Public Transport (2021) 13:579–595
https://doi.org/10.1007/s12469-021-00274-0

ORIGINAL RESEARCH

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

**Li He[1] · Martin Trépanier[2] · Bruno Agard[2]**

Accepted: 24 May 2021 / Published online: 21 June 2021


## Abstract

Smart card data from public transit systems has proven to be useful to understand the behaviors of public transit users. Relevant research has been done concerning: (1) the utilization of smart card data, (2) data mining techniques and (3) the utilization of data mining in smart card data. In prior research, the classification of user behavior has been based on trips when temporal and spatial classifications are considered to be separate processes. Therefore, it is of interest to develop a method based on users' daily behaviors that takes into account both spatial and temporal behaviors at the same time. In this work, a methodology is developed to classify smart card users' behaviors based on dynamic time warping (DTW), hierarchical clustering and a sampling method. A three-dimensional space–time prism plot demonstrates the efficiency of the algorithm.

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

✉ Li He
he.li@polymtl.ca

Martin Trépanier
martin.trepanier@polymtl.ca

Bruno Agard
bruno.agard@polymtl.ca

1 Autorité Régionale de Transport Métropolitain, 700 Rue De La Gauchetière O, Montréal, QC H3B 5M2, Canada

2 Département de Mathématique et de Génie Industriel, Polytechnique Montréal and CIRRELT, 2500 Ch. de Polytechnique, Montréal, QC H3T 1J4, Canada

# 1 Introduction

Data from smart card fare collection systems is very useful for public transit planners (Pelletier et al. 2011). Smart card data from a public transit system assists in understanding smart card users' behaviors. This knowledge is helpful in improving the services provided by a public transit authority. Many efforts have been made using data mining to classify users' transactions. In particular, some methodologies were proposed to classify smart card users' temporal and spatial behaviors by using diverse distance metrics and a classification method. In this paper, we present a method to classify public transit users according to the time and the location of their trips during the day.

This article will be organized as follows. In the next section, a literature review will focus on relevant work; mainly, the data mining methods that will be used. Then, the pragmatics and the objectives of this paper will be introduced. To solve the problem of classifying spatiotemporal behaviors, a methodology is developed in part 4. The next section will introduce the cases studied and important takeaways when testing the algorithms. Then, the results and their analyses will be in part 6. The end of the article presents a conclusion that contains the contributions, limitations and perspectives of this work.

# 2 Literature review

The literature review consists of four parts. Firstly, the use of public transit smart card data is introduced. Secondly, data mining techniques are presented; in particular, the description of the dynamic time warping distance is shown in detail because it is an important part of the methodology. Thirdly, starting from Sects. 2.1 and 2.2, we show the use of data mining in smart card data. Finally, some limitations are found in the literature, to show the need for the proposed method.

## 2.1 Utilization of smart card data

Over the years, work has been done with smart card data in the public transit sector. Data from a smart card system enables a better understanding of users' behaviors and offers users a better service through the development of a public transit optimization method (Pelletier et al. 2011).

In terms of data preparation and completion, relevant articles introduce the description of smart card data (Trépanier et al. 2004), data management (Covic and Voß 2019), and enriching the data, including a destination estimation method (Trépanier et al. 2007), an unlinked trip destination estimation using kernel density estimation (He and Trépanier 2015), a method to improve the accuracy of the destination estimation method (He et al. 2015), and so forth. Furthermore, some methods based on transfer detection (Chu and Chapleau 2008) and trip purpose inferences (Lee and Hickman 2014) have been developed. These research works construct the base of public transit user behavior analysis.

In terms of smart card user behavior detection, smart card data can be used to analyze user behavior; for example, characterizing users from temporal information (such as transaction time, travel duration, delay, etc.), (Morency et al. 2007; Bunker 2018), by spatial information (origin–destination, trajectory, etc.) (Shi and Hangfei 2014), by mode choice (Kurauchi et al. 2014; Viggiano et al. 2017) and also by the personalities of passengers (such as user loyalty) (Imaz et al. 2015), network characterization (Sun et al. 2016), analysis of external factors that influence the utilization of the network (Briand et al. 2017), and data prediction (Ceapa et al. 2012).

In particular, methods help to estimate and understand a change in behavior (Asakura et al. 2012) of various improvement strategies on transit service reliability (Diab and El-Geneidy 2013). This research aims at understanding and analyzing public transit user behavior. A better understanding of user behavior helps to improve the services provided by public transit. For example, work has been done concerning the optimization of public transit timetables (Nishiuchi et al. 2018), bus stop optimization (El-Geneidy and Surprenant-Legault 2010), and so forth.

The amount of smart card transactions can be in the multi-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 the one hand, there is a range of methods, including partition algorithms (Chevalier and Le Bellac 2013), hierarchical algorithms (Rokach and Maimon 2005), and algorithms based on density (Kriegel et al. 2011). On the other hand, several metrics can be used to evaluate the dissimilarity of two vectors, including Euclidean distance (Deza and Deza 2009), Manhattan distance (Black 2006), cross correlation distance (Mori et al. 2016), and dynamic time warping distance (Giorgino 2009).

Figure 1 illustrates the dynamic time warping method. Dynamic time warping (DTW) 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 $A$ allows the cumulative distance between $A$ and $B$ to be minimal (Fig. 1a).

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

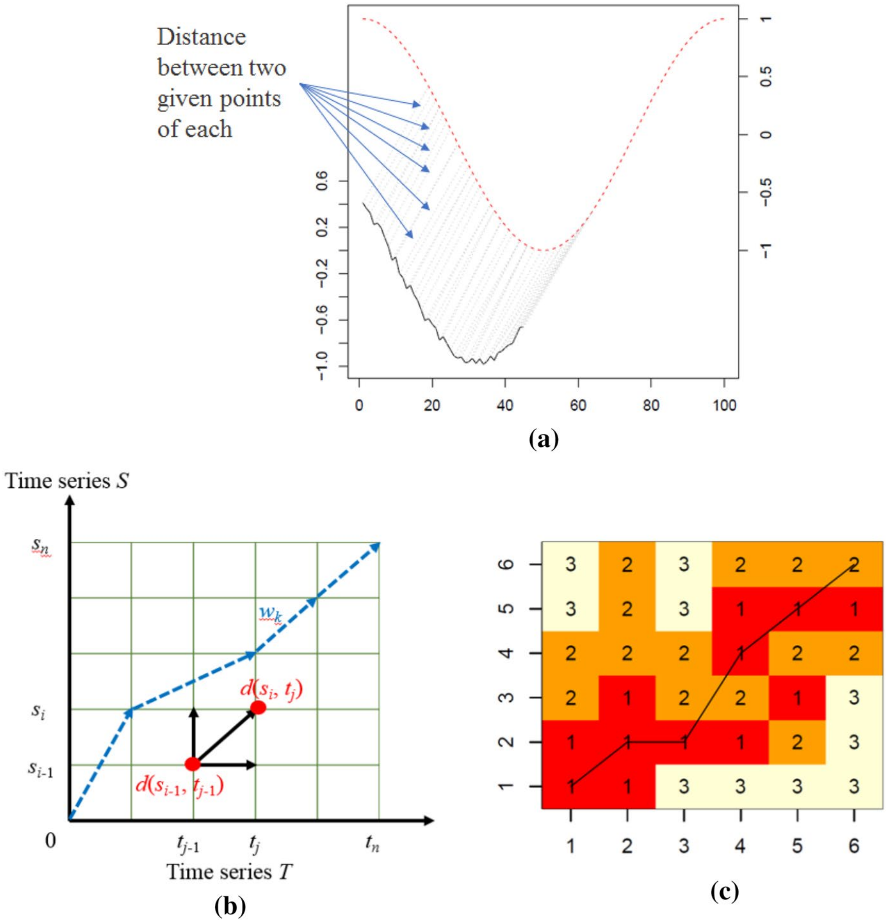


**Fig. 1** Dynamic time warping method

### 2.3 Utilization of data mining in smart card data

An issue of great interest to public transit researchers involves partitioning passengers into clusters based on their trips. The classical data mining technique (k-means and hierarchical clustering) has been used to classify users' general behavior over a period of 12 weeks (Agard et al. 2006). Some other works have been done based on k-means (Morency et al. 2006), neural networks (Ma et al. 2013) and DBSCAN (density-based spatial clustering of applications with noise) (Kieu et al. 2014), which were bound to identify regular passengers, or propose clustering according to their behaviors. Moreover, classification methods can also be developed to analyze the quality levels of transit service (de Oña and de Oña 2015) and to estimate the land-use influence (Nagy et al. 2017).

It is also of great interest to analyze users' behaviors temporally and spatially, based on the temporal data mining method (Ghaemi et al. 2017) and spatial data mining method (Ghaemi et al. 2015); the public transit card user's temporal patterns and spatial patterns have been analyzed separately.

At the end, to verify the efficiency of space–time clustering algorithms, a space–time prism 3D (three-dimensional) plot (Farber et al. 2015) helps show the profile of each cluster. As illustrated in Fig. 2, this 3D plot shows a user's (or a user group's) location for one day. It illustrates not only the difference in a public transit user's individual behavior, but also the difference in a user group.

### 2.4 Limitations of the current methods

Different papers have presented a pertinent methodology on public transit smart card users' behavior detection, the diversity of the classification method, and the application of data mining methods to smart card data. Some articles present original classification methods by using other sources of data (Yoon and Shahabi 2008; Yuan and Raubal 2014). Their methods are based on a time-referenced distance and a spatio-temporal edit distance, which permit to classify trajectories. However, as users' behaviors are considered by means of time series, few articles present time series classification as a means to describe public transit passengers' temporal and spatial behaviors. Compared to an analysis in which the users' behaviors are treated separately at each time point, the time series classification should have contained more information about users' characteristics.

However, time series classification is a special issue because of the limitation of the classical classification method, and the research is based on each individual smart card user's transactions instead of a daily behaviour time series. For example, when clustering using k-means, the algorithm considers only the value of vector elements, not the position of these elements in the vector. The interest in transportation planners is to consider the time of the day in the boarding sequence. To solve this problem, a temporal classification method has been developed based on cross-correlation distance metrics (He et al. 2020), but that method did not consider spatial aspect of the travels, which we include here.

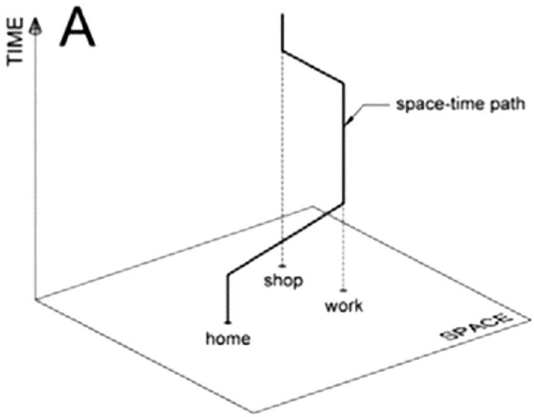


**Fig. 2** Example of space–time prism (Farber et al. 2015)

Even though a temporal classification enables a classification of a user's temporal behaviors into clusters, few articles present a pertinent method on how to classify public transit smart card users' daily behaviors spatially or spatio-temporally. In this article, these problems would be resolved by reconstructing dynamic time warping distance and an application of the sampling hierarchical clustering method. The results would enable transit authorities to offer better service and to satisfy the daily requirements of their passengers.

## 3 Problem setting and objective

By nature, a public transit path is characterized by both the time of day when a boarding activity occurs and the location where it occurred. The most intuitive way of clustering users would be to consider space and time at the same time. In this article, user behavior 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 et al. 2020). Now, we propose integrating the spatial dimension. At each time stamp considered, we calculate and use the physical (Euclidian) distance between actual position in space of the two smart cards.

The aim of this paper is to propose a methodology to classify users' spatiotemporal behaviors using pertinent classification algorithms and distance metrics. The behavior is composed of the sequence of the bus stop locations at each hour. To demonstrate the method, Fig. 3 presents an example of three of the users' daily behaviors:

- the first user leaves home at 06:30 to go to school and returns home at 16:00;
- the second user leaves home at 07:30 to go to work and returns home at 18:00.

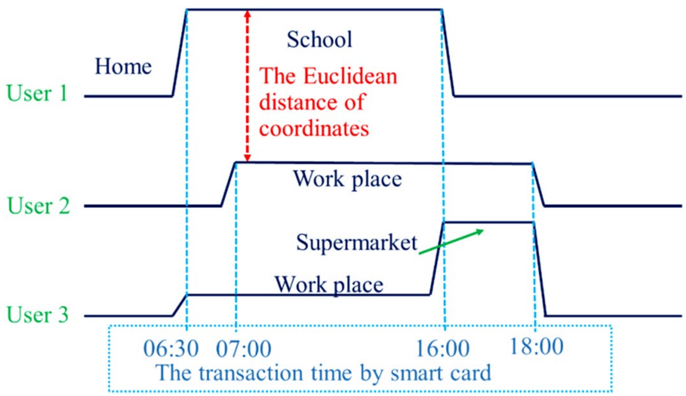


**Fig. 3** Brief example showing three user behaviors to be classified

- the third one also leaves at 06:30 to go to work, but before going home at 18:00, the user goes 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 a few clusters. In this case, if we measure the behaviors of users 1 and 2, which are "more similar" than user 3, then a cluster will be created with users 1 and 2, and user 3 will be in another cluster.

In the spatiotemporal classification, when measuring the dissimilarity of two users' profiles, we consider not only the time of the smart card transaction, but also the real distance between bus stops, serving as proxies for the user's 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.

## 4 Methodology

Figure 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, which are described in the next section. The methodology contains seven steps that are described hereafter.

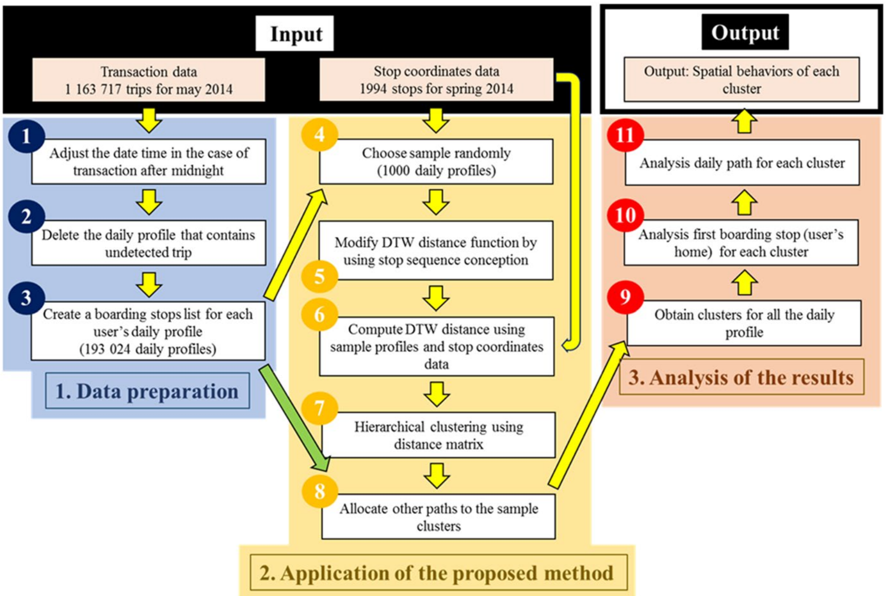


**Fig. 4** Proposed method

## 4.1 Preparation of data and design of spatial and spatiotemporal classification

Regarding preparation of data, first, the smart card transactions are formatted and pre-processed. The trips that occurred after midnight are adjusted so that the trips remain in the same user journey, using a 24+ hour system (step 1 in Fig. 4). For example, a trip that occurred at 1:00 AM the next day is changed to 25:00 the same day.

Secondly, for trajectory classification, we have to use the destinations of the smart card transactions. Smart card data used for this paper do have a 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 that are not estimated) are removed (step 2 in Fig. 4).

Regarding the design of a spatial and spatiotemporal classification, 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). Figure 5 presents three methods to build the time series in this case. The main idea is to link all of the stops in a sequence of given moments (…stop 1 at 11:00, stop 2 at 12:00, stop 3 at 13:00 …until the end of the day). Bullets 2 and 3 of the figure present the method chosen in this paper to build time series for a spatial and spatiotemporal classification. Table 1 presents the characteristics of each approach. In this paper, we use the latter two approaches.

For classical, spatial and spatiotemporal DTW, the differences are in the definition of the point and the metric used to calculate the distance between observations. The classical DTW cannot be used for a spatial classification, because it does not take into account the locations in space, and the spatial DTW cannot be used for a through spatiotemporal classification, because it only takes into

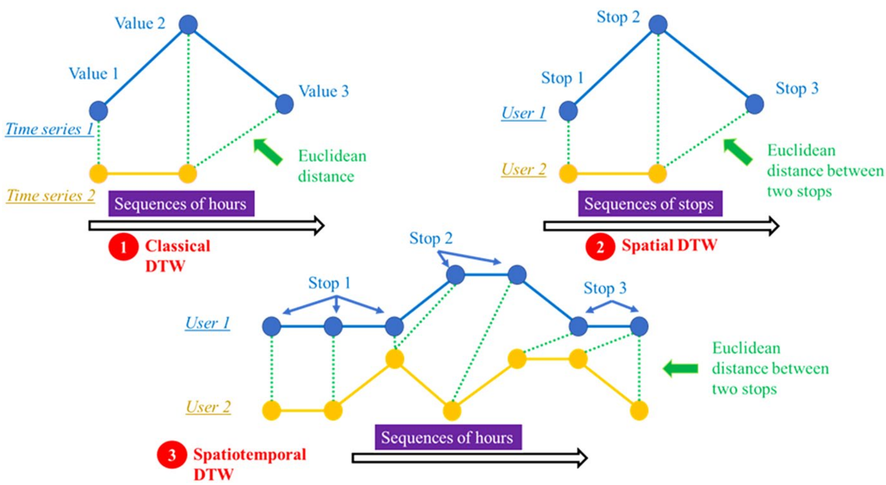


Fig. 5 Comparison of the three DTW methods

**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 relation to time) | Stop sequence (uneven relation to time) |
| Distance between grid points | 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) |

consideration the sequence of stops visited during a day, and not the time of the day where these visits occur.

The usual computation process of DTW is divided into two parts: (1) computation of dissimilarity of two series, and (2) classification of series. In the first part, the dissimilarity is measured by DTW using a provided distance metric. In the end of this procedure, a matrix that contains the dissimilarity of any two is obtained. In the second part, hierarchical clustering is applied to this matrix to group those series into a few clusters.

### 4.2 Application of the sampling method and analysis

The clustering of more than a hundred thousand users' daily profiles is a time-consuming process. The calculation time (when feasible) is way too long, and the amount of computer memory needed would be far too much because of the size of the dissimilarity matrix. To do the clustering, we propose using a sampling approach. This section explains steps 4 to 8 of the methodology.

Figure 6 shows the overall sampling process (He et al. 2020). At first, all observations are provided in Fig. 6a. The red points in Fig. 6b are the randomly selected points. Then, we apply dynamic time warping and hierarchical clustering algorithms to these sample points. Figure 6c presents the clusters created in this example. We used the dendrogram showing the distance between observations to cut a number of clusters suitable to the needs of the analysis.

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

A random sampling of 1000 daily profiles is used in this experiment. Our tests show that such a sample can represent users' behavior from a medium-size city like Gatineau by a proposed indicator that is computed by combining inter-cluster and intra-cluster distance variances. In the test, for a given number of clusters, we gradually increase the size of sampling data to see when the indicator stabilizes.

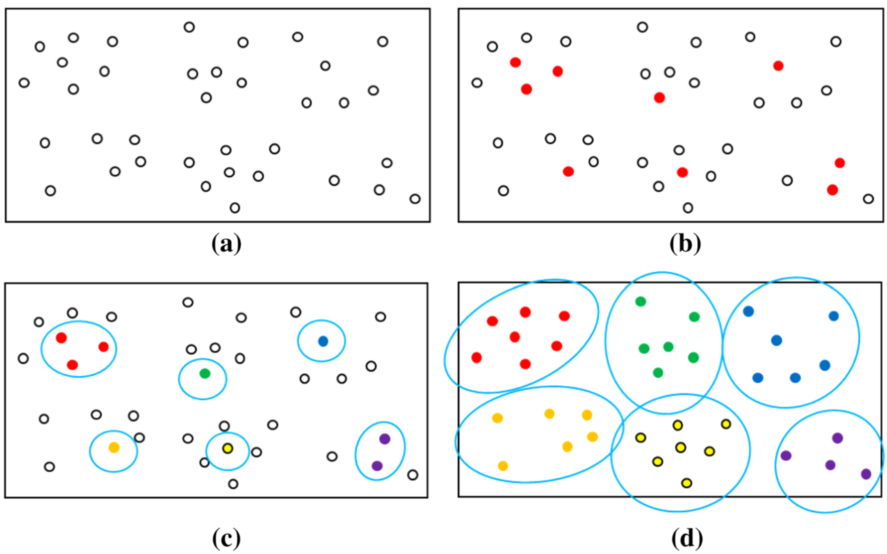


**Fig. 6** Allocation method

Further experiments (see He et al. 2019) show that when the number of clusters is around 10, then the value of the indicator stabilizes when the size of sampling data is around 1000.

Based on the results obtained, we analyze smart card users' behaviors, specifically looking at the boarding stops, daily profiles and space–time path for each cluster (steps 9–10–11 in Fig. 4).

## 5 Case study implementation

The dataset was provided by the *Société de Transport de l'Outaouai*s (STO), a transit authority serving the 280,000 inhabitants of Gatineau, Quebec. The STO authority is a Canadian leader in using fare collection with public transit smart cards. 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 of the weekday transaction data from May 2014 has been used to test the proposed method of spatial classification. This dataset contains 1,163,717 trips.

The method is programmed in Python, which enables us to deal with such a large database.

During implementation, the number of clusters should have been determined by cutting dendrogram branches. Figure 7 shows the dendrogram of a spatial classification algorithm. We cut it into 10 clusters because:

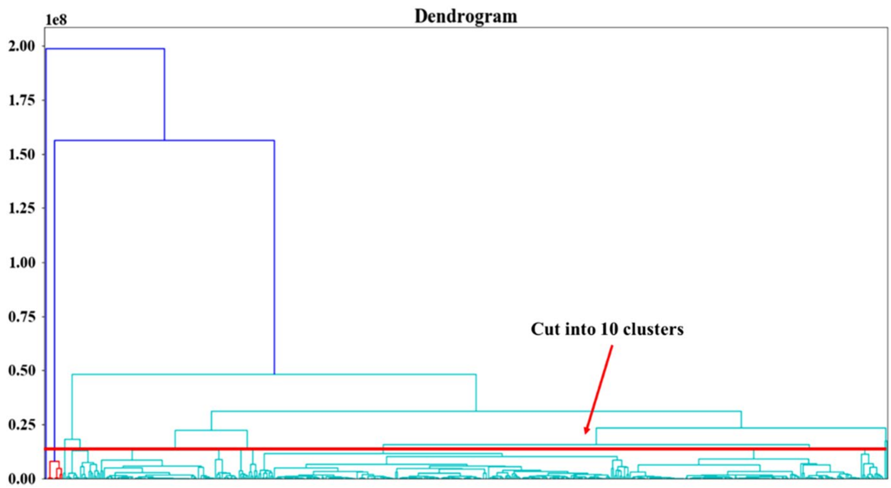


**Fig. 7** Dendrogram of hierarchical clustering of spatial classification algorithm

- We attempted to obtain as even-sized clusters as possible, even though this is not mandatory (user behaviors may not be balanced evenly). We can compare more different behaviors if we were to obtain more clusters.
- In this case, if we increased the number of clusters from 10 to 11, there would be a cluster with too small of a size. Then, after the allocation process, this cluster size would be negligible compared to other clusters. In the analysis, we preferred to keep a larger cluster size.

It is worth noting that using the sampling method allows decreasing the computational time. For example, for the dissimilarity matrix process, the computational time of 1000 user-day profiles is 100 times faster than for 10,000 user-day profiles. In the case study, instead of having to calculate a 193,024 × 193,024 distance matrix ($3.7 \times 10^{10}$ cells), we calculate a 1000 × 1000 matrix ($1 \times 10^6$ cells).

## 6 Results and analysis

### 6.1 Results

An excerpt of the results of the spatial classification is presented in Table 2. Each row in Table 2 is a daily profile, where the column "Stop_list" presents the sequence of stops visited for a daily profile given, and the column "Cluster" presents the classification result for that daily profile.

The daily profile is represented by a combination of "smart card number + date" (card-day). "1185321492030080_2014-05-01" represents the profile of card "1185321492030080" for May 1st, 2014. The card holder visited stop 2060 and

**Table 2** Spatial classification results

| 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 |

5034 in sequence. Then, this daily profile belongs to cluster 7, according to the classification result.

We could find that for the combination "1292322417029248_2014-05-20", many trips were in this daily profile. One of the advantages of DTW is that it can deal with a different number of trips during the day. We can also find that for the user "1000309", the user's spatial behaviors have not been changed, even though there is a minor difference in the boarding stop.

### 6.2 Analysis according to boarding stop

Figure 8 shows the analysis by the first boarding stop for the spatial classification. Every color 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 colors 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 in the orange clusters are located in Hull or Gatineau.

If we used a baseline approach such as K-means, the classification of the first boarding stop (probably home location) would be based solely on coordinates (longitude and latitude). It would not separate orange stops (cluster 2) from other stops of Aylmer (cluster 1). This shows that our method also considers the temporal dimension at the same time.

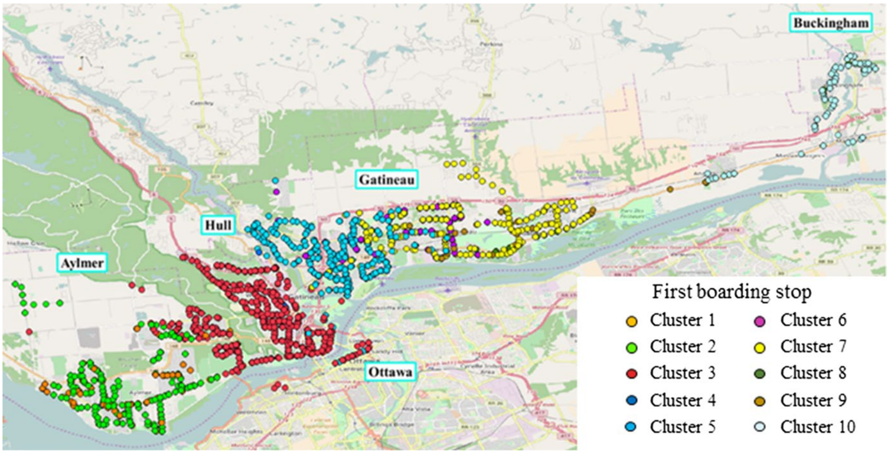


**Fig. 8** Analysis by first boarding stop

### 6.3 Analysis by daily trajectory

Figure 9 shows the daily trajectory of each cluster obtained through spatial classification. By watching the colors, we can see an overview of the characteristics of each cluster. For example, the users from the cyan cluster live 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 behaviors (whether they 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 behaviors helps characterize the demand. Based on this result, we may suggest to the public transit authority to implement new lines or

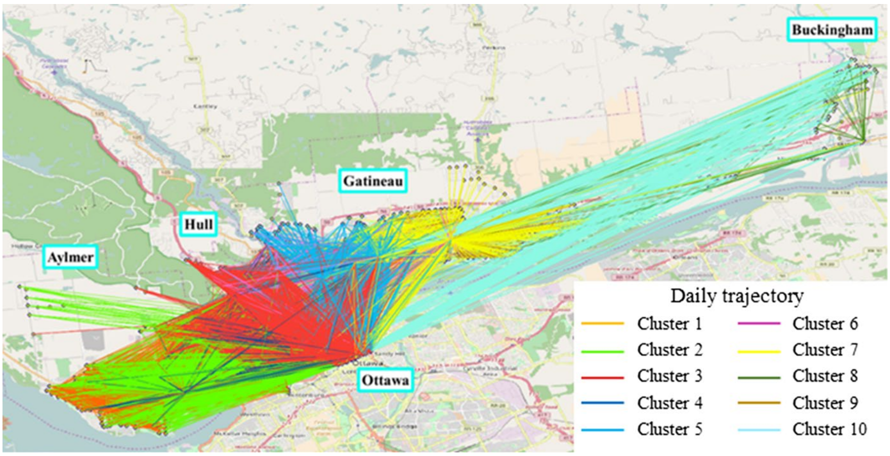


**Fig. 9** Analysis by daily trajectory

enhance the bus service so that the people can travel directly and easily from Buckingham to Ottawa.

### 6.4 Analysis by space–time path

Based on the spatiotemporal classification result, a 3D space–time path prism of each cluster is plotted. Figure 10a shows all profiles individually, and Fig. 10b shows the average path for each cluster. The Z axis of each figure is the hour within the day (the 25th hour is for a 1 a.m. transaction).

In Fig. 10b, even though users of the green cluster live closer to their work location than those of the light blue cluster (both from east of downtown), the green cluster leaves home earlier and returns home later than the light blue cluster. This may be due to an express bus line that links the origin and destination of the light blue cluster. Therefore, it would be possible to suggest to the public transit authority to implement an express bus line to serve the users in the green cluster so that they could save more time when commuting.

It is also possible to find that the behavior of the light green cluster is stable during work hours (during 9:30–15:00, the location of the light green cluster does not change a lot). That means these users only travel locally. It would be possible to suggest to the public transit authority to implement a special bus line for these users. This new bus line could link the origin and destination of the light green cluster, and this would only operate during peak hours, but it could still respond well to the demand of this cluster.

## 7 Conclusion

Regarding contribution, in this paper, a new methodology based on dynamic time warping, hierarchical clustering and the sampling method is proposed to classify public transit smart card users' spatiotemporal behaviors. The results demonstrate that the behaviors can be distinguished well. Based on the results, it is possible to suggest enhancements to the public transit authority to better serve customers from specific clusters.

Regarding limitations, first, the dynamic time warping algorithm is quadratic; therefore, the computation time is long. Secondly, the separation criterion is based on distance, so different behaviors can still stay in the same cluster because their dissimilarity of other factors is not considered (for example, the purpose of travel is not a dissimilarity criterion in this case). Other limitations come from the data: the estimation of destinations may not be perfect (it was not validated) and therefore this might hamper the results of the clustering method.

Regarding perspective, in the future, some works are proposed 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 reduce computation time of the dynamic time warping method. Finally, more methods are

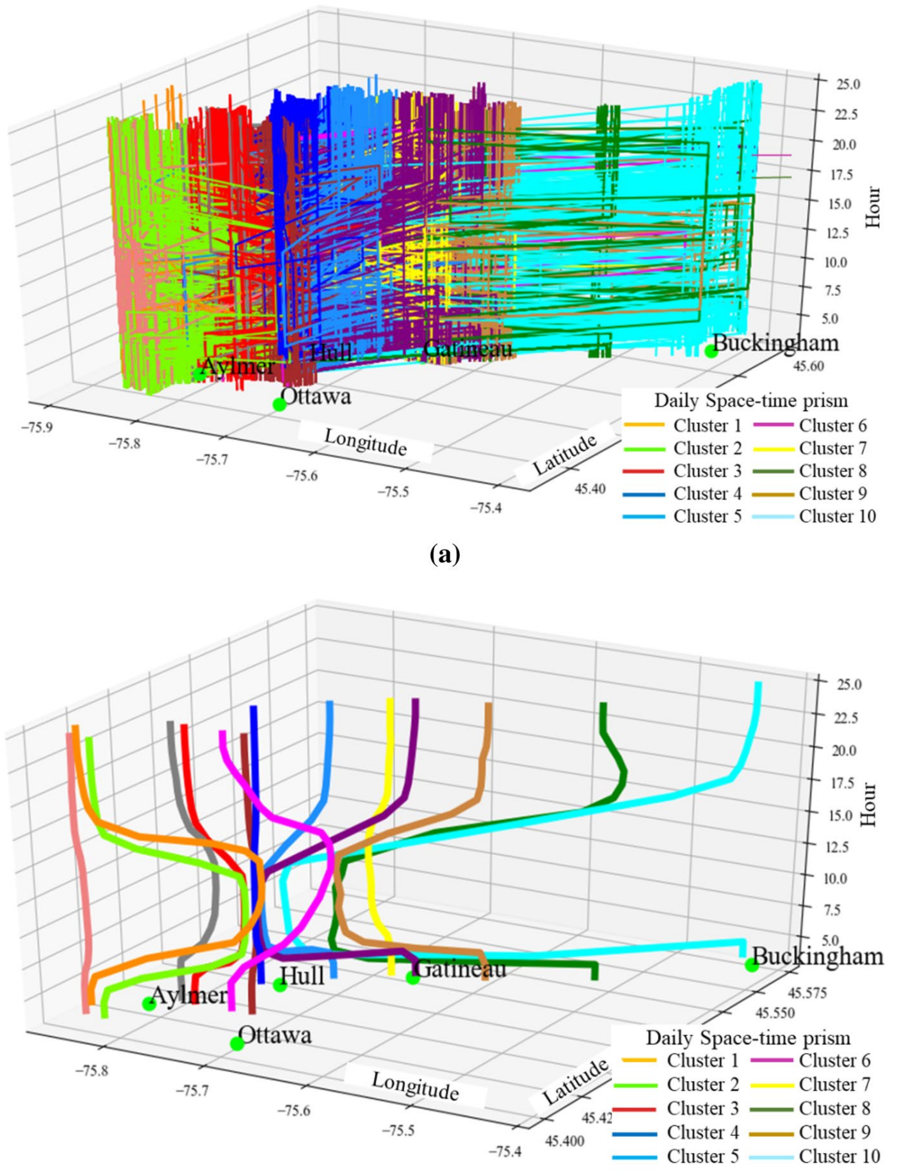


**Fig. 10** Space–time prism of clusters **a** all observations, **b** center of groups

needed to cross-validate the clustering results even though these results seem intuitive as they were validated by the public transit planners. A statistical measure could be developed to calculate whether inter-cluster dissimilarity is as large as possible, and whether intra-cluster dissimilarity is as low as possible. In addition, by using smart card data, the users' travel distance of different zones can be computed, which may be used to check whether inter-zone and intra-zone behavior are well separated.

**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 Natural Science and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

## References

Agard B, Morency C, Trépanier M (2006) Mining public transport user behaviour from smart card data. IFAC Proc Vol 39(3): 399–404
Asakura Y, Iryo T, Nakajima Y, Kusakabe T (2012) Estimation of behavioural change of railway passengers using smart card data. Public Transp 4(1):1–16
Black PE (2006) Manhattan distance. Dictionary of algorithms and data structures. http://xlinux.nist.gov/dads//
Briand AS, Côme E, Trépanier M, Oukhellou L (2017) Analyzing year-to-year changes in public transport passenger behavior using smart card data. Transp Res Part C Emerg Technol 79:274–289
Bunker JM (2018) High volume bus stop upstream average waiting time for working capacity and quality of service. Public Transp 10(2):311–333
Ceapa I, Smith C, Capra L (2012) 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, Le Bellac J (2013) La classification. Université de Rennes. https://docplayer.fr/13650741-La-classification-2012-2013-fabien-chevalier-jerome-le-bellac.html
Chu KA, Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. Transp Res Rec J Transp Res Board 2063:63–72
Covic F, Voß S (2019) Interoperable smart card data management in public mass transit. Public Transp 11(3):523–548
de Oña R, de Oña J (2015) Analysis of transit quality of service through segmentation and classification tree techniques. Transportmetrica A Transp Sci 11(5):365–387
Deza MM, Deza E (2009) Encyclopedia of distances, 3rd edn. Springer, Berlin
Diab EI, El-Geneidy AM (2013) Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. Public Transp 4(3):209–231
El-Geneidy AM, Surprenant-Legault J (2010) Limited-stop bus service: an evaluation of an implementation strategy. Public Transp 2(4):291–306
Farber S, O'Kelly M, Miller HJ, Neutens T (2015) Measuring segregation using patterns of daily travel behavior: a social interaction based model of exposure. J Transp Geogr 49:26–38
Ghaemi MS, Agard B, Nia VP, Trépanier M (2015) Challenges in spatial-temporal data analysis targeting public transport. IFAC-PapersOnLine 48(3):442–447
Ghaemi MS, Agard B, Trépanier M, Partovi Nia V (2017) A visual segmentation method for temporal smart card data. Transportmetrica A Transp Sci 13(5):381–404
Giorgino T (2009) Computing and visualizing dynamic time warping alignments in R: the dtw package. J Stat Softw 31(7):1–24
He L, Trépanier M (2015) Estimating the destination of unlinked trips in transit smart card fare data. Transp Res Rec J Transp Res Board 2535:97–104
He L, Trépanier M, Hickman M, Nassir N (2015) Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems (No. CIRRELT-2015-52). CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport= Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation
He L, Trépanier M, Agard B (2019) Sampling method applied to the clustering of temporal patterns of public transit smart card users. CIRRELT research paper, CIRRELT-2019-30, p 18

He L, Agard B, Trépanier M (2020) A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. Transportmetrica A Transp Sci 16(1):56–75

Imaz A, Habib KMN, Shalaby A, Idris AO (2015) Investigating the factors affecting transit user loyalty. Public Transp 7(1):39–60

Kieu LM, Bhaskar A, Chung E (2014) Transit passenger segmentation using travel regularity mined from Smart Card transactions data. In: Proc., Transportation Research Board 93rd annual meeting, National Research Council, Washington, DC

Kriegel HP, Kröger P, Sander J, Zimek A (2011) Density-based clustering. Wiley Interdiscipl Rev Data Min Knowl Discov 1(3):231–240

Kurauchi F, Schmöcker JD, Shimamoto H, Hassan SM (2014) Variability of commuters' bus line choice: an analysis of oyster card data. Public Transp 6(1–2):21–34

Lee SG, Hickman M (2014) Trip purpose inference using automated fare collection data. Public Transp 6(1–2):1–20

Ma X, Wu YJ, Wang Y, Chen F, Liu J (2013) Mining smart card data for transit riders' travel patterns. Transp Res Part C Emerg Technol 36:1–12

Morency C, Trépanier M, Agard B (2006) Analysing the variability of transit users behavior with smart card data. In: Intelligent transportation systems conference, 2006. ITSC'06. IEEE, pp 44–49

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

Mori U, Mendiburu A, Lozano JA (2016) Distance measures for time series in R: the TSdist package. The R Journal 8(2):451–459

Nagy V, Horváth B, Horváth R (2017) Land-use zone estimation in public transport planning with data mining. Transp Res Procedia 27:1050–1057

Nishiuchi H, Kobayashi Y, Todoroki T, Kawasaki T (2018) Impact analysis of reductions in tram services in rural areas in Japan using smart card data. Public Transp 10(2):291–309

Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C Emerg Technol 19(4):557–568

Rokach L, Maimon O (2005) Clustering methods. Data mining and knowledge discovery handbook. Springer, Boston, pp 321–352

Shi X, Hangfei L (2014) The analysis of bus commuters' travel characteristics using smart card data: the case of Shenzhen, China. In: 93rd annual meeting of the Transportation Research Board, Washington, D.C. (No. 14-2571)

Sun Y, Shi J, Schonfeld PM (2016) Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: a case study of Shanghai Metro. Public Transp 8(3):341–363

Trépanier M, Barj S, Dufour C, Poilpré R (2004) Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain. In: Congrès de l'Association des transports du Canada

Trépanier M, Chapleau R, Tranchant N (2007) Individual trip destination estimation in a transit smart card automated fare collection system. J Intell Transp Syst 11(1):1–14

Viggiano C, Koutsopoulos HN, Wilson NH, Attanucci J (2017) Journey-based characterization of multimodal public transportation networks. Public Transp 9(1–2):437–461

Yoon H, Shahabi C (2008) Robust time-referenced segmentation of moving object trajectories. In: 2008 eighth IEEE international conference on data mining, pp 1121–1126. IEEE

Yuan Y, Raubal M (2014) Measuring similarity of mobile phone user trajectories—a spatio-temporal edit distance method. Int J Geogr Inf Sci 28(3):496–520