UTTRI

Research Report

Analysis of Montevideo Smartcard Data, Preliminary Report

Report 3, iCity SOUTH

Catalina Parada Hernandez, Eric J. Miller

September 2017

iCITY-SOUTH:
Urban Informatics for Sustainable Metropolitan Growth in Latin America

REPORT 3:
ANALYSIS OF MONTEVIDEO SMARTCARD DATA, PRELIMINARY REPORT

A report to CAF, the Development Bank of Latin America.

By:

Catalina Parada Hernandez
M.A.Sc. Candidate
Department of Civil Engineering

Eric J. Miller, Ph.D.
Professor, Department of Civil Engineering
Director, UTTRI

September, 2017

UTTRI

EXECUTIVE SUMMARY

The purpose of this report is to investigate public transportation traditional and new data collection methods in Montevideo, Uruguay. The data collected the public transportation system STM (Sistema de Transporte Metropolitano) of Montevideo can be processed to provide powerful tools for planning purposes, monitoring of the system, and understanding of the travel behaviour of public transit users in Montevideo.

This report describes the methodology and presents preliminary results of the analysis and processing of boarding records of smartcard users in public transit. The methodology estimates the alighting locations and daily trips for smartcard users with multiple daily transactions. Due to validation errors, the methodology was applied to only 43% of the smartcard transactions corresponding to 125,401 users.

The success rate for estimating the alighting locations is of 89.5%. For 91,544 smartcard users the methodology estimated all the alighting locations for their daily trips. Using this trip data, Origin-Destination matrices are illustrated for the morning and evening times per Census block group (Segmentos Censal).

This report concludes by outlining improvements in the validation and processing procedures to improve the results and understand temporal variations of passenger behaviour.

ACKNOWLEDGEMENTS

This study was funded by CAF. The unstinting support and patience of CAF, and, in particular, Nicolas Estupiñan and Andres Alcala has been greatly appreciated. The collaboration with Diego Hernandez, Universidad Católica del Uruguay, and Antonio Mauttone, Universidad de la República de Uruguay, has been most welcome and helpful. Verónica Orellano Chiazzaro, Smart Cities Technology group, City of Montevideo, has provided invaluable support in accessing and interpreting the smartcard data. And thanks to Brendan Reilly, University of Toronto - Travel Modelling Group, for technical support and guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF MAPS

LIST OF TABLES

CHAPTER 1: STUDY PURPOSE & MOTIVATION

Urban regions with Latin America (and elsewhere) face enormous challenges in terms of the provision of transportation infrastructure and services to meet the travel needs of their growing population in a cost-effective, equitable and sustainable manner. High quality, comprehensive information concerning travel behaviour and transportation system performance is a fundamental prerequisite for successful urban transportation planning and decision-making to address these pressing, first-order needs.

In recognition of this need, CAF established the Urban Mobility Observatory (OMU, *Observatorio de Movilidad Urbana*)[1] to assemble and utilize standardized transportation-related data for Latin American cities. 29 cities are currently members of OMU. Collecting consistent, time-series data for these cities, however, is a difficult and costly task for CAF and its partner cities.

At the same time, exciting, new transportation data collection sources are emerging to complement or even replace the traditional methods used to collect the OMU data. These include:

- The pervasive penetration of cellphone and smartphone technology within urban populations.
- The widespread adoption of smartcard systems by public transit agencies in many cities.
- Extensive deployment of many types of sensors (video, thermal, Bluetooth, etc.) for monitoring travel flows.
- Increasing availability of very large (typically crowd-sourced) datasets collected in a variety of ways by private sector companies (Google, Waze, Inrix, etc.) that can provide travel information.
- Web-based survey methods to complement/replace traditional survey methods such as home-interviews, telephone interviews, etc.

In 2015, the University of Toronto Transportation Research Institute (UTTRI) launched the *iCity* research program, which is dedicated to applying modern *urban informatics* (the combination of data collection, data science, modelling, visualization and high-performance computing methods) to the promotion of sustainable metropolitan growth. As one component of CAF's strategy for promoting its urban sustainable mobility objectives, it has partnered with UTTRI to create the *iCity-South* research program to apply the *iCity* urban informatics vision and capabilities in Latin American cities.

Two initial projects were chosen to launch the *iCity-South* research program. One involves the demonstration of agent-based microsimulation methods for modelling urban travel demand in terms of developing a prototype microsimulation model for Asunción, Paraguay.[2] The second is investigating traditional and new data collection methods in Montevideo,

[1] https://www.caf.com/es/temas/o/observatorio-de-movilidad-urbana/

[2] This project was completed in April, 2017. See Miller, et al., (2017a, 2017b) for the results of this project.

Uruguay. This report is the third in a series of reports documenting the Montevideo project results.

This report presents quantitative and qualitative analysis of the boarding records for the public transit system of Montevideo and a methodology for estimating alighting locations and destinations of trips made by smartcard users. This report presents the preliminary results of the methodology applied to a sample of smartcard data and provides insights of the capabilities of smartcard data for planning purposes and transit system operation and performance measures.

In addition to this brief introduction, this report consists of 5 chapters that are organized as follows. Chapter 2 summarizes previous work with smartcard data for planning purposes, with a focus on methodologies that are relevant to this report and to the data available for Montevideo. Chapter 3 then presents a quantitative description of all the data and a thorough analysis of the data as of August 20, 2017. Chapter 4 describes the procedures for processing and validating data that are used to estimate the alighting locations for smartcard transactions. Moreover, this chapter contains the results of the alighting estimation methodology. Lastly, Chapter 5 concludes the report by outlining improvements for validation procedures and the methodology.

CHAPTER 2: LITERATURE REVIEW

Many cities and regions have adopted smartcard systems that have benefits for the public transport operators and the passengers. Smartcard systems promote efficiency in fare collection (Trépanier, Tranchant , & Chapleau, 2007) and are a convenient method of payment for passengers. Smartcard systems passively collect details of smartcard transactions and this information is useful to transportation planners as it contains a large sample of transit trips every day (Hickman, 2017). This data has a variety of uses, including serving short- and long-term planning strategies and complementing transit system operation, development, and evaluation strategies (Schmocker, Kurauchi, & Shimamoto, 2017).

This literature review includes prior work by researchers on transportation systems that we aim to improve and apply to Montevideo. The first part of this literature review describes methodologies proposed to identify the destinations of public transit users based on their smartcard transactions. The approaches included on this part are described in detail as they pertain to the methodology and results presented on this report.

The second and third parts include an overview of previous work that has been done as post-processing of the data from the methodologies discussed in the first part or available from transportation systems where users validate their card as they alight, to incorporate smartcard data with survey data, monitor operations of the system, and understand passenger behavior.

2.1 SMARTCARD DATA FOR ESTIMATION OF PUBLIC TRANSIT TRIP DESTINATIONS

One of the most interesting applications of smartcard data for transportation planning is the determination of Origin-Destination (OD) matrices for public transit. For public transit systems where passengers only validate their card while boarding (tap-on systems), researchers have proposed methodologies for estimating alighting locations (Trépanier, Tranchant , & Chapleau, 2007) and creating transit OD matrices (Munizaga & Palma, 2012). Trépanier et al. (2007) estimate the alighting stop for a passenger by identifying the stop of the route that is closest to the boarding stop on the subsequent route the passenger takes, as illustrated in Figure 1.

In addition to identifying the alighting stop, transfers and destinations are distinguished based on the time and distance between transactions. For smartcards with single daily transactions, Trépanier et al. (2007) inspect previous trips of the card that have similar boarding location and time to the single trip and for which alighting location can be identified, to assign the alighting stop to the single trip. The application of this methodology estimated 66% alighting locations and 80% in peak hours.

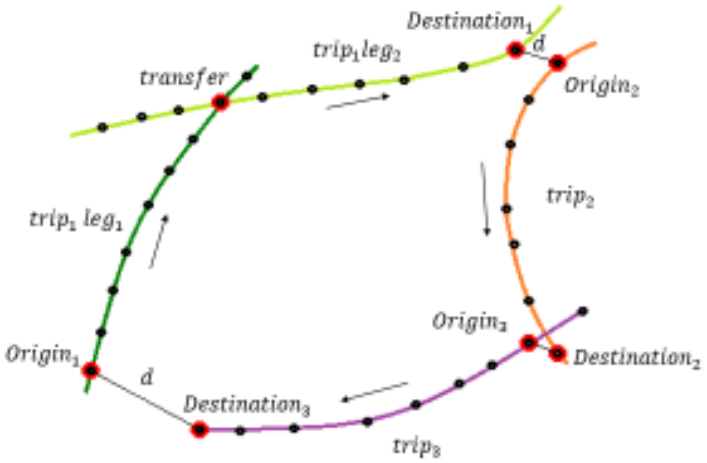


Figure 1.Trips destinations and legs of trips.

This methodology was further developed by Munizaga & Palma (2012) to be implemented on multimodal transit systems and create OD matrices. The major contributions are estimating the alighting location by minimizing the generalized time (on-board and walking time) instead of distance between alighting stop and next boarding, and building OD matrices with this data.

Over 80% of the alighting locations could be estimated and used to create OD matrices; these matrices can be aggregated at any level as the boarding and alighting data is on the disaggregate stop level. Munizaga & Palma (2012) built OD matrices for 6 zones in Santiago de Chile (North, West, East, Downtown, South-East) and applied expansion factors to account for transactions for which there was missing information regarding the boarding and/or alighting.

This methodology was later validated using three data sources: the smartcard data used in the methodology, an OD survey for metro users, and a group of volunteers (Munizaga, Devillaine, Navarrete, & Silva, 2014). This validation revelealed that the methodology proposed correctly estimates 84.2% of alighting locations and distinguishes 90% of the legs of trips from trips.

Based on these results, Munizaga et al. (2014) propose four improvements to the methology: allowing a walking distance greater than 1 kilometre between the alighting location and the next boarding, considering the start of a day at the time period with the lowest transactions (4:00:00 am for this case) instead of midnight, estimating the alighting location for single day transactions by using the subsequent day trips, and recognizing separate trips by comparing the Euclidean distance between the board and alight stops with the on-route distance travelled. This last proposition allows to identify trips that were previously considered as legs part of the same trip, but that are separate trips instead. This is because the on-route distance travelled exceeds the Euclidean distance, as shown in Figure 2.

Figure 2. A three-legged trip comparison between Euclidean and on-route distance (from Munizaga et al., 2014)

2.2 APPLICATION OF SMARTCARD DATA IN TRANSIT OPERATION AND PERFORMANCE MEASURES

Data about passenger boarding and alighting locations at the stop and route level that is obtained from the estimation procedures previously discussed can be used for a myriad of operations and performance measures. Some of these include creating load profiles of individual buses and bus routes (Trépanier et al., 2007) (Beltran et al., 2011), identifying spatiotemporal demand variations of bus routes, analyzing on-route travel times and distances (Trepanier, Morency, & Agard, 2009), and recognizing transfer points, volumes, and transfer times for passengers (Jang, 2010). These measures can be aggregated at any spatial and temporal level to monitor, evaluate, and/or propose improvements to the transit network.

Moreover, by merging smartcard data with scheduling and AVL (Automated Vehicle Location) data it is possible to compute commercial speeds (Beltran et al., 2011) (Trepanier et al., 2009), identify headway variation (Beltran et al., 2011), schedule adherence and passenger performance indicators (Trepanier et al., 2009). The latter indicators include trip duration, distance, and speed, can only be calculated to the passenger level using smartcard data. These operation and performance measures can be analyzed per car type (e.g. student) and for any area and/or time of the day (Trepanier et al., 2009).

Note that these measures can be computed with confidence for transactions using smartcards, but cannot be applied for passengers without cards without understanding their travel behaviour first. Smartcard users could have very different travel behaviours than no-card users depending on the fare structure and incentives available to smartcard users. As the incentives differ among transportation systems (Schmocker, Kurauchi, & Shimamoto, 2017), the travel behaviour for smartcard and non-card passengers should be compared or studied independently to prevent obtaining biased results (Park, Kim, & Lim, 2008).

2.3 DATA FUSION OF SMARTCARD DATA WITH TRANSPORTATION SURVEYS

The household survey data from the Encuesta de Movilidad Area Metropolitanda de Montevideo 2016 can be used to validate and/or apply data fusion techniques with the smartcard data. Hickman (2017) highlights the need for integrating smartcard data with household surveys as only few authors have integrated smartcard data with surveys and travel diaries.

Kusakabe & Asakura (2017) estimate trip purposes for rail smartcard data by fusing this data with survey data using a Naïve Bayes classifier. The integration of data sets was based on boarding and alighting locations and times. This method correctly identified over 80% of the communiting and home trips but only over 20% of leisure trips; this is expected as leisure trips are less common and often underreported in surveys.

Another application of data fusion between trip data and smartcard data consists of evaluating the information provided in surveys about public transit usage. Spurr et al. (2015) proposed matching smartcard data with household travel survey data using spatiotemporal windows regarding boarding and alighting times and locations, as well as line numbers and subway stations. With this approach and a sample of survey responses, the daily journeys of 50% survey respondants that declared using public transit could be paired with at least one smartcard. The 50% paired journeys comprise three matching scenarios: exact matches, partial matches with undereporting of trips, and match with typical daily travel patterns instead of the day asked on the survey.

This results are fairly similar to those obtained by Riegel (2013). The difference resides in that Riegel (2013) obtained the smartcard ID linked to survey respondants volunteers and could pair exact survey responses to the transactions of a specific smartcard. For this study, there were only 44% exact matches between reported daily trips and the smartcard data for the card IDs.

CHAPTER 3: DATA

The data was provided by the Intendencia de Montevideo, the governmental agency that monitors, coordinates, and integrates the public transportation system in the metropolitan area of Montevideo, Uruguay. The integrated transportation system STM (Sistema de Transporte Metropolitano), is composed of buses from 4 different operators that serve the city of Montevideo and surrounding areas (Coetc, Comesa, Cutcsa, Ucot). This system has 144 bus lines with 107 different destinations, and 4,835 stops. [3]
There are 4 main components of the data:

1. Boarding records: 7 consecutive days of passenger boarding records, including the five weekdays and a weekend from August 15th to August 21st , 2016. These records belong to smartcard (STM card) and no-card passengers recorded by the system.
2. Lines and branches: Information about bus routes including the direction and order of stops. Each bus run or trajectory in one direction, is labeled with a unique identification number that can be paired with this data to obtain the run's line and branch.
3. Stops: Number, coordinates, and description of the closest intersection from the stop.
4. Automatic Vehicle Location (AVL): Position and speed of 295 bus runs without timestamps.

This chapter provides qualitative and quantitative descriptions of the boarding records. This section begins by providing an overall description of the 7 days of data and explaining the differences between trips made with STM cards and without them. And then it closes with an in-depth analysis of the data for Monday, processed for overall understanding of travel patterns and temporal distribution of trips.

3.1 DATA DESCRIPTION

The passenger boarding records correspond to smartcard and non-smartcard users during a complete week (Monday-Sunday). The total boarding records for smartcards is 5,077,674 and for no cards is 2,371,815, representing a 68% to 32% split.

Table 1 shows the volumes and some descriptive statistics for the boarding records. For smartcards, the weekday average is 867,269 with Thursday having the highest volume of 872,403 records. The weekend has significantly lower volumes with 455,977 records on Saturday and 285,351 on Sunday. For records with no cards, the weekday average is 394,640 with Monday having the highest volume of 401,328 records. The weekend has significantly lower volumes with 240,139 on Saturday and 158,472 on Sunday.

[3] http://www.montevideo.gub.uy/transito-y-transporte/stm-sistema-de-transporte-metropolitano/el-sistema

Table 1. Boarding records and descriptive statistics for August 15-21, 2016

	Boarding records	
Weekdays	Smartcard	No card
Monday	866,469	401,328
Tuesday	869,439	392,366
Wednesday	868,190	388,364
Thursday	872,403	395,202
Friday	859,845	395,944
Weekday total	4,336,346	1,973,204
Average	867,269.2	394,640.8
Standard deviation	4,681.6	4,777.7

	Boarding records	
Weekend	Smartcard	No card
Saturday	455,977	240,139
Sunday	285,351	158,472
Weekend total	741,328	398,611
Average	370,664	199,306
Standard deviation	120,650.8	57,747.3
Week total	5,077,674	2,371,815

There are several differences for the passengers that use a smartcard and those who do not. Smartcard users benefit from being able to transfer between buses within 1 hour or 2 hours, depending on the trip type they choose and they also pay reduced fares. Smartcard users can use their card for people they travel with and benefit from fares and transfers between buses, as long as they travel together. This is a unique characteristic of this transportation system, as most of the transportation systems with smartcards permit only one card per person. On the other side, passengers without cards cannot make transfers and pay higher fares than the users that have smartcards.

The passengers that do not have cards pay the fare as they board the bus and the system records the time of boarding, ticket number, boarding stop, bus run unique identification number and bus destination, fare details, and number of passengers. The users that have smartcards tap their STM card on readers that are mounted on the buses and the system records the number of the card, time of boarding, boarding stop, bus run unique identification number and bus destination, fare details, card type and fare discount if applicable, ordinal of trip, and whether the tap is considered a transfer (ordinal of trip≥1) or a new trip (ordinal of trip=1).

Furthermore, the system records the transactions considered as part of the same trip (trip with 2 or more trip legs) and assigns them a common trip ID. This information is essential to understand the methodology proposed in Section 4.3. There is no record of where passengers alight as the system is designed for tap-on only.

For smartcard users the fare discounts are associated to the different card types. These types distinguish ordinary users from other user groups that benefit from reduced or subsidized trip fares (see Appendix A).

Table 2 shows the boarding records for each smartcard type on Monday August 15. Note the high percentage of boarding records made by students (Student A and Student Free).

Table 2. Boardings per STM card type for Monday, August 15th

STM card type	Boardings	Percentage
Standard	397,034	45.8%
Student A	170,134	19.6%
Student B	21,448	2.5%
Student Free	142,712	16.5%
Retired A	44,317	5.1%
Retired B	16,235	1.9%
Social Work	29,330	3.4%
Prepaid	23,608	2.7%
Others	21,651	2.5%

3.2 DATA ANALYSIS FOR AUGUST 15

For the subsequent contents of this report, the data corresponding to Monday, August 15th is used. This is done with the aim of providing an in-depth description and validation of daily data, testing procedures and assumptions, and developing methodologies that can be used for any other day. For this selected day, there are 1,267,798 records with a split of 68% to 32%, corresponding to 866,469 and 401,328 STM card and no card records respectively. Moreover, the smartcard records correspond to 302,516 STM cards with an average of 2.86 transactions per card.

Data is processed for overall understanding of travel patterns and temporal distribution of trips. Smartcard and no card data is processed separately to identify differences in travel patterns; moreover, the smartcard users can be analyzed according to the card type. The temporal distributions in Figure 3 and Figure 4 aggregated by 30 minute intervals reveal interesting and different travel patterns for smartcard and no card transactions.

There are three evident peak times for STM cards between 7 am and 8 am, 1 pm and 2 pm, and 5:30 pm and 6:30 pm. Interestingly, the midday peak exceeds the morning and evening peak volumes and the volumes after this peak are similar or higher than morning volumes until 7 pm.

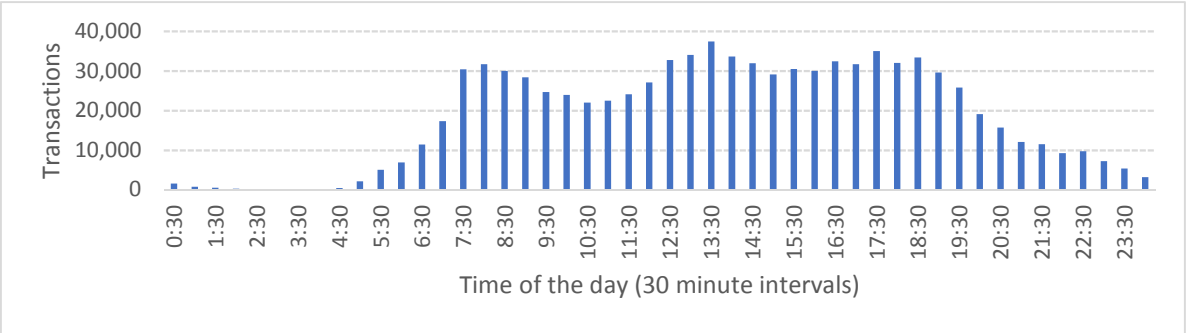


Figure 3. Temporal distribution of STM card transactions

On the other hand, for no card transactions there are two evident peaks between 8 am and 9 am, and 5:30 pm and 6:30 pm. There is no noticeable midday peak, instead there are high transaction volumes starting at 12:30 pm until the evening peak. The volumes at midday and evening times are relatively higher than morning ones.

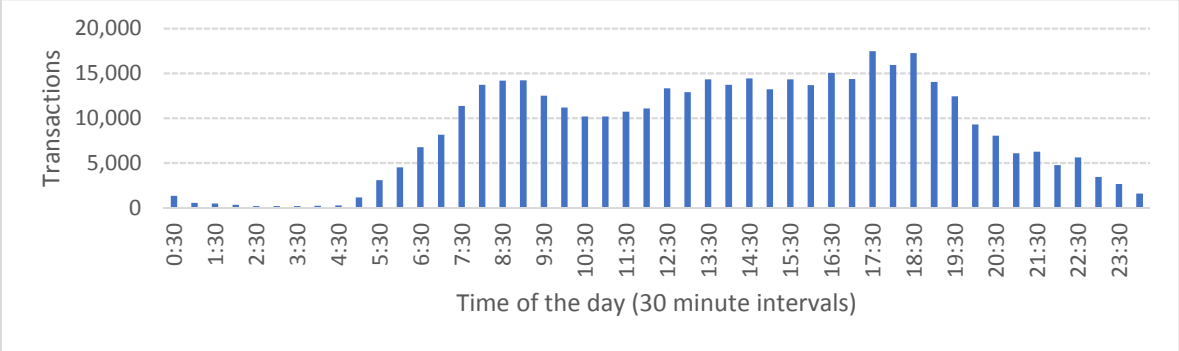


Figure 4. Temporal distribution of transactions without card

These distributions are compared for statistical similarity using the Kolmogorov-Smirnov test. Using a 90% confidence level, the hypothesis that the distributions are similar can be rejected($D_n = 0.74$).

From the previous analysis, the transactions are aggregated into four time periods that differentiate volumes between the peaks: AM from 4 am to 11 am, Midday from 11 am to 3:30 pm, PM from 3:30 pm to 10pm, and Overnight from 10pm to 4 am. The midday period is short (4.5 hours) compared to the other three, to prevent including typical morning home-to-work and evening work-to-home trips. And even though it is short, Figure 5 and Figure 6 illustrate that almost a third of daily transactions occur during Midday.

The total number of passengers boarding buses with STM cards is 884,018 and these are shown per time period in Figure 5 with the highest volume occurring during the PM period (317,422) followed by Midday (280,454). Note that this number exceeds by 17,549 the number of STM cards boarding records. As previously discussed, this occurs as smartcard users can use their cards for the trips of others.

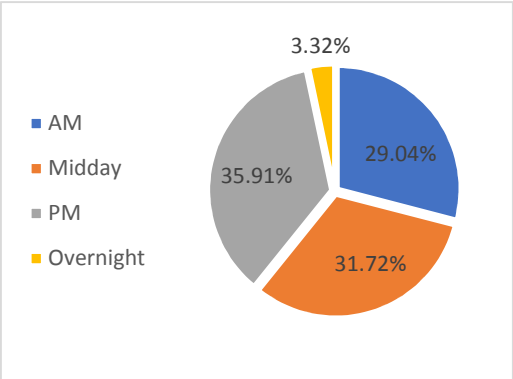


Figure 5. Transactions for STM cards per time period

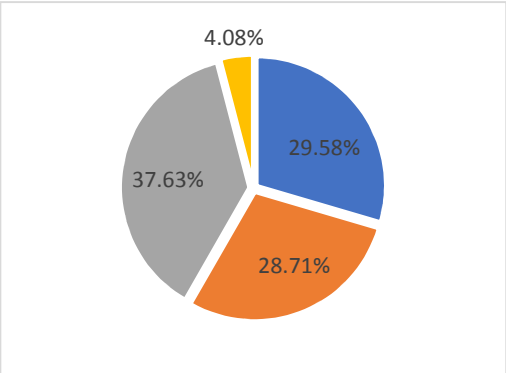


Figure 6. Transactions for no cards per time period

The total passenger transactions without STM cards is 411,156 and are distributed in the four time periods as portrayed in Figure 6. The highest volume occurs during the PM hours (154,722) followed by AM volumes (121,605).

In addition to the temporal travel pattern analysis, for STM card users the daily transactions and transfers per card can be identified. Figure 7 shows the transactions per card. Just above half of the cards (53.7%) have one or two transactions per day and 99.6% of the cards make 9 or less transactions on this day.

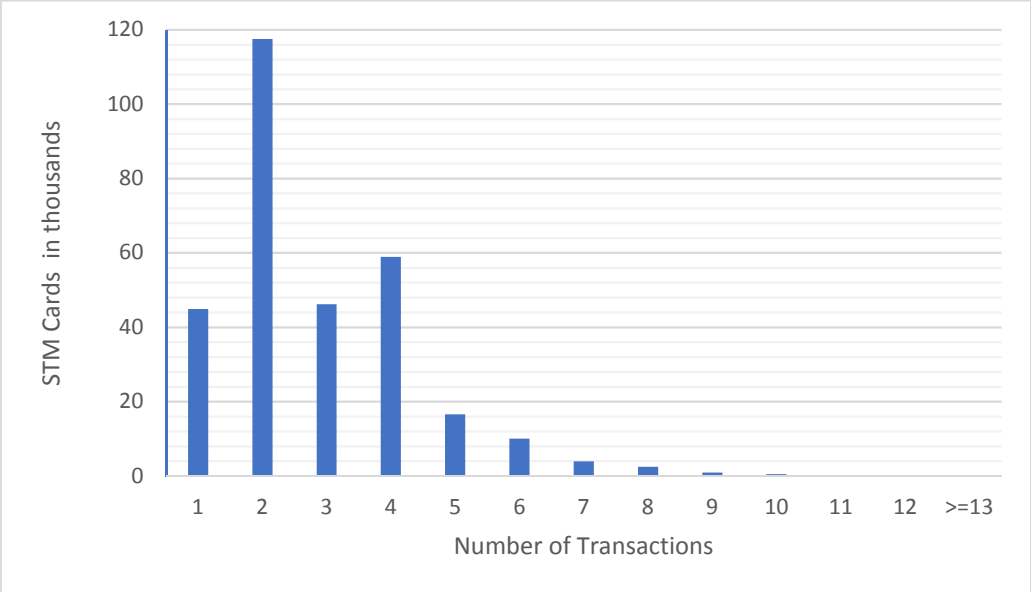


Figure 7. Transactions per STM card

85% of STM cards have more than one transaction and for these cards, the transfers per card are shown in Figure 8. 94.3% of the users transfer 1 or 2 times per day and 99.6% of the cards make 4 or less transfers.

Figure 8. Transfers per STM card

CHAPTER 4: METHODOLOGY FOR ORIGIN AND DESTIMATION ESTIMATION

The passenger transactions recorded by the system include the boarding location and time of passengers, but their alighting location and time are unknown. This section presents the preparation, processing, and methodology applied to the smartcard data to estimate the alighting locations and times of the transactions.

This chapter contains the methodology and main results of this report and is organized as follows: First, all the transactions (smartcard and no card) are used to create itineraries for the bus runs; second, the smartcard data undergoes cleaning and validation procedures, and lastly, the methodology is proposed and applied with successful preliminary results.

4.1 ITINERARIES FROM BOARDING DATA

Schedules are used to determine the time buses arrive at a certain location, which in turn can be used to estimate the alighting times for passengers at that location. In the absence of schedule data, the itineraries can be created using the data available. The passenger boarding records and the characteristics of the bus routes (lines and branches) are used for this purpose.

Each bus run has a unique identification number (UID) that is attached to the passenger transactions when they board the bus. Both, smartcard and no card records, can then be grouped by UID and stop number to build the itinerary for each individual bus run. As all the boarding records have valid UIDs and boarding information, the 1,267,797 records are used for this purpose.

For stops with multiple boardings the time of arrival is taken as the average boarding time. Also as alighting is not recorded, the time of arrival at a stop is used as the alighting time. This is similar to the approach used by (Trépanier, Tranchant , & Chapleau, 2007) but instead of using monthly boarding records, only the daily records are used to determine the time of arrival.

Because passengers might not board at every stop on a route, the itineraries created from passenger records are concatenated with the sequence of all the stops assigned to the bus route based on its line and branch number. The sequences of stops reveal that there are many intermediate stops for which there are not arrival times. Table 3 shows an example of the built itinerary for a bus route with some stops that do not have arrival times highlighted in blue.

The arrival time for these stops are calculated by using simple interpolation between previous and subsequent stops that have arrival times as shown in the column "interpolated arrival time" in Table 3. Interpolation is only applied between the first and last stops for which arrival time is available. This interpolation technique is adapted from the technique used by Fourie, et al. (2017) to reconstruct bus trayectories using smartcard and GPS data.

Table 3. Sample itinerary built from passenger transactions

UID	Branch	Line	Stop ID	Arrival time	Interpolated arrival time	Stop Ordinal
1.653E+10	1763	130	2521	4:13:26 PM	4:13:26 PM	1
1.653E+10	1763	130	6153	0	4:14:22 PM	2
1.653E+10	1763	130	2522	4:15:17 PM	4:15:17 PM	3
1.653E+10	1763	130	2523	4:15:34 PM	4:15:34 PM	4
1.653E+10	1763	130	2524	4:16:40 PM	4:16:40 PM	5
1.653E+10	1763	130	2022	4:17:25 PM	4:17:25 PM	6
1.653E+10	1763	130	2023	4:18:29 PM	4:18:29 PM	7
1.653E+10	1763	130	2525	4:19:39 PM	4:19:39 PM	8
1.653E+10	1763	130	2526	0	4:20:32 PM	9
1.653E+10	1763	130	2527	4:21:24 PM	4:21:24 PM	10
1.653E+10	1763	130	2528	0	4:21:55 PM	11
1.653E+10	1763	130	2529	0	4:22:25 PM	12
1.653E+10	1763	130	2530	0	4:22:56 PM	13
1.653E+10	1763	130	2531	0	4:23:26 PM	14
1.653E+10	1763	130	2532	0	4:23:57 PM	15
1.653E+10	1763	130	2533	0	4:24:28 PM	16
1.653E+10	1763	130	2534	0	4:24:58 PM	17
1.653E+10	1763	130	2535	4:25:29 PM	4:25:29 PM	18

An algorithm was developed using Spyder (Python 3.6) to build and output these itineraries in a CSV file format. The algorithm's output raised an issue related to the concatenation of the itineraries created from passenger records with the information from the bus routes.

In theory, each bus run UID can be paired with a branch number that has a sequence of stops. However, the current data available for lines and branches only contains the stop sequences for 466 of the 918 bus branches that can be retrieved from the passengers boarding records. For the remaining 452 bus runs, considered here as invalid runs, the following methodology is proposed:

Methodology for invalid bus runs

The goal of this methodology is to determine if the invalid bus runs can be matched with any valid run that contains all the stops in the invalid run. This is done by grouping the stops where passengers board the bus for each invalid run and determining which of the valid runs contain all the stops of the invalid run. Figure 9 shows the stops associated with an invalid bus run and the valid run that can be assigned to it. As there is **one** valid run that contains all the stops from the invalid run, the characteristics of the valid run (line, branch, and stop sequence) are assigned to the invalid run.

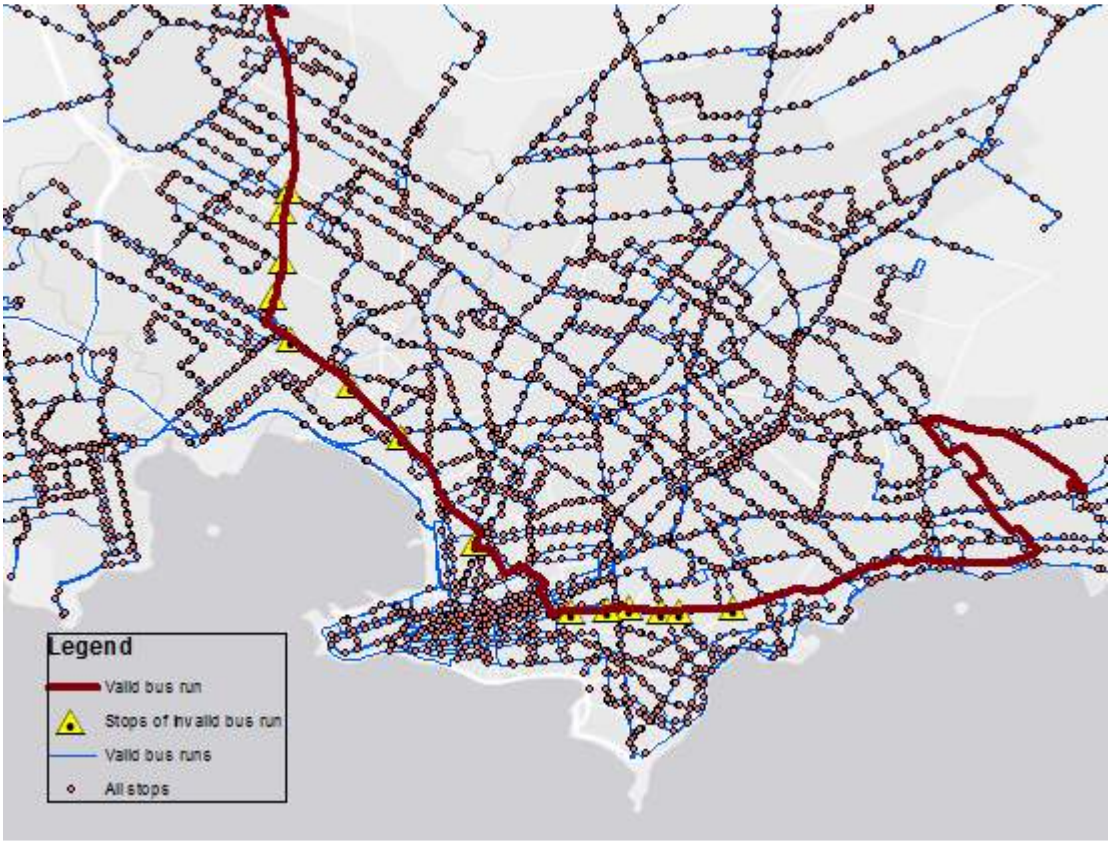


Figure 9. Valid bus run assigned to invalid run

If the invalid run can be matched with one, **and only one**, of the valid runs, the invalid run is given the information of the respective valid run. The condition of pairing an invalid run with only one valid run allows to identify with certainty the run and discard the runs for which passengers board in a few stops that are common to many valid runs.

This methodology was applied to the 452 invalid bus runs and for 125 of them a valid bus run could be identified. Further data is required to be able to identify the remaining runs.

4.2 DATA CLEANING AND VALIDATION

The data needs to undergo a process to remove invalid records and cards with abnormal travel behaviour. Due to the differences between smartcard and non-smartcard records, the cleaning process differs. For non-smartcard records, the only records that can be removed are those that are null. For these records 0.4% are null, leaving **399,834** boarding records to work with.

The smartcard data is queried based on the criteria observed and described in Section 3.2. The query criteria are included in Table 4

Table 4. Query criteria for smartcard data

Query criteria	Value
Void	No
Number of transfers	<=9
Number of passengers	<5
Transactions per day	≠1 or <=9

The query criteria for number of transfers accounts for 99.6% of the data. The criteria for the transactions represents standard user travelling behaviour and removes records that are either abnormal (more than 9 boardings per day) or need to be processed as unlinked trips (one boarding per day). [4] 85% of the records on August 15 meet the query criteria. Of the records that did not meet the criteria, 14% correspond to cards with only one daily transaction.

These records are then validated with the bus route data to identify and remove the records with invalid stop numbers (e.g. stop does not belong to a valid bus run) or invalid bus runs. As this report is written, the validated runs are being assessed and have not been incorporated into this process; therefore, the smartcard records that have at least one boarding on one of the 452 invalid bus runs must be removed. From this validation only **346,645** of the **800,519** (43%) smartcard transactions corresponding to 125,401 users can be used for the methodology described on the following section.

4.3 METHODOLOGY

The methodology has two goals: 1. Estimate the alighting locations and times for STM card transactions 2. Identify the origin and destination of trips for STM users. These are similar to the goals proposed by Trépanier, et al. (2007) and Munizaga & Palma (2012); therefore, the methodology of this section is also similar to the methodologies proposed by these researchers. The difference of this methodology is the availability and quality of data that is reflected in the results.

[4] To process cards with a single daily transaction it is necessary to analyze weekly or monthly travel patterns as proposed by Trépanier, et al. (2007).

First, some terms are defined to help understand the goals of this methodology:

- A trip is defined as the travel from an origin (e.g. home) to a destination for a specific purpose (e.g. work).
- Trips can have one or multiple legs, identified by the transfers between bus services, and can have a walking portion during the transfers.
- The daily trips made by a smartcard user that start and end around the same location constitute a tour.

The STM card transactions can be either trips or legs of trips. These are differentiated by the trip ordinal and the trip ID fields assigned by the system. Transactions that are trips have unique trip IDs that are not shared with any other transactions; while the transactions that are legs of trips share trip IDs with the other legs of the trip (transactions) and their ordinals of trip are labeled chronologically with an ordinal of 1 for the first trip leg and so on. Figure 10 shows a schematic example of trips, legs of trips, and a tour for a smartcard, where the variables and indices refer to:

$n = Trip\ number\ (The\ first\ trip\ is\ n = 1)$
$l = Leg\ of\ trip\ number$
$O_n = Origin\ of\ trip\ n$
$D_n = Destination\ of\ trip\ n$
$a_{n,\ l} = alighting\ location\ for\ trip\ n, leg\ of\ trip\ l$
$b_{n,\ l} = boarding\ loccation\ for\ trip\ n, leg\ of\ trip\ l$
$d = distance\ between\ stops$
$\rightarrow Direction\ of\ travel\ and\ sequence\ of\ stops\ for\ a\ bus\ run$

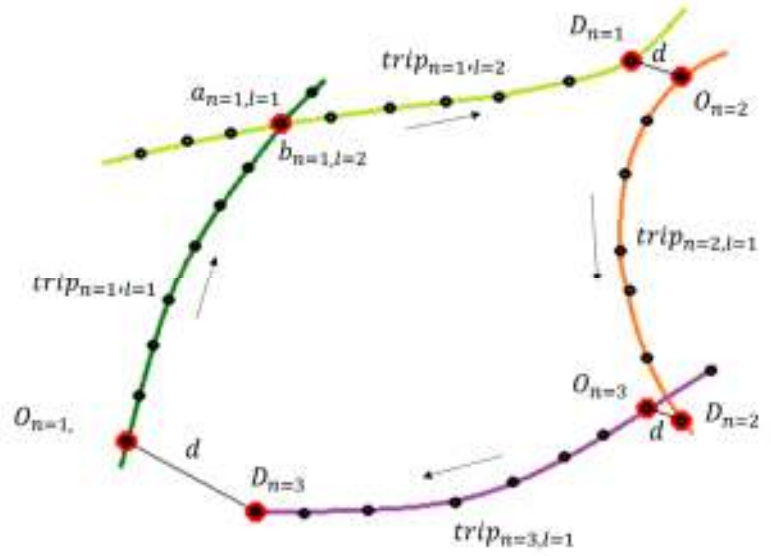


Figure 10. Schematic example of transactions for a smartcard.

From Figure 10 one can infer the data needed to estimate the alighting location for transactions: the boarding location for the transactions, whether they are trips of legs of trips; the direction and stop sequence for the routes that correspond to the transactions, and the

geographic location of the stops to obtain the distance between alighting and boarding stops. Additionally, the time of alighting can be retrieved from the bus routes itineraries.

The methodology is an algorithm that integrates and organizes these data sources for the transactions of each STM card. For a transaction, the algorithm analyzes which of the subsequent stops of the bus route is closest to the next transaction's boarding stop. The closest stop is estimated as the alighting stop. For the last transaction of the day, the algorithm considers the first boarding stop of the day to estimate the alighting stop for this last transaction. When the alighting stop is estimated the algorithm retrieves the time or arrival of the bus at this stop.

After all the transactions of a STM card are processed the algorithm identifies the origins, destinations, and transfer locations for the trips. For technical details of the algorithm refer to Appendix B.

4.4 RESULTS

Two versions of the algorithm were implemented that differ on the maximum allowable distance[5] between alighting and next boarding stops. The first version allows a 500 metre walking distance while the second version allows a 1000 metre distance. The results for these versions are shown in Table 5. There is an increase of 6.8% in estimation of alighting stops when the walking range increases from 500m to 1000m.

Table 5. Algorithm results

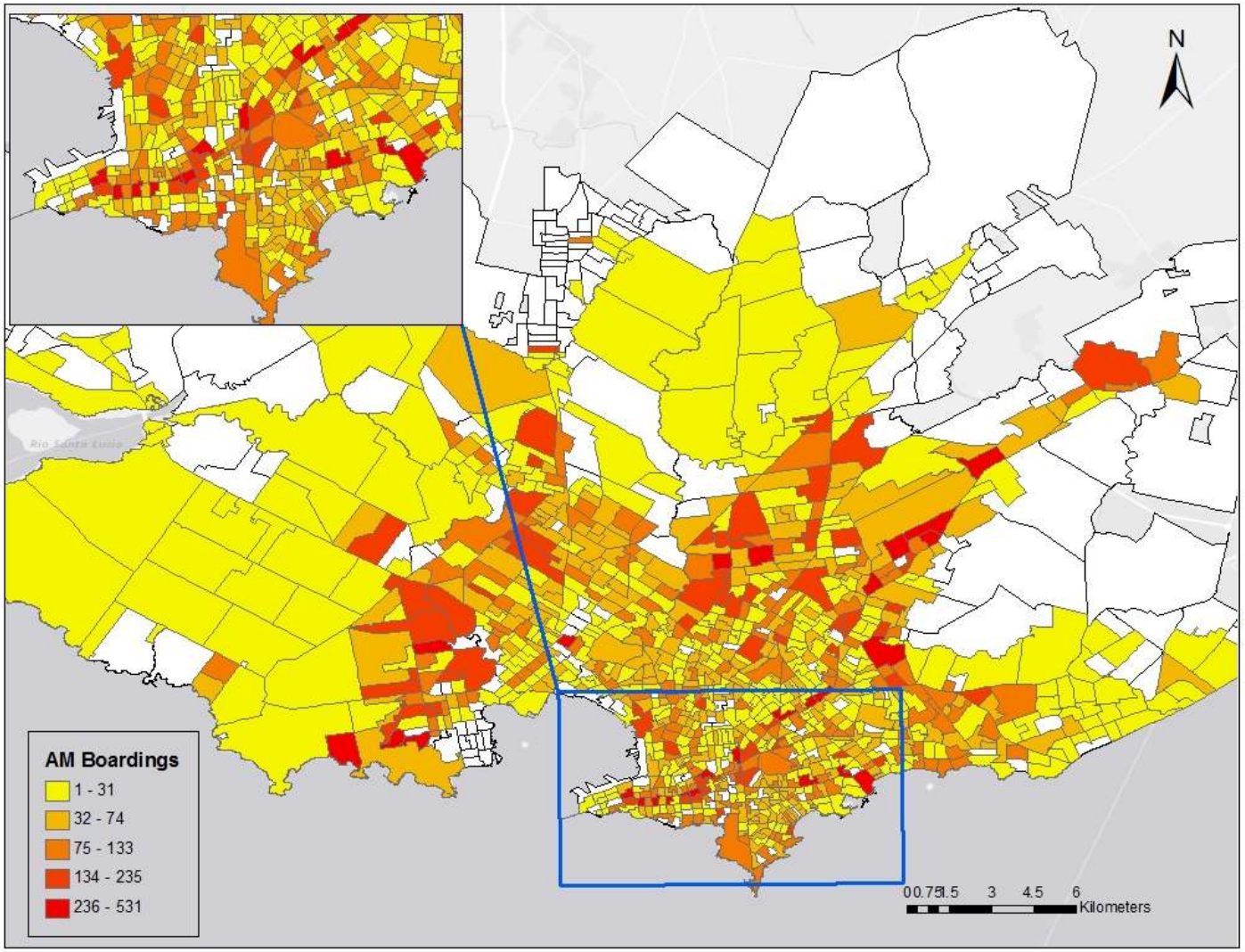
Result Indicators	500m walking range	1000m walking range
Alighting location identification	286,807 transactions (82.7%)	310,497 transactions (89.5%)
Average Euclidean distance between alight and next boarding	91.43m (or 112 m without including 0 m distances)	130.5 m (or 148.2 m without including 0 m distances)
Number of STM Cards with estimated alighting location	73,817 (200,244 transactions)	91,544 (254,053 transactions)
Number of STM Cards with estimated alighting location and time	12,085 (26,982 transactions)	14,840 (33,407 transactions)

The 254,053 transactions of the highlighted cell in Table 5, are used in the following analysis of passenger travel and transfer behaviour. There are 189,034 total passenger trips and 69,180 passenger transfers. These trips and transfers constitute 91,544 tours that correspond to one tour per card.

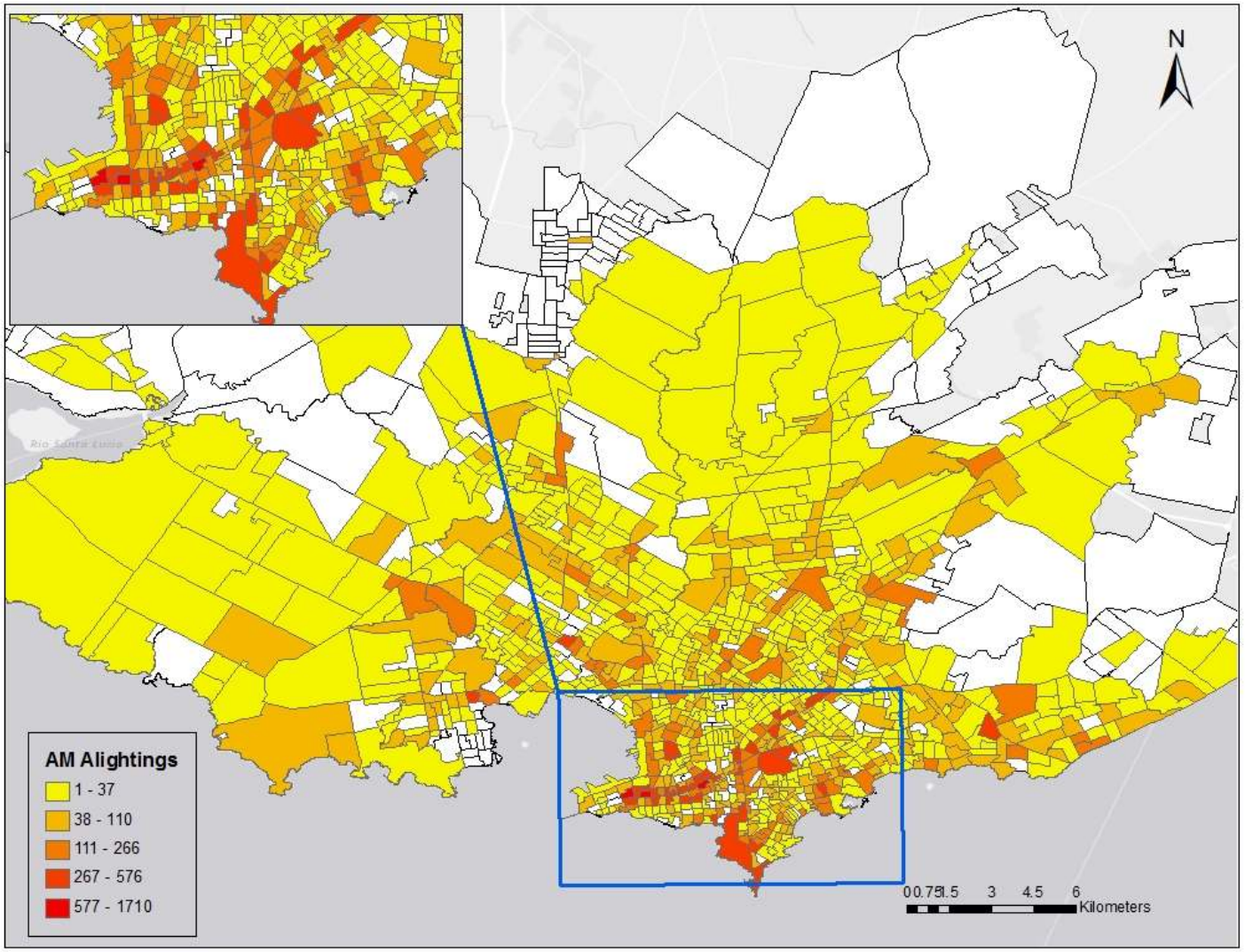
[5] The maximum allowable distance represents actual walking distance determined using the road network for Montevideo on ArcMap 10.2.2

The origins, destinations, and transfers from these transactions can be visualized at any level of spatiotemporal aggregation. In this report, trips are aggregated per Census block groups (Segmentos Censales) and transfers are analyzed at the disaggregate stop level. On the temporal dimension, both trips and transfers are aggregated at the four time periods discussed previously: AM from 4 am to 11 am, Midday from 11 am to 3:30 pm, PM from 3:30 pm to 10pm. Using ArcMap 10.2.2 the following maps are produced:

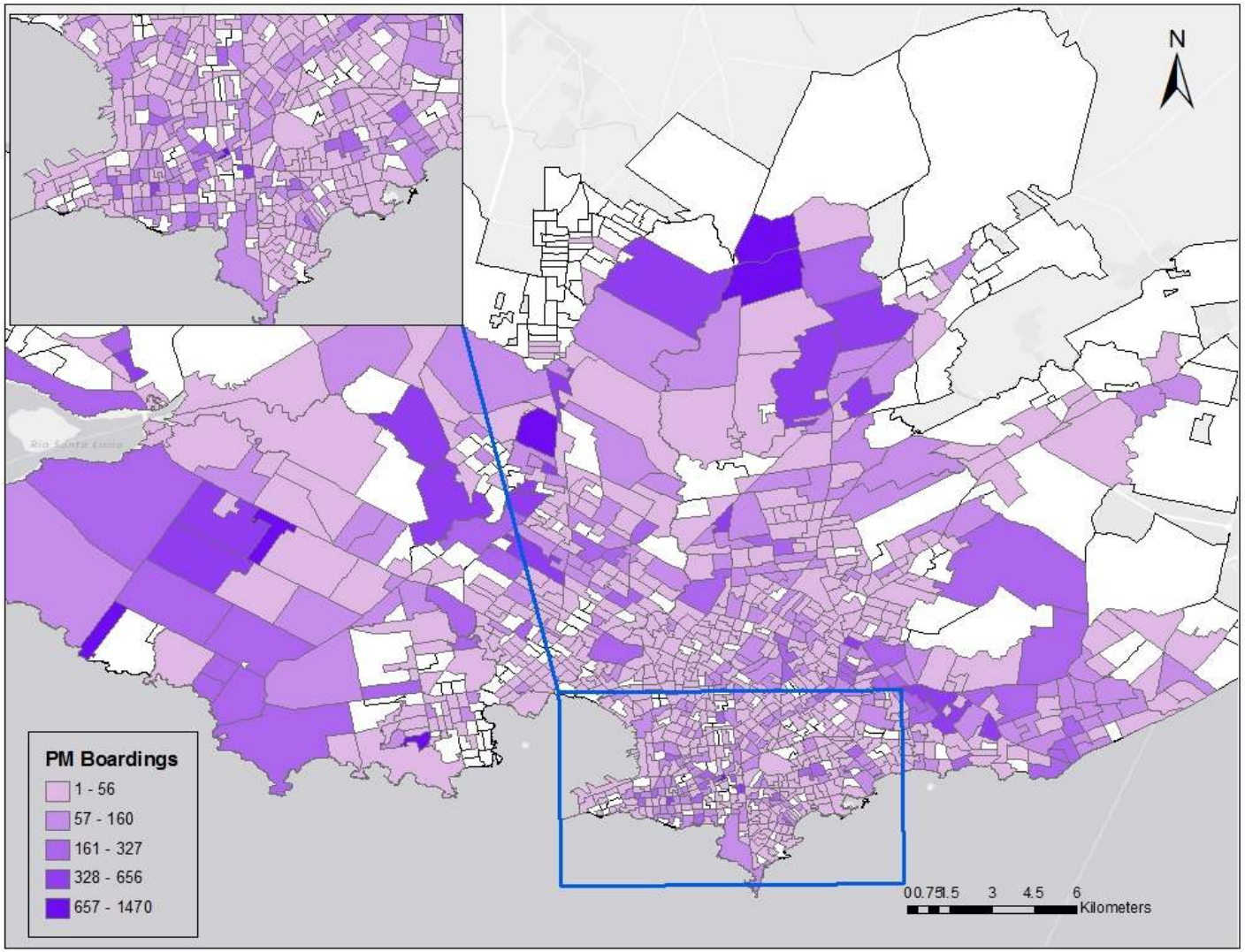
- Map 1- AM Trip Origins
- Map 2 - AM Trip Destinations
- Map 3 - PM Trip Origins
- Map 4 - PM Trip Destinations
- Map 5 - AM Transfers
- Map 6- PM Transfers**Error! Reference source not found.**



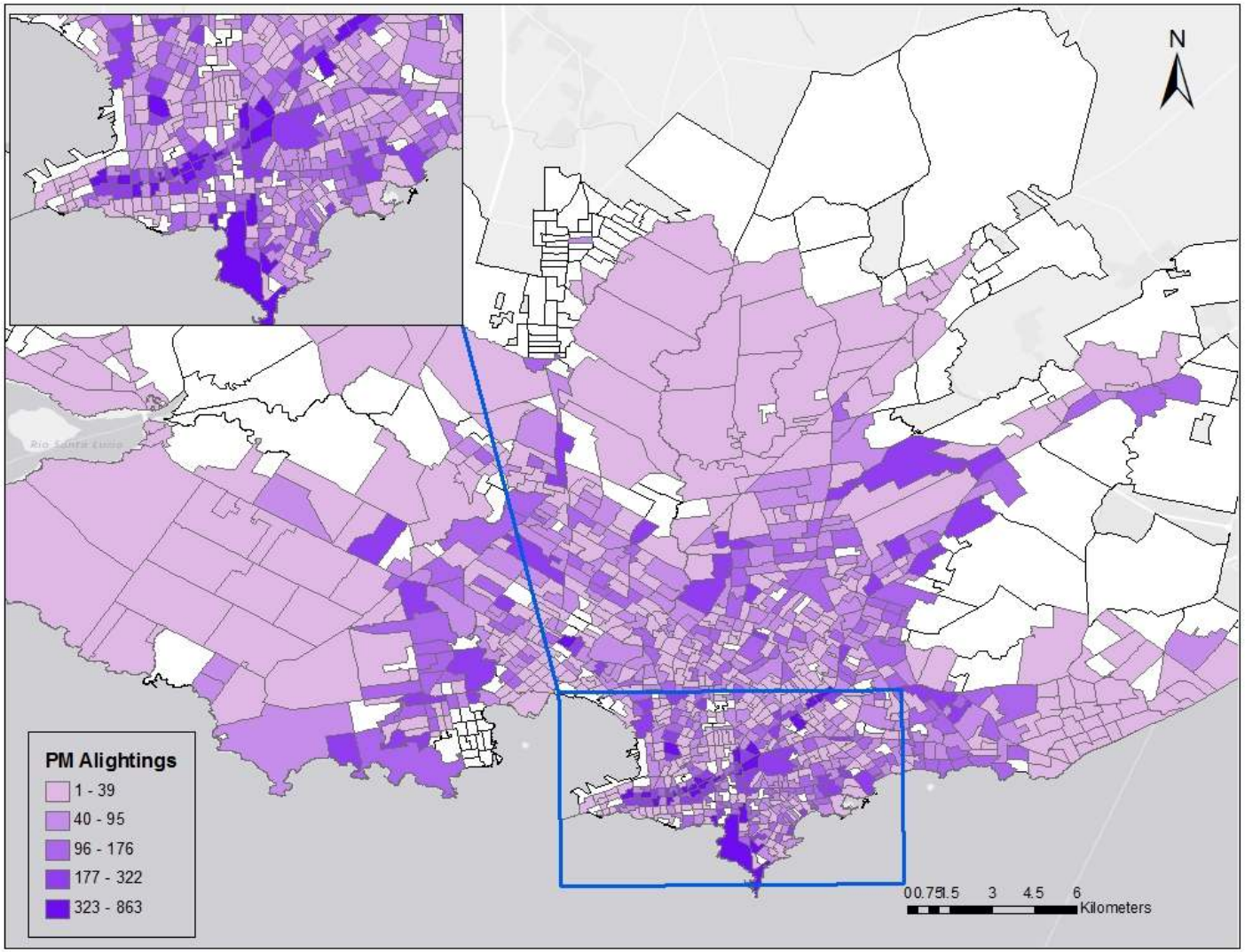
Map 1- AM Trip Origins



Map 2 - AM Trip Destinations



Map 3 - PM Trip Origins



Map 4 - PM Trip Destinations

Map 5 - AM Transfers

Map 6- PM Transfers

The origins of trips in the AM period occur around the urban periphery of the city, as shown in Map 1, and along major axes, such as Avenida de las Instrucciones and Avenida 8 de Octubre. There are also many trip origins within and close to the downtown area. However, the trip destinations exceed the trip origins in the downtown, as can be seen in Map 2. The destination volumes in the downtown and areas surrounding the Avenida 18 de Julio are high with volumes of over 1,000 person-trips per Census block group. While the trip origin volumes are between 300 and 500 person-trips per Census block group in the downtown.

The origins for trips in the PM period, in Map 3, occur across the city, and there are several areas with high volumes on the outskirts of the city. This is interesting, as these areas correspond to Census block groups that are rural areas. An inspection of these using Google maps reveals that there are multiple hotels, industrial parks, sports complexes, and farms in these areas, which could explain the passenger boardings. On the other hand, Map 4 identifies the destinations of the PM trips which are distributed across the city and along the Avenida 18 de Julio.

The transfers during both the AM and PM time periods in Map 5 and Map 6, occur at specific locations: along three major roads, the downtown area, bus terminals, and major stops. The roads with higher transfers are Avenida Agraciada on the West, and Avenida 8 de Octubre and Avenida 18 de Julio on the North-East of downtown. As expected, there are many transfers on the terminals: Terminal Colon, Terminal del Cerro, Terminal Paso de la Arena, Terminal Tres Cruces, and Terminal Belvedere. Moreover, there are stops with high transfer volumes such as the stop at Avenida 8 de Octubre & Comercio and Agraciada & Jose B. Freire. This latter has the highest transfer volume for the AM period.

CHAPTER 5: IMPROVEMENTS

There are many improvements that we are working on to enhance the data analysis, validation, and processing, that will be reflected in more comprehensive results and understanding of travel behaviour. These improvements are discussed here in order of relevance, beginning with the most important ones.

- The OD methodology was applied to only 43% of the smartcard transactions available, mainly due to validation errors (Refer to Sections 4.1 and 4.2 for details). Given the data available, we proposed and applied the methodology to the invalid bus runs, but the results are lower than we expected. This methodology can be improved by including transactions for more days or relaxing the assumptions of allowing only one match.
- The OD methodology does not estimate alighting locations and times for single transactions. To incorporate these transactions, we need to apply the algorithm to transactions for other days with the aim of identifying travel patterns. We have a week of data but access to more data would likely provide better insights of travel patterns.
- There is a significant share of the public transportation system users that do not use STM cards. We need to develop a formal procedure to identify the most likely travel behaviour of these passengers. Once this is accomplished, the trips of these passengers are analyzed with the smartcard trips to create transit OD matrices and compute operation and performance measures for the system.
- The data analysis of this report briefly mentions different STM card types. Data analysis and travel behaviour for the different STM cards (e.g. students) could reveal the peak hours, the busiest bus routes, and other valuable indicators.
- The recommendations from Munizaga et. Al (2014) described on Section 2.3 can be incorporated to the current OD methodology to improve the alighting estimation and trip differentiation methodologies.

By implementing these improvements, it will be possible to validate and/or evaluate the trip survey for public transit users and have complete data to compute operation and performance measures, as described in Sections 2.2 and 2.3.

REFERENCES

Beltran, P., Cortes, C. E., Gschwender, A., Ibarra, R., Munizaga, M., Palma, C., . . . Zuniga, M. (2011). Obtencion de informacion valiosa a partir de datos de Transantiago. *XV Congreso Chileno de Ingenieria de Transporte, 16*, 34-47.

Fourie, P. J., Erath, A., Ordonez, S. A., Charikov, A., & Axhausen, K. W. (2017). Using Smart Card Data for Agent-Based Transport Simulation. En F. Kurauchi, & J.-D. Schmocker, *Public Transport Planning with Smart Card Data* (págs. 133-160). Boca Raton: CRC Press.

Hickman, M. (2017). Transit Origin-Destination Estimation. In F. Kurauchi, & J.-D. Schmocker, *Public Transport Planning with Smart Card Data* (pp. 15-35). Boca Raton: CRC Press.

Jang, W. (2010). Travel Time and Transfer Analysis Using Transit Smart Card Data. *Journal of the Transportation Research Board*(2144), 142-149.

Kusakabe, T., & Asakura, Y. (2017). Combination of Smart Card Data with Person Trip Survey Data. En F. Kurauchi, & J.-D. Schmocker, *Public Transport Planning with Smart Card Data* (págs. 73-92). Boca Raton: CRC Press.

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

Munizaga, M., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smart card data. *Transportation Research Part C*, 70-79.

Park, J. Y., Kim, D.-J., & Lim, Y. (2008). Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea. *Journal of the Transportation Research Board*(2063), 3-9.

Riegel, L. K. (2013). *Utilizing Automatically Collected Smart Card Data to Enhance Travel Demand Surveys (Master's thesis).* Massachusetts Institute of Technology. Cambridge: MIT.

Schmocker, J.-D., Kurauchi, F., & Shimamoto, H. (2017). An Overview on Opportunities and Challenges of Smart Card Data Analysis. En F. Kurauchi, & J.-D. Schmocker, *Public Transport Planning with Smart Card Data* (págs. 1-12). Boca Raton: CRC Press.

Spurr, T., Chu, A., Chapleau, R., & Piche, D. (2015). A smart card transaction "travel diary" to assess the accuracy of the Montréal household travel survey. *Transportation Research Procedia, 11*, 350-364.

Trepanier, M., Morency, C., & Agard, B. (2009). Calculation of Transit Performance Measures Using Smartcard Data. *Journal of Public Transportation, 12*, 79-96.

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

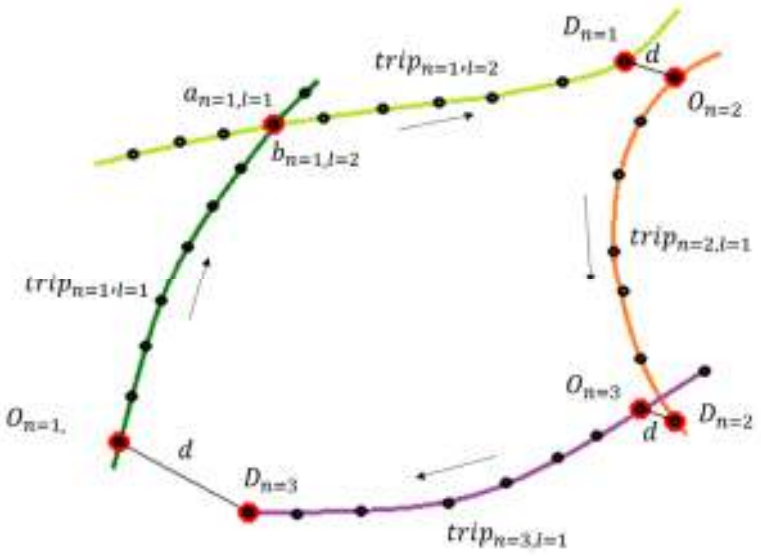
APPENDICES

APPENDIX A: STM Card Types

Group	Description	User Code	Description
1	Normal	01/11	Normal
2	Student	21/121	Student A
		22/122	Student B
		23/123	Student FREE
3	Retired	31/131	Retired A
		21/132	Retired B
4	Social Work	41	Special schools
		42	Social benefits
5	Conventions organisms	51	Entity with quotes
		52	Employee with quotes
		53	Entity without quote validation
		54	Quote without quote validation
		320/07	Ministry of National Defense (Special characteristics)
6	Prepaid	61	Employee of authorized private companies and public organizations
7	Vinculation	71	Employee with quotes
		72	Retired
		73	Investor without quotes
		74	Relative of employee/investor
		75	Employee of transport system

APPENDIX B: Details of algorithm

Recall the definition of variables and indices:
$n = Trip\ number\ (The\ first\ trip\ is\ n = 1)$
$l = Leg\ of\ trip\ number$
$O_n = Origin\ of\ trip\ n$
$D_n = Destination\ of\ trip\ n$
$a_{n,\ l} = alighting\ location\ for\ trip\ n, leg\ of\ trip\ l$
$b_{n,\ l} = boarding\ loccation\ for\ trip\ n, leg\ of\ trip\ l$
$d = distance\ between\ stops$
$\rightarrow Direction\ of\ travel\ and\ sequence\ of\ stops\ for\ a\ bus\ run$



The transactions are identified by the index i, with the first transaction labelled $i = 1$. The algorithm consists of two parts. The steps for each part are outlined as follows:

First part: Estimation of alighting location and time

1. Identify all the transactions for a smartcard and organize them chronologically. Label the transactions as $i, i + 1, \ldots, k$. Starting with i =1 and $k \leq 9$.
2. For transaction i, retrieve the bus UID and match with the corresponding valid bus run to obtain the sequence of stops following the boarding location.
3. Pair the boarding location of the next boarding transaction $(i + 1)$ with one of the stops from step 1 that minimizes the Euclidean distance (d) between the stops and label this as the alighting stop $(a_{n,\ l})$ for transaction i. Thus, minimizing the walking distance for passengers between the alighting stop and the next boarding.[6]

[6] The pairing process can be done by minimizing the distance between alighting and boarding stops (Trépanier, Tranchant , & Chapleau, 2007) or the generalized time (Munizaga & Palma, 2012). The methodology proposed here considers minimizing the distance between stops and sets a maximum walking distance of 500m and 1000m.

4. Retrieve the time of arrival for the alighting stop ($a_{n,\ l}$) from the bus UID itinerary.
5. Repeat steps 2 through 4 for transaction $i = i + 1$ until reaching transaction k*.

*For transaction k, which is the last transaction of the day, use the boarding location for the first transaction of the day ($i = 1$) as the next boarding location for step 3.

Second part: Estimation of trip origin and destination

1. Set variables $n = 1$, $l = 1$, count=0
2. Identify trip IDs for the transactions for a smartcard:
    a. If transaction i has unique trip ID:
        i. Assign label n
        ii. O_n = Boarding stop transaction i
        iii. D_n = Alighting stop transaction i
    b. If transaction i shares trip ID with transaction $i + 1$:
        i. Retrieve and count subsequent transactions with shared trip ID and assign them label n. Assign label l for the first transaction, $l + 1$ for the second, and so on until all transactions are labeled.
        ii. O_n = Boarding stop transaction i
        iii. $a_{n,\ l}$ = Alighting stop transaction i (Note that the alighting is not the trip destination as this is the first leg of the trip n)
        iv. If transaction labeled $l + 1$ is last transaction with shared n:
            1. $b_{n,\ l+1}$ = Boarding stop
            2. D_n = Alighting stop
        v. If transaction labeled $l + 1$ is not last transaction with shared n:
            1. $b_{n,\ l+1}$ = Boarding stop (transfer boarding stop for leg $l + 1$)
            2. $a_{n,\ l+1}$ = Alighting stop (transfer alighting stop for leg $l + 1$)
        vi. Repeat steps iv and v for subsequent transactions with shared n. Update $l = l + 1$.
3. Set variables $n = n + 1$, $l = l + 1$. Repeat step 2 transaction $i = i + 1$
4. Repeat steps 1 and 2 for next smartcard

This algorithm is coded in Spyder (Python 3.6)