CIRRELT

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

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Towards Transit Trip Itinerary Inference from Smartphone Data: A Case Study from Montreal, Canada

Seyed Amir H. Zahabi
Zachary Patterson

February 2016

CIRRELT-2016-07

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

Towards Transit Trip Itinerary Inference from Smartphone Data: A Case Study from Montreal, Canada

Seyed Amir H. Zahabi*, Zachary Patterson

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Geography, Planning and Environment, Concordia University, 1455 de Maisonneuve W., H 1255-15 (Hall Building), Montreal, Canada H3G 1M8

Abstract. Recently, a myriad of emerging technologies have been developed to supplement and contribute to conventional household travel surveys for transport-related data collection. While a great deal of research has concentrated on the inference of information from GPS and mobile phone-collected data (e.g. trip detection, mode detection, etc.), to our knowledge, methods for inferring transit routes have not received much attention. This paper describes research whose aim is to work towards transit route inference based on data collected from the smartphone travel survey application, DataMobile. More specifically, we focus on trying to infer transit route itineraries by combining smartphone-collected GPS with geographically precise data on transit routes in Montreal, Canada. The data was collected as part of a mobility study of Concordia University in November of 2014. Since transit route information was not validated in the data collection, our approach is not to compare our inferred routes with reported routes. Instead, as a first step towards inferring transit route itineraries, we have begun by trying to establish the degree to which it is difficult to infer transit itineraries from GPS data on transit trips. That is, since transit routes can overlap on significant portions of their routes, any attempts to associate GPS data to routes, when routes overlap, will necessarily result in "ambiguity" with respect to which routes were actually used. Using this notion of ambiguity, we calculate the proportion of transit trips whose associated transit routes are ambiguous (i.e. cannot be associated with only one route) under different simple assumptions, rules and eventually a simple algorithm. We find that using relatively simple rules, 77% of transit trip distance can be assigned to one route.

Keywords. Transit trip itinerary, GPS, GIS, itinerary inference, ambiguity, smartphone travel surveys, mobile technologies.

Acknowledgements. We would like to acknowledge the financial support provided by Fonds de recherche du Québec – Nature et technologie (FRQNT) under the post-doctoral fellowship program scholarship, Fonds de recherche du Québec - Société et culture (FRQSC) through their "Nouveaux chercheurs" program, the Canada Research Chairs Program, the Canadian Foundation for Innovation, the Concordia University postdoctoral top-up program, and thank the STM and Tram (Transportation Research at McGill) for providing us with the GIS data necessary for this research.

---

* Corresponding author: Seyed.Zahabi@cirrelt.ca

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

1. INTRODUCTION & BACKGROUND

The workhorse for urban travel data collection has long been household travel surveys. These surveys not only represent significant costs, but also are facing increasing challenges due to decreased response rates and data quality issues such as under reporting of short trips (*1, 2 & 3*). As a result, a myriad of emerging technologies are being developed to supplement and contribute to conventional data collection processes. A particularly fertile area of research relates to the use of mobile phones for data collection. There are two broad categories of data collection related to mobile phones. The first involves the passive collection of mobile phone movements recorded by telecommunication companies. The second is the use of GPS (as well as other movement sensor)-enabled smartphones and their associated applications that can be used to collect locational data to observe individual movements during daily travel. With ever-increasing proportions of people owning mobile phones and smartphones in particular, trip recording mobile phone applications and telecommunication cell tower data are becoming hot topics in research related to transportation data collection. Recent studies in the literature focusing on deriving personal trip data are mostly focused in Europe and North America (*2, 4-10*).

Passive collection of mobile phone movements by telecommunication companies uses data generated by cell phone usage from cellular towers, which provides information such as people's location, and can be used to infer typical trips and movement habits (*16*). The advantage of this approach is the incredible amount of data being continually collected. On the other hand this approach doesn't provide any detail on the linkage between persons' characteristics and their travel behaviour (*16*). Smartphone travel surveys and data collection on the other hand can provide a more spatially and temporally precise picture of the travel behaviour of individuals compared to traditional surveying methods (*11, 12, and 13*), as well as compared to passive mobile phone data methods. The main challenge of this type of data collection is recruiting, and retaining users primarily as a result of the battery consumption typically required by these types of applications – the result has tended to be small sample sizes (*17*).

In addition to the data collection tools themselves, a great deal of effort in the literature relating to the collection of data with mobile phones has been dedicated to inferring various types of information about people's trips. Studies like Akin and Sisipiku (*14*) and Sohn (*15*) focus on O-D matrix calculations using cell tower data. The main objective of these studies has been to look at the effectiveness of the methods to obtain accurate OD matrices and trip characteristics. The accuracy in these studies has been obtained by reducing the number of individuals and focusing closely on tracking smaller samples of trip makers. In a recent study by Çolak et al. (*16*) the authors discuss how raw telecommunication cell phone data can be processed to implement a four step transportation model, focusing on the different limitations and strengths of this type of data. With respect to data collected using smartphone apps, and not through telecommunications companies, areas of research receiving the greatest attention have been: stop detection (*2, 6*), trip

breaking (*6*), travel mode inference (*5, 6, 8, 17, and 18*), travel time estimation (*3, 11*), congestion detection (*3*), and real time transit tracking (*19*).

On the topic of mode choice inference, Chung and Shalaby (*6*) developed an algorithm to classify changes in the mode choice into walk, bicycle, bus and passenger car. They did this using data collected via wearable GPS loggers and a written trip report. It's worth mentioning that this study has become the foundation for many other researchers seeking to build mode classification models (*17*). In another study, Reddy et al. (*18*) developed a transportation classification framework that employs a three axis accelerometer and GPS. The classifier used a combined decision tree-discrete hidden Markov model to classify 5 modes from the data set. In a more recent study done by Nour et al. (*17*) the authors present a data-driven classification model to infer mode choice using data collected with Smartphones (GPS equipped). They employed an optimization method to objectively produce a series of classifier components and methods. Thiagarajan et al. (*19*) focus on real-time transit tracking using smart-phones. They developed a method to determine if the person was riding the vehicle, and whether the person is on a bus or another vehicle, and also tracking underground vehicles.

As such, while research concerning inference related to transit trip information has explored a number of different aspects of transit trips, to our knowledge, methods aiming to infer routes used during transit trips do not seem to have received much attention. As such, this paper describes research whose aim is to work towards transit route inference based on data collected from the smartphone travel survey application, DataMobile (www.datamobileapp.ca). More specifically, we focus on trying to infer transit route itineraries by combining smartphone-collected GPS with geographically precise data on transit routes in Montreal, Canada. The data was collected as part of a mobility study of Concordia University in November of 2014. The present research focuses on participants in this study who reported that they only used transit as their mode of travel between home and the university. Since transit route information was not validated in the data collection, our approach is not to compare our inferred routes with reported routes. Instead, as a first step towards inferring transit route itineraries, we have begun by trying to establish the degree to which it is difficult to infer transit itineraries from GPS data on transit trips. That is, since transit routes can overlap on significant portions of their routes, any attempts to associate GPS data to routes, when routes overlap, will necessarily result in "ambiguity" with respect to which routes were actually used. Using this notion of ambiguity, we calculate the proportion of transit trips whose associated transit routes are ambiguous (i.e. cannot be associated with only one route) under different simple assumptions, rules and eventually a simple algorithm.

The rest of the paper is organized as follows. The next section briefly describes the case study region. This is followed by a description of the methodology and the approach used for the calculation of transit route ambiguity, which includes data collection, and processing of the data

used in the analysis. Section four presents the main results obtained and is followed by a short discussion, general conclusions and future work.

2. CASE STUDY REGION – MONTREAL, CANADA

Montreal is the largest city in the province of Quebec, and the second largest city in Canada covering 4,258.31 square kilometers (1,644.14 sq mi) and a population of 4,027,100 (*21*). Montreal has an extensive public transit system comprising bus, heavy rail (Metro) and commuter rail lines, and as a result also has one of the highest transit mode shares in North America (*20*). While getting comparable measures of transit network complexity and density is difficult across many cities, Walk Score (https://www.walkscore.com/) recently developed their ``Transit Score'' - a measure of how well locations are served by public transit. It is calculated based on proximity of locations within a city to transit routes. Based on this Lerner (*22*) reports that with a Transit Score of 77, Montreal is just below Toronto and above all US cities apart from New York and San Francisco. **Fig. 1** shows the Montreal transit network (bus and metro and commuter train lines).

3. METHODOLOGY

The primary approach taken in this paper is to establish for each point of collected GPS data from a transit trip, the degree to which transit route is ambiguous, and in particular, the degree to which it is possible to assign only one route to a given portion of the trip. If only one route is assigned to a GPS point, it is said to be unambiguous. The main purpose of the research was simply to establish to what extent it is difficult to unambiguously establish transit route use from GPS data. The process of establishing the degree of ambiguity can be summarized in four main stages: (1) GPS and other data collection; (2) transit trip data extraction; (3) calculating transit route ambiguity under different assumptions, rules and finally a simple transit route inference algorithm; and (4) evaluating overall transit route ambiguity.

GPS Data Collection and Other Data Sources

A number of data sources were used in this research. GPS data related to transit trips were collected as part of a travel survey using the smartphone application DataMobile (www.datamobileapp.ca) developed in the Transportation Research for Integrated Planning (TRIP) Lab of Concordia University in Montreal. The survey was conducted at Concordia University in Montreal, Canada in November of 2014. All 44,000 members of the Concordia community (students, faculty and staff) were invited by e-mail to download the application. The application included a short survey on respondent socio-demographics, residential location and travel mode between home and Concordia. After completing the survey, respondents could allow the application to run in the background for up to two weeks. While the app ran in the

background it collected locational information when respondents were in transit between destinations. 891 people downloaded the app, completed the survey and had locational data recorded from at least one day.

A number of other sources of GIS data were also used in the research: transit and road network data, metro station locations, location of Concordia campuses, and postal code shape file of Montreal. The shape files of Montreal's public transit network were obtained from the transit agency that operates public transit on the Island of Montreal (Société de Transport de Montreal, STM) for the period of the data collection. This file was then geocoded in ArcMap in order to be used in the next steps for identifying the transit routes taken. GTFS data in Montreal only include the location of the stops associated with route(s) and as a result do not always offer a geographically faithful representation of the routes themselves. The .shp files provided by the STM on the other hand provided geographically accurate representations of the entirety of the routes. Road network and postal code files from DMTI Spatial, metro station locations from the STM, commuter station location and rail lines from the Agence métroplitaine de transport (AMT) were obtained from the Transport Research at McGill (TRAM) archive. The Concordia Campus maps were digitized in the TRIP Lab.

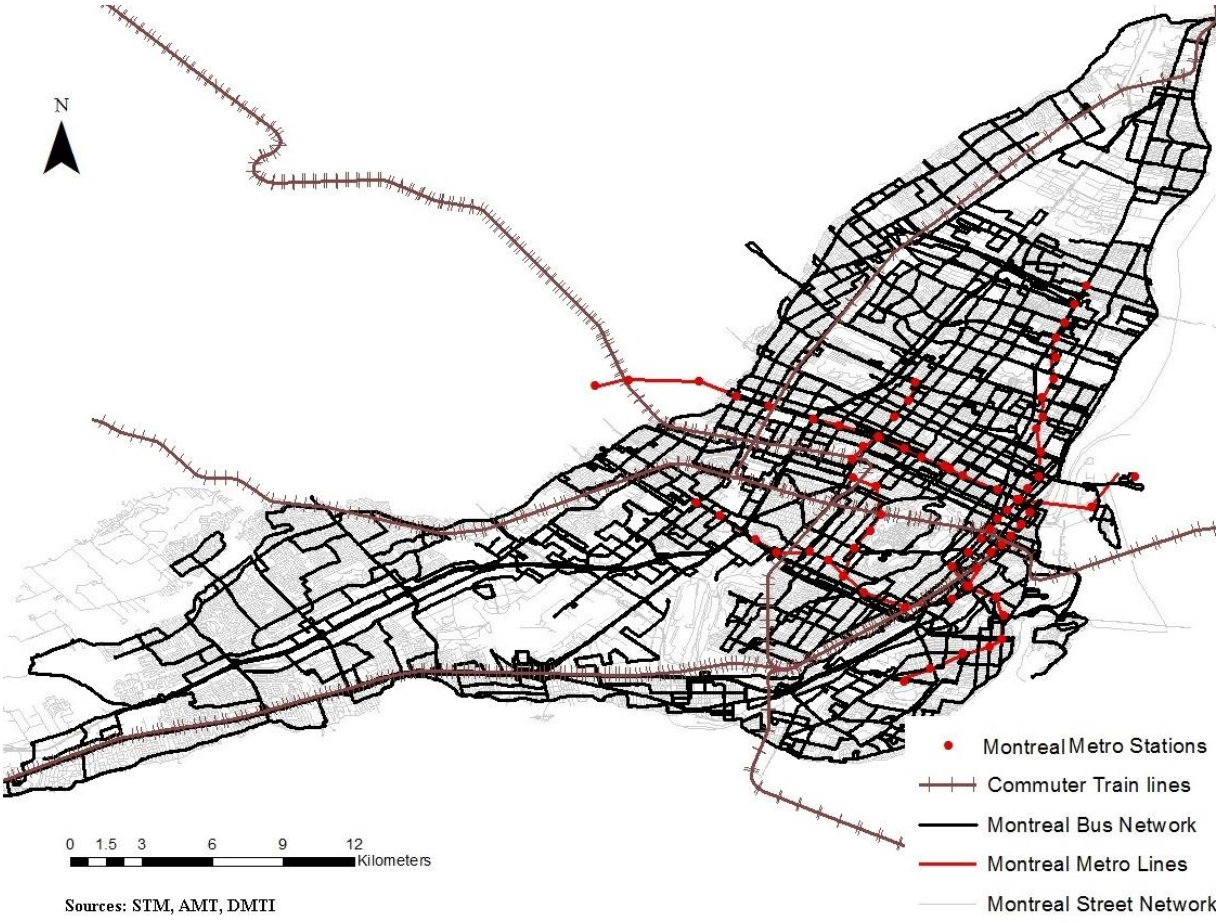

Fig. 1: Montreal transit network

Data Processing

Location data from DataMobile contained fields with user ID, coordinates and a time stamp, in addition to other information such as horizontal accuracy that is not used in this analysis. Since the analysis focused on transit route inference, any trips done by other modes were filtered out. In order to achieve that, we used the declared mode of transportation by the individual as reference. As part of the app setup, there is a small set of questions asked of the respondent, and one of these asks what mode of transportation (and alternative mode) is used between home and Concordia, as well as the postal code of residence. This information was used to select respondents making trips by transit between home and Concordia. Once respondents declaring only transit trips between home and Concordia were separated, these users' data were broken in to trips. Only iOS users were included in locational data on iPhones was collected more frequently.

Trips were identified using a relatively simple trip-breaking algorithm. Time gaps of greater than 5 minutes were classified as stops. Because many trips are done by Metro, and because data collection while in the Metro is sparse, it was necessary to account for longer gaps in time in the case of Metro trips. As such, if two consecutive points were collected within a 250m buffer of a Metro station, a gap of 40 minutes (maximum travel time on the network) was allowed before identifying a stop.

Because spatially accurate transit route information (i.e. a faithful description of routes, and not just the location of stops) was only available for the Island of Montreal, only transit trips that had their origins and destinations on the island were included. Among these trips, the ones of interest were home-based trips done to and from Concordia University. In order to identify these trips, we needed to establish user home location. To locate the home location, we compared the declared postal code of residence against the first and last points of individual's daily trips. This was done by setting a 500m buffer around postal code centroids. Then, by spatially joining these buffers with the first and last daily trip points, the individuals that had their first and last daily trip points falling in the buffer corresponding to the postal code declared as their home, were kept for the next steps of analysis. In the next step, using the first and last point for all trips done by the individuals isolated in the previous step, two qualities were checked and both had to be satisfied for a trip to be considered a home-based Concordia trip: (i) if the first point of a trip was on Concordia University campuses, or the first point is within the declared postal code buffer; and (ii) if the last point of the trip was falling on Concordia University campuses, or the last point is within the declared postal code buffer. This was done using STATA, and it was verified that if the start of the trip were at home, then the end had to be Concordia and vice-versa.

Using these criteria 324 trips were available for analysis. So to summarize, the trips referred to from this point on, are home-based to or from Concordia transit trips on the Island of Montreal. **Fig. 2** presents the data points and the transit lines used in this analysis.

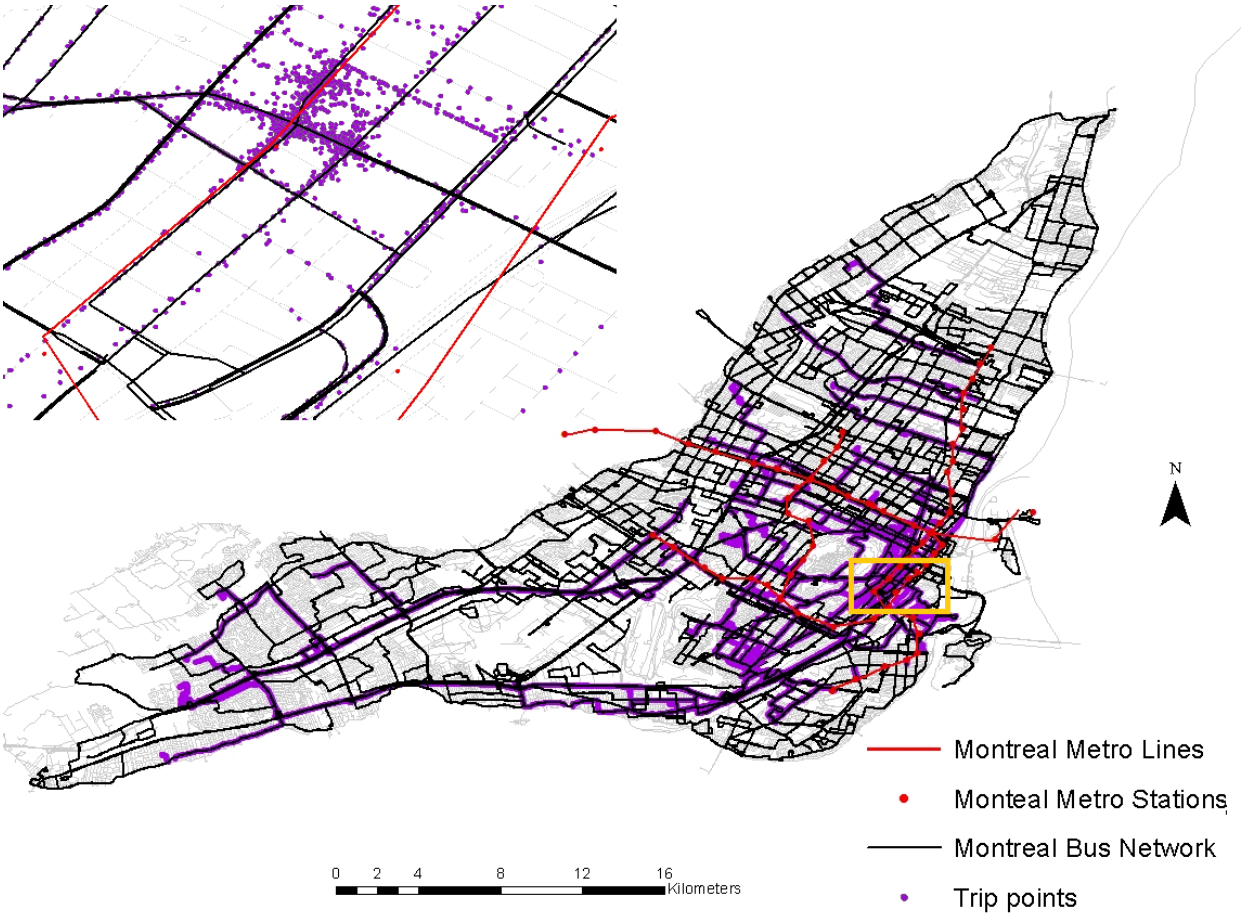

Fig. 2: Home-based Concordia GPS trip points and transit lines

Transit Ambiguity Processing and Route Inference

After preparing the trip data, the next step of the methodology of this research was to calculate the proportion of trips for which there was ambiguity in transit route. Ambiguity was calculated using progressively more (simple) rules and finally a relatively simple algorithm. As such, the following steps were taken:

i. In order to capture candidate bus lines in the vicinity of each trip point, a 15m buffer was set around the bus lines. The 15m distance was chosen after testing several different buffer sizes, and picking the size that didn't capture too many false lines, but also was big enough to capture the trip point on arteries and highways. Then these buffers were

spatially joined to the trip points, recording all the bus line buffers each point intersected with.

ii. The outcome of the previous step, multiple rows (each row representing a bus line) for each trip point was concatenated in one field to represent all the bus lines possible for each point. This was done using scripting in STATA. This is referred to as "baseline processing," and its results were used calculate "baseline ambiguity."

iii. In order to infer if more than one bus route was associated with points for the segments of the trip, a set of codes were developed in MATLAB, in which the trip points were sorted based on timestamp, then bus lines associated with each point were added to a "mother set." Iterating over the points sequentially, the bus lines for each point sharing lines with the mother set was kept and finally recorded when there was no common line. After this, a new mother set was initiated using the same method and the same procedure was followed over and over again. The benefit of using this technique is to isolate the portions of the trip where only one line is available and therefore using that line number, the algorithm moves back up the trip points searching for a point that cannot be associated with the unique line. The points in between where there were only one possible route and the first point for which that line was found in the mother set were all associated with the unique route and considered "unambiguous." This is what is referred to as "bus route processing."

iv. The next step was to establish which points belonged to the portion of the trip done by metro. For this, two filters were used. First, consecutive points of a trip that had a time gap of over 5mins were identified and flagged. The second filter was to flag the points that were in a 250m buffer around metro stations. The points having both qualities were flagged as having been done by metro. This is what is referred to as "metro processing."

v. In order to eliminate any bus lines not operating at the time the GPS point was recorded, the operation time of each remaining associated bus line was checked against the time-stamp of the GPS point. This will filter for night bus lines and express bus lines which function in specific periods of the day, and day buses for trips done off their operating time. This step is referred to as "bus time processing".

vi. The output of the previous step shows all proposed lines for segments of the trip. In the final step, walk trips had to be identified. First, distance between two consecutive GPS recording points is calculated. If the total length traveled at a segment (part of the trip with the same bus line) were less than 200m, and the points on the segment were not a metro trip point, they were set as walking. Also trip points where no transit line was found in the 15m buffer vicinity (because of not being in the proximity of bus lines reported as "999"), those segments were also considered walking trips, which were most commonly observed at the beginning and end segments of the trips. In a case of having such points (with no bus line in their proximity) in the middle of a section where a bus line is found before and after these points, the algorithm in step iii assigns this bus line (the same bus line before and after these points) as the probable line taken for these

points. This is referred to as “walk processing.” Bus route, metro and walk processing taken altogether is referred to as “final processing.”

This procedure is summarized in an algorithm in **Fig. 3.**

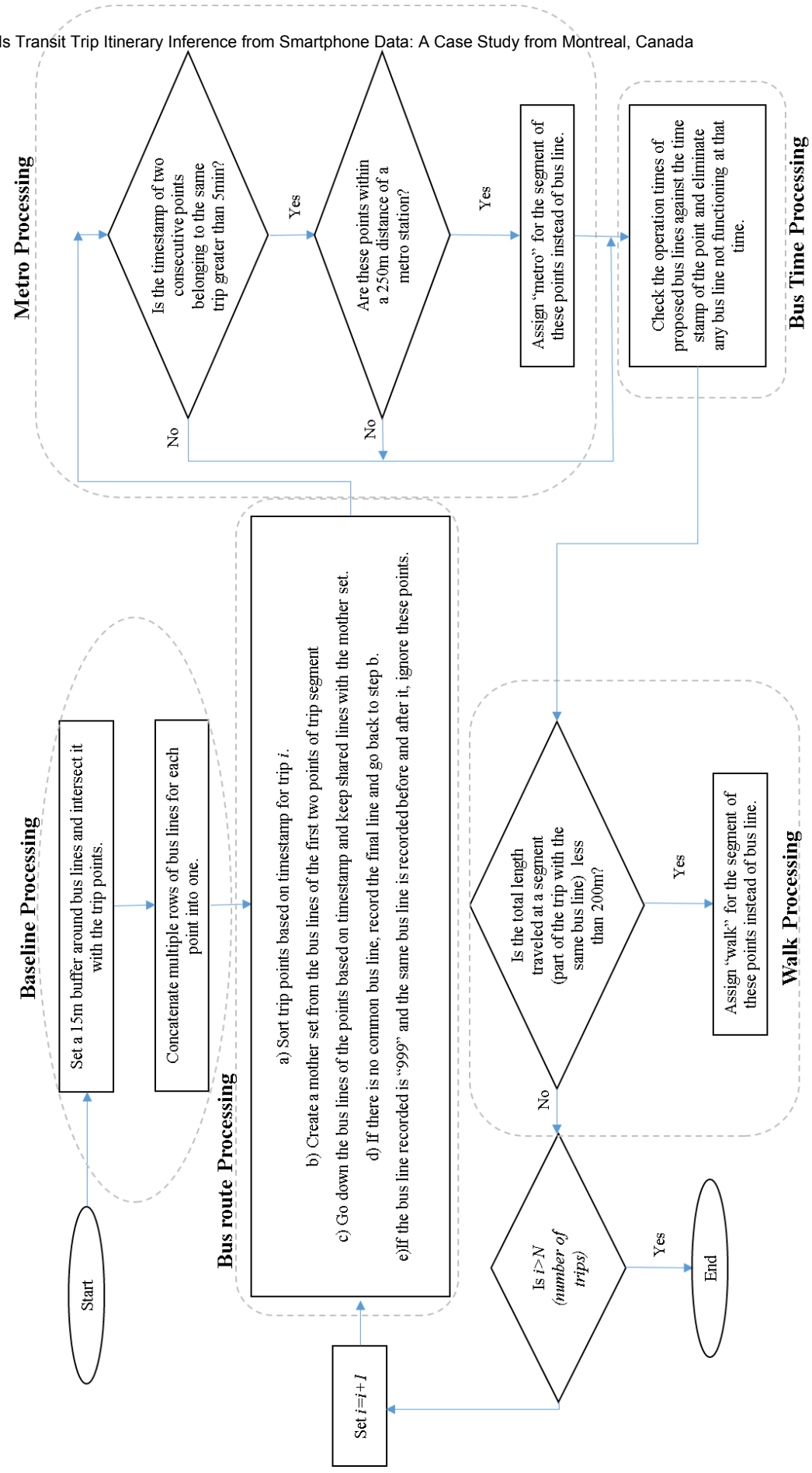

Fig. 3: Algorithm outlining the transit route inference procedure

Measures of Transit Route Ambiguity

To summarize the degree to which transit route was ambiguous over the course of a transit trip, a distance-based measure was used. As such, transit route ambiguity was summarized as the proportion of a trip's distance for which the transit line used was ambiguous. This is referred to as percent ambiguity. In order to demonstrate how the different rules and algorithm reduce ambiguity, percent ambiguity was calculated after each stage of transit route processing. That is, it was calculated when no rules were used to evaluate ambiguity (baseline processing), after the bus route ambiguity detection algorithm was used (bus route processing), and after the metro and walking adjustments were also included (final processing). The portion of distance where ambiguity is caused by having two and three or more candidate lines were also calculated to provide a sense of just how ambiguous the ambiguous portions of trips were.

4. RESULTS

In this section, the main results of ambiguity detection under the different stages of processing are presented. **Fig. 4** presents the buffer analysis to capture the bus line(s) in the vicinity of GPS points for a home based trip from Concordia. As one can observe on the bottom image in **Fig. 4**, the 15m buffer around the bus network shows the option of line(s) in the vicinity of each GPS observation.

Fig. 4: Trip GPS points and the bus line buffer analysis

To better understand the output of the transit ambiguity processing, **Figure 5 and Table 1** present an example trip. **Fig. 5** presents a map that shows trip GPS points and how different transit modes and lines (metro, and bus) are accessible at each point. This figure provides a visual representation of what the data fed to the algorithm looks like for one trip, which then turns out to produce an output table similar to **Table 1**. In **Fig. 5**, the blue points show the trip GPS points, the purple polygon represents one of Concordia's campuses, and the red lines are bus lines. The "M" signs stand for metro stations. We see that this individual walks to the metro station, (top right) and takes the metro and gets off and takes the bus (bottom left), and at the end of the trip because of not having a line passing till the end point recorded, the person walks from the bus stop to his/her destination, which is home in this case.

Table 1 briefly demonstrates output after final ambiguity processing. Due to lack of space, only the top two and bottom two points of each trip segment are presented. As for the points not being in the vicinity (15m buffer distance) of a transit line, the line "999" has been recorded. The last two columns of this figure show if the algorithm is assigning walk or metro trips for those segments. If either of these columns is 1, that overwrites the line number in the proposed line column. Once the algorithm has finished coding and generating results for all the points, the evaluation phase is initiated. The point-to-point distance for consecutive trip points is calculated and used as to calculate percent ambiguity. Trip segments that were neither chosen as walk nor metro, and had at least two bus lines proposed, are considered as ambiguous. The bold text in the table (lines or 1s for the walk and metro) show the final line, or mode proposed by the algorithm. To see the effect of the algorithm on this sample trip, we report the ambiguity at each step. At the baseline ambiguity level, there exists 77% ambiguity, after bus route processing percent ambiguity is reduced to 10%, and after factoring the walk and metro section of the algorithm too, the ambiguity goes down to only 9%.

Table 1 – Output of final transit ambiguity processing for sample user

Point ID	Bus lines in vicinity	Proposed Bus Line(s)							Walk	Metro
1	427-57-165-66-166-15-435-369	15							**1**	0
2	15	15							**1**	0
3	15	15							**1**	0
4	15	15							0	**1**
5	225-215-170-216-213-378-70-468-177-64-368-174-380	**174**							0	**1**
6	174-70-380-368-70-177-216-225-215-213-378-213	**174**							0	0
…	…	…	…	…	…	…	…	…	…	…
63	196-174-475	**174**							0	0
64	196-475-174	**174**							0	0
65	475	**475**							0	0
66	409-220-475-376-216	**475**							0	0
…	…	…	…	…	…	…	…	…	…	…
99	225-409-216-475-215-376	**475**							0	0
100	215-409-216-475-376	**475**							0	0
101	485-217-72	**72**	**217**						0	0
102	485-72-217	**72**	**217**						0	0
…	…	…	…	…	…	…	…	…	…	…
115	215-202-207-217-419-219-200-72	**72**	**217**						0	0
116	207-218-225-206-217-202-202-215-72	**72**	**217**						0	0
117	207	**207**							0	0
118	207	**207**							0	0
…				…	…	…	…	…	…	…
142	407-207	**207**							0	0
143	407-207	**207**							0	0
144	999								**1**	0
145	999								**1**	0
…	…	…	…	…	…	…	…	…	…	…
155	999								**1**	0
156	999								**1**	0

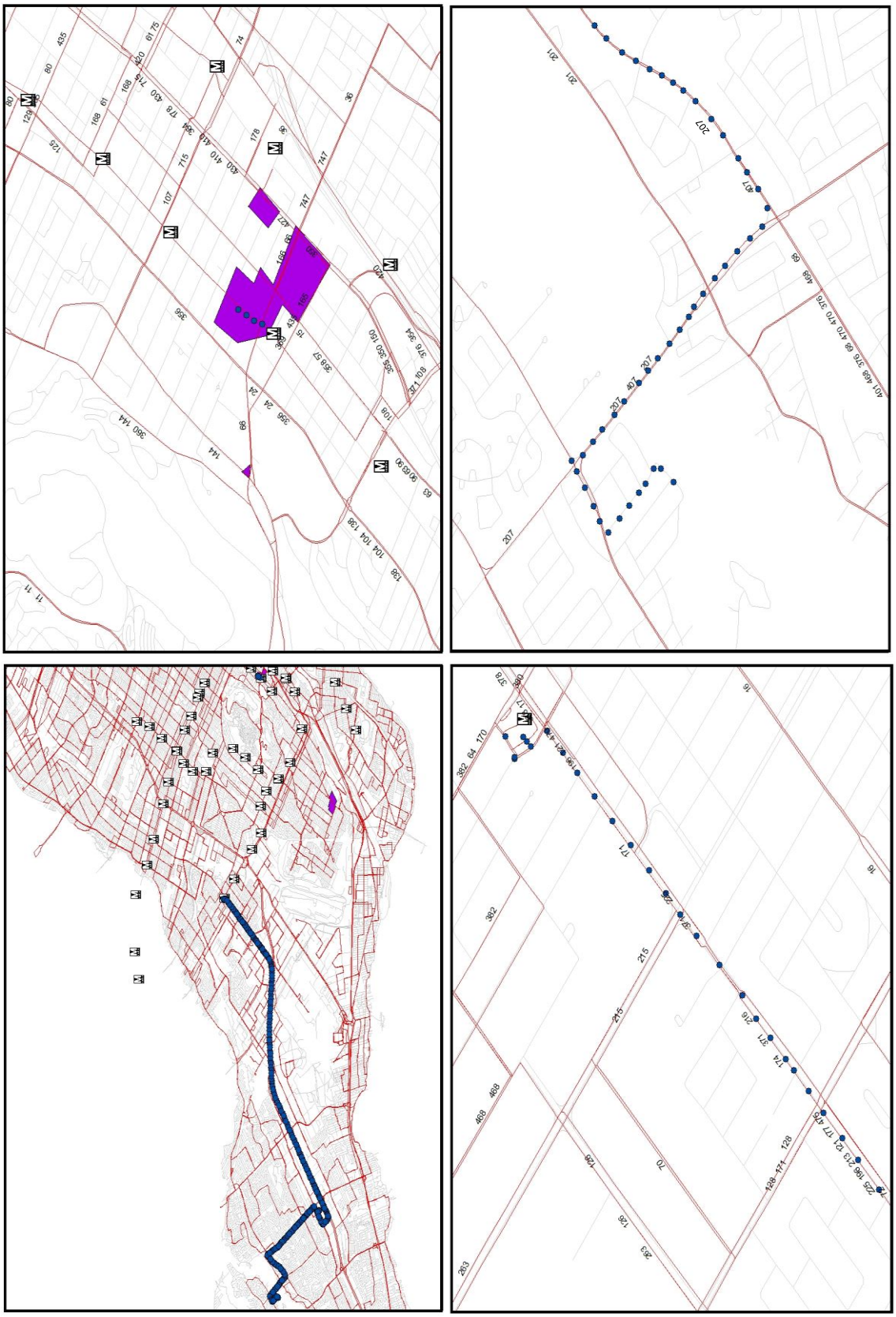

Fig. 5: Visual representation of a sample trip's GPS points and possible transit options

Table 2 presents summary statistics for all of the 324 trips in the dataset. The maximum percent ambiguity without processing (baseline ambiguity) associated with a given trip is 100% whereas the minimum is 21%. The table also shows summary characteristics of the transit trips themselves with an average 7.4 km, a min of 1.1 km and a maximum of 29 km.

Table 2. Summary statistic of the trip data and relative baseline ambiguity

	Number of trips	Trip Distance (km)	Trip time (min)	Baseline Ambiguity (%)	Final Ambiguity (%)
Average	-	7.4	49.2	94	6.8
Min	-	1.1	7.3	21	0
Max	-	29	115.8	100	98
Total	324	-	-	-	-

Table 3 presents the results of ambiguity processing summarized across the entire dataset for each of the different stages of ambiguity processing. It also breaks down percent ambiguity according to the number of lines causing the ambiguity. The first level of output that was evaluated was before doing any analysis on the data and simply focusing on the level of non-ambiguous points before bus route, metro, or walk processing (i.e. points that only have one line in their vicinity as raw data). This is called ***baseline processing***. Notice that after baseline processing 56% of the distance of transit trips are associated with more than three lines, while 5% are associated with three lines, and 11% with only two lines. Percent ambiguity is then reported after the bus line selection portion of the algorithm, without walk and metro processing. This is indicated as ***bus route processing***. As can be seen in **Table 3**, after bus route processing, percent ambiguity is reduced from 72% to 30%. Finally, percent ambiguity is reported after processing for walk and metro trips and bus route and time- ***final ambiguity processing***. We can see that this correction decreases percent ambiguity by another 25% to as low as 4.65%. We can also observe that percent ambiguity with more than 3 lines is reduced significantly after applying the different stages of processing.

Table 3. Ambiguous and non-ambiguous proportion of trips

Processing Stage	**% ambiguity**	**% 2 lines**	**% 3 lines**	**% > 3 lines**
Baseline	72%	11%	5%	56%
Bus route	30%	15%	10%	5%
Bus route + metro	24%	13%	9%	2%
Bus route + metro + walk	23%	13%	8%	2%
Final ambiguity processing	4.65%	4.06%	2%	1%

Table 3 helps us evaluate how hard (or easy) it is to establish what transit route has been taken for each trip. We observe that final ambiguity processing reduces percent ambiguity from 72% to 4.65%, that is by 67% overall.

5. DISCUSSION

Since the analyses presented here are not based on validated (respondent reported) data, caution must be used in the strength of the conviction when reporting these results. First, there are different ways in which the processing of the data could be improved to reduce transit route ambiguity. For example, it would likely be possible to reduce ambiguity further if timetable information were included. This would likely reduce the number of candidate bus routes that could be associated with a given point since not all buses operate during all times of the day. It would also be possible to build in line frequency that could further reduce percent ambiguity if a probabilistic measure were used. Second, this analysis was done in one case city and one might wonder how applicable it would be to other cities.

On both these counts, we feel that the results we have reported are, if anything, conservative. With respect to processing improvements, we believe that any additional improvements to processing would be more likely to reduce ambiguity than to increase it. With respect to applicability to other cities, as mentioned above, Montreal has a very dense transit network by North American standards (although not by the standards of Europe or some Asian cities). As well, the transit network on the Island of Montreal is even denser still. As a result, percent ambiguity calculated as we have in this paper would likely be even higher for many cities – at least in North America. As such, we feel our percent ambiguity calculations are likely to be upwardly biased and as a result, even these simple processing rules and algorithms that we have used show potential to help in transit route inference.

6. CONCLUSION

An important piece of information required for transportation planners is understanding trip-maker's travel behaviour in urban areas. As mentioned in the introduction, two main sources of data can be distinguished when it comes to travel data collected using mobile phones. The second source of data as mentioned before is GPS-based travel surveys and data collection which can provide a more spatially and temporally precise picture of the travel behaviour of individuals compared to traditional surveying methods. In this paper, as a first step towards inferring transit route itineraries, we have tried to establish the degree to which it is difficult to infer transit itineraries from GPS data on transit trips. That is, since transit routes can overlap on significant portions of their path, any attempts to associate GPS data to routes, when they overlap, will necessarily result in "ambiguity" with respect to which routes were actually used. Using this notion of ambiguity, we calculated the proportion of transit trips whose associated transit routes are ambiguous (i.e. cannot be associated with only one route) under different simple assumptions, rules and eventually a simple algorithm using smartphone-collected GPS data for transport survey at Concordia University in Montreal, Canada.

The methodology used in this paper for calculating transit route ambiguity follows a set of GIS based analysis and processing using different software packages. The calculations demonstrate that in a city such as Montreal with a relatively dense transit network, transit route ambiguity without any processing is quite high, making it difficult to unambiguously infer transit routes. At the same time, we have shown that by applying a relatively simple algorithm we find a significant reduction in the transit route ambiguity. More precisely, we find that without any processing, 72% of transit trip distance cannot be unambiguously associated with a single route, but that after processing this is reduced to 23%. We also observe that percent ambiguity associated with situations in which more than 3 potential lines are present is reduced to 2% (from 56% after baseline processing). Finally, to our knowledge, this is a rare attempt to infer transit route from GPS data, and hopefully will help to contribute to the development of research in this area.

Since this paper is a starting point to work towards transit route inference, there are a number of areas in which the analysis could be improved to further reduce ambiguity. These include the addition of timetable information on bus lines as well as the directionality of the roads could help identify walk segments of the trip (if the person GPS recordings were moving the opposite direction of traffic).

7. ACKNOWLEDGMENTS

We would like to acknowledge the financial support provided by FQ-RNT under the post-doctoral fellowship program scholarship, FQ-RSC Nouveaux chercheurs program, the Canada Research Chairs Program, the Canadian Foundation for Innovation, the Concordia University postdoctoral top-up program, and thank the STM and Tram (Transportation Research at McGill) for providing us with the GIS data necessary for this research.

8. REFERENCES

1. Yang, F., Yao, Z., & Jin, P. J. (2015). Multi-mode Trip Information Recognition Based on Wavelet Transform Modulus Maximum Algorithm by Using GPS and Acceleration Data. In *Transportation Research Board 94th Annual Meeting* (No. 15-1411).

2. Pearson, D. (2004). A comparison of trip determination methods in GPS-enhanced household 26 travel surveys. Presented at 84th Annual Meeting of the Transportation Research Board, Washington, D.C.

3. Wolf, J., M. Oliveira, and M. Thompson. (2003). Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1854, No. 1, pp. 189-198.

4. Stopher, P. R., Q. Jiang, and C. FitzGerald. (2005). Processing GPS data from travel surveys. Presented at the 2nd International Colloqium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, Toronto.

5. Tsui, S. Y. A., and A. S. Shalaby. (2006). Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1972, No. 1, pp. 38-45.

6. Chung, E.-H., and A. Shalaby. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology,* Vol. 28, No. 5, pp. 381-401.

7. Rasmussen, T. K., J. B. Ingvardson, K. Halldórsdóttir, and O. A. Nielsen. (2013). Using wearable GPS devices in travel surveys: A case study in the Greater Copenhagen Area. In Proceedings of *the Annual Transport Conference at Aalborg University*. pp. 1603-9696.

8. Bohte, W., and K. Maat. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies,* Vol. 17, No. 3, pp. 8 285-297.

9. Wolf, J., S. Schonfelder, U. Samaga, M. Oliveira, K. W. Axhausen, and Trb. (2004). Eighty weeks of global positioning system traces-Approaches to enriching trip information. In *Data and Information Technology*, Transportation Research Board Natl Research Council, Washington. pp. 46-54.

10. Gong, H., C. Chen, E. Bialostozky, and C. T. Lawson. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems,* Vol. 36, No. 2, pp. 131-139.

11. Stopher, P., C. FitzGerald, and M. Xu. (2007). Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation, Vol., No. 6, pp. 723-741.

12. Forrest, T. L., and D. F. Pearson. (2005). Comparison of trip determination methods in household travel surveys enhanced by a Global Positioning System. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1917, No. 1, pp. 63-71.

13. Lee-Gosselin, M. E., S. T. Doherty, and D. Papinski. (2006) Internet-Based Prompted Recall Diary with Automated GPS Activity-Trip Detection: System Design. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C.

14. Sohn, K. (2004). *Dynamic estimation of origin–destination flows using cell phones as probes.* SDI 2004-R-04, Department of Urban Transportation, Seoul Development Institute, Korea.

15. Akin, D., and Sisiopiku, V.P. (2002). *Estimating origin–destination matrices using location information from cell phones*. Proc. 49th Annual North American Meetings of the Regional Science Association Int, Puerto Rico.

16. Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R., & González, M. C. (2015). ANALYZING CELL PHONE LOCATION DATA FOR URBAN TRAVEL: CURRENT 2 METHODS, LIMITATIONS AND OPPORTUNITIES 3. In Transportation Research Board 94th Annual Meeting (No. 15-5279).

17. Nour, A., Casello, J., & Hellinga, B. (2015). Developing and Optimizing a Transportation Mode Inference Model Utilizing Data from GPS Embedded Smartphones. In *Transportation Research Board 94th Annual Meeting* (No. 15-5027).

18. Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN), 6*(2), 13.

19. Thiagarajan, A., Biagioni, J., Gerlich, T., & Eriksson, J. (2010, November). Cooperative transit tracking using smart-phones. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems* (pp. 85-98). ACM.

20. Ahmed El-Geneidy, Zachary Patterson, and Evelyne St. Louis. *Transport and land-use interactions in cities: Getting closer to opportunities*, chapter 10, pages 175–193. Canadian Cities in Transition. University of Oxford Press, fifth edition, 2015.

21. http://journalmetro.com/actualites/montreal/719530/grand-montreal-maintenant-4-millions-de-personnes/

22. Matt Lerner (2014) Best Canadian cities for public transit, http://blog.walkscore.com/ 2014/03/best-canadian-cities-for-public-transit/#.VbkkzkV-pEZ. Accessed: 2015-08-29.