CIRRELT

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

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Understanding Transit use Patterns in Montreal

Aurélie Dubos-Golain
Martin Trépanier
Catherine Morency

October 2017

CIRRELT-2017-64

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

Understanding Transit use Patterns in Montreal

Aurélie Dubos-Golain[1], Martin Trépanier[2,3,*], Catherine Morency[2,4]

[1] École Nationale des Travaux Publics de l'État (ENTPE), Department of Transport, 3, rue Maurice Audin, 69518 Vaulx en Velin, France

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

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

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

Abstract. The aim of this paper is to use smart card data to understand the variability of public transit use. Weather variables are used to understand the variability of transit use, estimated by the daily distribution of boardings. This paper uses smart card data from the Montreal transit authority. Data collected during 202 days of 2016 are processed for this purpose. Clustering techniques are used to segment daily patterns of boardings on each bus line and subway station of the network. Results suggest that the variability in transit use is correlated with spatial location, weather and line purpose. Daily use of subway stations is mainly determined by the station's location. Namely, downtown stations have specific patterns: they have higher evening peak than the morning peak. Temperature is correlated with different bus lines with morning and evening peaks patterns as well as with boarding intensity at subway stations.

Keywords: Public transit, smart card fare collection systems, travel behaviour.

Acknowledgements. This work was made possible to the support and collaboration of the *Société de transport de Montréal*.

---

* Corresponding author: Martin.Trepanier@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2017

INTRODUCTION

Public transit planning is often based on household surveys and passenger counts. In the recent years, new data sources have become widely available namely because an increasing number of public transit authorities have implemented automated fare collection systems. The Montreal Transit Authority (called the STM) adopted such system in 2008. The data they gather constitutes an immense potential to develop more precise and complex transit forecasting demand models [1]. Many studies have focused on transit trips [2] and on the characteristics of the users [3]. To complement these studies, we propose to examine the use of the network itself, at the bus line and subway station level, on a daily basis. Analyzing the variability of transit usage across these network objects is useful to understand the typical usage patterns and the features related to such patterns and provide insights into how network design is correlated with transit use.

In this paper, we propose a method to characterize the use of the Montreal public transit network, though the proposed methodology can be applied to other public transit networks. The goal is to understand how network features are related to daily patterns of use and to provide significant elements to include in a transit forecasting demand models, at the bus line and subway station levels. For that purpose, our interest is to look at the influence of weather and other factors on the variability of transit daily usage.

The paper is organized as follows. Section 2 provides a review of the literature and positions our research. Section 3 presents the data sources used and the methods developed to construct typical daily patterns of use. The results are discussed in Section 4. Finally, section 6 concludes the paper.

BACKGROUND

In this section, we present an overview of previous work undertaken with smart card data in public transport planning, data mining tools and applications, analysis of transit use and the influence of weather on travel behavior.

Smart Card in Public Transport Planning

The technology of smart cards emerged in 1968. Implementations by transit agencies began in the mid-1990s [4]. Its use in public transit is perceived as a secure method for user validation [5].

Furthermore, the data are beneficial for transit planning purposes. Indeed, smart card systems collect get colossal datasets, for a large part of transit users over a continuous observation period. Three categories of transit planning purposes have been identified for smart card data. At the strategic level, the focus is on long-term network planning, and demand forecasting. Tactical-level studies are associated with schedule adjustment. Finally, operational-level analyses are related to supplying transit quality of service indicators [6]. This study is focused on the strategic level.

At this level, a significant amount of research has been done on user characterization and classification [7]. Researchers have no personal information on users [6]. However, the large amount of data collected provides a more precise understanding of user behavior (card level), as this system possesses more observations in space and time than any other data collection method [2]. However, smart card data do not provide any information about trip purposes or socio-economic data on the user, for confidentiality reason. To enrich this data,

studies need to also include data from traditional data collection methods such as origin-destination surveys [8] [9]. [10] [11] [12] help to adapt networks to user needs. Other studies provide information in real time on performance of the network [13].

Data Mining Tools and Applications

Data mining is a compilation of tools and techniques aimed to find previously unknown and potentially useful patterns from large datasets [14]. It may be divided into three sets: classification, segmentation and description [15]:

- classification arranges new data compared to the knowledge extracted from historical data;
- segmentation (or clustering) creates groups in a population;
- description and visualization are used to extract patterns from data.

Applications of data mining techniques are found in many areas such as marketing [16], manufacturing processes [17] and business process reengineering [18].

Analysis of Passenger Behavior

To understand users' behavior in a more comprehensive was, researchers have been studying day-to-day variability of travel behavior for forty years [19]. In the previous years, analyses were complicated to conduct because the available data only reported travels on a single day. Now, some studies follow passengers during years [20]. Passengers' habits are meaningful to transit operators because it can help predict ridership and its variability, providing grounds for better fitting of supply over demand. Some studies have sought to identify factors that influence users' habits, such as weather conditions [21], spatio-temporal dynamics [22] and change in passenger activity [23]. Most researches in this topic group the users with clustering approaches such as the K-means algorithm or the Hierarchical method [24] [2]. [25] classify travelers in groups of similar daily pattern.

Our paper proposes to contribute to the available body of knowledge by providing insights from another point of view: the transit use variability among network objects such as bus lines and subway stations.

Influence of Weather on Travel Behaviors

It is useful to understand the influence of weather on travel behaviors. [26] summarizes the major classes of meteorological variables likely to influence the use of all modes of transport: freezing conditions, rainfall, major thunderstorms, extreme temperatures, visibility, strong winds and others (air quality ...). The main weather variables are rainfall, snow fall and temperature. Various approaches to include such variables in correlation analysis are possible, namely using binary variables [27] and classes based on thresholds [28].

Studies show that weather conditions can lead to a shift in transportation modes, schedule or itinerary [29], [30]. [31] shows that traffic tends to decrease on the Irish public transport system during rainy days. Furthermore, a survey demonstrates that users change their mode of transportation due to bad weather conditions (snow, cold, rain, fog, storm) in Flanders, Belgium. 586 respondents completed the survey. [29].

METHODOLOGY

This section presents the data processing procedure. The study leans on transaction data from smart cards supplied by the STM. To see more clearly the influence of weather, we focus the analysis only on business days of 2016, excluding holiday periods: February 29 to March 4, 2016 (School period), July 1 to August 27, 2016 (typical Holiday period in Quebec), and December 23, 2016, to December 31, 2016 (Christmas / New Year Holiday). The analysis relies on 202 days of data. Furthermore, the daily operation hours are coded between 4 am and 28 am (12 AM + 4 hours).

In 2016, STM exploited 1771 buses and 828 subway cars across 68 subway stations and 233 bus lines. The bus network is divided into 5 sub-networks (related to type of service provided): Shuttle, Express, Gold, Local and Night. Shuttle is a bus service for tourists or provided during special events. Express lines are available during peak periods, often with more service in one direction. Gold routes are specially design bus services covering points of interest for elderly people. Local is the typical bus route.

Information System

The data were made available by the STM. In 2016, there was a total of 416 million trips on the transit network. This research uses 329,820,267 trips, which is the number of trips related to lines or stations during the selected business days (as mentioned previously). Since 2008, STM uses OPUS, a smart card system developed between 2003 and 2006. This system records the transactions made by cards and tickets and generates data which are further anonymized to preserve the confidentiality of the travelers.

Weather data come from Environment and Climate Change Canada Climate Services. The weather station used is located on the Montreal Island and generates daily data reports since 1992. For each day, the station provides: maximum, minimum and average temperature, heating and cooling degrees, rainfall, snow fall and total precipitation, snow on ground, direction and speed of maximum wind. This paper uses mean temperature, rainfall and snow fall.

Clustering Approach

Clustering techniques offer the opportunity to segment data in different groups. Data of each group share common characteristics. The general principle behind clustering is to minimize the distance between objects, as described by vectors of attributes. Many methods are available. Our clustering approach uses the k-means method, a data-mining technique partitioning vectors into k clusters. The Lloyd algorithm obtains a solution by assigning case i to the closest cluster according to a particular distance metric. When all cases are assigned, centroids are recalculated. These iterations are repeated until the centroids stop changing [32]. To avoid keeping a sub-optimal solution, the algorithm runs with 10 random initial conditions and the best solution is selected.

For all days and all lines, the number of daily boardings is transformed to approach a normal distribution using $f(x) = \ln(x + 10^{-17}x^5)$. The distributions of daily boarding before and after the transformation are presented in Figure 1. Data has been transformed into “line-day” vectors, where every record is the number of boarding on each line for each hour of each day. Each vector contains the daily intensity of boarding, normalized, in one variable and the distribution throughout the day using 24 variables, in proportion of the day (Table 1). Without this normalization process, the daily boarding weighed too much in the clustering process,

resulting in groups mostly based on daily boarding, hence neglecting the distribution throughout the day. Table 2 presents examples of vectors used for the clustering.

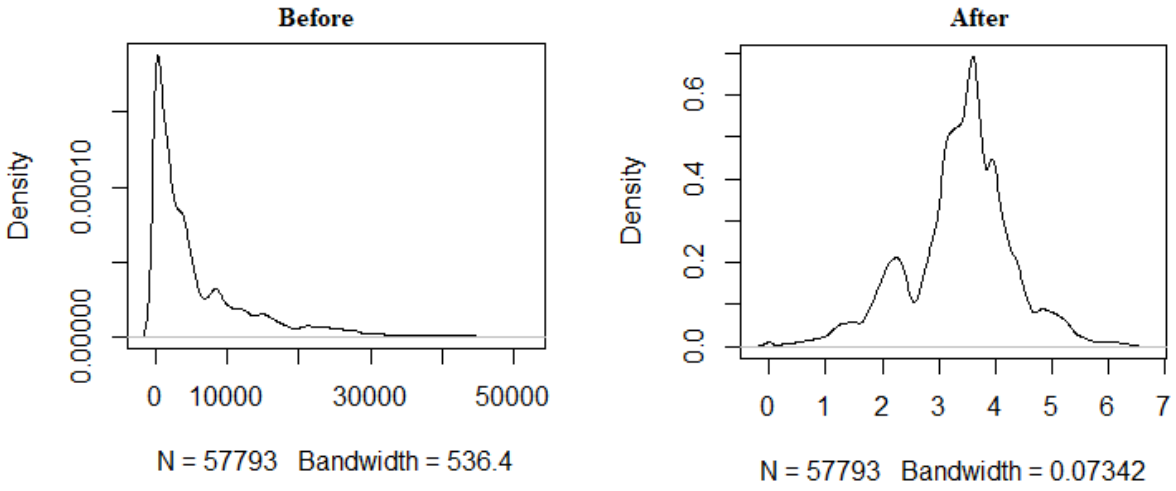

Figure 1. Distribution of the daily number of boarding before (left) and after (right) transformation

Table 1. Vector composition

	Variables	**Details**	**Example**
Classification variables	Daily boarding, normalized	Continuous variable, normalized by: $\frac{\frac{f(x)}{(1.5 * IQR)}}{MAX}$ Where: • x: total boarding per day • IQR: interquartile range of all daily boarding • MAX: maximum of $\frac{f(x)}{(1.5 * IQR)}$	Mean value: 0.517 Minimum value: 0 Maximum value: 1 Standard deviation: 0.62
	24 variables containing the proportion of daily boarding per hour	One variable for each hour of the day, continuous value between 0 and 1	0, 0, 0.02, 0.12, 0.31, 0.17, 0.26, 0.09, 0.02, 0.13, 0.32, 0.16, 0.25, 0.09

The main issue is to agree on a relevant number of clusters k. Selecting this number usually has some arbitrariness component. The choice is made in two steps. First, the k-means method is calculated on the dataset, with the number of clusters set to 20. Then, a dendrogram is constructed with the mean centers measured in the previous step. The choice of the best k for the clustering is done considering the largest step of this resulting dendrogram.

Table 2. examples of day-line vectors

Line/Stat	Date	4	5	6	7	8	9	10	11	12	13	14	15
10	04/01/2016	0	0.01	0.03	0.06	0.06	0.05	0.03	0.05	0.06	0.06	0.06	0.07
JOLIETTE	09/09/2016	0	0.02	0.04	0.13	0.15	0.08	0.05	0.04	0.04	0.05	0.05	0.06
809	08/12/2016	0	0.01	0.02	0.22	0.18	0.02	0.01	0.06	0.07	0.01	0.10	0.16
BERRI-UQAM	19/01/2016	0	0	0.01	0.02	0.03	0.02	0.02	0.03	0.09	0.06	0.05	0.07

16	17	18	19	20	21	22	23	24	25	26	27	Intensity
0.10	0.11	0.07	0.04	0.03	0.04	0.04	0.02	0.01	0	0	0	0.51
0.06	0.05	0.05	0.04	0.03	0.03	0.02	0.01	0.01	0	0	0	0.62
0.06	0.03	0.01	0.01	0.01	0.01	0	0	0	0	0	0	0.56
0.13	0.15	0.07	0.05	0.07	0.08	0.03	0.02	0.01	0	0	0	0.97

Weather Variables

The study focuses on the influence of mean temperature, rainfall and snow on daily patterns of transit usage. The objective is to validate whether these variables are correlated with changes in typical patterns of usage for specific lines (changes in cluster belonging). Actually, the clustering process generates clusters representing typical daily patterns. Hence, each line*day is associated with a particular cluster. Some lines may always belong to the same cluster (for all the days of observations) while others may change clusters during the observation period. To facilitate the identification of correlation between typical pattern of usage (i.e. a particular cluster) and weather-related variables, we focus on lines belonging to two different clusters during the year. We seek to observe whether this change in cluster belonging can be explained by weather-related variables. The temperature, rainfall and snowfall distributions of the six most important changes are contrasted between dominant (cluster with more days observed) and secondary clusters.

RESULTS

Number of Clusters

After various attempts, 8 clusters have been identified. In the following dendrogram (Figure 2), the first separation, in two groups, is made by the intensity value. The average of one leaf is 0.24 while the other is 0.54. The intensity values vary between 0.80 and 0. Various numbers of clusters could have been chosen in the light of this diagram, but we selected 8 clusters because that number was providing the desired granularity, allowing to get patterns of “day-line” in a sufficient number of clusters.

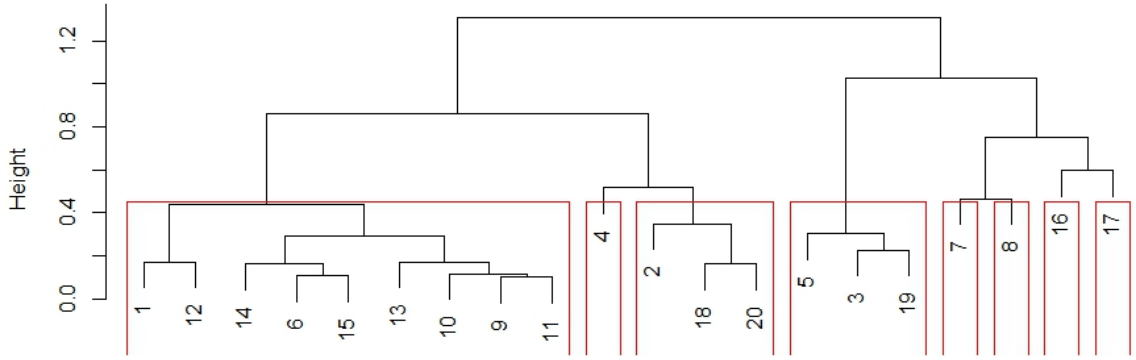

Figure 2 Clustering dendrogram

Cluster patterns

Figure 3 presents the daily patterns of each cluster as well as the analyzing the daily patterns reveals the emergence of different patterns of use. For each cluster, the daily distribution, in proportion, is presented along with the intensity indicator. The cluster with the highest intensity is Cluster 1 that presents a dispersed pattern of boarding throughout the day. Cluster 8 has quite a stable level of usage between 8AM and 3PM with a lower boarding intensity. Cluster 7 corresponds to evening activities with the peak at 11h pm. Cluster 6 represents the night-time activity with higher proportions of boarding between 2h-4h am.

Some clusters reveal commuting patterns with two important peaks. However, some differences can be noted. First, Cluster 5 has the highest morning peak and the lowest evening peak. Cluster 4 presents the typical two-peak usage pattern. Clusters 2 and 3 have almost the same pattern. It looks like a two-peak usage pattern but with some boarding between.

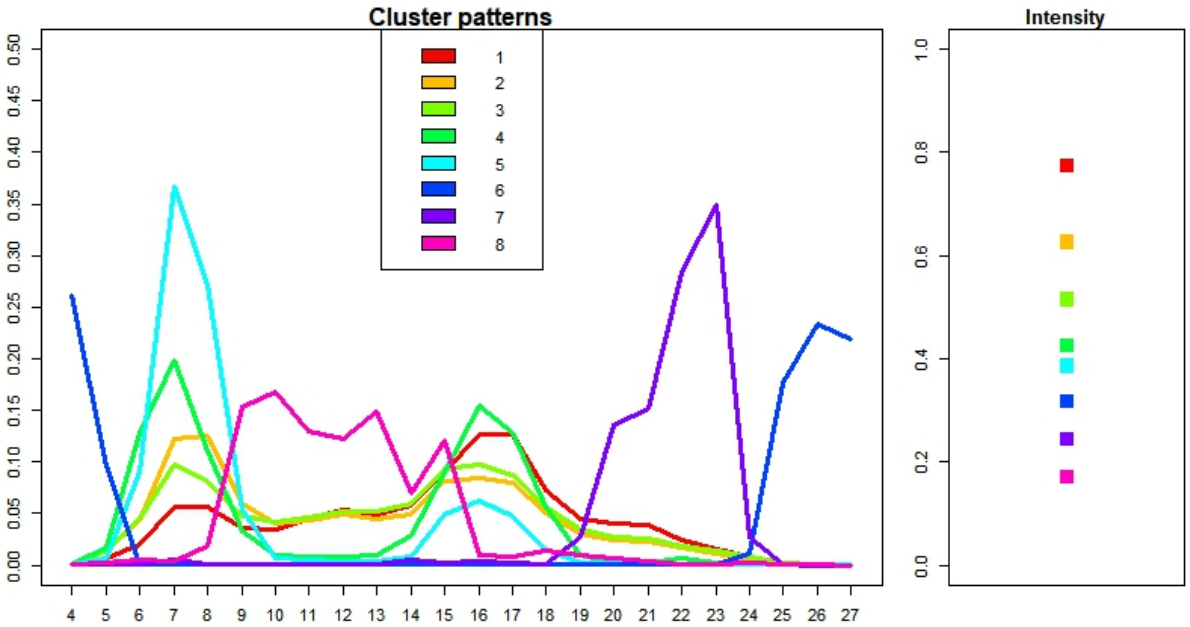

Figure 3. Daily patterns and boarding intensity for each cluster

Key Features of the Clusters

Table 3 presents the indicators used to describe the clusters while Table 4 presents the key features of these clusters, sorted per the normalized intensity in decreasing order .

Table 3 Indicators

Indicator	Explanation
% day-line	Percentage of day-line
Number of different lines	Number of lines and stations with at least one day in the cluster
Avg. number of days per line	Average number of days where lines and stations (which are at least one day in the cluster) are in the cluster
Avg. number of lines per day	Average number of lines and stations in a day for each cluster
% exclusive lines	% of lines and stations which are always in the same cluster

Table 4 Key features of clusters and networks

		Boarding Intensity	% day-line	Number of different lines	Avg. of days per line	Avg. of lines for each day	% exclusive lines
Clusters	**C1**	0.77	6.9%	32	124	19.7	34%
	C2	0.62	25.3%	107	137	72.4	34%
	C3	0.51	39.7%	172	133	113.5	39%
	C4	0.42	13.5%	75	104	38.5	14%
	C5	0.38	3.3%	41	46	9.3	2%
	C6	0.31	8.1%	28	167	23.1	64%
	C7	0.24	0.7%	13	34	2.1	7%
	C8	0.17	2.5%	29	51	7.3	27%
	AVG.	**0.43**	**12.5**	**62.1**	**99**	**35.8**	**28%**
Networks	**Or**	0.18	2.0%	10	118	6	70%
	Night	0.32	8.0%	23	202	23	100%
	Navette	0.36	3.0%	18	97	9	22%
	Express	0.50	10.8%	32	195	31	38%
	Local	0.52	52.3%	150	202	150	47%
	Subway	0.64	23.8%	68	202	68	65%
	AVG.	**0.42**	**17%**	**50**	**169.33**	**47.83**	**57%**

Figure 4 presents the frequency distribution of lines and stations based on the number of clusters to whom they belong throughout the year. It also represents how this distribution is distributed among the types of lines. Figure 5 presents the distribution of days-line among networks and clusters.

Clusters have different percentages of exclusive lines: they vary between 2% and 64% (see Table 4). Figure 4 shows that 154 lines (approximately 51%) are always in the same cluster (for all the days observed) and more than 90% are in two clusters or less. It confirms that usage is quite stable for some services (in terms of temporal distribution and intensity level). Lines allocated to many different clusters (5 and 6 different clusters) belong to the Shuttle network. In the light of Figure 5, Shuttle lines appear in all clusters except the one with the highest daily

intensity: these lines provide very specific services, either for touristic activity or for special events (soccer game for instance) and have more variable levels of usage.

Figure 5 presents to which clusters the various types of service belong.

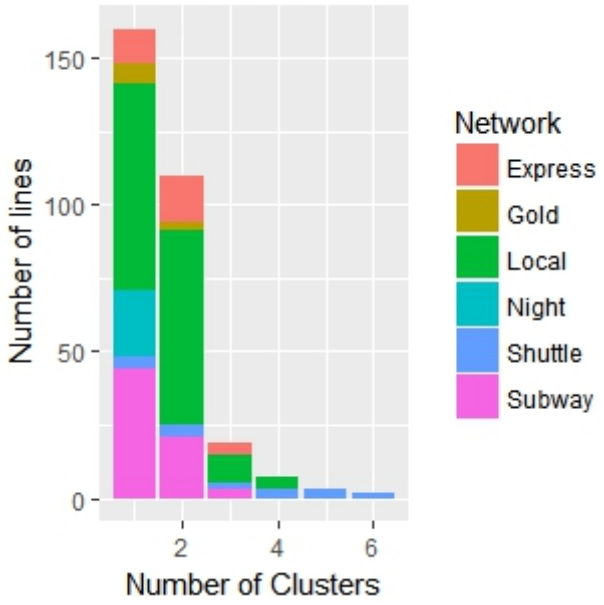

Figure 4. Number of lines in different clusters

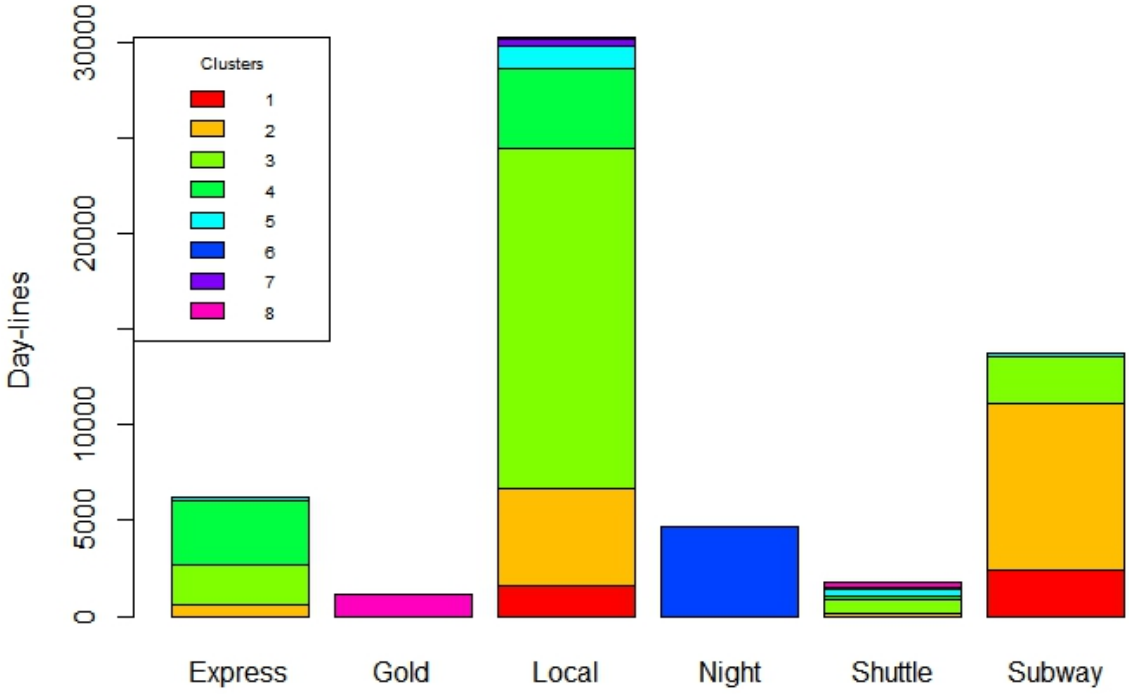

Figure 5. Clusters distribution by service type

Some clusters are specific to a particular type of service as seen for clusters 8 and 6. Cluster 8 represents the Gold network. Only 27% of the Gold lines always belong to the same cluster. Figure 5 shows that Cluster 8 is made of Gold and Shuttle services. Gold network has 70% of exclusive lines so the low percentage of Cluster 8 is due to the Shuttle lines. As

explained previously, Shuttle lines have variable levels of usage. Cluster 8 has the lowest boarding intensity. According to Table 4, Gold network also has the lowest boarding intensity, so it seems that elderly people have lower travel intensity and also rely on other types of service for their travel needs. Cluster 6 represents the Night network; it is composed of Night and Shuttle services (see Figure 5). It has the highest percentage of exclusive lines (64%). This high percentage is due to the Night network with 100% of exclusive lines.

Cluster 7 has a lower percentage of exclusive lines (7%) and it is composed of 13 Local and Shuttle lines according to Table 4. It has the lowest average of days per line (34 days) and average of lines per day (2). As this cluster is mainly composed of two lines, which run only at 10h pm and 12h pm, its pattern is mostly created by the structure of the transit supply. In the light of Figure 5, Local, Express and Shuttle lines mostly compose Cluster 5. It has the lowest percentage of exclusive lines (2%) and lines have this pattern only 46 days out of 202 observed, so these lines do not have this high intensity morning peak every day.

Cluster 2, 3, 4 represent more than half of Express, Local and subway network, respectively. Cluster 3 has the highest number of different lines and exclusive lines. Even if the percentage of exclusive lines is higher for Cluster 3 (39%) than Cluster 2 (34%), the latter has a higher average of days per line (137 compared to 133 days), so lines from Cluster 2 are more stable (in terms of temporal distribution and intensity level) than lines from Cluster 3, so it seems that subway stations have more stable patterns that Local lines.

Cluster 1 is composed of Local and subway stations. It is the highest boarding intensity cluster (0.77). It is higher than Local and Subway boarding intensity, so it seems that Cluster 1 Local lines have a higher intensity than some subway stations.

The Express network has a lower boarding intensity (0.50) than the Local one (0.52), which is related to the fact that express lines usually only operate during peak periods while local lines are in operation the entire day. Subway stations are almost exclusively in the three highest intensity clusters, although some bus lines are also in these clusters. These lines therefore have the same boarding intensity as subway stations and hence play a major role in the provision of transit service in Montreal. This is also explained by the fact that only boarding patterns are accounted for, and not alighting ones.

Impact of Weather on Clusters

As indicated by Figure 4, there are 109 lines which are in two clusters. For each line, a dominant cluster and a secondary cluster are identified, the dominant one containing most of the days. Table 5 shows the average value for each of the weather variables, for both the dominant and the secondary clusters. The p-values are determined by the Wilcoxon signed-rank test. He allows to test the hypothesis that the distribution of the data is the same in two groups. Some changes are only for one day-line but the six changes with the highest number of day-lines have more than 200 secondary day-lines. All changes have different average weather variables between dominant and secondary cluster. But they are not all significant.

Table 5. Weather variables for each change (DOM=dominant, SEC=secondary

Cluster number		Number of secondary day-lines	Average temperature (°C)			Average Rainfall (mm)			Average Snowfall (cm)		
DOM	SEC		DOM	SEC	p-value	DOM	SEC	p-value	DOM	SEC	p-value
3	2	705	6.70	1.45	4.01E-39	1.82	2.93	0.290	0.56	1.09	1.25E-32
2	3	551	5.46	10.23	2.18E-28	2.04	1.49	0.533	0.66	0.44	2.31E-10
4	3	338	5.49	8.95	1.42E-08	2.02	1.73	0.463	0.67	0.45	0.00034
2	1	261	6.27	3.73	8.11E-05	2.05	1.52	0.240	0.65	0.58	0.04206
4	5	225	5.66	7.16	0.04159	2.05	1.70	0.635	0.67	0.48	0.34042
3	4	210	6.10	4.27	0.01602	1.95	2.32	0.886	0.63	0.75	0.14004
1	2	166	5.54	8.92	6.95E-06	1.93	2.37	0.157	0.65	0.56	0.00152
1	3	47	3.44	14.19	6.55E-11	2.31	0.91	0.888	0.81	0.09	0.00054
5	4	36	5.62	11.17	0.00277	2.05	0.99	0.591	0.64	0.57	0.1623
4	2	9	5.92	6.81	0.67866	2.02	0.51	0.766	0.65	0.04	0.43135
3	1	5	6.00	1.26	0.32289	1.94	5.32	0.791	0.64	0.32	0.97593
3	8	3	5.83	12.97	0.20276	1.82	13.13	0.003	0.65	0.00	0.38763
3	5	1	5.98	-1.50	0.39118	1.99	0.00	0.486	0.62	4.00	0.03172
3	7	1	19.19	4.70	0.12036	1.64	3.00	0.288	0.00	0.00	NA
4	8	1	19.55	-8.40	0.66667	4.20	0.00	0.405	0.00	0.00	NA
8	3	1	5.97	17.50	0.25861	1.34	0.00	0.513	0.67	0.00	0.66481
8	5	1	5.85	17.60	0.24613	1.41	0.00	0.513	0.66	0.00	0.66481
8	6	1	5.91	1.60	0.599	1.57	0.40	0.383	0.66	6.40	0.02054

One change comes with the rainfall (Clusters 3 to 8). This rainfall involves a strong drop in boarding intensity. Snowfall has impacted 8 changes. The impact is not the same for all changes. Further, six changes are due to the temperature and snowfall and three are only due to temperature. All changes between Cluster 1 and lower boarding intensity clusters are associated to an increasing temperature. The same happens with Cluster 2. As Clusters 1 and 2 represent mainly subway stations, this may mean that subway stations are more used during the winter period.

Figure 6 represents the weather distribution between dominant and secondary cluster of the six changes with the highest number of secondary day-lines with rainfall distribution because Table 5 shows that rainfall is insignificant for these six changes. It highlights the fact that clusters changes are also differentiated by weather variables. The temperature is a significant variable for the six changes and snowfall for the first four changes. The six changes are from the five highest boarding intensity clusters. They have a pattern with morning and evening peaks. Their main difference is the boarding intensity. All the changes show a drop-in intensity when moving from the dominant to the secondary cluster, related to an increase in average temperature except for the change between C4 and C3. This drop might be due to the availability of bikesharing service and the use of active modes during the warmer months. Indeed, some users might prefer to use these modes instead of transit or people simply reduce their trip rates. The change between clusters C3 and C4 could be explained by the fact that with cold temperature, users might use public transit only for home to work travels and limit their other types of activities.

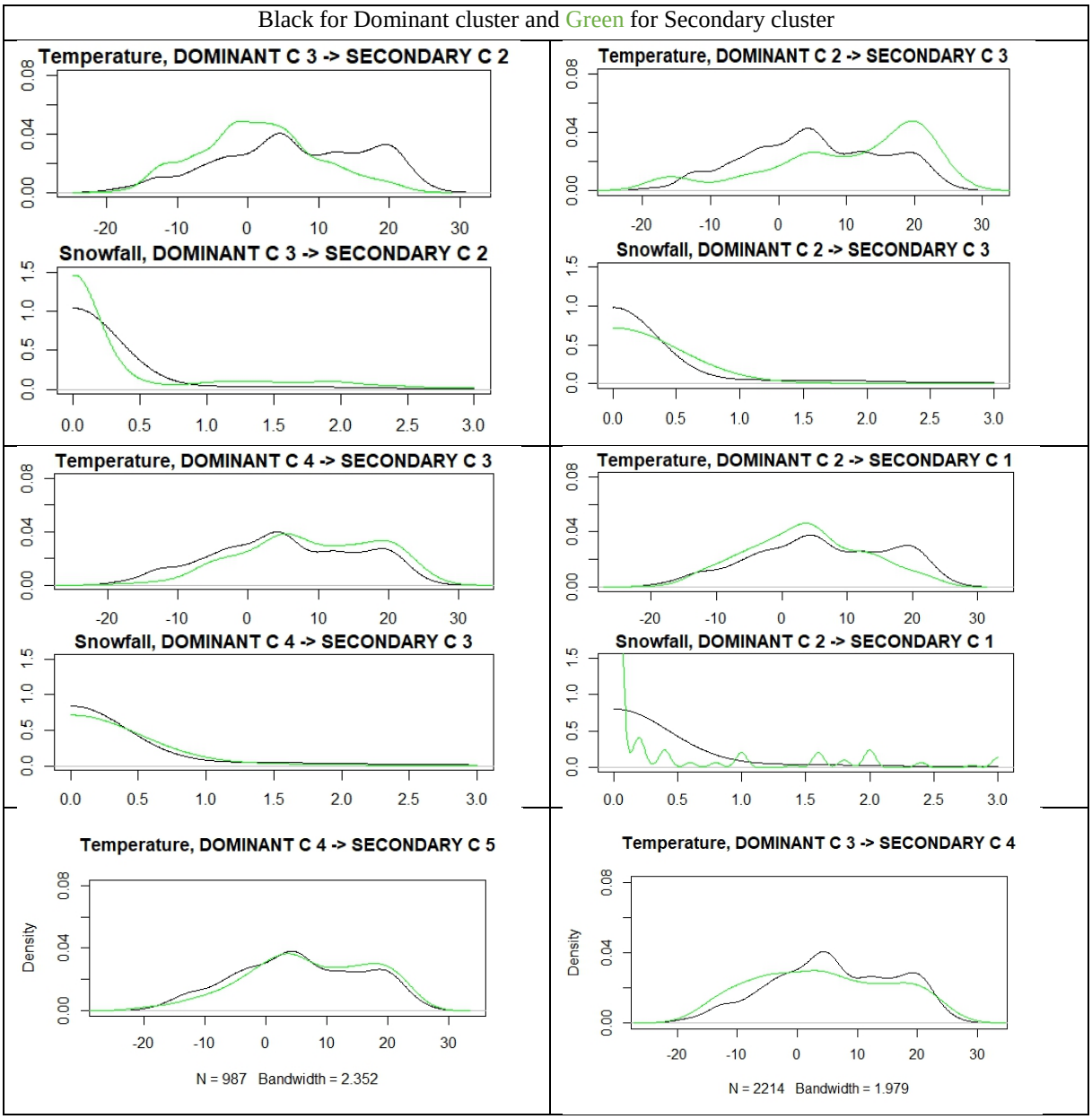

Figure 6. Weather distributions

Spatial Distribution

As noted previously, Gold and Night lines belong mostly to one cluster. Shuttle lines are clustered because of their inherent purpose, so their clusters are not due to the spatial dispersion of use. Subway station, Local and Express networks could be influenced by spatial distribution. Figure 7 shows the spatial distribution of subway clusters for one day of April 2016. The cluster with the highest intensity is related to subway stations located in the downtown area while most of the other subway stations are in the cluster with the second-highest intensity level (Cluster 2). Cluster 2 stations have different boarding intensities so they are in the same cluster because they have the same pattern. Downtown stations have higher levels of use during the evening peak while travelers board the subway network to head back home and this is the opposite for Cluster 2 stations. Some Cluster 3 stations have the same boarding intensity as some Cluster 2 stations so these C3 stations may be located in work and residential areas. They are used with the same intensity in morning as in evening peak time.

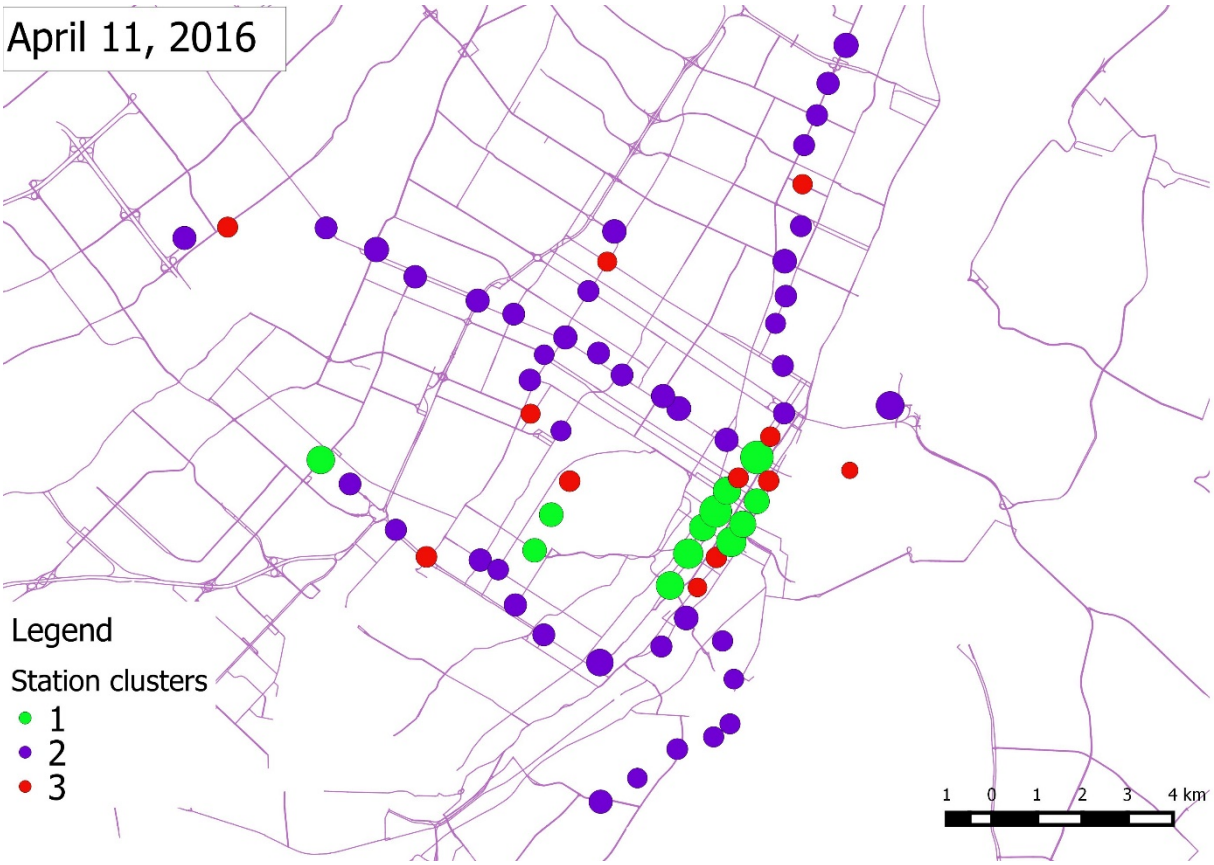

Figure 7 Subway stations by clusters and boarding intensity of April 11

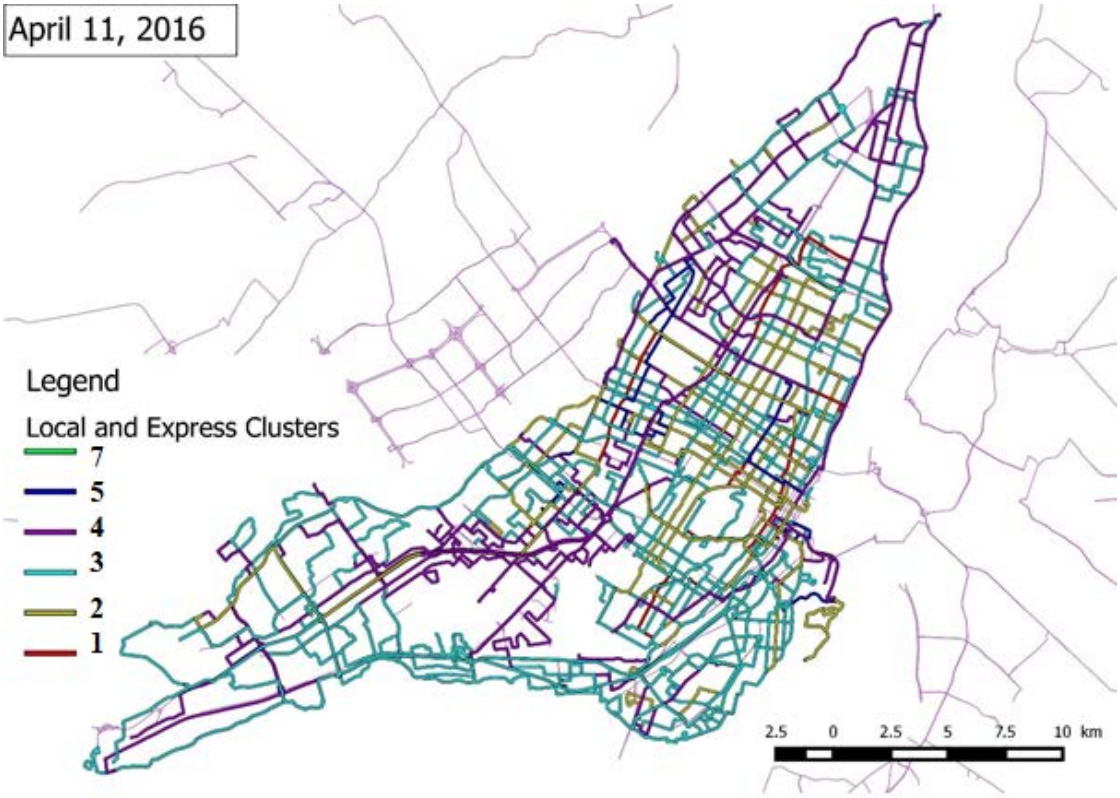

Figure 8 Bus lines by clusters and boarding intensity of April 11,2016

Figure 8 represents the spatial distribution of Local and Express line clusters for one day of April 2016. Cluster 4 lines are lines starting from the ends of the island so their pattern is probably due to the commuting patterns of people traveling to work.

CONCLUSION

This paper presented an analysis involving the identification of clusters for daily patterns of transit use in Montreal during business days. The clustering was performed based on boarding intensities and daily patterns of usage on bus lines and subway stations. A weather characterization was also performed. This characterization allowed for a better understanding of the relation between typical usage patterns and weather conditions.

A detailed analysis of the clusters shows that typical commuting patterns (with AM/PM peak periods) clearly appear only in few clusters, the peak more often showing different intensities. Other clusters show more dispersed daily distribution of the boardings as well as unequal intensity at peaks. Clusters also vary according to the daily intensity of boardings and some lines switch clusters during the year namely moving to lowest or highest intensity clusters with similar daily patterns. Some clusters have totally different patterns, which represent trips during specific periods (evening, night), from particular segments (elderly people) or to specific locations (touristic activities and events).

The weather distributions reveal that patterns of usage of bus lines and subway stations change due to temperature and rainfall, although not all of them are affected and mostly, not of them showing similar influence. Subway seems to be more used in winter and patterns of some bus lines will change of daily patterns when it is colder, showing a more restricted two-peak pattern.

Downtown subway stations belong to the same cluster with the highest boarding intensity. Stations are clustered by their location, their area. The outermost bus lines are used only for home to work travels. Highest boarding intensity bus lines deserve the downtown and they are used for home to work travels and for leisure travels.

Further analyses are needed to better understand the public transit use. The clustering approach gives an objective way of characterizing public transit network based on the daily patterns and intensity of boardings. Still, there are some limitations: the number of clusters is partly arbitrary as well as the classes used to define weather features. It would be interesting to investigate the choice of the number of clusters if we intend to highlight public transit network with atypical use. Furthermore, using several years of data would help to better understand the influence of weather as well as longer-term evolution in the level and patterns of usage. With more data, the weather variables could be chosen with objective values. To this end, a model should be developed to quantify the influence of all these variables.

ACKNOWLEDGMENT

This work was made possible, thanks to the support and collaboration of the *Société de transport de Montréal*.

REFERENCES

[1] M. Bagchi et P. R. White, « The potential of public transport smart card data », *Transp. Policy*, vol. 12, nº 5, p. 464-474, sept 2005.

[2] B. Agard, C. Morency, et M. Trépanier, « MINING PUBLIC TRANSPORT USER BEHAVIOUR FROM SMART CARD DATA », *IFAC Proc. Vol.*, vol. 39, nº 3, p. 399-404, janv. 2006.
[3] B. Agard, V. Partovi Nia, et M. Trépanier, « Assessing public transport travel behaviour from smart card data with advanced data mining techniques », in *World Conference on Transport Research*, 2013, vol. 13, p. 15-18.
[4] P. T. Blythe, « Improving public transport ticketing through smart cards », in *Proceedings of the Institution of Civil Engineers-Municipal Engineer*, 2004, vol. 157, p. 47-54.
[5] M. Trépanier, S. Barj, C. Dufour, et R. Poilpré, « Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain », *Congrès Assoc. Transp. Can.*, 2004.
[6] M.-P. Pelletier, M. Trépanier, et C. Morency, « Smart card data use in public transit: A literature review », *Transp. Res. Part C Emerg. Technol.*, vol. 19, nº 4, p. 557-568, 2011.
[7] E. I. Vlahogianni, M. G. Karlaftis, et J. C. Golias, « Short-term traffic forecasting: Where we are and where we're going », *Transp. Res. Part C Emerg. Technol.*, vol. 43, p. 3-19, 2014.
[8] M. A. Munizaga et C. Palma, « Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile », *Transp. Res. Part C Emerg. Technol.*, vol. 24, p. 9-18, 2012.
[9] M. Trépanier, C. Morency, et B. Agard, « Calculation of Transit Performance Measures Using Smartcard Data », *J. Public Transp.*, vol. 12, nº 1, mars 2009.
[10] Y. Wei et M.-C. Chen, « Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks », *Transp. Res. Part C Emerg. Technol.*, vol. 21, nº 1, p. 148-162, 2012.
[11] K. Alfred Chu et R. Chapleau, « Enriching Archived Smart Card Transaction Data for Transit Demand Modeling », *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2063, p. 63-72, déc. 2008.
[12] K. Chu, R. Chapleau, et M. Trépanier, « Driver-Assisted Bus Interview », *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2105, p. 1-10, oct 2009.
[13] E. Deakin et S. Kim, « Transportation Technologies: Implications for Planning », *UC Berkeley Univ. Calif. Transp. Cent.*, 2001.
[14] S. S. Anand et A. G. Büchner, *Decision support using data mining*. Financial Times Management, 1998.
[15] C. Westphal et T. Blaxton, « Data mining solutions: methods and tools for solving real-world problems », 1998.
[16] M. J. Berry et G. Linoff, *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.
[17] B. Agard et A. Kusiak, « Data Mining for Selection of Manufacturing Processes », in *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 2005, p. 1159-1166.
[18] C. da Cunha et B. Agard, « Business Process Reengineering with Data Mining in Real Estate Credit Attribution: a Case Study », in *Intl. Conf. on Information Systems, Logistics and Supply Chain–ILS 2006*, 2006, p. 15-17.
[19] P. Jones et M. Clarke, « Significance and Measurement of Variability in Travel Behaviour: A Discussion Paper », 1987.
[20] A.-S. Briand, E. Côme, M. Trépanier, et L. Oukhellou, « Analyzing year-to-year changes in public transport passenger behaviour using smart card data », *Transp. Res. Part C Emerg. Technol.*, vol. 79, p. 274-289, 2017.

[21] P. Arana, S. Cabezudo, et M. Peñalba, « Influence of weather conditions on transit ridership: A statistical study using data from Smartcards », *Transp. Res. Part Policy Pract.*, vol. 59, p. 1-12, janv. 2014.
[22] S. Tao, D. Rohde, et J. Corcoran, « Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap », *J. Transp. Geogr.*, vol. 41, p. 21-36, déc. 2014.
[23] K. K. A. Chu, « Two-year worth of smart card transaction data–extracting longitudinal observations for the understanding of travel behaviour », *Transp. Res. Procedia*, vol. 11, p. 365-380, 2015.
[24] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, et J. Liu, « Mining smart card data for transit riders' travel patterns », *Transp. Res. Part C Emerg. Technol.*, vol. 36, p. 1-12, 2013.
[25] E. I. Pas et F. S. Koppelman, « An examination of the determinants of day-to-day variability in individuals' urban travel behavior », *Transportation*, vol. 14, nº 1, p. 3-20, 1987.
[26] A. Handman, « Weather implications for urban and rural public transit », présenté à Second Annual Users Conference, 2004.
[27] S. A. Changnon, « Effects of summer precipitation on urban transportation », *Clim. Change*, vol. 32, nº 4, p. 481-494, 1996.
[28] J. H. Hogema, « Effects of Rain on Daily Traffic Volume and on Driving Behaviour (Effecten van regen op verkeersvolume en op rijgedrag). », HOOFDGROEP MAATSCHAPPELIJKE TECHNOLOGIE TNO DELFT (NETHERLANDS), 1996.
[29] A. J. Khattak et A. De Palma, « The impact of adverse weather conditions on the propensity to change travel decisions: a survey of Brussels commuters », *Transp. Res. Part Policy Pract.*, vol. 31, nº 3, p. 181-203, 1997.
[30] M. Cools, E. Moons, L. Creemers, et G. Wets, « Changes in Travel Behavior in Response to Weather Conditions », *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2157, p. 22-28, sept. 2010.
[31] M. Hofmann et M. O'Mahony, « The impact of adverse weather conditions on urban bus performance measures », in *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, 2005, p. 84-89.
[32] L. Morissette et S. Chartier, « The k-means clustering technique: General considerations and implementation in Mathematica », *Tutor. Quant. Methods Psychol.*, vol. 9, nº 1, p. 15-24, févr. 2013.