Transportmetrica A: Transport Science

ISSN: 2324-9935 (Print) 2324-9943 (Online) Journal homepage: www.tandfonline.com/journals/ttra21

A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method

Li He, Bruno Agard & Martin Trépanier

To cite this article: Li He, Bruno Agard & Martin Trépanier (2020) A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method, Transportmetrica A: Transport Science, 16:1, 56-75, DOI: 10.1080/23249935.2018.1479722

To link to this article: https://doi.org/10.1080/23249935.2018.1479722

Published online: 12 Jun 2018.

TRANSPORTMETRICA A: TRANSPORT SCIENCE
2020, VOL. 16, NO. 1, 56–75
https://doi.org/10.1080/23249935.2018.1479722

A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method

Li He, Bruno Agard and Martin Trépanier

Department of Mathematics and Industrial Engineering and CIRRELT, Polytechnique Montréal, Montréal, Québec, Canada

ABSTRACT

A classification of the behavior of smart card users is important in the field of public transit demand analysis. It provides an understanding of people's sequence of activities within a period of time. However, classical metrics such as Euclidean distance is not appropriate when dealing with time-series classification. To solve this problem, in this article a method for the classification of public transit smart card users' daily transactions, which are represented in time series, is presented. The chosen approach uses cross-correlation distance (CCD), hierarchical clustering, and subgroups by metric parameter to understand the users' temporal patterns. The clustering results are compared with dynamic time warping (DTW) distance (a common method to measure time-series distance). After a brief pedagogical example to explain the DTW and CCD concepts, a program is developed in R to validate the method on a real dataset of smart card data transactions. The dataset concerns the use of the public transit system in the city of Gatineau in September 2013. The results demonstrate that CCD performs better than DTW to classify the time series, and that the classification method identifies different public transit users' daily behaviors. The results will help transit authorities to offer better services for smart card users from diverse groups.

ARTICLE HISTORY
Received 2 June 2017
Accepted 18 May 2018

KEYWORDS
Public transportation data; smart card users' behavior; time-series classification; cross-correlation; dynamic time warping

1. Introduction

The extraction of customer behavior in public transit systems is of great interest in scientific communities (Jou, Lam, and Wu 2007). Having a better understanding of travelers' behavioral patterns is helpful in assessing the demand for transportation services (Joh, Timmermans, and Arentze 2006). We can now take advantage of automated payment-based smart card technology that generates and stores gigabits for data on the day-to-day activities of users. The time-series technique is widely used in customer behavior forecasting (Chen, Chang, and Chang 2009). A time series represents a collection of values obtained from sequential measurements over time (Esling and Agon 2012). The segmentation of

CONTACT Li He li.he@polymtl.ca Department of Mathematics and Industrial Engineering and CIRRELT, Polytechnique Montréal, 2500 ch. de Polytechnique, Montréal, Québec, Canada H3T 1J4

transit users allows user activity to be synthesized into a limited number of groups of typical behaviors. Knowledge from those groups may then be used to improve the service.

Data from automatic fare systems of public transit can be analyzed in many directions. For example, the data help evaluate and predict public transit users' demand (Kurauchi and Schmöcker 2017). Some analyses have been done to cluster smart card users' temporal behaviors using data mining: Agard, Morency, and Trépanier (2006) and Nishiuchi, King, and Todoroki (2013) designed a method to mine the patterns of card users' daily frequency. Many authors have suggested definitions (Das and Pandit 2015), metrics, tools (Bordagaray et al. 2014), models (Li, Schmöcker, and Fujii 2015), algorithms (Chang et al. 2010), and methods (Del Castillo and Benitez 2013) to help better understand the mobility of users during different time periods.

In data mining, most classification methods are based on distance metrics between observations. Traditional methods used to measure the distance of samples includes Euclidean distance (Berkhin 2006) or Manhattan distance (Bakar et al. 2006). Other derivative methods such as Minkowski distance (Jain, Murty, and Flynn 1999) is a generalization of both the Euclidean and Manhattan distances (Lhermitte et al. 2011). However, these methods do not adhere to a conception of time process, which prevents methods for analyzing smart card users' behavior series. Some other distance measure methods can deal with near-time points, such as cross-correlation distance (CCD) (Liao 2005) or dynamic time warping (DTW) distance (Berndt and Clifford 1994). All of these distance-calculating methods have been implemented in R (Meyer and Buchta 2015). The selection of a pertinent metric is critical for clustering methods. Various performant clustering methods are actually available, such as partition methods and hierarchical clustering methods (Langfelder, Zhang, and Horvath 2008).

The classification of transit users into subgroups is of great interest to transportation planners. Schedules and transit networks can better suit travelers' needs if it is possible to identify those needs. It can also serve to define fare strategies and offer market-oriented services. This could be an advantage for both transit users and authorities. The objective of this paper is to propose a transit user classification approach to classify time series from a smart card user's profile, by combining time-series metrics and data mining methods. To this end, Section 2 provides a state of the art on the application of data mining on public transit smart card data, and then classification methods are outlined as well as distance metrics between time series. In Section 3, the method uses CCD and DTW distance combined with hierarchical clustering to extract users' temporal behavior. Then, in Section 4, the performance of the developed algorithms is compared. Finally, in Section 5, a real database is tested to see how it performs.

2. State of the art

2.1. Use of data mining in public transit smart card data

Data collected from an automatic collection system (in this case, smart card data) can be used to understand characteristics of public transit card users (Pelletier, Trépanier, and Morency 2011). Much research has been done to explore the potential information from smart card data.

(1) Data Completion and Preparation

Due to the characteristics of smart card data, some preparation must be made before an analysis of a user's behavior can be done. An algorithm has been developed to estimate the alighting location given a smart card user's boarding location (Trépanier, Tranchant, and Chapleau 2007). This algorithm has been improved using kernel density estimation to estimate the unlinked trips (He and Trépanier 2015). Furthermore, this algorithm has been calibrated to obtain a more accurate estimation of destinations (He et al. 2015). In addition, some research focuses on transfer detection (Chu and Chapleau 2008), trip purpose inferences (Lee and Hickman 2014), etc.

(2) Classification of Transit Smart Card Users' Behavior

An issue of great interest to transport operators involves partitioning network passengers into groups based on their transportation network activity. Clustering approaches are used, such as the Hierarchical Ascendant Classification (HAC) or k-means algorithm (Agard, Morency, and Trépanier 2006). Based on temporal analysis (Ghaemi et al. 2017) and spatial analysis (Ghaemi et al. 2015), the public transit card user's temporal patterns and spatial patterns are discovered and analyzed. Data mining even helps predict user demand (Nuzzolo and Comi 2016). Moreover, by using data mining, especially the classification technique, a methodology has been developed to analyze the quality level of transit service (de Oña and de Oña 2015). Langlois, Koutsopoulos, and Zhao (2016) perform analyses of multi-week activity patterns using clustering, and these authors propose a representation of longitudinal activity sequences using temporal and spatial activity. The groups created by the clustering are associated with distinct sequence structures, allowing for better knowledge of several weeks of passenger activity (Briand et al. 2017).

(3) Limitation of the Current Methods

These papers present a pertinent methodology to explain user behavior using smart card data. However, the research is based on each individual smart card user's transactions instead of daily behavior 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 of transportation planners is to consider the time of the day in the boarding sequence. In fact, the current classification methods are not suitable to solve this problem, because they are not designed to measure the dissimilarities between time series. The introduction of the time-series classification technique helps develop a method in which to analyze a smart card user's daily profile, so that public transit authorities can offer better service that will satisfy passengers' daily requirements.

2.2. Classification

In this document, classification and clustering will be used as synonyms. A cluster is a collection of data objects arranged so that an object is similar to another within the same cluster, and dissimilar to the objects in other clusters. A classification method groups a set of data objects into clusters. A good clustering method will produce high-quality clusters with high intra-class similarity and high inter-class dissimilarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation (Subbiah 2011). There are several major approaches to classification.

(1) Partitioning Algorithms

Partitioning algorithms construct various partitions and then evaluate them according to certain criteria. The major idea is to find a partition of *k* clusters that optimizes the chosen partitioning criterion given to *k* number of partitions. The two main heuristic methods are k-means and k-medoids (Subbiah 2011). In the first, each cluster is represented by the center of the cluster, while in the second each cluster is represented by one of the objects in the cluster. Partitioning algorithms have both advantages and limits when treating a time series. For example, the k-means can deal with large datasets, but it uses traditional distance metrics between vectors.

(2) Hierarchical Algorithms

Hierarchical clustering (also called hierarchical cluster analysis) is a method of cluster analysis that seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types (Rokach and Maimon 2005): agglomerative and divisive. For the first, each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. This is a ‘bottom-up’ approach. For the second, all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. This is a ‘top-down’ approach. Compared to partitioning algorithms, hierarchical clustering is available for a variety of distances but it cannot deal easily with large datasets due to high computational costs.

(3) Other algorithms

There are other methods that were used in the case of transit data:

- DBSCAN is used to classify regular passengers based on the dissimilarity of the last alighting stop and first boarding stop of the day and boarding time (Kieu, Bhaskar, and Chung 2014).
- Neural networks can also be used to classify regular passengers; this is based on boarding stops, then boarding time (Ma et al. 2013).
- Naïve Bayesian network is used to estimate trip purpose based on time and location of boarding and alighting stops (Kusakabe and Asakura 2014).
- Continuous hidden Markov model subgroup puts all users into eight groups based on their start time and duration of activity, land use around stops (Han and Sohn 2016).

2.3. Distance between time series

A time series is a set of observations x_t, with each recorded at a specific time t. A discrete-time series (the type to which this article is primarily devoted) is one in which the set T_0 of times at which observations are made is a discrete set (Brockwell and Davis 2002).

Comparing it to other vectors, a time series contains a relationship among the time t itself. For example, for a time series $x_1, x_2, x_3, \ldots, x_n$, with the corresponding specific time $t_1, t_2, t_3, \ldots, t_n$, we know that t_1 is closer to t_2 than to t_n regarding time, regardless of the value of x_1, x_2 and x_n.

Various distance metrics exist to measure the (dis)similarity between two vectors (He, Trépanier, and Agard 2017). In this part, four types of distance are presented: Euclidean distance, Manhattan distance, CCD, and DTW distance. The first two distances are basic ones traditionally used in the classification methods presented earlier. Even if those metrics are not dedicated to measuring the distance between time series, it is still a common practice in transportation research (Agard, Morency, and Trépanier 2006; Morency, Trepanier, and Agard 2007; Ghaemi et al. 2015). On the other hand, CCD and DTW are designed to compare the (dis)similarity of time series but are not actually incorporated in classification methods.

2.3.1. Euclidean and Manhattan distances

Euclidean distance is the straight-line distance between two points in Euclidean space (Deza and Deza 2009). Let x_i and v_j each be a P-dimensional vector. The Euclidean distance is computed as (Liao 2005):

$$d_\mathrm{E} = \sqrt{\sum_{k=1}^{P} (x_{ik} - v_{jk})^2}. \tag{1}$$

Manhattan distance is computed between the two numeric series using the following formula (Mori, Mendiburu, and Lozano 2016):

$$d_\mathrm{M} = \sum_{k=1}^{P} |x_{ik} - v_{jk}|. \tag{2}$$

According to functions (1) and (2), for both distances the result of the distance would not be changed if the order of k is changed; for example, if the positions of k_1 and k_2 are exchanged, the distance remains the same. However, a time series contains relationship among the time t itself; this is a characteristic that makes time series different from other vectors. For a time series, if the values of k_1 and k_2 are exchanged, the distance result should change. Therefore, the Euclidean distance and Manhattan distance are not suitable for time series. Besides, some effort (Ghaemi et al. 2017) has been made to classify smart card users' daily transaction times.

2.3.2. Cross-correlation distance

This distance is based on the cross-correlation between two time series (Mori, Mendiburu, and Lozano 2016). The similarity of two time series is measured by shifting one time series to find a maximum cross-correlation with another time series. The CCD between two time series at lag k is calculated as:

$$CC_k(X, Y) = \frac{\sum_{i=0}^{N-1-k} (x_i - \bar{x})(y_{i+k} - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_{i+k} - \bar{y})^2}}, \tag{3}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the series. Based on this, the distance measure is defined as:

$$CCD(X,Y) = \sqrt{\frac{(1 - CC_0(X,Y))}{\sum_{k=1}^{\max} CC_k(X,Y)}}. \tag{4}$$

In R, the distance measure can be calculated by using a function. This function will return the distance between two time series by specifying two numeric vectors (x and y) and maximum lag.

2.3.3. Dynamic time warping

DTW is a popular technique for comparing time series, providing a distance measure that is insensitive to local compression and stretches and warping, which optimally deform one of the two input series on the other (Giorgino 2009). The method to calculate the DTW is as follows (Berndt and Clifford 1994):

$$S = s_1, s_2, \ldots s_i, \ldots, s_n, \tag{5}$$

$$T = t_1, t_2, \ldots t_j, \ldots, t_m. \tag{6}$$

The sequences S and T can be arranged to form a n-by-m plane or grid, where each grid point (i,j) corresponds to an alignment between elements s_i and t_j. A warping path, W, maps or aligns the elements of S and T, such that the ‘distance’ between them is minimized.

$$W = w_1, w_2, \ldots w_k, \ldots, w_P. \tag{7}$$

That is, W is a sequence of grid points, where each w_k corresponds to a point $(i, j)_k$.

To formulate a dynamic programming problem, a distance measure between two elements is indispensable. Two possible distance measures are usually used for a distance function d. They are the magnitude of the difference (8) or the square of the difference (9),

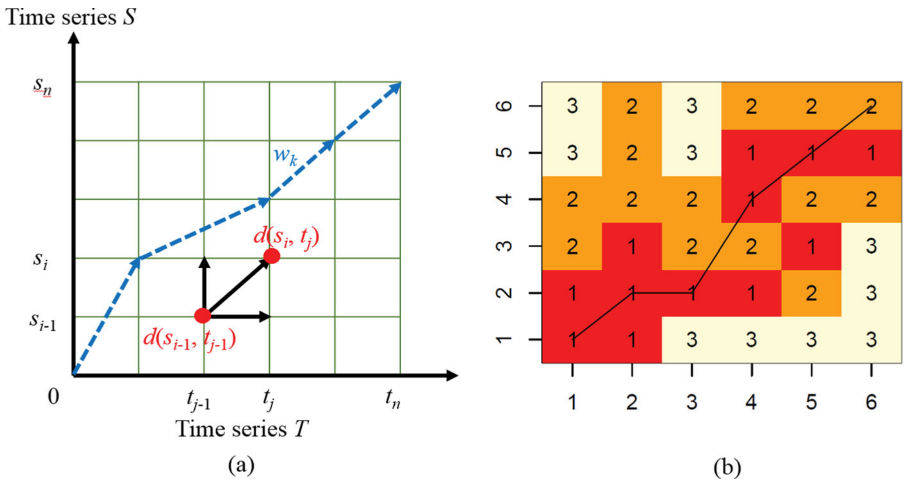
$$d(i,j) = |s_i - t_j|, \tag{8}$$

$$d(i,j) = (s_i - t_j)^2. \tag{9}$$

Once a distance measure is selected, the DTW problem can be defined as minimization over potential warping paths based on the cumulative distance for each path, where d is a distance measure between two time-series elements.

$$DTW(S,T) = \min w \left[\sum_{k=1}^{P} d(w_k) \right]. \tag{10}$$

Figure 1 illustrates the DTW method. In Figure 1(a), to obtain a minimum cumulative distance, the time series can be warped to the next time point (moment). For example, grid point (s_{i-1}, t_{j-1}) can be warped to (s_i, t_{j-1}), (s_{i-1}, t_j), (s_i, t_j) to compute each distance. A sequence of grid points w_k can be a path from (s_0, t_0) to (s_m, t_n). On every grid point, the distance between two time points (moments) d (s_i, t_j) should be computed, as shown in Figure 1(b). Then, all possible paths from grid point (1, 1) to (6, 6) are calculated, to find the path with the minimum cumulative distance. In this grid in Figure 1(b), the distance of DTW is 7.



(a)

(b)

Figure 1. Dynamic time warping example (Giorgino 2009).

2.4. CCD and DTW parameters

CCD parameters:

(1) Correlation coefficient. The Pearson correlation coefficient measures the strength of the linear association between two variables (Sedgwick 2012). A correlation coefficient close to $+1$ or -1 represents a strong correlation.

(2) Max lag. This argument represents the maximum delay accepted to compare one time series to another. Figure 2(a) illustrates parameter 'lag' for CCD. The first time series can be shifted to the right one unit, so that the first and second time series will better correspond. The lag can be explained as the number of units needed to shift, so that one time series will be aligned to another. In Figure 2(a), if the max lag is 1, then the first time series can be shifted 1 unit to align second time series. In this way, the second time series can be accepted by the first time series, so the first and second time series will be in the same group, and the third time series will be in another group.

DTW parameter:

The parameter window represents the maximum time that a time series can be warped. Figure 2(b) illustrates parameter 'window' for DTW. The values of elements in time series are warped, so that two time series can be compared. For example, if value of the fourth time point of the first time series is warped from 0 to 1, then the first and second time series will be the same. The parameter window can be explained as the maximum change allowed caused by warping. In Figure 2(b), if the window is 1, then the first time series can be warped 1 unit so that the first time series and the second will be the same. This way, the second time series can be accepted by the first time series, so that the first and second time series will be in the same group, and the third time series will be in another group.

To avoid having the first point of a time series compared to the last point of another time series, the parameters (max lag, window) have been set. The points of two time series will

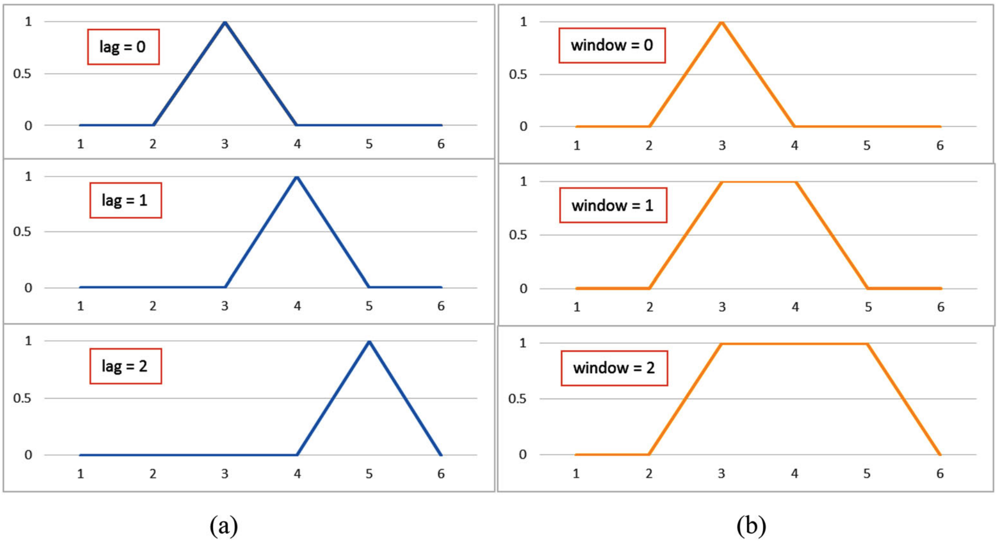


Figure 2. Time-series metrics parameter – (a) lag for CCD and (b) window for DTW.

be computed, and it will ensure that the dissimilarity of two users' behaviors is measured as much as possible.

3. Proposed methodology for the classification of a time series

3.1. Algorithm design

The following three steps are proposed for the time-series classification as presented in Figure 3.

- Step 1 The input data are time series on transit smart card activities' data. On the first part, pairwise distances are computed with CCD. The output is a distance matrix between any two time series. On the other part, DTW is used to perform this step.
- Step 2 Hierarchical clustering is computed to cluster the time series using the distance matrices of CCD and DTW. The results of hierarchical clustering are presented in a dendrogram, from which the clusters are selected. The output is the clusters, which includes all of the observation points. At the end of this step, we obtain the result of the series classification by CCD on the one side and by DTW on the other side. For CCD, a finer result will be obtained in step 3.
- Step 3 Each cluster from the CCD side is separated using the CCD parameters: correlation coefficient and maximum lag.

In conclusion, both ways (DTW and CCD) were used to make clusters of time series from a mathematical point of view. A comparison of results and performance will enable the most efficient method to be defined from a practical point of view in the classification of public transit smart card data temporal profiles.

The time series are first separated by the correlation coefficient. In this example, if series 1 and 3 are positive, then series 2 is negative. Therefore, the pattern of series 2 is opposite

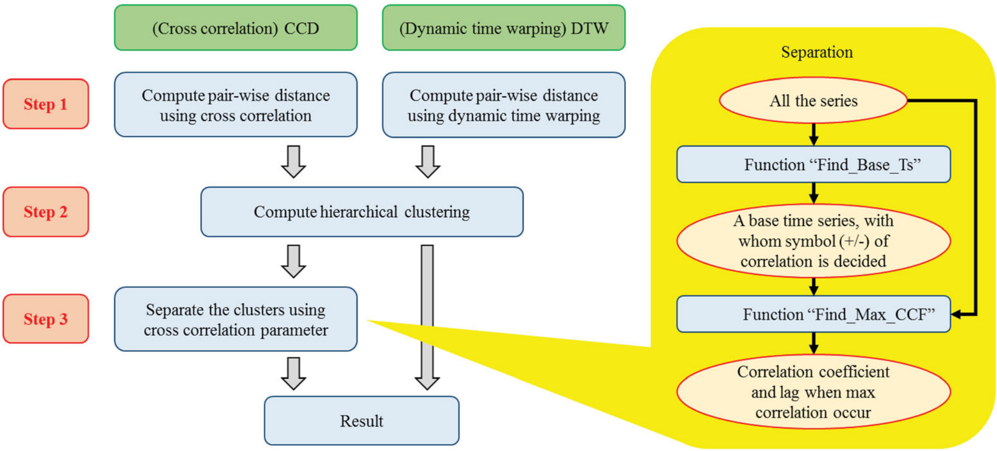

(Cross correlation) CCD
(Dynamic time warping) DTW
Step 1
Compute pair-wise distance using cross correlation
Compute pair-wise distance using dynamic time warping
Step 2
Compute hierarchical clustering
Step 3
Separate the clusters using cross correlation parameter
Result
Separation
All the series
Function "Find_Base_Ts"
A base time series, with whom symbol (+/-) of correlation is decided
Function "Find_Max_CCF"
Correlation coefficient and lag when max correlation occur

Figure 3. Proposed algorithm for the time-series classification.

to that of series 1 and 3. Then, a separation is done by maximum lag (k in the function (4)) in the middle left part of Figure 4. The bottom parts of Figure 4 show the results. If a lag of series whose peak occurs in the time point 3 is 0, then lag of series whose peak occurs in the time point 4 is 1. That means the series 1 has been shifted by 1 time unit (lag) to align to series 3, then the lag of the series 3 compared to the series 1 is 1. In this way, the relation between series 1, 2, and 3 in the cluster is obtained by CCD and the hierarchical method.

3.2. *Implementation*

Implementation contains three main steps as shown in Figure 3:

- Step 1 First, in some cases, a pre-treatment is needed to deal with the original database. For example, series whose values are all 0 or 'NA' are removed. However, giving all of the values a scale is not necessary; or, some values that are treated are not original. Then, the CCD and DTW are computed to calculate the dissimilarity of any two time series.
- Step 2 Compute hierarchical clustering method. At the end of this step, the clusters in which the correlation coefficient and lag are not separated are obtained.
- Step 3 Separate the clusters using the CCD parameter (correlation coefficient and lag). The most important part is in the right rounded corner rectangle:
  - Step 3.1 Firstly, a function 'Find_Base_Ts' is applied to all the series in a certain cluster. It will return a base time series whose correlation is positive and lag is 0.
  - Step 3.2 Based on this time series, another function 'Find_Max_CCF' is applied. This function will return correlation coefficients and lags relating to the base time series of all the other time series in a cluster. With the correlation coefficient and lag, a minimum CCD between the base time series and a given series in the cluster can also be obtained.

Finally, three values are obtained for a time series: (1) Cluster, in which this time series has the best correlation with the other time series in the same cluster. (2) Correlation, the symbol of the base series. (3) Lag, the best one with which a minimum CCD can be obtained.

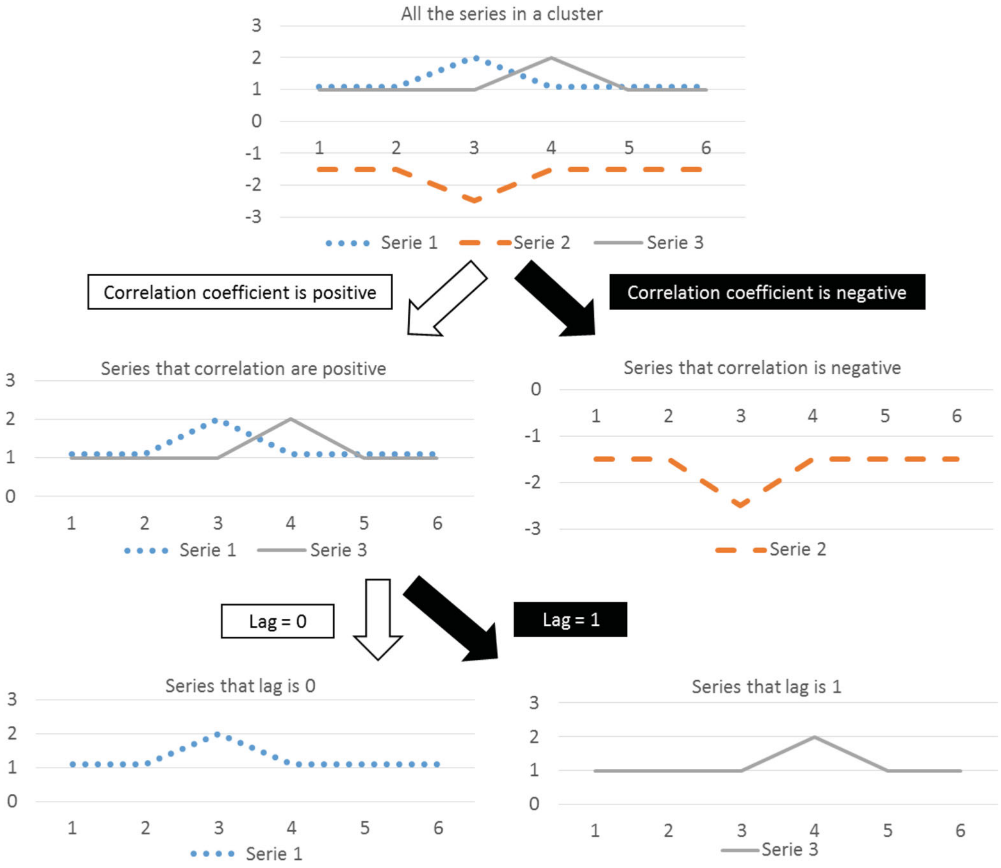


Figure 4. Separating by parameters in a cluster.

4. Comparison between CCD and DTW for classifying transit smart card data

4.1. A pedagogical example

In this section, we apply the methods of public transit smart card data coming from an automated fare collection system. We need them to classify the transit users according to their travel behavior. In the following section, an application to the real public transit smart card data in the city of Gatineau, in Canada, demonstrates the efficiency of the classification of smart card users' daily transaction time series. A sample, which contains 26 time series (user's daily behavior) of 7 time periods, is exemplified. By similarity with the real dataset, the values in these time series are 0 or 1, as shown in Table 1. '1' at a time period (TP*i*) means that a transaction has been registered during this time period while '0' means no transaction.

In this table, for example, the value of the first time series is $T_1 = (1, 0, 0, 0, 0, 0, 0)$, means that a transaction of the smart card number 1 happened in the first time period and nothing in the remaining periods. In this table, smart cards series 14–26 have symmetric behavior to the series 1–13. 'Symmetric' here means that if a value of a certain time point is 0 in a series, then the value in the same time period in the other series is 1.

Table 1. A pedagogical example.

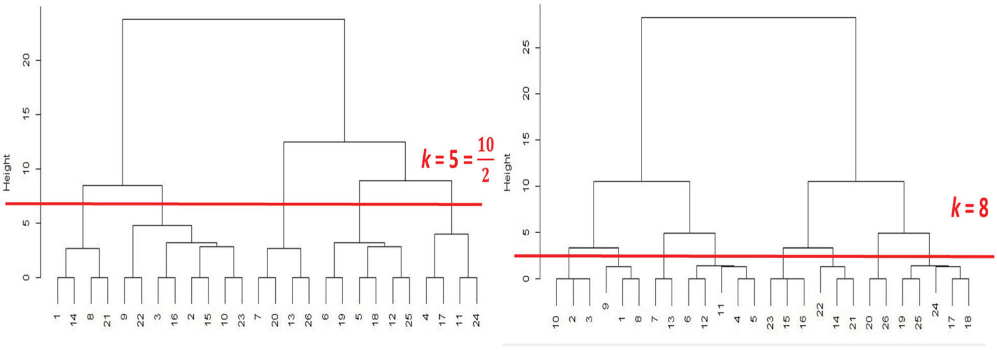
Sample								Sample result			
Smart Card	TP 1	TP 2	TP 3	TP 4	TP 5	TP 6	TP 7	Cross correlation		Dynamic time warping	
								max lag = 1	max lag = 2	window = 1	window = 2
1	1	0	0	0	0	0	0	1+0	1+0	1	1
2	0	1	0	0	0	0	0	2+0	2+0	2	2
3	0	0	1	0	0	0	0	2+1	2+1	2	2
4	0	0	0	1	0	0	0	3+0	3+0	3	2
5	0	0	0	0	1	0	0	5+0	5+0	3	3
6	0	0	0	0	0	1	0	5+1	5+1	3	3
7	0	0	0	0	0	0	1	4+1	4+1	4	4
8	1	1	0	0	0	0	0	1+1	1+1	1	1
9	1	0	1	0	0	0	0	2+1	1+2	1	1
10	0	1	1	0	0	0	0	2+0	2+0	2	2
11	1	0	0	0	1	0	0	3+1	3+1	3	1
12	0	0	0	0	1	1	0	5+0	5+0	3	3
13	0	0	0	0	0	1	1	4+0	4+0	4	4
14	0	1	1	1	1	1	1	1-0	1-0	5	5
15	1	0	1	1	1	1	1	2-0	2-0	6	6
16	1	1	0	1	1	1	1	2-1	2-1	6	6
17	1	1	1	0	1	1	1	3-0	3-0	7	6
18	1	1	1	1	0	1	1	5-0	5-0	7	7
19	1	1	1	1	1	0	1	5-1	5-1	7	7
20	1	1	1	1	1	1	0	4-1	4-1	8	8
21	0	0	1	1	1	1	1	1-1	1-1	5	5
22	0	1	0	1	1	1	1	2-1	1-2	5	5
23	1	0	0	1	1	1	1	2-0	2-0	6	6
24	0	1	1	1	0	1	1	3-1	3-1	7	5
25	1	1	1	1	0	0	1	5-0	5-0	7	7
26	1	1	1	1	1	0	0	4-0	4-0	8	8

Note: Left half: Sample 0–1 sample data (26 smart cards' data for 7 time periods TP*i*). Right half: Sample result. The number of groups of CCD (calibrated by 'lag') and DTW (calibrated by 'window').

Figure 5(a) is the dendrogram of hierarchical clustering resulting from CCD, in which lag is 2. All the series are separated into five clusters as shown in Table 1 (column 'lag = 2'). Figure 5(b) is the dendrogram of hierarchical clustering resulting from DTW distance, in which the parameter window is 2. All the series are cut into six clusters, as shown in Table 1 (column 'window = 2').

In Table 1, for CCD, the lag varies from 1 to 2, and for DTW, the window varies from 1 to 2. The calibration test helps the sensitivity of each parameter and metric to be understood. Each result consists of a first number, a plus or minus sign, and a second number. The first number indicates the cluster affectation, the sign means whether the correlation coefficient is positive or negative, and the second number is the lag (of this time series compared to lag 0 in the same group).

For example, for the 22nd smart card in the column 'lag = 2' of Table 1, the result of CCD with lag of 2 is cluster 1, with a negative correlation coefficient. Thus, it is presented as '1



(a)

(b)

Figure 5. Hierarchical clustering dendrogram (a) with CCD (max lag = 2) and (b) with DTW (window = 2).

(cluster)–(negative) 2 (lag)'. From this pedagogical sample containing 26 smart card users' behavior in a time series, four types of results have been obtained by using CCD, DTW, and each parameter.

(1) By using CCD, if the max lag is configured as '1', the 22nd smart card user's daily transaction series will be grouped in cluster 2. Compared to the other smart card user's daily transaction series in the same group, this behavior has a negative correlation coefficient, and a lag of 1 time period.
(2) By using CCD, if the max lag is configured as '2', the 22nd smart card user's daily transaction series will be grouped in cluster 1. Compared to the other smart card user's daily transaction series in the same group, this behavior has a negative correlation coefficient, and a lag of 2 time periods.
(3) By using DTW, if the window is configured as '1', the 22nd smart card user's daily transaction series will be grouped in cluster 5.
(4) By using DTW, if the window is configured as '2', the 22nd smart card user's daily transaction series will also be grouped in cluster 5.

To explain the lag parameter in depth, the second and third user's daily profile can be compared. They are in the same group and the lag of the second time series is 0, while the third one is 1. That is to say that the second time period of the second time series can be shifted by 1 time unit, to align to the third time period of the third time series. Then, the third time series can be accepted in the same group as the second group.

4.2. Comparison between cross-correlation and time warping

The classification result is impacted by the selected metrics and parameters: it is of interest to discover the difference with the influence of each. Based on Table 1, given a method (CCD or DTW), the size of a given cluster can be known, and the intersection size of two methods or parameters can also be known. For example, for the smart card data series in the cluster 1+ of CCD, there are two smart card data series that correspond in cluster 1 of DTW (with window = 1). In Table 1, these two smart card data series are series 1 and 8. Therefore, in Table 2(a), the horizontal axis is the size of each group of DTW. The parameter window is 1. The vertical axis is the size of each group by CCD. The parameter lag is also 1. Based on the same logic, Table 2(b–d) is built, for the comparison of other metrics and parameter values.

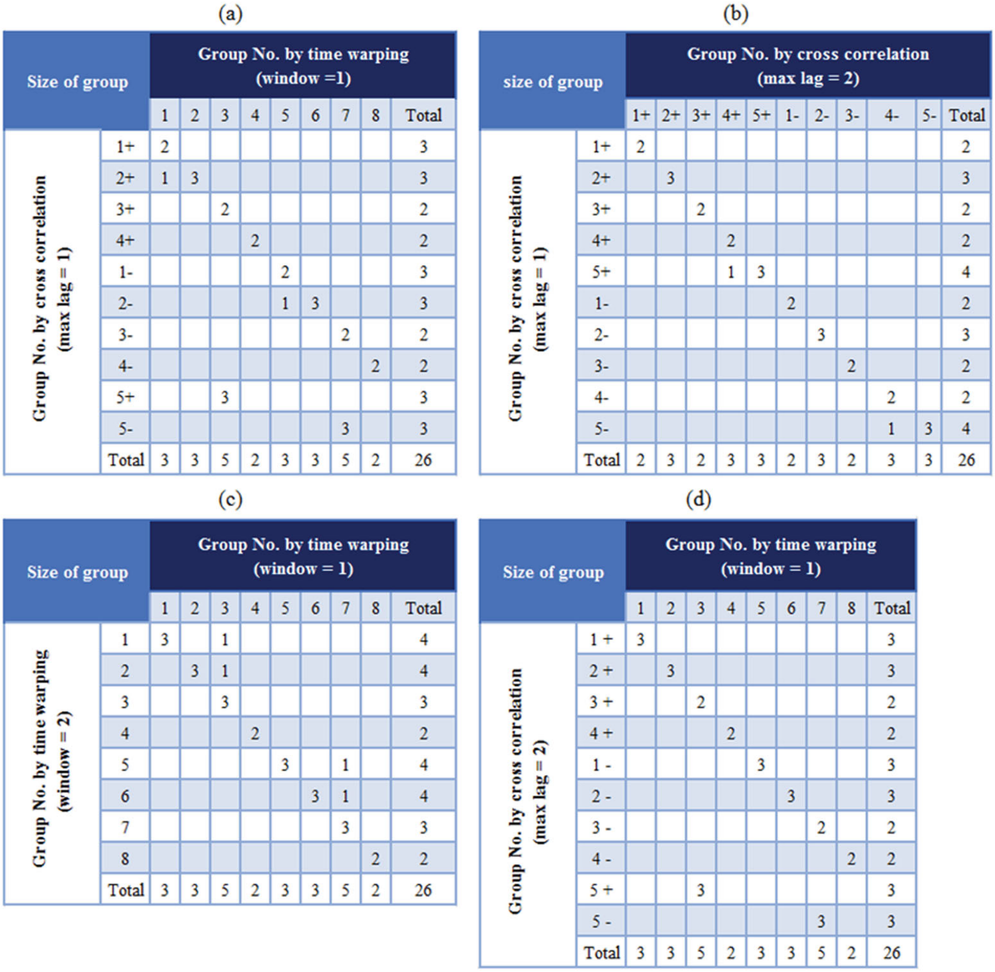
(1) Table 2(a) shows the comparison between CCD (max lag = 1) and DTW (window = 1). Almost the same result can be obtained by using these two methods. Except for group 3 and group 7 of DTW, these groups can be divided into two groups when using CCD. Groups 1 and 5 of DTW have minor differences with CCD. Moreover, the group sizes that are given by DTW contain large numbers; CCD divides these groups into smaller numbers, so that the group sizes are more uniform.
(2) Table 2(b) shows the comparison between CCD (max lag = 1 and 2). The results are almost the same. Moreover, the augmentation of max lag can lead to a more comparable size. This means that even though augmentation of the max lag will more significantly shift the time series, the best correlation coefficient should be matched when the max lag is equal to 1. Therefore, the max lag of 2 not only maintains satisfactory results that do not need to significantly shift a time series, but it also makes the group size similar. In conclusion, the result of a bigger lag has a minor change compared to the result of the smaller lag.
(3) Table 2(c) shows the comparison between DTW (window = 1 and 2). Unlike the comparison between CCD, almost all the groups have been changed if the parameter window has been changed. This means that DTW is more sensitive when changing parameters. Even though a substantial change in this case can make a 50% change in the group (four groups out of eight have been changed in Table 3).
(4) Based on (1), (2), and (3), the results from CCD (max lag = 2) and DTW (window = 1) are better. It is of interest to compare the results via these two conditions, as shown in Table 2(d). This demonstrates that if the parameter is carefully chosen for each of the methods, the results through the two methods will be nearly the same. However, CCD is easier to calibrate because it is less sensitive to parameter values.

4.3. Comparison results

After comparing the application of the CCD and DTW in 0–1 sample data, the CCD is determined to be better in the classification of smart card data for the following reasons:

(1) CCD is easier to calibrate. As presented in Table 2(b), when using CCD, the choice of parameters (lag) has a minor impact on the result, and almost the same result can be obtained when using 1 or 2 as the lag. However, as presented in Table 2(c), the choice of parameters (window) of DTW has a larger impact on the result than CCD.
(2) The result of CCD contains information on the correlation coefficient and lag. As presented from Table 2(a) to Table 2(d), besides the parameter 'lag', for each result of CCD, there is another factor, 'correlation coefficient' (positive or negative). This means the number of groups can be adjusted depending on our need. For example, group 1+ and 1− can be combined into group 1 if fewer groups are needed. However, the DTW has only one choice in the group number.

Table 2. Comparison between metrics and parameters.



(a)

Size of group		Group No. by time warping (window =1)								
		1	2	3	4	5	6	7	8	Total
Group No. by cross correlation (max lag = 1)	1+	2								3
	2+	1	3							3
	3+			2						2
	4+				2					2
	1-					2				3
	2-					1	3			3
	3-							2		2
	4-								2	2
	5+			3						3
	5-							3		3
	Total	3	3	5	2	3	3	5	2	26

(b)

size of group		Group No. by cross correlation (max lag = 2)										
		1+	2+	3+	4+	5+	1-	2-	3-	4-	5-	Total
Group No. by cross correlation (max lag = 1)	1+	2										2
	2+		3									3
	3+			2								2
	4+				2							2
	5+				1	3						4
	1-						2					2
	2-							3				3
	3-								2			2
	4-									2		2
	5-									1	3	4
	Total	2	3	2	3	3	2	3	2	3	3	26

(c)

Size of group		Group No. by time warping (window = 1)								
		1	2	3	4	5	6	7	8	Total
Group No. by time warping (window = 2)	1	3		1						4
	2		3	1						4
	3			3						3
	4				2					2
	5					3		1		4
	6						3	1		4
	7							3		3
	8								2	2
	Total	3	3	5	2	3	3	5	2	26

(d)

Size of group		Group No. by time warping (window = 1)								
		1	2	3	4	5	6	7	8	Total
Group No. by cross correlation (max lag = 2)	1 +	3								3
	2 +		3							3
	3 +			2						2
	4 +				2					2
	1 -					3				3
	2 -						3			3
	3 -							2		2
	4 -								2	2
	5 +			3						3
	5 -							3		3
	Total	3	3	5	2	3	3	5	2	26

(a) Comparison of CCD (max lag = 1) and DTW (window = 1).
(b) Comparison of CCD (max lag = 1 and max lag = 2).
(c) Comparison of DTW (window = 1 and window = 2).
(d) Comparison between CCD (max lag = 2) and DTW (window = 1).
Note: 1+ represents the group number 1 of CCD with positive correlation coefficient.

Table 3. Excerpts from the raw smart card dataset (He, Trépanier, and Agard 2017).

Card ID	Ticket type	Date	Time	Line	Direction	Weekday	Stop id
1150629967111800	140	03-09-2013	65232	44	Sud (South)	2	1140
1273590714804090	110	02-09-2014	71909	224	Sud (South)	1	2801
1273590714804090	110	02-09-2014	154607	224	Nord (North)	1	2610

(3) Group size is more similar and better defined when using CCD. Table 2(a) illustrates the group size of each method. For CCD, all the group sizes are 2 or 3, compared to the size of 2 to 5 for DTW. The grouping of CCD is more even than DTW in the case of temporal transaction profiles.

Table 4. Example dataset of users-day (0-1 table).

Combination	05_30	06_00	06_10	06_20	06_30	06_40	...
1150296033731200_2013-09-04	0	0	0	1	0	0	...
1150312817303160_2013-09-03	1	0	0	0	0	0	...
1150320729466490_2013-09-03	0	0	0	0	0	0	...

5. Application to real public transit smart card data

In this section, real smart card data are used and classified with the method proposed in section 3. The results are presented and analyzed to show how this method performs.

5.1. Presentation of the case study

This dataset has been provided by the *Société de Transport de l'Outaouais* (STO), a transit authority serving 280,000 inhabitants in Gatineau, Quebec. The STO authority is a Canadian leader in public transit using smart card fare collection (Morency, Trepanier, and Agard 2007). Table 3 shows an excerpt of the raw smart card dataset; it contains a variable of a user's trip information. Every line of Table 3 is obtained automatically once a transaction is made by a smart card user. Apart from the card identification (which has been made anonymous), there is the ticket type (fare categories such as junior, regular, and senior), the date and the time of the transaction, the line (route) number and the direction. All transactions are made on a bus network; the location of the transaction is also available (He, Trépanier, and Agard 2017). In this experiment, 100,000 transactions from 3095 cardholders have been tested.

The objective is to make clusters of smart card data that demonstrate similar daily behaviors. In each group, the boarding time of day for a user should be similar to that of another user in the same group: this might not mean exactly at the same time, but 'around' the same time period.

5.2. Results

First, the data from Table 3 are transformed into a 0–1 table (see Table 4), in which every line is a user's daily profile ('card id_date' combination), and every column is a time period; for example, the second column '05_30' means the period from 05:30 to 05:59. In the table, '1' represents that a transaction happened in this time period. For example, for the user whose card id is 1150312817303160, in 2013-09-03, he had a transaction in the time period 05:30–05:59.

With Table 4, the distance of every two combinations of lines is calculated by using the CCD and DTW. For the CCD, the parameter 'lag' is 2. For the DTW, the parameter 'window' is 2. Then, the CCD and DTW are computed and a distance matrix of any two combinations is obtained for each method. With that distance matrix, the combination ('user-date') is computed using hierarchical clustering.

By observing the dendrogram, we cut the dendrogram depending on the circumstances, to obtain the subgroups with a more even size. Then, 11 groups are cut for CCD and 6 groups for DTW. Finally, the sum of transactions for each cluster is calculated; this result is shown in Figure 6 (by using CCD) and Figure 7 (by using DTW).

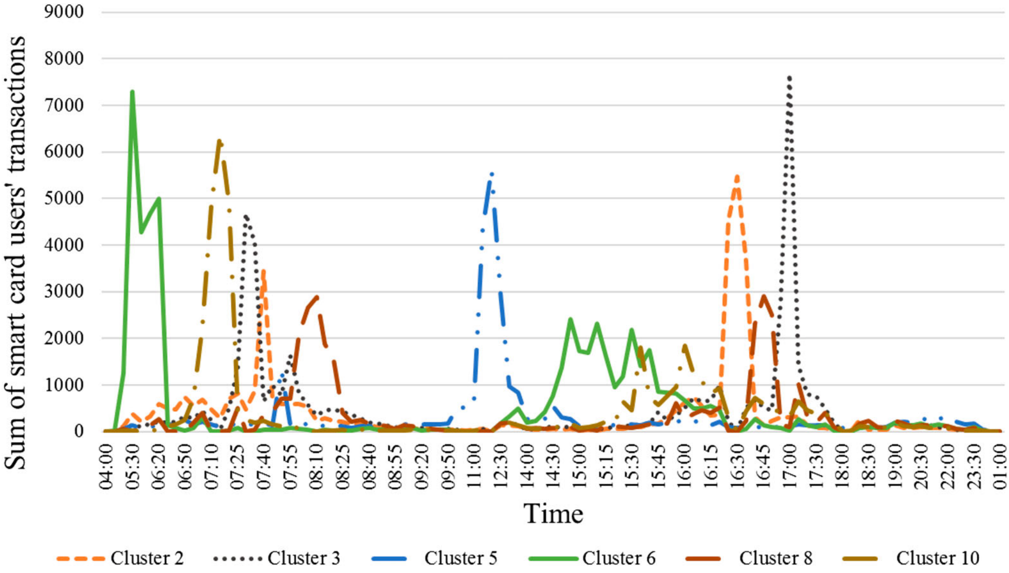


Figure 6. Sum of transaction time of each group (CCD).

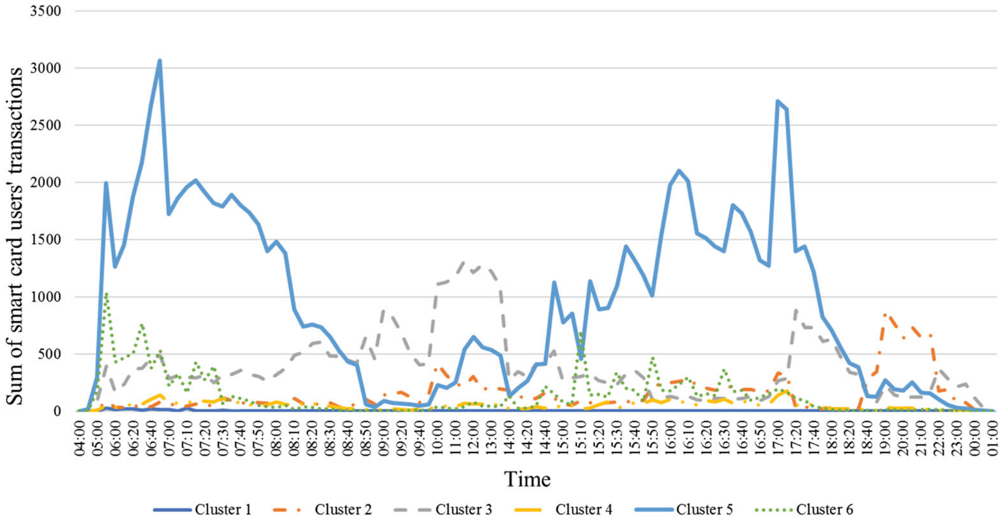


Figure 7. Sum of transaction time of each group (DTW).

By comparing Figures 6 and 7, the DTW is not effective in our case. Firstly, the size of cluster 5 is so large that cluster 5 contains most of the transaction profiles, which can lead to an uneven size between all of the clusters. Secondly, comparing cluster 5 and cluster 6 in Figure 7, even though the size is different, the 'peak hours' of these two clusters are almost the same. This means the users who have different behaviors cannot be separated by using DTW. Two criteria are used to judge which distance metric is better: exclusivity and homogeneity.

(1) Exclusivity:

This makes it possible to distinguish the different behaviors. It means that if one group occurs in one period, the other group will not occur in the same time period. In Figure 6, for CCD, the first smart card user transaction in Group 6 is between 05:30 and 06:20. In this period, the sum of transactions of group 6 always exceeds 4000. On the contrary, the sum of any other group is less than 600. The situation of the other groups is similar. In Figure 7, for the DTW, group 5 and group 6 appear in the period between 05:00 and 07:30, whose duration is 2.5 hours. This means that the DTW cannot separate the smart card users' behaviors very well. Therefore, with regard to the exclusivity criterion, it is preferable to classify user behaviors using CCD.

(2) Homogeneity:

When classifying, we try to obtain groups that are as homogeneous as possible. This means that the group size (the amount of user profiles of the smart card in the group) should be roughly uniform. In Figure 6, for CCD, the maximum sum of all groups is 2500–8000, and there is no group whose size represents more than 50% of the number of profiles of smart card users. However, in Figure 7, for DTW, group 5 represents more than 50% of all profiles of smart card users, which means that it is an unequal classification. This does not necessarily mean that users are distributed evenly among all groups. However, CCD provides more uniform size groups compared to DTW; therefore, in this case it is of greater interest, even though it is not absolutely necessary all of the time.

Finally, it is of great interest to discuss how the proposed method and classification results can be used to improve transit features. In general, the results will help transit authorities to offer better services for smart card users from diverse groups. Firstly, based on the behavior of different groups, the transit authority is able to optimize schedules to satisfy the demand of groups and save vehicle turns. For example, based on Figure 6, for the period 04:00–07:25, more vehicle turns may be scheduled between 04:00–06:20 and 06:50–07:25 to respond the demand of groups 6 and 10, respectively. Drivers could take a break during 06:20–06:50. Furthermore, if this optimization results in fewer total vehicle turns, it could help save energy and reduce exhaust that creates greenhouse gases. Secondly, even though activity episodes are not encoded, we know the boarding time of every group. This means that if a user's daily profile contains only two transactions, we can infer a user's home and work or place of study. In fact, in Figure 6, groups 2, 3, and 8 contain and only contain two peaks, which means that most of the users in these groups contain only two transactions in that day. Thirdly, taking advantage of off-peak periods, vehicles can be allocated to relieve burdens on other bus lines. For example, a vehicle that serves group 8 may finish the service at 09:00, then it can be allocated to serve group 5 at 11:00. In conclusion, the main idea is to look for different characteristics among the groups, then serve them differently.

6. Conclusion

An analysis of public transit smart card users' daily profiles needs a method that enables time series to be classified. Because of the limitations of traditional distance metrics, a method has been designed by combining a time-series metric and hierarchical clustering.

The results show that cross-correlation is better adapted for the classification of a public transit smart card data temporal profile. The test, using data from a mid-sized public transit association, shows a clear separation of card users' daily transaction profiles. This may bring forth better information about each subgroup of users and then improvements to the transit system could be made.

Regarding limitations and perspectives, the first limitation is the calculation time when trying different parameters (lag for cross-correlation and windows for DTW). To solve this problem, a new algorithm could be developed in order to avoid certain calculations in which the calculation of the CCD between certain vectors is canceled out by assuming that the distance between these two vectors is too large. The second limit is the choice of metrics when dealing with transportation issues. In this case, the CCD is suitable because the delay of a smart card user's transaction time is like the parameter 'lag' in the CCD. However, when dealing with other time series in transportation problems, other distances may be applied. For example, the Fourier transformation distance, based on the analysis of fluctuations, could be tested to explain the fluctuations in transactions, etc. Overall, the objective is still to find the best metric for a specific transportation issue.

Acknowledgements

The authors wish to acknowledge the support of the Société de transport de l'Outaouais (STO) for providing data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors wish to acknowledge the Thales Group and the Natural Sciences and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

References

Agard, B., C. Morency, and M. Trépanier. 2006. "Mining Public Transport User Behavior from Smart Card Data." *IFAC Proceedings Volumes* 39 (3): 399–404.

Bakar, Z. A., R. Mohemad, A. Ahmad, and M. M. Deris. 2006, June. "A Comparative Study for Outlier Detection Techniques in Data Mining." In *2006 IEEE conference on Cybernetics and intelligent systems*, 1–6. IEEE.

Berkhin, P. 2006. "A Survey of Clustering Data Mining Techniques." In *Grouping Multidimensional Data*, 25–71. Berlin: Springer.

Berndt, D. J., and J. Clifford. 1994, July. "Using dynamic time warping to find patterns in time series." *KDD workshop* 10 (16): 359–370.

Bordagaray, M., L. dell'Olio, A. Ibeas, and P. Cecín. 2014. "Modelling User Perception of Bus Transit Quality Considering User and Service Heterogeneity." *Transportmetrica A: Transport Science* 10 (8): 705–721.

Briand, A. S., E. Côme, M. Trépanier, and L. Oukhellou. 2017. "Analyzing Year-to-Year Changes in Public Transport Passenger Behavior Using Smart Card Data." *Transportation Research Part C: Emerging Technologies* 79: 274–289.

Brockwell, P. J., and R. A. Davis. 2002. *Introduction to Time Series and Forecasting*. 2nd ed. New York: Springer.

Chang, H., D. Park, S. Lee, H. Lee, and S. Baek. 2010. "Dynamic Multi-Interval Bus Travel Time Prediction Using Bus Transit Data." *Transportmetrica* 6 (1): 19–38.

Chen, C. F., Y. H. Chang, and Y. W. Chang. 2009. "Seasonal ARIMA Forecasting of Inbound Air Travel Arrivals to Taiwan." *Transportmetrica* 5 (2): 125–140.

Chu, K. A., and R. Chapleau. 2008. "Enriching Archived Smart Card Transaction Data for Transit Demand Modeling." *Transportation Research Record: Journal of the Transportation Research Board* 2063: 63–72.

Das, S., and D. Pandit. 2015. "Determination of Level-of-Service Scale Values for Quantitative bus Transit Service Attributes Based on User Perception." *Transportmetrica A: Transport Science* 11 (1): 1–21.

de Oña, R., and J. de Oña. 2015. "Analysis of Transit Quality of Service Through Segmentation and Classification Tree Techniques." *Transportmetrica A: Transport Science* 11 (5): 365–387.

Del Castillo, J. M., and F. G. Benitez. 2013. "Determining a Public Transport Satisfaction Index from User Surveys." *Transportmetrica A: Transport Science* 9 (8): 713–741.

Deza, M. M., and E. Deza. 2009. *Encyclopedia of Distances*. Berlin: Springer-Verlag.

Esling, P., and C. Agon. 2012. "Time-Series Data Mining." *ACM Computing Surveys (CSUR)* 45 (1): 12.

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

Ghaemi, M. S., B. Agard, M. Trépanier, and V. Partovi Nia. 2017. "A Visual Segmentation Method for Temporal Smart Card Data." *Transportmetrica A: Transport Science* 13 (5): 381–404.

Giorgino, T. 2009. "Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package." *Journal of Statistical Software* 31 (7): 1–24.

Han, G., and K. Sohn. 2016. "Activity Imputation for Trip-Chains Elicited from Smart-Card Data using a Continuous Hidden Markov Model." *Transportation Research Part B: Methodological* 83: 121–135.

He, L., N. Nassir, M. Trépanier, and M. Hickman. 2015. "Validating and Calibrating a Destination Estimation Algorithm for Public Transport Smart Card Fare Collection Systems." *Centre Interuniversitaire de Recherche sur les Reseaux d'Entreprise, la Logistique et le Transport (CIRRELT*), Montreal, QC, Canada, Tech. Rep.

He, L., and M. Trépanier. 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., M. Trépanier, and B. Agard. 2017. "*Evaluating the Impacts of a Bus-Rapid Transit on Users' Temporal Patterns Using Cross Correlation Distance and Sampled Hierarchical Clustering Applied to Smart Card Data* (No. 17-03711)".

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. "Data Clustering: A Review." *ACM Computing Surveys (CSUR)* 31 (3): 264–323.

Joh, C. H., H. J. P. Timmermans, and T. A. Arentze. 2006. "Measuring and Predicting Adaptation Behavior in Multidimensional Activity-Travel Patterns." *Transportmetrica* 2 (2): 153–173.

Jou, R. C., S. H. Lam, and P. H. Wu. 2007. "Acceptance Tendencies and Commuters' Behavior Under Different Road Pricing Schemes." *Transportmetrica* 3 (3): 213–230.

Kieu, L. M., A. Bhaskar, and E. Chung. 2014. "Transit passenger segmentation using travel regularity mined from Smart Card transactions data." Transportation Research Board 93rd annual meeting, Washington, DC, January 12–16.

Kurauchi, F., and J. D. Schmöcker, eds. 2017. *Public Transport Planning with Smart Card Data*. Boca Raton: CRC Press.

Kusakabe, T., and Y. Asakura. 2014. "Behavioural Data Mining of Transit Smart Card Data: A Data Fusion Approach." *Transportation Research Part C: Emerging Technologies* 46: 179–191.

Langfelder, P., B. Zhang, and S. Horvath. 2008. "Defining Clusters From a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R." *Bioinformatics (Oxford, England)* 24 (5): 719–720.

Langlois, G. G., H. N. Koutsopoulos, and J. Zhao. 2016. "Inferring Patterns in the Multi-Week Activity Sequences of Public Transport Users." *Transportation Research Part C: Emerging Technologies* 64: 1–16.

Lee, S. G., and M. Hickman. 2014. "Trip Purpose Inference Using Automated Fare Collection Data." *Public Transport* 6 (1-2): 1–20.

Lhermitte, S., J. Verbesselt, W. W. Verstraeten, and P. Coppin. 2011. “A Comparison of Time Series Similarity Measures for Classification and Change Detection of Ecosystem Dynamics.” *Remote Sensing of Environment* 115 (12): 3129–3152.

Li, Y. T., J. D. Schmöcker, and S. Fujii. 2015. “Demand Adaptation Towards New Transport Modes: The Case of High-Speed Rail in Taiwan.” *Transportmetrica B: Transport Dynamics* 3 (1): 27–43.

Liao, T. W. 2005. “Clustering of Time Series Data – A Survey.” *Pattern Recognition* 38 (11): 1857–1874.

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

Meyer, D., and C. Buchta. 2015. “Proxy: Distance and Similarity Measures.” R package version 0.4-15.

Morency, C., M. Trepanier, and B. Agard. 2007. “Measuring Transit Use Variability with Smart-Card Data.” *Transport Policy* 14 (3): 193–203.

Mori, U., A. Mendiburu, and J. A. Lozano. 2016. “Distance Measures for Time Series in R: The TSdist Package.” *R Journal* 8 (2): 451–459.

Nishiuchi, H., J. King, and T. Todoroki. 2013. “Spatial-temporal Daily Frequent Trip Pattern of Public Transport Passengers Using Smart Card Data.” *International Journal of Intelligent Transportation Systems Research* 11 (1): 1–10.

Nuzzolo, A., and A. Comi. 2016. “Advanced Public Transport and Intelligent Transport Systems: New Modelling Challenges.” *Transportmetrica A: Transport Science* 12 (8): 674–699.

Pelletier, M. P., M. Trépanier, and C. Morency. 2011. “Smart Card Data Use in Public Transit: A Literature Review.” *Transportation Research Part C: Emerging Technologies* 19 (4): 557–568.

Rokach, L., and O. Maimon. 2005. “Clustering Methods.” In *Data Mining and Knowledge Discovery Handbook*, edited by O. Maimon and L. Rokach, 321–352. Boston, MA: Springer.

Sedgwick, P. 2012. “Pearson’s Correlation Coefficient.” *BMJ* 345 (7): e4483.

Subbiah, K. 2011. *Partitioning Methods in Data Mining*. http://www.authorstream.com/Presentation/msusuresh-1133119-partitioning-methods/. Tirunelveli.

Trépanier, M., N. Tranchant, and R. Chapleau. 2007. “Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System.” *Journal of Intelligent Transportation Systems* 11 (1): 1–14.