CIRRELT

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

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Sampling Method Applied to the Clustering of Temporal Patterns of Public Transit Smart Card Users

Li He
Martin Trépanier
Bruno Agard

July 2019

CIRRELT-2019-30

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

UNIVERSITÉ LAVAL McGill Université Concordia University ÉTS UQÀM Université du Québec à Montréal HEC MONTRÉAL POLYTECHNIQUE MONTRÉAL Université de Montréal

Sampling Method Applied to the Clustering of Temporal Patterns of Public Transit Smart Card Users

Li He, Martin Trépanier*, Bruno Agard,

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Mathematics and Industrial Enginnering, Polytechnique Montréal, 2500, chemin de Polytechnique, Montréal, Canada H3T 1J4

Abstract. The study of temporal patterns has been applied to represent the various behaviours of transport users. Smart card data is useful for characterizing travel behaviours. The behaviours identified can be analysed and thereby transport services can be improved. For large datasets of transactions, the traditional method is to segment the data into several groups and use one unique pattern to represent each group. However, classifying very large datasets is still challenging. Here we propose a method to classify the temporal patterns of all the users of a public transit system. We recommend a clustering method that combines a sampling method and cross-correlation distances. This method was applied to classify the temporal patterns of public transport users from Gatineau, Canada. An indicator was developed to validate the efficiency of the proposed method. Compared to current methods, the proposed method is faster and better able to deal with very large datasets.

Keywords. Public transit big data, smart card user behaviour, time series classification, sampling method

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

Results and views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect those of CIRRELT.

Les résultats et opinions contenus dans cette publication ne reflètent pas nécessairement la position du CIRRELT et n'engagent pas sa responsabilité.

* Corresponding author: martin.trepanier@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2019

1 INTRODUCTION

Extracting temporal behaviours from large time series datasets has proven to be valuable for many applications in the domain of transport. In fact, large sets of transport data are analysed in order to determine the temporal behaviours that are used in numerous application domains. Analysing temporal behaviours allows public transit patterns to be understood (Liu & Cheng, 2018), user behaviour to be predicted (Yang et al., 2018) and crowding levels in public transit vehicles to be measured (Yap et al., 2018) so that transit authorities can intervene and improve levels of service.

Using data mining techniques on transport data helps us to better understand user behaviour (Mohamed et al., 2017). Data mining techniques make it possible to measure the travel time of different user groups and different fare types (Ma et al., 2013). It helps estimate the origin-destination demand by integrating pattern matching methods (Chen et al., 2015), and travel trajectory can be analysed in order to understand the behaviour of different traveler groups (Zheng, 2015). Data mining techniques help deduce/predict the purpose of trips (Kusakabe and Asakura, 2014) and they can also uncover traffic bottlenecks in the urban network through spatiotemporal analysis (Lee et al., 2011). Moreover, data mining can also help analyse the impacts of weather on public transport ridership (Zhou et al., 2017).

Data mining can be used with diverse sources of data. For example, when used with bike sharing data, the spatiotemporal behaviour of bike rentals can be understood (Bordagaray et al., 2016). A similar method can also be implemented in carshare systems (Morency et al., 2007). Data mining can also be applied to GPS data to understand freight trip chaining behaviour (Ma et al., 2016). Finally, smart cards is relevant to the various uses of data mining and its have long been proven to be effective and useful for analysing the mobility behaviour of transit riders (Pelletier et al., 2011).

The temporal pattern analysis of public transit users is challenging when dealing with large datasets. When using large datasets, many authors (Agard et al., 2006; Ghaemi et al., 2015; Ghaemi et al., 2017) propose classifying data into groups (user behaviours, for example) in order to deal with a limited, but still representative, set of generic behaviours that can be analysed. This practice makes it possible to simplify the data in a useful way and reveal the hidden patterns in a huge volume of individual data. The level of classification (number of groups) gives analysts the control to either get more or less detail on certain information, depending on the accuracy needed. However, the classification of large sets of temporal data (time series) is still a challenge and needs further development.

In this paper, we propose a new method for classifying temporal patterns, using cross-correlation distances and sampled hierarchical clustering. The method is used to analyse a large dataset of smart card transactions from a public transport service company and to classify the behavioural patterns of all the smart card users in the public transit system.

The paper is organized as follows. The following section presents the literature review. Section 3 describes the proposed classification method by first explaining the classical method, which combines cross-correlation distance and hierarchical clustering. Then, the framework for method we developed is described by combining the classical method, a sampling method and an assignment process. In section 4, the proposed framework is applied to a real case study. Furthermore, an indicator is designed to check which sample sizes are large enough. Finally, section 5 concludes the paper and introduces future research perspectives.

2 LITERATURE REVIEW

This section describes works relevant to the proposed method. Section 2.1 focuses on traditional classification methods. Section 2.2 points out the notion of distance. Two distance metrics, in the context of the analysis of smart card data, are highlighted. Section 2.3 outlines classification methods and distances in smart card data research.

2.1 Traditional Classification Methods

Data classification aims to solve the following problem: Given a set of training points with the associated training labels, determine the class label for a test instance that is also not labeled (Aggarwal, 2014). In other words, these methods aim to group a set of observations into clusters. There are many clustering approaches. The principle of intra-class distance (dissimilarity) is to maximize the similarities between objects in the same class. Inter-class distances aim to minimize the similarity between objects of different classes (Lakshmi & Raghunandhan, 2011). Some classification methods are presented as follows:

2.1.1 Partitioning Algorithms

Partitioning methods try to find the best partitions (k) from a given number of objects (n) (Ng & Han, 1994). The following introduces two popular partitioning algorithms:

K-means: K-means clustering (Lee & Hickman, 2014; MacQueen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting the initial cluster centers of the k groups and then iteratively refines them as follows: (1) Each instance (di) is assigned to its closest cluster center. (2) Each cluster center (Cj) is then updated as the mean of the instances (di) that were assigned to it (Wagstaff et al., 2001). This is one of the classification methods that best addresses the well-known problems in data clustering.

K-medoids: Each cluster is represented by one of the objects in that cluster (Park & Jun, 2009). The K-means algorithm is sensitive to outliers, such as inputs with extremely large values that may substantially distort the distribution of data. To solve this issue the K-medoids method uses a medoid instead of using the mean value of objects in a cluster as the reference point. A medoid is the most centrally located object in a cluster (Velmurugan & Santhanam, 2010).

2.1.2 Hierarchical Algorithms

Hierarchical algorithms create a hierarchical decomposition of the set of data (or objects) using some hierarchical criteria (Karypis et al., 1999).

Hierarchical clustering organizes objects into a dendrogram, which branches out into the desired clusters. In a dendrogram, pairs of vertices that are more closely related have common ancestors that are located lower in the tree. Those of more distantly related pairs are located higher in the tree (Clauset et al., 2008). The process of cluster detection is referred to as tree cutting, branch cutting, or branch pruning (Langfelder et al., 2007).

Hierarchical algorithms fall into two types. One is based on repeatedly merging two smaller clusters into a larger one (agglomerative, or bottom-up). The other is based on splitting a larger cluster into smaller ones (divisive, or top-down) (Ding & He, 2002). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters (Rokach et al., 2005).

There are two basic advantages for hierarchical algorithms. First, the number of clusters does not need to be specified a priori. Second, they are independent of the initial conditions (not necessary to initial cluster centers as partitioning algorithms) (Frigui and Krishnapuram, 1999). However, the disadvantage is that data-points may fail to separate overlapping clusters due to a lack of information about the global/overall shape or size of the clusters (Jain et al., 1999).

Furthermore, there are some derivative methods such as CURE (clustering using representatives) and BIRCH (balanced iterative reducing and clustering using hierarchies), among others. CURE employs a hierarchical clustering algorithm that adopts a middle ground between the centroid and all point extremes, in order to avoid problems relating to clusters that are not uniformly shaped or sized (Guha et al., 1998). For BIRCH, the advantage is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points (Zhang et al., 1996). In conclusion, hierarchical algorithms are a method of cluster analysis which seeks to build a hierarchy of clusters.

2.1.3 Density-Based

Density-based methods are based on connectivity and density functions (DBSCAN (density-based spatial clustering of applications with noise), OPTICS (ordering points to identify the clustering structure)) (Ester et al., 1996). In density-based clustering, clusters are defined as areas of higher density than the remainder of the dataset (Kriegel et al., 2011).

2.1.4 Other Methods

Other classification methods exist such as grid-based and model-based methods.

The grid-based method is based on a multiple-level granularity structure (Liao et al., 2004). Examples include STING (statistical information grid-based method) and CLIQUE (clustering in quest). For STING, the idea is to capture statistical information associated with spatial cells in a way in which whole classes of queries and clustering problems can be answered without recourse to the individual objects (Wang et al., 1997). CLIQUE identifies the dense units in the subspaces of high-dimensional data space, and uses these subspaces to provide more efficient clustering (Agrawal et al., 1998).

In the model-based classification method, a model is hypothesized for each of the clusters and the idea is to find the best way these models fit together (Yeung et al., 2001). The various classification methods help to choose the best one to solve a specific issue.

Each classification method has its advantages and disadvantages. K-means is a popular method in data mining classification that is applied to various fields (Cui et al., 2017). However, k-means usually uses Euclidean distance calculations. This method rarely uses other types of distances, and it needs a predefined number of clusters (*k*). Hierarchical algorithms can be used without a predefined number of clusters, and it can be used with most of the distance types. The disadvantage of hierarchical algorithms is that it is not suitable for large datasets because of its extended/long computational time.

In conclusion, there are four main factors to consider when choosing a classification method: whether the number of clusters is determined automatically, whether it can be applied to the entire dataset, the computational time and how the results of the algorithm are impacted by the choice of distance metric.

2.2 Distance Calculation Methods

Various distance metrics exist to measure the (dis)similarity between two instances (vectors related to observations). In this section, two types of distance are compared: Euclidean distance and cross correlation distance. Euclidean distance is largely used in various application domains. Cross-correlation distance is better adapted for time series but is much more time consuming, which limits its use for real and larger datasets.

2.2.1 Euclidean Distance

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

$$d_E = \sqrt{\sum_{k=1}^{P} (x_{ik} - y_{jk})^2} \qquad (1)$$

The Euclidean distance is widely used in many application domains. In the case of the transit system in Gatineau (Canada), the very large dataset collects about 600,000 entries each month. Data mining techniques have been used to analyse this data with valuable results (Agard et al., 2006).

The Euclidian distance, however, is not well adapted to time series analyses. According to the function (1), the result of the distance would not change if the order of k is changed. For example, when the values of *k* = 1 (x_{i1}) and *k* = 2 (x_{i2}) are exchanged, the distance will be the same. However, a time series represents the relationship between time (t) observations themselves, which is a characteristic that distinguishes it from other vectors. For a time series that represents the moment a transport system is used during a day, if the values of k = 1 and k = 2 are exchanged, the results of the distances should also change (He et al., 2018).

2.2.2 Cross-Correlation Distance

Cross-correlation distance is based on the correlation between two time series. The similarity between two time series is measured by shifting one time series to find a maximum cross-correlation with another time series. The cross-correlation between two time series at lag *k* is calculated as (Mori et al., 2016):

$$CC_k(X,Y) = \frac{\sum_{i=0}^{N-1-k}(x_i - \bar{x})(y_{i+k} - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_{i+k} - \bar{y})^2}} \qquad (8)$$

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

$$d_{CC}(X,Y) = \sqrt{\frac{(1 - CC_0(X,Y))}{\sum_{k=1}^{max} CC_k(X,Y)}} \qquad (9)$$

In the R software package (https://www.r-project.org), 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 Classification Methods and Distances in Smart Card Data Research

Through the years, several authors have proposed the use of data mining techniques to analyse smart card transaction data (Diab & El-Geneidy, 2013), such as by analysing the characteristics of smart card users (Sun et al., 2016), and the behaviour changes of travelers (Asakura et al., 2012). A previous study showed the variability of the travel behaviour of riders in a bus network using the k-means technique (Morency et al., 2007). In most cases, k-means is used when a mean value for each cluster needs to be computed (Vicente & Reis, 2016), making it difficult to consider smart card data as a time series. More recent works used DBSCAN (Density-based spatial clustering of applications with noise) to assess the mobility behaviour of smart card users (Kieu et al., 2015; Ma et al., 2013). Other research used dynamic time warping (DTW) to unify the time references of smart card transactions (Li & Chen, 2016). However, DTW is not well adapted to large sets of data, due to its computational complexity. Another study used a mix of techniques to analyse data from the city of Rennes, France (El Mahrsi et al., 2014). However, there are still challenges associated with characterizing behaviours based on temporal distribution, because the methods used to calculate the distances between observations are not well suited to the kinds of analyses that are required by transit authorities. To address these challenges, efforts have been made to incorporate spatial data in order to measure the spatial-temporal data dissimilarity (Ghaemi et al., 2015). Furthermore, a classification-related technique has also been used to analyse the quality of transit service (de Oña & de Oña, 2015), to gauge passenger type (Legara & Monterola, 2018) and to analyse transit network performance (Parzani et al., 2017).

In this study, we preferred using cross-correlation distance to Euclidean distance, for the following reasons. First, a time series should not be analysed using? Euclidean distance. Second, transaction times can be represented by a point with the value “1” at the time of the transaction, while the values are all “0” for other times. An earlier or later shift of a transaction time is represented by the shift of value “1”. This shift corresponds to the parameter “lag” seen in cross-correlation distance. Therefore, cross-correlation distance is preferred, as we have shown in previous work (He et al. 2018). Our study also prefers the use of hierarchical clustering to k-means in order to avoid using Euclidean distance between points and cluster centers (Jain, A. K. et al., 2010).

3 PROPOSED METHODOLOGY

This section introduces the method for clustering the temporal behaviour of public transport users with the help of smart card transaction data. The method consists of three elements: the distance metric (cross-correlation distance), the hierarchical clustering algorithm and the sampling method. Figure 1 presents the structure of the proposed method. Please note that steps 6, 7 and 8 are added here to validate the results of the method and therefore are not necessary when the method is implemented.

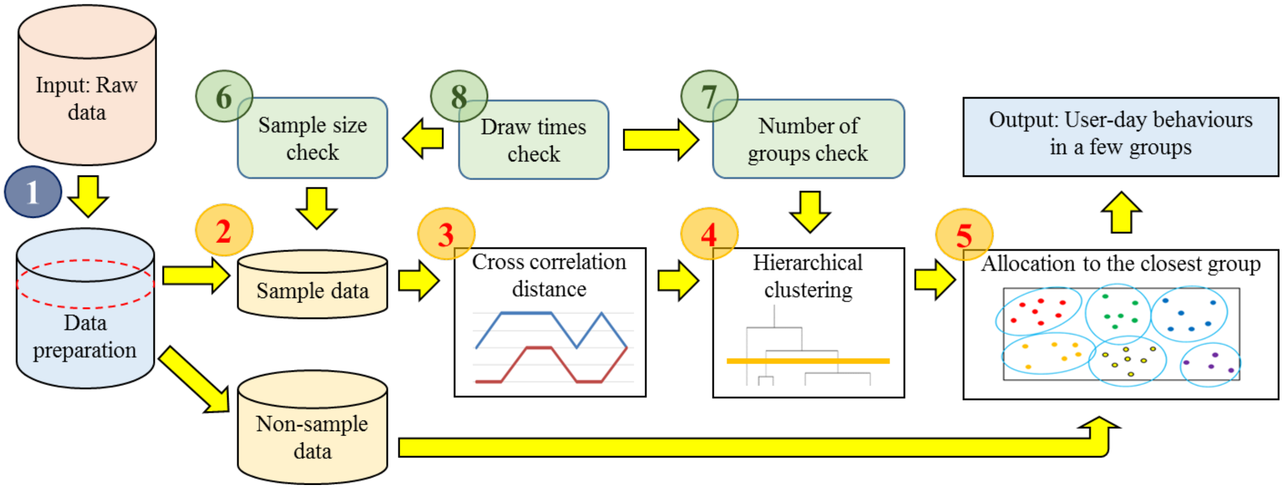


Figure 1. Overall proposed method

The unit of analysis is the temporal daily departure time profile of public transport users. For each user, and for each day, we created a vector of binary values stating whether or not the system was used. These “user-day” vectors were used as raw data. The following explains each of the steps in Figure 1.

1. The smart card transaction data is subjected to a series of pre-treatments in order to create the daily departure time profiles of users.
2. Some user profiles are selected for the sample dataset, while the other user profiles will be in the non-sample dataset.
3. Cross-correlation distance (CCD) is applied to the sample data to measure the dissimilarity between any two daily profiles.
4. A hierarchical algorithm is applied to regroup all the sample data profiles according to their similarities. At the end of this step, a dendrogram is used to arrange/separate the user profiles from the sample dataset into a number of groups.
5. Non-sample data observations are assigned to their nearest group. The data from each user-day is used to compute the distance to the closest cluster. The dissimilarity is also measured by cross-

correlation distance. Finally, a group consisting of the daily profiles of smart card users is obtained as the output in Figure 1.

6. For this paper, we experimented with several sample sizes to conduct a sensitivity analysis.
7. Because a dendrogram is used, the number of groups is also a parameter. We also changed the numbers of groups to test their effects on the results.
8. The contents of the samples were randomly selected. We also tested multiple numbers of draws in order to check the reliability of the results obtained by the sampling method.

Figure 2 presents an overview of the sampling method mentioned in steps 4 and 5.

a. Start with all observations (seen here as points).
b. Sample random points (red points).
c. Apply cross-correlation distance and hierarchical clustering algorithms to these sample points. Clusters are made with this sample.
d. Calculate the distance between each remaining point and all the other points of the sample group. Then, use these distances to allocate the remaining points to the nearest group. In the end, all the points are grouped.

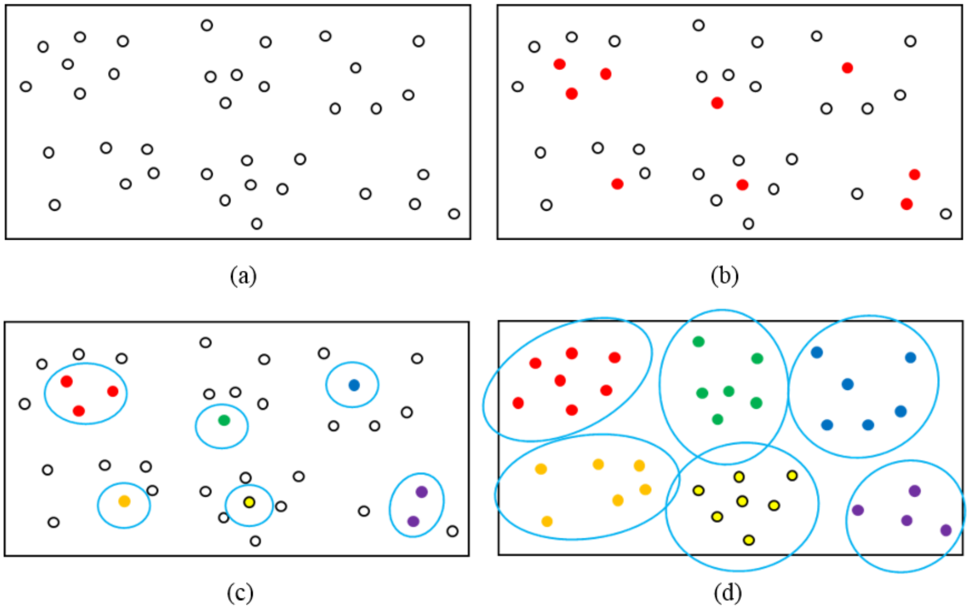


Figure 2. Sampling and allocation process

4 CASE STUDY

For the case study, the dataset was provided by the *Société de Transport de l'Outatouais* (STO), a transit authority serving the 280,000 inhabitants/residents of Gatineau, Quebec. The STO operates a bus-only service with a smart card automated fare collection system (Agard et al., 2006). This system has been in use since 2001, and a large proportion (over 80%) of STO users have a smart card. The following sections describe how the method was implemented.

4.1 Data preparation

This is the first step in Figure 1. The raw smart card transaction dataset is available in a flat file and contains the following information: card identification (anonymized), fare category, date and time of the transaction, location of the transaction (bus stop), route number and direction. Only the data from weekdays were selected in order to better characterize the travel behaviour of regular workers and students. The dataset contains 1,707,192 transactions that were registered in September and November 2013 from 26,320 cards.

4.1.1 Timeframes setting

In order to create temporal profiles based on transaction times, we organized the days that were studied into periods of time. For example, 09: 21 and 09:28 belong to the 09:20 - 09:29 time period, and 09:32 belongs to the 09:30 – 09:39 time period. The periods are of different lengths depending the time of day. From 04:00 to 01:30 (of the next day), there are a total of 35 time periods in each day. Table 1 shows the timeframes (time periods) for one day.

Table 1. Timeframes for the daily distribution of transactions

Time	Type of period	Interval
00:00 - 05:59	Off-peak hours	30
06:00 - 06:59	Regular hours	10
07:00 - 08:59	Peak hours	5
09:00 - 09:59	Regular hours	10
10:00 - 13:59	Off-peak hours	30
14:00 - 14:59	Regular hours	10
15:00 - 15:59	Peak hours	5
16:00 - 17:59	Regular hours	10
18:00 - 23:59	Off-peak hours	30

4.1.2 Vectors of Observations

In this step, we created a table that represents the temporal profile of each card (the transaction times for each card). The STO uses photo-ID cards, so each card represents one user. Each profile is thus called a "user-day". Unique IDs were created by concatenating the card number with the date. For example, in Table 2, the numerical identification displayed on the first line (1150296033731200_2013-09-04) represents the card user's profile followed by the date (04 September 2013). The number 1 that is shown in this row under column 06_20 indicates that a transaction occurred between 06:20 and 06:29 that morning. A total of 337,745 user-day profiles were created. When represented in its entirety, the table contains 337,745 × 35 elements (0 or 1), which is too large to be segmented by hierarchical clustering in a suitable amount of time on regular computers. For this reason, we suggest using a sampling method.

Table 2. Example dataset of user-days (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	…

4.2 Application of the Method

The following three sections present the proposed methods applied to STO data.

4.2.1 Sampling

This is step 2 in Figure 1. In the sampling process where user-days are randomly selected, each user was only present once in the sample. For example, if "user A date A" was chosen for the sample set, then "user A date B" was not added. We tested a range of sample sizes. More details about sample size will be discussed in Sections 4.3 and Section 5.

4.2.2 Sample Clustering

These are steps 3 and 4 in Figure 1. In this process, cross-correlation distances and hierarchical clustering algorithms were applied to the sample data. We then analysed the dendrogram obtained by the hierarchical clustering algorithm, and chose the number of groups to create the sample groups. A total of 25 groups was first tested. The branches cut by red line in Figure 5 present the results. There were still many groups that did not have enough points. Therefore, the number of groups was gradually reduced and the dendrogram was updated. In the end, 11 groups were selected (more detail about selecting the number of groups will be discussed in Part 4.3 and Part 5). Note that in the general method, fewer groups can be chosen. However, because this is a sampling method, we chose to conserve a higher number of clusters for the sample and to link the remaining observations to these clusters.

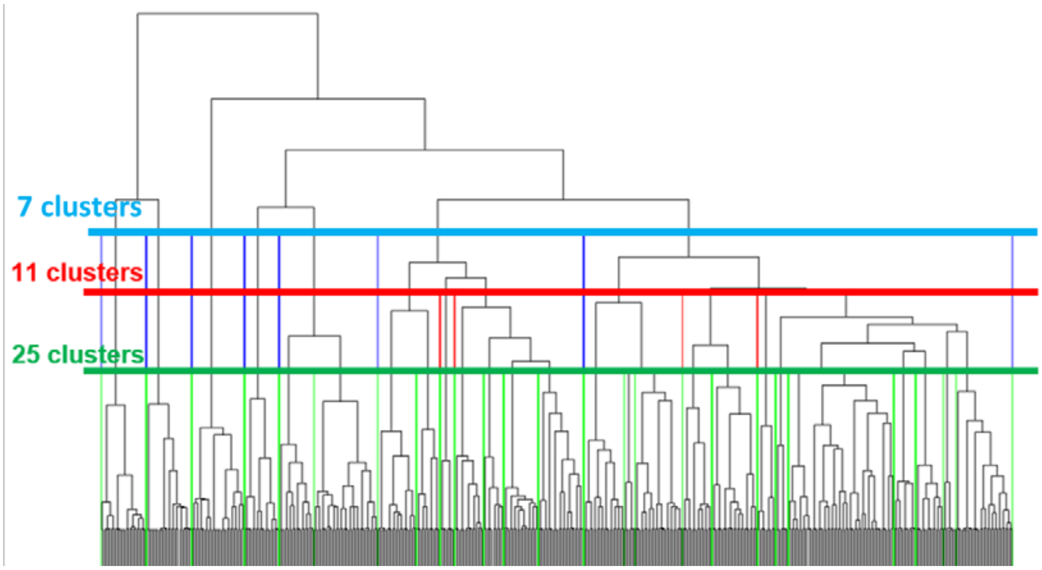


Figure 3. Dendrogram of sample data

4.2.3 Allocating the Remaining Observations

This is step 5 in Figure 1. In this process, the rest of the user-days were allocated to their nearest cluster. The allocation method is presented in section 3, Figure 2. The "nearest cluster" means the cluster with the minimum/shortest? average distance from a remaining point to all points in that cluster. After this step, all the daily transaction time profiles of the smart card users were added to clusters.

4.3 Evaluating Sampling Performance

In this section, we proposed a methodology to measure the effectiveness of different sample sizes and different numbers of groups. This approach proposes using the variance of inter-group distances (dissimilarity) and the variance of intra-group distances. The variance is based on a series of tests made using the same parameters. For example, if we choose 10 groups and a sample size of 1 000, we will execute the method a number of times (we later call it the number of draws, because each execution of the method will lead to a different draw in the sample). Ideally, the variance of inter-group distances (dissimilarity) and intra-group distances would decrease as the sample size increases. The objective is to determine a sample size with a small enough variance for inter-group and intra-group distances, so that the sample size does not influence the classification method results.

The following are the steps proposed to evaluate the sampling performance. Let us define S as the sample size, N as the number of groups and D as the number of draws:

1. Take a sample of S user-day and classify them into N groups.
2. Calculate the inter-group and intra-group distances with the method presented above.
3. Repeat steps 1 and 2, D number of times.
4. Based on the tests using D, calculate the variance of inter-group and intra-group distances. Next, calculate the combined inter-group and intra-group distance variance.

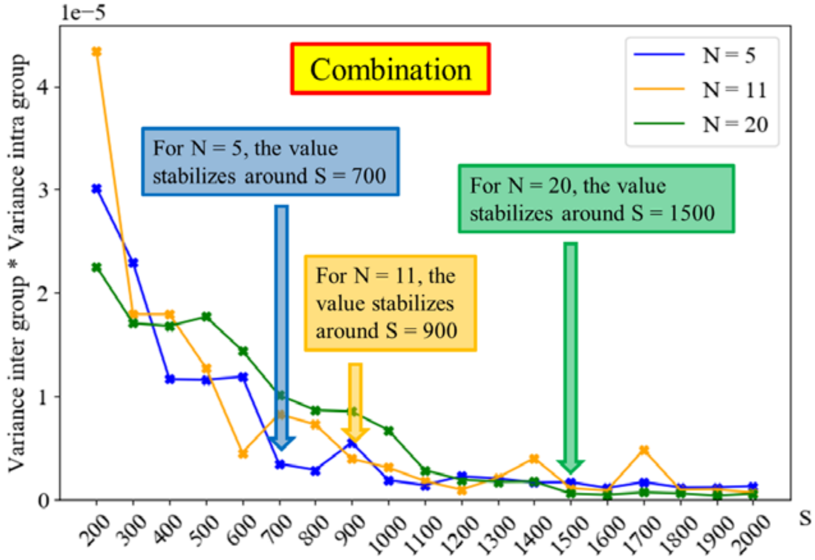
In the following section, we present the results determined using this approach with different values of S, N and D, to test the sampling method's performance in our case study.

5 RESULTS

In this section, we first present the results of the tests and then adjust the parameters to obtain the desired clustering of the temporal profiles.

5.1 Variance analysis by inter-group distance, intra-group distance, and their combined variances

Figure 4(a) illustrates the combined inter-group and intra-group distance variances for sample sizes from 200 to 2 000. The results show that for the groups where N=5, 11 and 20, the value of the indicator stabilizes around S=700, 900 and 1 500, respectively. For example, for N = 5, the variation of the indicator between S = 600 and 700 is about 10^{-5}. However, for S = 700 to 2 000, the difference between the maximum and the minimum value of the indicator does not exceed 5 × 10^{-4}. In this case, for N = 5, a sample size of 700 is sufficient to apply the clustering method. These tests were conducted using 20 random draws (D=20).

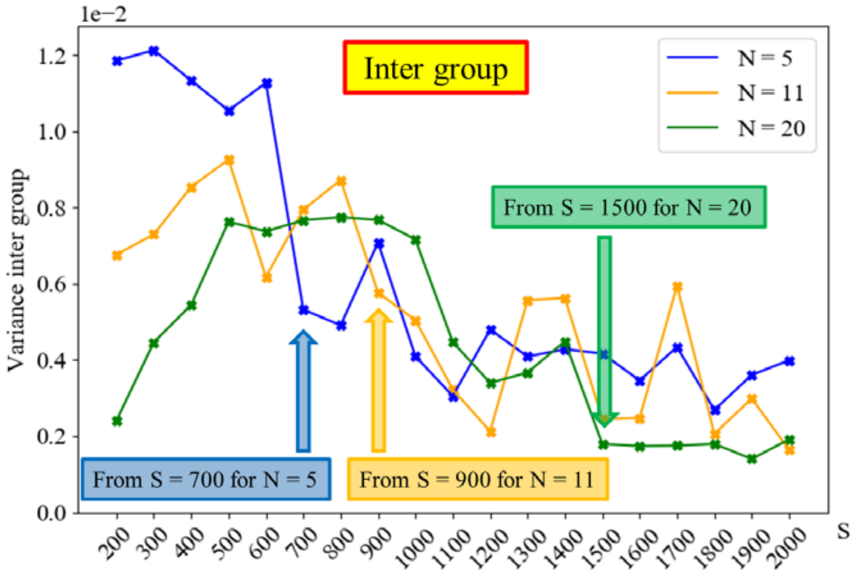


For N = 5, the value stabilizes around S = 700

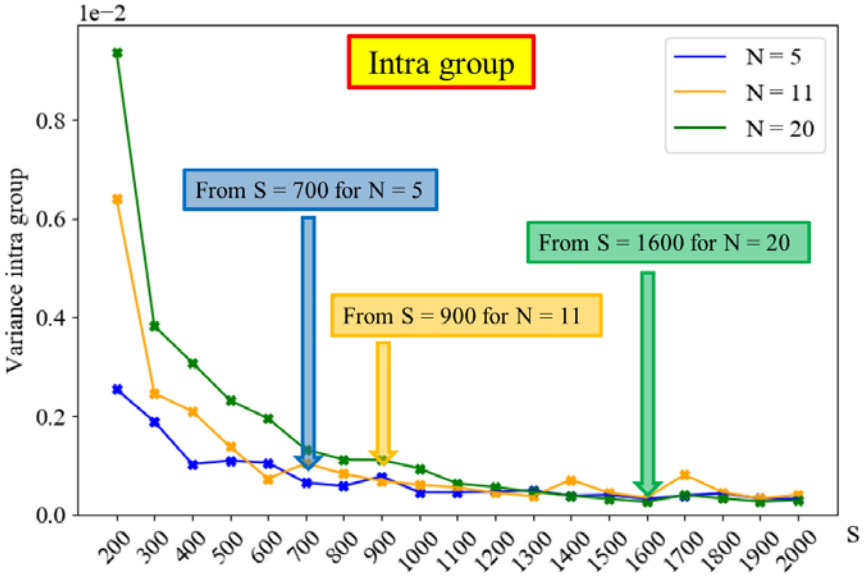
For N = 11, the value stabilizes around S = 900

For N = 20, the value stabilizes around S = 1500

(a)



(b)



(c)

Figure 4. Variance of (a) inter-group distance and intra-group distance combined (b) inter-group distance (c) intra-group distance (D=20)

We have to question whether the two combined variances can represent the tendency of both individual variances. Figure 4(b) and (c) illustrate the variances of inter-group distances and intra-group distances for N = 5, 11 and 20, respectively.

Unlike the almost monotonically decreasing function of the intra-group variances, the inter-group variances tend to demonstrate more complexity. First, it increases with the sample size. Then, the indicator stays high for a period determined by the number of groups (for N = 20, the indicator values stay high for longer than for N = 5). Finally, it decreases rapidly and reaches a stable range.

A possible explanation for this may be that in the beginning, the classification algorithm did not work well because the sample size was too small. An extreme case is to classify 6 elements into 6 groups. In this case, each group contains only one element. This would result in the inter-group distance being the same and therefore the variance of the inter-group distance would be 0. Then, as the sample size increases, the variance of the inter-group distances increases because the classification algorithm is not able to efficiently cluster the samples. In the end, the sufficient sample sizes are determined and other groups of similar sizes are created, thus rapidly decreasing the variance of inter-group distances.

In terms of choosing the sample size, we can see that the individual variances for the inter-group distances and intra-group distances show little change after S = 700, in the case of N=5. This matches the results observed for the variance of intra- and inter-distances combined (Figure 4(a)). Therefore, 700 was chosen for sample size for all three indicators (inter-group, intra-group and the combination of both). In the same way, S = 700 was chosen as the minimum sample size for N = 11. For N = 20, results show that the groups obtained become stable from around a sample size of 1 500 or 1 600.

5.2 Sensitivity analysis of the number of draws (D)

Because we are proposing a sampling process, we tested the sensitivity of the number of random draws to see if this method performs well in all cases. In addition to D=20 as presented in Figure 4(a), D=10 and D=50 are tested for 11 groups (N=11), as presented in Figure 5.

The comparison demonstrates that the results obtained from the different values for D are almost the same. For D = 10 and 50, the value starts to stabilize at around S=1 000. However, it becomes stable at S=900 when D = 20.

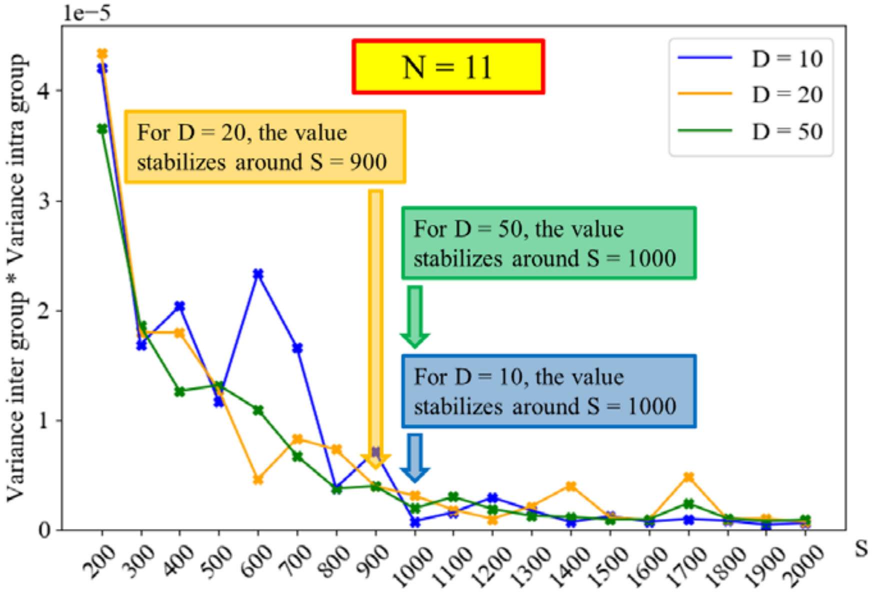


Figure 5. Variance analysis by the number of draws

It is surprising to see that a sample size of 2 000 is large enough to create clusters from 335,745 entries. We believe that this is due to the type of data and application used in this study?. Ultimately, temporal patterns remain relatively stable over time, especially for the same card.

5.3 Resulting Temporal Profiles

We applied the proposed method to the dataset using 11 groups, a sample size of 2 000 and 20 draws (N=11, S=2000, D=20). The daily temporal patterns of 6 of the 11 groups are presented in Figure 6. It shows that Cluster 2 and Cluster 3 display pendular behaviours related to commuters, however not at the same time. Clusters 6, 7 and 11 are characterized by single surges during peak hours. Cluster 9 regroups the people that are using the system in between the peak periods of the day. This type of analysis will help the transit authority identify the different types of customers in order to establish differentiated fares or service levels. Using this method also decreases the computational time. For example, for the dissimilarity matrix process, the computation time of 1 000 user-day profiles is 1% of 10 000 user-day profiles. In the case study, instead of having to use a 333,745 X 335,745 distance matrix (1,1 x 10^{11} entries), it used a 2 000 X 2 000 matrix (4 x 10^{6} entries).

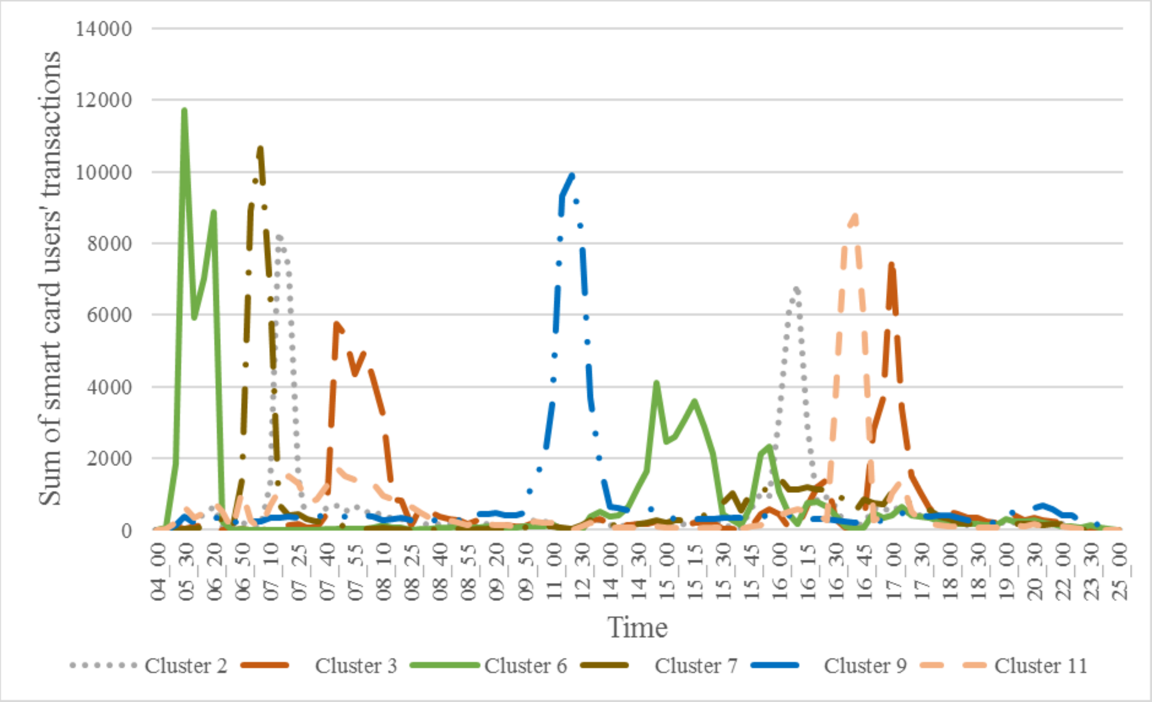


Figure 6. Resulting temporal profiles for some groups (N=11, S=2000, D=20), using STO data from Sep. and Nov. 2013

6 CONCLUSION

In this paper, we proposed a framework that combines cross-correlation distance, hierarchical clustering and a sampling method in order to characterize the temporal profiles of public transit travelers using data from smart card transactions. Applying this framework to the *Société de transport de l'Outaouais* transit network helped classify 333,745 user-days. We conducted a sensitivity analysis on the main parameters of this approach in order to test its validity with the dataset, analysing the number of groups, sample sizes and number of random draws for the sample.

The main limitation of this work is that the method for determining sampling efficacy is based on one? smart card transaction dataset. Transit users from the STO may have their own unique characteristics, and the sample size found here may not apply to data from another city. Therefore, we believe that by using our methodology?, the appropriate values of N, S and D can also be determined for other datasets.

We expect that this framework can also be applied to clustering the spatio-temporal profiles of transit users, simultaneously studying their location and time of use (He et al., 2019). We also look forward to improving computational times by proposing a strategy for calculating the distances between sample points and the remaining points, which is a time-consuming practice that actually requires less time than calculating the distance matrices between all points.

ACKNOWLEDGMENTS

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.

7 REFERENCES

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, *39*(3), 399-404.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications* (Vol. 27, No. 2, pp. 94-105). ACM.

Asakura, Y., Iryo, T., Nakajima, Y., & Kusakabe, T. (2012). Estimation of behavioural change of railway passengers using smart card data. *Public Transport*, *4*(1), 1-16.

Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In *KDD workshop* (Vol. 10, No. 16, pp. 359-370).

Bordagaray, M., dell'Olio, L., Fonzone, A., & Ibeas, Á. (2016). Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques. *Transportation Research Part C: Emerging Technologies*, *71*, 231-248.

Chen, Y., Mahmassani, H. S., & Hong, Z. (2015). Data mining and pattern matching for dynamic origin–destination demand estimation: Improving online network traffic prediction. *Transportation Research Record: Journal of the Transportation Research Board*, (2497), 23-34.

Cui, Q., Wei, Y. M., Li, Y., & Li, W. X. (2017). Exploring the differences in the airport competitiveness formation mechanism: evidence from 45 Chinese airports during 2010–2014. *Transportmetrica B: Transport Dynamics*, *5*(3), 325-341.

de Oña, R., & de Oña, J. (2015). Analysis of transit quality of service through segmentation and classification tree techniques. *Transportmetrica A: Transport Science*, *11*(5), 365-387.

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

Diab, E. I., & El-Geneidy, A. M. (2013). Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. *Public Transport*, *4*(3), 209-231.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

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

Ghaemi, M. S., Agard, B., Trépanier, M., & Partovi Nia, V. (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.

Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In *ACM Sigmod Record* (Vol. 27, No. 2, pp. 73-84). ACM.

He, L., Agard, B., & Trépanier, M. (2018). 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*, 1-20.

He, L., Trépanier, M., Agard B. (2019, submitted). Space-time classification of public transit users' activity locations from smart card data. *Public Transport.*

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*(8), 68-75.

Kieu, L. M., Bhaskar, A., & Chung, E. (2015). A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data. *Transportation Research Part C: Emerging Technologies*, *58*, 193-207.

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

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, 179-191.

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, *6*(1-2), 1-20.

Lee, W. H., Tseng, S. S., Shieh, J. L., & Chen, H. H. (2011). Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1047-1056.

Legara, E. F. T., & Monterola, C. P. (2018). Inferring passenger types from commuter eigentravel matrices. *Transportmetrica B: transport dynamics*, *6*(3), 230-250.

Li, H., & Chen, X. (2016). Unifying Time Reference of Smart Card Data Using Dynamic Time Warping. *Procedia Engineering*, *137*, 513-522.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857-1874.

Liao, W. K., Liu, Y., & Choudhary, A. (2004, April). A grid-based clustering algorithm using adaptive mesh refinement. In *7th Workshop on Mining Scientific and Engineering Datasets of SIAM International Conference on Data Mining* (pp. 61-69).

Liu, Y., & Cheng, T. (2018). Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science*, 1-28.

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

Ma, X., Wang, Y., McCormack, E., & Wang, Y. (2016). Understanding Freight Trip-Chaining Behavior Using a Spatial Data-Mining Approach with GPS Data. *Transportation Research Record: Journal of the Transportation Research Board*, (2596), 44-54.

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

Mirkes E.M. (2011). K-means and K-medoids applet. *University of Leicester*.

Mohamed, K., Côme, E., Baro, J., & Oukhellou, L. (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data. *UrbComp,(Seattle, WA, USA)*.

Mohamed, K., Côme, E., Oukhellou, L., & Verleysen, M. (2017). Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems*, *18*(3), 712-728.

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

Morency, C., Trépanier, M., Agard, B., Martin, B., & Quashie, J. (2007, September). Car sharing system: what transaction datasets reveal on users' behaviors. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE* (pp. 284-289). IEEE.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance measures for time series in R: The TSdist package. *R Journal*, *8*(2), 451-459.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications*, *36*(2), 3336-3341.

Parzani, C., Leclercq, L., Benoumechiara, N., & Villegas, D. (2017). Clustering route choices methodology for network performance analysis. *Transportmetrica B: Transport Dynamics*, *5*(2), 191-210.

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

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

Subbiah, K. (2011). *Partitioning Methods in Data Mining*.

Srimani, P. K., Mahesh, S., & Bhyratae, S. A. (2013, January). Improvement of Traditional K-means algorithm through the regulation of distance metric parameters. In *Intelligent Systems and Control (ISCO), 2013 7th International Conference on* (pp. 393-398). IEEE.

Sun, Y., Shi, J., & Schonfeld, P. M. (2016). Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: a case study of Shanghai Metro. *Public Transport*, *8*(3), 341-363.

Vicente, P., & Reis, E. (2016). Profiling public transport users through perceptions about public transport providers and satisfaction with the public transport service. *Public Transport*, *8*(3), 387-403.

Wang, W., Yang, J., & Muntz, R. (1997, August). STING: A statistical information grid approach to spatial data mining. In *VLDB* (Vol. 97, pp. 186-195).

Yang, C., Yan, F., & Ukkusuri, S. V. (2018). Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system. *Transportmetrica A: Transport Science*, *14*(7), 576-597.

Yap, M., Cats, O., & van Arem, B. (2018). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science*, 1-20.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, *17*(10), 977-987.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996, June). BIRCH: an efficient data clustering method for very large databases. In *ACM Sigmod Record* (Vol. 25, No. 2, pp. 103-114). ACM.

Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 29.

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation Research Part C: Emerging Technologies*, *75*, 17-29.