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

Sofia P. Cerqueira
Instituto Superior Técnico, Universidade de Lisboa
sofiacerqueira@live.com.pt

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).

Index Terms—**sustainable mobility, alighting stop inference , multimodality, georeferenced multivariate time series**

I. 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.

The remainder of the present research work is organized as follows. Section II describes previous related works to the researched area alighting stop inference and OD matrix estimation and shows the main contributions in the field. Section III describes the real-world problem, proposes a solution and outlines its advantages . Section IV shows the practicalities of the proposed solution, discusses its limitations and assesses the solution considering data visualisations and performance metrics. Section V of the document describes the set of instructions for using the developed tool.

II. Related work

A. Aligthing Inference

Whenever the smart card is used in the public transport vehicle, an electronic record is generated and registered in the AFC system. Until now, the public transport operator CARRIS only registers passengers' entries at each bus stop (boarding or entry-only count data). For future work on passenger flow analysis, the exit count data cannot be extracted from the available data and therefore has to be inferred. The literature on this topic outlines several implementations to different transport systems worldwide, which differ mainly by the set of assumptions implemented. To overcome this problem, the literature suggests several solutions 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.*

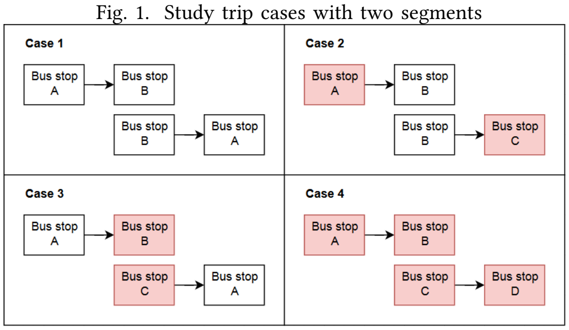
The table I outlines some papers where the authors applied these assumptions in different modes of transport and different cities in the world:

TABLE I
Literature review for alighting estimation

Literature	Mode	Location	Assumptions
Barry, et al. (2002) [5]	Subway	New York City	1 and 2
Barry, et al. (2009) [6]	Subway, bus, ferry	New York City	1 and 2
Nunes, et al. [7]	Bus	Porto	1, 2 and 7
Li, et al. (2011) [8]	Bus	Jinan, China	1 and 2
Zhao, et al. (2007) [9]	Rail System	Chicago	1,2 and 3
Munizaga, et al. (2012) [10]	Bus	Santiago, Chile	1, 2 and 3
Trépanier, et al. (2007) [11]	Bus	Canada	1, 3, 4 and 5
Farzin, et al. (2008) [12]	Bus	S. Paulo, Brasil	1, 2 and 3
Nassir, et al. (2011) [13]	Bus	S. Minneapolis-Saint Paul (USA)	1, 2, 3 and 6
Wang, et al. (2011) [14]	Bus	London, UK	1 and 2
Gordon, et al. (2013) [15]	Bus	London, UK	1, 2 and 3

To visualize the alighting stop problem, Barry, et al. [5] 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 (assumptions 1 and 2). Thus, it is represented the possible travel cases in figure 1, for a trip with two only segments, which may take place within 24 hours.

Fig. 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.

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 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.

There are several proposals and assumptions in the scientific community [10] [5] [17] [9] [18] to make inferences of exit stops for an only-control system. For example, Barry, et al. [5] compared the results with real station exit counting, which is extremely difficult in an entry only system control. Zhao, et al. [9] and Wang, et al. [14] compared the results with data from surveys. Alsger, et al. [18] conducted a sensitivity analysis, using the different assumptions, for example, the author validates the trip duration regarding the number of transfers. The research found with most complete validation work was performed by Munizaga, and Palma (2014) [10].

Munizaga, and Palma (2012) [10] 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) [19] follows the analysis methodology of Devillaine, et al. (2013) [17] 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.

B. *OD matrices Typology*

Various studies use the representation of OD matrices to explain the flow of passengers in a transport network. In the literature, we can find two types of trips used to build the OD matrices. For example, if a person decides to travel from A (home) to B(son's school, where spent its time less than 5 minutes) and then goes to C (workplace), can we say that the origin is A and the destination is C (one trip), or it should be origin A to destination B and also origin B to destination C (two trips). In fact, in the literature, many emphasize this distinction, however there is not a fixed denomination in order to distinguish the types of travel.

Cui, tel al. [20] 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. [21] 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 segments until reaching the destination 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. Segments of observed CDRs are merged to obtain the parts of a commute trip (for example, home-to-work and work-to-home).

C. *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. [22] 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 2) shows six segments summarized in only three journeys, one between home to work (as we can see in figure 2); work to shop after 17 pm; and then shop to home.

Fig. 2. Study trip cases with two segments

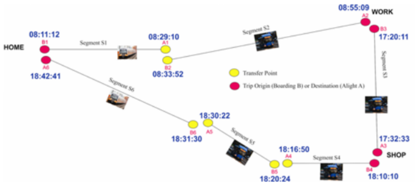


Fig. 3. Figure from [22]

Ali, et al. assumes that activity time, 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. 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). Nevertheless, [13] considers that activity should last at least 30 min, And, the maximum waiting time for a person to transfer cannot exceed 90 min.

III. Proposed Solution

In this section, two methodologies are described to approach estimating exits for transactions that occurred in the CARRIS system (public bus transport) and a model to generate commuting trips. Finally, it is described as the representation of origin and destination OD to show the flow of passengers and other metrics to represent the network's functioning. Therefore, solution development phases are as follows:

1) preprocessing of the public traffic data from the bus operator CARRIS;
2) consolidation smart-card transaction from AFC system from bus operator CARRIS and subway operator METRO;
3) alighting bus stop inference model is required to know passengers' disembark points in the bus network because CARRIS is a system that only requires to tap the smart card when the passenger enters (only-entry system);
4) alighting dual-mode stop inference model, that traces the passenger path in the transport mode bus and subways. This model is proposed to improve the limitations of the aforementioned bus model III-A1;
5) derive commute travels journeys from the trip segments produced by alighting stop inference model;
6) inference of OD matrices that explain the passenger flow between different network locations (stops, Taz, statistical sections);
7) inference of OD matrices that explain the past state of the network, according to with metrics such as time, distance, and the number of transfers spent between an origin and destination, time and distance spent in a journey, percentage of journeys that used subway within;
8) display dynamic OD in a graphical dashboard, with the possibility of filtering and parameterization, such as time window, location granularity, and passenger typology, among others;
9) description and corporation of the situational context within georeferenced multivariate time series.

The datasets with transactions under analysis come from the Lisbon bus transport (CARRIS) from the period of October 2018 and October 2019. The auxiliary dataset from the Lisbon subway operator (METRO) corresponds to the period of October 2019. CARRIS is the leading bus operator, and in 2018 and 2019 it served a total of 125684 and 139496 passengers, respectively. Due to the fact of having been able to consolidate the data for October 2019, from the METRO operator and CARRIS operator, the analysis will be more focused on this period (October 2019).

A. *Alighting Stop Inference*

This section aims to present a methodology for estimating the alighting of the passenger for each boarding transaction, collected from the entry-only AFC system of operator CARRIS. The first model - Bus Model- can only visualize the travel segments in the CARRIS systems while the second model - Dual-mode model - can trace the passenger's path in the bus (CARRIS operator) and metro (METRO operator) transport modes. Both models proposed can estimate the missing information, that is, the alighting stop and time of each transaction from the bus operator.

*1) **Bus Model:*** The algorithm acts on the travel segments of each passenger, on a given day. This means that the algorithm's first step is, precisely, to collect from the database and transactions occurred during one day. It is necessary to emphasize that the CARRIS operator's period of activity 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 October and the transactions, collected from each day, are sorted by card identifier and then chronologically. In short, the algorithm can be described in these following steps:

1) Collect transactions $\{T_1, T_2, ...T_p\}$ for passenger S, ordered chronologically.
2) If the passenger only made one trip segment during the day, the process ends, and we move on to step one to analyze the path of a new passenger.
3) If the algorithm is dealing with T_1, then information regarding the geographic location and time boarding is saved.
4) For each segment T_n where n belongs to $\{1, .., p\}$ is estimated the alighting stop and time.
    a) If n belongs to $\{1, .., p-1\}$, then for each stop that is upstream of the entry stop and on the route of the transaction T_n, the distance to the boarding stop of the next segment is calculated T_{n+1}. It is chosen as the alighting stop, the one that is closest to the boarding stop of the next trip segment T_{n+1}.
    b) If n is p , then for each stop that is upstream of the entry stop and on the route of the transaction T_p, the distance to the boarding stop of the next segment is calculated T_1. It is chosen as the alighting stop, the one that is closest to the boarding stop of the first segment performed during the day T_1.
5) As soon as the algorithm estimates the information needed for each transaction, other transactions from another passenger will be analyzed. The process ends when analyses all transactions that occurred during the mentioned interval.

*2) **Dual Mode Model:*** This new model proposed for alighting stop inference aims to fill the gaps of the previous model. The bus operator network does not have enough services to meet all passengers' needs, and therefore some passengers use more than one mode of transport to reach their destination. Since the METRO (public subway operator) provided subway transactions data from the same period of bus transactions data, this new algorithm will overcome inference errors produced by the previous model. It will be able to trace the path passenger in both modes of transport (bus and subway). The inferential errors in the estimation of bus stops will subsist, because some passenger uses other modes of transport such as a boat, train, bicycles, among others. In short, the algorithm can be described in these following steps:

1) Collect transactions $\{T_1, T_2, ...T_p\}$ for passenger S, ordered chronologically, where transactions can come from subway or bus operators.
2) If the passenger only made one trip segment during the day, the process ends, and we move on to step one to analyze the path of a new passenger.
3) If the algorithm is dealing with T_1, then information regarding the geographic location and time boarding is saved, whether subway or bus.
4) For each segment T_n where n belongs to $\{1, .., p\}$ is estimated the alighting stop and time, but if T_n corresponds to a transaction that occurred in the subway, then it isn't necessary to estimate anything, then the algorithm move on to the next transaction.
    a) If n belongs to $\{1, .., p-1\}$, then for each stop that is upstream of the entry stop and on the route of the transaction T_n, the distance to the boarding stop of the next segment is calculated T_{n+1}. It is chosen as the alighting stop, the one that is closest to the boarding stop of the next trip segment T_{n+1}. T_{n+1} can be a metro or bus transaction.
    b) If n is p , then for each stop that is upstream of the entry stop and on the route of the transaction T_p, the distance to the boarding stop of the next segment is calculated T_1. It is chosen as the alighting stop, the one that is closest to the boarding stop of the first segment performed during the day T_1. T_1 can be a metro or bus transaction.
5) As soon as the algorithm estimates the information needed for each transaction, other transactions from another passenger will be analyzed. The process ends when analyzing all transactions that occurred during the mentioned interval.

In the two estimation models, other attributes are calculated for each transaction besides the exit stop location and timestamp, such as walking distance, transfer time, path distance, travel time, next mode used (this last attribute is only calculated in the second model proposed).

B. *Trip Generation for commuting travel*

The generation of commuting trips aims to derive, from a set of travel segments, the origin and the proposed destination location and therefore, the travel segments between these two points are no longer described. The result of this derivation is interesting from the point of view of analysis, planning and improvement of the transport public system. An application of the output of this model can be applied in the following example: if there is a high demand between point A to point 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. The proposed model is based on the following ideas:

- A passenger who makes a commutative journey is willing to repeat the same route frequently. Therefore, a threshold is defined to eliminate passengers that do not reach this limit.
- Among the transfers made during the day, only the transfer with the most extended time interval is considered activity time (work, school) and should have a bigger time interval than the stipulated.
- If the distance between segments is above a certain threshold, previously defined, then there may exist trip segments incorrectly estimated by the inference model.

In short, the algorithm can be described by the following steps:

1) Collect trip segments $\{V_1, V_2, ...V_p\}$ for passenger S, ordered chronologically.
2) Applying the first rule described, we only assume that passengers have commuting trips if they have made at least 18 trips during the month.
3) Applying the third rule, the distance between all segments must be less than 1000 meters, so the algorithm will be filtering incorrectly inferred trips. The distance between the last stop of the day and the first stop of the day must be less than 700 meters.
4) Of all transfers performed by the passenger, the one with the most extended time interval is chosen (e.g. transfer from S_l to S_{l+1}). Applying the second rule described above, this interval must be more than one hour.
5) If all restrictions are respected then the two journeys are derived, the outward journey is from the boarding stop S_1 to alighting stop of S_l and the return journey is from boarding stop of S_l to alighting stop S_p.

C. *Origin-Destination Matrices*

Conventionally, origin-destination (OD) matrices are tables that describe people movement between locations, but other metrics are placed to represent the status of CARRIS transport network in this solution, such as the number of transfers needed between the origin and destination. OD matrices are extremely useful for planning and improve the public transportation system.

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

- **Time-based matrices(tOD)** Matrices that only show demand between origin and destination with the counting of trip segments.
- **Routine-based matrices (rOD)** Matrices that present demand between origin and destination with the counting of commuting trips (outward and return journey).

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

1) Select range of days;
2) Time window between 0 am and 12 pm;
3) Select one and more week days;
4) Filtering by title card;
5) Select origin/boarding routes or stops;
6) Select destination/alighting routes or stops;

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

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

1) **Passenger counting**
2) **Percentage of trips** when the passenger uses subway transport after performing a bus trip segment.

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

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 value that reflects the number of transfers needed for it from origin to destination.
   c) **Mean and median travel time** : application of media and median on the table attribute **Travel Time**;
   d) **Mean and median travel distance**: application of media and median on the table attribute **Path distance** ;
   e) **Percentage of journeys with metro segment** : when the passenger uses subway transport within a journey.
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. ;
   c) **Mean and median transfer distance** average and median value that reflects the walking distance between transfers.

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.
2) **Statistical sections**: a section is a territorial unit corresponding to a continuous area of a single parish with about 300 housing units.

IV. Results

A. *Data description*

Table II shows 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. And comparing October 2019 against the same month of 2018 there was an increase regarding all the attributes, described in the table, namely an increase of 12.9% of passengers and an increase of 13.6% of transactions in the AFC system.

TABLE II
Summary description of the 2018 and 2019 datasets

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

B. *Model's Comparison*

This subsection compares 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 can infer the largest possible number of transactions with its respective exit stop. From the transactions in which it was estimated an exit, it is necessary to understand which models best fulfils the following assumption: passengers will start their next trip near the stop alighting location of their previous trip.

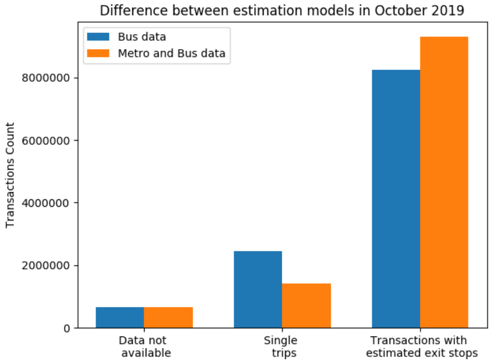


Fig. 4. Difference between estimation models in October 2019.

The orange bar plot in Figure 4 shows the results obtained for the dual-mode model (the model that traces the path of passengers within the metro and bus networks) whereas the blue bar plot represents the ones for the bus model (the model that traces the path only within the bus network). It can be observed an interesting phenomenon: part of the transactions that were not estimated (around 50% 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 to a point in the city and later used the metro to take off at 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.

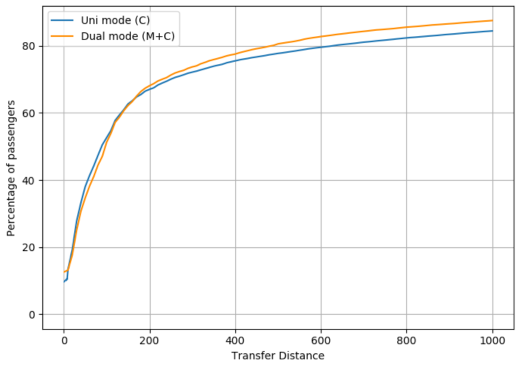


Fig. 5. Accumulative percentage of passenger walking in its transfers.

Figure 5.9 helps to understand which two models were able to fulfil the assumptions better. This graph represents the cumulative percentage of travel segments regarding the distance travelled in the transfer done after the trip segment. According to the assumption, a passenger tends to transfer with the shortest possible distance, so the model with more travel segments with a lower 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 increased 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. So, multimodal models 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 with greater precision (less walking distance in 80% of segments), because the travel segments travelled in the metro become visible. If we integrate data from other modes of transport such as trains, bicycles, boats, the model will be more robust.

C. *Dual mode data analysis*

Since the dual mode model achieves better performance, consequently we will proceed the investigation with the results of this model. So, this subsection will explore and discern the output generated by the dual mode model.

Figure 6 shows the proportion of the number of bus trip segments, in which the next segment was performed on the bus (left bar) and the metro 8(right bar). 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 used the metro at least once during the day, in order to complete its journey.

Figure 7 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 card types were chosen because they correspond to different age groups: 418/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. Observing the density of passengers entering during the day, we see that title 418/sub23 has the same density function as the title Navegante Metropolitano. 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 teaching institutions and returning home. Moreover, the entries in the other title may mean travelling to the workplace and returning home. The card type designated as "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.

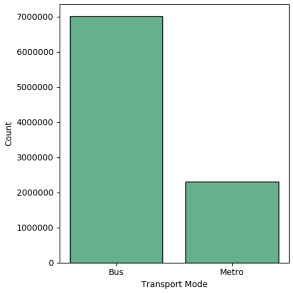


Fig. 6. Percentage of trip segment where the next segment was metro or bus, in October 2019

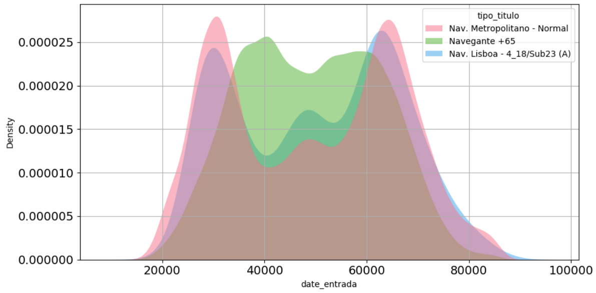


Fig. 7. Density of boarding by card title

D. *Trip Generation for commuting travel analysis*

In this section, an investigation on the commuting trips will be carried out. The process of generating commuting trips resulted in 3258394 journeys.

TABLE III
Counting of journeys by number of transfers performed during he journey

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 III shows the percentage and the absolute value of how many journeys made each number of transfers. As can be expected, we observed a higher number of trips without transfers (75%). About 17.22 % of commuting trips required a transfer. 7.19 % corresponds to trips that required more than two transfers.

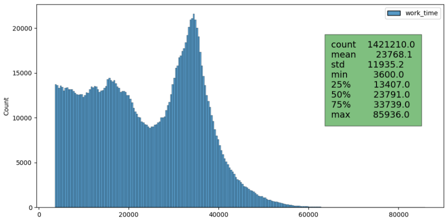


Fig. 8. Activity time distribution

During the generation of commuting movements, it was also possible to calculate the passenger's time at his destination. Therefore, figure 8 shows the distribution and statistics associated with the activity time (time spent after reaching the attracting destination). As you can see in the statistical data box, 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 around 9 hours of activity and then the passenger distribution suddenly decreases.

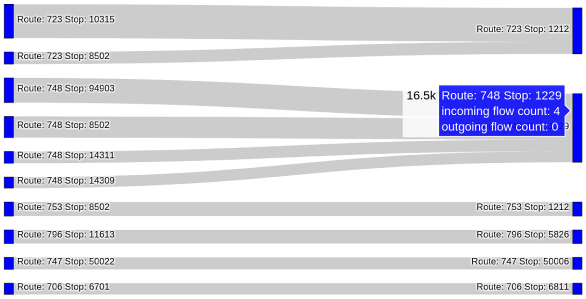


Fig. 9. Sankey representation for bus journeys for the purpose of later traveling on the metro, top 10 most frequent

The Sankey representation in figure 9 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, these bus plus metro connections can be revised to create direct bus connections, in order to avoid the overload on the metro stations.

E. *Analysis of OD Matrices*

This section explains and investigates the final product that dynamically visualizes the flow of passengers 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 operator will be able to 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.

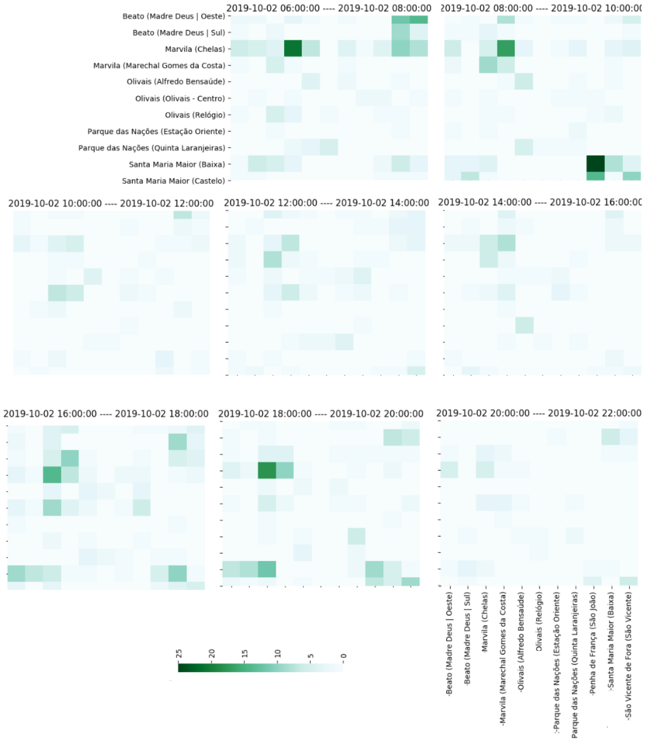


Fig. 10. Matrices with passenger flows in 2-hour periods during the 2nd of October, on route 759, with TAZ granularity

Figure 10, represents eight matrices referring to trips made on route 759, and the stops are grouped geographically by TAZ's. Each matrix contains trips made in 2 hours, on the 2nd of October. The first line with two matrices 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 759 route has stops.

Observing figure 10 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 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 will be useful to further understand which geographic areas in the city of Lisbon are attracting more passengers and generating more passengers' demand, along the different day periods.

F. *Situational Context 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 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.

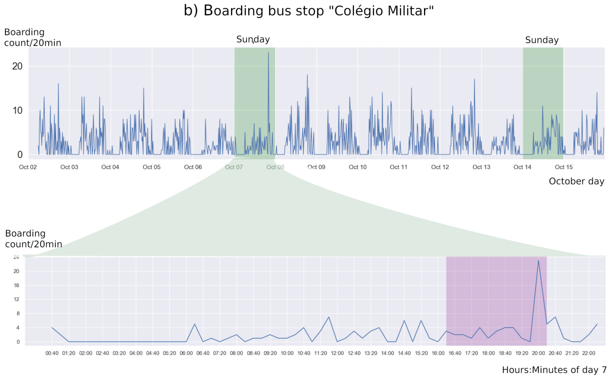


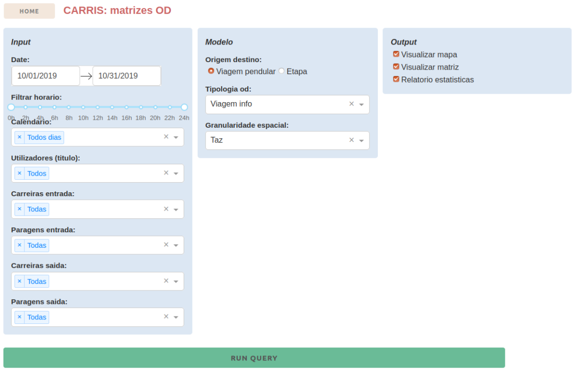
Fig. 11. Effect of the game event on bus stops

Figure 11 is a perfect example that shows the disruptive effect of an event planned in close stops. The 11 (a) provide graphical view of passenger boarding in the 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 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 and 50 minutes after of the end of the game.

V. Tool Visualization

This section presents the visualization tool implemented in the ILU project's scope, with the support of CARRIS and the Municipal Council of Lisbon. This tool was developed in Python with Plotly and Dash packages, and it allows the presentation of origin and destination matrices and complementary relevant information.



CARRIS: matrizes OD

Input

Date: 10/01/2019 → 10/31/2019

Filtrar horario:

Calendario: × 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

Fig. 12. Interface for matrix display

Figure 12 corresponds to the graphical interface responsible for parameterizing the matrices. The output elements that can be viewed on the page are: statistical report (fig. 13) ; the matrix (fig. 14); map with routes chosen (fig. 15).

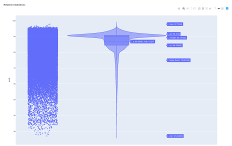


Fig. 13. Violin and stripplot with general matrix assessment

VI. Conclusion

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

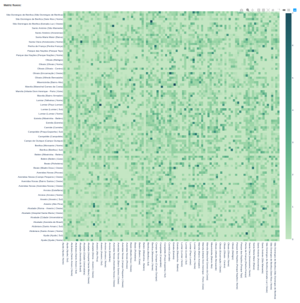


Fig. 14. Graphical visualization of the matrix, showing the journey demand between TAZ's

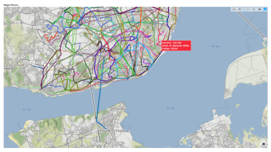


Fig. 15. Map displaying all the routes of CARRIS network

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.

References

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

[2] A. Ceder, "Urban mobility and public transport: future perspectives and review," *International Journal of Urban Sciences*, pp. 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] 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*, vol. 1817, no. 1, pp. 183–187, 2002.

[6] 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*, vol. 2112, no. 1, pp. 53–61, 2009.

[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*, vol. 17, no. 1, pp. 133–142, 2015.

[8] 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*, pp. 502–513, Springer, 2011.

[9] 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*, vol. 22, no. 5, pp. 376–387, 2007.

[10] 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*, vol. 24, pp. 9–18, 2012.

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

[12] J. M. Farzin, "Constructing an automated bus origin–destination matrix using farecard and global positioning system data in sao paulo, brazil," *Transportation research record*, vol. 2072, no. 1, pp. 30–37, 2008.

[13] 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*, vol. 2263, no. 1, pp. 140–150, 2011.

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

[15] 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*, vol. 2343, no. 1, pp. 17–24, 2013.

[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*, vol. 27, pp. 664–671, 2017.

[17] 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.

[18] 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*, vol. 2535, no. 1, pp. 88–96, 2015.

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

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

[21] M. Mamei, N. Bicocchi, M. Lippi, S. Mariani, and F. Zambonelli, "Evaluating origin–destination matrices obtained from cdr data," *Sensors*, vol. 19, no. 20, p. 4470, 2019.

[22] A. Ali, J. Kim, and S. Lee, "Travel behavior analysis using smart card data," *KSCE Journal of Civil Engineering*, vol. 20, no. 4, pp. 1532–1539, 2016.