TÉCNICO LISBOA

Integrative traffic flow analysis of public transport data in the city of Lisbon

Sofia Maria Pais Cerqueira

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Rui Miguel Carrasqueiro Henriques
Dr. Elisabete Maria Mourinho Arsénio

Examination Committee

Chairperson: Prof. Alberto Manuel Rodrigues da Silva
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Alexandra Sofia Martins de Carvalho

January 2021

Acknowledgments

First of all, I want to thank my advisors, Prof. Rui Henriques and Prof. Elisabete Arsénio, for their tireless support during the master's degree. I am grateful to participate in this fascinating project and even more grateful to have met my mentors. I must say that I never leave an orientation meeting without motivation; they always valued my work and motivated me for more.

Secondly, I want to thank CARRIS representatives who gave their time to participate in numerous meetings to give us their feedback on the work done and on possible new future solutions. Among the representatives of CARRIS, I want to thank Margarida Nunes and Engineer Salvador who have always been attentive and critical in developing my work.

This work was supported by the research fund from the LNEC institution and brought the guidance of Engineer Jose Barateiro, whom I must thank for all the times he intervened to guide me.

I also want to thank the Municipal Council of Lisbon to provide this innovative project that has also opened doors for new interests and research.

Finally, I want to thank my family, boyfriend, and friends who motivate me to achieve my ambitions and pursue my vocation. I was blessed by this work and by the people who accompanied me.

Resumo

As cidades mundiais enfrentam uma crescente massificação da população [1] e, consequentemente, novos desafios, incluindo a necessidade da mobilidade urbana sustentável. Com o aumento da população e da procura de veículos privados, os engarrafamentos de trânsito tornam-se recorrentes, afectando a mobilidade e contribuindo para a poluição atmosférica. Neste contexto, os modos de transporte público são essenciais para satisfazer as necessidades dos passageiros, contribuir para a qualidade de vida dos residentes, e oferecer modos de viagem convenientes e seguros para os não residentes. A oferta adequada depende da compreensão correcta da dinâmica real do tráfego dentro da cidade, que é geralmente desafiada pela necessidade de adquirir dados de viagem individuais, pela compreensão dos padrões de deslocação, e pela falta de vistas multimodais. Este trabalho visa abordar estes desafios propondo uma abordagem para inferir matrizes Origem-Destino (OD) a partir de validações de cartões inteligentes capazes de: i) detectar padrões de deslocações multimodais de viagens individuais, ii) detectar eficazmente vulnerabilidades na rede relativas a distâncias a pé e durações de viagem, e iii) decompor os fluxos de tráfego de acordo com regras calendericas e perfis de utilizadores, e iv) apoiar uma análise descritiva consciente do contexto. Além disso, dado o facto de que os sistemas de recolha automática de tarifas (AFC) podem assumir um controlo só de entrada ou saída, os modelos unimodais e multimodais para a inferência de paragens de saida de autocarros são ainda propostos nesta tese. A cidade de Lisboa é utilizada como caso de estudo, sendo as contribuições acima mencionadas avaliadas através da rede de transporte CARRIS e METRO. Os resultados recolhidos mostram que 70% das paragens de desembarque podem ser estimadas com elevado grau de confiança a partir dos dados do 'smart-card' CARRIS e na presença dos dados do 'smart-card' METRO ocorreu uma melhoria de 10% . As matrizes de OD inferidas permitiram a identificação de vulnerabilidades na rede, oferecendo à CARRIS novos conhecimentos e um meio para compreender a dinâmica multimodal e validar hipóteses de OD's. As contribuições do nosso trabalho foram desenvolvidas no contexto do projecto ILU, em estreita cooperação com o principal operador de autocarros em Lisboa, CARRIS, e a Câmara Municipal de Lisboa (CML).

Palavras-chave: mobilidade sustentável, inferência de paragens de desembarque, multimodalidade, series temporais multivariadas georreferenciadas.

Abstract

The worldwide cities face a growing massification of population [1] and, consequently new challenges arise, including the purse of sustainable urban mobility. With the growing population and increasing private vehicle demand, traffic jams become more prevalent, affecting mobility and creating air pollution. In this context, public transport modes are essential to meet travellers' needs, contribute to residents' quality of life, and offer convenient and safe travel modes for non-residents. The adequate offer is, nevertheless, dependent on the correct understanding of the real traffic dynamics within the city, which is generally challenged by the need to acquire individual trip data, understand commuting patterns, and lack of multimodal views.

This work aims at addressing these challenges by proposing an approach to infer Origin-Destination (OD) matrices from smart-card validations able to: i) detect multimodal commuting patterns from individual trips, ii) efficiently detect vulnerabilities on the network pertaining to walking distances and trip durations, and iii) decompose traffic flows in accordance with calendrical rules and user profiles, and iv) support context-aware descriptive analytics. In addition, and given the fact that automated fare collection (AFC) systems can assume an only-entry-or-exit control, unimodal and multimodal models for alight bus stop inference are further proposed in this thesis.

Lisbon city is used as the study case, with the aforementioned contributions being assessed over the CARRIS and METRO transportation network. The gathered results show that 70% alighting stops can be estimated with high confidence degree from CARRIS smart-card data and with the presence of METRO smart-card data constitutes an improvement of 10%. The inferred OD matrices allowed the identification of vulnerabilities in the network, offering CARRIS new knowledge and a means to understand multimodal dynamics and validate OD assumptions. The contributions of our work were developed in the context of the ILU project, in close cooperation with the primary bus operator in Lisbon, CARRIS, and the Lisbon City Council (CML).

Keywords: sustainable mobility, alighting stop inference, multimodality, origin-destination matrices, public transportation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the increasing population in urban cities and changing society lifestyles, the governances around the world are making an effort to become smart cities to satisfy the needs and improve the citizens' quality life. So, one of the strategic elements to become smart cities, is a sustainable urban mobility system [2], combined with policies to discourage the use of individual transport [3]. Indeed, the investment on intelligent transportation systems technologies can support transportation planning, improve the service given in public transports [2], and consequently increase the attractiveness to the use of collective transport. In this context, the Lisbon City Council (CML) is establishing efforts to collect the available traffic data and provide it to projects that can promote sustainable mobility.

In this context, this research aims to analyze multimodal public transport data in order to study passenger flow behaviour in a regular urban context, and as well to identify events of the situational context affecting the traffic demand. This work is being conducted in the context of the ILU project [4], an innovating and pioneering project that is committed to optimizing the urban mobility in the Lisbon city by combining multiple sources of traffic data. The Lisbon city is, in fact, used as the study case in this work, with traffic flow analysis being performed from raw smartcard validations gathered from the primary bus operator, CARRIS, and subway operator, METRO.

1.1 Challenges and Contributions

The adequate supply of public transportation depends on the correct modeling of the continuously changing traffic dynamics within the city. Understanding these dynamics is generally challenged by the need to acquire and process massive individual trip data, identify commuting patterns, estimate entry and exit smartcard validations whenever absent, and synergistically capture multimodal views.

To address this issues, this work aims contribute for its resolution by proposing an approach to infer dynamic Origin-Destination (OD) matrices from smart-card validations. Nevertheless, to reach this final contribution, it is enumerated the main sub contributions: i) estimate records corresponding to missing stop alighting; ii) detect multimodal commuting patterns from individual trips; iii) efficiently detect vulnerabilities on the network pertaining to walking distances and trip durations, and iv) decompose traffic flows in accordance with calendrical rules and user profiles, and v) support context-aware descriptive analytics.

Before enunciating a strategic plan for developing the models for alighting stop estimation and dynamic OD inference, it is important to identify challenges related to data extraction and preprocessing. Performing data collection and consolidation when some information sources are private, disperse or even unavailable in digital format, and, beyond these features, data sources are not always structured. Secondly, the massive amount of smart-card transactions requires algorithms, pragmatic programming data structures, and the incorporation of scalability principles [5] to efficiently query data and prevent high use of memory.

After addressing these issues, this work moves to answer the following major tasks:

1. consolidation of smart-card transaction from AFC system from bus operator CARRIS and subway operator METRO;

2. alighting bus stop inference model to estimate passengers' disembark points in the bus network as CARRIS relies on a fare collection system that only requires the validation of smartcards upon passenger entries (only-entry system);

3. alighting dual-mode stop inference model that traces the passenger path in across multiple modes of transport. This model is proposed to improve the limitations of the aforementioned bus model;

4. identification of commute travel journeys from trip segments produced by alighting stop inference model;

5. inference of dynamic OD matrices that explain the passenger flow between different network locations (stops, traffic analysis zones, and neighborhoods);

6. inference of dynamic OD matrices that explain the past state of the network, according to metrics such as time, distance, number of transfers spent between an origin and destination, walking distances and waiting times in transfers, time and distance spent in along a whole journey, and percentage of journeys that used subway within;

7. display dynamic ODs in a graphical dashboard, with the possibility of filtering and parameterizing the inference according to the desirable time period, calendrical rules, spatial granularity, passenger card profile, among others;

8. incorporation of the situational context by segmenting the periods for dynamic OD inference.

The contributions were validated with our stakeholders, CARRIS and CML, and have resulted in an accepted scientific manuscript accepted and presented in the European Transport Conference (ECT'2020), one extended abstract accepted in XIV Congreso de Ingeniería del Transporte (CIT'2020), one manuscript submitted in the European Transport Research Review (ETTR) journal, and four institutional presentations.

1.2 Thesis Outline

The work is structured as follows. **Chapter 2** provides a essential background on the concepts covered in the following sections. **Chapter 3** describes previous related works related to the researched area and show the main contributions in the field **Chapter 4** describes the proposed solution to answer the target problem. **Chapter 5** gathers and discusses the results produced from the application of the proposed OD inference methodology over the public transportation network in the Lisbon city. **Chapter 6** instructs the use of the tool for conducting the tasks of stop-alighting and dynamic OD inference. Finally, **Chapter 7** enumerates the major concluding remarks and presents possible future directions.

Chapter 2

Background

The background chapter introduces and formalizes essential concepts explored in the following chapters, including traffic data analysis using time series and OD data, commute analysis, the targeted public transportation network concept, and visualization principles used to represent dynamic passenger flow.

2.1 Traffic data analysis

Firstly, it is important to elucidate the reader, the different sources of data available to develop this work. Some of these sources are under privacy protocols, while others are open source. Three major sources are necessary to our study; i) smart-card transaction collected from AFC system only-control, ii) public traffic service data collected from general transit feed specification (GTFS), iii) other private fonts:

1. **Automated fare collection (AFC)** are used particularly in urban public transport systems. In this context, smart-card validation data are commonly denoted AFC data or individual trip data, which are stored when are validated in the automatic ticket retrieve system, including features such us boarding timestamp, boarding location, destination and timestamp disembark, passenger identifier, fare and route [6]. These features allow network monitoring decision-making support. Also, they are enough abstract to cross different public transport modes, and operators [7], powering future investigations on multimodal mobility. It is necessary to emphasize that it exists two types of AFC systems. The first is called entry-only configuration, that records only the entry ticket transaction [8], which the case of major bus operator where the main focus is to avoid alighting delays provoked by waiting times by the exit ticket controls, and the fare evasion is underestimated [6]. Since this system does not record alighting stops and exit time, it is necessary to spend time developing models to estimate each transaction's destination to provide better monitoring of the passenger flow. The second system configuration record boardings and alighting transactions, which is the case of subway operator where the ticket system control is out of the vehicles, and consequently does not affect its efficiency and provide better fare control.

2. **General Transit Feed Specification (GTFS)** is the most common data format adopted by operators to describe its fixed transit services. The available network data for a given carrier specified us-

ing the GTFS format is here informally termed GTFS data. This relation GTFS scheme, as shown in the Figure 2.1, was developed to broadcast the visualisation of possible routes and timetables information in the Google Maps [9], and consequently, users have efficient access to information regarding possible paths in public transport to travel to their destination. Since GTFS data are open source, with required skills, it is possible to combine different sources to power up new applications such as trip planning, multimodal visualisation, and document alighting estimation. In the case of CARRIS, GTFS data is updated every month.

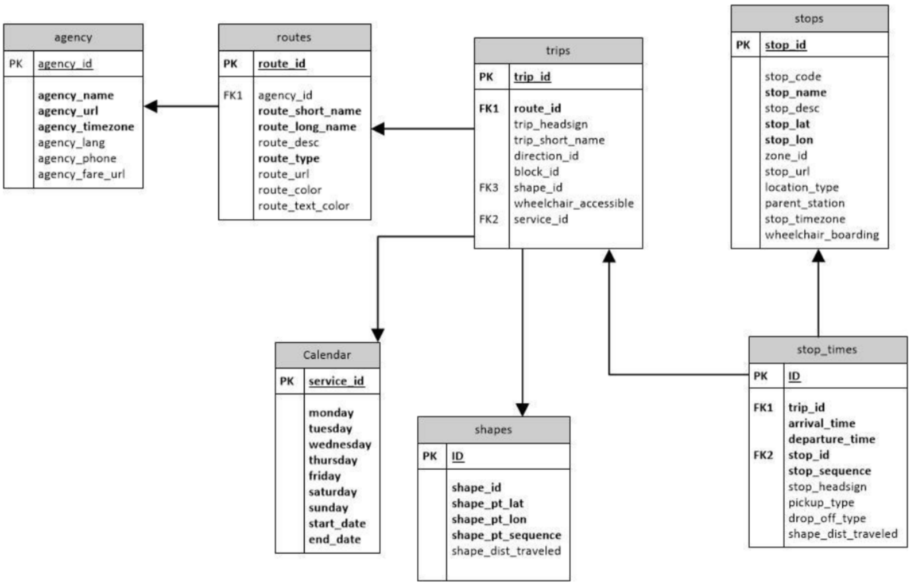


Figure 2.1: Illustrative General Transit Feed Specification schema [10].

3. **Other fonts** are required for this research because the consolidation between the data stored in AFC and GTFS has its limitations. In the chapter of Proposed Solution, it will be explained this new source given by CARRIS operator. In order to describe disruptive indicators, it was used the available sources of situational context consolidated by the City Council, including:

- *public events*, including: a) conventions, b) festivals, c) concerts, and d) sport events.
- relevant *occurrences* in the city, including: a) road accidents, b) medical emergencies, c) fires and floods, d) logistical help and falling structures, e) transport requests, f) conservation and complaints, and g) rescue and civil protection;
- ongoing and planned construction road works (*traffic conditioning events*) characterized by a set of trajectories with (possibly non-convex) interval of obstruction and accompaning details (including the number of affected ways and whether interruption is spasmodic);

2.2 Traffic Time series

As introduced, traffic records in public transportation systems are typically associated with smart-card validated at vehicles and/or stations, generally corresponding to events with temporal, spatial, modal and card-identity annotations. If we chronologically aggregate these events, delimited in a period of time, we are in the presence of a time series. A time series is thus a representation between a dependent variable (for example, daily passenger flows at a given location, vehicle or route) over an independent variable, time.

Time series can be *univariate*, $\mathbf{x} = (x_1, .., x_T)$, where $x_t \in \mathbb{R}$, or *multivariate*

$$\mathbf{x} = \begin{bmatrix} x_{11} & \cdots & x_{1T} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mT} \end{bmatrix}, \tag{2.1}$$

where $m > 1$ is the multivariate order (number of variables).

Variant formulations are available, including the possibly to accommodate non-equally spaced observations. However, discrete-time interval is generally assumed to be evenly spaced since it allows for a better understanding in terms of mathematics and functional programming.

Generally, time series are often explained by the decomposition of four components, including the following, which are also illustrated in the Figure 2.2:

- Trend: This feature describes a visible change upward or downward movement over a long period of time in the data, which might not be linear;
- Seasonality: describes the regular pattern in a fixed time period based, easily visible to the human eye (e.g. every month);
- Cycle: periodic variation (e.g. time interval such as every 4-5 years), and the average length of a cycle is longer than the seasonal pattern length;
- Residuals: a random variance left when other components are removed, as it does not present a recurring and persistent pattern [11].

Another subject that is important to understand it is stationarity and autocorrelation in a time series. Stationarity is present in the data when mean, and variance deviation remains constant over time. If the time series has trend or seasonal components, it is non-stationary. Autocorrelation is important to capture the correlation of time series data with its own lagged values.

Now that we have specified essential characteristics of a time series, it will be possible to characterize the times series presented in this research. First, we must understand the meaning of observations in the context of traffic time series.

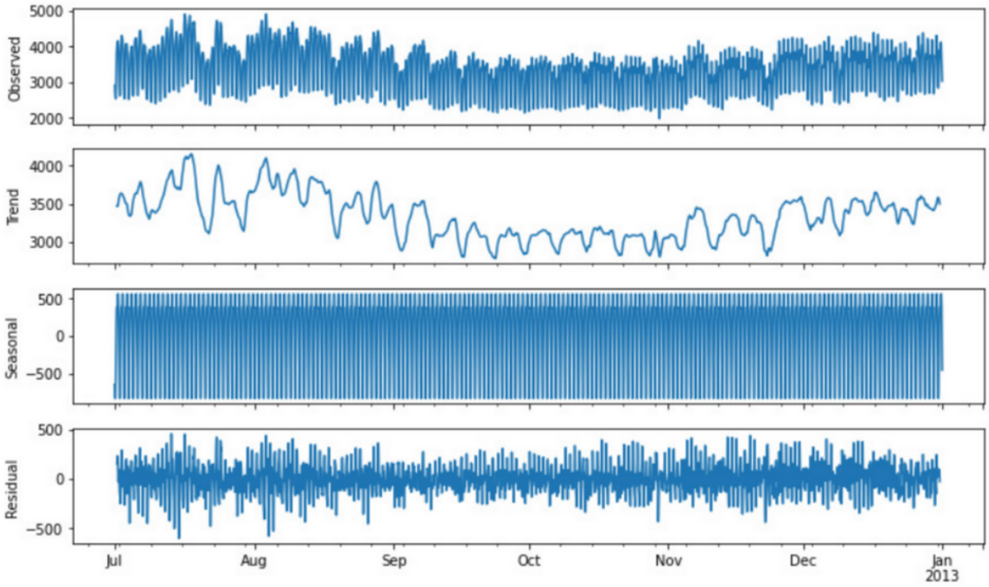


Figure 2.2: Representation of decomposed time series [11].

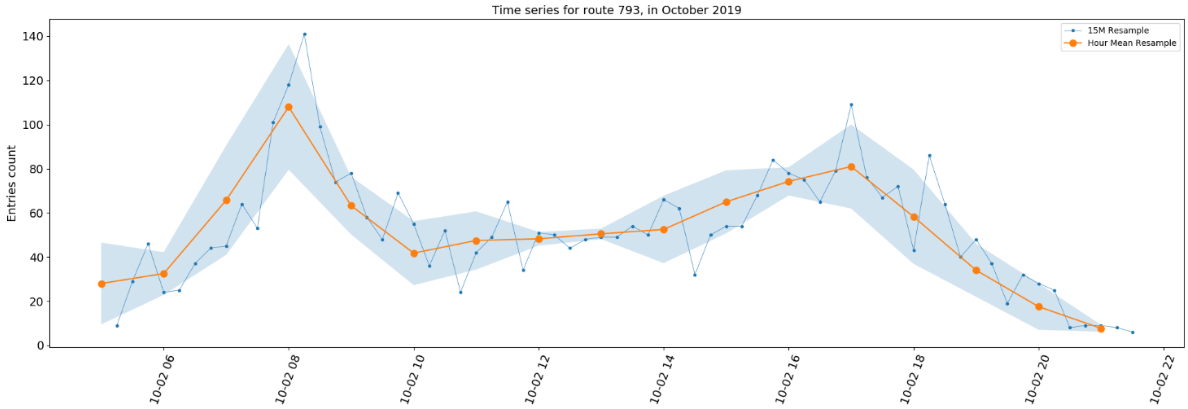


Figure 2.3: Real time series, representing a route 793, in 2 of October 2019.

The system CARRIS records a transaction, each time the passenger validates its card title. Therefore, in the AFC system, several transactions are stored, recorded over time, and also contain the location where they were transacted. Applying an aggregation to these transactions for each fixed time interval, a discrete-time series is obtained, as the Figure 2.3 shows. Formally, it is called a univariate, georeferenced and discrete-time series, because it is a set of discrete observations $\mathbf{x} = (x_1, .., x_t)$ varying in the time, that reference events that took place in a place (for example, Lisbon city, a route, or a bus stop). However, it is possible to dissociate a univariate time series into a multivariate time series, for example, a time series representing the passengers counting over each 1 hour in a route R_i can be transformed in multiple time series where the localization is each bus station in the Lisbon city.

2.3 Traffic Origin Destination Matrices

In the last recent year, researches in the field transport systems and smart cities have pay attention to the analysis of network mobility, especially regarding the public bus transport [12]. So, most of this research is using origin-destination (OD) matrices as a way to visualise the passenger flow over the network, and consequently, evaluating the transportation system [13]. Origin-Destination matrices can be used service planning, in operations analysis, before-and-after impact analysis [14]. Times series may show patterns over time, however, OD's matrices can show patterns over the space.

Classically, a destination matrix (OD) is a table containing quantified information that reveals passengers' traffic flows. There is a row for each origin locations, for each destination exits a column, and the cell of the matrix shows the demand between the two locations (OD pair).

Since most of the systems are entry-only control, it is necessary to develop a methodology to enrich raw data to complete Origin-Destination (O-D) matrices. For that in the related work section and Solution proposed section, it will be presented methodologies to estimate the missing exit stop and the timestamp for each transaction from the AFC system data. After applying the estimation methodology, it will be possible to construct O-D matrices at any level of geographic aggregation.

Route C, Time W to Z		Destination			
		Bus stop 1	Bus stop 2	Bus stop 3	Total boarding
Origin	Bus stop 1	2	7	4	13
	Bus stop 2		3	2	5
	Bus stop 3			5	5
	Total alighting	2	10	11	23

Table 2.1: Illustrative OD matrix for a route C.

Looking at the Table 2.1, it shows that the last row is the sum of column values, which corresponds to the total number of passengers alighting at the bus stop. Moreover, the last column is the sum of row values, which corresponds to the total number of passengers boarding at the bus stop. The sum of total boarding passengers and total alighting passengers must be the same because it is the number of passengers circulating on route C.

The example described in the table 2.1, the representation of people flow is restricted to one route, however, in the same matrix, it is possible to represent multiple routes (for example, representing passenger volume between all stops present in the mobility system). On the other hand, it is also possible to change the granularity of the OD matrix. Instead of using pairs of stops, it can be used passengers' traffic flows between pairs of micro or macro-areas, for example, spatial granularity in different metropolitan subareas. This type of granularity allows to discover passenger's travel trends, and consequently, it helps to define the transport service better to be provided.

Most existing techniques employed in the origin-destination matrix, and other mobility visualizations, focus on showing the demand between two locations, ignoring other factors such as riding time, transfer time, waiting time [15] however, in this urban mobility research this representation has been rethought.

In cooperation with the operator CARRIS mobility department, we can verify that there are other study needs besides counting passengers on the network. They expressed their interest in other metrics that could express their users' mobility, such as the average time spent travelling between two locations, the number of transfers between two locations on the network, among others that will be spelt out in the proposed solution section.

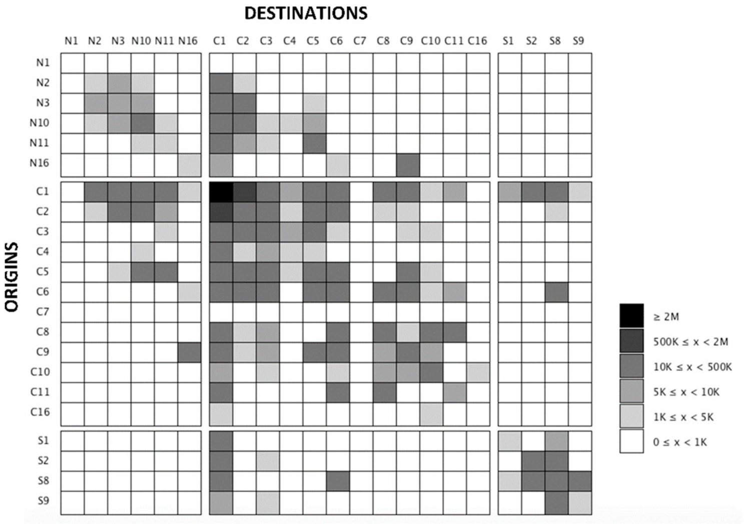


Figure 2.4: Figure of the OD matrix, from study in the Porto bus network [16].

Visualization is a crucial factor in the descriptive analysis, and therefore it was necessary to invest in a visualization that represented a large volume of data and at the same time describes patterns visible with the naked eye. Furthermore, for this reason, the heatmap has the perfect conditions to overcome these limitations, as shown in Figure 2.4. Instead of containing a numerical quantity, each cell in the matrix will be represented by a colour. The tones between the cells will allow a better perception in which pairs (OD) there is a greater demand, or in the case of other metrics where there is, for example, more transfers.

Alternative flow representations

Sankey diagram was created by I. Sankey [17], who used this diagram to publish a steam engine's energy efficiency in 1898. Like OD's matrices, it represents visual transfer between processes (nodes), and the link width is proportional to the flow quantity. The biggest advantage, concerning the previous visualization, is restricted to the fact that it can present intermediate nodes, and also, it has a greater capacity to represent flows. On the other hand, when there are many links and nodes, the visualization becomes impractical. Figure 2.5 corresponds to an illustrative example of a Sankey diagram, which contains four connections, two input nodes and three output nodes.

This type of diagram will be used to support descriptive analysis. The input and output nodes corre-

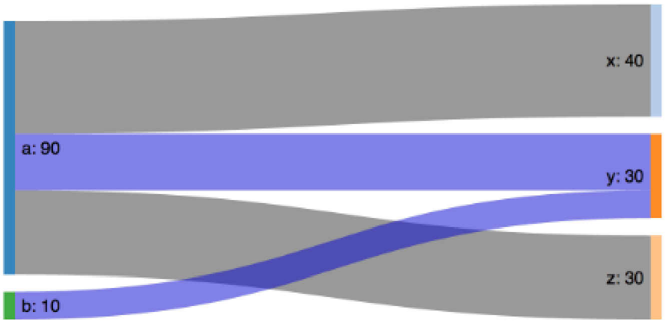


Figure 2.5: Illustrative example to represent a Sankey diagram [18].

spond to the loading and unloading locations, and the link will correspond to the number of passengers flowing between these two points in the bus network.

Chapter 3

Related Work

3.1 Alighting Stop Inference

Whenever the smart card is used in the public transport vehicle, an electronic record is generated and registered in the AFC system. To count and accelerate departures, the public transport operator CARRIS is concerned only with registering entries in the boarding area. For future work on passenger flow analysis, the exit count cannot be extracted from the available data and therefore has to be inferred. The literature on this topic counts with several implementations to different transport systems worldwide, which differ mainly by the set of assumptions implemented. To overcome this problem, the literature suggests several implementations to different transports systems worldwide, which may differ by the set of assumptions that the authors use. Therefore, it is presented a list of some important assumptions to this research:

1. *Passengers will start their next trip at or near the stop alighting location of their previous trip.*

2. *Passengers end the last trip of the day at the stop where they began their first trip of the day.*

3. *Passengers do not walk more than a certain threshold to transfer.*

4. *Using the second assumption, the alighting stop of the last segment trip is estimated considering the boarding stop of the first segment trip of the day if the route taken in the last segment trip is related with the previous first segment taken in the day. Otherwise, it is assigned the first stop boarding the next day.*

5. *If an alighting stop cannot be estimated, it is analyzed for certain period similar transactions to assign a successful alighting stop, for example, days when there is only one travel segment registration.*

6. *The time of candidate alighting stop must occur before the next registered boarding stop in the smart card.*

7. *If the maximum transfer distance between segments is exceeded, it means that the passenger has carried out an intermediate travel segment, in a different transport mode, and in this case, the alighting stop is not estimated.*

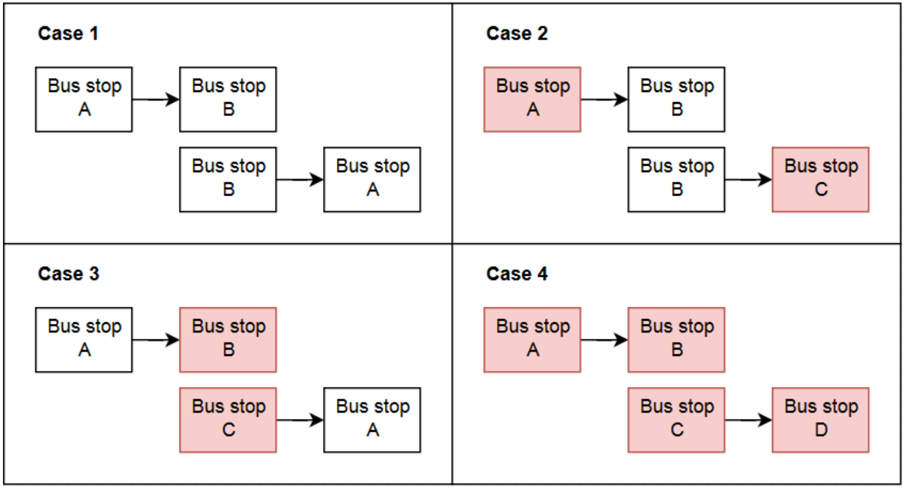
Table 3.1 summarizes some of the studies found that used the above mentioned assumptions. The table contains four columns, the first column indicates the authors of the study, the second shows the mode of transport studied, the third specifies the location of the case study, and the fourth column enumerates the assumptions used to develop the algorithm for alighting stop inference.

Literature	Mode	Location	Assumptions used
Barry, et al. (2002) [19]	Subway	New York City	1 and 2
Barry, et al. (2009) [8]	Subway, bus, ferry	New York City	1 and 2
Nunes, et al. [7]	Bus	Porto	1, 2 and 7
Li, et al. (2011) [20]	Bus	Jinan, China	1 and 2
Zhao, et al. (2007) [21]	Rail System	Chicago	1,2 and 3
Munizaga, et al. (2012) [22]	Bus	Santiago, Chile	1, 2 and 3
Trépanier, et al. (2007) [23]	Bus	Canada	1, 3, 4 and 5
Farzin, et al. (2008) [24]	Bus	S. Paulo, Brasil	1, 2 and 3
Nassir, et al. (2011) [25]	Bus	S. Minneapolis-Saint Paul (USA)	1, 2, 3 and 6
Wang, et al. (2011) [26]	Bus	London, UK	1 and 2
Gordon, et al. (2013) [27]	Bus	London, UK	1, 2 and 3

Table 3.1: Previous studies that use the above aforementioned rules.

To visualize the alighting stop problem, Barry, et al. [19] analyzes some possible travel cases in his study, where the destinations of the trip segments may be correct, or incorrectly inferred, putting into practice the assumptions cited above by him. Thus, it is represented the possible travel cases in Figure 3.1, for a trip with two only segments, which may take place within 24 hours.

Figure 3.1: Study trip cases with two segments



In short, in case 1, trips are correctly inferred, while in case 3 and case 4, stops may not be correctly inferred because they do not respect first assumption (the destination of the previous trip corresponds to the nearest stop). This may occur when the passenger chooses to walk or use another transport mode, such as the subway. For cases 2 and 4, since the first stop boarding does not match the last stop alighting, the second assumption will not allow the correct inferences of the final destination. Since these cases correspond to a time window of one day, it's neglecting the cases, in which a passenger begins the journey on a certain day and ends on the next day.

To implement an algorithm that obeys these restrictions, Nunes, et al. [7], suggest a methodology that estimates the exits, by connecting the trip segments for each passenger, for one day. Briefly explaining the algorithm proposed by Nunes, et al. [7]: firstly, the transaction records are ordered by their smart card identifier and chronologically. Then, for each passenger card, the carried transactions, along a day, are analysed. For the transaction in analysis, possible stop candidates who are upstream of that boarding stop are collected. Moreover, for each of these stops, choose the closest to the departure point of the next segment trip (the distance between the estimated landing stop and the boarding stop in the next segment is called the walking distance or transfer distance). For the last transaction carried out by the passenger, the stop near the departure point of the first transaction of the day is chosen. If the user made only one trip segment, then the algorithm cannot infer the exit stop.

In the same article, it is defined $d(s_x, s_y)$, to formalize the distance measure between bus stops s_x and s_y. Accordingly, the following formula $\hat{s}^a_{pjk}$ represent the first, and second assumption if $j < m_{pk}$ or $j = m_{pk}$ respectively, in order to infer the destinations for each transaction. Formalizing the first assumption $\hat{s}^a_{pjk}$ if $j < m_{pk}$, passengers will start their next trip at or near the stop alighting location of their previous trip:

$$\hat{s}^a_{pjk} = argmin\{d(s^b_{p(j+1)k}, s^a_{pjk}), s^a_{pjk} \in A^R_{pjk}\}, \quad j < m_{pk}, \tag{3.1}$$

and formalizing the second assumption $\hat{s}^a_{pjk}$ if $j < j = m_{pk}$, passengers end the last trip of the day at the stop where they began their first trip of the day:

$$\hat{s}^a_{pjk} = argmin\{d(s^b_{p1k}, s^a_{pjk}), s^a_{pjk} \in A^R_{pjk}\}, \quad j = m_{pk}, \tag{3.2}$$

where $\hat{s}^a_{pjk}$ is the estimated alighting bus stop of the jth trip segment of passenger p on day k; s^a_{pjk} is the alighting route stop candidate of the jth trip segment of passenger p on day k; s^b_{pjk} is the boarding route stop of the jth journey trip segment p on day k; m_{pk} is the number of daily trip segments of passenger p on day k; A^R_{pjk} is the set of candidate alighting stops along route of R, the jth trip segment of passenger p on day k.

Given a specific day, key assumptions are applied for each passenger and boarding transaction. Unless there is more than a single daily segment trip $(mpk = 1)$, the algorithm cannot infer destination. For each candidate alighting stops, along with a route R, is evaluated if the distance to the next boarding stop is within the limit, $d(s^b_{p1k}, s^a_{pjk}) \geq c$, where c is maximum walking distance between stops.

This distance travelled by foot between stops is calculated using the euclidean distance .According to the study by Nunes only considers as possible alighting stops candidates, those below the threshold c of 2000 meters. Hora, et al. [16] uses the same methodology, but was able to demonstrate that Manhattan distance, $D_{i,j}$, is a more realistic measure to represent walking distances, compared to Euclidian distance. In this case, the threshold used was 3000 meters.

The Manhatham Distance D_{ij} is defined by the formula:

$$D_{i,j} = \begin{cases} 2 * R_e * atan2(\sqrt{a(i,j)}, \sqrt{1-a(i,j)}), & \text{if i } \neq \text{ j,} \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

where R_e is the earth radius in meters; the function $atan2$ returns the arctangent of the quotient of its parameters; the part $a(i,j)$ is a function calculated by $hav(\delta\phi) + cos(\phi_i) * cos(\phi_j) * hav(\delta\lambda)$; The ϕ and λ are the latitude and longitude coordinates for each stop, previously converted to radians; the $hav(x) = sin^2(\frac{x}{2})$ is the haversine function.

There are several proposals and assumptions in the scientific community [22] [19] [28] [21] [29] to make inferences of exit stops for an only-control system. For example, Barry, et al. [19] compared the results with real station exit counting, which is extremely difficult in an entry only system control. Zhao, et al. [21] and Wang, et al. [26] compared the results with data from surveys. Alsger, et al. [29] conducted a sensitivity analysis, using the different assumptions, for example, the author validates the trip duration regarding the number of transfers, as shows the Figure 3.2. The research found with most complete validation work was performed by Munizaga, and Palma (2014) [22].

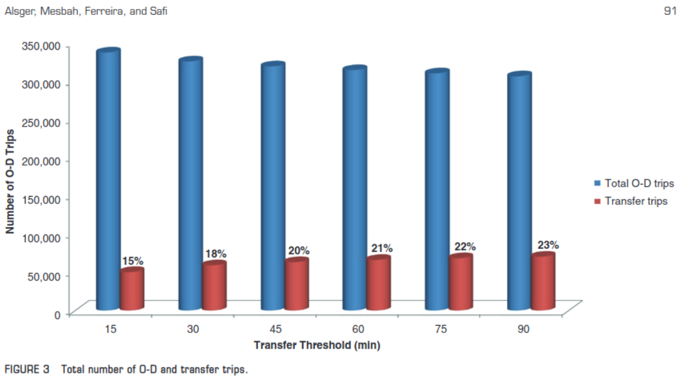


Figure 3.2: Figure from Alsger et al. were validated the transfer walking distance [29].

Munizaga, and Palma (2012) [22] proposes a methodology for alighting stop estimation in the public transport system, where it was estimated 80 per cent of the boarding transactions, and that percentage it was used to build origin-destination matrices. Later, Munizaga, and Palma (2014) [30] follows the analysis methodology of Devillaine, et al. (2013) [28] in order to validate the assumptions made in the last article (2012). The author performs an endogenous validation, which means analysing the data to verify each assumption accepted and detect anomalous behaviour, to propose new rules. These new rules were tested with an exogenous validation, with 53 recruited students volunteers.The records of its boarding transactions (made in a past week) were given to the students, and then they were asked to validate the results of the model performed over the student transactions—this validation showed that the model was able to estimate correctly 79 per cent of the cases.

3.2 Origin Destination Matrices

The background section introduced an OD matrix structure and each cell's content, including the demand between two locations. A matrix can render different formulations of trips, trip segment or journeys. Giving an example to illustrate the typology of the trips: if a person decides to travel from A (home) to B (son's school, and spent less than 5 minutes) and then goes to C (workplace), the matrice can describe a journey where the origin is A and the destination is C (one trip), or the matrice can describe two segments where the first segment has the origin A to destination B and also the segment where the origin is B to a destination C (two trips). In fact, in the literature, many emphasize this distinction, however there is no fixed denomination to distinguish the types of travel.

Cui, el al. [14] refer that bus passenger trips can be defined into two concepts **linked trip and unlinked trip**. The concept unlinked trips characterizes trips where the passenger uses only a bus between boarding and alighting. Moreover, linked trip means one or more unlinked trips compose that, and the origin is the first the boarding on the first unlinked trip and destination is the alighting stop from the last unlinked segment trip.

Mamei, Marco, et al. [31] estimate of individual trips from mobile phone positioning data call detail records (CDR), and it summarizes in the literature review of OD matrices data extraction, in two ways: (1) in time-based matrices (tOD) and (2) Routine-based matrices (rOD, or OD by purpose).

1. **Time-based matrices (tOD)** estimates the user's movements, observed within a given time window, without merging any segments into one. The observation of all movement segments can be advantageous and a disadvantage because it depends on the purpose of the research. These OD's can be advantageous when we are not focused on observing trip routines, but with the peculiarities of a given day.

2. **Routine-based matrices (rOD, or OD by purpose)** which means the analysis of commute trips, for example, routine trips, home-work commute, home-school commute derived from a trip generation model (like the generation model proposed in section Solution. Proposed). Segments of observed CDRs are merged to obtain the parts of a commute trip (for example, home-to-work and work-to-home).

In the area of public transport planning, it is logical the analysis of routine-based matrices, because the operators are interested in the origin and destination of its users to adjust its services. The disadvantage of the approach rOD happens when we observe only one mode service. Some routine trips can not be identified if the algorithm only visualizes the passenger path in a single transport, since passengers use more than one transport mode, besides bus (for example, subway, train, boat, taxi).

Commute trip generation

Commuting travel is relevant to the formation of rOD matrices, and many studies show how to summarize commuting travel from unlinked segments. Ali, et al. [32] develop a methodology based in some assumptions to extract commute trips from smart card data collected from a public transport. Basically, the leg segments and transfers are identified and unnecessary data is trimmed off. Figure 3.3) shows six segments summarized in only three journeys, one between home to work (as we can see in Figure 3.3); work to shop after 17 pm; and then shop to home.

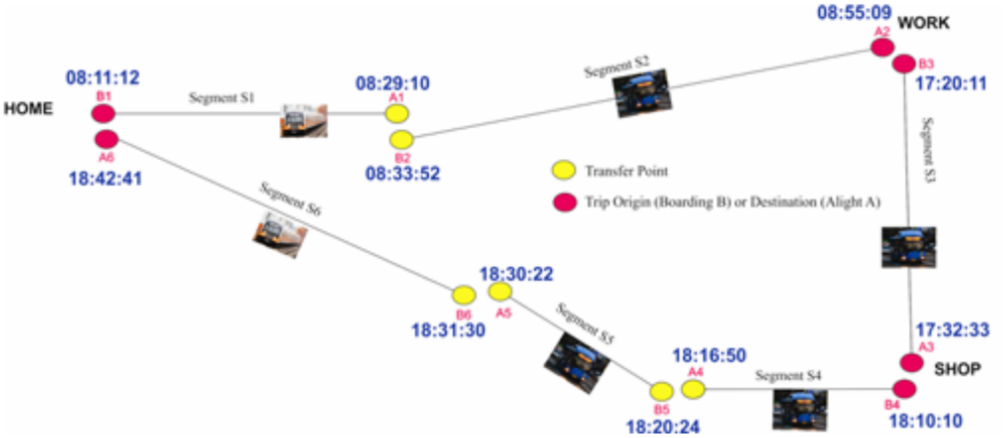
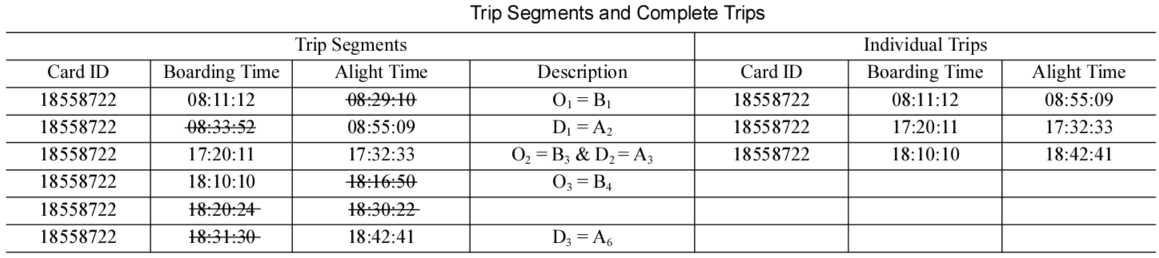


Figure 3.3: Illustrative example of passenger trips during a day [32].



Trip Segments and Complete Trips

Trip Segments				Individual Trips		
Card ID	Boarding Time	Alight Time	Description	Card ID	Boarding Time	Alight Time
18558722	08:11:12	~~08:29:10~~	$O_1 = B_1$	18558722	08:11:12	08:55:09
18558722	~~08:33:52~~	08:55:09	$D_1 = A_2$	18558722	17:20:11	17:32:33
18558722	17:20:11	17:32:33	$O_2 = B_3$ & $D_2 = A_3$	18558722	18:10:10	18:42:41
18558722	18:10:10	~~18:16:50~~	$O_3 = B_4$			
18558722	~~18:20:24~~	~~18:30:22~~				
18558722	~~18:31:30~~	18:42:41	$D_3 = A_6$			

Figure 3.4: Identification of commute journey's through extraction of trip records registered during a day made by a passenger [32].

Ali, et al. [32] assume that time spent in the activity, like work time, requires at least 30 min. So if two consecutive segments have more than 30 min, it is considered an activity; otherwise, it is a transfer point. So, it can be concluded that positive transfers that take more than 30 min can be considered an activity. The Figure 3.4 shows an example of commute trips identification with this parameterization - transfer time less than 30 minutes. However, this assumption cannot be applied in all type of public transport operators, because some waiting time for route services (especially bus) can be more than 30 min (is the case of some days and stop location where passenger density is low and the waiting time for a bus can last at least 1 hour). Nassir, et al. [25] considers that activity should last at least 30 min, And, the maximum waiting time for a person to transfer cannot exceed 90 min.

Chapter 4

Proposed solution

This chapter details the ILU project's contributions to exploring the data collected from the bus operator in Lisbon (CARRIS). In short, the following steps, also illustrated in the Figure 4.1, specify the process of developing the final functionality outlined in Chapter 7 (Visualization Tool):

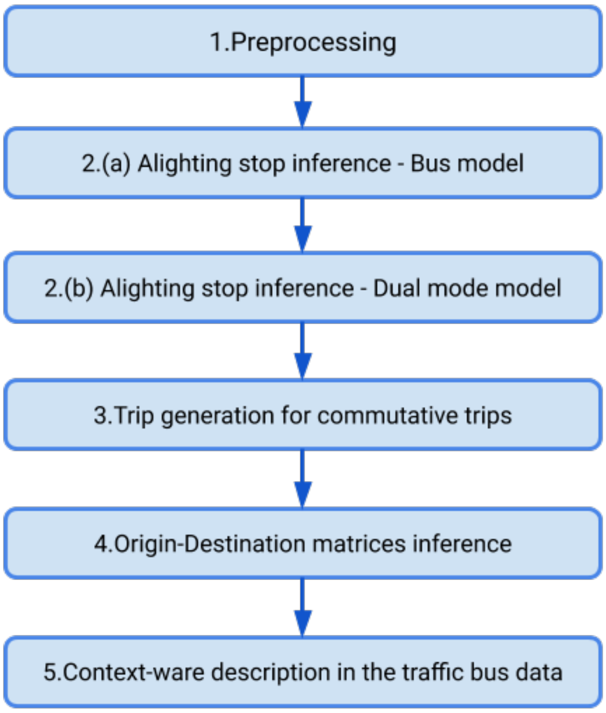

1.Preprocessing
2.(a) Alighting stop inference - Bus model
2.(b) Alighting stop inference - Dual mode model
3.Trip generation for commutative trips
4.Origin-Destination matrices inference
5.Context-ware description in the traffic bus data

Figure 4.1: Proposed methodology.

1. Pre-processing phase;

2. Alighting bus stop inference;

   (a) Alighting bus stop inference in the bus model for transactions validated by passengers only on the bus network, only using data from the CARRIS operator.

(b) Alighting bus stop inference in the dual-mode model for validations carried out by passengers on the bus network, using data from the CARRIS operator and data from the METRO operator.

3. Derive commuting trips from the output data set from alighting bus stop inference model mentioned previously;

4. Development of matrices that explain the demand on the bus network and other metrics of interest that arose from meetings with CARRIS officials;

5. Descriptive analysis of the resulting data with context incorporation.

During the preparation of these solutions, meetings were held with mobility officers at the company CARRIS to understand which features, in addition to the initial ones, would be interesting to develop. The phases described above will be explained in the following sub-chapters, in detail.

4.1 Problem Case Description

4.1.1 Raw data Description

The dataset under analysis comes from the Lisbon bus transport (CARRIS) and auxiliary data for post-result improvement comes from the Lisbon subway operator (METRO). Carris is the leader operator in the area of buses and in 2018 and 2019 it served a total of 125684 and 139496 passengers, respectively. Within the scope of this study, two datasets (from CARRIS) were made available, that correspond to different temporal windows, one is from October 2018 and the other is from October 2019. However, there will be a greater focus on the data from 2019, due to the fact of having been able to consolidate with the data for October 2019, from the METRO operator.

Dataset structure from CARRIS

The datasets extracted from AFC system CARRIS from October 2018 and 2019 are composed by transactions and its structure is shown in the Table 4.1.

Dataset structure from METRO

The datasets extracted from AFC system CARRIS from October 2019 are composed by boarding and alighting transactions. It was necessary to preprocess these dataset in order to merge boarding and alighting transaction. The result structure is shown in the Table 4.2.

Card ID	This identifier number is related with card and not to a individual person, it is used to cross records.
Datetime boarding	The date and time registered when the passenger let read the card, aboard of the bus vehicle.
Route code	Identifier of the route on which the bus is operating.
Direction	Direction of the route (ascending, descending, circular).
Variant	Some routes with a certain direction have variations due to the fact that it does not cover all stops on the original route.
Stop code boarding	Code that identifies the boarding bus stop, which is obtained with help of GPS device that locates the position of the vehicle.
Stop Name	Name of the boarding stop.
Stop sequence boarding	Sequence number of a stop on a given route (a stop may contain different sequence number, depending on the route that the vehicle is operating or even terminal has sequence number 0 or the last sequence number of a route.
Card type code	The fare code related to the card.
Card type name	The fare name related to the card (ex: Sub18/Sub23 is the student card).

Table 4.1: Columns of a dataset from CARRIS AFC system.

Card ID	This identifier number is related with card and not to a individual person, it is used to cross records.
Boarding datetime	The date and time registered on boarding station
Boarding Station	The code registered on boarding station
Alighting datetime	The date and time registered on alighting station
Alighting station	The code registered on alighting station

Table 4.2: Columns of a dataset from METRO system.

4.1.2 Preprocesssing

Selected sample

Before proceeding with the use of the dataset, it is necessary to carry out a data processing, namely to remove transactions that do not contain the card's identification number, because all the research it is based on the card id tracing. It is the case of on-board fares, purchased inside the vehicle, are validated in moment of the payment however the transaction doesn't contain a card id. Hardware problems can also be a cause for the loss of the card number.

Network support tables

In the process of estimating exits, some tables of the network of bus transponders are used to determine the bus stops and time of disembarkation. The exit estimation algorithm that will be explained in the next sub-chapter, can be explained by two phases: **phase one**, where the location of the landing stop is estimated; **phase two**, where the time for landing is estimated. And for each of these phases there is a different data source.

- **Table for location estimation**: has stop sequences for each route code. The columns of this dataset are **route code, variant, orientation, stop code, stop name, order number, and distance to the next stop** . This table is private and was provided by CARRIS
- **Table for time estimation**: has the columns attributes **stop sequence number, route code, stop code** (this columns referred until now are present in the sub section dataset structure from CARRIS), **route id, trip id,time (H:M:S), begin date, end date**. This new table is product of merge between some public GTFS tables . It is illustrated in the Figure 4.2.

These entire tables are loaded from the database and transformed into a dictionary. The complexity of insert a item from database is in the worst case $O(n)$, where n is the total records number of a dataset, and the average is $O(1)$, so to build a dict from database is $O(n * n)$ or in average case $O(n)$. At first sight can seems a bad solution, but the other solution possible it could be query the entire table and use the framework pandas that builds data frames. But build this data frame structure takes more time, more memory usage, and more and bigger time complexity in the search ($O(n)$), while a dictionary is the opposite, it is simple structure, less memory usage, and time complexity in the search ($O(logn)$). And the search in the data set is a crucial factor, because for each record of the sample data set a search will be made in these auxiliary data sets. So building the dictionary will be an undervalued cost.

On the other hand, there is also a reason to be collecting the entire table instead of making a query that collects the necessary data. After extracting the data from the database, they are immediately transformed into a pandas framework data frame. A search for data in the database for each record, implies the construction of a data frame, which has a very high cost in time and memory complexity.

The reader, at this point, may be asking, why not join these two tables into one, in order to save memory and have stop location and time in the same row. However in the first table the primary key is composed by three attributes which are route, variant, orientation, and stop code, but in the second table we don't have variant, orientation.

Table for time estimation

As previously mentioned, it was necessary to build a new table by joining some tables from the GTFS data source. This merge of tables was accomplished through the following query:

```
SELECT distinct c.id_rota, a.id_viagem, a.num_sequencia,
    c.cod_rota, a.cod_paragem, a.tempo, d.data_inicio, d.data_fim
FROM public.carris_sequencia_paragens as a,
    public.carris_viagens as b,
    public.carris_rotas as c,
    public.carris_viagem_calendario as d
WHERE a.id_viagem = b.id_viagem
```

```
AND a.id_viagem = d.id_viagem AND b.id_rota = c.id_rota
```

Table 4.2 illustrates the junction of the GTFS tables to create the new support table.

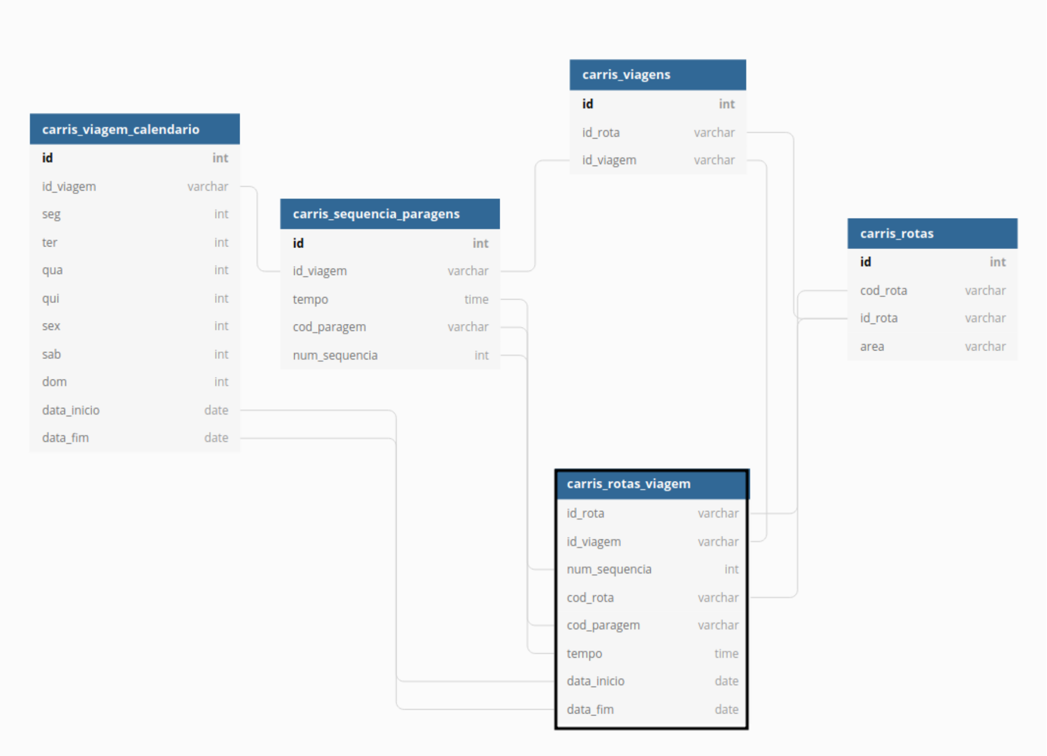


Figure 4.2: Merge tables for new table called carris rotas viagens.

Subway and bus data consolidation

The second model proposed for the alighting bus estimation relies on the use of the data set that contains records of transactions in the public transport subway METRO (Table 4.2), from the time interval of October 2019. And therefore it is necessary to merge these two datasets.

The data set from the metro has about 30 million records however, only a part of these transactions are important for the study. It is only necessary to select, from metro data set, the records of passengers who used both modes (metro and bus).

In this way, we proceeded with the intersection of the card ids of each dataset, to obtain passengers who used both modes. This intersection was made using a SQL query, because the database has an efficient mechanism for this task, which it is illustrated by a SQL process in the Figure 4.3), such us hash functions. The resulting common card ids between the two datasets are used to extract, from the METRO data sample, the records containing these card ids. The result of this operation is inserted into a new table in the database.

Finally, as soon as the important part of the METRO transactions are extracted, we can proceed with the union with the data set that contains CARRIS transactions. The new filtered sample from METRO validation system has around 19.5 million of transactions and the processed sample from CARRIS

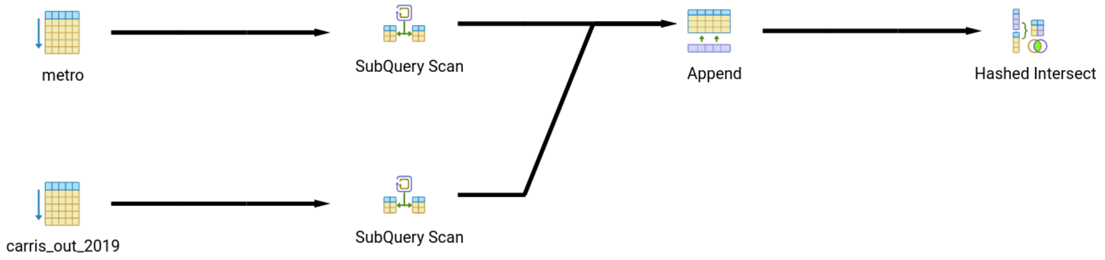


Figure 4.3: Intersection of tables, in order to obtain common transport modes (subway and bus) users.

validation system has 11.3 million transactions. Load all the entire tables from the database is not a solution, because it is not supportable to load so much data into memory.

To overcome this limitation, the sample of data corresponding to the time interval of one day of each dataset is loaded into memory and inserted in a new table in the database. This procedure is performed for each day of the month of October 2019. It should be noted that the period of one day corresponds between day X at 04:00:00 until day X + 1 at 03:59:59.When it is necessary to collect the data ordinarily chronologically, in the output estimation algorithm, it will already be a little sorted.

4.2 Alighting Stop Inference

The goal of this chapter is to present a methodology for estimating the alighting of the passenger for each boarding transaction, collected from entry-only AFC system of operator CARRIS. Two models are proposed to face this challenge. The first model only uses data described is the subsection Selected Sample, which means and therefore will only be observing transactions that occurred in the AFC system of CARRIS. The second model uses data described in the subsection Subway and bus data consolidation, which means that the model will be estimating the exits for each transaction of the CARRIS data sample, but nevertheless the model will be observing the passenger's journey that took place in the metro and bus transport modes. For short, the first referred model will be called bus model, and the second referred model will be called dual mode model.

4.2.1 Bus model

The algorithm acts on the travel segments of each passenger on a given day. Which means that the first step of the algorithm is, precisely, to collect from the database, transactions that took place during a period of time of one day. It is necessary to emphasize that the period of activity of the CARRIS operator occurs between day X at 04:00:00 t and day X + 1 at 03:59:59, and therefore will be that interval to be extracted. This process is carried out for all days of the month of October and the transactions, collected from each day, are sorted by card id and chronologically. Each step of the algorithm is represented in the flowchart of the Figure 4.4 and it will be explained in the follow enumeration:

- **Step 1**:If there are no more transactions to read then the algorithm ends, otherwise it goes to step 2.

- **Step 2**: Since there are more transactions to read, the card id of the passenger of the next transaction is read, which it will be denominated as s. Next step 3.

- **Step 3**: It is collected a new transaction T_n, s where n is order number of the segment and s is denomination for the actual passenger. The transaction has the columns described in the Dataset structure from CARRIS subsection. Next step 4.

- **Step 4**: Check if this transaction $T_{n,s}$ is the only one for the passenger s. If it is unique transaction then we skip step 5. Otherwise, step 6.

- **Step 5**: In this case, it was not possible to identify a landing stop. So, we go back to step 1 in order to infer the stop exits of the next transactions, for a new user s.

- **Step 6**: If $n = 1$, then this is the first transaction that occurred on the day, by passenger s, and therefore we jump to strep 7. If not, jump to step 10.

- **Step 7**: We arrived at this step, because we are dealing with the first transaction of the day, and the action to be taken will be to save the information of this transaction (location of the stop) to use later. After completing this step we skip to step 8.

- **Step 8**: Before determining the location and time of disembarkation of this transaction, it is necessary to obtain the location and time of the next transaction $T_n + 1, s$ (embark on the following travel segment). After completing this step we skip to step 9.

- **Step 9**: In this step, the landing stop and exit time for this travel segment $T_{n,s}$ are inferred. This step is explained in detail by the flowchart in the Figure 4.5. Then we go back to step 3, to help infer the stop exits of the next transaction.

- **Step 10**: If the transfer $T_{n,s}$ is the last one performed during day by passenger s, then we proceed to step 11. Otherwise, step 8.

- **Step 11**: When the first transaction of the day was observed, we kept the information of the place of departure and time of boarding the bus. This information will now be collected. Next step is 12.

- **Step 12**: With the information from transaction T1s collected in step 11, the stop and departure time for transaction $T_{p,s}$ will be determined, where p is the total number of transactions carried out by the passenger. Then we go back to step 1, to help infer the stop exits of the next transactions, for a new user s.

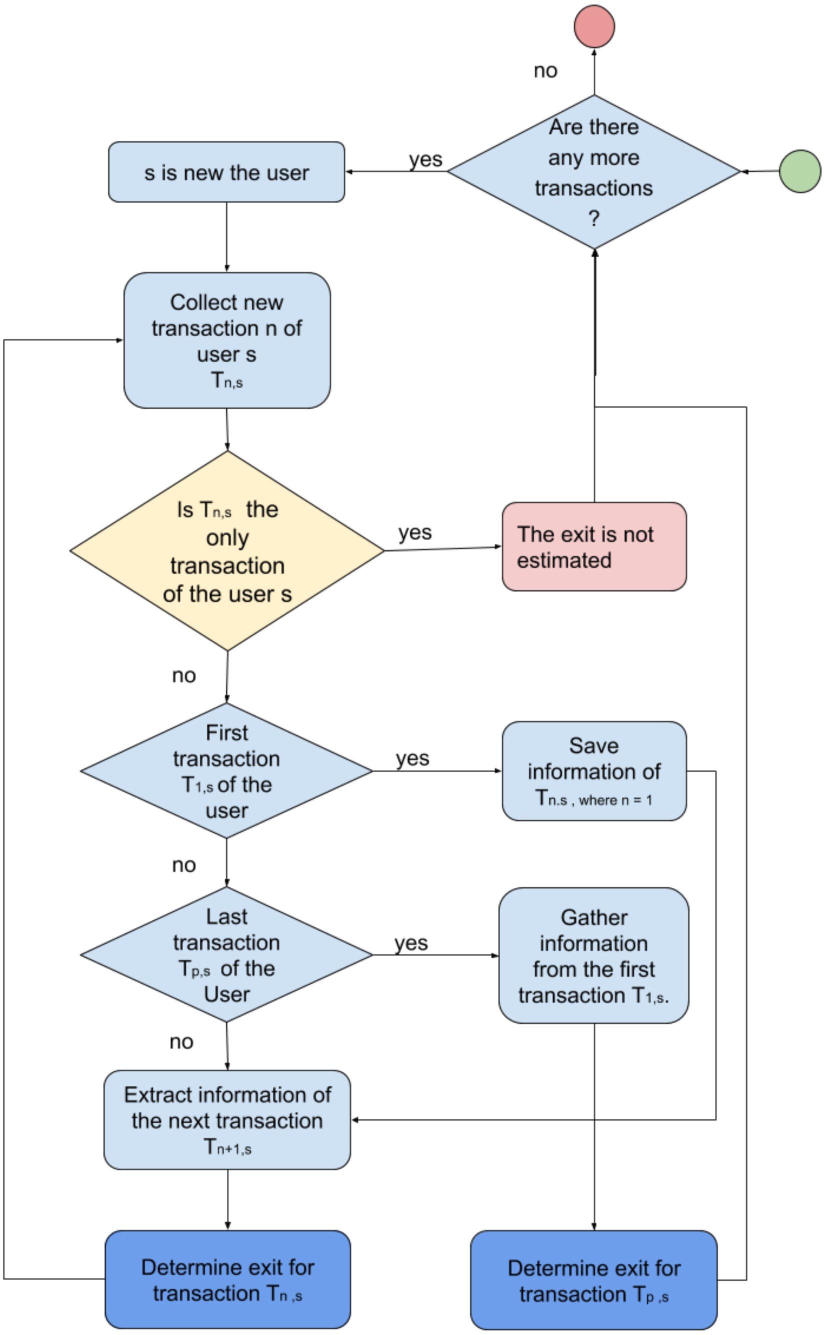

no
Are there any more transactions ?
yes
s is new the user
Collect new transaction n of user s Tn,s
Is Tn,s the only transaction of the user s
yes
The exit is not estimated
no
First transaction T1,s of the user
yes
Save information of Tn.s , where n = 1
no
Last transaction Tp,s of the User
yes
Gather information from the first transaction T1,s.
no
Extract information of the next transaction Tn+1,s
Determine exit for transaction Tn ,s
Determine exit for transaction Tp ,s

Figure 4.4: General exit estimation flowchart.

Steps 9 and 12 of the previous algorithm described, have the function of inferring the exit stop and the time in a given transaction. These steps are described in detail by the flowchart represented in the Figure 4.5. Basically, we can divide the algorithm into two phases: **phase A**, where we determine the exit stop geographically; **phase B** we determine chronologically the time at which the passenger disembarked, after boarding and recording the $T_{n,s}$ transaction.

Phase A (described by steps 1-6, 7) In this phase of the algorithm the Table for location estimation will be consulted described in Network support tables (chapter Preprocessing).

- **Step 1**: The $T_{n,s}$ transaction has the following attributes, route code, variant and orientation. So with these attributes we will check if there is any route in the table with these conditions. If not, then the process ends and the exit stop is not successfully estimated (we skip to step 7). Otherwise, we continue to step 2.
- **Step 2**: We extract the sequence number of the boarding stop (P_n where P is the stop and n the sequence number), from the transaction $T_{n,s}$. Next step 3.
- **Step 3**: With the sequence number extracted earlier, we consult the table to extract the stops that have an order number greater than n. We are currently extracting stops on the route c that are upstream from the stop n. Next step 4.
- **Step 4**: For each stop collected in the previous step, it is computed the distance between the candidate stop location and the boarding stop location in the $T_{n+1,s}$ segment will be calculated. If $T_{n,s}$ is the last travel segment performed by the passenger, then the distance of each stop will be calculated in relation to the $T_{1,s}$ travel segment. Next step 5.
- **Step 5**: The stop with the shortest distance to the boarding point in the $T_{n+1,s}$ segment is chosen. Next step 6.
- **Step 6**: Sometimes the passenger gets on the last stop and therefore there is no stop upstream. And in this rare exception, the landing stop is not successfully determined (step 7). Otherwise, continue to phase B, to determine the time at which the passenger disembarked (step 8).
- **Step 7**: Landing stop bus is not inferred.

Phase B (described by steps 7, 8-13)

- **Step 8**:In the new table that contains timetable and routes, candidate routes that contain the stop of boarding and the deferred stop code previously inferred will be searched. It is possible to find multiple travel identifiers because a route is performed at different times of the day and each of these trips has a different travel identifier.
- **Step 9**: If candidate travel identifiers exist, proceed to step 10. If not, then proceed to step 7. Which means that the landing stop is not successfully inferred. Since we are using two tables, the data may not be synchronized, and as a result there may be stops that are not present in one of the auxiliary datasets (auxiliary tables described in the section Network support tables).
- **Step 10**:Of the travel candidates, we select the one that contains the departure stop with the time closest to the time recorded in the transaction T_n, s.
- **Step 11**: From the trip chosen in the previous step, the time associated with the estimated stop code is extracted.
- **Step 12**: After extracting the two times associated with the entry stop and inferred exit stop, the time interval between these two times is calculated.

- **Step 13**: The time interval calculated in step 1, is added to the boarding time recorded in transaction T. In this way, we are better expressing the reality of delays and advances that may arise on routes.

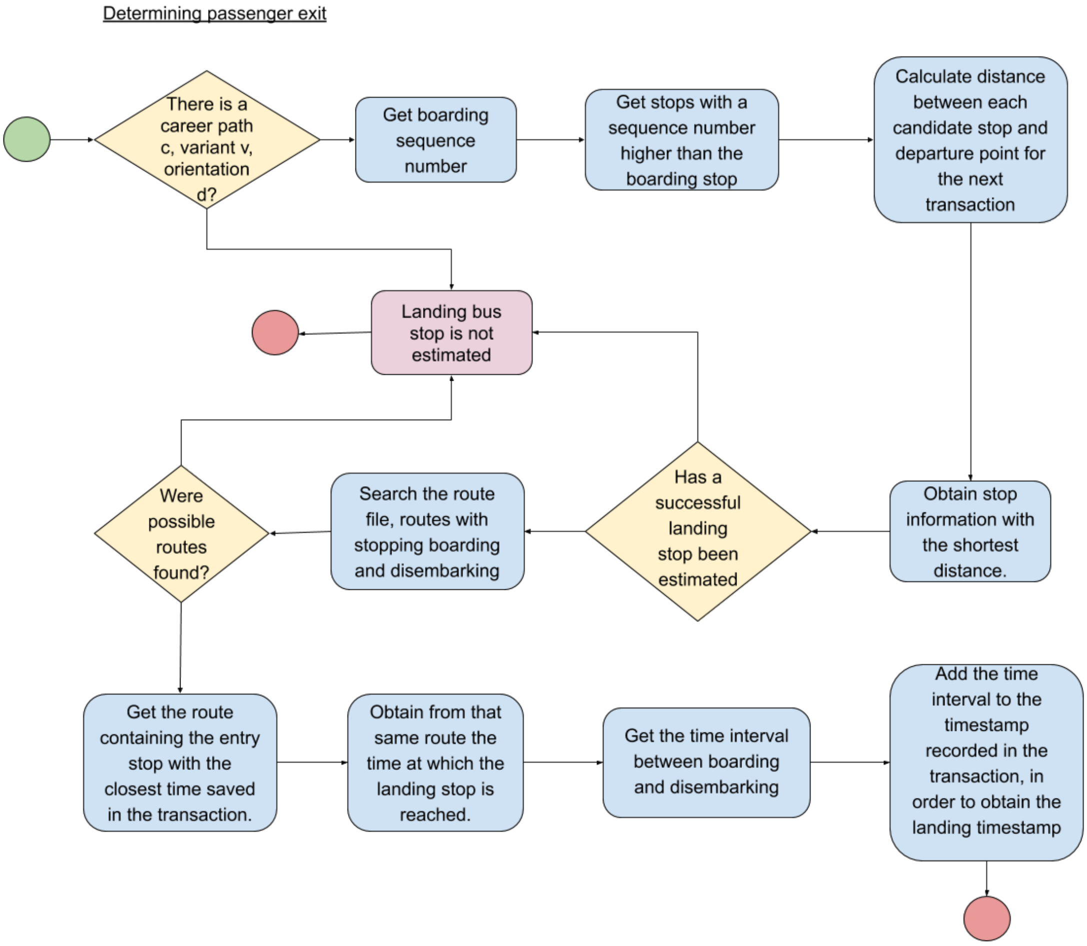

Determining passenger exit
There is a career path c, variant v, orientation d?
Get boarding sequence number
Get stops with a sequence number higher than the boarding stop
Calculate distance between each candidate stop and departure point for the next transaction
Landing bus stop is not estimated
Were possible routes found?
Search the route file, routes with stopping boarding and disembarking
Has a successful landing stop been estimated
Obtain stop information with the shortest distance.
Get the route containing the entry stop with the closest time saved in the transaction.
Obtain from that same route the time at which the landing stop is reached.
Get the time interval between boarding and disembarking
Add the time interval to the timestamp recorded in the transaction, in order to obtain the landing timestamp

Figure 4.5: Flowchart showing in detail the determination of exit information.

4.2.2 Dual mode model

The aforementioned algorithm has limitations for situations in which user uses more than one transport mode in their path. For example, the following fictitious situation may occur: passenger S_1 takes off from stop P_{20} to stop P_{29} on route R_4 at 8 am. However, at the end of the day, it travels by subway transport between station E_2 and station E_4 on line Z_1. The previous algorithm, as it only receives information about the paths taken on buses, so it will only see a segment of travel (bus segment) and consequently is not able to estimate the P_{29} exit stop, performed on route R_4.

To overcome these limitations, a dataset for the period of October 2019 was made available by the operator METRO (public transport subway). In the preprocessing sub chapter, in the section Subway and bus data consolidation, it is shown how we merged the two datasets (metro and bus transactions) from the same period. This final dataset will allow better tracing of multimodal passenger paths and as a result it will be possible to correctly infer the landing stops in the bus travel segments (when the next travel segment is a metro or bus route).

In cases where the following travel segment $T_{n+1,s}$ corresponds to another travel mode (taxi, boat, bicycle) it will not be possible to correctly estimate the departure stop of the transaction $T_{n,s}$. In Figure 4.6, we describe the process of estimating outputs for transactions that took place in the CARRIS transport network, through a flowchart.

The estimation process is practically the same as the previous one, with the exception of some rules such as:

1. If the algorithm is estimating the departure to a bus transaction $T_{n,s}$, and the next travel segment $T_{n+1,s}$ corresponds to a subway segment transaction, the algorithm will use the boarding station location and the boarding time $T_{n+1,s}$ to determine the bus exit of transaction $T_{n,s}$.

2. If the first transaction $T_{1,s}$ is a metro route, the relevant information is saved (to infer the bus station of the last segment), and the algorithm continues the inference process on the following transactions, by user s.

3. Transfers that take place in the metro do not require the inferencing process. They are only needed to collect information on the geographic location of the passenger and time of boarding at the metro station.

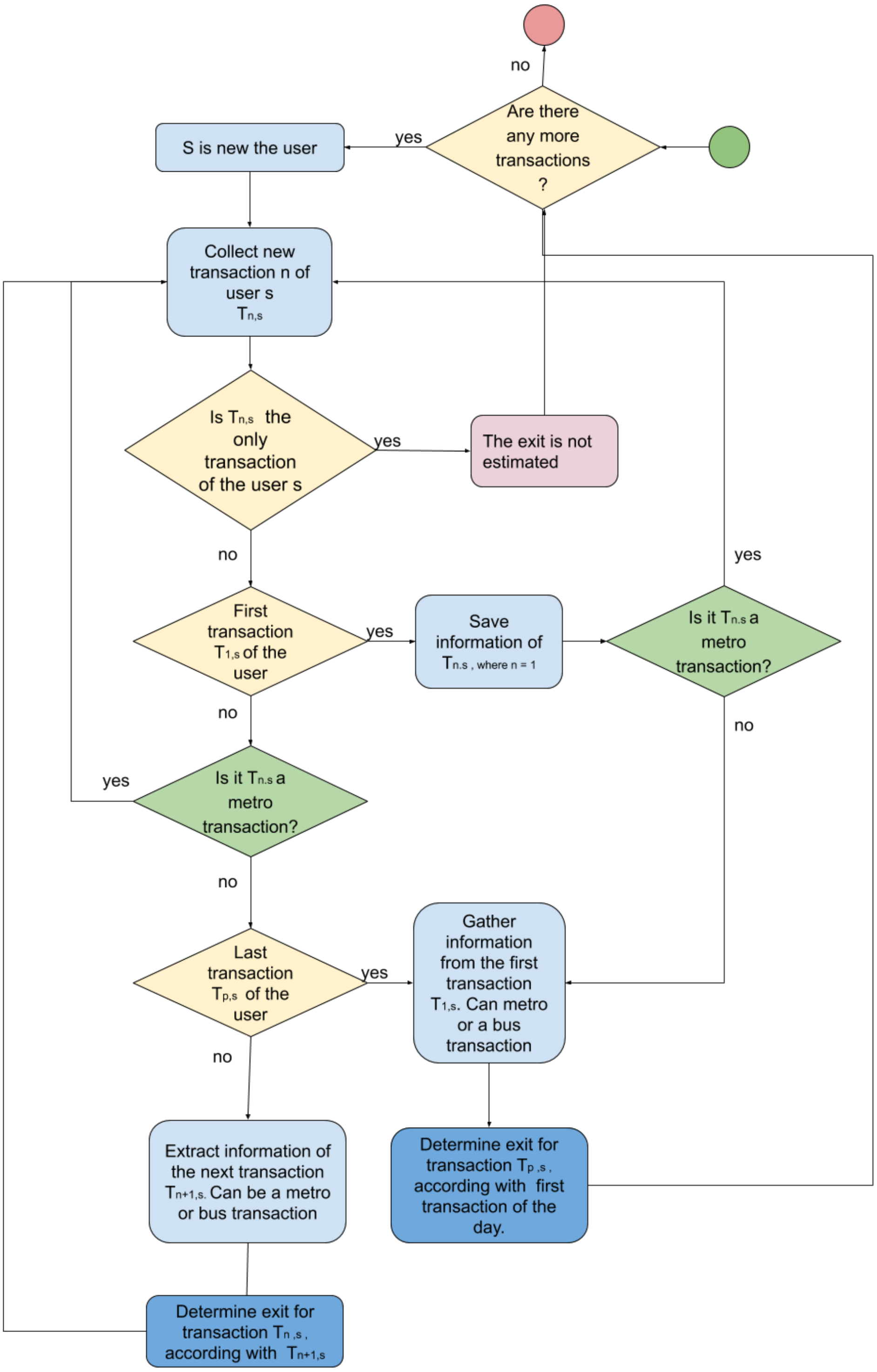

no
Are there any more transactions ?
yes
S is new the user
Collect new transaction n of user s $T_{n,s}$
Is $T_{n,s}$ the only transaction of the user s
yes
The exit is not estimated
no
First transaction $T_{1,s}$ of the user
yes
Save information of $T_{n,s}$, where n = 1
Is it $T_{n,s}$ a metro transaction?
yes
no
no
Is it $T_{n,s}$ a metro transaction?
yes
no
Last transaction $T_{p,s}$ of the user
yes
Gather information from the first transaction $T_{1,s}$. Can metro or a bus transaction
no
Extract information of the next transaction $T_{n+1,s}$. Can be a metro or bus transaction
Determine exit for transaction $T_{p,s}$, according with first transaction of the day.
Determine exit for transaction $T_{n,s}$, according with $T_{n+1,s}$

Figure 4.6: Flowchart of dual mode model to exit estimation.

Transactions that are the only ones carried out during the day and it was not possible to estimate an exit are also kept in a table for future investigations. Transactions in which an exit has been successfully estimated are stored in a table in the database.

Attributes	Description
Card ID	Attribute copied from the $T_{n,s}$ transaction of the input dataset. Mentioned in the Table 4.1.
Route	Attribute copied from the $T_{n,s}$ transaction of the input dataset . Mentioned in the Table 4.1.
Stop code boarding	Attribute copied from the $T_{n,s}$ transaction of the input dataset. Mentioned in the Table 4.1.
Datetime boarding	Attribute copied from the $T_{n,s}$ transaction of the input dataset. Mentioned in the Table 4.1.
Stop sequence boarding	Attribute copied from the $T_{n,s}$ transaction of the input dataset. Mentioned in the Table 4.1.
Stop code alighting	Attribute inferred by the algorithm. It indicates the code of exit stop in the transaction $T_{n,s}$
Datetime alighting	Attribute inferred by the algorithm. It indicates the exit time stop in the transaction $T_{n,s}$.
Stop sequence alighting	Attribute inferred by the algorithm. It indicates the sequence number of the exit stop in the transaction $T_{n,s}$.
Trip identifier	Attribute inferred by the algorithm. Identifier of a trip that depends on the route code, orientation, variant, and time.
Card type code:	Attribute copied from the $T_{n,s}$ transaction of the input dataset.
Walking distance	Attribute inferred by the algorithm. Distance between the alighting location of transaction $T_{n,s}$ to boarding location in transaction $T_{n+1,s}$. Inferred by the algorithm.
Transfer time	Attribute inferred by the algorithm. Interval time between the alighting location of transaction $T_{n,s}$ to boarding location in transaction $T_{n+1,s}$. Inferred by the algorithm.
Path distance	Attribute inferred by the algorithm. Distance traveled by the passenger in the travel segment $T_{n,s}$.
Travel time	Attribute inferred by the algorithm. Travel time spent in the travel segment $T_{n,s}$.
Next mode	Attribute inferred by the algorithm. It indicates which mode was used after transaction $T_{n,s}$ (bus or metro). This attribute only belongs to the dataset produced by the second model.

Table 4.3: Attributes from output dataset computed by alighting inference model.

After filling in the new table with the attributes mentioned above, it will then be updated with other attributes that express and characterize passengers and their routes. That is, for each row in the new table are the following attributes.

Added attributes	Description
Segment Count	Number of times the passenger made the segment from stop A to stop B, during the entire month
Route Count	Number of times the passenger used route R, during the entire month
Total segments	Number of travel segments the passenger has completed.
Eval	Assessment that conveys the level of confidence with which the exit stop was inferred, which ranges from 0 to 100. This assessment depends on the following attributes already mentioned: Segment Count, Total segments, Walking distance, which has the following weights in the evaluation respectively, 0.1, 0.2, 0.7. This metrics is subjective and will not be given great importance to research.

Table 4.4: Added attributes to the output dataset from alighting inference model.

4.3 Trip Generation for commuting travel

Trip generation is crucial for evaluating the transportation impacts where they act. One of the applications of this process is the generation of commute trips, for example routine trips, home-work commute, home-school commute. The models of generation of commuting trips aim to derive, from a set of travel segments, the origin and the proposed destination location and therefore the travel segment between these two points are no longer described. This derivation is interesting from the point of view of analysis, planning and improvement of the transport network, which is the object of analysis. An application of the result of this model can be applied in the following example: if there is a high demand between point of origin A and destination C, and there is a transfer point at B, then the operator can rethink the routes, in order to take passengers from point A to C, without transfers.

In this section, we propose a model that makes travel segments resulting from the exits inferences multimodal model (previously stated) and derives origin-destination trips.

The proposed model for the generation of commuting trips is based on rules:

1. A passenger who makes a commuting journey is willing to repeat the same route frequently.Therefore, a threshold is defined to eliminate passenger travel that does not reach this limit.
2. Among the transfers between travel segments, which took place during the day, only the transfer with the longest time interval is considered as activity time (work, school) and should have a time interval greater than the stipulated.
3. If the distance between segments is above a certain threshold previously defined, then there may be present segments of travel incorrectly estimated by the model of inference of exits.

The main algorithm is described by the follow steps and illustrated by the flowchart in the Figure 4.7. Then the sub process of derive/identify commute trips is explained by the flowchart shown in the Figure 4.8. Notice, that the process described is performed for each day of October.

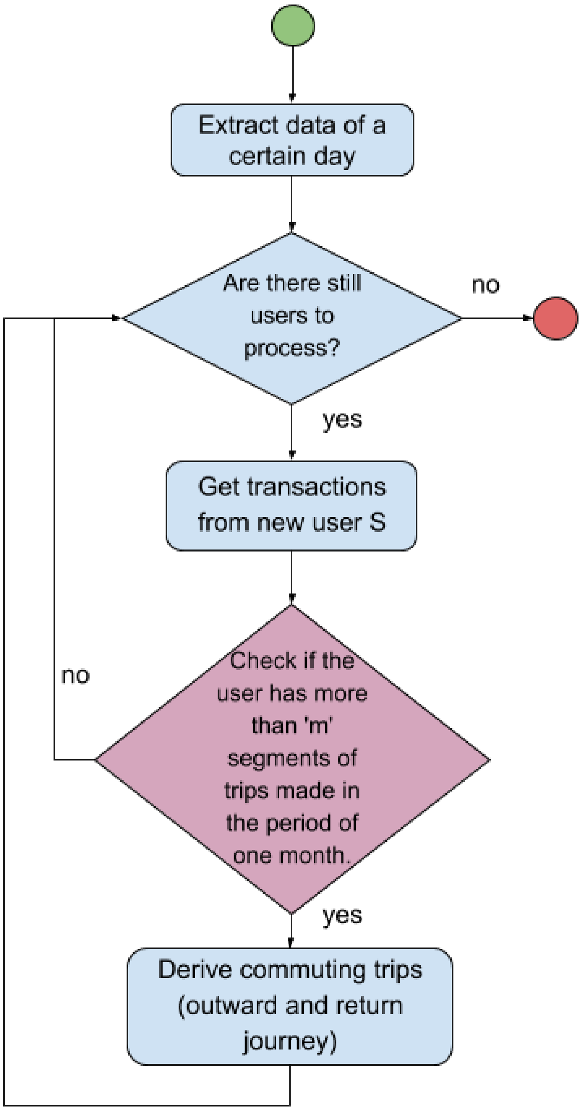

Extract data of a certain day
Are there still users to process?
no
yes
Get transactions from new user S
no
Check if the user has more than 'm' segments of trips made in the period of one month.
yes
Derive commuting trips (outward and return journey)

Figure 4.7: Determining commute journeys.

- **Step 1**: Extract data from the dataset, Extract travel segments between the time period from day X at 04:00:00 to day X+1 at 03:59:59. Continue to step 2.

- **Step 2**: Check if there is any travel segment to read, if exists continue to step 3, otherwise the process ends. The travel segment is described by the columns of the last referred dataset.

- **Step 3**: Get travel segments from new passenger, which we will call as S. Continue to step 4.

- **Step 4**: Checks whether a passenger has completed more than 'm' travel segments. 'm' is the defined limit to consider that the passenger makes commuting trips. Checks whether a passenger has completed more than 'm' travel segments. 'm' is the defined limit for considering that the passenger makes commuting trips. If yes, then proceed to step 5, otherwise go to step 3 to read travel by another passenger.

- **Step 5**: In this step, commuting trips are derived, that is, from a set of segments, a one-way trip to the place of activity and a return to the home are defined. This process is explained in detail by the flowchart in the Figure 4.8.

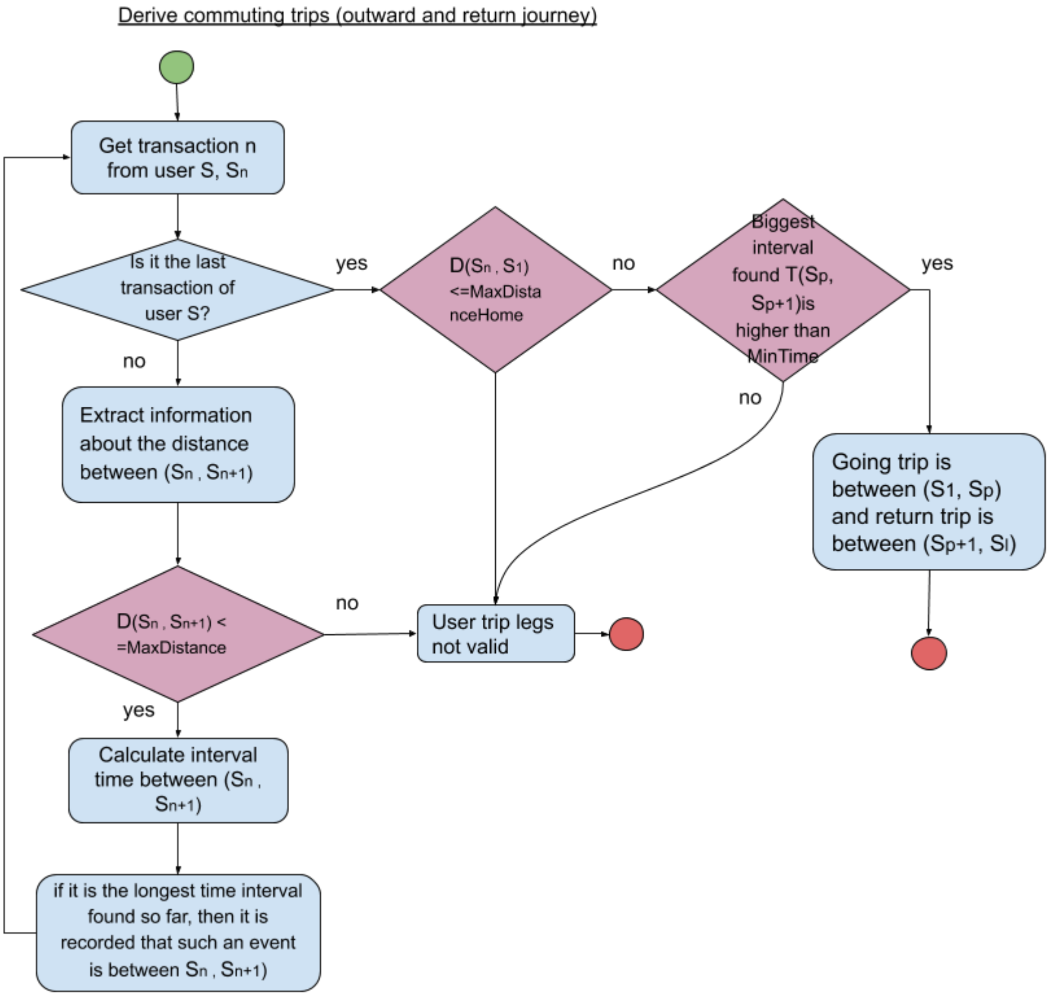

Derive commuting trips (outward and return journey)
Get transaction n from user S, Sn
Is it the last transaction of user S?
yes
no
D(Sn , S1) <=MaxDistanceHome
no
Biggest interval found T(Sp, Sp+1)is higher than MinTime
yes
no
Extract information about the distance between (Sn , Sn+1)
Going trip is between (S1, Sp) and return trip is between (Sp+1, Sl)
D(Sn , Sn+1) < =MaxDistance
no
User trip legs not valid
yes
Calculate interval time between (Sn , Sn+1)
if it is the longest time interval found so far, then it is recorded that such an event is between Sn , Sn+1)

Figure 4.8: Derive commuting trips, in detail.

- **Step 1**: The process deals with passenger S and in this step we extract a travel segment S_n. Next step 2.
- **Step 2**: If S_n is the last travel segment of the passenger then continuous to step 7, otherwise the continue to step 3 .
- **Step 3** : Extract the information about the distance between segments S_n and S_{n+1} (transfer distance). Next step 4.
- **Step 4**: If transfer distance is higher than MaxDistance (threshold for distance in intermediate transfers) then the process ends and concludes that the passenger has not made a valid commuting trip (step 10). Otherwise continues to step 5.
- **Step 5**: Calculate interval time between segments S_n and S_{n+1} (transfer time). Next step 6.

- **Step 6**: Save the time calculated in step 5 and the transfer information (from S_n to S_{n+1}), if it is the longest interval time found so far. Then go back to step 1

- **Step 7**: If the algorithm reached step 7, it means that it has already read the last travel segment of the day, made by the passenger. And it is necessary to check if the distance from the exit at S_n to the point of embarkation on the first trip segment of the day S_1, is below the threshold MaxDistanceHome. If affirmative, continue to step 8. Otherwise 10, the process ends and concludes that the passenger has not made a valid commuting trip (step 10)

- **Step 8**: In step 6 the transfer with the longest break was saved, because it corresponded to the time of activity (rule number two). However, it is still necessary to check if this time interval is longer than desired (MinTime threshold). If it is true then between the segments S_p and S_{p+1}, (where p defines the travel segment that is destined for the location of the activity), an activity has occurred and therefore it proceeds to step 9. Otherwise, the process ends and concludes that the passenger has not made a valid commuting trip (step 10)

- **Step 9**: In this step, trips from home to activity (S_1 to S_p) and activity travel to home (S_{p+1} to S_l, where l is the last trip segment) are generated.

- **Step 10**: The process ends and concludes that the passenger has not made a valid commuting trip (step 10)

During the trip generation process, the following attributes are calculated:

Column	Description
Transfer Count	Indicates the number of transfers found between the source and the destination.
Activity time	Time interval indicating the time spent on the activity . Th segment were the destination is home, doesn't have interval time, instead it isn't filled
Sum Transfer Time	Sum of transfer times found between origin and destination
Sum Transfer Distance	Sum of transfer distances found between origin and destination
Travel Time	Time spent on the route between origin and destination. The time interval for transfers is discounted.
Path distance	Distance traveled by the passenger between the point of origin and destination, but travel in transfers does not count.
Eval	Average of the evaluation of the travel segments between origin and destination. This assessment is based on the description of the segment.
Metro used	I is a flag that indicates if between the origin and the destination metro was used. It is possible to compute this attribute if the input dataset is the result of the multimodal inference model. Otherwise it will not appear.

4.4 Origin Destination Matrices Inference

Conventionally, origin destination (OD) matrices are tables that describe people movement between locations, as it shows in the Figure 4.9, where the entry and exits locations are bus stops, and the tone of cells indicates the passenger volume. OD matrices are extremely useful for planning and improve a good public transportation system.

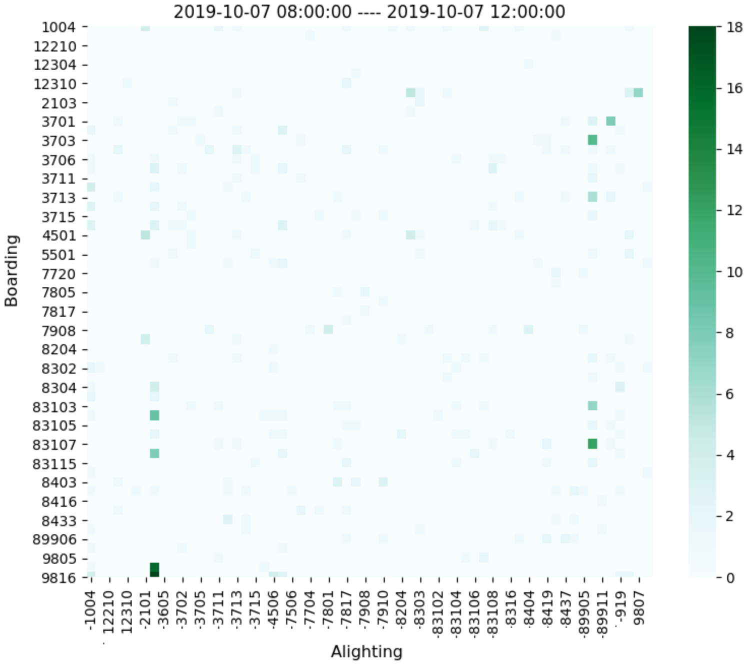


Figure 4.9: An illustrative Figure of a rOD matrice with stop granularity, with trips travelled in route 759, in the period of 8 am to 12 pm at the day 7 of October.

In this study, matrices with two different contents were developed:

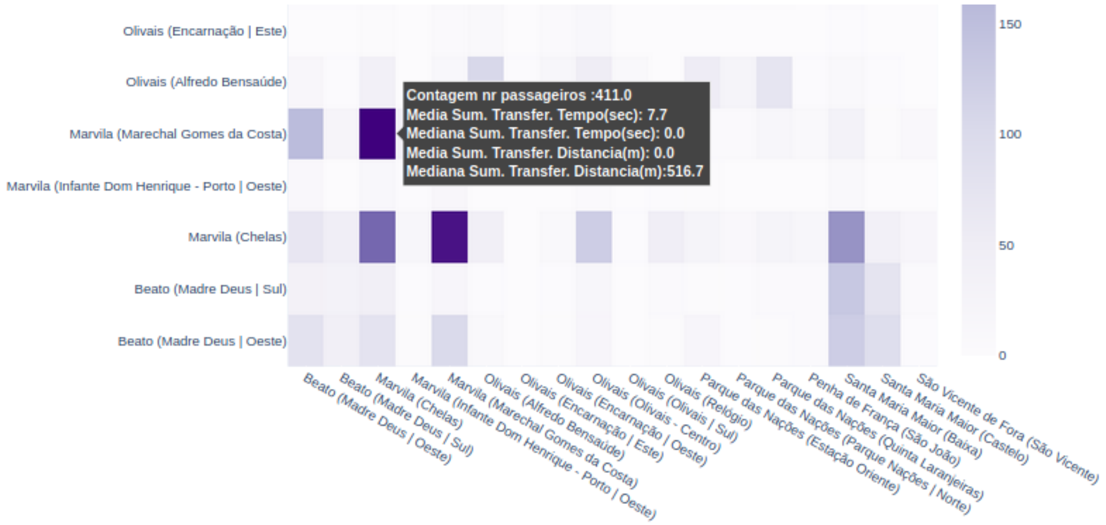
- **Time-based matrices(tOD)** Matrices that only show demand between origin and destination using travel segments.
- **Routine-based matrices (rOD)** Matrices that present demand between origin and destination in commuting trips.

This solution allows **filtering** the content of the matrices through the following **parameters**:

1. Select range of days
2. Time window time window between 0 am and 12 pm
3. Select one and more week days
4. Filter by title card
5. Select origin/boarding routes or even stops

6. Select destination/exit routes or even stops

As previously mentioned, in this study it was concerned with studying other metrics besides the passenger flow between points. And therefore, for each type of matrix (rOD or tOD) metrics are provided, in addition to the demand between origin and destinations.



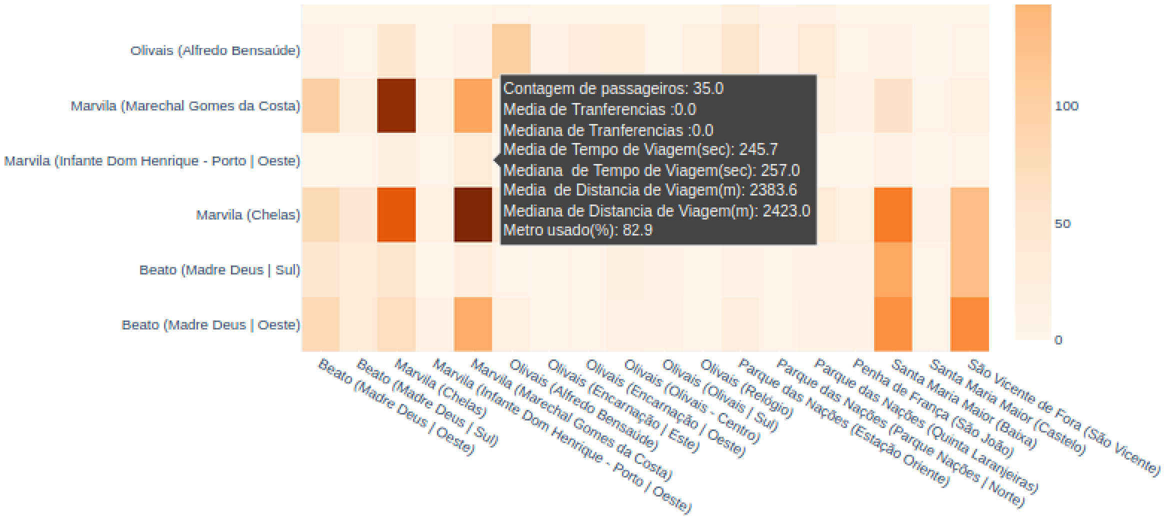
Contagem nr passageiros :411.0
Media Sum. Transfer. Tempo(sec): 7.7
Mediana Sum. Transfer. Tempo(sec): 0.0
Media Sum. Transfer. Distancia(m): 0.0
Mediana Sum. Transfer. Distancia(m):516.7

Figure 4.10: Matrices with granularity TAZ, displaying a tool tip with information regarding the transfers.

In the **tOD matrix** it is possible to view the following metrics in the cells:

1. **Passenger counting**

2. **Percentage of passengers who**, after traveling in the trip segment, **used metro**, in the next segment.



Contagem de passageiros: 35.0
Media de Tranferencias :0.0
Mediana de Tranferencias :0.0
Media de Tempo de Viagem(sec): 245.7
Mediana de Tempo de Viagem(sec): 257.0
Media de Distancia de Viagem(m): 2383.6
Mediana de Distancia de Viagem(m): 2423.0
Metro usado(%): 82.9

Figure 4.11: Matrice with granularity TAZ, displaying a tooltip with information regarding the travels.

In the **rOD matrix** it is possible to view regarding the travel, as it shows in Figure 4.10 and transfers information, as it shows in Figure 4.11:

1. **Travel information**: information on the path taken inside the buses is displayed in the cells of the matrix. Information regarding transfers are discounted.
   (a) **Passenger counting**: Count of passengers who made the journey between origin and destination.
   (b) **Mean and median Transfers**: Average and median value that reflects the number of transfers needed for it from origin to destination. Metric calculated through the attribute **Transfer count** described in the Table 4.3.
   (c) **Mean and median travel time** : application of media and median on attribute **Travel Time** described in the Table 4.3.
   (d) **Mean and median travel distance**: application of media and median on attribute **Path distance** described in the Table 4.3.
   (e) **Percentage of journeys with metro segment** : when the passenger uses subway transport within a journey. Metric calculated from the attribute **Metro used** described in the Table 4.3.
2. **Transfer information**: information regarding time and distance spent between travel segment transfers.
   (a) **Passenger counting**: Count of passengers who made the transfer between destination and origin.
   (b) **Mean and median transfer time**: average and median value that reflects the time spent walking and/or waiting between transfers. Metric calculated from the attribute **Sum Transfer Time** described in the Table 4.3.
   (c) **Mean and median transfer distance** average and median value that reflects the walking distance between transfers. Metric calculated from the attribute **Sum Transfer Distance** described in the Table 4.3.

Finally, the matrices can assume different granularities in origin and destination location. In other words, instead of origin and destination bus stop, it can be an aggregation of stops located in a geographic area. The spatial aggregations of stops considered are as follows:

1. **Traffic analysis zones (TAZ)**: A traffic analysis zone is a geographical unit used in conventional transport planning models. Figure 4.12 illustrates a matrix where it shows the passenger volume between TAZ sections.
2. **Statistical sections**: a section is a territorial unit corresponding to a continuous area of a single parish with about 300 housing units.

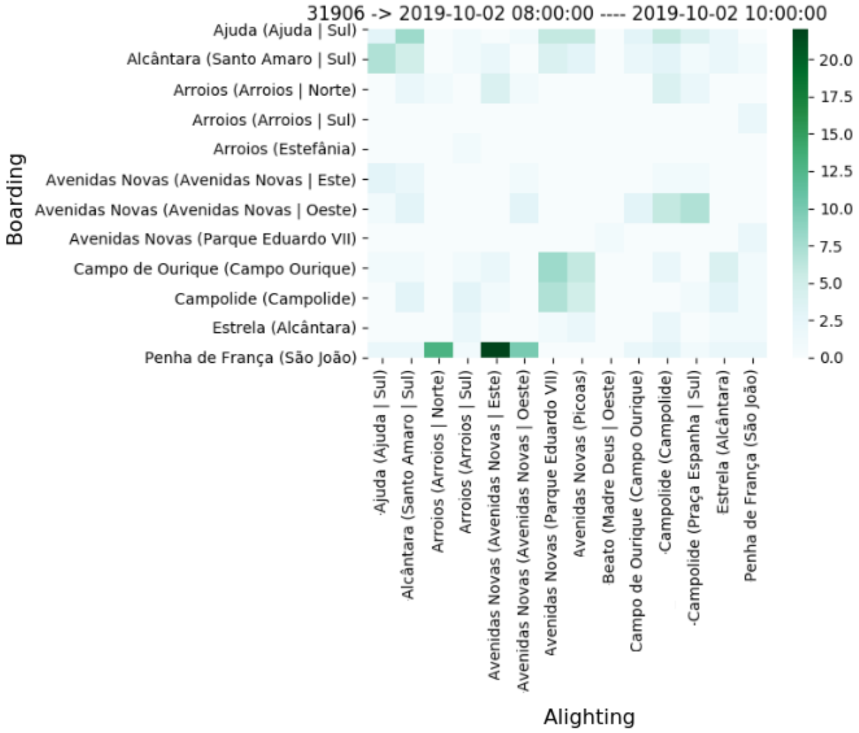


Figure 4.12: An illustrative Figure of a rOD matrice with TAZ granularity and filtered for a card title, with trips travelled in route 722, in the period of 8 am to 10 pm at the day 2 of October.

Chapter 5

Results

This chapter consists of five sections that make a descriptive and exploratory analysis. Each chapter represents a sub-goal to achieve the final contribution, which is the presentation of the support tool for mobility in the CARRIS public transport, whose function is to display the inferred dynamic Origin-Destination matrices.

The first section is called Exploration Data Analysis of boarding datasets and it compares two samples of data that correspond to the validations that occurred in the CARRIS AFC system in different periods and (October 2018 and October 2019). This study aims to demonstrate the similarity of mobility between different months. The next sections continue the study with the dataset corresponding to the period of October 2019, because one of the proposed algorithm uses an auxiliary dataset of the same period but from other public transport mode - subway.

The second section called Alighting stop inference results shows the performance of the two models proposed in the Proposed Solution section and defines through the results which of the models presented the best robustness to deal with the inference problem of stop exits in the CARRIS network validations. It is recognized that the Dual mode Model presents better performance and a sensitive analysis is made to the attributes resulting from the inference algorithm, such as times and distances spent on transfers. The presentation of this sensitive analysis allows us to see if the results are in accordance with the assumptions found in other studies mentioned at the Related work chapter. Afterwards, thi section is enriched with an exploratory analysis on the mobility behavior in relation to title cards that relate to different age groups.

The third section called Trip Generation for commute travel analysis presents a sensitive analysis of the attributes resulting from commutative trips identified in the period of October 2019. In this section the behavior of commute mobility is studied, through characteristics such as the percentage of trips with zero or more transfers. As in the previous section, an exploratory analysis on title cards is carried out.

The fourth section called Origin Destination Matrices Description presents figures related to the inference of dynamic origin-destination matrices. In the previous sections the graphs were tailored for each purpose under study, however these next matrix graphs were taken from the visualization tool that is explained in the Visualization Tool chapter. This research explores the matrices in accordance with calendar, granularity, titles, among others. It aims to show a small fraction of researches over on mobility behavior in the CARRIS network.

Finally, the fifth section called Situational Contextual Discovery in Data aims to clarify the impact of the context on public transport, namely buses. The figures presented were elaborated within the scope of an article submitted and presented at ECT'2020.

5.1 Exploration Data Analysis of boarding datasets

This first section aims to present an overview of the CARRIS transport network between two periods. A descriptive analysis of the datasets' characteristics (table with passengers transaction in the network) for the periods October 2018 and October 2019, including a summarized and comparative analysis of the network between these periods.

Samples Description

As mentioned in the preprocessing section's subsection, the datasets suffered from filtering, namely with the withdrawal of transactions that did not contain a passenger card identifier number.

.	October 2018	October 2019
Stops	2070	2152
Routes	85	93
Passengers	724703	818297
Boarding Count	9 993 762	11 360 893

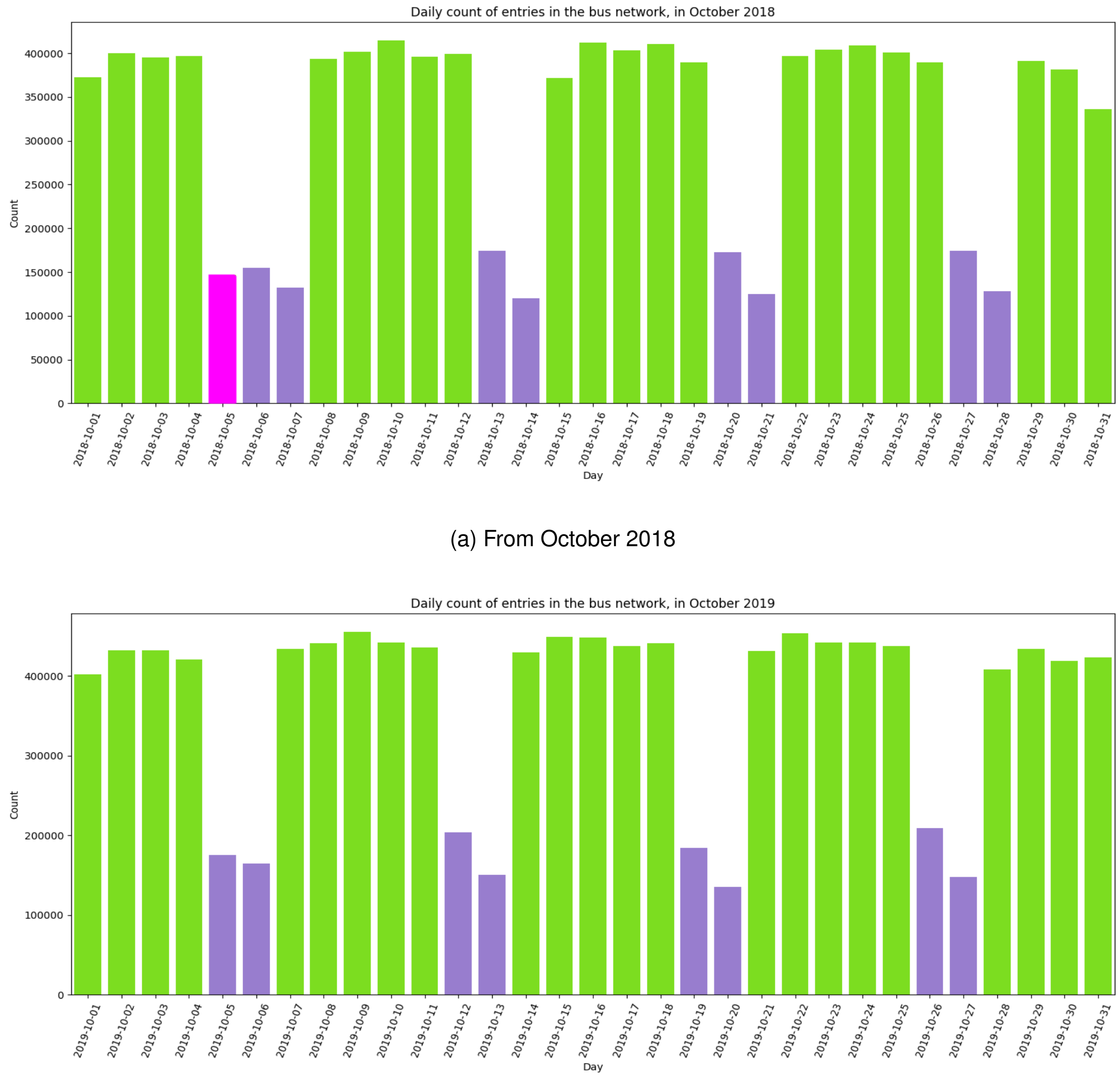
Table 5.1: Summary description of the 2018 and 2019 datasets.

Table 5.1 contains information that characterizes the datasets that contain the transactions carried out by passengers in the periods of October 2018 and 2019. The characteristics under analysis are the number of stops, routes, passenger identifiers, and transactions found in the datasets. It concludes that October 2019 concerning the same month of the year 2018, there was an increase regarding all attributes described in the table. Namely an increase of 12.9% of passengers and an increase of 13.6% of transactions the AFC system.

Comparative analysis of datasets

The exploratory analysis over the bus network can range from several points of view. We can do a comparative study that assesses passengers' entry over time and at different granularities, such as routes or stops. Therefore, in this section, various views of the network status are presented in the two periods mentioned, in the figures 5.1, 5.2, 5.3.

Figure 5.1 compares the daily count of entries in the transport network between the two periods (the month of October 2018 and 2019). The green bars correspond to the week's working days, and



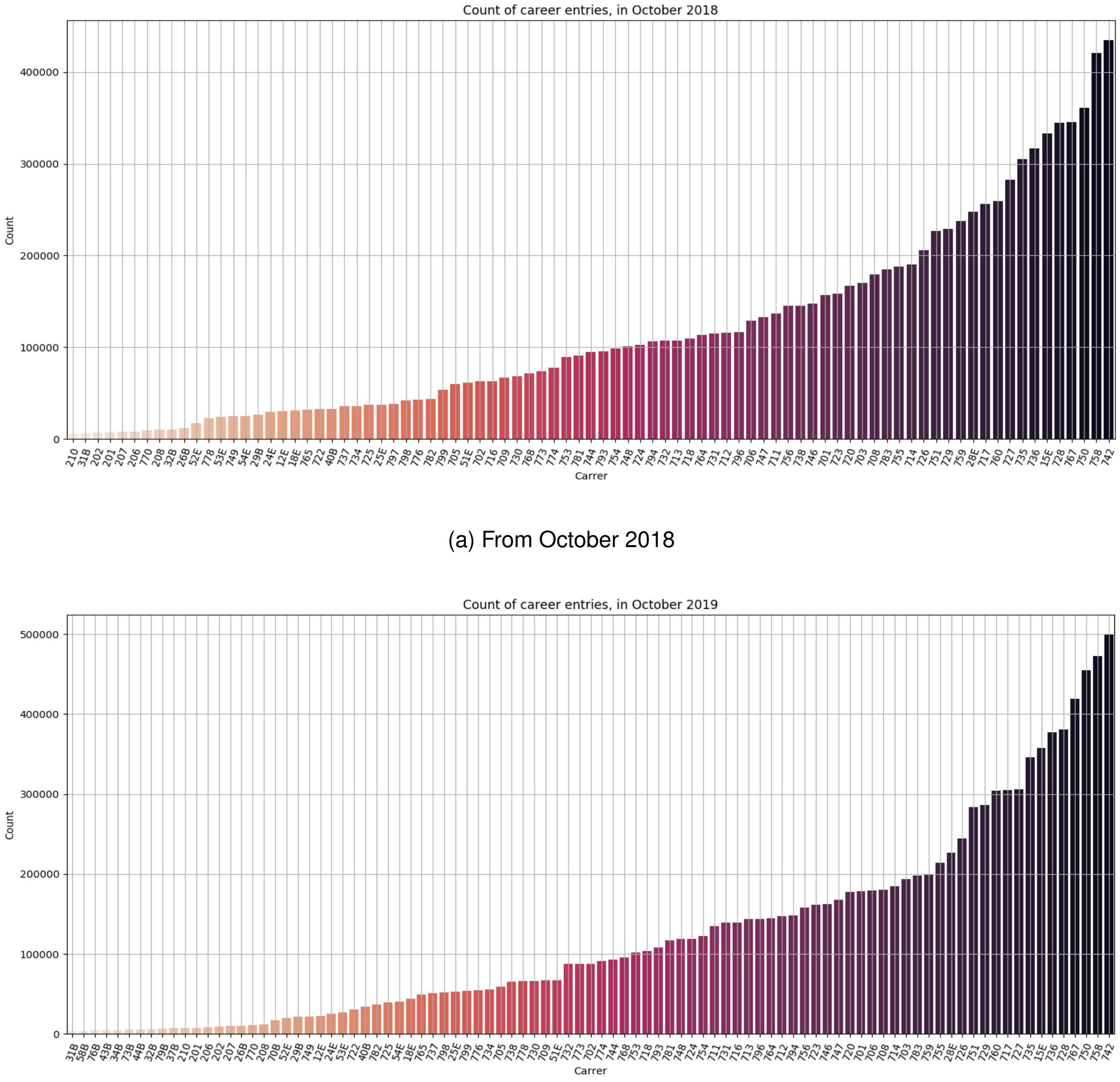
(a) From October 2018

(b) From October 2019

Figure 5.1: Daily count of entries in the bus network, of month October, in different years.

the violet-grey bars represent the weekend. Examining the subfigures, the following sentences can be affirmed: it is possible to verify that there is a greater influx (around the double) of entries in the network during working days against the weekend days. But if we look more closely, we notice that on October 5th, 2018 (pink bar), which corresponds to a working day (Friday) behaves like a weekend day. This situation may be explained by the fact that the 5th is a holiday in Portugal, consequently, a large part of the population does not work on that day. In 2019, the expected demand for October 5th was hidden because it matches to a weekend day. In the two subgraphs, all Saturdays show affluence greater than Sunday. As shown in Table 5.1, there is also a greater daily demand for entries in the period of 2019 compared to 2018.

Figure 5.2 also shows a count of entries but at the career level. In other words, it presents the count



(a) From October 2018

(b) From October 2019

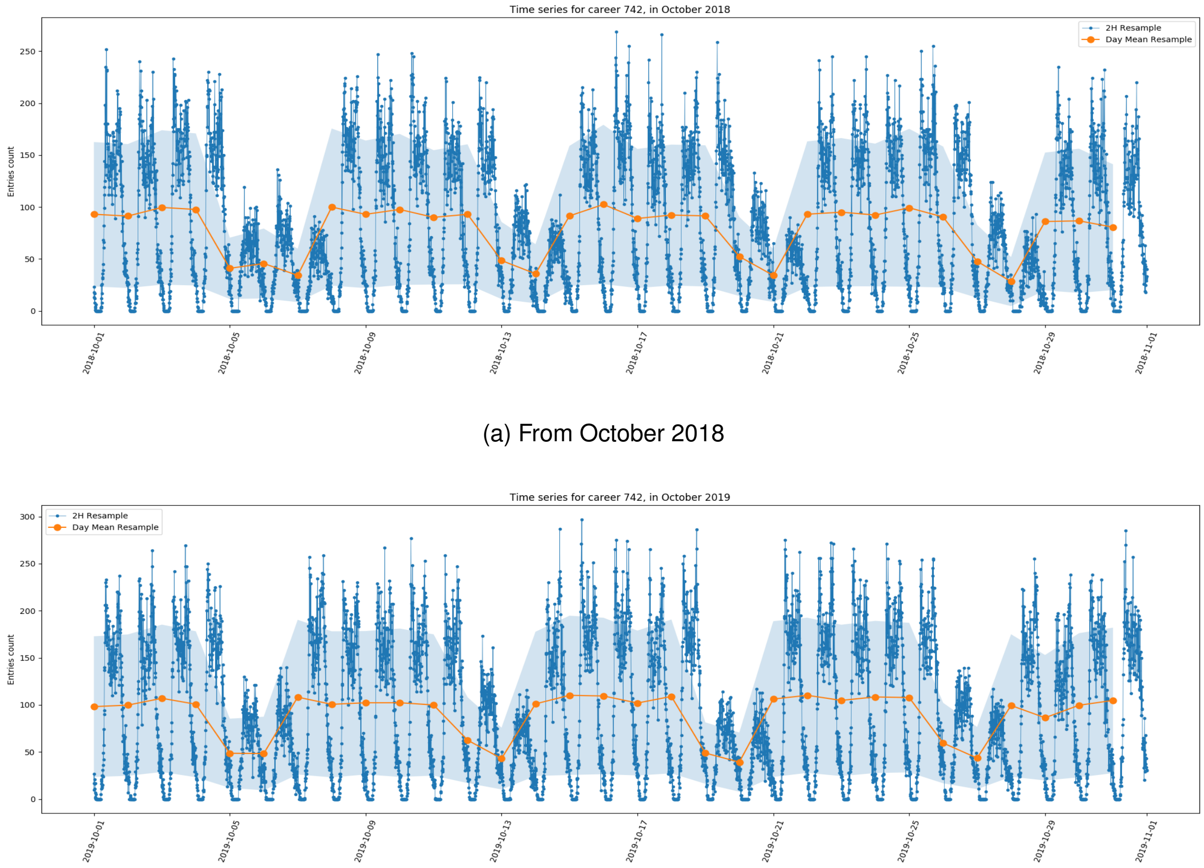
Figure 5.2: Count trips made in the route, of month October, in different years.

of entries in the bus network in ascending order.

Character of the route code	Description
Ends with an E	Network of trams and lifts, and works in daytime service
Starts with a 2	The bus network in the early morning hours includes routes that operate from Monday to Sunday, between 11:30 pm and 5:35 am.
Starts with a 7	Buses providing service during the day.
Ends with a B	Bus network that runs within neighborhoods, are generally shorter routes, and works in daytime service

Table 5.2: Description of routes through the first character of the code.

Figure 5.2 demonstrates that the neighbourhood routes (code end with a B) are those that have less influx of passenger entries and then there are the dawn routes (only integer character and starts with the number 2). The large influx of entries is mainly present in the daytime routes, especially route 742. In the top 6 and 12 of 2018 and in the top 7 and 15 of October 2019 appears the 15E and 28E respectively, which are trams. These trams are usually overloaded because they have an appealing route for tourists, as they pass through the historical locations, shopping and dining area of Lisbon, such as Belem, Chiado, Praça do Comércio, Jerónimos Monastery.



(a) From October 2018

(b) From October 2019

Figure 5.3: Time series representing the boarding's in the route 742, in the October month of 2018 and 2019.

Figure 5.3 shows two times series referring to the count of boarding in the most frequented route of the network (742) over October 2018 and 2019. The darker blue is exhibited a discrete-time series with a 15 min resample. The orange line represents the average daily boardings. Moreover, the light blue shade represents the variance of a day. Comparing and analyzing these subgraphs, we can conclude that: firstly, there is a similar pattern between the two periods, except for the 5th of October 2018. It corresponds to a holiday, reflecting on a working day; secondly, the average and the variation between working days are practically the same. On weekdays it is possible to verify that the average of a Saturday is almost always higher than Sunday.

5.2 Alighting stop inference results

5.2.1 Bus Model and validation results

This section discusses the results of the evaluation of the existing Bus Model for alighting stop inference, described in the Figure 4.4, from Chapter 4. To recapitulate, the algorithm is responsible for discovering the stops and departure times for each transaction of the dataset and can only visualize passengers' movements within the bus network. The results that will be shown below correspond to the period of October 2019. From this moment on, the results of the October 2018 data will no longer be discussed. October 2019 to carry out a comparative study with the results from the second proposed model that consolidates the above mentioned dataset with the dataset that contains the transactions that occurred in the metro system, in the same period.

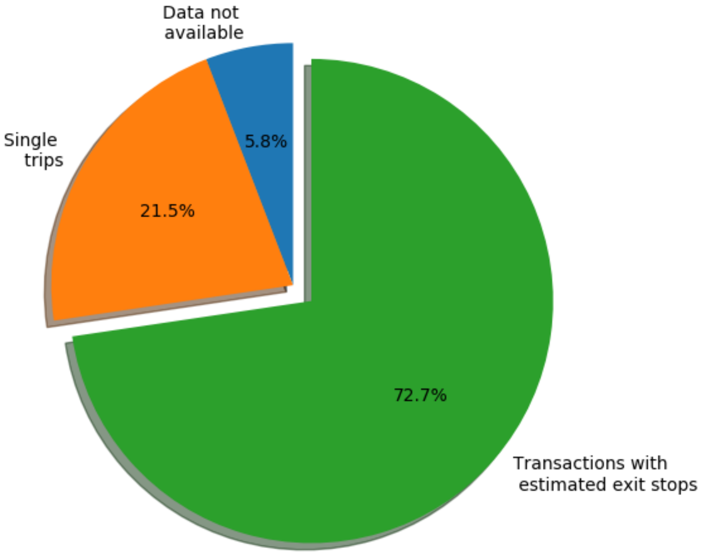


Figure 5.4: Percentages with and without success on alighting stop inference performed by bus model, on the data from October 2019.

Sector of pie chart 5.4	Count
Total segment trips trips	11 360 893
Single trips	2 441 080
Data not available	664 635
Transactions with estimated stops	8 255 178

Table 5.3: Absolute values of pie chart at the Figure 5.4.

Significant results to report, about the progress of the bus model, are: from the Table 5.3 and the pie chart, in the Figure 5.4, we can see that about 72% of the total transactions occurred in October 2019 (around 11 million), it was estimated the exit bus stop and its time. The remaining 27.3% of the transactions was not possible to estimate an exit stop since 21.5% of that percentage above corresponds to transactions in which the passenger only performed one transaction on the day (in these cases, the algorithm is not able to track the path of the passenger). The remaining 5.8% correspond to lack of synchronization between auxiliary tables data.

Distance interval (meters)	Count	Percentage	Accumulative Percentage
0 - 100	4 342 387	52.6	52.6
101 - 200	1 189 236	14.4	67
201 - 300	423 756	5.13	72.13
301 - 400	276 721	3.35	75.48
401 - 500	183 828	2.23	77.71
501 - 600	151 081	1.83	79.54
601 - 700	124 671	1.51	81.05
701 - 800	105 254	1.28	82.33
801 - 900	89 020	1.08	83.41
901 - 1000	83 367	1.01	84.42

Table 5.4: Distribution of travel segments regarding the distance traveled on foot, after a travel on a segment.

In transactions with a successful exit estimate (72%), are nominated with concept trip segments, because they now contain an entry and exit stop. Each journey segment is associated with a transfer distance, walked after arriving at the exit stop. Table 5.4 shows the travel segments' distribution against the walking distance mentioned earlier. Examining the results from the Table 5.4, it can be affirmed that 52% of the travel segments passengers, the exit stop was less than 100 meters from the next travel segment to be carried out. This fact is hugely positive and agrees with the assumption that the passenger tends to travel as little as possible on foot between transfers. Between 101 meters and 1000 meters of distance covered on foot, there is a gradual rise to 84% of travel segments. Finally, above 1000 meters, 15.58 % of the trip segments, it can be considered that the algorithm was not able to estimate the exit correctly. There may be another transport mode (boat, train, bike, subway) between the next visible segment (bus path).

Number single trips made by passenger	Count Passengers	Total trips in the month
1	257 405	257405
2-10	339 771	1 544 351
11-20 -	46 257	618 903
21-29	61 188	20 421

Table 5.5: Number of single trips performed by passengers during the month .

The results of Table 5.5 correspond to information about the orange section of the pie chart, that is, it transmits information about passengers who only transacted their card once during the day, at the CARRIS operator.

In the first line of the Table, we can see the number of passengers who made only one trip during the entire month: 257 405 passengers and 257 405 trips. The following Table lines reveal how many passengers made trips between the interval described in the column “Number single trips made by passenger” during the whole month.

107 445 passenger cards validated between 11 and 29 trips during the month (one per day). If the passenger makes a trip regularly over the month without turning back, it indicates that there is another mode of transport, for example, the passenger goes to his job on the bus and the return returns by metro.

We also consider calculating the number of card ids that made a single trip in one day and other days made more than one trip on the CARRIs operator, due to this suspicion. Furthermore, it is concluded that 383059 card ids traded one day a single trip and on others days validated more than one trip on the bus network. These referred passengers performed 1882372 single trips during the month. A multimodal model will trace the path of these card ids correctly if these passengers are using more than collective mode transport. So in the next subsection, it will be obtained an answer to this question.

5.2.2 Dual mode and validation results

This section discusses the results of the evaluation of the existing Dual mode Model for alighting stop inference,described in the Figure 4.6, from Chapter 4. The model can ultimately trace the passengers' path if it uses the bus and metro transport modes. Consequently estimates the bus stops for each transaction that takes place at the bus operator.

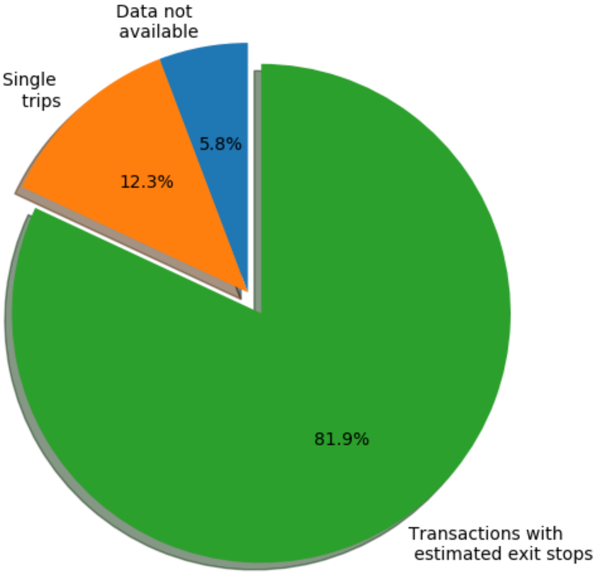


Figure 5.5: Percentages with and without success on alighting stop inference performed by dual mode model,on the data from October 2019.

Sector	Count
Total segment trips trips	11 360 893
Single trips	1 394 709
Data not available	654 709
Transactions with estimated stops	9 311 460

Table 5.6: Major results from dual model for alighting stop inference.

From Table 5.6 and Figure 5.5, the following conclusions. Firstly, the proposed new model manages to estimate exit stops for 81.9% of transactions, which corresponds to absolute values, around 9 million. Secondly, of the approximately 18% of transactions that were not inferred, 12.3 % are transactions that were not possible to trace a path, due to the fact that the user only transacts once in the AFC system during the day. Furthermore, the remaining 5.8 % correspond to asynchrony between supplementary data tables.

Transfer distance interval (meters)	Count	Percentage	Accumulative Percentage
0 - 100	4 770 192	51.23	51.23
101 - 200	1 570 248	16,86	68.09
201 - 300	523 334	5.62	73.71
301 - 400	355 913	3.35	77.06
401 - 500	283 516	3.82	77.71
501 - 600	203 003	2.18	80.88
601 - 700	144 406	1.55	82.43
701 - 800	116 251	1.25	83.68
801 - 900	94 132	1.01	84.69
901 - 1000	89 657	0.96	85.65

Table 5.7: Distribution of travel segments regarding the distance traveled on foot, after making a travel on a segment (dual mode model).

Although the algorithm could estimate outputs for more transactions, there is a similar inference behaviour in percentage terms over the number of successfully estimated transactions. Again 51% (1% less than the previous model) of the travel segments passengers, the boarding stop in the following segment is below 100 meters from the inferred exit stop. Between 101 meters and 1000 meters of walking distance, there is a gradual rise from 51% to 85% of travel segments (more 1% than the previous model). It is assumed that the remaining 14.35% may be incorrectly inferred because most passengers travelled short distances.

5.2.3 Comparison between models

In this subsection we compare the performance of the two proposed models (bus model and dual mode model). The most successful model with the best performance is the one that is able to infer the largest possible number of transactions with its respective exit stop. Of the transactions in which it was possible to estimate an exit, it is necessary to understand which of the models best fulfills the following assumption: passengers will start their next trip at or near the stop alighting location of their previous trip.

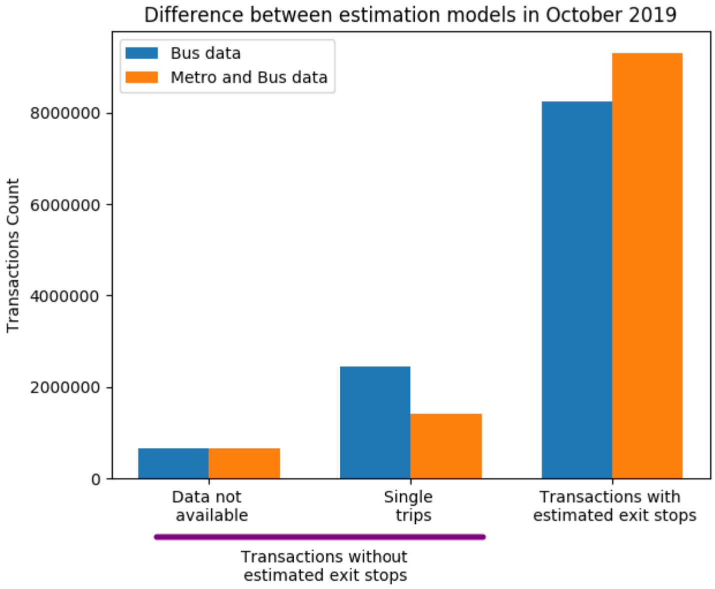


Figure 5.6: Difference between estimation models in October 2019.

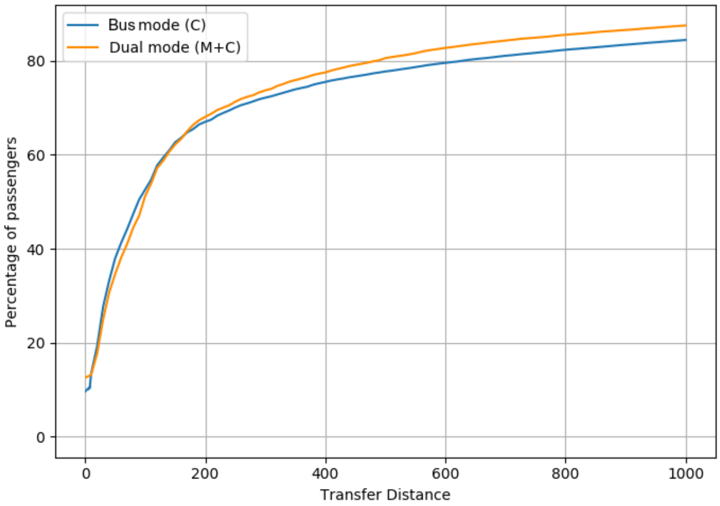


Figure 5.7: Accumulative percentage of passenger walking in its transfers.

Looking at figures 5.6 and 5.7 we can draw the following conclusions and lessons: In the bar plot of Figure 5.6, we can see in orange the results of the dual-mode model (the model that traces the route of the card ids within the metro and bus networks) and the results of the bus model (the model that traces the route only within the bus network) are in blue. It can be observed an interesting phenomenon: part of the transactions that were not estimated (that belong to the group of single trips) by the bus model, are now estimated by the dual-mode model. These transactions correspond to situations in which the passenger used the bus to travel city point and later used the metro to take off for another point, or the opposite, the passenger boarded the metro and later the bus. In short, the dual-mode model manages to infer the exit stop for a more significant number of transactions.

Figure 5.7 helps to understand which two models could better fulfil the above assumption. As already mentioned, each segment has a transfer associated with it. Therefore, this graph represents the cumulative percentage of travel segments against the distance travelled in the transfer. According to the assumption above, a passenger tends to transfer with the shortest possible distance, so the model with more travel segments with a certain distance of transfer will have higher accuracy.

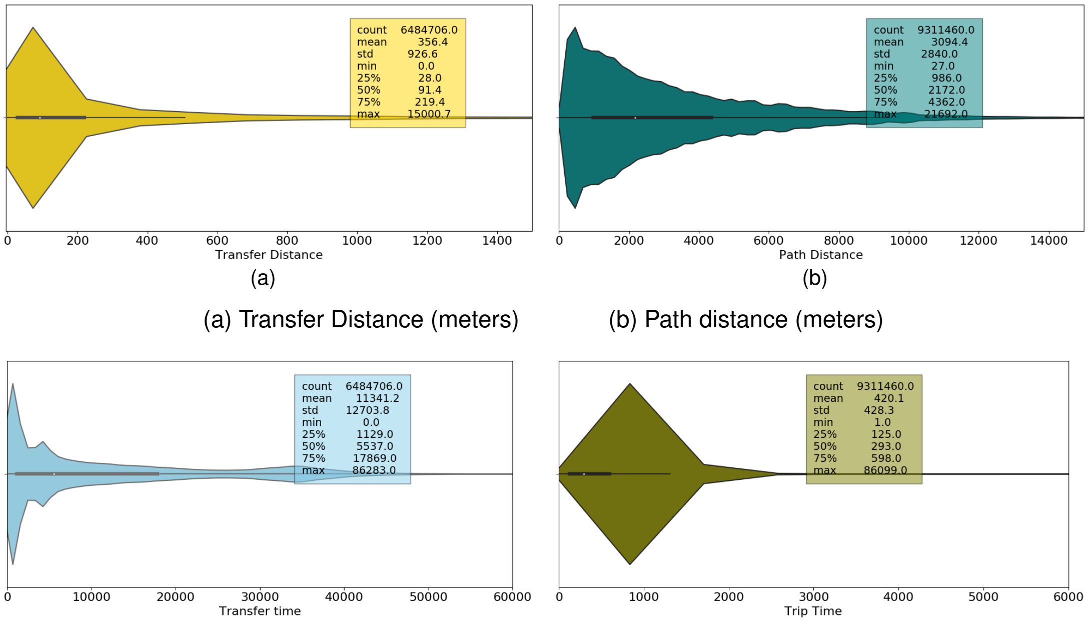
Up to 200 meters of distance walked, the number of segments tends to be the same, however after passing this limit, the dual-mode model tends to have more travel segments with the increase of the transfer distance. The dual-mode model converges more quickly to 80% of travel segments with a transfer distance below 500 meters while the bus model only reaches this percentage when it reaches 600 meters of transfer distance.

We can draw the following conclusion from these observations: multimodal models, that is, that use more than one transport operator to trace the path of a passenger, tend to be more accurate in estimating the exits to the only-control system (in the case of buses). The dual model estimated more travel segments and greater precision because the travel segments travelled in the metro became visible. If we integrate data from other modes of transport such as trains, bicycles, boats, the model will be more robust.

There will always be transactions in which it will not be possible to estimate a departure correctly because the next travel segment corresponds to non-traceable means of transport, such as car rides, taxi.

5.2.4 Dual mode data exploration

Since the dual-mode model achieves significant performance, we will proceed with the investigation with this model's results. This subsection will explore and discern the output generated by the dual-mode model. Next, a statistical description of the columns present in the dataset described in Table 4.3 will be made, such as the time and distance of transfers, and the route travelled on the buses.



(a)

(b)

(a) Transfer Distance (meters)

(b) Path distance (meters)

(c)

(d)

(c) Transfer Time (seconds)

(d) Trip Time (seconds)

Figure 5.8: Statistical assessment over the attributes.

In Figure 5.8, four plot violins are shown. This representation is similar to the box plot, represents numerical data, but additionally, it shows the probability density. Each violin is accompanied by a text box where some statistical metrics are presented, such as average, median, min, max, count, and quartiles.

Looking at the figure, we can describe the following general properties of all travel segments from October 2019, through the three significant affirmations.

Firstly, the violins in the left column refer to transfers between segments of travel. However, the sample under analysis does not observe all the supposed transfers. The distance between the last stop of the day and the day's first stop is not considered a transfer. Consequently, this non-transfers will be assessed on another analysis.

Secondly, we can affirm that the transfer distance is predominantly between 28 and 219 meters, with a median of 91 meters. Moreover, the time interval for a transfer is between 18 minutes and 4 hours, a median of 1 hour and 15 min. Quartiles are too far apart because some transfer times correspond to working time. It will only be possible to distinguish activity transfers with the help of the travel generation model.

Thirdly, the column on the left shows two violins related to the routes travelled by passengers. In this violin, the total sample was used, and we can say that the distance of a bus route is between 986 and 4362 meters, and the time is between 2 minutes and 9 minutes. The median distance and time correspond to 2172 meters and the time 4 minutes and 50 seconds, respectively, which means that, theoretically, the median speed of a bus would be 27 km / h (however the times when the bus is also stopped influences).

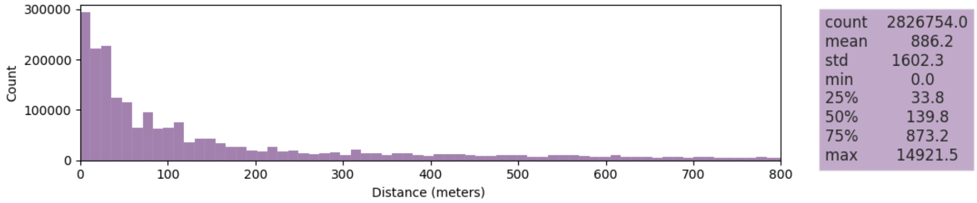


Figure 5.9: Distance distribution between the last stop of the day and the first stop of the day.

Figure 5.9 studies the distance between the last stop of the day and the first stop of the day, for all month. The study of this distance is critical to observe if passengers disembark close to the place where they boarded at the beginning of the day. According to statistics, this distance is between 33.8 and 873.2 meters and a median of 139.2. Once again, the quartiles are far apart due to outliers' presence, caused by the 15% exit stop of segments that may be incorrectly inferred.

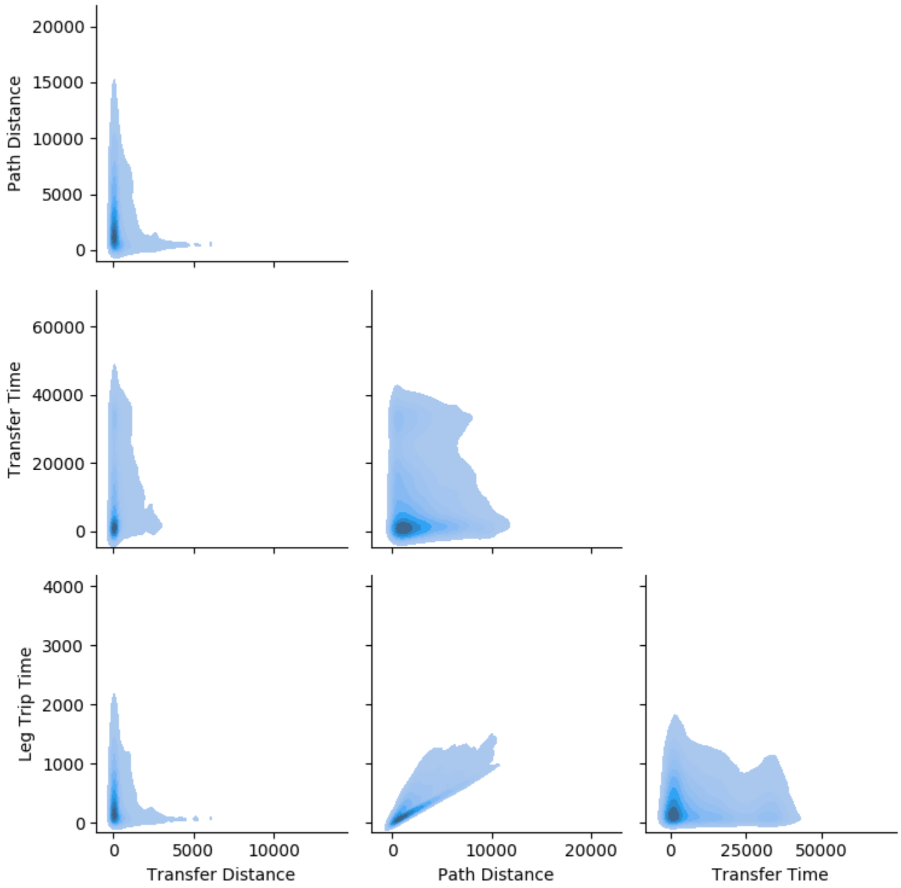
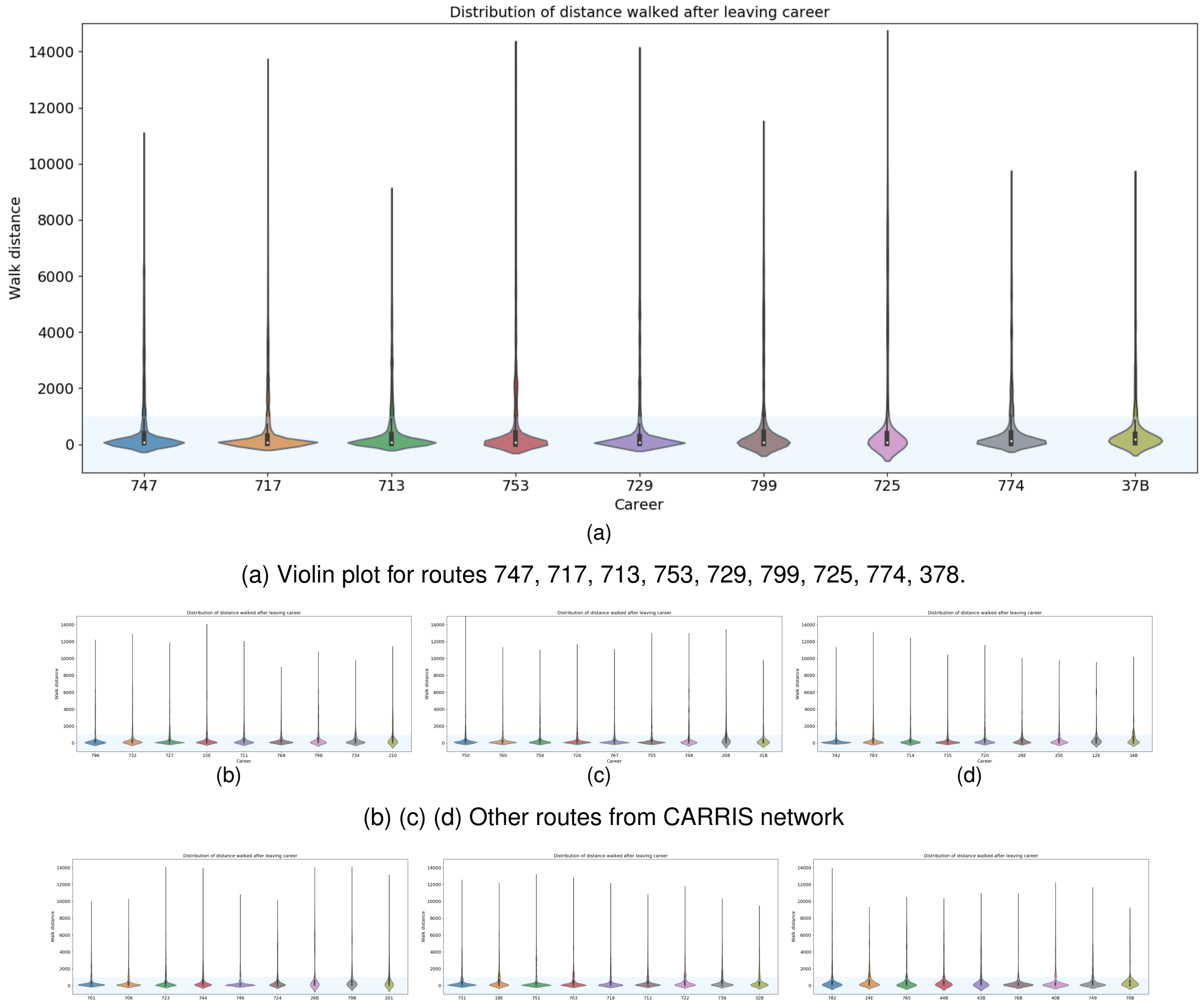


Figure 5.10: KDE relation between variables.

In Figure 5.10, it shows a kernel density estimate (KDE) plots to visualize the distribution of the observations of a dataset. Each plot is represented by a continuous probability density curve in two dimensions, giving a high insight into the shape of the data distribution, with much less variance. The observations related to this Figure were extracted from October 1, from 8 am to 8 pm. From the established relationships, we can draw the thre following conclusions:

Transfer distance does not grow proportionally with the variables trip time, transfer time, and path distance; the transfer time variable is sparsely distributed regarding the trip time and path distance variable; the variables path distance and trip time can be described by an increasing linear relationship.



(a)

(a) Violin plot for routes 747, 717, 713, 753, 729, 799, 725, 774, 378.

(b)

(c)

(d)

(b) (c) (d) Other routes from CARRIS network

(e)

(f)

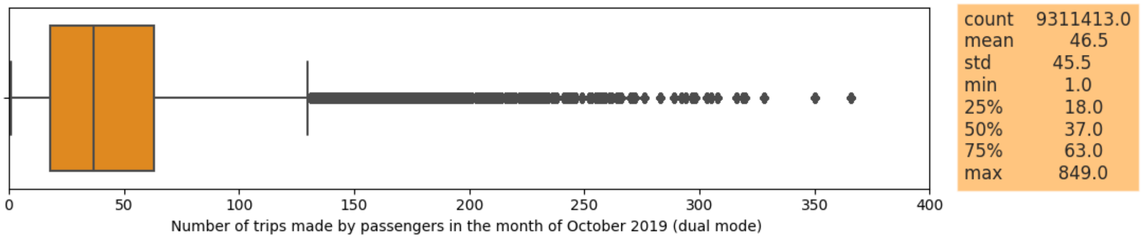
(g)

(e) (f) (g) Other routes from CARRIS network

Figure 5.11: Violin distribution for the attribute transfer/walking distance, for each route of the CARRIS network.

Figure 5.11 shows a different line of investigation regarding the transfer distance variable. Each sub-figure has several violins that show a transfer distance distribution after disembarking from a given route. We wanted to see any incorrect inferences on the dual-mode model for any particular route with this study. However, there were similarities between the violins. All have a greater distribution below 1000 meters and contain outliers (transfer distance deviating from the expectation).

Figure 5.12 is a box plot to depict the distribution of the number of trips performed by a user. The



```
count  9311413.0
mean        46.5
std         45.5
min          1.0
25%         18.0
50%         37.0
75%         63.0
max        849.0
```

Figure 5.12: Distribution of the number of trips made by passengers in the month of October 2019.

number of trips lies between 18 and 63 trips (two trips per day) per user. The sample also presents a significant number of outliers, actually, the maximum number of trips in the sample made by one of the passengers was 849 trips (28.3 trips).

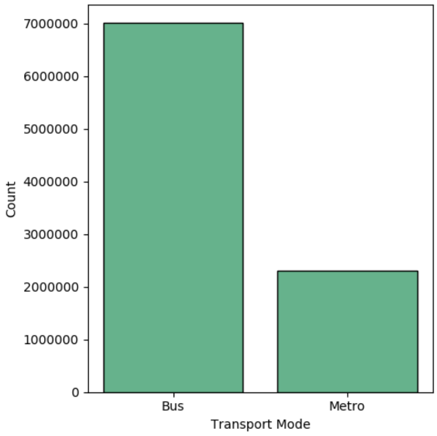


Figure 5.13: Percentage of trip segment where the next segment was metro or bus, in October 2019.

Figure 5.13 shows the proportion of the number of trip segments, in which the next segment was performed on the bus and the metro. It is concluded that about one-third of the travel segments, the next segment was performed in the metro. Of the total number of passengers, 64% passengers needed to use the metro at least once during the day to complete its journey.

Titles exploration

So far, when exploring the dataset, we envision characteristics and measures of passenger movements without describing them by groups. In the dataset we do not have access to characteristics such as gender, age, or profession, however we do have access to the card title that is associated with a fare, that indicates some clues of the age group. Profiling a user group is especially relevant for dedicating and improving carrier services to promote equity.

In October 2019, 85 fares titles were registered in the system. However, the next figures follow a line of investigation on the 15 most used titles.

The following Table 5.8 summarizes and describes the 15 titles. Some titles have derivations that correspond to titles for low-income groups (for example Social +). The title 4_18 /Sub23 has derivations depending on the valid travel area, but all are intended for students.

Title Name	Valid time period	Valid transport service	Target group age
Navegante metropolitano	Fixed monthly	CARRIS, METRO, others	Any person
Zapping	1 hour	CARRIS	Any person
Viagem CA/ML	1 hour	CARRIS and METRO	Any person
Ticket 24 Hours CA/ML	24 hours	CARRIS and METRO	Any person
Navegante Urbano 3.ª idade	30 days	CARRIS, Metro and CP (train)	Elderly +65
Familia Metropolitano	Fixed monthly	CARRIS, METRO, others	Any person
Navegante 65+	Fixed monthly	CARRIS, METRO, others	Elderly +65
Navegante Lisboa	Fixed monthly	CARRIS METRO CP FERTAGUS (boat)	Any person
4_18/Sub23	Fixed monthly	CARRIS, METRO, others	Students from 5th grade to higher education, aged 23 or less

Table 5.8: Titles card description.

The next Figure 5.14, describes the function and density of the boarding for each of the three titles chosen in analysis, during the period of 1 October 2019. The following titles were chosen because they correspond to different age groups: 4_18/Sub23 is used by a target group age between 11 and 23 years old e; Navegante +65 is used by the elderly over 65 years old; and the Navegante Metropolitano can be any age group, except those previously mentioned.

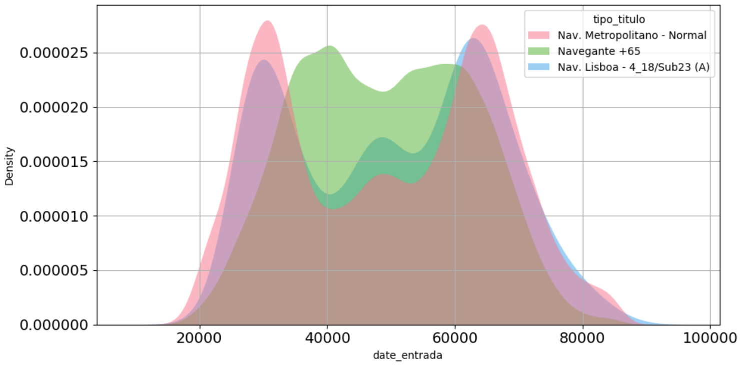


Figure 5.14: Density of boarding by card title.

Observing the density of passengers entering during the day, we see that title 4_18/sub23 has the same density function as Navegante Metropolitano's title. There is a peak density of boarding into the network at around 8 am and 6 pm, for both titles. We can suppose that it corresponds to students going to and from teaching institutions and returning home. Furthermore, the entries in the other title may mean travelling to the workplace and returning home. The title Navegante +65, directed to the elderly, presents a higher density of entries during the period from 11 am to 4 pm, without relevant peaks, and the function curve avoids the peaks of the other mentioned titles.

The graph that represents the density of exits in the network is not shown in the study because it shows similarity with Figure 5.14, but with a time lag of minutes.

The next Figures 5.15, 5.16, 5.17 explore the following characteristics of each 15 most used titles: transfer distance, number of trips made during the month, and distance in a travel segment.

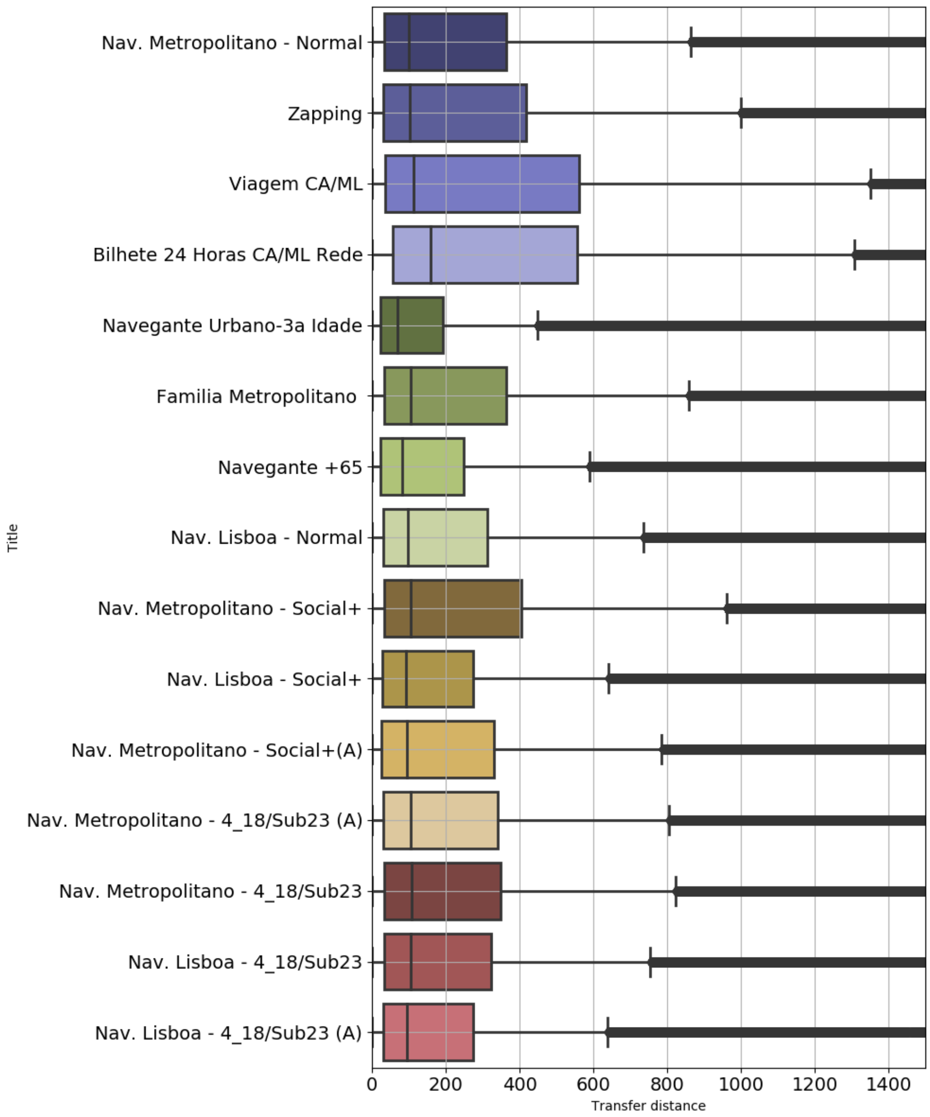


Figure 5.15: Box plot of attribute transfer distance for each most used 15 title cards, in October 2019.

From Figure 5.15, we can infer the most interesting conclusions. First, the median value for all titles is below 200 meters. , occasional titles with less time of validation use such as zapping (1hour), ticket 24h CA / ML (24 hours), Viagem CA / ML (1hour), have a greater dispersion of the transfer distance than the titles with more prolonged time validation. Conclusively, the title Navegante Urbano 3ª Idade, oriented to the age group over 65, has a dispersion of the transfer distance lower than all other titles. It is presumed that the elderly tend to walk shorter distances.

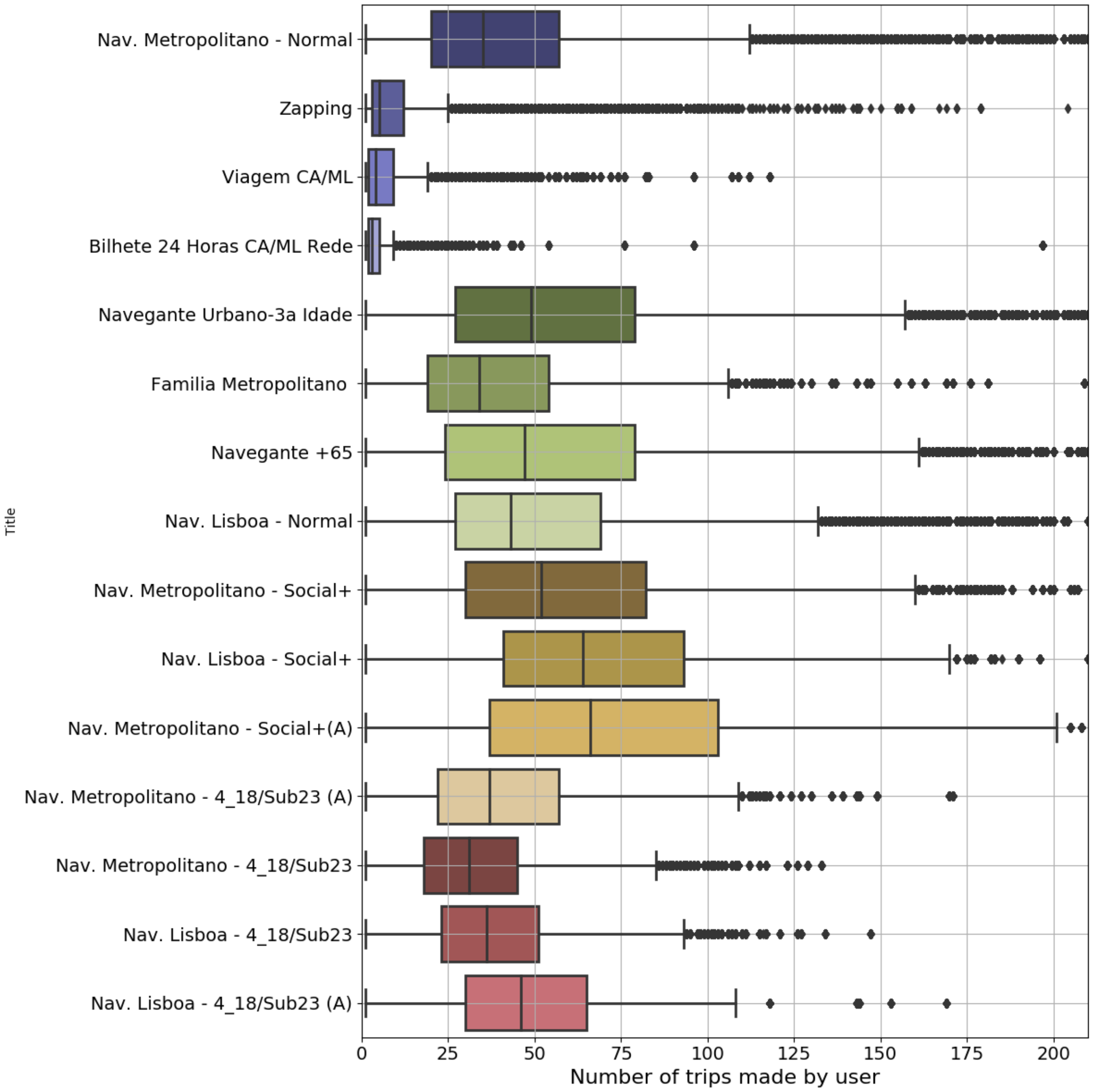


Figure 5.16: Box plot of attribute number of user's trip for each most used 15 title cards, in October 2019.

From Figure 5.16, we can infer the most interesting conclusions about the variable number of trips made in all month, by the users: as expected, occasional titles such as zapping (1hour), ticket 24h CA / ML (24 hours), Viagem CA / ML (1hour), are used to make few trips during the month. Titles associated with Social+ terminology, targeted for passengers with low wage income, tend to make more trips during the month, regarding the other titles.

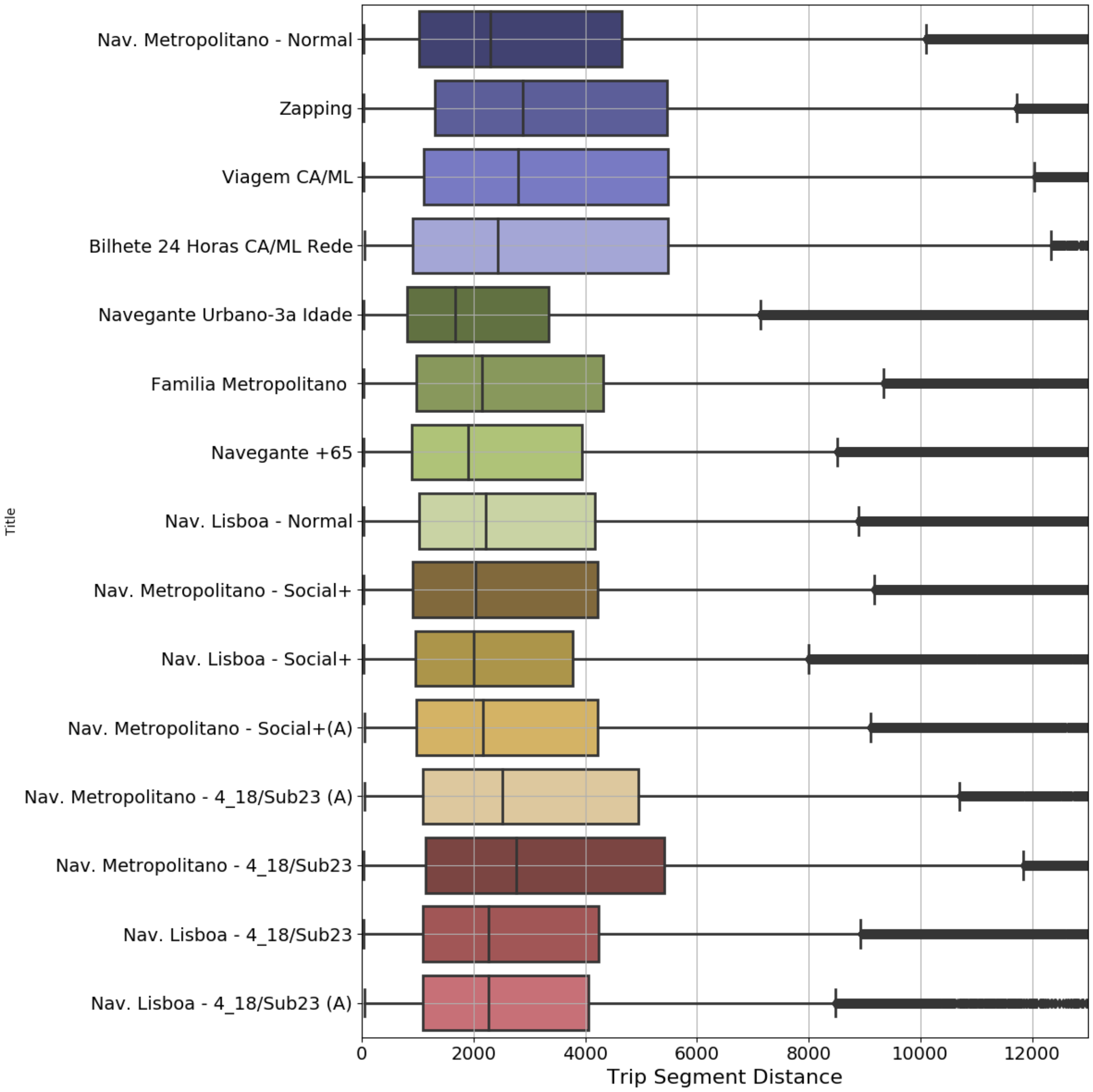


Figure 5.17: Box plot of attribute segment trip distance for each most used 15 title cards, in October 2019.

Figure 5.17 observes the distances travelled in a travel segment for each title and verifies that there is no group highlighted. The median of all titles is between 2km and 3km. We can affirm that Navegante Urbano 3rd Idade's title, oriented to the age group over 65, converges for a shorter segment distance than the other titles. In other words, the elderly also tend not to travel long distances.

5.3 Trip Generation for commuting travel analysis

In the section Trip Generation for commuting travel analysis, a solution is proposed to derive travel segments to journeys. The journey is a movement from an origin location to a destination, and the designation commuting travel implies an outward journey (home to school) and return journey (school to home). The generation of these trips and a subsequent study of the results, allows answering questions such as: how many trips are generated from this area? Or how many passengers have travelled to this point and from where?

Knowing the generating points and attractors of movement allows transport operators to improve their network to offer an appropriate service at any time of the day.

The proposed solution for generating commutative trips presents some parameterizations to determine the commute trips, so the following enumeration recapitulates the parameters with the chosen values:

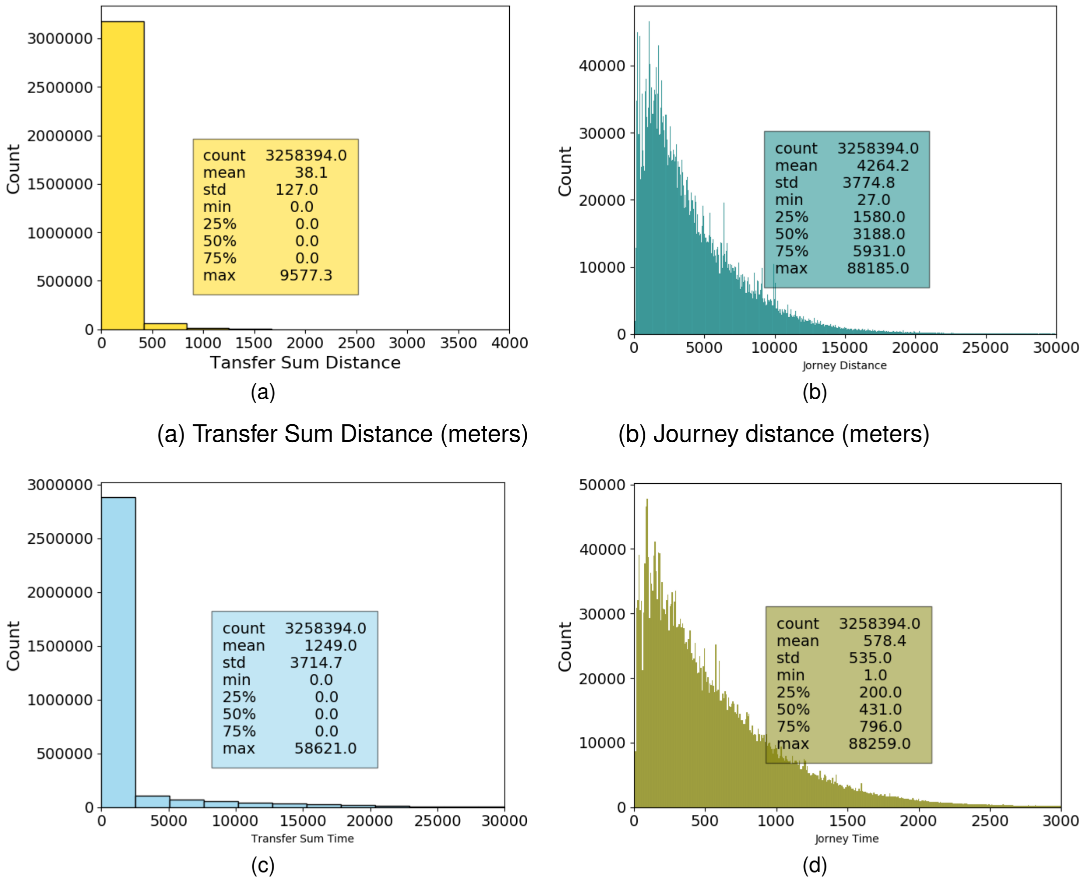
1. A passenger who does not have a minimum of trips cannot contain commutative trips. The minimum threshold is chosen according to the value of the 1st quartile of the box plot in Figure z, which is 18 minimum trips per card id.
2. The distance between travel segments must be less than stipulated. This limit avoids the presence of incorrectly inferred segments. Therefore, the maximum distance between segments corresponds to 1000 meters.
3. The distance between the last stop of the day and the day's first stop must be less than the stipulated, so the chosen value is 700 meters. In this last parameter, he decided to be more restricted, because as we intend commutative trips, we assume that the passenger will embark and disembark at the point closest to home.
4. The duration of an activity should be more than an hour.

The values of the parameters mentioned above were adopted so that the algorithm is relaxed and extracts the maximum commutative trips. Subsequently, in future investigations, the parameters of this model may be changed in order to avoid overfitting or underfitting.

As the dual-mode model presented more trustworthy inferences results, these travel segment results were used to derive commutative trips. In short, 3 258 769 journeys were generated from the commutative travel generation process. Moreover, 160 198 distinct card IDs were identified, that is, 24% of passengers coming from the input dataset (670 303 card ids).

Commuting travel features exploration

The same way that the exploration and endogenous analysis was executed on the output from the dual-mode model, the results from commutative trips model, will be analyzed. The following dataset properties will be interpreted: the sum of time and distance from transfers, the sum of distance and time from passenger journeys, time spent on an activity, titles present in the data, proportion of metro usage, number of transfers within a journey, among others.



(a) Transfer Sum Distance (meters)

(b) Journey distance (meters)

(c) Transfer Sum Time (seconds)

(d) Journey Time (seconds)

Figure 5.18: Variables distance and time distribution for transfers and journeys.

Figure 5.39 shows the distribution for the following features of the dataset. Through the subfigures, we can affirm the conclusions about the variables under analysis: Subfigure (a) indicates that 75 % of the journeys have in their sum of transfer distances equal to 0 meters, which means that there are no transfers in those journeys. There are still some outliers since the maximum value is 9 km. Subfigure (b) shows the distribution of the distances covered during the journeys (inside the bus) and shows a gradual decrease in journeys as the distance increases. According to the statistical data, this distance referred lies between 1.5km and 5.9 km with the median at 3km. Subfigure (c) shows the distribution of the sum of time spent on transfers, and one more time presents a behaviour similar to subfigure (a). As most journeys are trips without transfers, the time spent will also be zero. The average value corresponds to 1249 seconds, that is, 20 minutes. Subfigure (d) shows the distribution of the sum of time spent travelling on journeys (inside the bus only). And we can see that as travel time increases, the number of journeys decreases. Moreover, this behaviour complies with subfigure (b), which is to be expected. The time spent on a trip lies between 200 seconds and 796 seconds, which means between 3 minutes and 13 minutes. The median value corresponds to about 7 minutes.

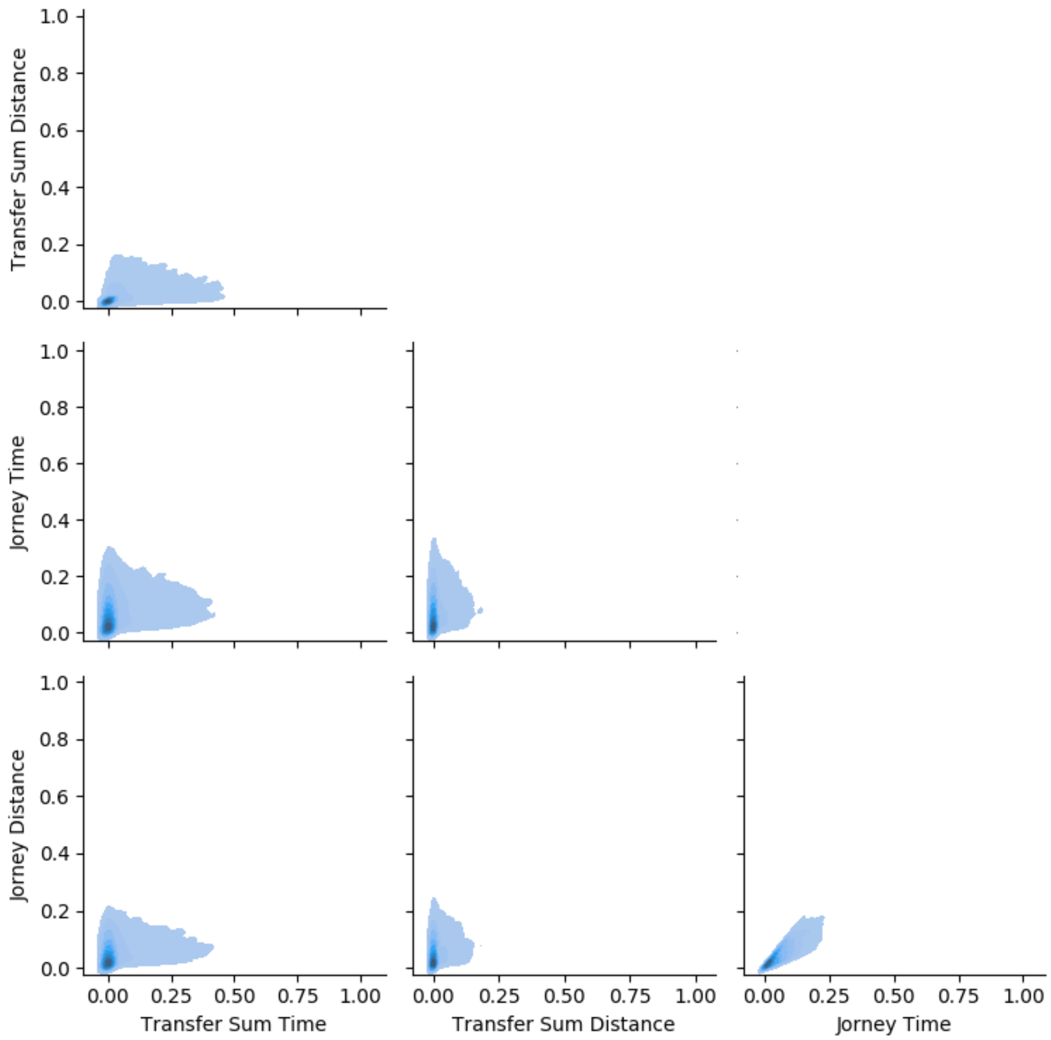


Figure 5.19: Density relation between variables.

In Figure 5.19 shows the kernel density estimate (KDE) again, but for observation from the commutative trips dataset. Each plot gives a high insight into the shape of the data distribution, with much less variance, between two dimensions. The observations related to this Figure were extracted from October 1, from 8 am to 8 pm, and each plot's scales are normalized between the values 0 and 1. The relationships established are practically the same as those in Figure 5.12, however, in this case, the time and distances in the transfers and journeys correspond to the sum of one or more segments.

It is noted that there is a greater centrality and proportionality between the dimensions, while in Figure 5.12, the relations showed distant dispersion. The reason hidden behind these results is due to the parameters that restrict the output data. Consequently, it avoids the presence of distances and times outliers. Basically, if there is a correct parameterization, the data becomes coherent.

During the generation of commutative movements, it was also possible to calculate the passenger's activity time at his destination. Therefore, Figure 5.20 shows the distribution and statistics associated with the activity time (time spent after reaching the attracting destination). As you can see in the sta-

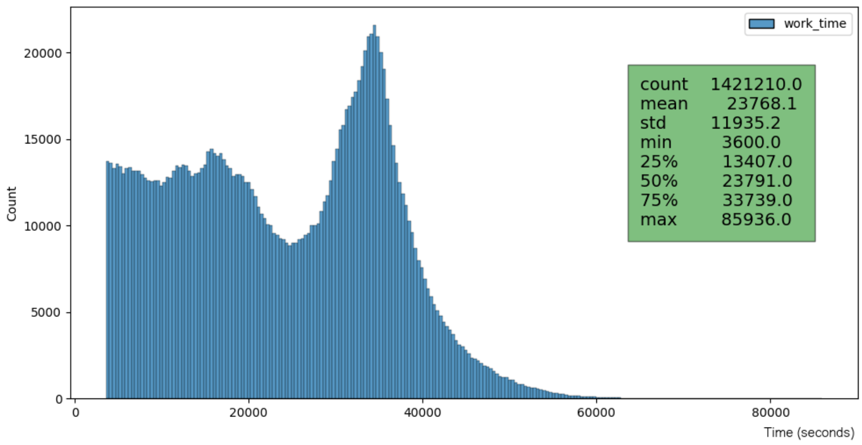


Figure 5.20: Activity time (seconds) distribution.

tistical data, the minimum time was 3600 seconds, that is one hour because it was the minimum time imposed on the algorithm to perform an activity. The activity time lies between 3 hours, 43 minutes and 9 hours 22 minutes. On the other hand, the median corresponds to 6 hours and 36 minutes, which is expected for most workers or students. In the distribution figure, we can see a peak of around 9 hours of activity and then the passenger distribution suddenly decreases.

Number of transfers	Count of journeys	Percentage (%)
0	2 463 307	75.5
1	561 212	17,22
2	150 696	4,62
3	52 066	1,6
equal and more than 4	31488	0,97

Table 5.9: Counting of journeys by number of transfers performed during he journey.

Table 5.9 shows the percentage and the absolute value of how many journeys made a certain number of transfers. As can be expected, we observed a higher number of trips without transfers (75%) compared to trips with transfers.

Figure 5.21 shows the density curve for each number of transfers and the effect it has on the distance travelled by the passenger, on the bus. Looking at the peaks of each curve, we can conclude that: when making 0 transfers, the most of the passenger travels around 1500 meters; for one transfer, the most of the passenger travels about 2500 meters; for two transfer, about 5000 meters are covered; for three transfers around 7000 meters;

Figure 5.22 demonstrates a bar plot, where the first bar is the number of journeys that in their path used only buses and the second bar is the number of journeys that used buses but also metro within the path. We observed that on commutative trips, about a third of the journeys had a travel segment on the metro during its journey. The dataset containing these described journeys contains 160198 different passengers, and 70.7% (113224 card ids) used the meter at least once during the month. Therefore,

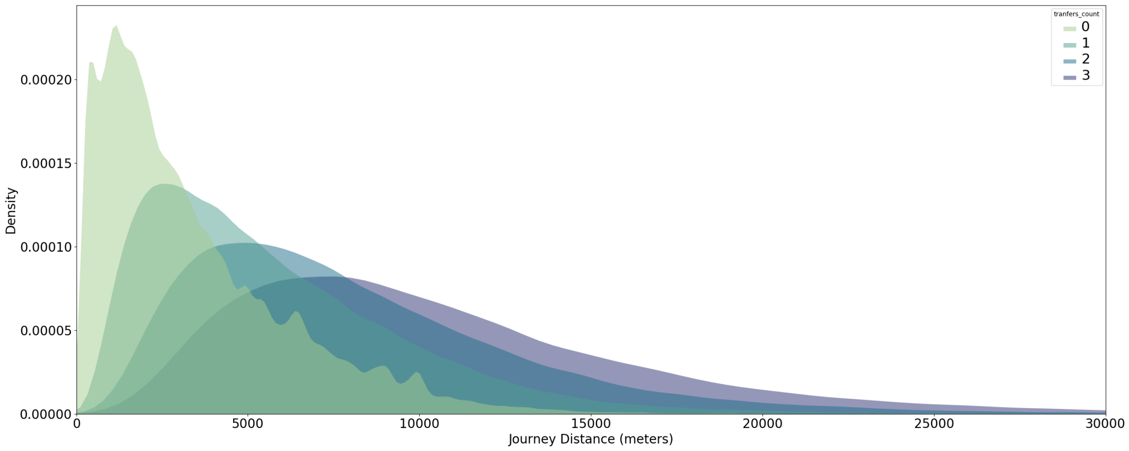


Figure 5.21: Travel density curve for each number of transfers in relation to the distance traveled on the bus.

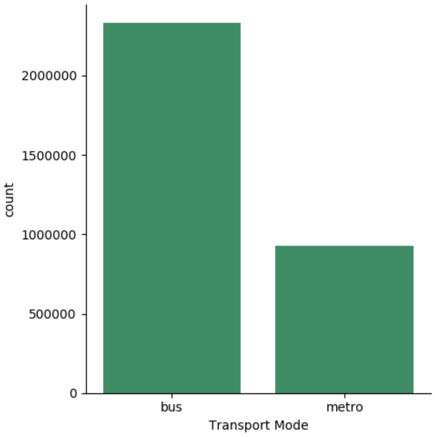


Figure 5.22: Percentage of journeys within metro trip segment or just bus segments, in October 2019.

restricting the sample to commutative trips only, passengers' percentage, using the metro at least once during the month, increased.

Sankey representation for analysis

The Sankey representation is a strong ally in the representation of flows. In summary, the node represents the location, and the connecting links are the amount of flow between these nodes. In the next figures, the nodes and links are the numbers of passengers travelling between points on the CARRIS transport network. This representation is not implemented in the visualization tool. However, for future works, it should be considered. The only python package that presents an interactive and appealing visualization belongs to the plotlly package. This feature is appealing because the information related to the elements (nodes and link) only appears when hovering, making reading comprehensible.

Figure 5.23 focuses on showing the top 10 journeys in which the entry route is not the same as the exit route, which meant that, within the journey, there was at least one transfer. During October 2019,

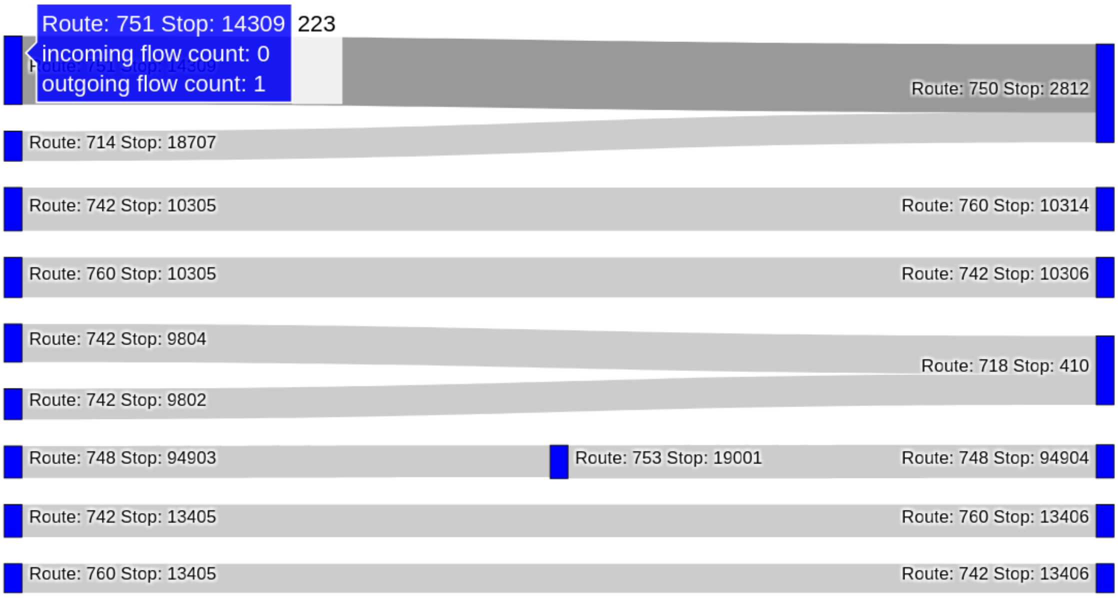


Figure 5.23: Sankey representation for journey link with transfers, top 10 most frequent.

223 journeys were carried out between stop 14309 of route 751 and stop 2812 of route 750, and this connection was the most frequent among the top 10. The Figure shows an intermediate point at a stop in row 753, which means, that exits trips from row 748 to 753 and then from 753 to 748. It should be noted that these two links do not require that the trips are from the same passengers.

In short, these ten connections indicate that passengers must at least make a transfer between origin and destination, and therefore it is a starting point for CARRIS to review these connections and analyze whether it is possible and worthwhile to make direct connections.

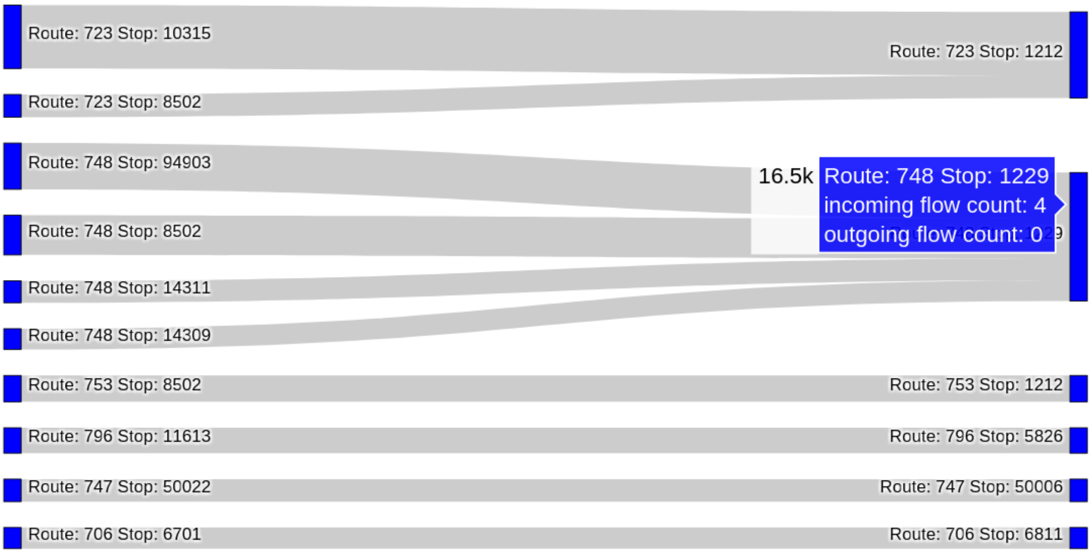


Figure 5.24: Sankey representation for bus journeys for the purpose of later traveling on the metro, top 10 most frequent.

The Sankey representation in Figure 5.24 shows the top 10 connections in which passengers travelled by metro after making one of the connections (bus journey) shown in the figure. Mentioning the main results, we can say that stop 1229 of row 748 receives four connections from others the stops of the same route, 16500 journeys. It should be emphasized that this destination stop is located in ”Marques de Pombal”, and in that same place there is a metro connection through two lines (yellow and blue), so we can assume that passengers go to that stop to later make a trip on the yellow or blue line. If in the future, CARRIS and METRO join forces to consolidate the service network for the population benefit, then these multimodal connections can be revised to create direct bus connections to not overload the metro service.

Titles exploration in commutative trips

Finally, to finish the analysis of the properties of commutative trips, we analyze what kind of titles are present in the journeys and how they are distributed in the boarding during the day.

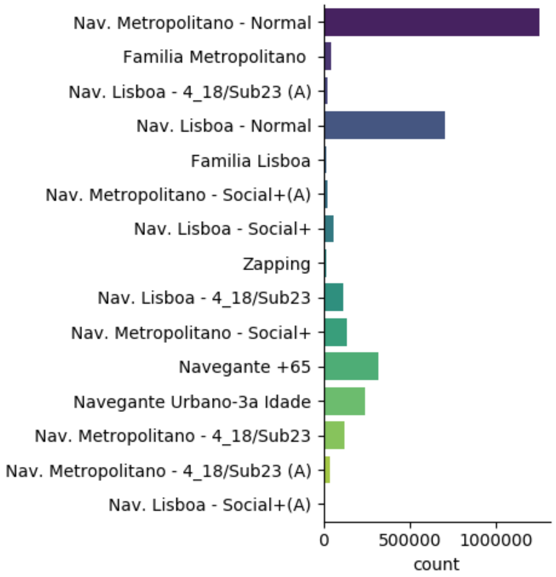


Figure 5.25: Titles top 15.

Figure 5.25 shows the number of journeys made by each title during October 2019. We can see that the ”Navegante Metropolinato Normal” (which can be used by anyone except by the elderly over 65 and young people under 23 years), is the most frequently used. In second place is the title Navegante Lisboa Normal, where the difference against the previous one, is that the service area is restricted to the Municipality of Lisbon, while the other comprises 18 municipalities divided by the two sides of the Tejo River (district of Lisbon and Setubal). The third most frequent one is the Navegante +65, intended for the public of over 65 years old.

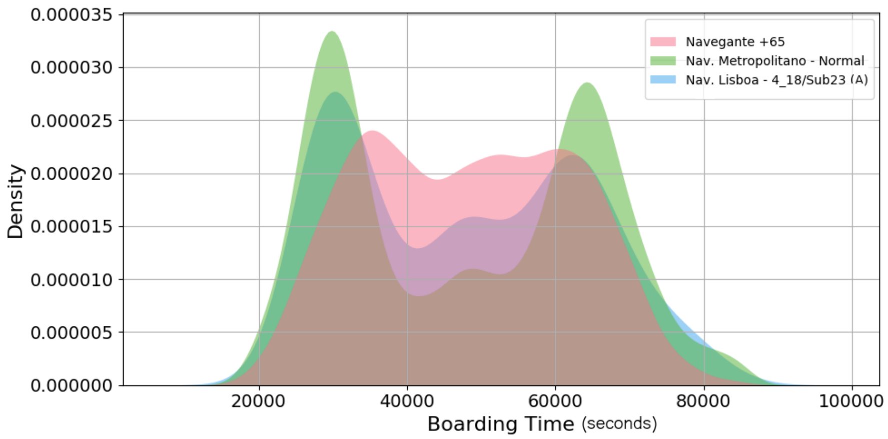


Figure 5.26: Density of boarding by card title, for commuting trips.

The titles shown in Figure 5.25, ”Navegante Metropolitano”, ”Navegante Lisboa 4_18/Sub23”, ”Navegante +65” were chosen to analyze the adult, young age groups and elderly, respectively, within one working day. For the same reasons explained in the Figure 5.16, there is a peak density of boarding into the network at around 8 am and 6 pm, for the titles ”Navegante Metropolitano”, ”Navegante Lisboa 4_18/Sub23”. Again the title Navegante +65, directed to the elderly, presents a higher density of entries during the period from 11 am to 4 pm, the curve density function avoids the peaks of the other mentioned titles.

5.4 Origin Destination Matrices Description

This section presents and investigates the final product that dynamically visualizes passengers' flow and other metrics in the network, which is called origin-destination matrices.

Thanks to this new visualization implemented in the ILU project's scope, the CARRIS transporter will trace the path of the passengers; find out how many passengers are heading to a specific location at a particular time; check for bus overloads; among other applications.

The subsequent investigation aims to show the different applications of the matrix parameterization. Basically, a matrix inventory will be carried out with specific parameters to show this tool's potential, and the novelty factor to the transport network's planning.

The visualization tool presents several filters to parameterize the matrices, but only a few of these filters will be explored. On the other hand, we can view boarding and exits in the entire network, but it will not be possible to show such visualizations due to the greatness of existing stops, therefore, routes will be chosen for entry and exit that contain few stops to facilitate the visualization and understanding.

In short, the following factors will be explored:

- Time-based Matrices vs Routine Matrices

- Time Horizon Exploration
- Granularity horizon Exploration
- Titles horizon Exploration
- Transport mode percentage
- Other metrics

It is necessary to emphasize that a matrix is multivariate and therefore has several variables to depend on. Therefore, in the above categories, there will be relationships, however we focus on one specific analysis related to the category.

5.4.1 Time-based Matrices vs Routine Matrices

The first category that is spelt out in this work is the type of travels present in the origin-destination matrices. As I mentioned, there are two types of matrices: time-based matrices and routine matrices. Time-based matrices is a microscopic analysis, because it presents all stages of the trip, while routine -matrices present a macroscopic view of the trips, only the origin and destination (ignoring the segments of trips that may exist). Both are important because they have different applicabilities. The first can be used for forecasting, while the second type can be used for route planning and review.

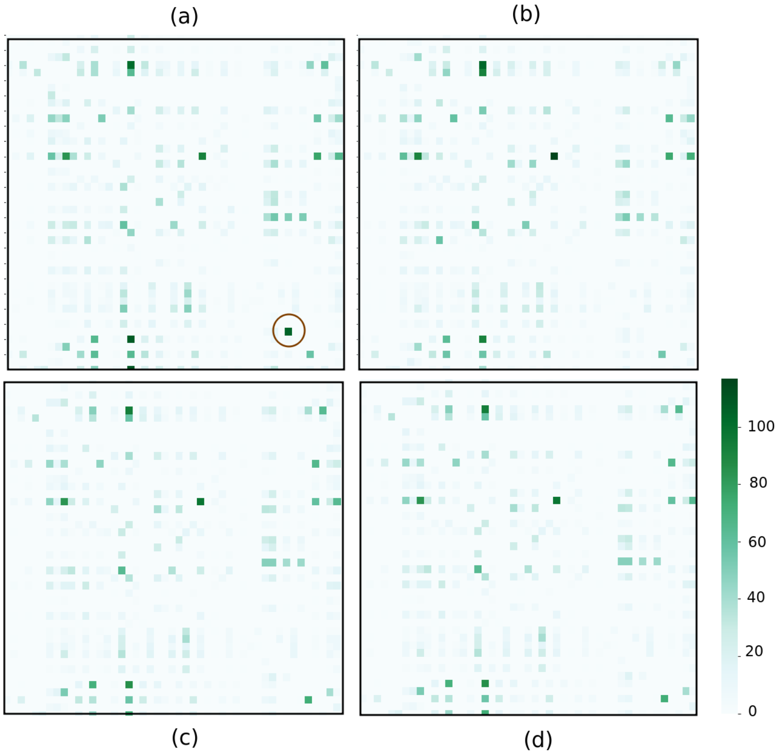


Figure 5.27: Time-based matrices for each Monday of October 2019 with outliers, granularity stop, boarding's in 756 route.

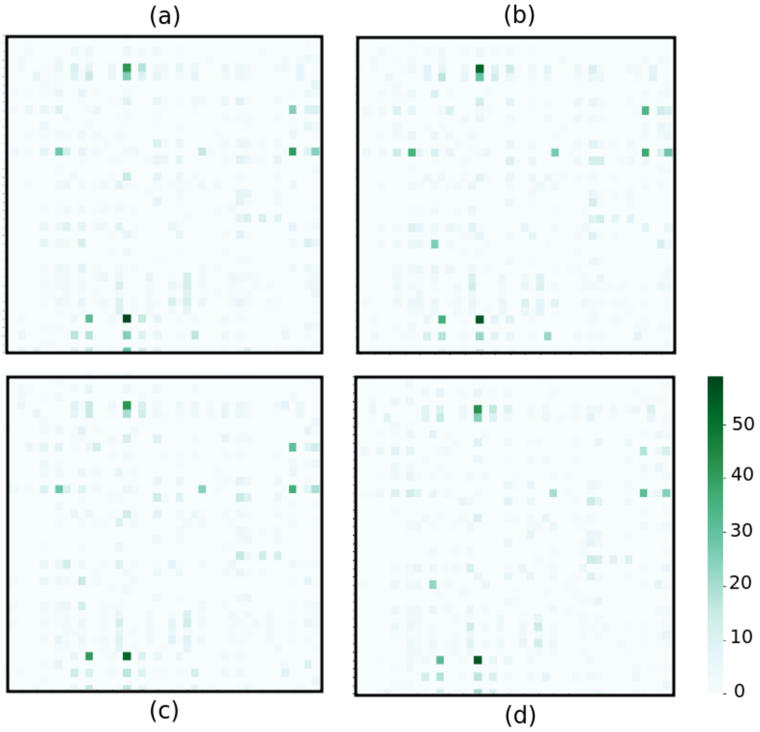


Figure 5.28: Routine-based matrices for each Monday of October 2019, granularity stop, boarding's in 756 route.

The source dataset for the time-based matrices corresponds to the output generated by the dual-mode model, while the source dataset of the routine matrices will be the dataset output of the commutative travel generation model. Consequently, routine matrices will have less demand against the other typology, since they only present commutative trips. In fact, this consequence is demonstrated by the Figures under study 5.27 and 5.28.

Figures 5.27 and 5.28 show the flow of passengers between stops on route 756, and each sub-figure within the Figure represents a different day, in this case, they represent the 7th (a), 14th (b), 21st (c) and 28th (d) of October (every Monday of the month). The first figure's matrices correspond to time-based matrices while the matrices in the second Figure correspond to routine-based matrices. Comparing the subfigures to each other, and comparing the two figures, we can state the following:

- Between the matrices of each Monday there is a similarity, except in Figure 5.27. (a) A cell with a dark green hue could be an outlier (cell surrounded by a circle). It will be investigated later.

- The scale of the first Figure ranges from 0 to 100, while the second ranges from 0 to 50, indicating that the maximum value of trips in the first matrix, about 50 per cent, are commutative trips.

- In the two figures 5.27 and 5.28 there are two cells that present the greatest demand on every Monday. These cells have different origins, however they see each other at the same place that corresponds to the ”Campo Pequeno” stop. There is a metro station close to this stop, so the high demand generated may be due to this factor.

We follow the investigation line regarding the possible outlier present in the Figure 5.27. Therefore, first, we calculate the residues between the four matrices, as shown in Figure 5.29.

In most cells the residue is between the interval between -20 and 20, in other words, the passenger flow in this route is practically similar on any Monday of October. As it was suspected, there is a non-standard cell in which the residue compared to the remaining matrices is much higher, around 100 trips. The next step was to investigate which travel segments are present between these two locations.

Firstly, it is suspicious that the origin and destination points are two adjacent stops on the route. Secondly, the walking distance average after travelling the segments present in the cell corresponds to 3k, that is, after travelling between the origin and destination of the cell under study, passengers walked 3km, which is not normal. Thirdly, on this route under study, one of the stops is very close to a train station. All of these statements lead us to believe that the algorithm is incorrectly estimating these segments. So, this little investigation serves to underline the importance of consolidation between various sources of transport to trace passengers’ path.

Another aspect of being highlighted is the trade off of filtering incorrect segments or not to do it. We can filter the incorrectly inferred segments through, the property walking/transfer distance associated with each segment.

In other words, if there are segments where the passenger, after exit, had to walk more than a certain number of meters, the segments may be poorly inferred.

However, this approach ignores the problem and only contributes to not transmitting the real number of passengers on the network. We chose not to filter the time-based matrices because each computed matrix has an overall assessment calculated by the transfer walking distance property that allows the assessment of the matrix’s reliability.

If the assessment is low, we can investigate microscopically which cells are contributing to it. If the matrix has a positive evaluation, then the matrix is correctly translating the absolute value of the passenger flow. In routine matrices, the poorly inferred segments are filtered because these matrices contribute to network planning.

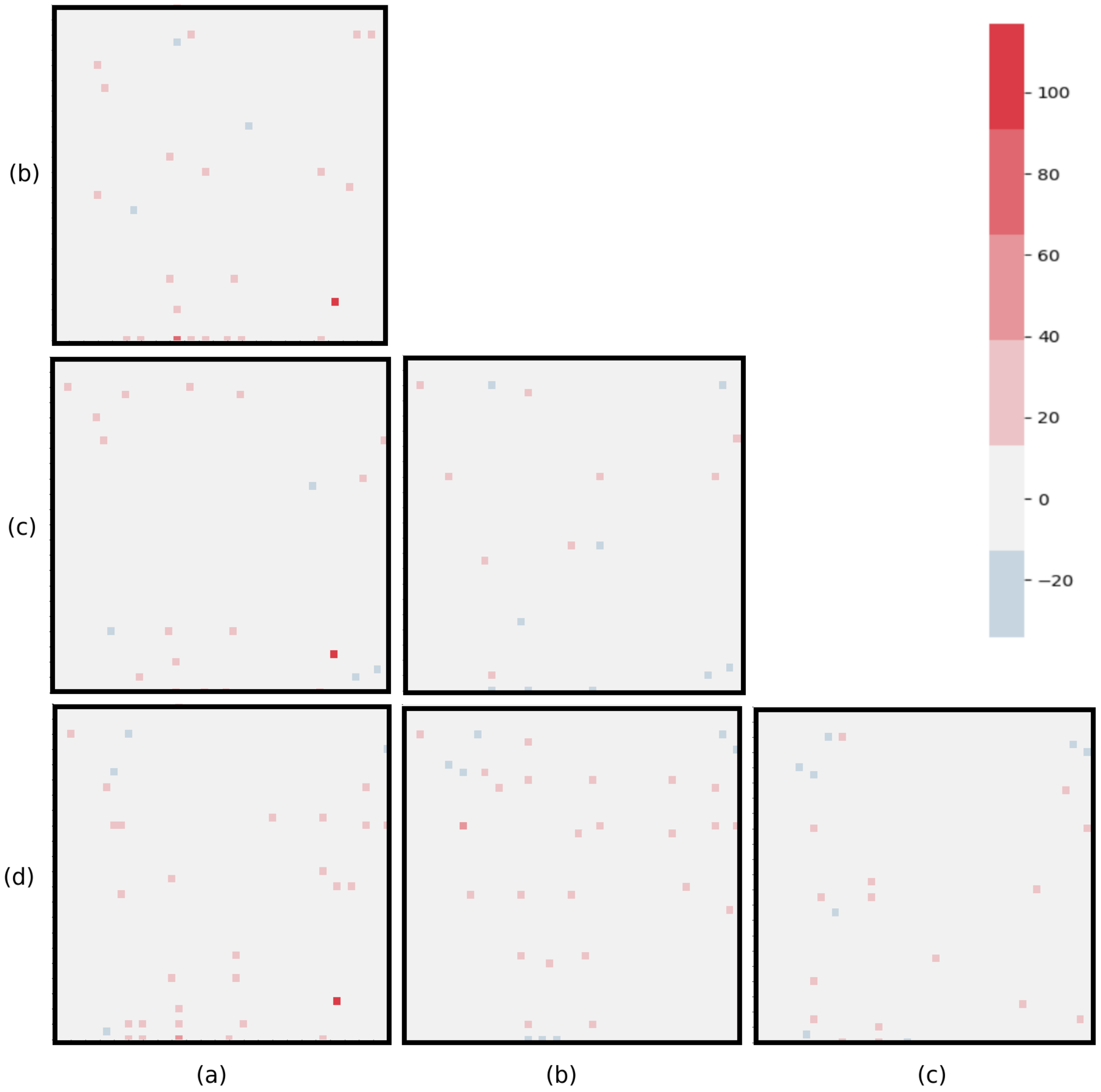


Figure 5.29: Residual difference between each time-based matrix from Figure 5.27

5.4.2 Calendar Horizon Exploration

The next topic to be addressed is the variation in passengers' flow on commuting trips during different days of the week. Therefore, the matrix matrices that will be shown corresponding to the routine-matrices typology. The next Figure intends to show how the flow of people who go to their routine activities varies.

Figure 5.30 has 31 matrices that show the passenger flow between some stops on route 765 and an exit stop on the same route (called 'Colegio Militar (Metro)'), and each matrix corresponds to each day of October between 8 am and 10 am. Note that this route does not work on weekend days and therefore those days are not displayed. This route was chosen because it contains a few stops. The matrix layout is organized by weeks, that is, each week is a line of matrices from Monday to Friday. In short, the month of October 2019 had five weeks.

Looking at the matrices in Figure 5.30, only one exit point was chosen, since every day of the month October there was an outflow, probably because that stop has a close connection to the subway. It can

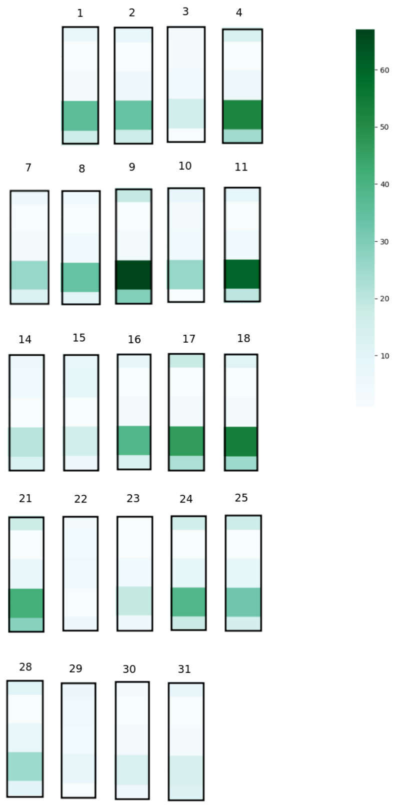


Figure 5.30: Routine-based matrices for each day of October 2019, boardings stop from route 756 with one exit stop on ”Colégio Militar”, granularity stop.

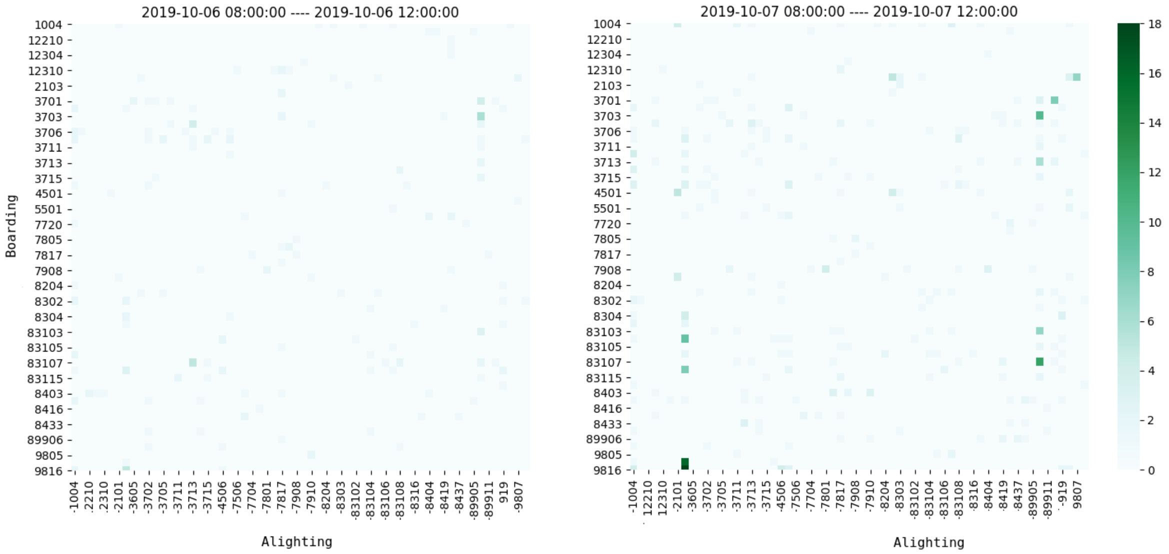
be stated that the shade scale ranges from 0 to 60, in the Figure 5.30 and in all the matrices there is a boarding stop that stands out because it is the one that generates substantially flow of passengers, in direction to exit stop ’Colegio Militar’. Future investigations on these matrices would be important to consider the analysis on different months, to compare periodically. We can conclude that this type of calendar analysis allows the operator to adjust.

Figures 5.31 and 5.32 state something that is quite predictable at the routes 759 and 728, respectively. These figures analyze the difference in the flow of commutative trips, in the period of 8 am, and 10 am, between a working day Monday (matrix on the right) and a weekend day Sunday (matrix on the left). There is a greater flow of routes on weekdays, as commuting passengers tend to travel to work or school on working days.

5.4.3 Granularity horizon Exploration

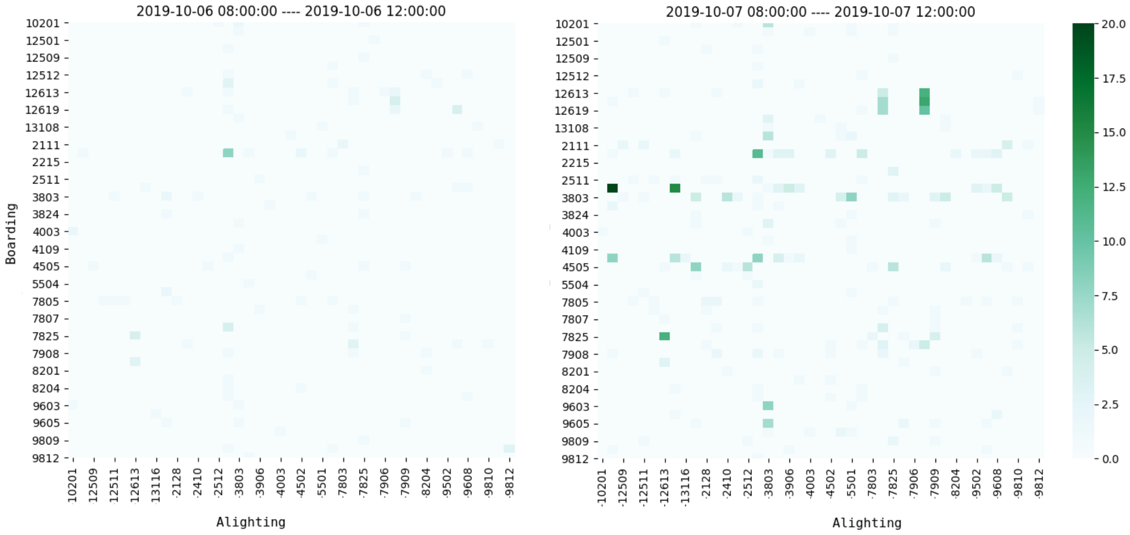
As mentioned in the Solution Proposal section, a matrix can assume different granularities, that is, the rows and columns can be entry and exit stops, or aggregations of entry and exit stops. In the visual dashboard, we present two types of granularities that are TAZ and statistical section. The next image is an example of TAZ granularity. However, in the figure, we do not limit to showing the effect of the matrix granularity, it is also compared matrices referring to the same route at different times of the day.

Figure 5.33, has eight matrices referring to trips made on route 759, and the stops are grouped


2019-10-06 08:00:00 ---- 2019-10-06 12:00:00

2019-10-07 08:00:00 ---- 2019-10-07 12:00:00

Figure 5.31: Passenger flow difference between Sunday and Monday on the 759 route.


2019-10-06 08:00:00 ---- 2019-10-06 12:00:00

2019-10-07 08:00:00 ---- 2019-10-07 12:00:00

Figure 5.32: Passenger flow difference between Sunday and Monday on the 728 route.

geographically by TAZ's. Each matrix contains trips made in 2 hours, on the 2nd of October. The first line with two matrixes corresponds to the following periods: from 6 am to 8 am and 8 am to 10 am. The second line contains three arrays from the following periods: 10 am to 12 am, 12 pm to 2 pm, and 2 pm to 4 pm. the third line contains the matrices of the following periods: 4 pm to 6 pm, 6 pm until 8 pm, 8 pm until 10 pm. All matrices share the legend of boarding TAZ's that is in the first line, in the horizontal. Furthermore, all matrices share the legend of landing TAZ's that is in the last column, vertically. Notice that the Figure shows only a few TAZ's in Lisbon, where route 759 has stops

Observing Figure 5.33 and its matrices, we can draw the following interesting events: the largest passenger flow occurred from 8 am to 10 am, with boardings at the TAZ Santa Maria Maior to exit TAZ Penha de França; between 6 am to 12 am and in the afternoon between 4 pm and 8 pm, there is a

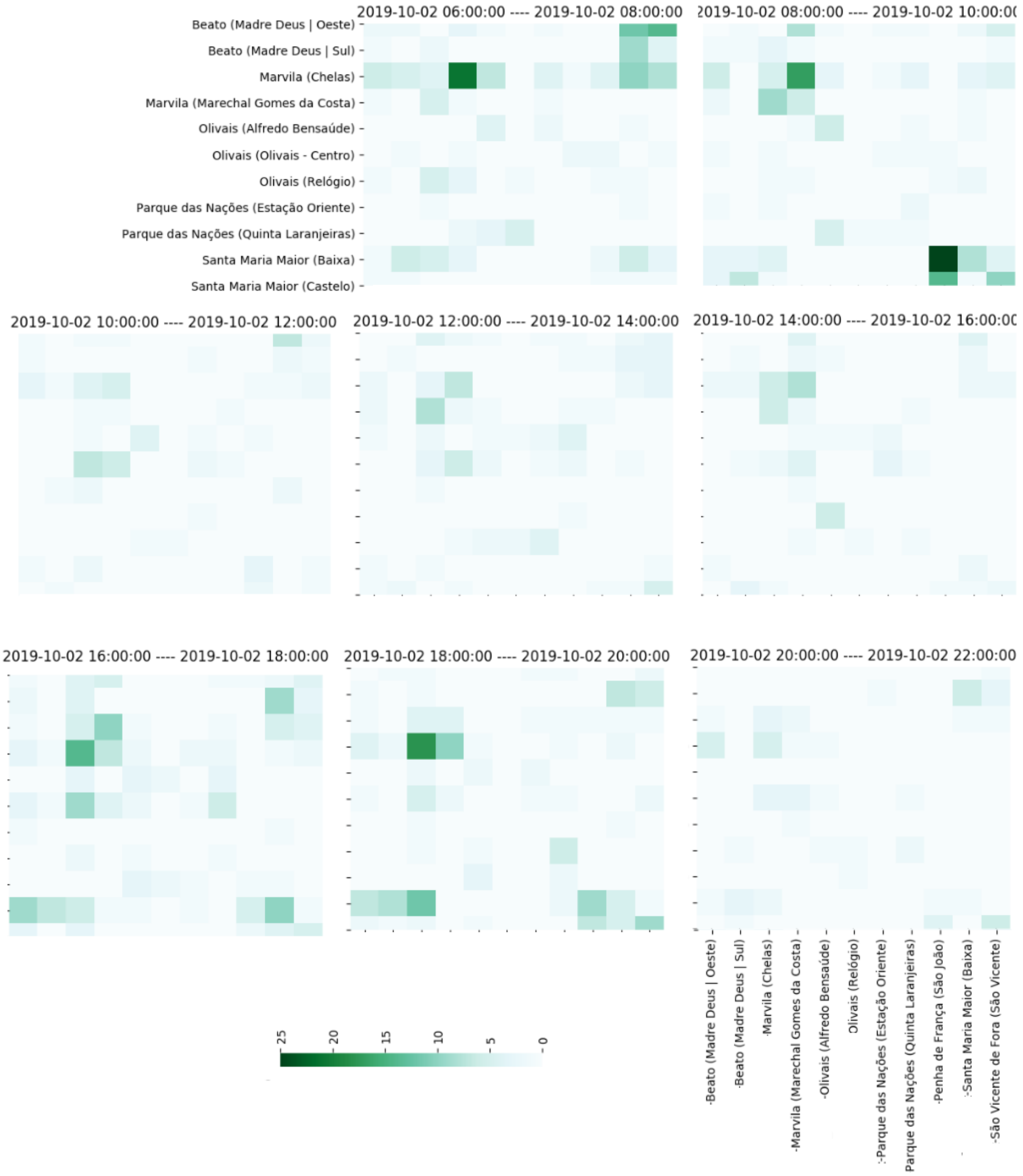


Figure 5.33: Matrices with passenger flows in 2-hour periods during the 2nd of October, on route 759, with TAZ granularity.

greater flow of passengers between TAZ; there is passenger flow boarding and alighting between stops on the same TAZ; it is possible to envision an interesting event: in the period from 8 am to 10 am, the largest passenger flow occurs between Marvila (Chelas) to Marvila (Marechal Gomes da Costa). In this last referred TAZ there is a subway station. And at the end of the day from 6 pm to 8 pm, the largest flow of people occurs on the reverse route, from Marvila (Marechal Gomes da Costa) to Marvila (Chelas).

This research line is beneficial to discover which geographic areas of the city of Lisbon are attracting zones and producing zones at different times of the day, to later analyze and compare with socioeconomic indicators.

5.4.4 Titles horizon Exploration

One of the parameters of the dashboard to filter the matrices corresponds to the title, for example, we can choose matrices where we only view the flow of passengers with one or more titles. This approach is mainly interesting for making a flow's profile at certain titles.

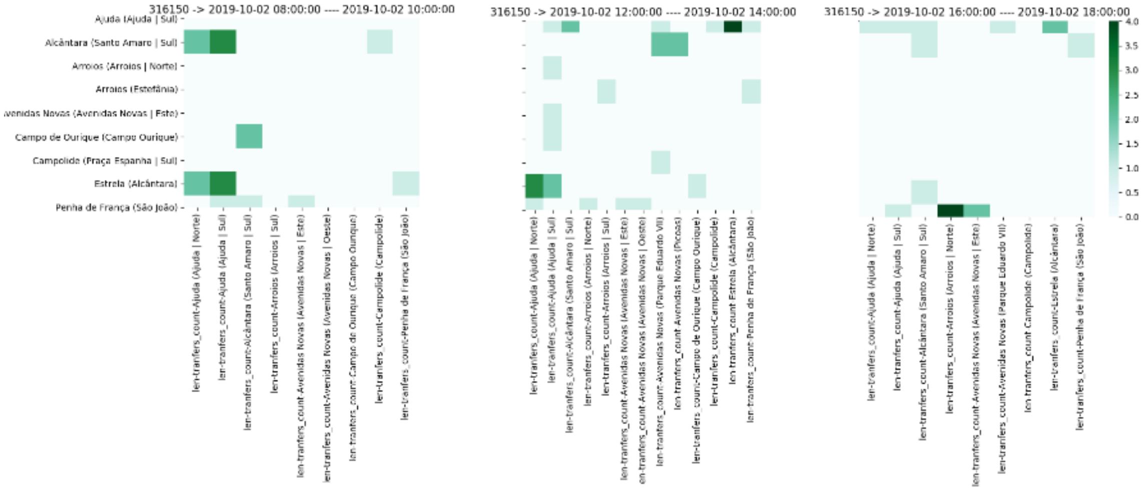


Figure 5.34: Routine Matrices filtered for card title Navegante 4_18 / Sub23 for different periods of the 2nd of October, boarding and alighting in the route 722.

Figure 5.34, is a practical example of how we can assess the flow of passengers who carry the title Navegante 4_18 / Sub23 (intended for students under 23 years old). Each matrix in the Figure represents a different time window from October 2, 2019, with TAZ granularity and flow only belonging to route 722 (that is, they embarked and disembarked on this route). The first matrix corresponds to the time of 8 am until 10 am, the second corresponds to the time window between 12 am and 2 pm, and the third matrix is between 4 pm and 6 pm.

From the three matrices, we can observe that, during the morning, this profile demarcates on the way from Alcantara and Estrela to Ajuda Sul, while in the period from 12 am we start to see the reverse way, that is from Ajuda Sul to Estrela.

5.4.5 Other metrics

In the proposed solution, we present other metrics in addition to the demand between origin and destination (OD). It is also possible to envisage other relevant variables on the network, such as: time or distance between OD, time and distance traveled in transfers, destination-origin (DO), or percentage of passengers who had to use the metro at least one travel segment between two locations on the network.

For the CARRIS operator, these variables are extremely interesting to evaluate their services and for this reason we provide this solution. In order not to lengthen the investigation, we only present some matrices on some variables that view the entire bus network in general, without paying attention to particular places. Since CARRIS has greater insight into its services it will make better use of this part of the solution.

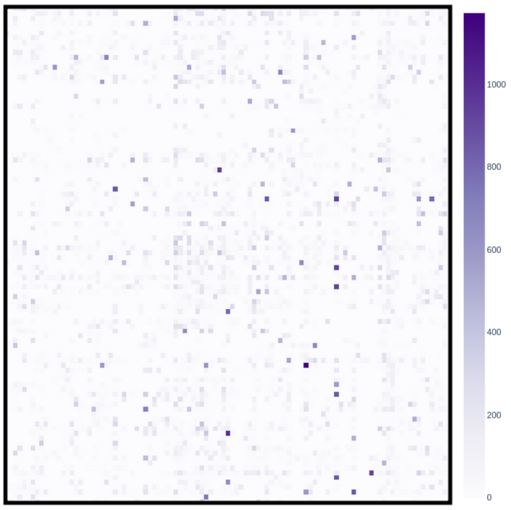


Figure 5.35: Routine matrix showing the pattern of median time spent between transfers between TAZ's, across the network, between the 1st and 7th of October.

One of the matrices presented is in Figure 5.35, which shows the pattern of the median time spent between transfers between TAZ's, between all days from day 1 to 7 of October 2019. It appears that on the scale the interval goes from 0 to 1000 seconds, that is, between transfers is spent at most about 16 minutes. In addition to these metrics concerning time, it also has the average time spent on transfers. The median is more robust to deviations from the pattern caused by the incorrect inference of the landing times, therefore it is preferable to observe the median.

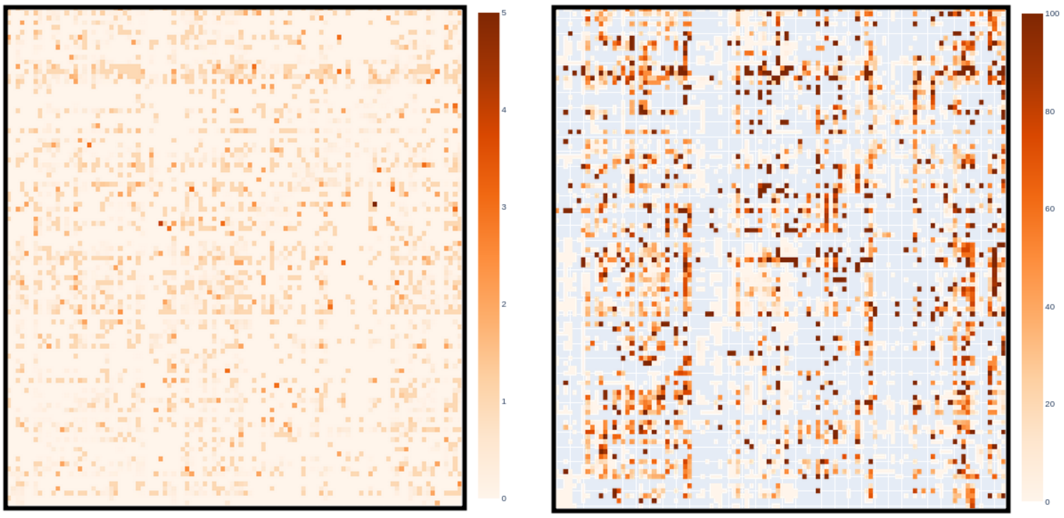


Figure 5.36: Routine matrices showing the pattern of average number of transfers and percentage of passenger journeys with metro within, between transfers between TAZ's, across the network, between the 1st and 7th of October.

The variable time and distance spent between transfers belong to the category of Information on Transfers, and it is necessary to emphasize that the lines correspond to places of disembarkation and the columns correspond to places of embarkation, therefore the information regarding the transfers has a destination-origin disposition. (DO). Besides, to the category mentioned above, there is also information on variables regarding travel information between locations, origin-destination (DO), such as the number of transfers required between two locations, time spent and distance travelled between two locations, percentage of passengers who had to use the subway at least one travel segment between two locations on the network.

The Figure 5.36 shows two matrices that contain information about two variables regarding the category Travel Information. The matrices correspond to all the TAZ's in Lisbon, during the 1st and 7th of October 2019. The first matrix (left) represents the average number of transfers needed between a place of origin and destination. And the second matrix (right) represents the percentage of passengers who needed to use the metro and the bus between two points in the network.

The matrix on the left, which shows the number of transfers required between TAZ's, is relevant for determining whether the transport service between these two locations is efficient. An ideal matrix would be if there were 0 transfers between any TAZ, however there are trips between points in the network where it is necessary to carry out more than one transfer. Combining this matrix and the one that shows the flow demand, it will be possible to identify new priority routes to satisfy citizens' needs.

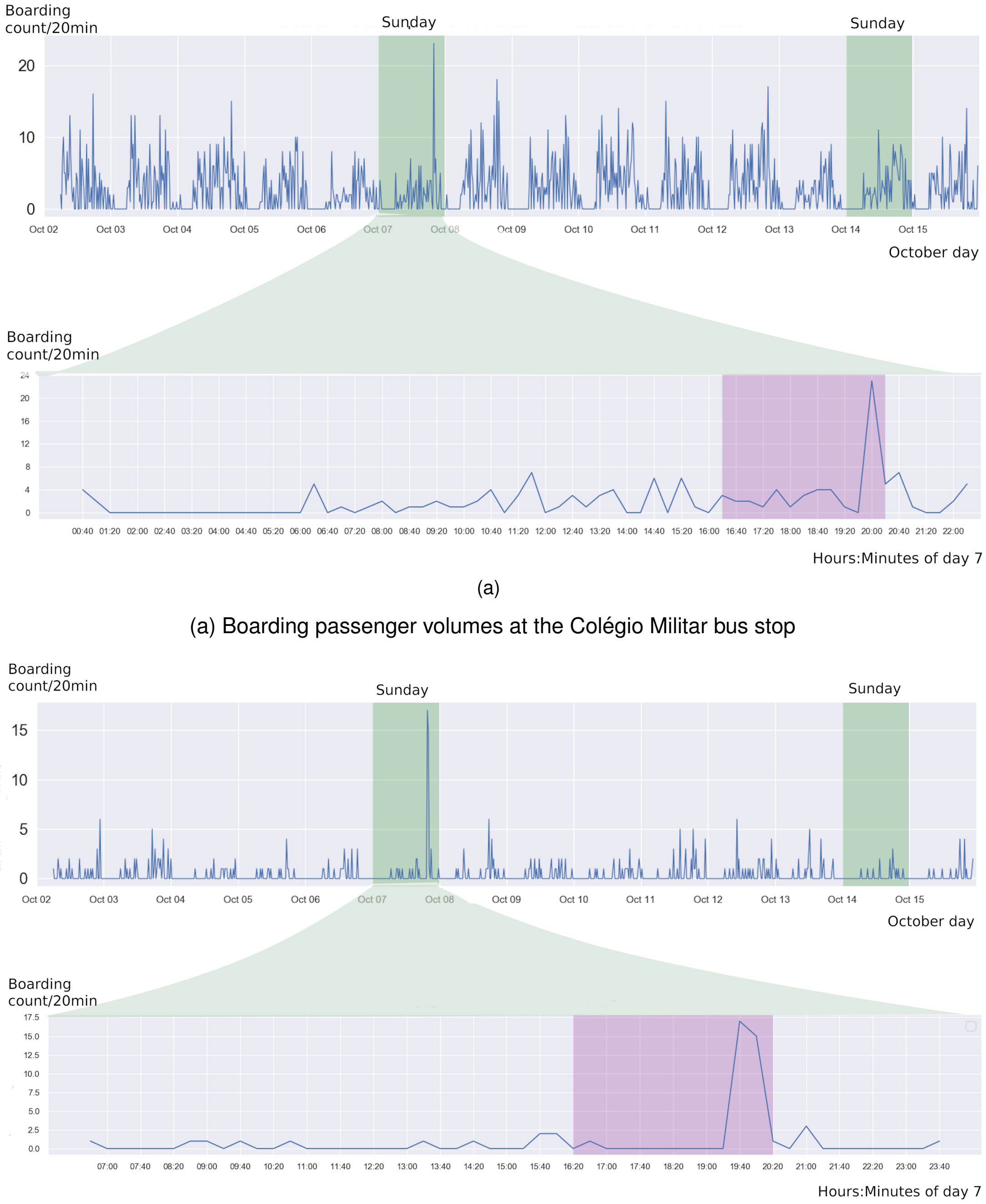
The matrix on the right, which shows the percentage of passenger journeys in which it was necessary to use the metro, is especially interesting to analyze the bus network's limitations. In the matrix, there are trips between TAZ's in which bus and metro are always used (100%), which shows that the CARRIS operator does not have adequate and sufficient service. If in the future, the METRO and CARRIS operators join forces to promote balanced and efficient traffic of citizens, it will have as its starting point this matrix, which gives light on how the two modes of transport are being used over the city of Lisbon.

5.5 Situational Contextual Discovery in Data

During the development of this research work, the effect of the situational context on urban public transport data was also analyzed, and written in the article ”On the Need to Combine Sources of Situational Context in Public Transport Data Analysis”, within the scope of the European Transport Conference 2020 (ETC).

The analysis carried out relates to boardings in the bus and metro network, that is, boarding transactions are aggregated and transformed into discrete multivariate time series. The aggregation of data in a time series may in some cases transmit the demand over a long time on a route or even a stop. The following images will show disruptive indicators on the bus traffic dynamic, caused by the situational context, such as events, calendar context, and road closures.

Figure 5.37 is a perfect example that shows the disruptive feat of an event planned in close stops. The 5.37 (a) and 5.37 (b) subfigures provide graphical view of passenger boarding in the two bus stops near to the ”Luz” stadium in the period of 2 of October and 15 of October of 2019. In the Sunday day, 7



(a)

(a) Boarding passenger volumes at the Colégio Militar bus stop

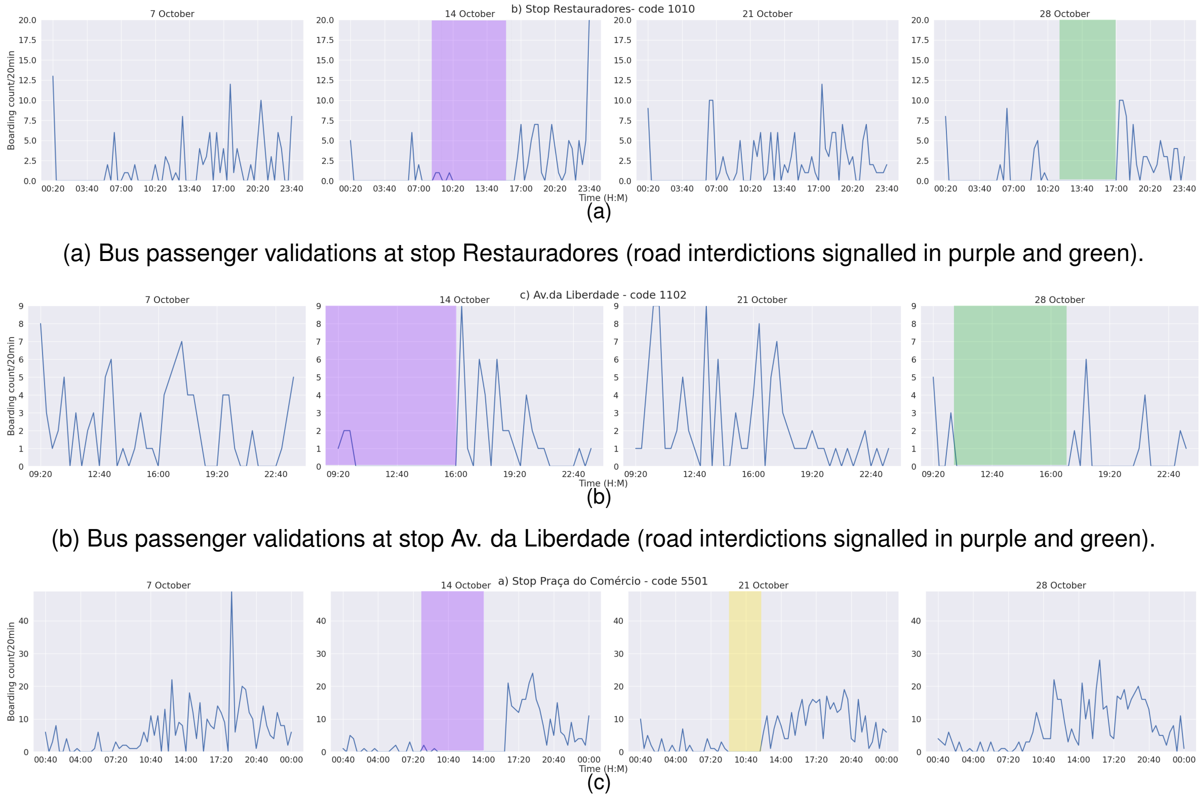
(b)

(b) Boarding passenger volumes at the Estádio da Luz bus stop

Figure 5.37: Impact of a soccer match at Luz Stadium (Oct 7, 17h30).

of October, between 17h30 and 19h10 a soccer game was played between the ”Sport Lisboa e Benfica” and ”Futebol Clube do Porto” teams and both attract thousands of fans in the country to the stadiums.

In order to correlate same week day, we highlighted the Sundays, and we zoom the day where the event occurred. The zoomed visualization is also highlighted in period between 70 minutes before and after. A regular Sunday such day 14 of October has the expected behaviour which it is high flow in working days than the weekends. However in the day 7 there is a strong evidence of disruptive event due to the presence of a outlier 30 minutes after (Fig 1a) and 50 minutes (Fig.1b) after of the end of the game.



(a) Bus passenger validations at stop Restauradores (road interdictions signalled in purple and green).

(b) Bus passenger validations at stop Av. da Liberdade (road interdictions signalled in purple and green).

(c) Bus passenger validations at stop Praça do Comércio (road interdictions signalled in purple and yellow).

Figure 5.38: Impact of a road interdictions on bus stops.

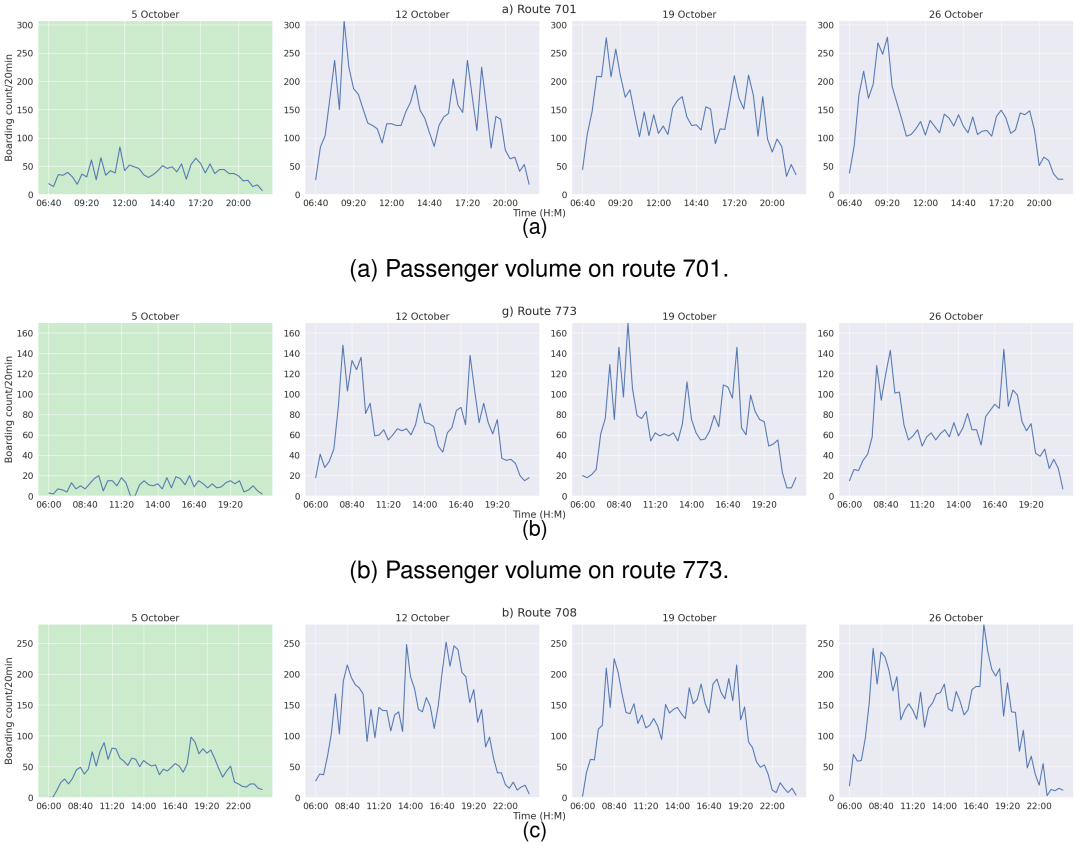
We also explored the effects of road closures in bus passengers' traffic. To analyse stops that are in the vicinity of the critical road segments, we created a tool that places closure segments and bus stops in the Lisbon map. The majority of road interdictions were settled in streets weren't any bus route pass by. However, in three consecutive Sundays occurred events in the city, which are assigend with a color in graphic visualizations. At the day 14, between 8am and 14pm, occurred a marathon event, placed near to the bus stop ”Restauradores” , the bus stop “Av da Liberdade” and with finish line near to the bus stop ”Praça do Comércio”, that is represented with purple color; at the day 21 occur at 10 am, a bike event placed in ”Praça do Comércio”, that it is represented with yellow color; at the day 28 occurred at 12 am a event which we don't has knowledge of the nature event (it was not described in the database) , in the locations ”Restauradores” and ”Av. da Liberdade”, that are represented in green color .

In this case we leverage a time series related with closures on a stops at the location stop at the ”Restauradores” Figure 5.38.A, ”Praça do Comércio” Figure 5.38.B, and ”Av. Liberdade” Figure 5.38.C Likewise the visual graphic shown in Figures 1A and 1B , we explored traffic pattern similarities between all October Sundays. Comparing with 7th of October, (where an event occurred) with the three bus stop locations presented (Figures 5.38.A, 5.38.B and 5.38.C), we can easily observe a strong evidence for disruptive traffic impacts as follows:

- At day 14, we observe zero boardings in the three locations around at the period of 10 am and 16:00 h. Such graphical overview is aligned with what is expected when a marathon event blocks the vehicles in the presented road segments. On the other hand the race participants were able to travel free of charge in transport to reach the places where the marathon started or to return home at the end. Therefore, there is a evidence of a peak point in the graphical view of day 14, after the end of the event.

- At day 21, we observe zero boardings on ”Praça de Comércio” around 10 am. It shall be noted that the bus top ”Restauradores” has a time window of zero boarding around 00h20 and 5 am because buses do not run in that period.
- At day 28, it is again expected zero boardings after the 12am since it started the aforementioned event, in the stops at the ”Restauradores” and ”Av. da Liberdade”

.



(a) Passenger volume on route 701.

(b) Passenger volume on route 773.

(c) Passenger volume on route 708.

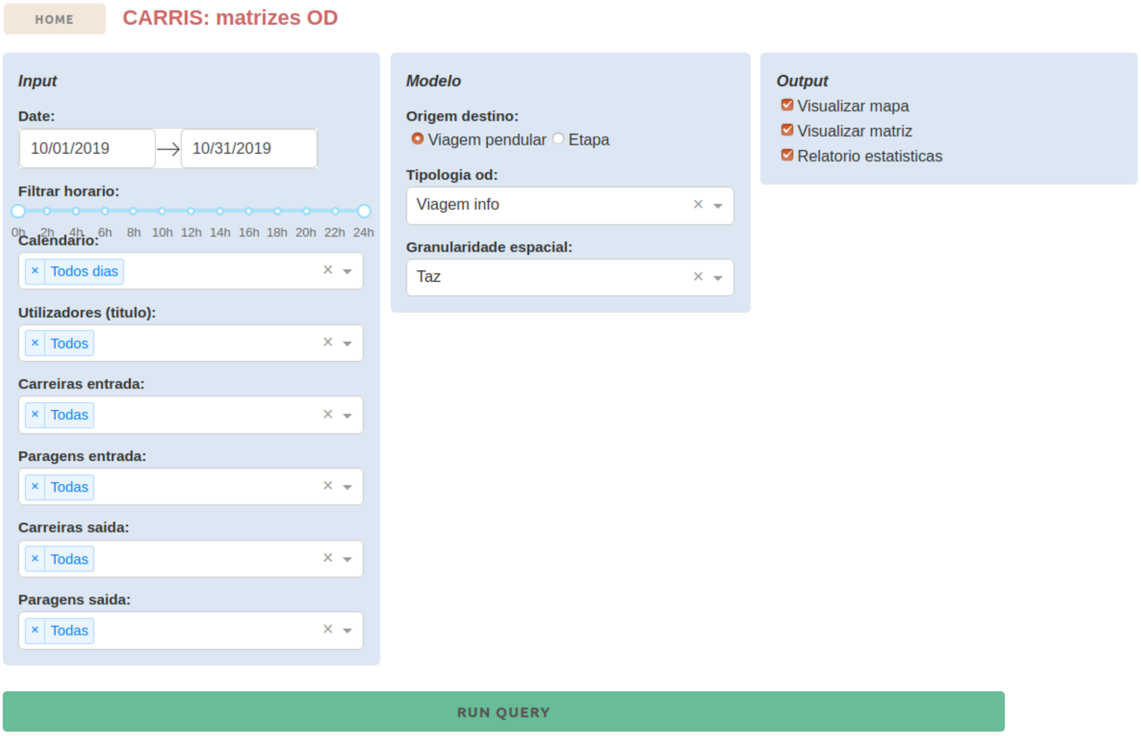
Figure 5.39: Effect of the national holiday (Oct 5 - Friday) and other October Fridays, on passenger volume along three different bus routes.

The following visual graphics in the 5.39 are an example of the impact of atypical day in the bus routes, such a national holiday . Each route visualisation is subdivided into four time series, representing the same week day (Friday) and the national holiday (5 of October) is highlighted with green color. It is extremely evident, that it exits a similar pattern between the last three Fridays and comparing with the Friday of 5 of October we observe a drop in demand.

Chapter 6

Visualization Tool

This chapter presents some prints of the visualization tool implemented in the ILU project's scope, whose functionalities were identified with the support of the stakeholders - CARRIS and CML. This tool was developed in Python with the use of Plotly and Dash packages, and it allows the presentation of origin-destination matrices and complementary relevant information.



CARRIS: matrizes OD

Input

Date: 10/01/2019 → 10/31/2019

Filtrar horario:

Calendário: Todos dias

Utilizadores (titulo): Todos

Carreiras entrada: Todas

Paragens entrada: Todas

Carreiras saida: Todas

Paragens saida: Todas

Modelo

Origem destino: ● Viagem pendular ○ Etapa

Tipologia od: Viagem info

Granularidade espacial: Taz

Output

☑ Visualizar mapa
☑ Visualizar matriz
☑ Relatorio estatisticas

Figure 6.1: Dashboard for matrix displaying.

Figure 6.1 corresponds to the graphical interface responsible for parameterizing the matrices. Contains 3 sections for parameterization, which are:

1. **Input**: Section where it is possible to filter the content of the matrices, through the following parameters
    - **Date**: Selects the range of days
    - **Filter time**: Option to choose a time window
    - **Calendar**: The calendar allows you to select specific days, for example working days, weekend, Monday, etc.
    - **Title card**: It allows specifying one or a more specific group of titles, for example, ”Navegante +65” (the title used by people over 65 years).
    - **Boarding Routes**: It allows filtering routes were passengers have boarded.
    - **Boarding Stops**: It allows filtering stops were passengers have boarded.
    - **Exit Routes**: It allows filtering routes were passengers disembarked
    - **Exit Stops**: It allows filtering stops were passengers disembarked.
2. **Model**: Section that chooses the model to view, through the following parameters:
    - **Origin Destination**: There are two options for this category which are the two types of time-based matrices (displays matrices regarding segment trips) and routine-matrices (display matrices regarding journeys).
    - **OD typology**: Selects as metrics that you want to view in the matrix cells. For each OD mentioned in the previous point, there are specific metrics.
        (a) For time-based matrices, the following metrics are available:
            - Passenger demand
            - Percentage of the number of segments where the next segment was subway
        (b) For routine matrices, the following metrics are available:
            - Travel information between two locations (entrance to exit).
                i. Journeys number
                ii. Average and median time
                iii. Average and median distance
                iv. Percentage of metro trips
            - Information regarding transfers between two locations (exit to entrance)
                i. Number of transfers
                ii. Average and median time spent on the transfer
                iii. Average and median distance walked on the transfer
    - **Granularity**: The columns and rows of the matrices can correspond to entry and exit stops, but they can also be geographic aggregation of stops, such as TAZ, or statistical section.

3. **Output**: Determines which elements can be viewed on the page. The available options are:

- **Statistical Report**:A statistical evaluation of the filtered observations is presented. A trip is an observation that is associated with a subjective assessment (described in chapter mm). The total evaluation of all the observations, that are aggregated in the matrices, is shown through the metrics presented on the violin (average, median, etc.) and the strip plot shows the distribution of the observations, within the scale of 0 to 100%, like it is shown in the figure6.3

- **Visualize Matrix**: By enabling this option, the graphical representation of the desired and parameterized matrix will be displayed. Each colour is associated with the main metric, and when we are hovering over the cell, more metrics are presented (for example, median time, distance ...). It is possible to zoom in to areas of the matrix.

- **Visualize Map**: It provides a map that shows the filtered entry and exit routes. When hovering a stop on the map, the respective code appears and what other routes pass, as shown in the figure 6.4.

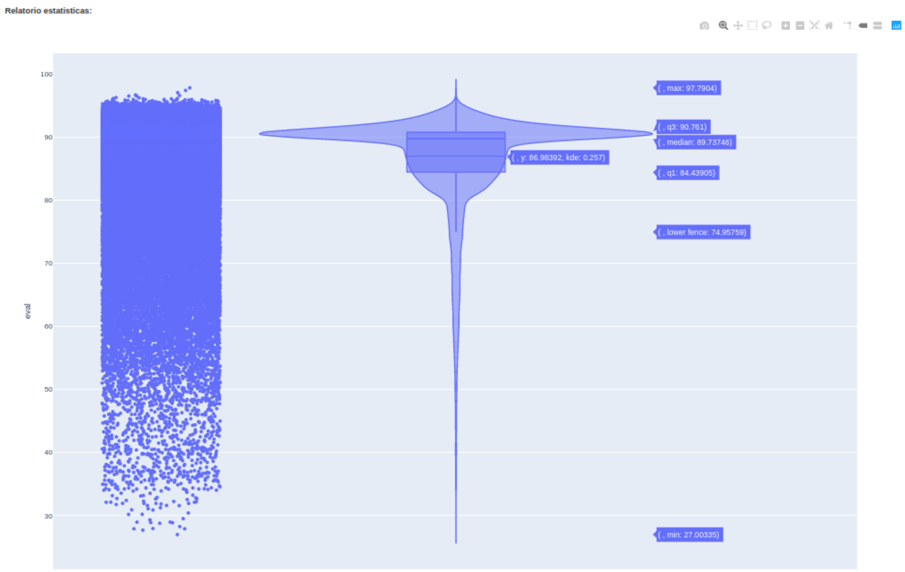


Figure 6.2: Violin and strip plot for statistical assessment of matrix' content.

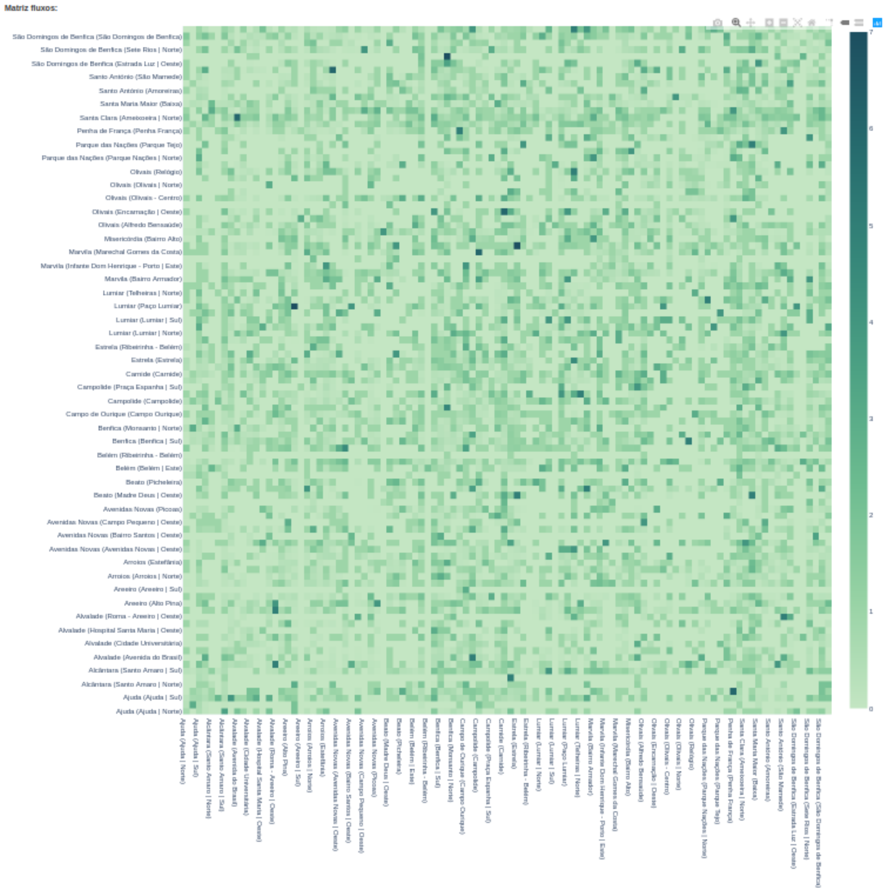


Figure 6.3: Graphical visualization of a matrix, showing the journey demand between TAZ's.

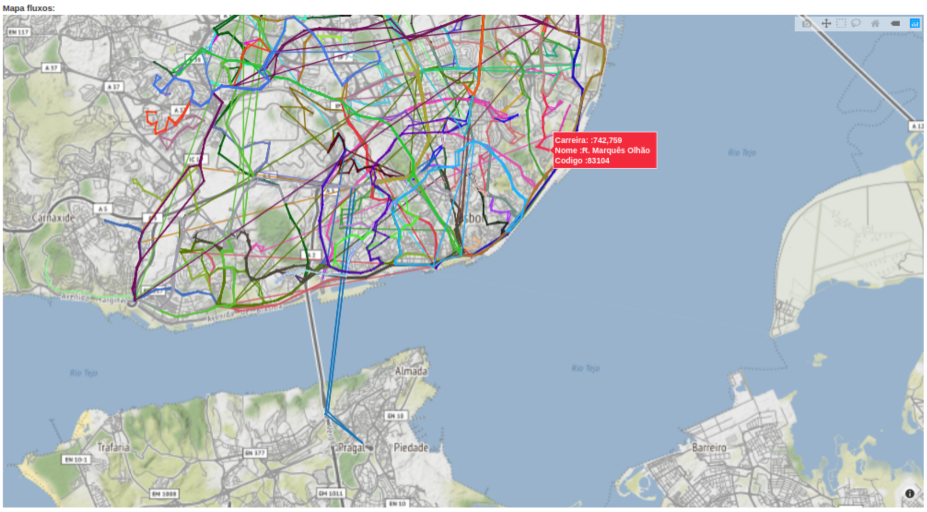


Figure 6.4: Map displaying all the routes of CARRIS network, selected on boarding and alighting routes.

Chapter 7

Conclusions

7.1 Achievements

This work provides novel contributions to the field of OD matrix estimation using passengers' boarding count data. In summary, the present dissertation provides both useful insights from theoretical and practical perspectives.

Firstly, we propose alighting stop inference models over the passengers' paths in the absence and presence of multimodal views, and further extended classical assumptions. The multimodal model demonstrated that alighting stop could be more accurately inferred for each transaction, in overall it was able to estimate the exits for 82% transactions from the input dataset (more 10% than the unimodal model), and 85% transactions corresponding to entire segments. In addition, the alighting stop inference is easily parameterizable to comprise assumptions on the maximum walking distances and waiting times on route transfers.

In addition, the proposed approach for inferring origin-destination matrices yields five unique contributions. First, we allow inference to consider multimodal commuting patterns, detecting individual trips undertaken along different operators. This was shown to be an essential step since nearly 20% of journeys in the Lisbon's transportation network require one or more transfers.

Second, we support dynamic OD inference along parameterizable time intervals and calendrical rules, and further support the decomposition of traffic flows according to the user profile. Moreover, we allow user to parameterize the desirable spatial granularity and visualization preferences.

Third, our solution efficiently computes several statistics that support OD analysis, helping with the detection of vulnerabilities throughout the transport network. In particular, statistics pertaining to commutation needs, walking distances and trip durations are supported.

Fourth, and finally, we show that the proposed solution is compliant with context-aware descriptive analytics by segmenting the periods in accordance with the available situational context and inferring context-specific OD matrices.

The contributions were validated with our stakeholders, CARRIS and CML, and have resulted in an accepted scientific manuscript accepted and presented in the European Transport Conference (ECT'2020), one extended abstract accepted in XIV Congreso de Ingeniería del Transporte (CIT'2020), one manuscript submitted in the European Transport Research Review (ETTR) journal, and four institutional presentations.

7.2 Future Work

There are numerous interesting work fronts to pursue this investigation, such as:

- alighting stop inference with additional modes, including boat, bicycle, train;
- context-aware inference of origin-destination matrices ;
- automatic anomaly detection in origin-destination matrices, including inference error, outliers provoked by context ;
- traffic flow prediction from recurrent origin-destination matrices in the absence and presence of context and situational context;
- extended visualization of traffic flow data using Sankey representations to visualize transfers more efficiently.

Bibliography

[1] N. Soares and A. Domingues. Consolidação e maturidade demográfica de uma área metropolitana. *Consultado a*, 27:2016, 2007.

[2] A. Ceder. Urban mobility and public transport: future perspectives and review. *International Journal of Urban Sciences*, pages 1–25, 2020.

[3] C. S. A. d. Almeida. *Planos de mobilidade no contexto da melhoria da qualidade do ar em Lisboa*. PhD thesis, FCT-UNL, 2010.

[4] I. Leite, A. Finamore, and R. Henriques. Context-sensitive modeling of public transport data. 01 2020.

[5] P. Russom et al. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34, 2011.

[6] A. Cyril, R. H. Mulangi, and V. George. Bus passenger demand modelling using time-series techniquesbig data analytics. *The Open Transportation Journal*, 13(1), 2019.

[7] A. A. Nunes, T. G. Dias, and J. F. e Cunha. Passenger journey destination estimation from automated fare collection system data using spatial validation. *IEEE transactions on intelligent transportation systems*, 17(1):133–142, 2015.

[8] J. J. Barry, R. Freimer, and H. Slavin. Use of entry-only automatic fare collection data to estimate linked transit trips in new york city. *Transportation research record*, 2112(1):53–61, 2009.

[9] A. Antrim, S. J. Barbeau, et al. The many uses of gtfs data–opening the door to transit and multimodal applications. *Location-Aware Information Systems Laboratory at the University of South Florida*, 4, 2013.

[10] E. Mishevska, B. Najdenov, M. Jovanovik, and D. Trajanov. Open public transport data in macedonia. In *Proceedings of the 11th Conference for Informatics and Information Technology*, pages 161–166, 2014.

[11] F. Lazzeri. *Machine Learning for Time Series Forecasting with Python*. John Wiley & Sons, 2020.

[12] C. Celes, A. Boukerche, and A. A. Loureiro. Towards understanding of bus mobility for intelligent vehicular networks using real-world data. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2019.

[13] M. Tanaka, T. Kimata, and T. Arai. Estimation of passenger origin-destination matrices and efficiency evaluation of public transportation. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 1146–1150. IEEE, 2016.

[14] A. Cui. *Bus passenger origin-destination matrix estimation using automated data collection systems*. PhD thesis, Massachusetts Institute of Technology, 2006.

[15] W. Zeng, C.-W. Fu, S. M. Arisona, A. Erath, and H. Qu. Visualizing mobility of public transportation system. *IEEE transactions on visualization and computer graphics*, 20(12):1833–1842, 2014.

[16] J. Hora, T. G. Dias, A. Camanho, and T. Sobral. Estimation of origin-destination matrices under automatic fare collection: the case study of porto transportation system. *Transportation Research Procedia*, 27:664–671, 2017.

[17] S. Explained. Sankey diagrams forenergy balance, 2000. URL `https://ec.europa.eu/eurostat/statistics-explained/pdfscache/50452.pdf`.

[18] A. Chaudhuri. A visual technique to analyze flow of information in a machine learning system. *Electronic Imaging*, 2018(1):380–1, 2018.

[19] J. 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*, 1817(1):183–187, 2002.

[20] D. Li, Y. Lin, X. Zhao, H. Song, and N. Zou. Estimating a transit passenger trip origin-destination matrix using automatic fare collection system. In *International Conference on Database Systems for Advanced Applications*, pages 502–513. Springer, 2011.

[21] J. Zhao, A. Rahbee, and N. H. Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22 (5):376–387, 2007.

[22] M. A. 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*, 24:9–18, 2012.

[23] 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*, 11(1):1–14, 2007.

[24] J. M. Farzin. Constructing an automated bus origin–destination matrix using farecard and global positioning system data in sao paulo, brazil. *Transportation research record*, 2072(1):30–37, 2008.

[25] N. Nassir, A. Khani, S. G. Lee, H. Noh, and M. Hickman. Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. *Transportation research record*, 2263(1):140–150, 2011.

[26] W. Wang, J. P. Attanucci, and N. H. Wilson. Bus passenger origin-destination estimation and related analyses using automated data collection systems. 2011.

[27] J. B. Gordon, H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci. Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. *Transportation research record*, 2343(1):17–24, 2013.

[28] F. Devillaine. Towards a reliable origin-destination matrix from massive amounts of smartcard and gps data: application to santiago in: Zmud, j., lee-gosselin, m., munizaga, ma, carrasco, ja (eds.). transport survey methods; best practice for decision making, 2013.

[29] 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*, 2535(1):88–96, 2015.

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

[31] M. Mamei, N. Bicocchi, M. Lippi, S. Mariani, and F. Zambonelli. Evaluating origin–destination matrices obtained from cdr data. *Sensors*, 19(20):4470, 2019.

[32] A. Ali, J. Kim, and S. Lee. Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, 20(4):1532–1539, 2016.