CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Data Fusion of APC, Smart Card and GTFS to Visualize Public Transit Use

Antoine Giraud
Félix Légaré
Martin Trépanier
Catherine Morency

October 2016

CIRRELT-2016-54

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

Data Fusion of APC, Smart Card and GTFS to Visualize Public Transit Use

Antoine Giraud[1], Félix Légaré[1], Martin Trépanier[1,2,*], Catherine Morency[2,3]

[1] Department of Mathematics and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

[2] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

[3] Department of Civil, Geological and Mining Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

Abstract. Data from smart card fare collection systems can be used to have a better knowledge of travel behavior of transit users. However, some systems do not collect the geographical location of transactions ("tap-in"), nor the destination ("tap-out"). In this paper, we propose an approach to merge data from smart card, automated passenger counting systems (APC) and network supply data (GTFS). In the case study of the Réseau de transport de Longueuil, the method finds 92% of the transaction locations, and 71% of the destinations for these transactions. The paper also presents a web interface specially developed to map the load profiles of the routes taken by users interacting in specific areas of the network.

Keywords: Public transit, public transit, automated passenger counting system, smart card fare collection systems, GTFS, visualization.

Acknowledgements. The authors wish to acknowledge the support of the Réseau de transport de Longueuil (RTL) for providing data, the Thales group, the PROMPT Quebec research fund and the Natural Sciences and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

---

* Corresponding author: martin.trepanier@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2016

INTRODUCTION

Even though smart card automated fare collection systems have become increasingly popular among transit authorities, they are designed to collect revenue, not to monitor public transit use. Hence, there is still a need to develop methods and tools to process and value its transaction data on a regular basis. Automated passenger counting systems (APC) are dedicated to monitor the ridership aboard vehicles. However they do not provide much information on the full path of the user (first origin to the final destination). We need to combine data from other sources to be able to overcome the lack of information. In addition, it is not enough to have complete data: we need to have adequate data analyzers. The use of business intelligence (BI) cubes and other tools is not suitable for the analyses that are conducted by transit operators: there is a need to have transit-adapted tools that will show trip sections and itineraries on the transit network, not only cross-tabulated variables.

The aim of this paper is twofold:

- first, it proposes a data fusion method aimed at completing transactional data from smart card (SC) systems, adding data from automated passenger counting (APC) systems (that give the location of vehicles in time) and General Transit Feed Specifications (GTFS) file (that describes the network geometry and schedule); the method being used to determine the origins, the destinations and the full path of users on the network;
- second, the paper presents a visualization interface that was developed for a transit operator in order to present the analyses made from this data in a “transit-planner-acceptable” way.

The paper begins with a literature review that recalls the most relevant work in the area. Then, the data fusion process is exposed in the methodology section, showing its application to the case study of the *Réseau de transport de Longueuil*, Canada. The results section presents the statistics related to the data fusion and show the visualization interface that has been developed.

BACKGROUND

In this section, we overview previous work done with smart card and APC data, and then emphasizes on visualization tools in public transit.

Smart card & APC data

Numerous research works have shown the potentialities of exploiting smart card data for strategic, tactical and operational purposes *(1)*. Unfortunately, not all smart card systems were designed at first to support such analyses. In many cases, the destination of the trips is unknown (“tap-in” only); methods were developed to impute the alighting stop *(2)* and have been improved since *(3)*. Moreover, in other cases, the location of the boarding transaction is unknown. Lomone *(4)* has proposed a method to merge APC and smart card data in the case of a single bus route of the Montreal transit network.

Recent works demonstrated the usefulness of such data to characterize travel patterns with data mining techniques like DBSCAN *(5, 6)* and dynamic time warping *(7)*. Finally, some works have tried to overcome the lack of socio-demographics information on users in smart card data. Kusakabe and Asakura *(8)* have proposed a data fusion approach to link smart card data to person trip survey. Munizaga et al. *(9)* have linked smart card data to the Santiago survey to validate responses. An activity-based approach has also been proposed by Grapperon et al. *(10)*.

Visualization Tools in Public Transit

Dedicated graphics and visualizations have been developed in the past decade to answer the planning needs of transit agencies. Single route load profiles *(11, 12, 13)* and time-space diagram *(4, 14, 15*) are typically in use. More precise OD matrix can be obtained from "tap-in / tap-out" smart card data and can be represented thanks to geomatics on a map as suggested by Tao et al. *(16)* for Brisbane trip flows and by Gordon *(17)* who showed the origins of all trips sequence that used a given line of London for a given day.

Nowadays, web tools have enabled new ways of visualizing data and interacting easily with them. Barry and Card *(18)* used this medium to represent Boston's subway system OD transfer times. Côme and Oukhellou *(19)* developed an online tool to visualize data from numerous bikesharing systems. Other web products worth of mentioning were developed by private companies and research bodies. They provide advanced visualization of public transit ridership *(20)* or features to design transit networks *(21, 22)*.

METHODOLOGY

Information system

The case study is the *Réseau de transport de Longueuil (RTL)*, a mid-size authority located on the south shore of the St. Lawrence River near Montréal, Canada. RTL is a 350-bus network connected to the Montreal subway and the Montreal Central Business District (CBD). The study covers the month of March 2013. Three datasets were made available to us.

The first dataset is the log of the automated passenger counting (APC) system of RTL. At the time of the study, only a third of the buses were equipped with an APC system that locates the buses over space and time. Fortunately, APC-buses are randomly assigned to the different routes over time to ensure a maximum coverage. There are 2,193,000 APC records that contain: APC event (door opening and closing, counts), date and time of the event, longitude and latitude, vehicle number, route, direction.

The second dataset contains 2.4 million smart card (SC) transactions that were recorded aboard buses and 618,000 additional transactions that were recorded at the three Montreal's subway stations that are fed by RTL. Smart card transaction records report on: smart card number, date and time, vehicle number, route & direction (bus transaction), station (subway transaction). No bus stop location is recorded in the Montreal region smart card system.

The last dataset groups the General Transit Feed System (GTFS) files that describe the service for March 2013. GTFS lists the routes, geometry of routes and bus stops locations. In addition, GTFS gives the schedule of passage at each stop for all month. There was a total of 404,000 bus passages during March 2013.

Data Processing

Data processing steps are presented in Figure 1. There are three parts: location estimation of the bus smart card transactions (A), estimation of the destinations ("tap-out") for smart card transactions (B), and data preparation for the web visualization interface (C).

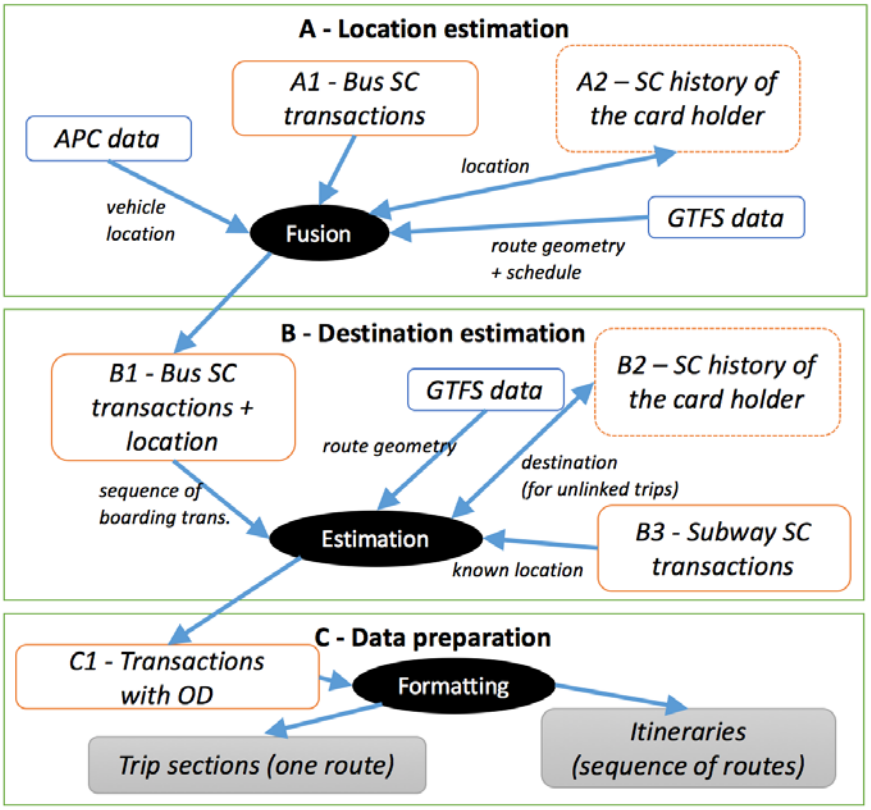

Figure 1 – Methodology

The method to locate boarding stops is decomposed into three phases:

- SC transactions are linked to APC using the bus identification and the time of the transaction, with different time difference thresholds, to be able to transfer the geographical coordinates of the APC event and thus find the nearest bus stop (A1);
- SC history (incrementally built) is scanned to impute stop locations to boarding that occur on the same route and at the same time as spatially located boarding stops (A2);
- at last, SC transactions are also associated with GTFS schedule data to complete stop locations (assuming that there is no major disruption in schedules for that suburb network).

This part of the method has been implemented on a *SQL Server* database.

After, a destination estimation algorithm developed by He and Trépanier *(23)* is applied to the dataset to retrieve the alighting stops. The algorithm is an enhancement of the one proposed by *(2)*. It is based on the sequence of transactions that occurred during a given day, assuming that the alighting stop is the nearest to the next boarding stop (B1). It also takes into account the history of the card to be able to retrieve the destination of unlinked trips (not embedded within a sequence) (B2). This algorithm has been optimized and adapted to work with a GTFS network. New error codes have been added to better categorize the trip destination that could not be estimated (Figure 2). Metrics of distance travel, time spent, rank of the transaction in the day (first, second, last) have been added to the result log file. Subway transactions have been used in order to better estimate the destination of the bus trip transactions in the case of multimodal trips (B3). This part has been developed in *Python*.

Finally, for each smart card, the transactions (C1) are converted into trip sections (one route taken) and transit itineraries (from the first stop to the last, including transfers). Sections and itineraries are imported into a *Elasticsearch* database that is used by the visualization tool (http://elasticsearch.org).

Visualization Tool: Load Profiles Between Two Zones

For a public transit operator, it is critical to understand the mobility of its users in order to better plan the network design, schedule and routes. Often, in order to get the visualization of a given time period and particular place or line of the network, several operations and filters need to be manually done by the analyst. Moreover, more complex maps showing the state of a network at a given time and place need the use of GIS software.

The proposed visualization tool, called “VisuLignes” (*VisuRoutes*) provides an easy-to-use web interface that maps the routes and stop activities over the network. It also shows a metric summary of all the routes – and their geometries – which includes diagrams of the load profile and the distribution of transactions by hours of the day. This tool also enables the analyst to do more complex operations like showing only the trips going in and out of a given area or between two given areas. The map and the route metric summary are impacted by this spatial filter and updated live.

RESULTS

Data fusion results

On the 2.4 million bus transactions of our dataset, 91.6% of the boarding stops have been located (part A). For 79% of them, a destination has been estimated (part B). Figure 2 shows the distribution by the code of the transactions at the end of the data fusion. Overall, 20.8% of the transactions destinations could not be estimated. This is due to several factors:

- step A of the algorithm lacked locating 8.4% of the boarding stops (“tap-in”), so “tap-out” could not be found;
- the GTFS network definition did not include all the service (school and taxi routes were missing);
- previous data processing made by the APC and SC vendors may have impeded on the vehicle ID matching for some transactions.

code	count	%	category	description
11	1 145 542	46,5	success linked trip	linked trip
12	561 286	22,8	success linked trip	return home
13	65 323	2,6	success linked trip	next day first trip
21	141 112	5,7	success unlinked trip	success unlinked trip
22	50 220	2,0	success unlinked trip	success unlinked trip
30	3 833	0,2	fail	fail
31	68 075	2,8	fail	no similar past tap-in
400	182 726	7,4	empty stop route	empty stop ID
404	78 483	3,2	unkown stop error	unknown stop ID
41	84 443	3,4	route error	unknown route ID
42	36 174	1,5	route error	route / stop mismatch
43	48 129	2,0	route error	tap in at end of route
412	13 460	0,5	route error	can't find closest stop
45	*618 868*		*other*	*subway tap-in*

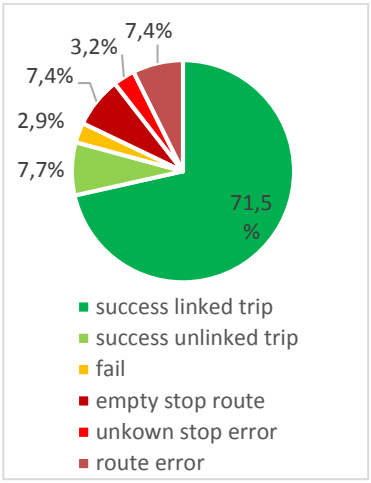

Figure 2 – Destination algorithm results (for bus tap in) distributed by code

Interface description

The main screen of the web interface is presented at Figure 3. Section (A) is the general filtering by date. Setting a date filter will change the dataset used for the map. The daily ridership is presented at (B). Section (C) is an interactive map that can be panned and zoomed to enhance the cartographic details. The map displays the load profile of all the routes that are part of spatial selection, plus the ridership entering and exiting at stops or stations (C1). As any map interface, layers can be selected (C2). In (C4), users can create small areas on the maps to filter the dataset. Here are the options for filtering provided (C3):

- display all trips that are going out of a specific area;
- display all trips that are entering a specific area;
- display all trips that travel between two areas;
- display all internal trips of an area.

For all given filter, the map will display the route load profile using the thickness applied to the geometry of the route (C5). The map can also show the transactions in error (C6). Section (D) is the legend of the colors used in the maps, showing a range of colors for the number of trips entering or exiting bus stops and subway stations.

The bottom section (E) shows a table with the load profiles of all routes that are part of the dataset that is filtered by the map (E1). The section also shows the profile for the different geometries (variants) of the routes. The table also displays the total ridership (“charge”), ridership per direction (“A/R”), percentage of the trips that are between two zones (“O/D”), distribution per day (“par jour”) and per hour (“par heure”), and the load profile (“profil détaillé”).

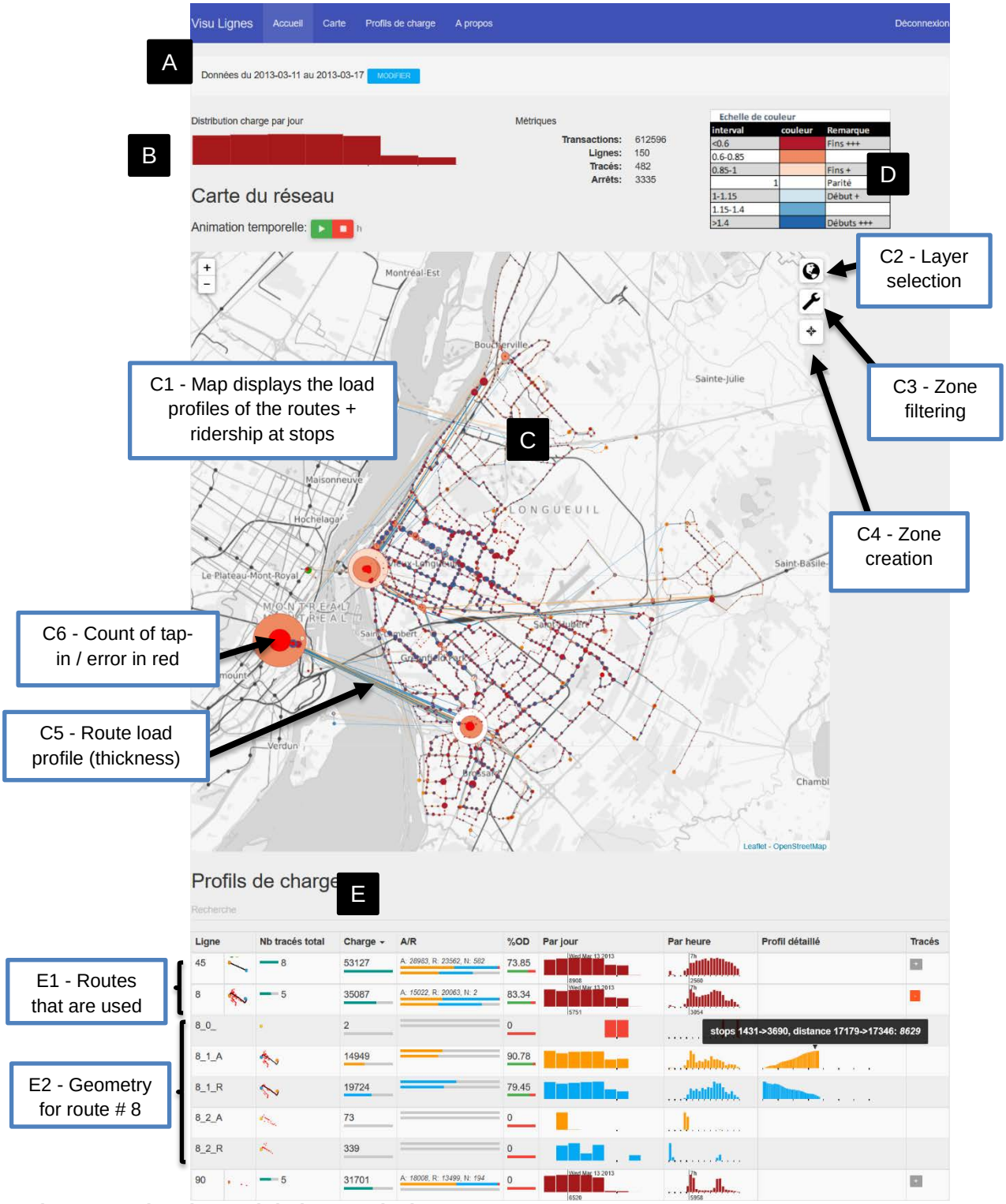

Figure 3 – VisuLine – global network view

Trips that are Bounded to a Single Area

Figure 4 presents the case where a single area is selected (A). The map shows the routes taken by the users who entered or exited the area during the whole period (trip sections). The map clearly shows that two paths are taken to access Montreal CBD. First, route 5 is taken to go to the RTL bus terminal (also linked to Bonaventure subway station). Second, route 19 is used to go to the Longueuil-Université de Sherbrooke subway station that also

links to the CBD. The map also shows the “tap-in” and “tap-out” that occurred along these routes, by the users entering or exiting area (A).

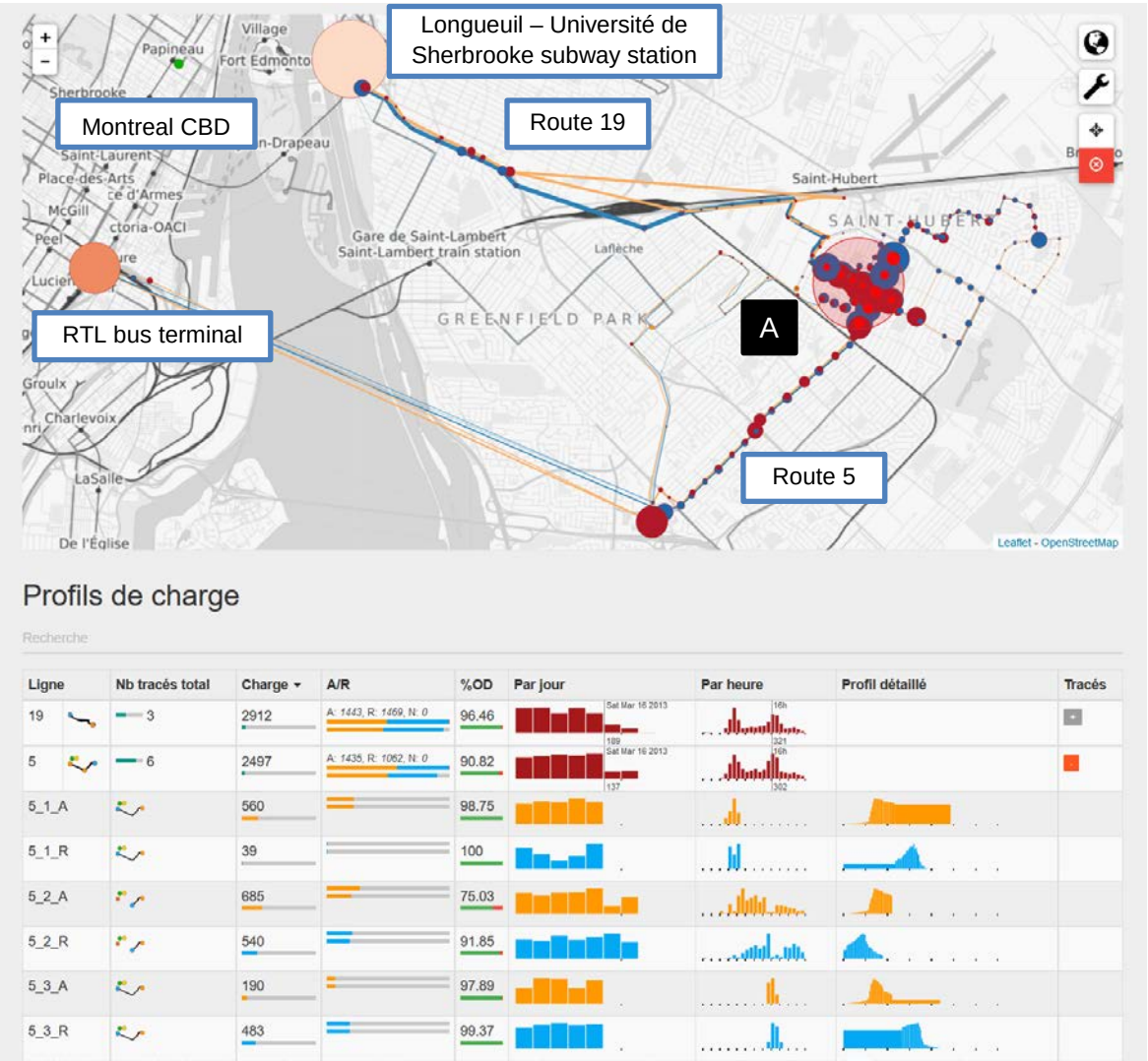

Figure 4 – VisuLine – incoming and outgoing of a given area

Trips Between Two Areas

Figure 5 maps the case where two areas are drawn (A and B). In this case, the load profiles related to the flows going from A to B and vice-versa (this option selected). The bottom of the figure shows that route 73 is mostly used at peak hours, while routes 8/88 are taken more regularly, having a better frequency all through the day.

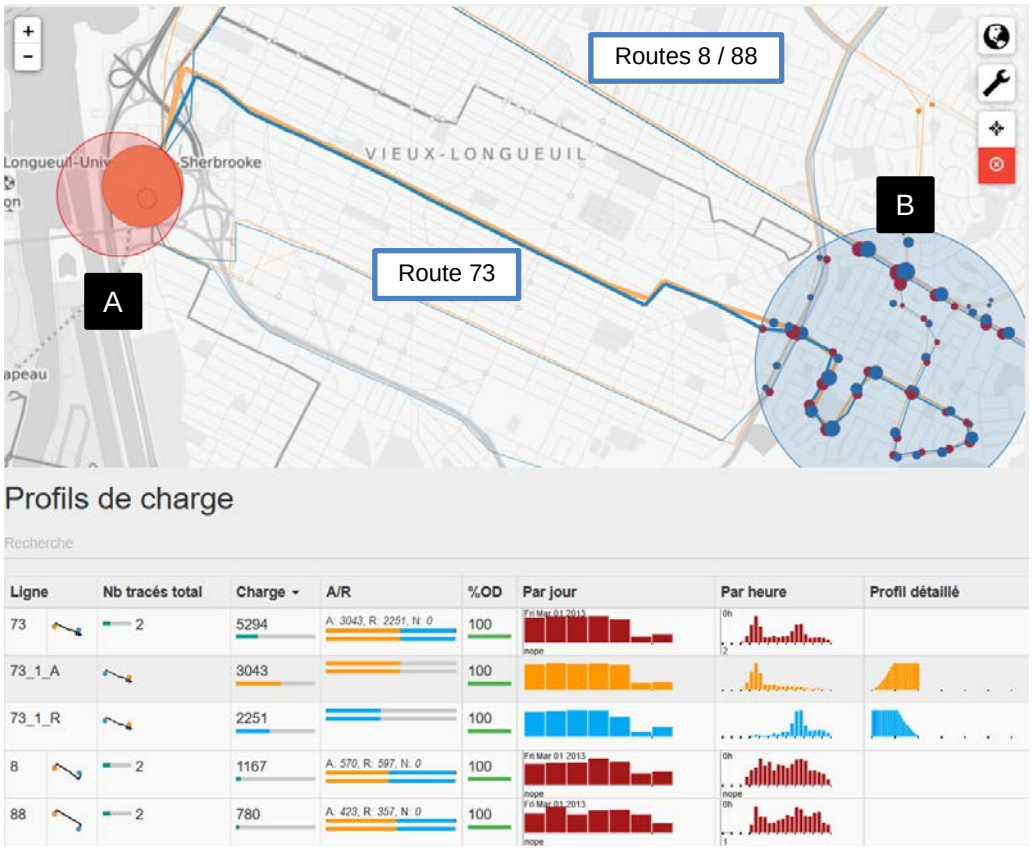

Figure 5 – VisuLine – flow between two given areas

Animated Visualization

It may not be easy to visualize such a quantity of information through time. To overcome this, an animated visualization has been put in place. For any map drawn, the viewer can animate it through time. Figure 6 presents screenshots of the animation of all data for weekdays. It shows the increasing ridership in the morning going to the CBD (red means entering), and then the exit in the PM peak (blue dots). The clear red dots show errors at RTL terminal. This is because of the omission of a route in the GTFS dataset. The error has been detected, thanks to the interface, and was corrected after.

Figure 6 - Network transaction pulse made per hour of the day

A video has been made to show the demonstration of this web interface. It can be seen at https://youtu.be/NSGI5X1SiE0.

CONCLUSION

Contribution

This paper proposed a method to merge data from smart card automated fare collection systems, automated passenger counting systems and GTFS. This is aimed to complete the information provided in the smart card system, where locations of the transactions, nor the destinations were available at first. The approach has been applied to data from the *Réseau de transport de Longueuil*. The method found a location in 92% of the cases, and a destination for 79% of them. The paper also presents a web interface developed to map the load profiles of the routes for specific period of time, and for specific areas in the territory. The interface helps the transit planners to see in detail the movements of the transit users within the network.

Limitations

There are some limitations. First, there is no way to directly validate, at this moment, the locations found, nor the destinations. However, the results are consistent with other studies done by RTL for limited areas and periods of time. Second, the map interface only shows the transactions where locations and destinations were found. When possible, all transactions are shown in accompanying tables and charts (when location is not needed for the analysis). Finally, smart card payments do not account for all the payments at RTL, although they represent a large majority (up to 90%). Hence, the load profile are not completely showing the ridership.

Perspectives

Our next research works will try to overcome the limitations. Data mining approaches can be developed to assess the stability of the transit use and the behavior of the users over time, to reassure the use of the history of the card in the method. In addition, an expansion method can be developed to present load profile with 100% of ridership. Next steps will also be the improvement of the web interface proposed by our research partners: filter trips going through certain routes instead of areas, be able to select and characterize bus stops, enhance filters, etc.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the *Réseau de transport de Longueuil (RTL*) for providing data, the Thales group, the PROMPT Quebec research fund and the National Science and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

REFERENCES

1. Pelletier, M.-P., Trépanier M., Morency C. (2011). Smart Card Data Use in Public Transit: A Literature Review. *Transportation Research Part C*, Vol. 19, 557–568.
2. 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.
3. Munizaga M., Palma C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C*, 24, 9-18.
4. Lomone A. (2014). *Exploration et traitement multidonnées appliqués à des corridors d'autobus*. Mémoire de maîtrise en Génie Civil (MScA) de l'École Polytechnique de Montréal.

5. Kieu L.-M., Bhaskar A., Chung E. (2015), A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data, *Transportation Research Part C*, Volume 58, Part B, 193-207.
6. Ma X., Wu Y.-J., Wang Y., Chen F., Liu J. (2013). Mining smart card data for transit riders' travel patterns, *Transportation Research Part C*, Volume 36, 1-12.
7. Li H., Chen X. (2016). Unifying Time Reference of Smart Card Data Using Dynamic Time Warping, *Procedia Engineering*, Volume 137, 513-522.
8. Kusakabe T., Asakura Y. (2014), Behavioural data mining of transit smart card data: A data fusion approach, *Transportation Research Part C*, Volume 46, 179-191.
9. Munizaga M., Devillaine F., Navarrete C., Silva D. (2013). Validating travel behavior estimated from smartcard data Marcela Munizaga, *International Choice Modelling Conference*, Sydney, Australia.
10. Grapperon A., Farooq B., Trépanier M., (2016). Activity-Based Approach to Estimation of Dynamic Origin-Destination Matrix Using Smartcard data. *TRISTAN IX - Triennal Symposium on Transportation Analysis*.
11. Tranchant N (2005). *Modèle de dérivation des déplacements en transport collectif à partir de données de cartes à puce.* Mémoire de maîtrise en Génie Industriel (MScA) de l'École Polytechnique de Montréal.
12. Trépanier M., Vassivière F. (2008), Democratized Smartcard Data for Transit Operators, *15th World Congress on Intelligent Transport Systems*, New York.
13. Gaudette P., Chapleau R., Spurr T. (2016). Bus Network Microsimulation with GTFS and Tap-in-Only Smart Card Data, *95th Annual Meeting of the Transportation Research Board*, Washington, D.C.
14. Suchkov B., Mikhail B., Reddy A. (2014). Development of a New, Lightweight GTFS Real Time Stringlines Tool to Visualize Subway Operations and Manage Service at New York City Transit. *93rd Annual Meeting of the Transportation Research Board,* Washington, D.C..
15. Anwar A., Odoni A., & Toh N. (2016). BusViz: Big Data for Bus Fleets. *95th Annual Meeting of the Transportation Research Board*, Washington, D.C.
16. Tao S., Corcoran J., Mateo-Babiano I., Rohde D. (2014). Exploring Bus Rapid Transit passenger travel behaviour using big data. *Applied Geography, 53*, 90-104.
17. Gordon J. B. (2012). *Intermodal passenger flows on London's public transport network: automated inference of full passenger journeys using fare-transaction and vehicle-location data.* M.Sc.A. thesis from Massachusetts Institute of Technology.
18. Barry M., Card B. (2014). An interactive exploration of Boston's subway system. from *Visualizing MBTA Data*: http://mbtaviz.github.io/.
19. Côme E., Oukhellou L. (2012). Model-based count series clustering for Bike-sharing system usage mining, a case study with the Vélib'system of Paris, *ACM Transactions on Intelligent Systems and Technology (TIST) - Special Section on Urban Computing archive*, Volume 5, Issue 3.
20. Senseable City Lab, SMART. (2012). *Touching Bus Rides*. Récupéré sur Visual explorations of urban mobility. See: http://senseable.mit.edu/visual-explorations-urban-mobility/
21. Remix. (2016). Transit planning for the 21st century. See: http://getremix.com/
22. Urban Engines. (2016). *Cities*. See: https://www.urbanengines.com/cities/
23. He L., Trépanier M. (2015). Estimating The Destination Of Unlinked Trips In Public Transportation Smart Card Fare Collection Systems. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2535, 97–104.