00 a.m.–7.00 a.m.

(b) 7.00 a.m.–8.00 a.m.

Fig. 10: 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.

Fig. 11: Choropleth map of STM transactions in the evening during May 2015

1) 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, Fig. 12 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.

2) 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. Fig. 13 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

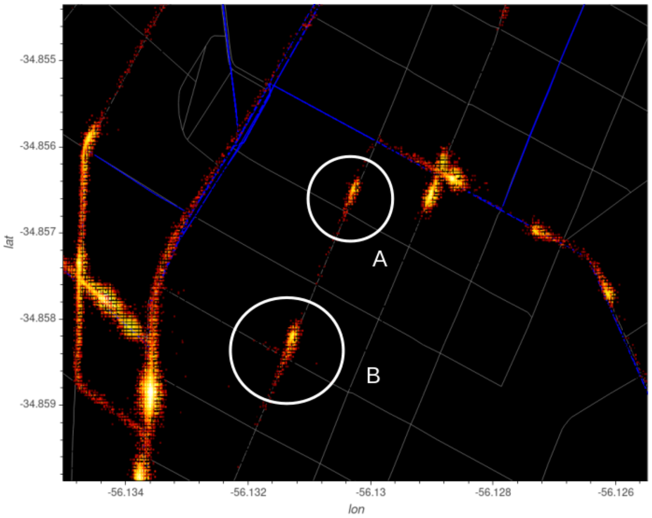


Fig. 12: Anomaly detection: example of detour. The blue lines represent bus routes. A and B are two clusters of transactions which occurred outside of the bus network.

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.

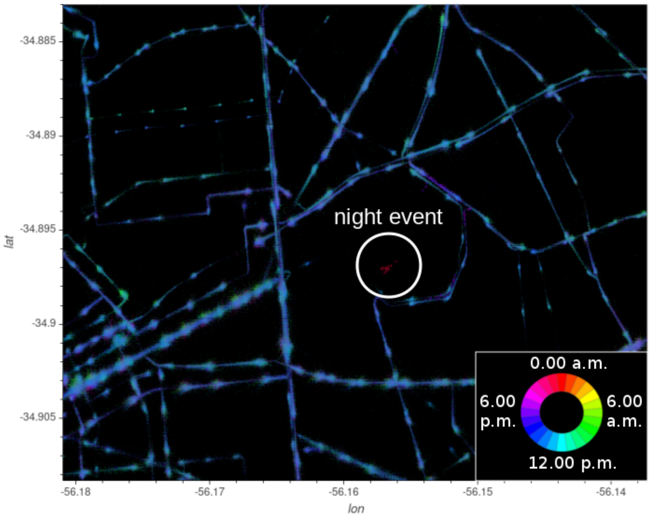


Fig. 13: Anomaly detection: example of event taking place at midnight near an outdoor venue.

3) 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. Fig. 14 shows a heatmap of transactions and bus stops in the surroundings of a roundabout. 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. 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.

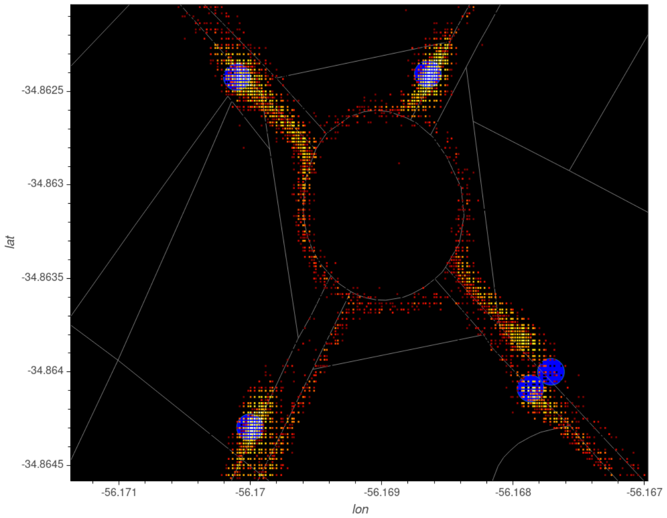


Fig. 14: Spatial distribution of transactions in a roundabout

IV. OD MATRIX ESTIMATION

This section describes the methodology and discusses the results of computing OD matrices using data from the ITS in Montevideo, Uruguay.

A. Destination estimation algorithm

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. [17] and later applied by other researchers, as outlined in Section II.

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. Fig. 15 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.

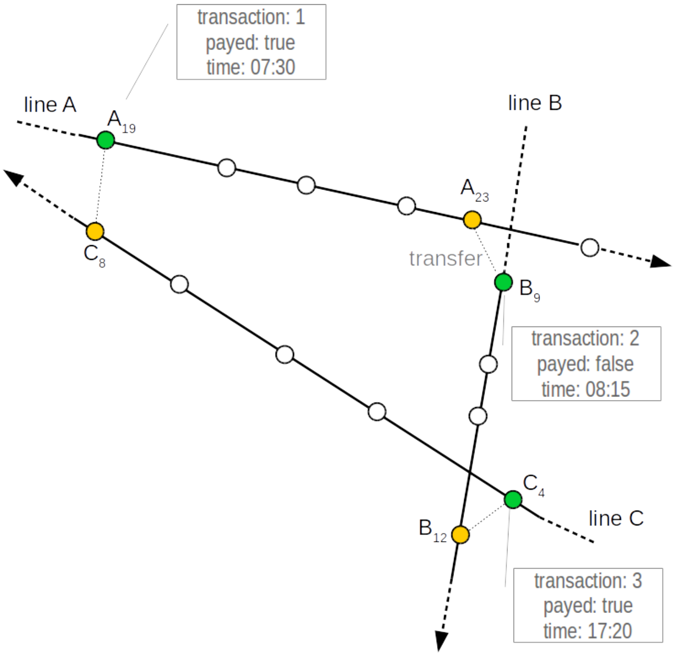


Fig. 15: Example of the trip chaining algorithm

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.

The destination estimation algorithm was implemented using the Python programming language. The 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. [22]. 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 III-E3. 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. [23].

The destination estimation algorithm limits the search of a possible destination bus stop to a configurable radius. 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 works of the related literature, several values were found for this parameter: 800 m [43], 1000 m [19], [21], and 2000 m [18]. 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 [23].

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 bus location 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 proposed using a threshold of 30 minutes for the transfer [21]. 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 bus location 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.

B. Experimental results

After the cleansing process described in Section III-D3, 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% [19], 66% [18], 80% [21]. 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 III-E2. 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 www.fing.edu.uy/~renzom/msc in CSV files with their corresponding metadata. For the sake of visualization, results are discussed at a coarser grain in this document, aggregating the computed OD pairs by municipality. Table II outlines the estimated OD matrix corresponding to the studied dataset of May 2015. Each municipality is represented by its identifying code.

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.

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 II) were used. 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 be assessed in Fig. 16, 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 Fig. 16a whereas those derived from the mobility survey are presented in Fig. 16b.

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 Fig. 16a than in Fig. 16b. 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.

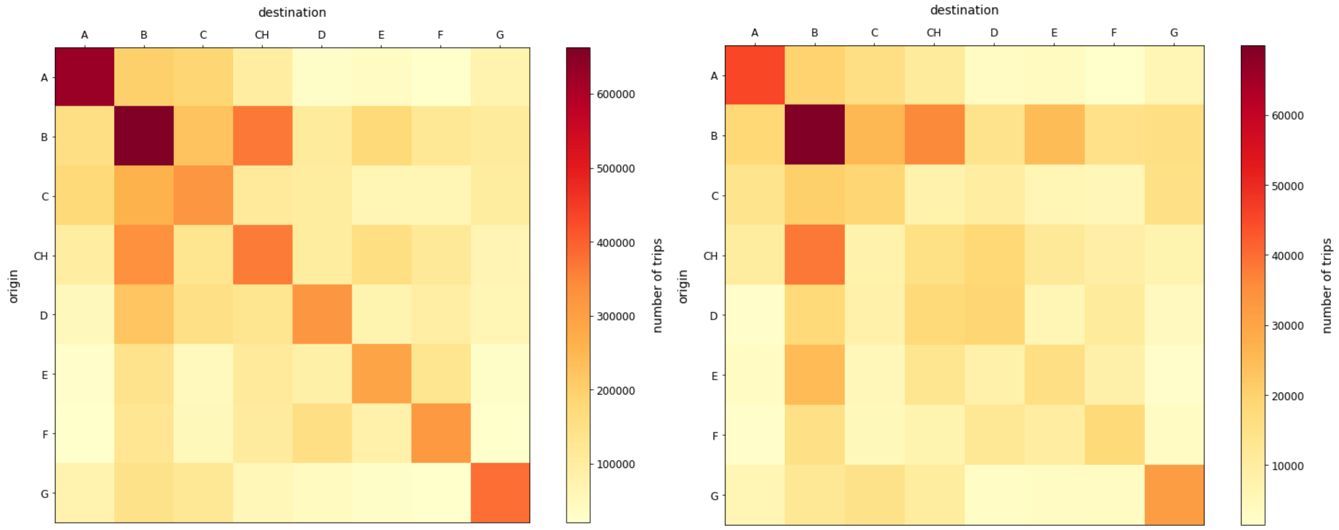
The Pearson correlation coefficient was applied to quantify the similarities between the OD matrix estimated from ITS data and the OD matrix from the mobility survey. To compute this coefficient, matrices were vectorized in row-major order, without losing information, since proximity in the matrix does not imply geographical proximity between municipalities. Results show that the estimated OD matrix and the OD matrix corresponding to the mobility survey have a Pearson correlation coefficient of 0.90. This value indicates a strong correlation between both results, thus validating the proposed approach for OD matrix estimation based on ITS data.

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 research, 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, Fig. 17 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 admin-

TABLE II: 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	



(a) estimation for May 2015 using ITS data

(b) results from the 2016 mobility survey

Fig. 16: Comparison of OD matrices.

istrations 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, 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 conducting them periodically. Firstly, because they serve as a ground-truth for other methodologies, such as the one proposed in this article. 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 [44]. Another example is leg identification in trips involving bus transfers. In the approach proposed in this article, 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 data 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.

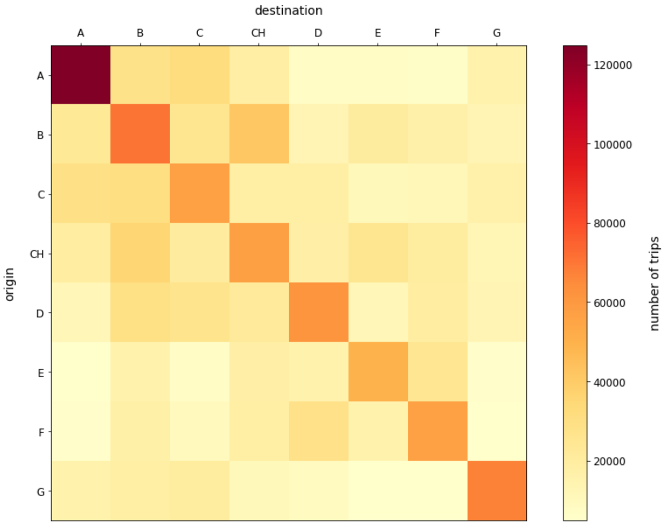


Fig. 17: OD matrix heatmap by municipalities for weekends in May 2015

C. OD matrix visualization tool

As outlined in Section III-D, 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 [31]. 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 web application is freely available at www.fing.edu.uy/~renzom/msc. Fig. 18 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 shown on the right to quantify the information visually displayed. Additionally, the application 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.

V. Conclusions and future work

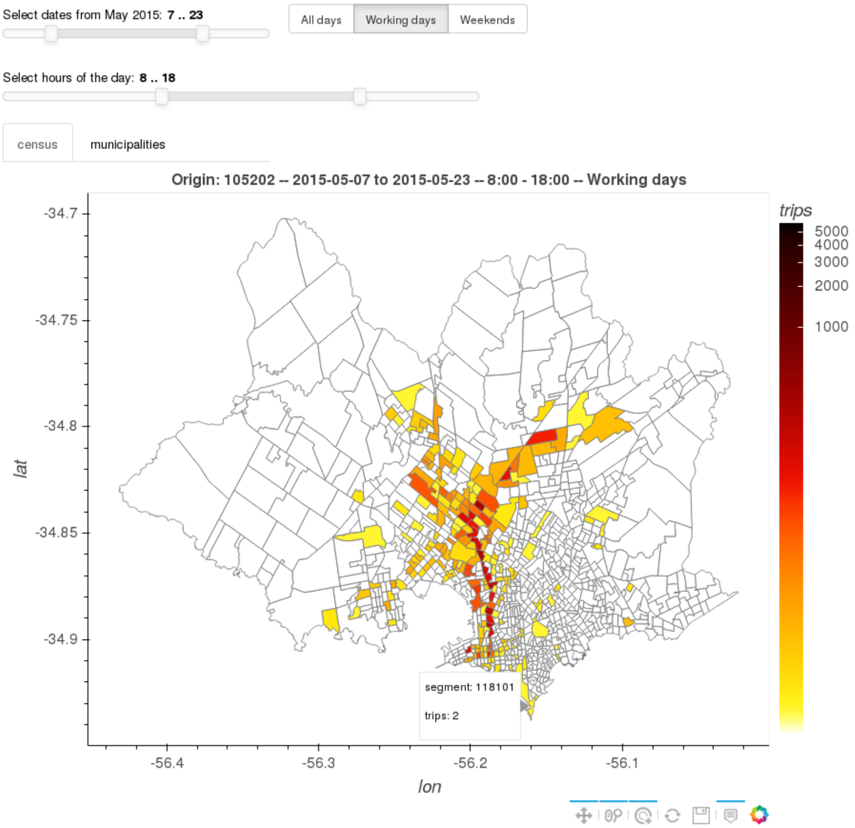
This closing section outlines the main findings and conclusions resulting from the research reported in this article, along with the main lines of future work.

A. Conclusions

This article presented an urban data analysis approach to study mobility using ITS data. As a case study, the ITS in Montevideo, Uruguay, was analyzed by studying a dataset of GPS bus location data and smart card ticket sales data accounting for over 150 GB. 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, and number of bus transfers. A temporal analysis of ticket sales was performed, identifying three peak hours during working days, namely, morning, midday, and afternoon. Then, a spatiotemporal analysis 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 a pure descriptive utilization of ITS data to characterize a public transportation system, a methodology for building OD matrices was proposed and implemented. A trip chaining algorithm was developed, based on previous works in the related literature, for estimating the destination of trips. The algorithm links trips done by the same cardholder, considering reasonable assumptions about origin and destination. The implemented algorithm was able to estimate the destination for 81.62% of trips in the studied dataset, a highly competitive result when compared to the ones 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 OD matrix computed for Montevideo was compared against the one from the 2016 urban mobility survey. Results showed a Pearson correlation coefficient of 0.90 between both matrices, suggesting that the proposed approach is a valid alternative to understand mobility in a city. Both methods


Select dates from May 2015: **7 .. 23**

Select hours of the day: **8 .. 18**

Origin: 105202 -- 2015-05-07 to 2015-05-23 -- 8:00 - 18:00 -- Working days

segment: 118101

trips: 2

Fig. 18: User interface of the OD matrix visualization tool

identified the same largest generator and attractor of trips in the city, as well as a large number of intra-municipality trips. Several advantages are highlighted from the proposed approach. Taking advantage of ITS data allows studying mobility at a finer grain, obtaining OD matrices between pairs of bus stops (4096×4096) and census segments (1063×1063), whereas the OD matrix built using data from the mobility survey only applies to municipalities (8×8). 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 deployed, as in the case of Montevideo. 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. The visualization tool allows selecting a geographical area and displays a heatmap indicating the number of passengers traveling from the selected area to all other in 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 (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.

B. *Future work*

The work reported in this article is a first proposal towards using ITS data to understand mobility in a smart city. Several lines of research remain to be explored in order to extract more and richer information that can be used to improve a public transportation system.

The data analysis mainly focused on understanding the interaction between passengers and the transportation system. However, the available data sources allow studying other very interesting aspects of mobility in the city. For instance, GPS bus location data 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 OD matrices estimation, the proposed approach can be extended by considering data from tickets sold without smart cards, to account for all passengers of the transportation system. The destination estimation algorithm can be further refined by using historical passenger data and machine learning techniques to infer frequent destinations when trip chaining fails. Furthermore, GPS bus location data could be used to discriminate between short individual trips using the same ticket from multi-leg trips involving transfers.

The proposed approach should be applied to recent ITS data when it becomes available. In this regard, this work contributes towards authorities opening up more data. Results of the analysis can be applied to solve optimization problems, e.g., synchronization of bus schedules, demand-based fleet size optimization, bus stops location, and bus line network redesign.

Acknowledgements

The author would like to thank PEDECIBA and ANII (Uruguay) and Fundación Carolina (Spain) for their support.

References

[1] R. Camagni, M. C. Gibelli, and P. Rigamonti, "Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion," *Ecological economics*, vol. 40, no. 2, pp. 199–216, 2002.

[2] O. D. Cardozo and C. E. Rey, "La vulnerabilidad en la movilidad urbana: aportes teóricos y metodológicos," in *Aportes conceptuales y empricos de la vulnerabilidad global*, A. Foschiatti, Ed. Editorial Universitaria de la Universidad Nacional del Nordeste, 2007, pp. 398–423.

[3] S. Grava, *Urban Transportation Systems*. McGraw-Hill Professional Publishing, 2000.

[4] C. Benevolo, R. P. Dameri, and B. D'Auria, "Smart mobility in smart city," in *Empowering Organizations*, T. Torre, A. M. Braccini, and R. Spinelli, Eds. Springer International Publishing, 2016, pp. 13–28.

[5] L. Figueiredo, I. Jesus, J. A. T. Machado, J. R. Ferreira, and J. L. M. de Carvalho, "Towards the development of intelligent transportation systems," in *IEEE Intelligent Transportation Systems*, 2001, pp. 1206–1211.

[6] J. D. D. Ortzar, J. Armoogum, J. Madre, and F. Potier, "Continuous mobility surveys: The state of practice," *Transport Reviews*, vol. 31, no. 3, pp. 293–312, 2011.

[7] X. Zheng, W. Chen, P. Wang, D. Shen, S. Chen, X. Wang, Q. Zhang, and L. Yang, "Big data for social transportation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 3, pp. 620–630, 2016.

[8] M.-P. Pelletier, M. Trpanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[9] M. Bagchi and P. White, "The potential of public transport smart card data," *Transport Policy*, vol. 12, no. 5, pp. 464–474, 2005.

[10] P. Furth, B. Hemily, T. Muller, and J. Strathman, "Using archived AVL-APC data to improve transit performance and management," Transit Cooperative Research program-Transportation Research Board, Tech. Rep. 113, 2006.

[11] D. Lu, "Route level bus transit passenger origin-destination flow estimation using apc data: Numerical and empirical investigations," Ph.D. dissertation, The Ohio State University, 2008.

[12] H. Wang, F. Calabrese, G. D. Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records," in *13th International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 318–323.

[13] J. Doyle, P. Hung, D. Kelly, S. F. McLoone, and R. Farrell, "Utilising mobile phone billing records for travel mode discovery," in *22^{nd} IET Irish Signals and Systems Conference*, 2011.

[14] C. Anda, A. Erath, and P. J. Fourie, "Transport modelling in the age of big data," *International Journal of Urban Sciences*, vol. 21, no. 1, pp. 19–42, 2017.

[15] V. Kostakos, T. Camacho, and C. Mantero, "Wireless detection of end-to-end passenger trips on public transport buses," in *IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 1795 – 1800.

[16] T. Li, D. Sun, P. Jing, and K. Yang, "Smart card data mining of public transport destination: A literature review," *Information*, vol. 9, no. 1, p. 18, 2018.

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

[18] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

[19] W. Wang, J. Attanucci, and N. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, pp. 131–150, 2011.

[20] Transport for London, "London travel demand survey," 2018, 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).

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

[22] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

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

[24] Instituto Nacional de Estadstica, Uruguay, "Resultados del censo de poblacin 2011: poblacin, crecimiento y estructura por sexo y edad," 2012, 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).

[25] Servicio de Geomtica - Intendencia de Montevideo, "Personas por zona 2011," 2014, [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).

[26] J. J. Calvo, "Atlas sociodemogrfico y de la desigualdad del Uruguay," Unidad Multidisciplinaria, Facultad de Ciencias Sociales, Tech. Rep., 2012.

[27] Servicio de Geomtica - Intendencia de Montevideo, "Hogares con NBI por segmento 2011," 2014, [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).

[28] P. Abreu and J. F. Vespa, "Plan de Movilidad," Intendencia de Montevideo, Tech. Rep., 2010.

[29] Servicio de Geomtica - Intendencia de Montevideo, "Lneas de Transporte," 2012, [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).

[30] ——, "Ejes de Calles," 1996, [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).

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

[32] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

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

[34] E. R. Tufte, *The Visual Display of Quantitative Information*. Cheshire, CT, USA: Graphics Press, 1986.

[35] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, and et al., "Jupyter notebooks - a publishing format for reproducible computational workflows," in *20th International Conference on Electronic Publishing*, 2016, pp. 87–90.

[36] S. Nesmachnow, "Computación científica de alto desempeño en la Facultad de Ingeniería, Universidad de la República," *Revista de la Asociación de Ingenieros del Uruguay*, vol. 61, no. 1, pp. 12–15, 2010.

[37] R. D. Peng, “Reproducible research in computational science,” *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.

[38] Presidencia de la Repblica, “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, 2010.

[39] G. R. Jagadeesh, T. Srikanthan, and X. D. Zhang, “A map matching method for GPS based real-time vehicle location,” *The Journal of Navigation*, vol. 57, no. 3, pp. 429–440, 2004.

[40] R. Massobrio and S. Nesmachnow, “Anlisis de datos de movilidad del transporte pblico de montevideo,” in *XIX Congreso Latinoamericano de Transporte Pblico y Urbano*, 2016, pp. 1–11.

[41] Intendencia de Montevideo, “El corazn de montevideo se renueva,” 2017, retrieved from: http://www.montevideo.gub.uy/institucional/noticias/el-corazon-de-montevideo-se-renueva (Last accessed: 2018-08-30).

[42] S. Nesmachnow, S. Baa, and R. Massobrio, “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*, vol. 2, no. 5, pp. 1–18, 2017.

[43] A. A. Alsger, M. Mesbah, L. Ferreira, and H. Safi, “Use of smart card fare data to estimate public transport origin–destination matrix,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2535, pp. 88–96, jan 2015.

[44] N. Nassir, M. Hickman, and Z.-L. Ma, “Activity detection and transfer identification for public transit fare card data,” *Transportation*, vol. 42, no. 4, pp. 683–705, 2015.