Research Report

Analysis of Montevideo Automated Fare Collection Data, Final Report

Report 5, iCity SOUTH

Catalina Parada Hernandez,Eric J. Miller

May 2018

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

REPORT 5:
ANALYSIS OF MONTEVIDEO AUTOMATED FARE COLLECTION DATA, FINAL 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

May, 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATIONS

AFC – Automated Fare Collection
STM – Sistema de Transporte Metropolitano (In English, Metropolitan Transportation System)
MHMS – Montevideo Household Mobility Survey
AMMON – Area Metropolitana de Montevideo (In English, Metropolitan Area of Montevideo)

1. INTRODUCTION

Automated Fare Collection (AFC) and smartcard systems have been rapidly adopted by transit systems all over the world, including Montevideo's transit system, STM. The datasets produced by these systems are extensive and can be analyzed for planning purposes and system evaluation.

This report presents a summary of three strategies applied to STM AFC data, aiming to provide insights about transit usage and operations. These strategies are reconstruction of itineraries, identification of alighting locations of transactions, and understanding of travel behaviour of smartcard users. The summaries of the strategies include a brief description of the methods and a sample of results.

The travel behaviour of smartcard users resulting from the analysis of AFC data is further analyzed using the transit trips observed in the Montevideo Household Mobility Survey. These two sources are compared first at an aggregate level and then by pairing smartcards with individuals from the survey. As the AFC transactions and the travel survey datasets have different variables, spatial and temporal windows were used to enable matching.

This comparison as well as the three strategies provide insights on travel behaviour from the customer perspective. Moreover, this study provides a glimpse into the potential of incorporating AFC data analysis into transportation planning and evaluation of transit systems.

2. DATA

The data was provided by the Smart Cities Technology group and the Intendencia de Montevideo; the governmental agency that monitors, coordinates, and integrates the public transportation system in the metropolitan area of Montevideo (AMMON), Uruguay. The integrated transportation system STM (Sistema de Transporte Metropolitano) serves the city of Montevideo and surrounding urban areas in blue as shown in Figure 2-1.

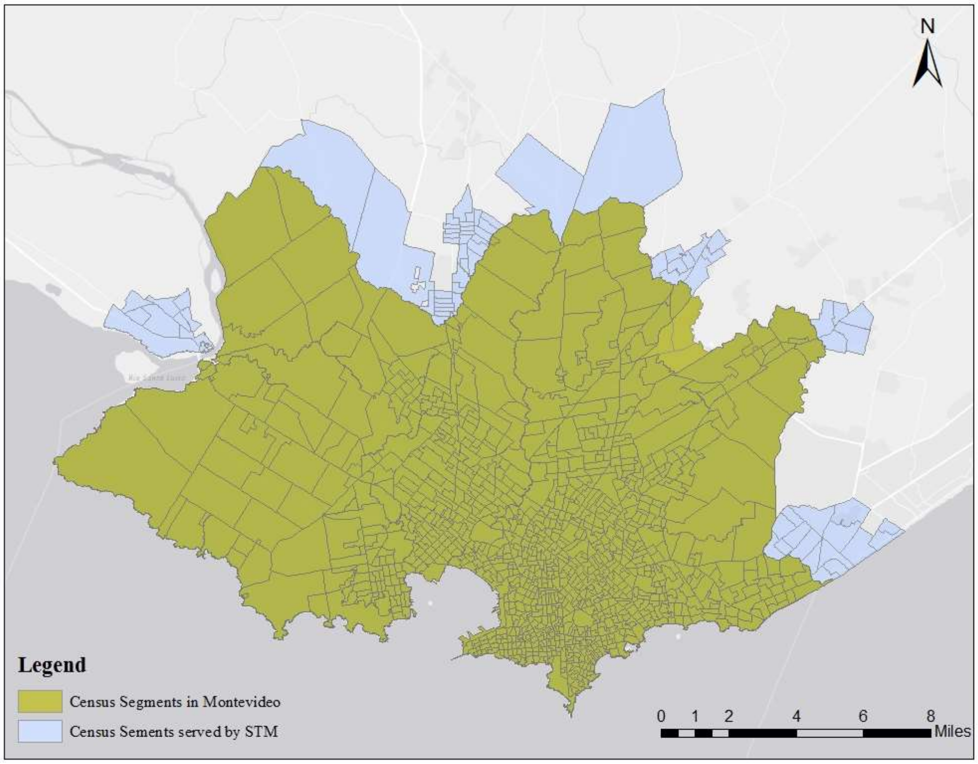

Figure 2-1 Census segments served by STM

This system has 144 bus lines with 107 different terminals, and 4,835 stops. [1]
There are four main components of the data:

1. Boarding records (tap-ons): Seven 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[2] recorded by the system.[3]
2. Bus 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.

A fifth additional source of data is the 2016 Montevideo Home Mobility Survey (MHMS). This is a household survey that collects trips by individuals from a sample of households in

[1] http://www.montevideo.gub.uy/transito-y-transporte/stm-sistema-de-transporte-metropolitano/el-sistema
[2] Transit users without a smartcard pay a cash fare upon boarding the bus. These transactions are electronically recorded and so are available for analysis, along with the smartcard transactions. These transit passengers are referred to in this report as "no-card passengers".
[3] The term smartcard is used interchangeable with STM when referring to boarding transactions and passengers.

the AMMON. The trips by bus are of interest in this study and the survey results can be used to evaluate the OD method results.

For details about the data sources (except the AVL data), the data collected, and methods to prepare it for further analysis, please refer to (Parada Hernandez, 2018). This document also contains aggregated and detailed metrics of the data and travel patterns of smartcard users.

3. ITINERARIES FROM BOARDING TRANSACTIONS

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. 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 (sequence of stops in bus lines and branches).

The itineraries are created using a sequence of Python scripts that combine these data sources. The boarding records are clustered for each stop of their corresponding bus run and the average boarding time is used as the time for the itineraries. The times for stops with no passenger records are interpolated using previous and subsequent stops. The arrival times are forecasted for stop after the last stop with boarding transactions.

The procedure identifies outliers in the transaction records, stops with high dwell times[4], and outputs the itineraries in text and CSV files. For further details about the algorithm, refer to (Parada Hernandez, 2018).

About 1% of all transactions are considered as outliers. To identify the stops with high dwell time, the passenger service time is computed. The service time is the time per boarding passenger and, based on all the transactions, the threshold for acceptable service time is set at 10 seconds per passenger.

The stops with unusually high dwell time can be analyzed spatially and temporally to identify the locations where they occur. Considering all time periods, there are 2,260 stops (48.0% of all stops), shown in Figure 3-1. They occur especially along major corridors and in downtown (inset map). There are also few stops with high dwell times on the outskirts and outside Montevideo.

[4] The term dwell time refers to the passenger boarding flow time, disregarding the doors opening and closing times. The flow time is computed as the time between the first and last passenger of all the passengers boarding at a stop.

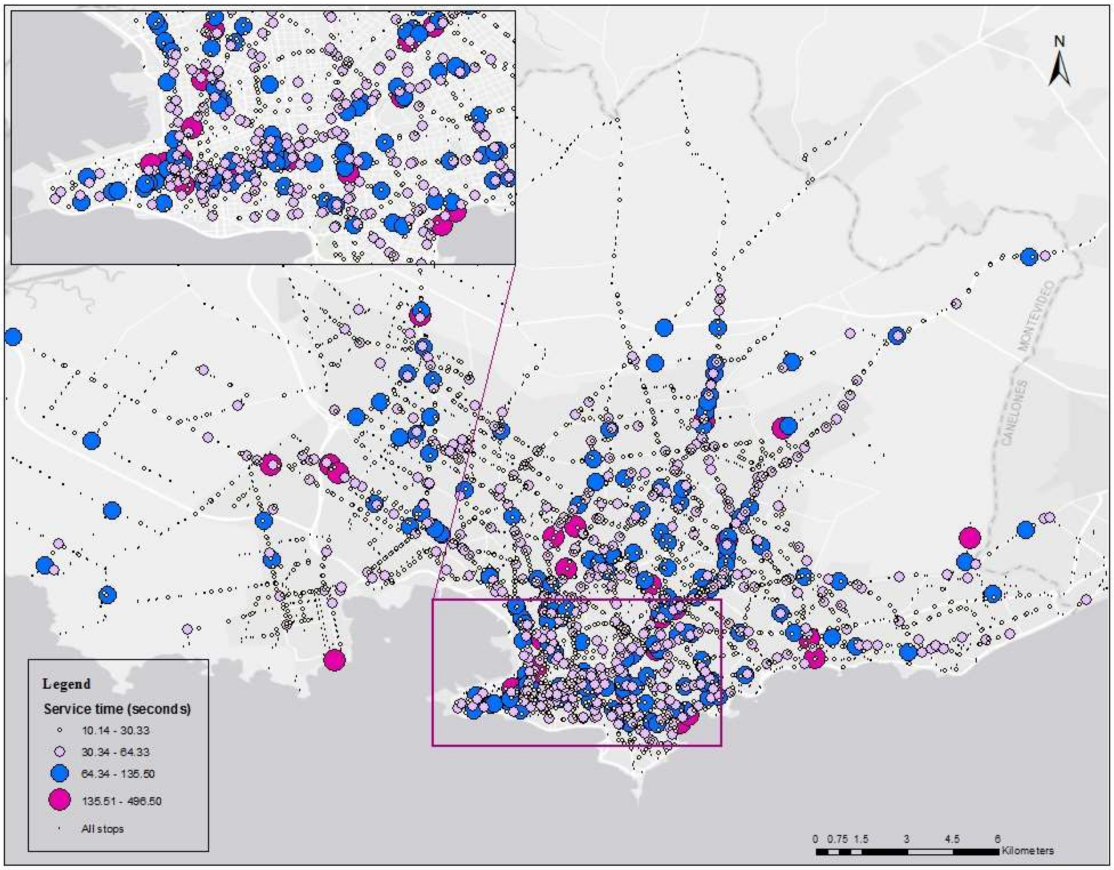

Figure 3-1 Stops with high dwell time

Having identified the stops with high dwell times, itineraries are built for the bus runs. The itineraries are built for over 97% of the daily bus runs. Figure 3-2 shows an example of the itinerary for five buses serving the bus line 19, branch number 205.

The times highlighted in blue correspond to stops with no passenger boardings and those in grey, to stops with high dwell times. The arrival times for these cells were interpolated, using the arrival times from the previous and subsequent stops that are not highlighted. Note that the last stops of runs, highlighted in yellow, are forecasted using the time step of the previous interpolation.

Bus line 19		Run 1	Run 2	Run 3	Run 4	Run 5
Branch 205		Start time	Start time	Start time	Start time	Start time
Monday, August 15		5:23:15	5:56:48	6:19:47	6:26:06	6:42:01
Stop ordinal	Stop	Arrival time	Arrival time	Arrival time	Arrival time	Arrival time
1	3079	-	-	-	-	-
2	3017	05:31:44	06:10:43	06:21:35	06:32:48	06:46:28
3	3019	05:33:21	06:11:36	06:22:09	06:33:33	06:47:17
4	3020	05:33:51	06:12:29	06:22:44	06:34:10	06:48:07
5	3021	05:34:22	06:13:15	06:23:19	06:34:53	06:48:43
6	3022	05:35:04	06:14:00	06:23:54	06:35:32	06:49:19
7	3023	05:35:47	06:15:02	06:24:31	06:36:21	06:50:11
8	3024	05:36:52	06:15:44	06:25:10	06:37:45	06:50:52
9	3025	05:37:58	06:16:27	06:25:50	06:38:11	06:51:32
10	3026	05:39:06	06:17:09	06:26:55	06:39:03	06:52:13
…		…	…	…	…	…
60	3733	06:16:16	06:59:23	07:08:07	07:22:21	07:37:15
61	3734	06:17:00	07:00:10	07:08:52	07:23:21	07:38:06
62	3735	06:17:43	07:00:58	07:09:38	07:24:20	07:39:07
63	563	06:18:27	07:01:46	07:11:06	07:26:00	07:40:59
64	564	06:19:12	07:02:35	07:11:37	07:26:42	07:41:23
65	565	06:19:58	07:03:30	07:12:13	07:27:25	07:42:02
66	4578	06:20:44	07:04:28	07:12:50	07:28:08	07:42:41
67	566	06:21:30	07:05:27	07:13:27	07:28:51	07:43:20
68	3924	06:22:16	07:06:25	07:14:04	07:29:33	07:44:00
69	570	06:23:02	07:07:24	07:14:40	07:30:16	07:45:00
70	3925	06:23:48	07:08:22	07:15:17	07:30:59	07:46:00
71	4580	06:24:34	07:09:21	07:15:54	07:31:42	07:47:00
72	4615	06:25:20	07:10:19	07:16:35	07:32:29	07:48:06
73	4909	06:26:05	07:11:18	07:17:15	07:33:16	07:49:12
74	4040	06:26:51	07:12:16	07:17:56	07:34:03	07:50:19
75	4041	06:27:37	07:13:15	07:18:36	07:34:50	07:51:25
76	4763	06:28:23	07:14:13	07:19:17	07:35:37	07:52:31
77	4764	06:29:09	07:15:12	07:19:57	07:36:24	07:53:37
78	4765	06:29:55	07:16:10	07:20:38	07:37:11	07:54:43
79	5086	06:30:41	07:17:09	07:21:19	07:37:59	07:55:50
80	4766	06:31:27	07:18:07	07:21:59	07:38:46	07:56:56
81	4767	06:32:13	07:19:06	07:22:40	07:39:33	07:58:02

Figure 3-2 Example of itinerary

4. ORIGIN AND DESTINATION ESTIMATION

The STM system records the boarding transactions. The data collected includes the boarding location and time of passengers, but their alighting location and time are unknown. This section presents an overview of the method to estimate the alighting locations and times of the boarding transactions, and the results. The method has three goals:

1. Estimate the alighting locations and times of transit transactions (from STM and no-card users).
2. Identify the origin and destination of trips for STM users.[5]
3. Compute travel behaviour metrics such as travel times, transfer walking distance, location, and time for STM users.

Due to the differences between STM cards and no-card users, the method has different components. For STM cards with more than one daily transaction an algorithm is proposed to estimate the alighting locations. This algorithm is explained in the following paragraphs. The results are used to estimate the alighting locations of STM transactions for which the method cannot be applied to, and to no-card transactions.

The algorithm analyzes the transactions of each STM card on the bus network. For a card's boarding 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 transaction of the day to estimate the alighting stop for this last transaction. When the alighting stop is estimated the algorithm retrieves the time of arrival of the bus at this stop from the itinerary.

After all the transactions of a STM card are processed the algorithm identifies the origins, destinations, and transfer locations for the trips as well as travel and transfer times. The cards for which all alighting locations can be estimated are considered as having complete trip chains.

To distinguish transfers the algorithm considers the trip ordinal and trip ID fields[6], but does not solely rely on these as passengers can pay one fare and make more than one trip. The algorithm considers as different trips those transactions with transfer time above 30 minutes and when a passenger transaction is on the same line as the previous transaction. Taking the same line in the same direction indicates that the passenger had two destinations on the same path; taking the same line in the opposite direction indicates that the passenger went to a destination and returned.

---

[5] 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 be composed of one or multiple transactions or legs, identified by transfers between bus services.

[6] 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.

The algorithm is implemented for the seven days of data (all weekdays and the weekend). Highlights of the results are included in Figure 4-1. Note the alighting estimation and trip chain rates are significantly lower for the weekend; this can be attributed to the different and irregular travel behaviour expected on weekends.

87.8% average alighting location estimation for weekdays (84.3% for weekend)

67.5% average cards with complete trip chains for weekdays (60.9% for weekend)

178.8 metre average walking distance for weekdays (171.6 for weekend)

18.7 minutes average on-board time

30.2 minutes average trip time

Figure 4-1 Results of algorithm for STM cards with multiple daily transactions

The alighting location estimation results are used to assign alighting locations of two transaction types: those for which alighting could not be estimated using the algorithm, and single daily transactions. This is done by observing the transactions for each card and identifying similar transactions in other days for which alightings could be estimated. Moreover, the alighting locations of passengers that do not use a card are estimated based on the travel patterns of card users. Note that this is considering that passengers that do not use a card have a similar behaviour than passengers that have STM cards.

The alighting location estimation can be analyzed and visualized at any spatiotemporal level and for specific bus lines and/or STM card types. The results for Monday, August 15 2016 are shown here. For instance, Figure 4-2 and Figure 4-3 show the trip origins and destinations for STM users for the AM period (4 a.m. to 11 a.m.) aggregated by census segment. The volumes of transfers can also be identified for different time periods. Figure 4-4 shows the transfer volumes of stops in the PM period. Moreover, bus load profiles can be created for different bus lines, time periods and corridors. Figure 4-5 shows the loading profile of one of the morning bus runs for branch number 205.

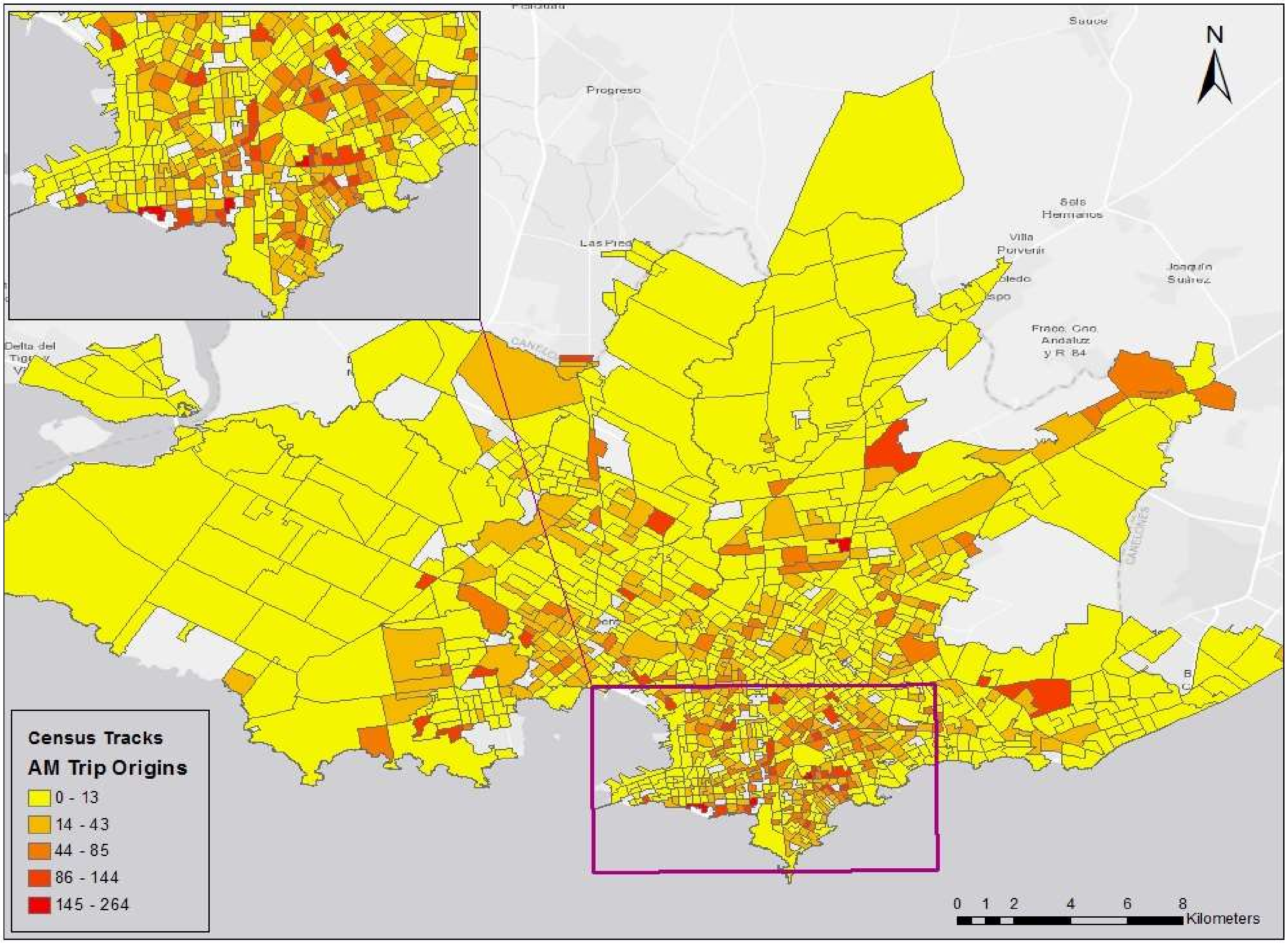

Figure 4-2 AM Trip origins of STM users

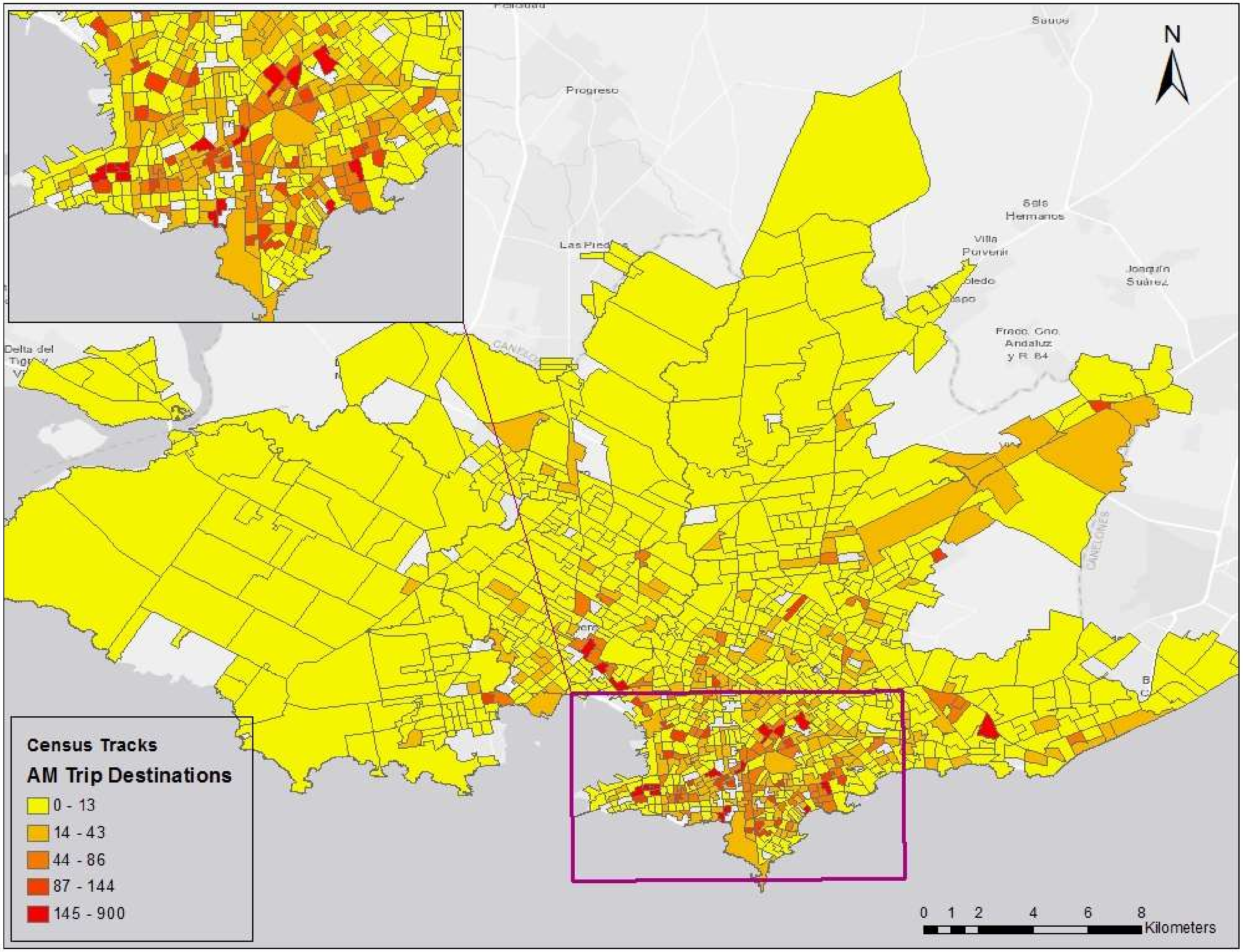

Figure 4-3 AM Trip destinations of STM users

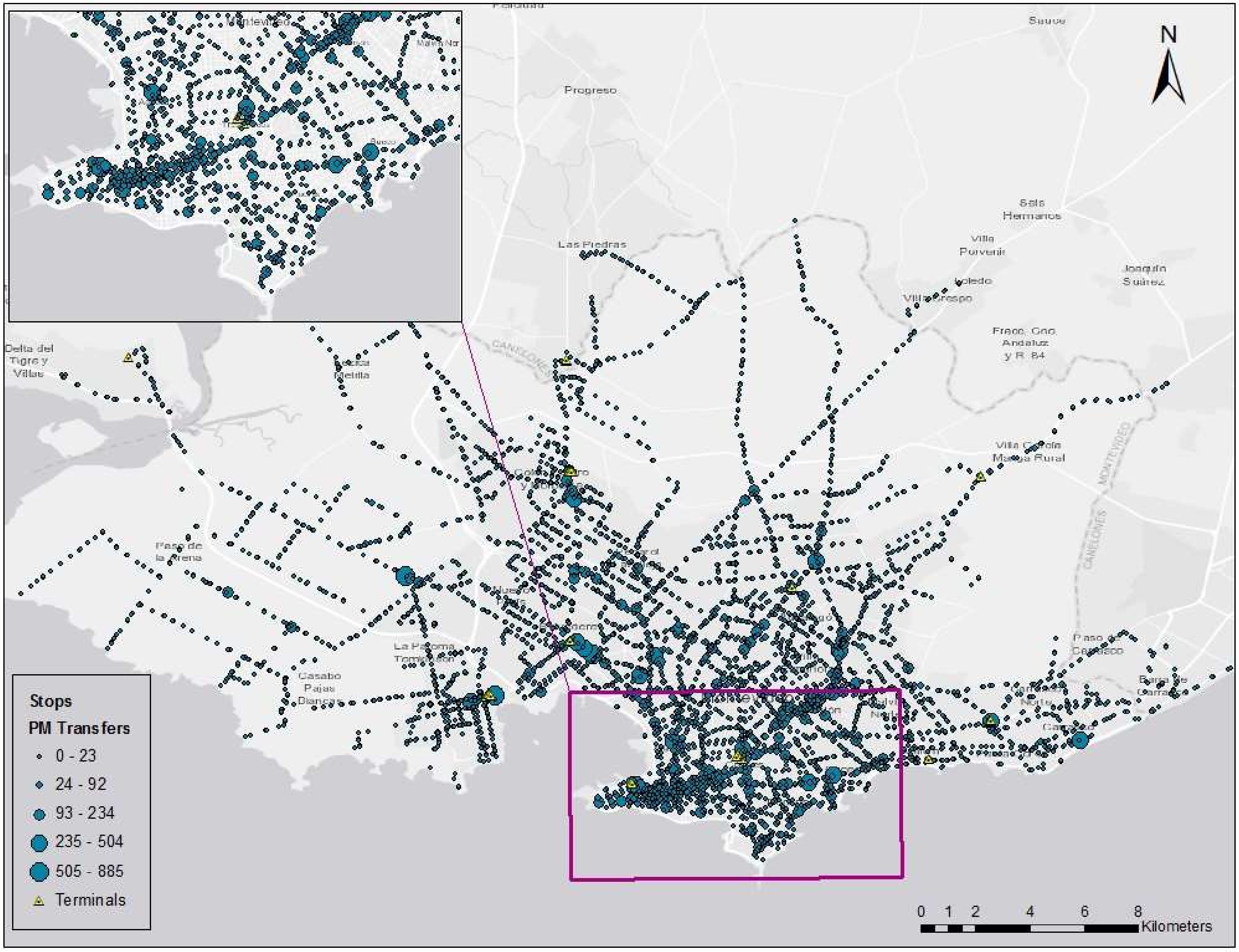

Figure 4-4 PM Transfers

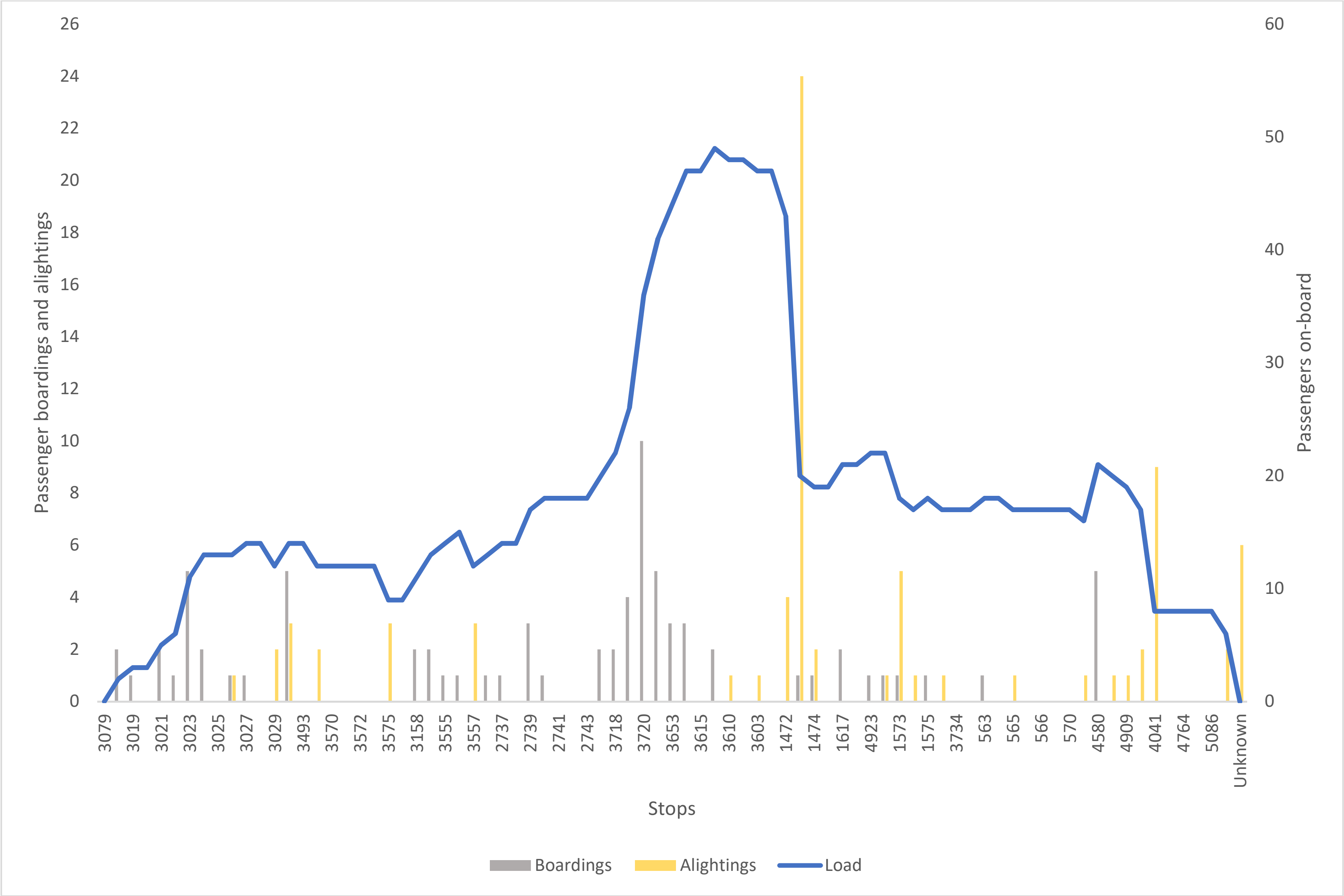

Figure 4-5 Bus loading profile for all passengers

Some of the most important findings from the analysis of the results and the visualizations include:

- The travel behaviour of passengers can be analyzed at any spatiotemporal level. Figure 4-2 and Figure 4-3 show the origins and destinations of trips by all STM users for the AM period, and, as expected, the origins occur around the urban periphery and the high urbanized areas in the northeast and northwest of Montevideo. There are also many trips that originate in the downtown. However, the destinations of these trips occur in few census tracks, particularly in or close to downtown. There are also some clusters of census segments in the east and northeast part of the city with moderate volumes of trips destination.
- The transfers during the AM time, shown in Figure 4-4, occur at specific locations: along major roads, the downtown area, terminals, and major stops. As expected, there are many transfers on the terminals, identified on the maps as yellow triangles. Additionally, there are few stops with high transfer volumes in the periphery.
- The share of boardings per card type differs from the share for which the trip chains can be estimated. The standard users represent 45.8% of cardholders, but 39.4% of the cards for which trip chains are estimated. In contrast, for students (particularly Student A and Student free) and for retired cardholders, the trip chain percentage is 1 to 3% higher than their percentage as cardholders. These differences indicate more traceable travel patterns for students and retired users, who make all legs and trips of their daily travel by transit; and less traceable patterns for standard users, which means that these users are more likely to use multiple modes (e.g. car, taxi, car-pooling) on their daily travel.

Even though these are some observations from the analysis of results, more observations can be done of particular lines, user types, and/or time periods. The transaction times are recorded to the closest second and the locations to the stop level, therefore any spatiotemporal window could be used.

5. ANALYSIS OF TRAVEL SURVEY TRANSIT RIDERS AND STM CARD USERS

The estimation of ODs for STM card users, particularly the trip chains, are of interest as they can be joined with the transit trips of individuals collected by the MHMS. Even though these two datasets are different, they are compared based on aggregate metrics such as legs and trips per person, and at a disaggregate level by pairing the survey individuals with smartcard users using the trips' locations and times.

The individuals from the MHMS who are selected for comparison are those with transit transactions in the STM. Some of the aggregate metrics that are compared between the MHMS and AFC datasets are shown in Table 5-1 and Figure 5-1.

Table 5-1Comparison of legs and trips for MHMS and STM data

Comparison condition		MHMS	STM data
Bus trips per person (percentage)	1	15.09	21.23
	2	66.53	60.51
	3	10.53	11.72
	4	6.85	5.61
	>4	0.99	0.92
Average		2.13	2.04
Total		2,150	248,605
Legs per trip (percentage)	1	79.78	71.50
	2	17.73	26.85
	3	2.33	1.33
	4	0.15	0.30
	>4	0	0.01
Average		1.25	1.30
Total		2,572	304,397

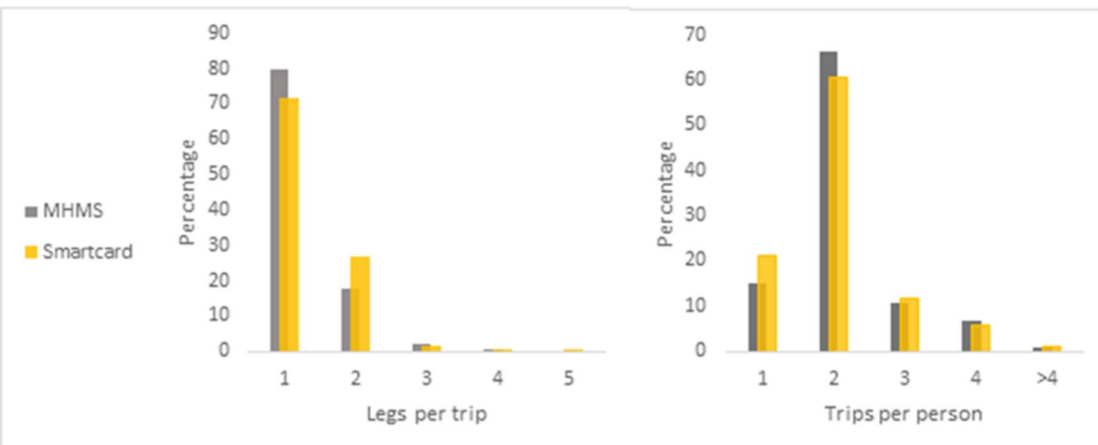

Figure 5-1Histogram of legs and trips comparing MHMS and STM

The average trips per person and legs per trip are similar for both datasets. There are significant differences in the cells highlighted in blue: the share of single trips from the STM users is higher than the reported trips in the survey, but the share for two trips is lower. Conversely, the one-legged trips represent a higher percentage on the survey but the two-legged ones a lower one.

There are some potential explanations for these differences. The most important difference consists on the different samples of population captured. The MHMS only captures less than 1,000 individuals, while over 150,000 STM cards make transactions per day. Also, there could be underreporting of trips, particularly unusual trips such as single trips.

For the disaggregate identification of MHMS individuals in the STM dataset, a method uses spatial and temporal windows to enable the matching. The MHMS data is spatially assigned to the closest census segment and temporally reported by individuals to the closest 5 or 10-minute mark. In contrast, the STM dataset is collected spatially at the stop level and temporally to the closest second. For this reason, the method includes spatial and temporal windows to capture the differences.

As the date MHMS individuals are surveyed is not available, the method is applied for each weekday and two approaches to identify MHMS individuals are considered:

1. Pairing based on origin location and time
2. Pairing based on the origin and destination locations and the start time (Using the output from the alighting estimation method for STM cards)

The results are shown in Table 5-2 and the column "Increase rate" shows the percentage increase in pair identification when only the boardings are matched.

Table 5-2 MHMS identification of individuals for different temporal windows

Day	Time window (minutes)	1. Board only	2. Board and alight	Increase rate
Monday	20	57	39	46.15%
	30	72	60	20.00%
	40	87	78	11.54%
	60	137	92	48.91%
Tuesday	20	41	28	46.43%
	30	62	59	5.08%
	40	92	78	17.95%
	60	133	94	41.49%
Wednesday	20	61	42	45.24%
	30	71	69	2.90%
	40	96	88	9.09%
	60	146	107	36.45%
Thursday	20	42	31	35.48%
	30	74	62	19.35%
	40	89	89	0.00%
	60	130	106	22.64%
Friday	20	43	30	43.33%
	30	78	66	18.18%
	40	85	82	3.66%
	60	126	97	29.90%

The number of individuals identified increases as the time window is expanded but the match rate of the 864 individuals is low: less than 10% with the 20 and 30-minute windows, and around 15%

with the 60-minute one. The days with higher matches are Monday and Wednesday; making these days as the most likely days when individuals were interviewed. Another interesting observation is the decline of the increase rate at the 40-minute window for most days.

Even though the method using only the boarding information capture more pairs, 94% of the pairs are observed in one and two days. Meanwhile, for the board and alight method 32% of the pairs are observed in more than two days. This difference is striking and evidences that considering boarding and alighting information could help identify regular transit riders.

6. KEY FINDINGS AND STRATEGIC USES OF AUTOMATED FARE COLLECTION DATA ANALYSIS

The previous chapters have provided a look into the potential of analysing AFC data for planning purposes, analysis of the STM transit system, and integration with travel survey data. The methods in this study can provide metrics and results for these objectives, as shown on the results and examples, and can be further analyzed for strategic studies.

These results could be obtained from the collection methods and data available for the STM system at the moment of the study. AFC system of the STM collects high quality data for passenger transactions, which include the bus runs and the boarding locations and times. The location and bus run are not usually collected on bus AFC systems and this is an advantage reflected in all the methods: from computing dwell times of each occurrence to identifying the origins and destinations of trips.

The itineraries of buses are built from the AFC records. To do this, boarding transactions were clustered per bus run and stop and it was possible to identify stops with high service times. Around 48% of the stops experienced high service times. In this report, the stops with high dwell times were analyzed spatially to identify the locations of stops and corridors were this occurs. These stops could also be identified on specific bus lines and/or time of the day for specific analyses.

Identifying the locations and potential causes of high dwell time can aid in reducing passenger service time. This reduction would improve travel times, bus operation times and reduce stopping time for buses at stops. Moreover, these itineraries can be enhanced and validated using the schedule data (recently standardized and digitalized) and AVL data. These sources of data can also be used to measure on-time performance and adherence to schedules.

The method to estimate the alighting locations of transactions for smartcard users has similar assumptions to methods previously proposed in other transit systems (Munizaga & Palma, 2012; Trepanier, Tranchant, & Chapleau, 2007); but the estimation results in this study are significantly higher: 88% for weekdays and 84% for weekends. Being able to capture this large percentage of transactions' alighting locations is extremely valuable for understanding Origin-Destination of transit trips, transfer locations, and utilization of the bus network.

From the maps shown in this report, morning and evening transit flows can be observed and analyzed between areas of the city. Also, the stops with high transfer volumes that are not terminals or first/last stops of bus routes can be further studied to improve infrastructure or propose bus routes that minimize transfers.

The high percentage of users with complete trip chains (67.5% for weekdays and 61% for weekends) reveals the transit culture and the passenger behaviour, and regularity. As briefly mentioned before, students and retired STM cards have a higher trip chain rates than standard users, indicating more traceable travel patterns for students and retired users, who make all legs and trips of their daily travel by transit; and less traceable patterns for standard users, which means that these users are more likely to use multiple modes (e.g. car, taxi, car-pooling) on their daily travel.

The differences between the transactions and results of weekdays and weekend show the variable transit patters and could be quantified by location and flows between areas, and the routes used. Even though the results for the methods applied to each weekday were similar, observations across different weeks of the year could provide a better picture of the variability in user behaviour.

Finally, the joint analysis of AFC and the MHMS individuals highlighted the difficulty in identifying individuals in the smartcard database. This could happen due to misreporting of trips or users than do not have a STM card. There are many ways in which the matching process could be improved through additional questions in the survey such as reporting survey date, collecting STM card IDs and card type. Also, another way to gather travel characteristics of transit riders (e.g. trip purpose or demographics), would be to conduct on-board surveys.

This study has continued to reinforce the potential of smartcard data as a powerful source of data for transit studies. The methods proposed in this study use and incorporate the data sources available, taking into consideration the data limitations. Even though the methods and their assumptions have limitations and weaknesses, the results reveal the usefulness of these methods for processing AFC data for transit planning purposes and computing and evaluating the system operations and the transit network.

7. FUTURE WORK

There are weaknesses on the method used to process the AFC data and the assumptions made. Future work needs to be done to ensure the methods are calibrated and used accordingly. This could be done by conducting on-board surveys, collecting information about boarding and alighting locations, travel time, and trip regularity. The collection of STM card IDs could also enhance understanding of travel patterns. Analysis of smartcard data over longer time periods (i.e., multi-week samples) would also assist in developing a more complete description of transit usage.

8. REFERENCES

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9–18. https://doi.org/10.1016/j.trc.2012.01.007

Parada Hernandez, C. (2018). *Analysis of Automated Fare Collection Data from Montevideo , Uruguay for Planning Purposes*. University of Toronto.

Trepanier, M., Tranchant, N., & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated, *11*(1), 1–14. https://doi.org/10.1080/15472450601122256