POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Yearly patterns of transit usage: a cumulative clustering approach using 7 years of smart card data
Auteurs: Authors:	Jean-Simon Bourdeau, & Catherine Morency
Date:	2020
Type:	Communication de conférence / Conference or Workshop Item
Référence: Citation:	Bourdeau, J.-S., & Morency, C. (mai 2020). Yearly patterns of transit usage: a cumulative clustering approach using 7 years of smart card data [Communication écrite]. 12th International Conference on Transport Survey Methods, Lisbon, Portugal. Publié dans Transportation Research Procedia, 76. https://doi.org/10.1016/j.trpro.2023.12.035

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/57274/
Version:	Version officielle de l'éditeur / Published version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	CC BY-NC-ND

Document publié chez l'éditeur officiel
Document issued by the official publisher

Nom de la conférence: Conference Name:	12th International Conference on Transport Survey Methods
Date et lieu: Date and Location:	2020-05-31 - 2020-06-05, Lisbon, Portugal
Maison d'édition: Publisher:	Elsevier
URL officiel: Official URL:	https://doi.org/10.1016/j.trpro.2023.12.035
Mention légale: Legal notice:	

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 76 (2024) 25–36

Transportation Research Procedia

www.elsevier.com/locate/procedia

12th International Conference on Transport Survey Methods

Yearly patterns of transit usage: a cumulative clustering approach using 7 years of smart card data

Jean-Simon Bourdeau [a] and Catherine Morency[a]*

[a] *Polytechnique Montréal, Chaire Mobilité and Chaire de recherche du Canada sur la mobilité des personnes*

Abstract

This paper presents results of the application of clustering methods to report on the yearly patterns of transit ridership, at metro stations and on bus lines, as observed using 7 years of smart card validations (2015-2021). Results from the k-means approach are presented in this paper while various methods were tested using the same vectors of data. The results confirm the ability of the proposed approach to understand the global changes in ridership patterns. The cumulative clustering approach used allowed to understand the evolution of both bus and metro across the seven years from a temporal and spatial perspective.

© 2023 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the International Steering Committee for Transport Survey Conferences (ISCTSC)

Keywords: smart card data; longitudinal data; clustering; yearly ridership; transit usage

1. Introduction

Smart card data have been around for quite some years now and many researchers have demonstrated their usability in the context of network analysis and planning. While survey data allow to capture more in-depth understanding of the personal travel behaviours, passive data such as those provided by smart card validations allow to monitor either variability of personal use of transit services (when individual cards are analysed) or variability of transit use at the system level. Understanding the analytical opportunities of time-series data at the micro (card), meso (network objects) and macro (system) levels is key for transportation researchers and operators.

The Montreal transit systems is equipped with the OPUS card which is a tap-in only system: hence, users validate their cards when they board a bus or when they enter a subway station but do not need to validate it when they exit.

* Corresponding author. *E-mail address:* cmorency@polymtl.ca

2352-1465 © 2023 The Authors. Published by ELSEVIER B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the International Steering Committee for Transport Survey Conferences (ISCTSC)
10.1016/j.trpro.2023.12.035

Moreover, while it is possible in some systems to impute the destination stop using multiple days of observation and spatial-temporal logics, there is still no usable GPS data available for the Montreal bus system, limiting the possibility to impute destination points.

With these features in mind, this paper focuses on the processing of boarding validations and aims at identifying typical yearly patterns of transit usage. With the increasing expectations towards transit in the war against climate change, it is important to understand the evolution of transit usage and the usage patterns of its various components, to be able to anticipate how transit usage will fluctuate. Hence, in this paper, the focus is on the structure of yearly usage of the subway and bus network objects (stations and lines) and the contribution of each week in the overall yearly ridership. Seven years of data are used to observe the weekly contribution of boardings to the yearly ridership.

This paper first proposes some background elements regarding research on the variability of transit use as well as on various clustering methods that can be useful to process time-series data. It then proposes some information on the methodology developed namely the data used to analyse yearly ridership as well as the data preparation steps. Results of the clustering process (only from k-means) are presented and discussed. A conclusion closes this short paper.

2. Background

Smart cards have been used in various research in the last 20 years. The paper by Pelletier et al. (2011) is still one of the most structured and exhaustive review of the use of such data to support transit network analysis. There are plenty more recent papers demonstrating how to value such longitudinal data to better understand the variability of transit usage. For instance, Deschaintres et al. (2019) used 51 weeks of transit usage to analyse the variability of user behaviors; clustering methods were used to create typologies of usage at the card level. Viallard et al. (2019) also used clustering method to analyse the evolution of travel patterns among card users. Tao et al. (2014) also used smart card data to examine the spatio-temporal dynamics of bus passengers. Smart card data has also been used to develop visual segmentation methods (Ghaemi et al., 2018) or to classify public transit users (Ortega, 2013; Ma et al., 2013). Egu and Bonnel (2020) propose an analysis of the variability of transit usage based on smart card data using clustering and similarity metric; their dataset covers 6 months. Manley et al (2018) also examine variability of transit usage but benefit from tap-in/tap-out data that allows them to examine variability over space and time. DBSCAN is used to identify clusters of travel activities from cards. Goulet-Langlois et al. (2018) provide another example of transit use variability analysis based on smart-card data. They propose a regularity metric based on entropy able to identify both typical and atypical routines.

For this research, several clustering methods are experimented. First, two partitioning algorithms are used with a predefined number of clusters: k-means clustering (MacQueen, 1967) and the PAM clustering, also known as k-medoids (Kaufman & Rousseeuw, 1990). Hierarchical clustering is also used, since it does not require the number of clusters to be specified (Hartigan, 1975). More advanced methods are also explored: fuzzy clustering and model-based clustering (Fraley & Referty, 2002, Fraley & al., 2012) which are soft clustering methods, and density-based clustering that can be used to identify any shape of clusters (Ester & al., 1996). While all these methods were tested, only results obtained using the k-means clustering approach are discussed in this paper due to space constraint.

3. Methodology

3.1. Data

The data available consists in the number of validations for intervals of 1 hour for each day of every year, and for each metro station or bus line. The available data covers a 7 years' period, from January 1st, 2015, to December 31st, 2021. There is a total of 206 bus lines and 68 metro stations over the analysis period (that were in operation in the full set of years). The number of validations for the bus and metro networks are presented in Table 1. The volume of bus validations is slightly decreasing from 2015 to 2019 (this may partly be since some important lines connected to subway stations are not requiring validations anymore to allow entering by multiple doors), while the volume of metro validations is slightly increasing. After 2020, the effects of the COVID-19 pandemic appeared on both bus and metro validations. It also appears that metro validations have seen a sharper reduction than bus validations since 2020.

Table 1. Number of bus and metro yearly validations over the analysis period

Year	Number of bus validations (% vs previous year)	Number of metro validations (% vs previous year)	Total number of validations (% vs previous year)
2015	234 906 013	245 575 946	480 481 959
2016	(-3.4%) 226 970 476	(+0.6 %) 247 038 482	(-1.3 %) 474 008 958
2017	(-1.4 %) 223 886 876	(+2.9 %) 254 255 849	(+0.9 %) 478 142 725
2018	(+0.5 %) 225 114 021	(+6.0 %) 269 401 649	(+3.4 %) 494 515 670
2019	(-2.7 %) 218 953 067	(+4.3 %) 280 953 740	(+1.1 %) 499 906 807
2020	(-60.9 %) 85 651 566	(-59.1 %) 115 005 432	(-59.9 %) 200 656 998
2021	(+31.4 %) 112 512 634	(+0.4 %) 115 507 685	(+13.6 %) 228 020 319
Total	1 327 654 753	1 527 738 783	2 855 733 436

3.2. Data Preparation

The first step is to aggregate the smart card data. For each metro station and bus line, an annual cumulative vector is computed at the week (j) level, with equation (1):

$$cumulative\ validations_{j,y} = \frac{\sum_{j=1}^{j} validations_j}{validations_y}$$ (1) where j is the week number and y is the year

This means that for each line/station, the value of the vector is 0 on the first week of the year and 1 on the last week of the year. Some years have 52 weeks, and some years have 53 weeks, the week definition is based on the ISO week format.

Another vector is created to integrate the relative intensity of each week, with the following equation (2). The second week is chosen as reference since the first week of the year is not typical with new year's celebrations and holidays. If the intensity is lower than 1, there was a lower number of validations during this week than during the first week.

$$weekly\ intensity_j = \frac{\sum validations_j}{\sum validation_2}$$ (2) where j is the week number and varies between 1 and 52

This means that for a given week, if the weekly intensity is greater than 1, there was a greater number of validations during this week than during the second week of the year.

The dataset created for the clustering is shown in Table 2. It consists of the metro station (or bus line), year, weekly intensity (52 columns) and cumulative validations (52 columns). For the metro stations, 99.4 % of the vectors were complete, meaning that every week of the year had validations, while for the bus network 93.8 % of the vectors were complete. There is a total of 99 The intensity of week #2 is removed when performing the clustering since it is a vector with null distances between all observations. Also note that for the cumulative validations, week 52 might not be exactly 1 since some years have 53 weeks. Week 53 is not included in the dataset since it contains missing data (incomplete week).

Table 2. Example of the dataset created for the cluster

		intensity				cumulative		
metro station	year	week 1	week 2	…	week 52	week 1	…	week 52
ACADIE	2015	0.30	1	…	0.79	0.01	…	0.99
ACADIE	2016	0.21	1	…	1.03	0.00	…	0.99
ACADIE	2017	0.06	1	…	1.33	0.00	…	0.99
ACADIE	2018	0.63	1	…	0.57	0.01	…	1.00
ACADIE	2019	0.56	1	…	0.71	0.01	…	1.00
ACADIE	2020	0.37	1	…	0.25	0.02	…	0.99
ACADIE	2021	0.23	1	…	1.41	0.00	…	0.99

Once the vectors of cumulative validations and monthly intensity are produced, they can be used as input for each of the clustering methods described in the background. These methods are summarized in Table 3. Soft algorithms imply that each record inherits a probability of belonging to every cluster instead of a single assignment to one of the clusters. Also, some methods require the desired number of clusters to be specified as an input. Finally, the DBSCAN method is the best clustering method to use when the shape or the size of the clusters is not constant.

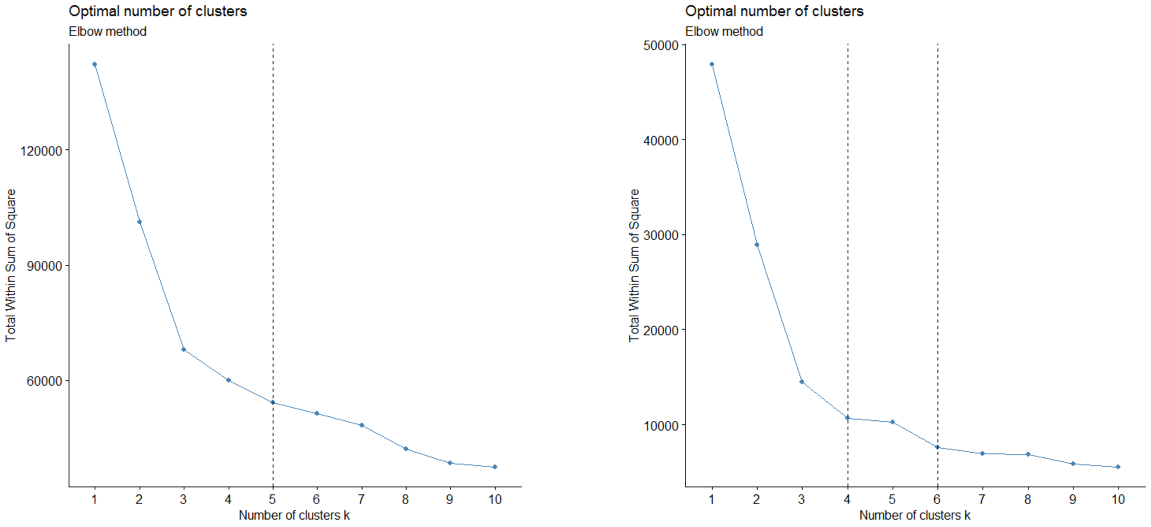
Table 3. Summary of the clustering methods tested

Clustering method	**Soft/hard algorithm?**	**Predefined number of clusters**	**Shape of the clusters**	**Use for large datasets**
K-means	Hard	Yes	Spherical/convex	Yes
PAM / k-medoids	Hard	Yes	Spherical/convex	Yes, see CLARA algorithm
Hierarchical	Hard	No	Spherical/convex	No
Fuzzy	Soft	Yes	Spherical/convex	Yes
Density-based clustering (DBSCAN)	Hard	No	Any	Yes

4. Results

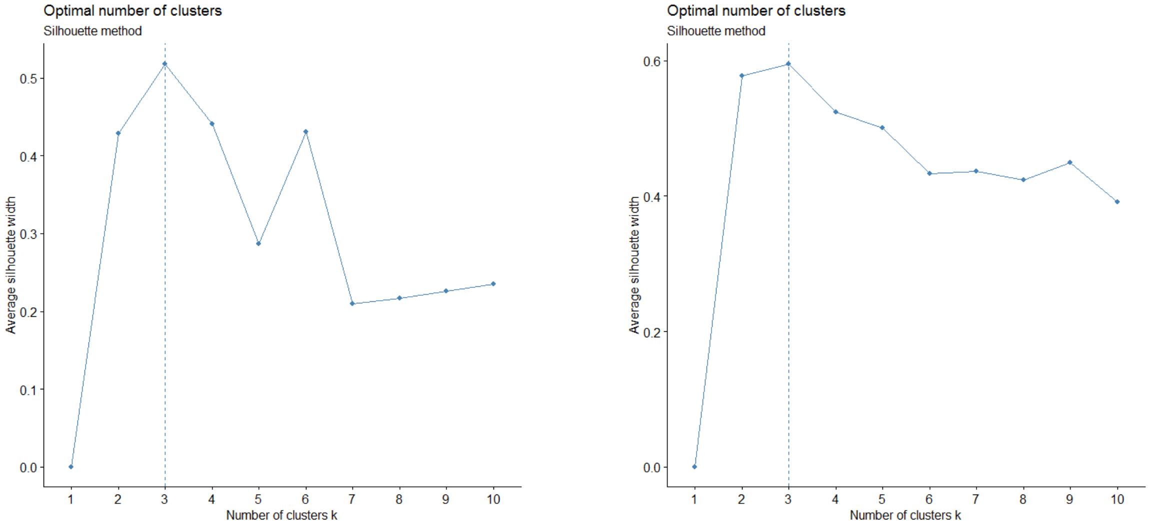
4.1. Setting the number of clusters

To determine the optimal number of clusters for the k-means and PAM methods, two methods are used, the elbow method and the silhouette method. For the elbow method (see Figure 1) , the optimal number of clusters is 5 for the bus lines and 4 or 6 for the metro stations. For the silhouette method (see Figure 2), the optimal number of clusters is 3 for the bus lines and 4 for the metro stations. At the view of these results, we tested 3 and 5 clusters for bus and 4 to 6 clusters for metro stations to see which ones would be more informative with respect to the research objectives. Results for 5 and 6 clusters respectively for bus and metro and presented in the rest of the paper.



(a) (b)

Figure 1. Elbow method to determine the optimal number of clusters for (a) bus lines (b) metro stations

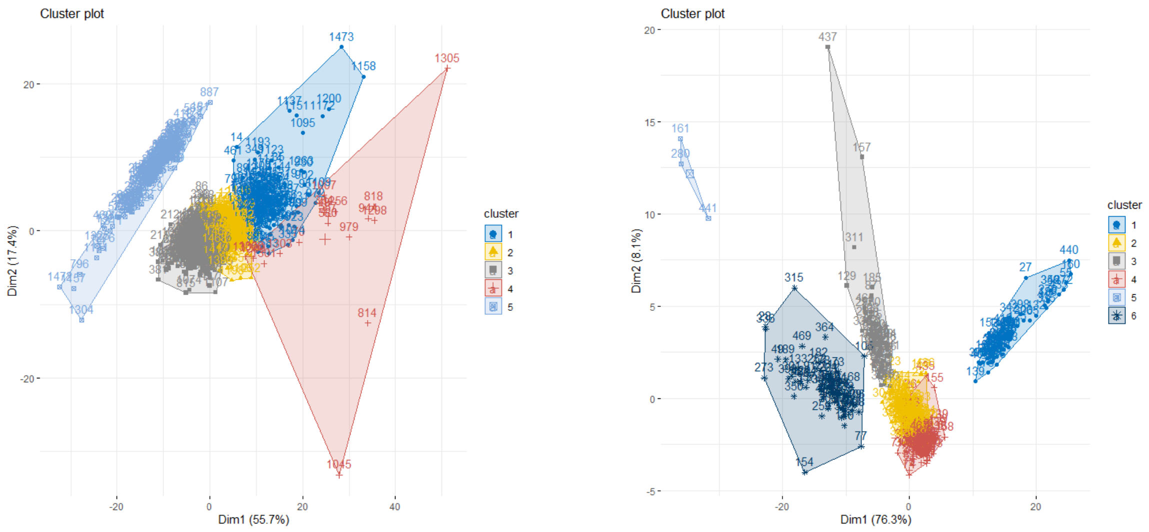


(a) (b)

Figure 2. Silhouette method to determine the optimal number of clusters for (a) bus lines (b) metro stations

4.2. Clustering results

The first clustering method used is the k-means. The results are shown in Figure 3 both for bus lines (5 clusters) (a) and metro stations (6 clusters) (b). The two dimensions on the cluster plot represent the principal components obtained after the reduction of the initial number of dimensions of the dataset (51 + 52 = 103 dimensions). The percentage along the two principal dimensions represent the percentage of the variance of the entire dataset, for example for the bus lines dimension 1 represents 55.7% of the total variance of the dataset and dimension 2 represents 17.4% of the variance, meaning that these two dimensions represent 73.1% of the total variance of the dataset. For the bus lines, the three first clusters are very closely located along the cluster plot, with cluster number 2 being in the middle of the cluster plot. There is a slight overlap of the clusters as represented by convex hulls. For the metro stations, clusters 2, 3 and 4 are very closely located while clusters 1, 5 and 6 are more distant from the other clusters.



(a) (b)

Figure 3. K-means clustering results for (a) bus lines (b) metro stations

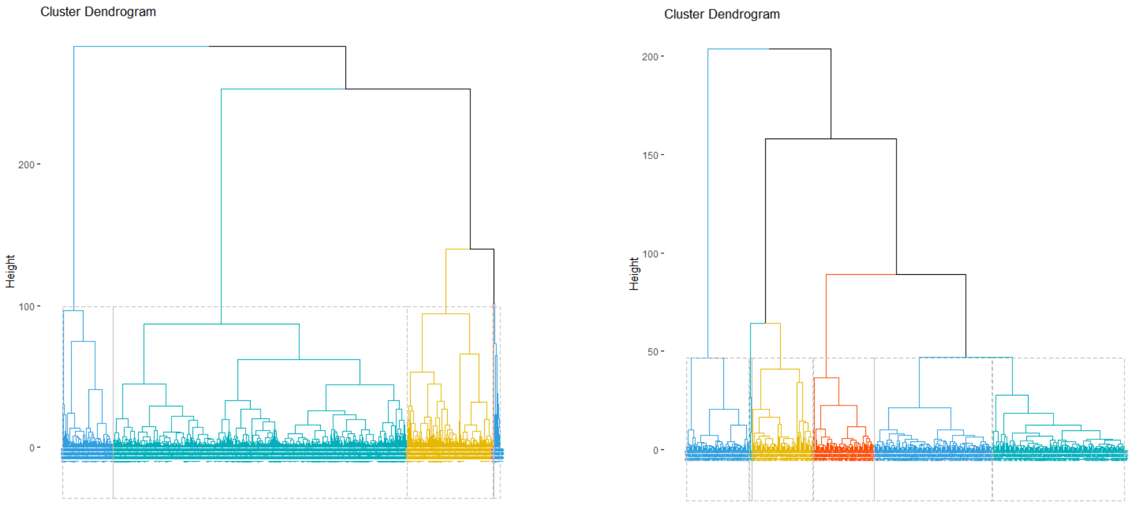
The second clustering method used is the k-medoids, also known as PAM (partitioning around medoids). For the bus lines, the results are very similar to the results obtained with the k-means clustering technique except for clusters 4 and 5 However, for the metro stations, there are differences, clusters 5 and 6 are merged and cluster 6 is divided into two new clusters (3 and 6).

The third method is the hierarchical clustering technique; results are shown in Figure 4 using dendrograms. This tool allows to identify the importance of each cluster in terms of observations (bottom) as well as the coherent numbers of clusters (with horizontal lines). For the bus lines, the height reduction (height represents the distance between clusters or observations) is not so important to get to the creation of 5 groups, while for metro stations the height reduction is greater, meaning that the gain from adding clusters is more important for the metro stations than for the bus lines. Hence, at the view of this tool, 4 clusters may be better for bus validations than 5.

The fourth method is fuzzy clustering technique which is soft method. It produces an ellipse encompassing most of the points belonging to the clusters. 3 clusters are obtained for both bus and metro stations.

The last clustering method is DBSCAN. The number of clusters is not determined by the user; however, the user must provide two parameters: the eps value and the minimal number of points. The eps value is determined using the kNNdistplot function: looking at the inflection point in the graph the user can determine the appropriate eps value. For the bus lines, the eps value is 8 and the minimal number of points is 5. For the metro stations, the eps value is 6 and the minimal number of points is also 5. For the bus lines, only two clusters where created. When comparing these results with the results from the k-means method, it seems that the cluster 5 is the same, but clusters 1, 2 and 3 all seem to be included in cluster 1 of the DBSCAN method, and cluster 1 is not included in the DBCAN clusters. For the metro stations, two clusters were created, and the results are like the k-means method: the DBSCAN cluster 1 is made of the k-means clusters 2, 3, 4, and 6 while the DBCSCAN cluster 2 is the same as the k-means cluster 1, and k-means cluster 5 is not included in the DBSCAN clusters.

At the view of the results obtained with the various clustering methods, we focus on the analysis of the clusters obtained using the k-means methods, with 5 and 6 clusters respectively for bus and metro stations.



(a)

(b)

Figure 4. Hierarchical clustering results for (a) bus lines (b) metro stations

4.3. Summary of the k-means clustering results

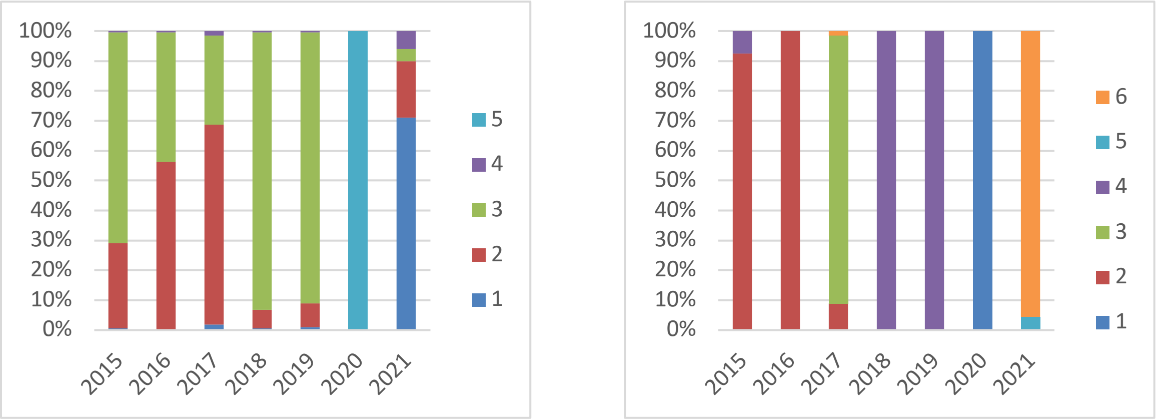
A summary of the clustering results is presented in Table 4 with the min, max and average intensity per cluster, as well as the percentage of observations represented for each cluster. For the bus lines, the most intense cluster is C4 while the least intense cluster is C5, and the most frequent cluster is C3. For the metro stations, the most intense cluster is C5 while the least intense is C1, and the most frequent cluster is C4 followed closely by C2.

Table 4. Summary of the k-means clustering results

	Bus lines				Metro stations			
	min	max	average	% of observations	min	max	average	% of observations
C1	0,00	8,32	1,76	10,7%	0,01	1,38	0,43	14,4%
C2	0,00	5,97	1,24	27,4%	0,02	1,68	1,06	28,8%
C3	0,00	6,20	0,98	48,9%	0,02	3,34	1,47	12,9%
C4	0,05	19,48	3,32	1,4%	0,07	2,33	1,02	29,4%
C5	0,00	3,71	0,39	11,6%	0,13	8,96	3,59	0,6%
C6	-				0,00	5,31	1,85	13,9%

4.4. Yearly occurrence of clusters

It can be interesting to look at the clustering results per year, since each bus line/metro station vector is constructed for every year. For the k-means clustering results, the distribution of the cluster membership per year is shown in Figure 5. For the bus lines, C2 and C3 are found between 2015 and 2019, while 2020 is only composed of C5. C1 only appears in 2021 but this year also has patterns similar to those observed from 2015-2019 which suggests a return to pre-COVID yearly patterns of ridership. For the metro stations, the two first years are dominated by C2, then year 2017 is dominated by C3 and then the two years preceding COVID are fully represented by C4. 2020 is solely composed of C1 while 2021 is dominated by C6 with some C5; there is not clear return to pre-COVID patterns as can be seen for bus lines. C5 is specific to the metro stations closely located to universities (McGill, Édouard-Montpetit and Université-de-Montréal stations).



(a)

(b)

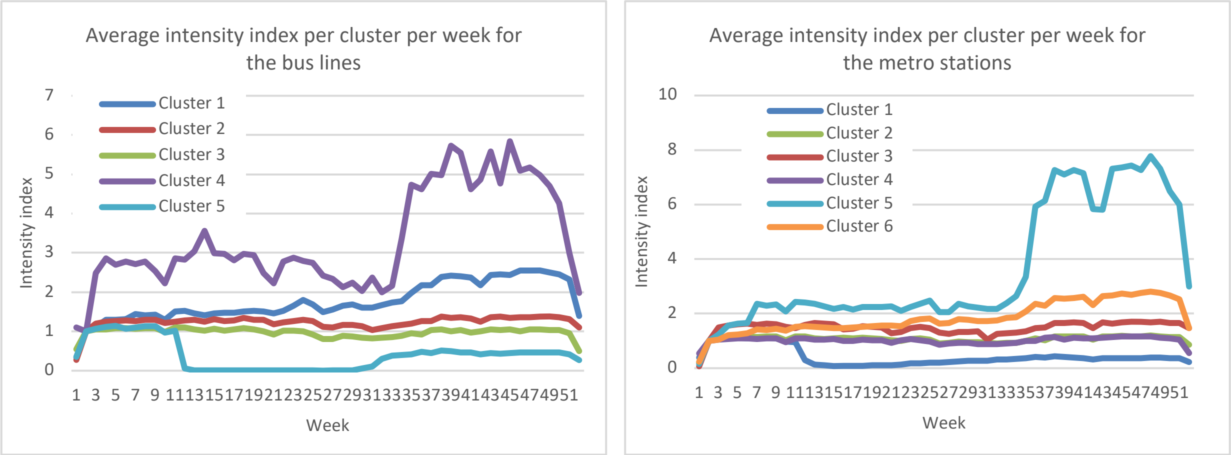
Figure 5. K-means cluster distribution per year for (a) bus lines (b) metro stations

4.5. Cluster Centers

Differences between clusters are examined through cluster centers.

4.5.1. Average intensity

The average intensity index per week per cluster is shown in Figure 6. For the bus lines, shows that C1, C2 and C4 are the most intense, while C5, dominant in 2020, has a very distinct shape : it is similar to C1 and C3 for the first 11 weeks, then, not surprisingly, there is almost no intensity between weeks 12 and 29 (since some important lines connected to subway stations did not require validations at the beginning of the pandemic and mobility was highly affected). Finally, there is an increase of intensity after week 30. For C1 and C4 representing most of 2021, the intensity is very high through the year and much less constant than for the pre-pandemic C2 and C3. For the metro stations, C1 representing 2020 is intense for the first 11 weeks, then rapidly drops under 0.1, and rapidly starts increasing until it reaches an intensity of 0.42 at week 38. C5 and C6 representing all of 2021 are respectively the two most intense clusters through the year.


Average intensity index per cluster per week for the bus lines

Average intensity index per cluster per week for the metro stations

(a)

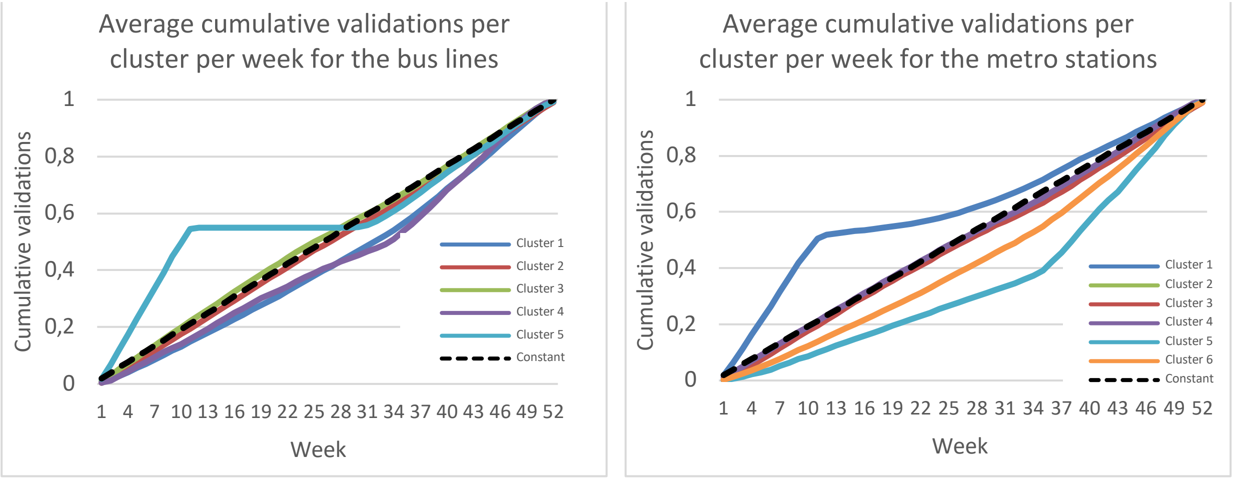
(b)

Figure 6. Average intensity index per k-means cluster per week for (a) bus lines (b) metro stations

4.5.2. Cumulative validations

The average cumulative validations per week per cluster is shown in Figure 7. For the bus lines, clusters 2 and 3 are almost the same, while cluster 1, representing most of 2020, has a very distinct shape : a sharp slope up until week 11, then it is horizontal until week 29, and after week 30 it is similar to C2 and C3. C2 and C3 are much more stable and closer to the constant line (line representing the situation where each week would represent 1/52 % of the validations of the year). For C1 and C4 representing most of 2021, the line is always under C2 and C3 meaning that there were more validations at the end of the year. For the metro stations, C1 representing all of 2020 has the sharpest slope for the first 11 weeks, then there is a break and a sharp decrease of the slope, and it increases until the end of the year. C2, C3 and C4 have much more stable slopes, while C6 (representing most of 2021) and C5 have slopes under the constant line, meaning that there were much more validations at the end of the year for these clusters.

As expected, year 2020 and 2021 present quite different profiles than the previous years for both bus lines and metro stations. Hence, bus lines are more or less segmented in two dominant clusters (C2 and C3) for the first 5 years with shares changing over time, with a high dominance of C2 in the 2 years preceding the pandemics. The main difference between these two clusters is a lower share and intensity of validations during the summer months, suggesting a shift to other modes maybe active ones. For the subway stations, the patterns are a bit different with dominance of C2 in 2015 and 2016, shift to C3 in 2018 and then to C4 for the 2 years before the pandemics (this later cluster has lower share and intensity of validation during summer which also suggest a shift in mode during the summer while the annual ridership is increasing). Bus is showing signs of returns to pre-pandemics patterns for some 25% of the lines (belonging to C2 and C3) but nothing similar is observed for metro stations.

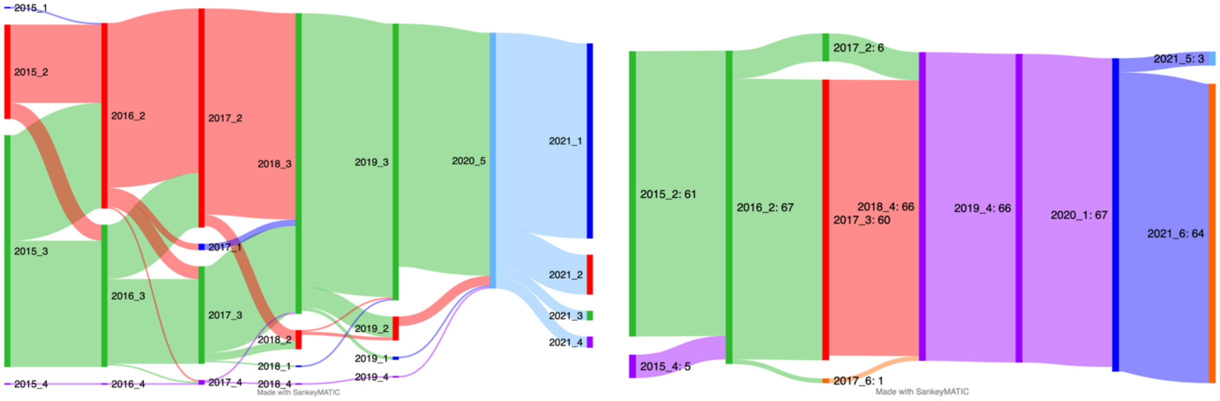


(a) (b)

Figure 7. Average cumulative validations per k-means cluster per week for (a) bus lines (b) metro stations

4.6. Transitions of Objects across the Years

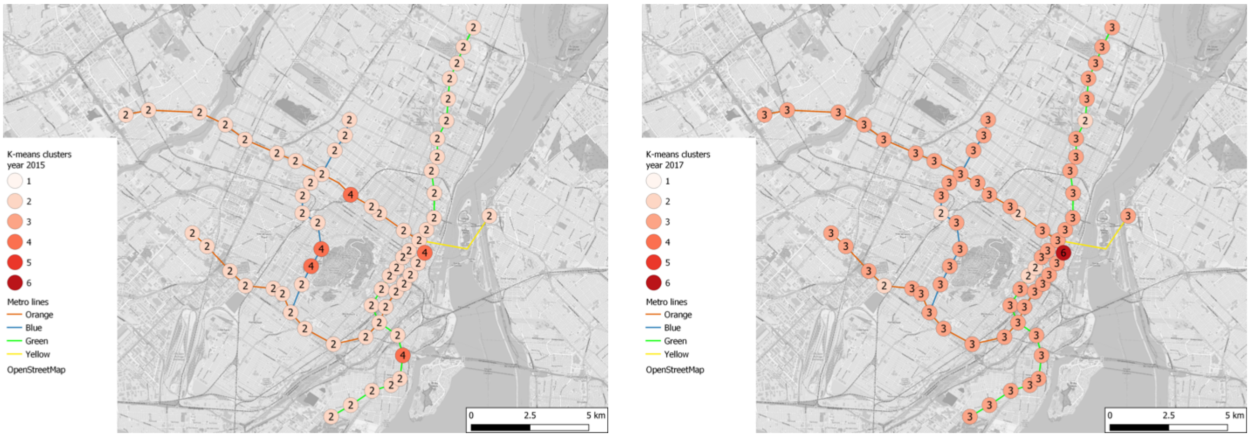
This section examines how the yearly assignment of the various examined objects to clusters (metro stations and bus lines) has changed over time. Sankey diagrams were produced to evaluate the transition of objects across the years and are shown in Figure 8. For the bus lines, there is a lot of evolution between the years, especially in the first three years between clusters 2 and 3, and in the last two years because of the pandemic with the dominance of clusters 5 in 2020 and 1 in 2021. For the metro stations, the years are much more monolithic with each year having a dominant cluster: 2 in 2015 and 2016, 3 in 2017, 4 in 2018 and 2019, 1 in 2020 and 5 in 2021.



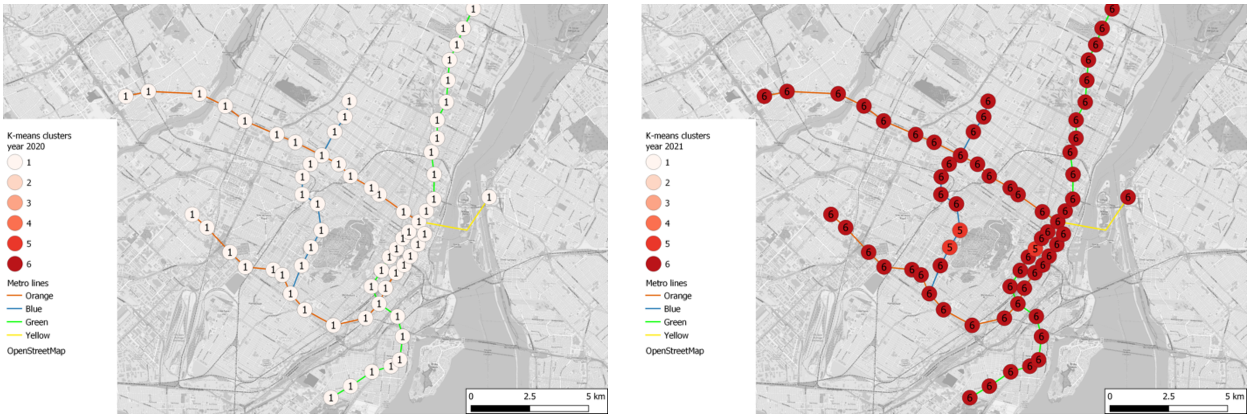
(a) (b)

Figure 8. Sankey diagrams of the evolution of k-means clusters for (a) bus lines (b) metro stations

The spatial evolution of the clusters is also evaluated and is shown in Figure 9. The first year, 2015 (a), most stations were allocated to C2, except a few stations allocated to C4. The C4 stations were stations close to the University of Montréal on the blue line and a few other stations on the orange and green lines. In 2017 (b), most stations were allocated to C3, however 6 stations were allocated to C2 (two of them downtown on the green line), and 1 station was allocated to C6 (downtown on the orange line). In 2020 (c), all the stations belonged to C1. Finally, in 2021, most of the stations were allocated to C6 including two stations close to the University of Montréal on the blue line and one station close to McGill university on the green line. The spatial distribution of the various clusters will need further investigations as well as the interaction between bus lines and metro stations' patterns.



(a) (b)



(c)

(d)

Figure 9. Maps of the k-means clusters for the metro stations in (a) 2015 (b) 2017 (c) 2020 (d) 2021

5. Conclusion

With the aim to monitor the evolution of transit usage across 7 years and namely the contribution of weekly validation to yearly transit ridership, this paper has provided an illustration of clustering methods applied to cumulative vectors including weekly shares of yearly ridership and weekly intensity indicators. The 7 years of data cover years 2020 and 2021 when the COVID-19 pandemic had important influence on transit ridership as estimated using smart card validations. This paper was done as part of a research program on the innovative processing of administrative time-series data. Only results from the k-means approach were presented in this paper while various methods were tested using the same vectors of data. Previous research in the same program also compared cumulative and non-cumulative vectors and showed that the cumulative ones were more fitted to the use of simple distance matrix. Such an approach borrows from the logic behind KS test with differences estimated between cumulative distributions.

Results confirm the ability of the proposed approach to understand the global changes in ridership patterns. Further work is needed to provide explanatory components to the analysis. Currently, only total weekly ridership (both in terms of yearly share and relative intensity) is used but some lines and metro stations support higher shares of commuting trips while others have more diverse usages (shopping and leisure destinations). The data used in this paper could easily be used for processing weekdays and weekend days separately or looking into peaks and non-peaks periods may better assist in understanding the transformation with respect to usage patterns. It may also be relevant to examine the resulting clusters with respect to some features of the metro stations (features of the neighborhoods for instance) or bus lines (operational features such as length, number of bus stops, frequency, type of line, etc.). Hence, there is currently no link with transit supply. In this perspective, comparison with cumulative service level, for lines and metro stations, will provide insights into the adequacy between planned services and observed demand as reflected by validations. The methodology developed in this paper can be used by transit agencies to identify long term trends in ridership data as well as the system-wide detection of breaking points. It may also assist in better adjusting the level of supply with the trends with respect to weekly ridership.

Acknowledgements

The authors wish to acknowledge the Société de Transport de Montréal (STM) for providing data and supporting the research program of the Mobility Chair.

References

Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.

Deschaintres, E., Morency, C., Trépanier, M. (2019). Analyzing transit user behavior with 51 weeks of smart card data, Transportation Research Record, Volume 2673, Issue 6, pp. 33-45.

Egu, O., Bonnel, P. (2020). Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon, Travel Behaviour and Society 19 (2020) 112–123

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In, 226–31. AAAI Press.

Fraley, Chris, and Adrian E Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation." Journal of the American Statistical Association 97 (458): 611–31.

Fraley, Chris, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca. 2012. "Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation." Technical Report No. 597, Department of Statistics, University of Washington. https://www.stat.washington.edu/research/reports/2012/tr597.pdf.

Ghaemi M. S., Agard B., Trépanier M., Partovi Nia V. A Visual Segmentation Method for Temporal Smart Card Data. Transportmetrica A: Transport Science, Vol. 13, No. 5, 2017, pp. 381–404.

Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, Z., Zhao, J. (2018). Measuring Regularity of Individual Travel Patterns, IEEE Transactions on Intelligent Transportation Systems, Vol.19, No. 5, May 2018.

Hartigan, J.A. (1975). Clustering Algorithms. New York: Wiley.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis, 344, 68-125.

Ma X., Wu Y. J., Wang Y., Chen F., Liu J. Mining Smart Card Data for Transit Riders' Travel Patterns. Transportation Research Part C: Emerging Technologies, Vol. 36, 2013, pp. 1–12.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Manley, E., Zhong, C., Batty, M. (2018). Spatiotemporal variation in travel regularity through transit user profiling, Transportation (2018) 45:703–732.

Ortega-Tong M. A. Classification of London's Public Transport Users Using Smart Card Data. Massachusetts Institute of Technology, 2013.

Pelletier, M.-P., Trépanier, M., Morency, C. (2011). Smart card data use in public transit: A literature review, Transportation Research Part C: Emerging Technologies, Vol. 19, no. 4, pp. 557-568.

Tao S., Rohde D., Corcoran J. Examining the Spatial–Temporal Dynamics of Bus Passenger Travel Behaviour Using Smart Card Data and the Flow-Comap. Journal of Transport Geography, Vol. 41, 2014, pp. 21–36.

Viallard, A., Trépanier, M., Morency, C. (2019). Assessing the evolution of transit user behavior from smart card data, Transportation Research Record, Volume 2673, Issue 4, pp. 184-194.