Space-time classification of public transit smart card users' activity locations from smart card data

Li He · Martin Trépanier · Bruno Agard

Abstract Smart card data from public transit system has proven to be useful to understand public transit users' behaviours. Many relevant research has been done concerning: (1) the utilization of smart card data (2) data mining techniques (3) the utilization of data mining in smart card data. In these researches, the classification of users' behaviours is based on trips where temporal and spatial classifications are seen as separated process. Therefore, it is interesting to develop a method, based on the users' daily behaviours, that take into account spatial and temporal behaviours at the same time. To do so, in this article, a methodology is developed to classify smart card users' behaviours based on a dynamic time warping, hierarchical clustering and sampling method. A 3-dimension space-time prism plot shows the efficiency of the algorithm.

Keywords: Public transit · Smart card data · Dynamic time warping · Spatiotemporal classification · Activity locations

Li He, M.Sc.A., Ph.D. student
Polytechnique Montréal and CIRRELT
Département de Mathématique et de Génie Industriel
2500 ch. de Polytechnique Montréal (Québec), Canada, H3T 1J4
Tel.: (514) 340-4711 ext. 4914, Fax: (514) 340-4173
Email: li.he@polymtl.ca li.he@polymtl.ca

Martin Trépanier, Ph.D., Full Professor
Polytechnique Montréal and CIRRELT
Département de Mathématique et de Génie Industriel
2500 ch. de Polytechnique Montréal (Québec), Canada, H3T 1J4
Tel.: (514) 340-4711 ext. 4911, Fax: (514) 340-4173
Email: martin.trepanier@polymtl.ca martin.trepanier@polymtl.ca

Bruno Agard, Ph.D., Full Professor
Polytechnique Montréal and CIRRELT
Département de Mathématique et de Génie Industriel
2500 ch. de Polytechnique Montréal (Québec), Canada, H3T 1J4
Tel.: (514) 340-4711 ext. 4914, Fax: (514) 340-4173
Email: bruno.agard@polymtl.ca bruno.agard@polymtl.ca

1 Introduction

Data from smart card fare collection systems is very useful to public transit planners [Pelletier and al., 2011]. The smart card data from public transit system helps to better know the smart card user's behaviours. This knowledge is helpful to improve the service of public transit authority. Many efforts have been made by using data mining to classify users' transactions. In particular, some methodologies were proposed to classify smart card users' temporal and spatial behaviours by using diverse distance metrics and classification method. In this paper, we present a method to classify public transit users accordingly to the time and the location of their trips during the day.

This article will be organized as follows. In the next part, the literature review will focus on relevant work, mainly the data mining methods that will be used. Then, the pragmatic and the objective of this paper will be introduced. To solve the problem of classifying spatiotemporal behaviours, a methodology is developed in the part 4. After, the "Implementation" section will introduce the case studied and something important during the test of algorithms. Then, the result and their analyses will be in the part 6. The end of the article is a conclusion which contains the contribution, limitation and perspective.

2 Literature review

2.1 Utilization of smart card data

Over the years, several works have been done with smart card data in public transit. Relevant articles introduce the description of smart card data [Trépanier et al., 2004], enriching of the data (including estimation of destinations [Trépanier and al., 2007; He et al., 2015]), and prediction in using the data [Ceapa et al., 2012]. Smart card data can be used to analysis users' behaviours, for example: users' characterization [Morency et al., 2007], networks' characterization [Tranchant, 2005], analysis of extern factors that influent the utilization of the network [Briand et al., 2017]. The amount of smart card transactions can reach multiple millions for a typical city, it is therefore relevant to use data mining techniques to be able to analyze data in a meaningful way.

2.2 Data mining techniques

Many data mining techniques can be used to process data. Two elements must be foreseen. On one side, there is a choice of methods amongst partition algorithms [Chevalier et al., 2013], hierarchical algorithms [Rokach et al., 2005], and algorithms based on density (Kriegel et al., 2011). On the other side, several metrics can be used to evaluate the dissimilarity of two vectors, including Euclidean distance [Deza et al.,

2009], Manhattan distance [Black, 2006], cross correlation distance [Mori et al., 2016], and dynamic time warping distance [Giorgino, 2009].

Figure 1 illustrates dynamic time warping method. Dynamic time warping is a popular technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches and the warping which optimally deforms one of the two input series onto the other [Giorgino, 2009]. We can formally define the dynamic time warping problem minimization over potential warping paths based on the cumulative distance for each path, where d is a distance measure between two time-series elements. Warping the last moment of time series B to the last moment of time series A, in order that the cumulative distance between A and B is minimum (Fig. 1(a)).

To obtain a minimum cumulative distance, the time series can be wrapped to the next time point (moment). For example, grid point (M -1, N -1) can be wrapped to (M, N -1), (M - 1, N), (M, N) to compute each distance (Fig. 1(b)). Calculate all the possible paths from grid points (1, 1) to (6, 6), find the path with minimum cumulative distance. In this grid above, the distance of DTW is 7 (Fig. 1(c)).

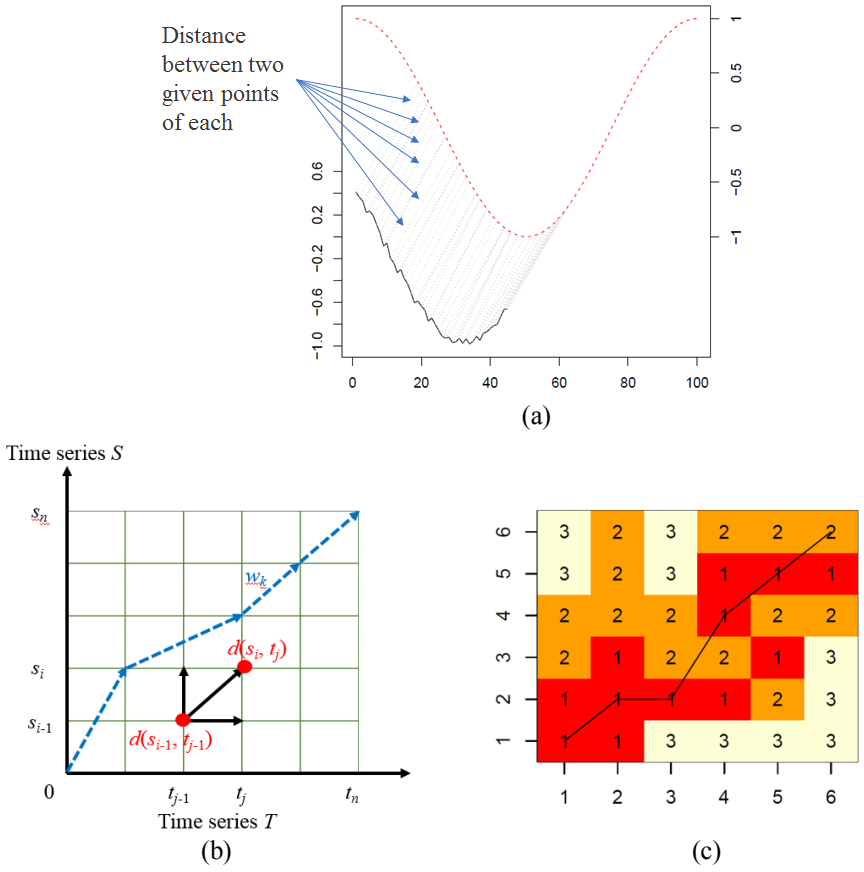


Fig. 1 Dynamic time warping method

2.3 Utilization of data mining in smart card data

A more classical data mining technique (k-means) has been used to classify users' general behaviour over a period of 12 weeks [Agard et al., 2006]. In another work, the spatial and temporal data mining methods have been applied separately [Ghaemi et al., 2015]. Furthermore, other works based on k-means [Morency et al., 2006], neural networks [Ma et al., 2013] and DBSCAN [Kieu et al., 2014] were bound to identify regular passengers, or propose a clustering accordingly to their behaviours.
At the end, to verify the efficiency of space-time clustering algorithms, it is interesting to show the profile of each cluster in a space-time prism 3D plot, as proposed by Farber et al. (2015) (see Fig. 2).

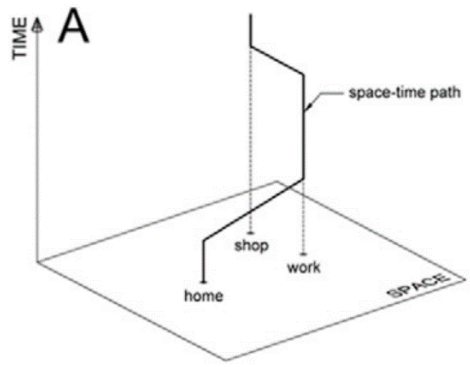


Fig.2 Space-time prism example (Farber et al., 2015)

3 Problematic and objective

3.1 Problematic

Due to its nature, a public transit path is characterized both by the time of the day where the boarding activities occurred, and the location where they occurred. The most intuitive way of clustering users would be to consider space and time at the same time. In this article, users' behaviours will be treated as a time series of spatial locations. The classification technique will therefore take into account space and time at the same time, using a specific dissimilarity metric.
In our previous works, cross correlation distance and dynamic time warping distance have been integrated with hierarchical clustering to create time series segmentation methods [He and al., 2017]. But now, we propose to integrate the spatial dimension.

3.2 Objective

The aim of this paper is to propose a methodology to classify users' spatiotemporal behaviours using pertinent classification algorithms and distance metrics. The

behaviour is composed of the sequence of bus stop locations at each hour. This includes two phases.

To demonstrate the method, Fig. 3 presents an example of 3 users' daily behaviours:

- the first user leaves home at 06:30 to go to school and return home at 16:00;
- the second user leaves home at 07:30 to go to work and return home at 18:00.
- the third one leaves on also at 06:30 to go to work, but before going home at 18:00, the user went to the supermarket at 16:00.

The objective of spatiotemporal classification is to group these daily profiles in terms of time and location, in order to separate them into some clusters. In this case, if we measure that the behaviours of users 1 and 2 are "more similar" than user 3, then a group will be created with users 1 and 2, and user 3 will be in another group.

In the spatiotemporal classification, when measuring the dissimilarity of two users' profile, we consider not only the time of transaction by smart cards, but also the real distance between bus stops, serving as proxies for user location during the day (the Euclidean distance between school of user 1 and work of user 2, for example). The objective is to have a measure of dissimilarity that takes the two dimensions (space and time) into account, in order to proceed to clustering.

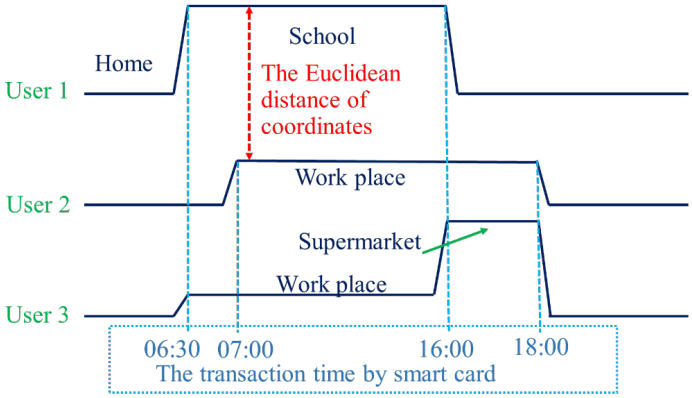


Fig.3 Short example showing 3 user behaviours to classify

4 Methodology

Fig. 4 shows the methodology developed to put the proposed dissimilarity metric and clustering methods in action. The figure shows the number of records for data that were used in the case study, described in the next section. The methodology contains seven steps that are described hereafter.

4.1 Data preparation

First, the smart card transactions are formatted and preprocessed. The trips that occurred after midnight are adjusted so that the trip remains in the same user journey,

using a 24+ hour system (step 1 in the Fig. 4). For example, a trip that occurred at 1:00 AM the next day is changed to 25:00, same day.

Secondly, for trajectory classification, we have to use the destinations of the smart card transactions. Smart card data used for this paper does have tap-out, so the destinations were estimated using the method proposed in (He and Trépanier 2015). Therefore, the transactions that do not contain destinations (destinations not estimated) are removed (step 2 in Fig. 4).

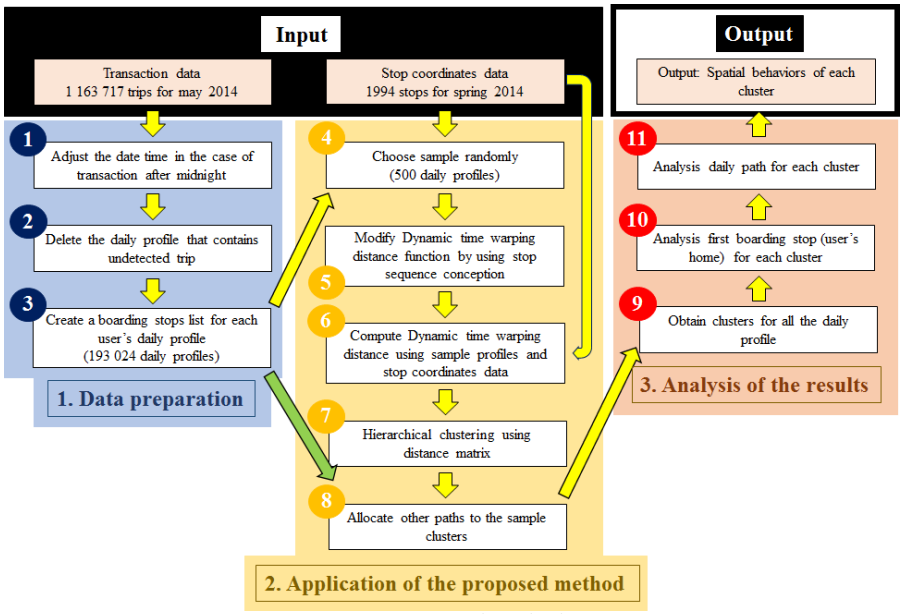


Fig.4 Proposed method

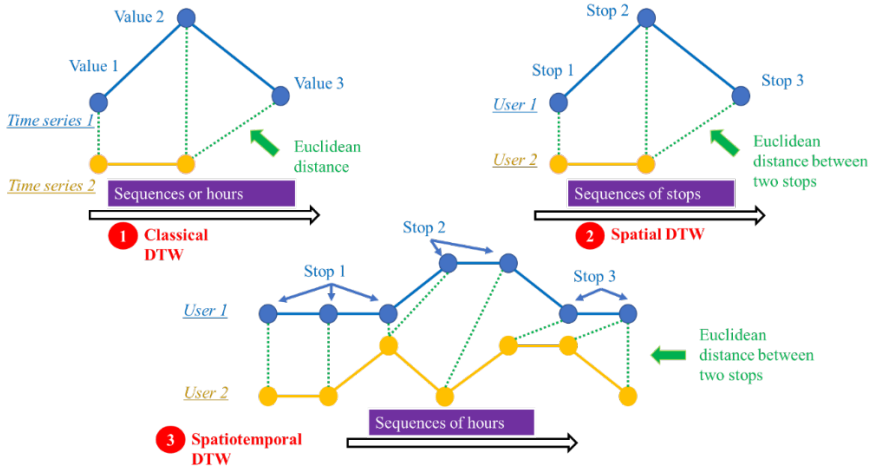


Fig.5 Comparison of the three DTW methods

Table 1 Conception of three types of DTW

Conception	Classical DTW	Spatial DTW	Spatiotemporal DTW
Object to be treated	Time series	User path in daily profile (Stop sequences)	User path-hour in daily profile (Stop sequences at every given moment)
Point	Time point (moment)	Stop	Stop at every given moment
Sequence (time series)	Time sequence	Stop sequence (Uneven relating to time)	Stop sequence (Uneven relating to time)
Distance between grid point	Can be defined as Euclidean distance, Manhattan distance, etc.	Distance between two given stops (Only Euclidean distance)	Distance between two given stops (Only Euclidean distance)
Euclidean distance	In sense of time (X: time; Y: value in x)	In the sense of geography (X: longitude; Y: latitude)	In the sense of geography (X: longitude; Y: latitude)

In the final step of data preparation, for each card and for each day, a list of bus stops is created, showing the hourly sequence of stops where the user is located during the day (step 3 in Fig. 4). Fig. 5 presents three methods to build the time series in this case. The main idea is to link all the stop at a sequence of given moment (…Stop 1 at 11:00, stop 2 at 12:00, stop 3 at 13:00 …until the end of the day). The bullets 2 and 3 of the figure presents the method chosen in this paper to build time series for spatial and spatiotemporal classification. Table 1 presents the characteristics of each approach. In this paper, we use the last two approaches.

4.2 Application of the proposed method

The clustering of more than a hundred thousand user daily profiles is a time-consuming process. Calculation time (when feasible) is way too long, and the amount of computer memory needed will explode because of the size of the dissimilarity matrix. To do the clustering, we propose to use a sampling approach. Our tests showed that a sample of 500 daily profile (over 100,000) is sufficient here. This section explains the steps 4 to 8 of the methodology.

Fig. 6 shows the overall sampling process (He et al., 2018). At first, all observations are provided in Fig. 6(a). The red points in Fig. 6(b) are the randomly selected points. Then, we apply dynamic time warping and hierarchical clustering algorithms to these sample points. The Fig. 6(c) presents the clusters created in this example. We used the dendogram showing the distance between observations to cut a number of groups suitable to the analysis needs.

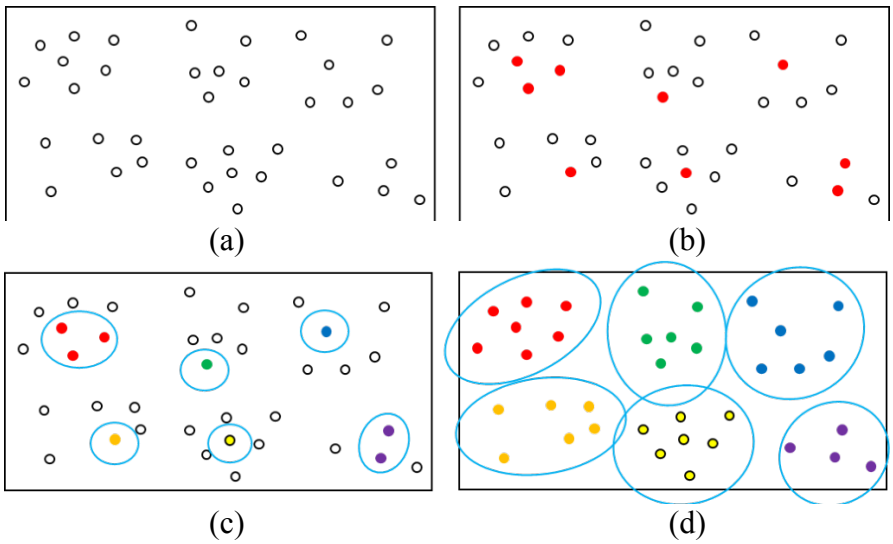


Fig.6 Allocation method

We then calculate the distance between any other point and all the points of a sample group, and then allocate them to the nearest group. Finally, obtain the groups for all the points (time series), as shown in Fig. 6(d).

4.3 Result analysis

Based on the results obtained, we analyze smart card users' behaviours looking at boarding stops, daily profile and space-time path for each cluster (steps 9-10-11 in the Figure 4).

5 Implementation

The dataset is provided by the *Société de Transport de l'Outatouai*s (STO), a transit authority serving the 280,000 inhabitants of Gatineau, Quebec. The STO authority is a Canadian leader in using public transit smart cards fare collection. This system has been in use since 2001, and a substantial proportion (over 80%) of STO users have a smart card [Pelletier et al., 2011].
In this study, all the weekday transaction data of May 2014 have been used to test the proposed method of spatial classification. This dataset contains 1 163 717 trips.
The method is programmed in python, allowing to deal with such a large database.
During the implementation, the number of clusters should have been determined by cutting dendrogram branches. Fig. 7 shows the dendrogram of spatial classification algorithm. We cut it into 10 clusters, because:

- We tried to get even size clusters as possible, even though this is not mandatory (user behaviours may not be balanced evenly). We can compare more different behaviour if we get more clusters.
- In this case, if we increase the number of clusters from 10 to 11, there will be a cluster whose size is too small. Then after the allocation process, this cluster size will be negligible compared to other clusters. For the analysis, we prefer to keep larger cluster size.

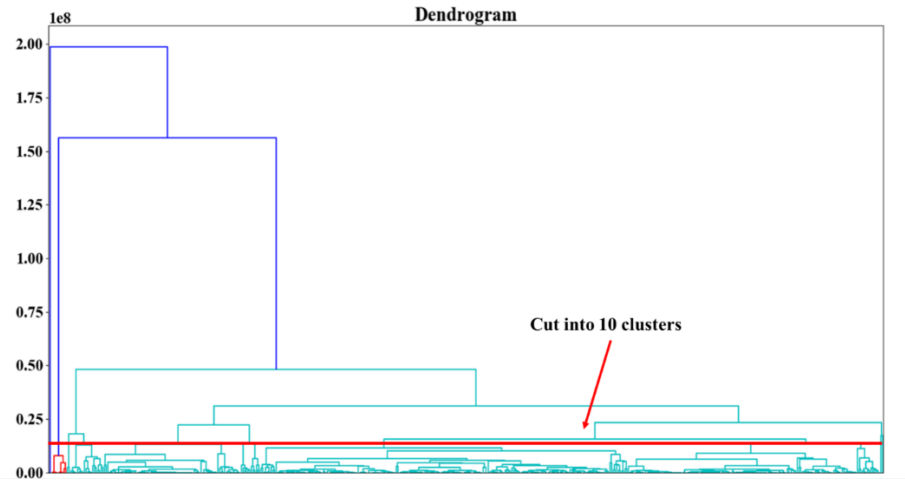


Fig.7 Dendrogram of hierarchical clustering of spatial classification algorithm

6 Results and analysis

6.1 Results

An excerpt of the results of spatial classification is presented in the Table 2. For each combination of "smart card number + date" (card-day), a stop list is generated, and a cluster is obtained.
We can find that for the combination "1292322417029248_2014-05-20", many trips in this daily profile. One of the advantages of dynamic time warping is that it can deal with different numbers of trips during the day. We can also find that for the user "1000309", the user's spatial behaviours haven't been changed even though there is a minor difference in boarding stop.

6.2 Analysis by boarding stop

Fig. 8 shows the analysis by first boarding stop for the spatial classification. Every colour represents a cluster and dots represent the first boarding stop only. In general, the clusters are grouped by the location (coordinates), however, there are some places where the case is more complicated. For example, in the "Aylmer" area, the orange

and green colour are mixed, because the destinations of these two clusters are different, even though the origins are similar. In this case, the destinations of green clusters are located in Ottawa, but those of orange cluster are located in Hull or Gatineau. This is an advantage of the proposed method compared to the classical ones.

Table 2 Spatial classification result

Daily_profile	Stop_list	Cluster
1185321492030080_2014-05-01	['2060', '5034']	7
1188606196918144_2014-05-05	['1425', '5030']	5
1162476560982656_2014-05-13	['8071', '2618', '8030']	8
1144962089103488_2014-05-22	['2822', '1377']	6
1256806531407488_2014-05-30	['2390', '2427', '2108']	7
1243736397129600_2014-05-23	['4631', '5030', '3307']	2
1159327275886208_2014-05-27	['4442', '8101', '2724', '3991']	4
1173514901724800_2014-05-12	['3991', '4772']	1
1214358820824960_2014-05-26	['8101', '2318']	10
1292322417029248_2014-05-20	['8101', '3501', '3496', '9735', '5022', '3991']	8
1000309_2014-05-02	['5022', '2604']	7
1000309_2014-05-06	['5022', '2604']	7
1000309_2014-05-15	['5022', '2604']	7
1000309_2014-05-16	['5022', '2604']	7
1000309_2014-05-28	['5022', '2625']	7

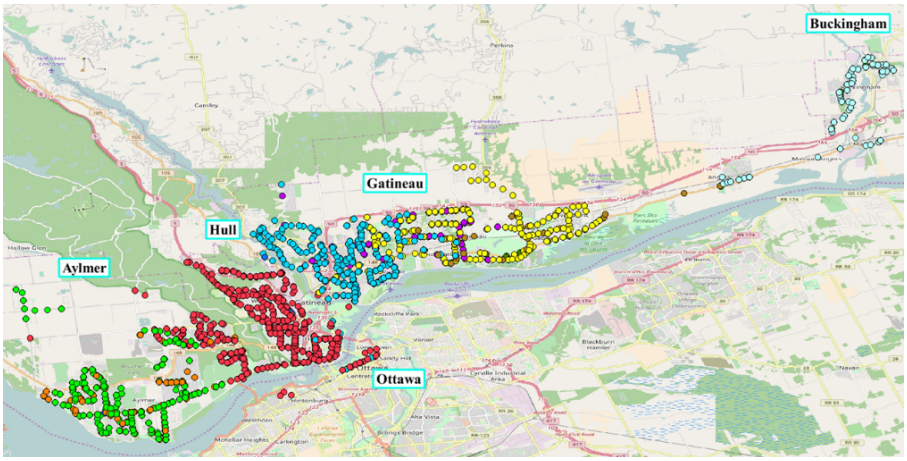


Fig.8 Analysis by first boarding stop

6.3 Analysis by daily trajectory

Figure 9 shows the daily trajectory of each cluster obtained through spatial classification. By watching the colours, we can overview the characteristics of each cluster. For example, the users of cluster cyan are living in Buckingham, and they go to work in Ottawa. Maybe they go there directly, or maybe they have a transfer in Gatineau. If we want to distinguish between these two behaviours (transfer or not),

we can cut the dendrogram into more clusters. This is an advantage of the proposed method, compared to the classical ones.
This separation of the two behaviours helps to characterize the demand. Based on this result, we may suggest to the public transit authority to implement new lines or enhance bus service, so that the people can travel directly and easily from Buckingham to Ottawa.

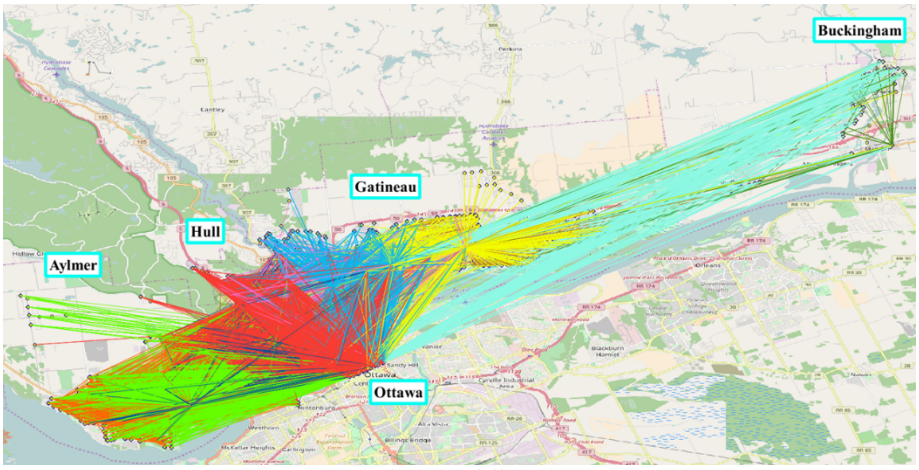
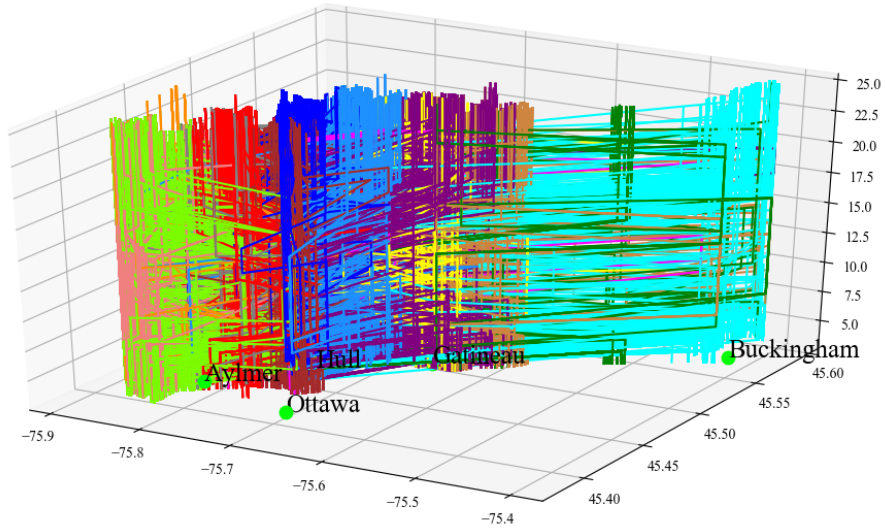


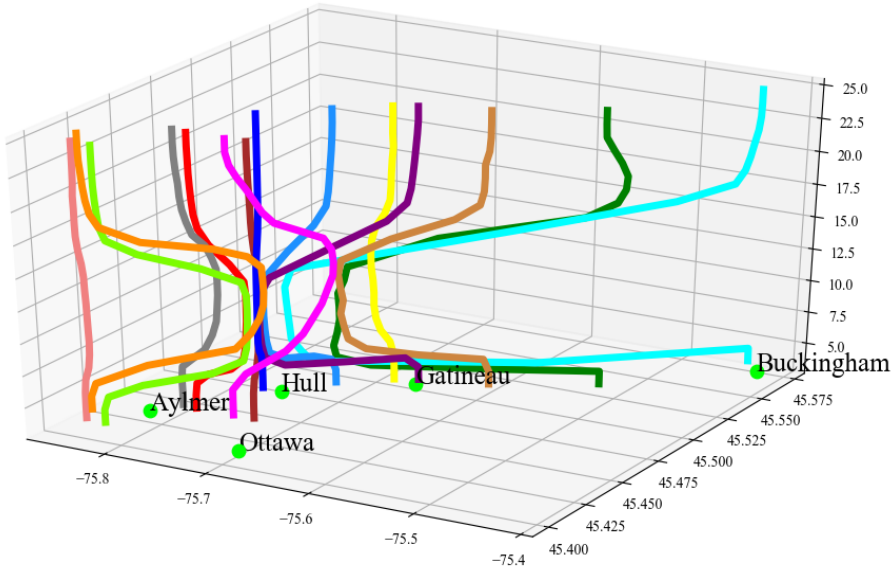
Fig.9 Analysis by daily trajectory

6.4 Analysis by space-time path

Based on the spatiotemporal classification result, a 3D space-time path prism of each cluster is plotted. Fig. 10 (a) shows all profiles individually, and in Fig. 10 (b) shows the average path for each cluster. The Z axis of each figure is the hour within the day (25^{th} hour is for 1AM transaction).
In the Figure 10 (b), even though users of the green cluster live closer to their work location than light blue cluster (both from east to downtown), green cluster leaves home early and return home later than light blue cluster. This may be due to an express bus line that links the origin and destination of cluster light blue. Therefore, it is possible to suggest public transit authority to implement an express bus line to serve the users in green cluster so that they can save time when commuting.
We can also find that the behaviour of cluster light green is stable during work hours (during 9:30 – 15:00, the location of light green cluster does not change a lot). That means these users' travel locally. It is possible to suggest public transit authority to implement a special bus line for these users. This new bus line will link the origin and destination of cluster light green, and will be operated only in peak hours, but it can well respond to the demand of this cluster.



(a)



(b)

Fig.10 Space-time prism of each cluster

7 Conclusion

7.1 Contribution

In this paper, a new methodology based on dynamic time warping, hierarchical clustering and sampling method is proposed to classify public transit smart card users'

spatiotemporal behaviours. The result shows that the behaviours can be well distinguished. Based on the result, it is possible to suggest enhancements to the public transit authority to better server customers of specific clusters.

7.2 Limitation

Firstly, the dynamic time warping algorithm is quadratic, therefore, the computation time is long. Secondly, the criterion of separation is distance, therefore, different behaviours may remain in the same cluster because the dissimilarity between them is smaller than their travel distance. Other limitations come from data: the estimation of destinations may not be perfect (it is not validated), this may hamper the results of the clustering method.

7.3 Perspective

In the future, some works are proposed to be done to improve this new method. Firstly, at this time, we judge the quality of the classification by watching the daily trajectory and the space-time path plot. A quantitative method is needed to measure the dissimilarity between each cluster, to prove that the proposed method works mathematically. Secondly, some work should be done to improve computation time, to limit the bound of dynamic time warping method. Besides, more suggestions can be proposed to public transit authority, to better respond the users' demand of specific cluster.

Acknowledgements: The authors wish to acknowledge the support of the *Société de transport de l'Outaouais (STO*) for providing data, the Thales group and the National Science and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

References

Black, P. E. (2006). Manhattan distance. Dictionary of Algorithms and Data Structures, 18, 2012.

Briand, A. S., Côme, E., Trépanier, M., & Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transportation Research Part C: Emerging Technologies, 79, 274-289.

Ceapa, I., Smith, C., & Capra, L. (2012, August). Avoiding the crowds: understanding tube station congestion patterns from trip data. In Proceedings of the ACM SIGKDD international workshop on urban computing (pp. 134-141). ACM.

Chevalier, F., et al. (2013). La classification. Université de Rennes, 1.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances. In Encyclopedia of Distances (pp. 1-583). Springer Berlin Heidelberg.

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behaviour: A social interaction based model of exposure. Journal of transport geography, 49, 26-38.

Ghaemi, M. S., Agard, B., Nia, V. P., & Trépanier, M. (2015). Challenges in Spatial-Temporal Data Analysis Targeting Public TransportÕ. IFAC-PapersOnLine, 48(3), 442-447.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. Journal of statistical Software, 31(7), 1-24.

He, L., & Trépanier, M. (2015). Estimating the Destination of Unlinked Trips in Transit Smart Card Fare Data. Transportation Research Record: Journal of the Transportation Research Board, (2535), 97-104.

He, L., Agard, B., & Trépanier, M. (2017). Comparing Time Series Segmentation Methods for the Analysis of Transportation Patterns with Smart Card Data (No. CIRRELT-2017-28).

He, L., Agard, B., & Trépanier, M. (2018). Measuring the Impact of the Implementation of a BRT on Individual Behavior of Transit User Based on Smart Card Data

Kieu, L. M., Bhaskar, A., & Chung, E. (2014). Transit passenger segmentation using travel regularity mined from Smart Card transactions data.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 231-240.

Ma, X., Wu, Y. J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies, 36, 1-12.

Morency, C., Trépanier, M., & Agard, B. (2006, September). Analysing the variability of transit user's behaviour with smart card data. In Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE (pp. 44-49). IEEE.

Morency, C., Trépanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. Transport Policy, 14(3), 193-203.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance Measures for Time Series in R: The TSdist Package. R JOURNAL, 8(2), 451-459.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, 19(4), 557-568.

Rokach, L., & Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook (pp. 321-352). Springer US.

Tranchant, N. (2005). Analyse des déplacements d'usagers à partir de données de cartes à puce.

Trépanier, et al. (2004). Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain. Congrès de l'Association des transports du Canada.

Trépanier, et al. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems, 11(1), 1-14.