**CASPT 2018 Full Paper**

# Analyzing Interpersonal and Intrapersonal Variability of Transit Use with Smart Card Data

**Élodie Deschaintres · Catherine Morency · Martin Trépanier**

**Abstract** Variations in public transport system use can be observed from one user to another (interpersonal variations), but also within the behaviour of a same user over time (intrapersonal variations). This variability should be considered forecasting public transit demand. However, a better understanding of travel behaviours and tools to measure their variability (or regularity) are first required. This paper processes smart card data over a period of 51 weeks to investigate public transport passengers both interpersonal and intrapersonal variability using data mining methods. Several clustering algorithms based on different distances (Euclidian and non-Euclidian) are employed to build user typologies and thus highlight interpersonal variability. A typology of weeks is also created to reveal intrapersonal regularity, defined by the repetition of the same types of week in the user behaviour. Furthermore, indicators are provided to quantify the two types of variability. The results show that, even for annual pass users, expected to be quite consistent, different types of users exist in terms of frequency of use, activity ratios and stability over time. Moreover, although the vast majority of the behaviours are found to be quite regular at the intrapersonal level, some of them are more atypical.

**Keywords:** Public transport · Smart card data · Variability · Clustering · Indicators

**Elodie Deschaintres, M.A.Sc Candidate (Corresponding author)**
Department of Civil, Geological and Mining Engineering
Polytechnique Montréal, C.P. 6079, succ. Centre-Ville
Montréal (Québec) Canada H3C 3A7
Tel: +1 (438) 501-1372
Email: elodie.deschaintres@polymtl.ca

**Catherine Morency, ing., Ph.D., P.Eng., Full Professor**
Mobilité Chair on sustainability, CIRRELT / Polytechnique Montréal
Department of Civil, Geological and Mining Engineering
Polytechnique Montréal, C.P. 6079, succ. Centre-Ville
Montréal (Québec) Canada H3C 3A7
Tel: +1 (514) 340-4711 ext. 4502
Email: cmorency@polymtl.ca

**Martin Trépanier, ing., Ph.D., P.Eng., Full Professor**
CIRRELT / Polytechnique Montréal, Mobilité Chair on sustainability
Department of Mathematical and Industrial Engineering
Polytechnique Montréal, C.P. 6079, succ. Centre-Ville
Montréal (Québec) Canada H3C 3A7
Tel: +1 (514) 340-4711 ext. 4911
Email: mtrepanier@polymtl.ca

# 1 Introduction

Variations in public transit use can be observed at the individual level in terms of frequency of use, times or locations of boarding (Morency et al. 2007). Despite the known existence of this variability, most current models consider an average use per person. This limitation is mainly due to a lack of longitudinal and individualized data on mobility behaviour. However, Smart Card Automated Fare Collection (SCAFC) systems now make it possible to fill these gaps. These systems provide large amounts of data at a high level of temporal and spatial granularity, and thus enable a better analysis of the transit use fluctuations. This could help to improve demand forecasting, and thereby make micro service adjustments possible, leading to reduced operating costs and optimized vehicle allocation on the network.

This work will confirm that smart card data are particularly relevant to better understand mobility behaviours. More specifically, it provides analytical tools to study their inter and intrapersonal variability based on data mining methods. These tools are applied on a year's worth (2016) of OPUS smart card data from the SCAFC system of Montreal, Canada. The technology was implemented in 2008 by the transit authority Société de Transport de Montréal (STM) and is now well adopted: the OPUS smart card is employed by more than 90% of the transit passengers. The STM operates a metro/subway network of 4 lines (a total of 71 km long) and a bus network of 220 regular lines (covering an area of 500 $km^2$). About 1.3 million trips are made every day on these networks. The SCAFC system of Montreal only gathers tap-in validations but each observation contained both temporal and spatial information on when and where the user boarded. However, spatial information is partial because only the line number is available for bus while metro stations are reached.

The paper is organized as follows. First, the literature review presents existing works related to using smart card data to analyze mobility behaviours. Research on behaviour variability is also introduced, especially with indicators and traveller segmentation. Then, the topic of this paper is clarified and some definitions are given. The data and methodology used in this paper are also described. The results are reported and analyzed in the fourth section: several typologies of different objects are presented and then compared. Finally, the paper is concluded and future work is discussed.

# 2 Literature review

## 2.1 Smart card data and mobility behaviours

The richness of smart card data makes it possible to analyze mobility behaviours at both temporal and spatial levels, in an aggregated or totally individualized way. Simple data mining methods, statistics and visualization tools may be leveraged to describe mobility patterns from this data.

On the one hand, time information in smart card data can be used to draw temporal profiles of travel by aggregating ridership at different time intervals, for all the customers or by the type of fare (White et al. 2010). The resulting graphs may depict the entire study period observing cyclical trends (Morency et al. 2007), or simply a few days to emphasize differences between weekdays and weekends, or between peak hours and off-peak hours (Huang et al. 2015; Liu et al. 2009). Moreover, statistical tests can be applied to underline those temporal disparities in users' travel behaviour (Nishiuchi et al. 2013; Zhong et al. 2015). Boarding times are also used as a classification criterion in many studies, to create clusters of cards-weeks (Agard et al. 2006), cards-days (Morency et al. 2007) or, on a finer scale, clusters of regular hours for each passenger (Manley et al. 2016).
On the other hand, considerable effort was put forward to add a spatial dimension to travel behaviour analysis. While Zhong et al. (2015) brought out travel daily profile by station and segmented them with Community detection, Chu and Chapleau (2010) created a paradigm shift by analyzing trips between anchorage points instead of stops. Manley et al. (2016) evaluated spatial travel regularity both at the station level and at the line level. They concluded that more regular trips originate from suburban areas and move downtown. A geo-visual technique called flow-mapping has also been used by Tao et al. (2014) to examine spatiotemporal dynamics from smart card data.

## 2.2 Variability indicators

To go further than just describing or visualizing mobility behaviours, some authors developed indicators to quantify their variability.
Jones and Clarke (1988) were among the first to propose activity-based graphical and numerical methods to figure out behaviour variability. Moreover, they rightly underlined that the more detailed behaviour is described, the more apparent its variations will be. Other indicators were subsequently suggested in the literature, mainly indicators based on the frequency of trips made at different stops and at different times of the day (Huang et al. 2015; Morency et al. 2007), or the co-occurrence of the same travel attributes, measured with a contingency matrix, and indicators based on time-budget allocation (Schlich and Axhausen 2003). In addition, the variance of some indices such as the number of trips per day (Pas and Koppelman 1987), the departure time of the first trip of the day (Kitamura et al. 2006), or the individual daily time use (Raux et al. 2016) was calculated and split it into interpersonal and intrapersonal variances.
Furthermore, intrapersonal variability is often captured by comparing sequences of activity-travel events of a same individual. Wilson (1998) was the first to apply Sequential Alignment Method (SAM) in a travel behaviour context. This method calculates the dissimilarity or distance, called Levenshtein distance, between two sequences of strings in terms of the minimal number of operations (deletion, insertion, substitution) required to equalize the two sequences. Multidimensional Alignment Method was then developed by Joh et al. (2002) to consider the different

dimensions (or attributes) of the activity-travel events. For instance, Xianyu et al. (2017) applied this method to measure the degree of dissimilarity between the daily activity-travel sequences of every individual and then studied the effect of his/her sociodemographic variables on his/her variability. Instead of computing the Levenshtein distance, other authors like Goulet-Langlois et al. (2017) took into account the order of travel events by calculating an entropy rate.

### 2.3 Traveller segmentation

Customer classification can also be undertaken to reveal interpersonal variability. Many clustering algorithms can be used (k-means, HAC, DBSCAN, EM clustering, etc.) and rely on different ways of describing mobility as input.

First, a typology of users may be completed based on the characteristics of their public transit use. This use may be depicted by summarizing the user's validations into a series of discrete time bins, for instance into weekly (El Mahrsi et al. 2017) or daily profiles (Agard et al. 2013). However, to avoid scalar aggregation, Briand et al. (2017) proposed a Gaussian mixture generative based model and then kept a continuous representation of time. Another way to report public transit use is to construct indicators as was done by Ortega-Tong (2013). For each user the author defined a vector of 20 variables regarding travel frequency, temporal and spatial attributes, activity duration, sociodemographic characteristics and mode choices.

Furthermore, some authors pointed out the need to consider the order and the organization of the user's journeys over time. That's why Goulet-Langlois et al. (2016) represented each passenger by a sequence of activities spanning four weeks before performing PCA and clustering on it. Similarly, Saneinejad and Roorda (2009) clustered 282 individuals based on their routine weekly activity sequences using Sequence Alignemement Method (SAM) and the iterative neighbour-joining algorithm. Moreover, Joh and Timmermans (2011) found a heuristic approach to apply SAM in a segmentation context to large data sets like smart card databases.

Another way to segment transit passengers is to evaluate their regularity in public transit use. Ma et al. (2013) gathered four indicators in a vector (namely the number of days travelled and the numbers of similar first boarding times, route sequences and stop ID sequences) to capture travel regularity of each user. Based on this vector, they employed k-means ++ and Rough Set Theory to classify users into different regularity levels. As for Kieu et al. (2014), they first used the DBSCAN algorithm to identify regular OD pairs and habitual travel times for each passenger. They then adopted an a priori market segmentation approach to segment transit passengers into four types of regularity.

All the previous works will serve as a basis for the following study which employs similar methods. However, the originality of this paper lies in the combination of several of these methods to propose a formal and analytical scheme to study both interpersonal and intrapersonal variability of a group of public transit users. Furthermore, unlike mostly of the works reviewed before, this paper will analyze variability over an entire year.

## 3 Problem, Data and Methodology

### 3.1 Problem description and hypotheses

The studies previously described in the literature review mainly focused on day-to-day variability. It means they assume that behaviours are regular on a daily cycle. However, variability could be visible at another than the daily level, for example monthly or weekly level, that is why we decided to study individual behaviours regarding different scales of a year in this paper. Note also that we use the word "regularity" as the opposite of variability.

On the one hand, interpersonal variability is analyzed at the monthly level. In this way, each user is characterized and dissociated from others by his monthly use of the transit network over the year. Different groups of users will be formed according to their transit use along the twelve months in 2016, and this typology will lead to big types of annual patterns specified with monthly profiles.

On the other hand, intrapersonal variability is examined at the weekly level. All the weeks of every transit user are considered separately, and the intrapersonal regularity of a given user is defined as the repetition of the same weekly patterns in his behaviour during the year. Thus, we will create a typology of weeks and look at the diversity of week types observed in the individual behaviour of each user. Sequences will also be built to consider the order and the organization of these weekly patterns over the year.

Moreover, in both cases, spatial regularity is stated by the use of the same metro stations and the same bus lines during the time scale studied.

### 3.2 Data sampling

The dataset used in this paper is a sample of the 2016 OPUS transactional database of the STM. This sample was selected among the smart card users only, who made 89% of the total validations in 2016. Moreover, only the users who did at least one validation during the 51 complete weeks (from Monday to Sunday) of 2016 were considered. This condition of completeness is necessary for the week typology explained later. Therefore, the dataset used in this paper spans from January 4th, 2016 to December 25th, 2016. This study period is much longer than the one

recommended by Schlich and Axhausen (2003) to investigate behaviour variably, namely two weeks.

Among the smart card users active at least once during the 51 weeks studied, we settled for passengers who used only one type of product during the year because we assume there is a link between the user behaviour and the product he used. These one-product type users represent almost 64% of the total smart card users. More specifically, we sampled annual pass users (12% of the one-product type users), with an amplitude of 12 months (39% of the annual pass users), which means they used their card at least once in January and once in December. At the end, the sample used in this paper is composed of 56,988 cards.

These card users were chosen because they are present on the transit network all the year with the same card and can thus be followed over a longer period. Moreover, **Table 1** reports the distribution of the cards according to their number of active months over the year (a month is said active if the card was used at least once during this month). This table allows us to conclude that the vast majority of the sampled users are active every month. In the following, we will propose a methodology to answer the question: are they regular too?

**Table 1** Card distribution according to the number of active months

| **Number of active months** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **% of cards** | 0.0% | 0.1% | 0.1% | 0.2% | 0.3% | 0.4% | 0.6% | 0.9% | 1.1% | 1.9% | 4.6% | 89.8% |

### 3.3 Methodology

The proposed methodology consists of three major steps. They are illustrated by the methodological scheme in **Fig. 1** and each of these steps is described below.

Step 1: Data preprocessing

The first step is to preprocess the sampled smart card data. First, validations are converted into trips to avoid a spatial bias which would have led to overestimation of long trips. This means that if we had used validations, at an equal number of trips, a user who would have made longer trips would also have had more transfers, therefore he would have made more validations and would have been seen as a more frequent user than one making short trips. Business rules based on the STM's fare policy are used to achieve this translation. For instance, a threshold of 120 minutes is applied, and a user cannot board the same bus line or enter the metro network twice in the same trip.

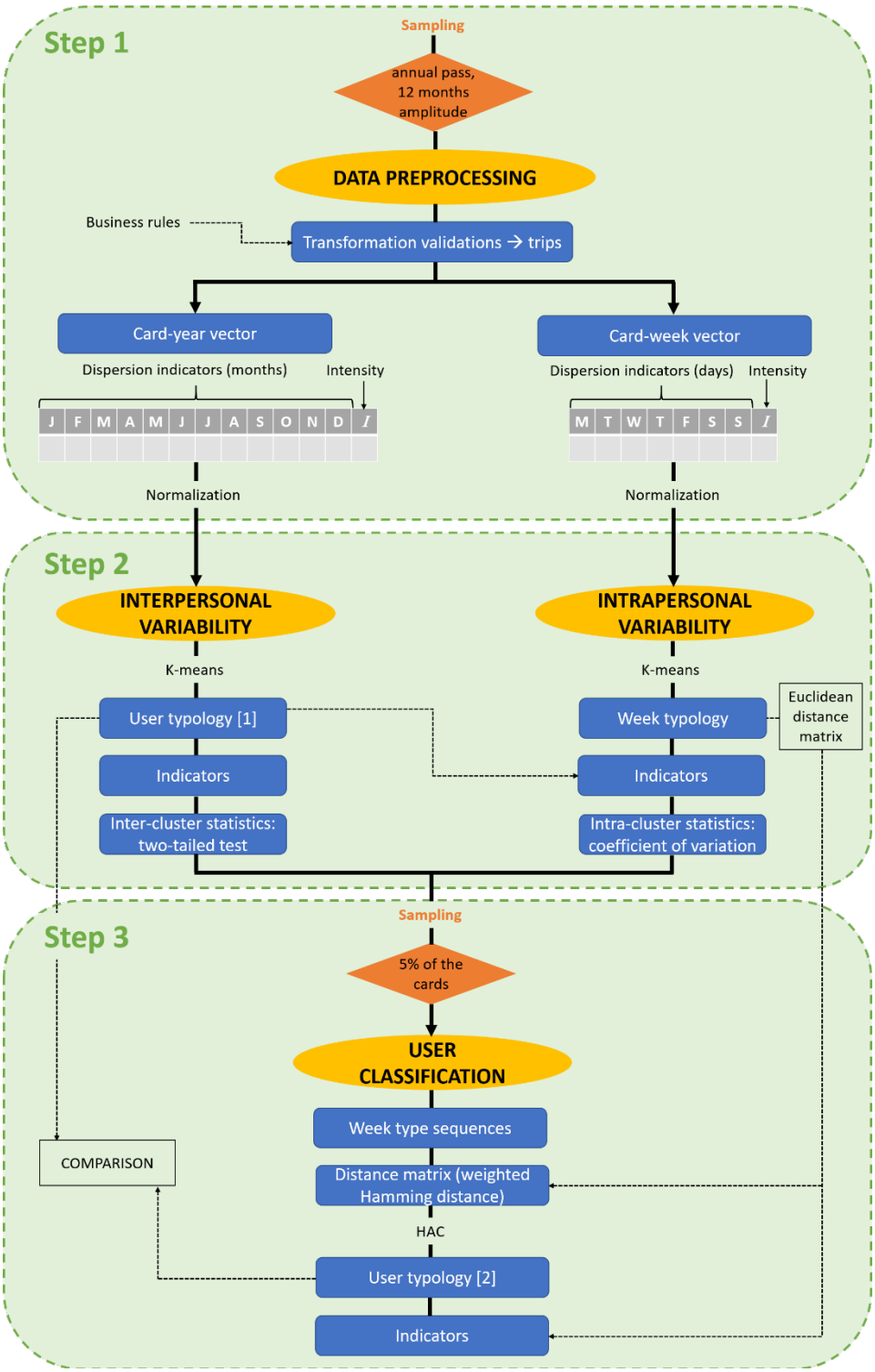



**Fig. 1** Methodological scheme

After that, travel activities of each passenger are summarized into vectors. Inspired by the format used in Morency et al. (2017) for bikesharing, "card-year" vectors are constructed to report the monthly use of the system by each card during the year, and "card-week" vectors are built to account for the daily use by each card for each week. Every vector is composed of intensity and dispersion indicators about public transit use. The "card-year" vector (*respectively the "card-week" vector*) holds the average number of trips per active month (*respectively day*) and the number of trips made each month (*respectively day*). This results in vectors of thirteen (*respectively eight*) continuous features. A month (*or a day*) is said active when the card user taped-in his/her card at least once during it. It means that we only consider the months (*or the days*) with more than 0 trips in the calculation of this average monthly (*or daily*) intensity. At the end, we obtain a 56,988 "card-year" vectors database and a 56,988×51 = 2,906,388 "card-week" vectors database, because one single card matches one "card-year" vector but 51 "card-week" vectors. **Table 2** a) and b) shows an extract of these two databases. Afterward, the vectors are normalized in order to give the variables comparable weights in the clustering process. We use the same normalization method as Morency et al. (2017), except for the daily intensity which is normalized with a logarithm function to reduce the influence of outliers.

**Table 2** Extracts of a) the cards-year database b) the cards-week database

**a)**

| Card_id | January | February | March | April | May | June | July | August | September | October | November | December | Monthly intensity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| i | 39 | 43 | 43 | 40 | 40 | 48 | 0 | 55 | 41 | 42 | 53 | 38 | 43.8 |
| ii | 51 | 57 | 62 | 46 | 58 | 63 | 71 | 74 | 60 | 49 | 50 | 47 | 57.3 |
| iii | 6 | 0 | 2 | 7 | 5 | 0 | 7 | 3 | 7 | 10 | 14 | 6 | 6.7 |
| iv | 74 | 80 | 93 | 92 | 78 | 63 | 72 | 78 | 97 | 69 | 66 | 61 | 76.9 |

**b)**

| Card_id | Week | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday | Daily intensity |
|---|---|---|---|---|---|---|---|---|---|
| i | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 2.0 |
| ii | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 |
| iii | 1 | 3 | 2 | 2 | 3 | 2 | 2 | 0 | 2.3 |
| iv | 1 | 4 | 5 | 2 | 0 | 2 | 2 | 4 | 3.2 |

Step 2: Interpersonal and intrapersonal variability

On the one hand, interpersonal variability is determined by segmenting the "card-year" vectors. This leads to groups of transit user cards according to their annual use of the transit network. In this way, the typology obtained allows to show differences in mobility behaviour between users during a year's worth.

The clustering approach employed here is the k-means method: it partitions the vectors into k clusters by minimizing the within-cluster sum of squares. The computed method relies on the Lloyd algorithm and the chosen distance is Euclidean. To avoid getting a local optimum instead of the global optimum, which is a traditional problem in the k-means algorithm (Steinley 2006), ten different initializations are achieved. Moreover, two criteria are tested to determine the value of the parameter k. We first look at the percentage of variance explained as a function of the number of clusters, and k is chosen by the "elbow criterion", hence at the point where an angle is formed in the graph. The second test is the two-step method used by Morency et al. (2017), which consists in producing 30 clusters with k-means and then drawing a dendrogram based on the 30 mean centres obtained in the previous step. **Fig. 2** reports the two methods used to select the most appropriate number of clusters k. Here, we choose k = 6 clusters.

Moreover, both temporal and spatial indicators of transit use are estimated at the individual level to characterize the travel features of each group and help the analysis of the clustering results. These indicators are the following:

- **TripsActM**, average number of trips per active month: average number of trips per month by considering only the months when the card was active at least once. It is the intensity variable used in the 'card-year' vector for clustering process. Only normalized results are presented in this paper for this indicator because the absolute value is confidential and used to calculate the STM's funding.
- **ActivityR**, average activity ratio: average ratio between the number of active months and the amplitude (here, 12 months).
- **%Business**, average proportion of trips made during business days: the average proportion of trips made during non-business days may then be calculated by subtracting this indicator to 100.
- **%MetroBus**, average proportion of trips made by metro and bus: the two modes are used in these trips.
- **%MetroOnly**, average proportion of trips made by metro only: only metro is used in these trips. Consequently, the average proportion of trips made by bus only may be calculated by:

  $$100 - \%\text{MetroBus} - \%\text{MetroOnly}$$

- **MetroStations**, average number of different metro stations used during the year: each metro station must be used at least once during the year to be considered.
- **BusLines**, average number of different bus lines used during the year: each bus line must be used at least once during the year to be considered.

Each indicator is calculated in each group by averaging the desired variable on the group members only. Furthermore, the Wilcoxon–Mann–Whitney U test is used to verify that the observed pairwise differences between the 6 groups are significant for each indicator. As a non-parametric test, the Wilcoxon–Mann–Whitney U test does not assume that the data follows a Gaussian law and is not affected by outliers (Cleophas and Zwinderman 2011). More specifically, a two-tailed test is applied and evaluates whether the two compared distributions have the same central tendencies (same centre, same average, same median). Because of the large data size, the U statistic follows approximately a normal distribution and an asymptotic Z statistic is calculated (Adjengue 2014). If the corresponding p-value is small, it means the null hypothesis may be rejected, hence the difference between the two groups is statistically significant.

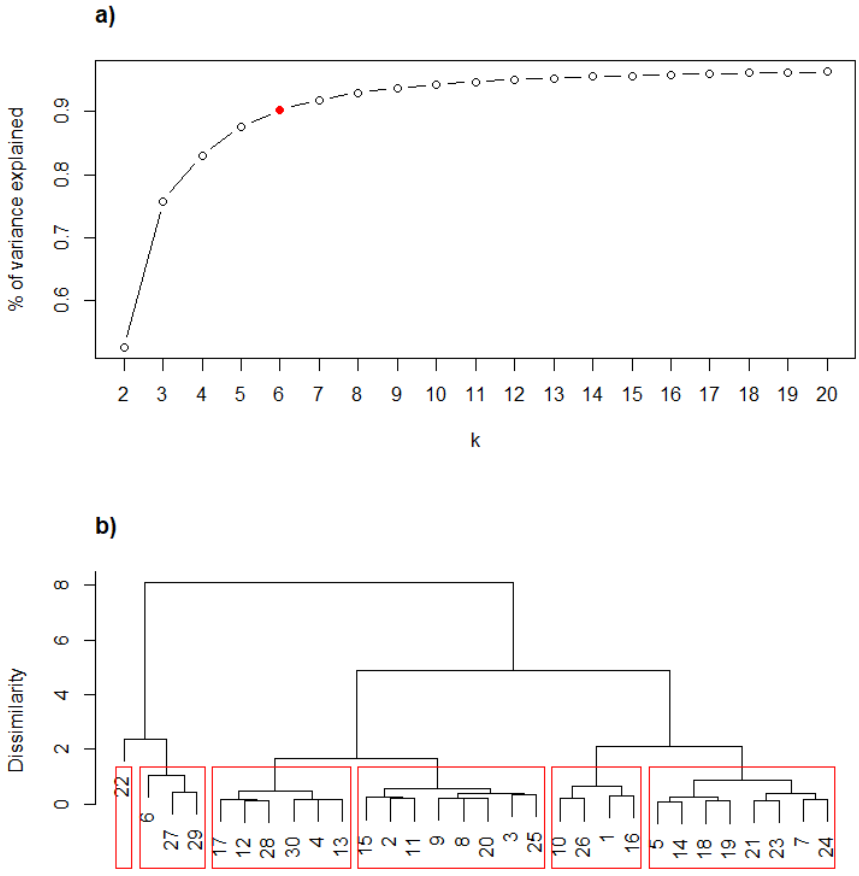



**Fig. 2** Choice of the number of user clusters k with two criteria: a) percentage of variance explained b) dendrogram

On the other hand, intrapersonal variability is determined by segmenting the "card-week" vectors with the same clustering methods as for interpersonal variability. This leads to k = 10 clusters of typical weeks for all the dataset. Then, the level of intrapersonal regularity of a given user is defined by measuring the repetition of the same weekly clusters in his transit use during the year. It means the lower the number of clusters to which the user's weeks belong, the more regular he is at the intrapersonal level. Some indices based on this definition are calculated (as a mean of the individual indicators) within the 6 user cards groups previously obtained. They allow to quantify intrapersonal variability of the members of these groups based on a weekly cycle. The chosen indicators are described below for one card user:

- **NbCL**, average number of week clusters: the average number of week clusters to which belong all the 51 weeks of the card user.
- **W0Trips**, average number of weeks without (0) trips: the number of weeks which belong to the cluster without trips (cluster 10 here). This indicator measures immobility.
- **%1stCL**, average proportion of weeks in the most used (the 1st one) week cluster: this indicator measures the concentration of the weeks in the most used cluster of each card user. The higher it is, the more regular the user is.
- **NbCL80%**, average number of week clusters in which there are 80% of the weeks. The lower it is, the more regular the user is.
- **Entropy**, average entropy of the proportions of weeks in each cluster: we use the Shannon entropy defined by the following equations.

$$H_i(X) = -\mathbb{E}\left[\log P\left(X = x_{ij}\right)\right] = -\sum_{j=1}^{n} P_{ij} \log P_{ij} \qquad \text{(Eq. 1)}$$

$$H_i^*(X) = \frac{H_i(X)}{\log(n)} \qquad \text{(Eq. 2)}$$

where $H_i$ is the entropy index for card user $i$, $H_i^*$ the corresponding normalized entropy, $n$ the number of different week clusters ($n = 10$ here) and $P_{ij}$ the proportion of the weeks of the card user $i$ which belongs to the week cluster $j$. The lower the entropy (probabilities $P_{ij}$ close to 0 or 1), the less diverse are the week types of the card user and then the more regular he is.

Moreover, a coefficient of variation is calculated to evaluate the variability of each indicator within each group. For every variable studied, the coefficient of variation measures the dispersion of the values around the calculated mean in the group. It is defined as the ratio between the standard deviation $\sigma$ to the mean $\mu$ :

$$CV = \frac{\sigma}{\mu} \cdot 100\% \qquad \text{(Eq. 3)}$$

Note that in the following, we will use the code W$x$ to designate the week cluster $x$, with $x$ ranging from 1 to 10.

Step 3: User typology based on week type sequences

In step 3, another method (non-Euclidean) is tested to leverage the previous results of the two segmentations done in the step 2. For each user, a new "card-year" vector is built as a sequence of weekly clusters: it means that every week of the user is associated with the cluster to which it belongs. **Table 3** presents an example of such a sequence.

**Table 3** Example of a week type sequence

| Card_id | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | … | 50 | 51 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| i | W2 | W3 | W5 | W5 | W5 | W5 | W9 | W9 | W1 | W3 | | W9 | W2 |

Then, a distance matrix is calculated between all these sequences. To achieve that, a weighted Hamming distance is used. The usual Hamming distance counts the number of characters that differ between two sequences of same length. However, instead of applying a substitution weight of 1 whenever there is a difference between the two sequences, a weight equal to the Euclidean distance between the two compared week clusters is applied here. This Euclidian distance is evaluated between the two centres of the week clusters. The equation below gives the mathematical expression of this specific distance (or dissimilarity) between two cards $i$ and $j$.

$$d(\text{card } i, \text{card } j) = \sum_{k=1}^{N} d_E(W_{(i,k)}, W_{(j,k)}) \qquad \text{(Eq. 4)}$$

with $N = 51$ weeks, $d(A,B)$ indicates the distance between A and B, $d_E$ refers more specifically to the Euclidian distance, and $W_{(i,k)}$ is the cluster to which belongs the $k$th week of the card user $i$. Moreover, an example is illustrated in **Table 4** with two sequences of 5 weeks. To obtain the distance (or dissimilarity) between these two cards, we compare two by two weeks of each sequence and we sum the pairwise distances between the clusters to which they belong. The resulting distance for this example is given by equation 5.

**Table 4** Example of two sequences over a period of 5 weeks

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| user 1 | W2 | W3 | W5 | W5 | W5 |
| user 2 | W9 | W9 | W9 | W2 | W2 |

$$d_E(W_{(1,3)}, W_{(2,3)}) = d_E(W_5, W_9)$$

$$d(\text{user } 1, \text{user } 2) = d_E(W_2, W_9) + d_E(W_3, W_9) + d_E(W_5, W_9) + 2 * d_E(W_5, W_2) \quad \text{(Eq. 5)}$$

As the calculation of such a distance matrix is time-consuming, a sample of 5% of the cards is selected, corresponding to 2850 card users. The produced matrix 2850x2850 serves as a basis to create a new card user classification by applying a Hierarchical Agglomerative Clustering (HAC). More precisely, the Ward (1963)'s method, with the application of the Ward's clustering criterion (Murtagh and Legendre 2014), is implemented recursively by the Lance–Williams algorithm. It was indeed proved that the Ward's method may be generalized to other distances than Euclidean distance because the coefficients of the Lance-Williams formula used to update the inter-cluster distances remain the same for any dissimilarity $d$ (Batagelj 1988; Strauss and Michael Johan von 2017).

Furthermore, we propose two indicators based on sequence analysis. The first one measures sequence variability within each cluster, hence an average intra-cluster dissimilarity. It is calculated by summing all the pairwise dissimilarities between the members of each cluster, and then dividing by the total number of members in that cluster. The second indicator is defined to quantify the average intrapersonal variability within each cluster. This individual indicator, given by equation 6, evaluates the average Euclidean distance between two successive weeks of a card user, thus the mean instability of the week cluster's membership. The smaller is this indicator, the more stable is the user's weekly behaviour over time. This index is computed for each user $i$ and a mean is assessed on all the users of each cluster.

$$V_i = \frac{1}{N-1} \sum_{k=1}^{N-1} d_E(W_{(i,k)}, W_{(i,k+1)}) \quad \text{(Eq. 6)}$$

Finally, the two card user typologies obtained, the first one based on transit use indicators and the second one based on week type sequences, are crossed by means of a confusion matrix. Two internal clustering validation measures are also calculated to compare the quality of the two classifications. First, the Dunn's index (D) is the ratio of the minimum distance between two cards not classified together (separation measure) and the maximum distance between two cards classified together (compactness measure). Secondly, the silhouette index (S) is summation-type index based on intra-cluster and inter-cluster pairwise dissimilarities. The optimal partition is obtained by maximizing these two indices. See (Arbelaitz et al. 2013; Liu et al. 2010) for more details.

## 4 Application

4.1 Interpersonal variability - User typology based on monthly use indicators

In this section we will present the results regarding the first typology of users made from the card-year vectors composed of intensity and dispersion indicators at the monthly level. The centres of the 6 clusters obtained are illustrated below in **Fig. 3**. The graph on the left shows the distribution of the average normalized monthly intensities measured in each cluster. Note that the 6 clusters were sorted in descending order of this average monthly intensity. Moreover, the plot on the right depicts the average profile of the annual trip distribution by month in each cluster. The dashed gray line represents a consistent use (same number of trips every month). In the following, we will designate by C$x$ the $x$th cluster of this first typology.

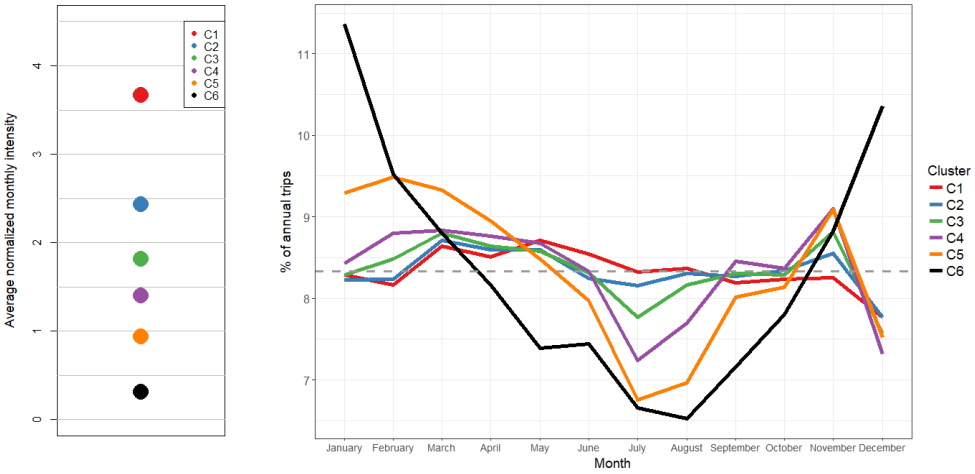



**Fig. 3** Normalized monthly intensity and trip distribution of the 6 clusters' centres

We observe that the average monthly intensity is clearly different from one cluster to another: the intensities plotted in the left graph are well distributed. A lower intensity seems to be related to a longer summer period during which the number of trips decreases. In the clusters C5 and C6, this trend is more apparent and turns into higher proportions at both ends of the year (because the sum of the 12 points must equal 1). Moreover, a peak of trips in November is visible for almost all clusters.

The size (in proportion of cards-year) and the trip distribution among the 6 clusters are given in the first part of **Table 5**. The two most frequent and constant clusters, namely C1 and C2, accounts for about 10% of the cards but they made almost 19%

of the total trips in 2016. The clusters C4, then C3, are the biggest ones and they represent together more than 65% of the trips.

Furthermore, some indicators (defined in the 3.3 Methodology section) were calculated within each cluster to characterize both their temporal and spatial use of public transit. Statistical tests were also applied for each indicator to verify if the differences between pairwise groups are significant. The results are reported in **Table 6** in terms of p-value.

**Table 5** Average usage indicators per cluster

| | Indicator | C1 | C2 | C3 | C4 | C5 | C6 | Total |
|---|---|---|---|---|---|---|---|---|
| DISTRIB-UTIONS | Cards-year | 1.4% | 8.1% | 19.3% | 38.8% | 18.6% | 13.8% | 100% |
| | Trips | 3.7% | 14.6% | 25.9% | 40.1% | 12.7% | 3.0% | 100% |
| TEMPORAL INDICATORS | TripsActM | 3.66 | 2.43 | 1.82 | 1.40 | 0.93 | 0.31 | 1.36 |
| | ActivityR | 0.998 | 0.999 | 0.999 | 0.997 | 0.985 | 0.873 | 0.978 |
| | %Business | 75.2% | 78.5% | 86.6% | 94.5% | 92.3% | 86.6% | 89.9% |
| SPATIAL INDICATORS | %MetroBus | 26.2% | 31.4% | 29.5% | 20.7% | 16.5% | 11.3% | 21.2% |
| | %MetroOnly | 57.6% | 50.1% | 56.0% | 68.6% | 65.3% | 71.5% | 64.3% |
| | MetroStations | 32.2 | 26.6 | 21.3 | 14.2 | 13.5 | 10.7 | 16.2 |
| | BusLines | 25.4 | 22.2 | 15.2 | 8.0 | 7.7 | 4.8 | 10.3 |

**Table 6** P-values resulting from statistical tests between pairwise clusters for each indicator

| Test | TripsActM | ActivityR | %Business | %MetroBus | %MetroOnly | MetroStations | BusLines |
|---|---|---|---|---|---|---|---|
| C1 v C2 | 0.000 | 0.019 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C1 v C3 | 0.000 | 0.107 | 0.000 | 0.019 | 0.047 | 0.000 | 0.000 |
| C1 v C4 | 0.000 | 0.382 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C1 v C5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C1 v C6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C2 v C3 | 0.000 | 0.144 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C2 v C4 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C2 v C5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C2 v C6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C3 v C4 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C3 v C5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C3 v C6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| C4 v C5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.143 |
| C4 v C6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.906 | 0.000 | 0.000 |
| C5 v C6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

As observed previously in **Fig. 3**, travel intensity is significantly different between all pairs of clusters, ranging from 3.66 normalized trips per active month to 0.31 normalized trips per active month on average. Activity ratio is similar between the first four groups but a little lower in the last two. This is because the card users of these two groups tend to be active less time, especially in the middle of the year. The proportion of trips made in business days is very high in the clusters C4 and C5, so the card users of these two groups may mostly use public transit to commute to work. This proportion is lower in C1 and C2 who use public transit during non-business days too. At the modal and spatial levels, the majority of trips are made by metro only, especially for the cluster C6. The clusters C1 and C3 seem to be similar in their modal choices (high p-values). Moreover, we remark that the average number of metro stations and bus lines used during the year decreases with the frequency of use, from C1 to C6. The boarding locations set of C1 and C2 is particularly diversified: they used at least once almost half of the metro stations in the network (consisting of 68 stations). Therefore, a higher number of trips seems to encourage a higher acquisition rate of the network.

4.2 Intrapersonal variability - Repetition of the same types of weeks in behaviours

Remember that in this paper we define intrapersonal regularity of a card user by the repetition of the same types of weeks in his/her behaviour. First, the week typology we work with is presented in **Fig. 4**. We note W$x$ the $x$th cluster of this typology.

The first cluster W1 is the most common one, hence the typical regular work week. More trips are made at both ends of this working week. On the contrary, no trips are made on Friday in the cluster W2 and on Monday in the cluster W3. In the cluster W4 trips are concentrated in the weekend. From the cluster W5 to W9 there is only one travelled day (from Monday to Friday respectively). Finally, no trips at all are recorded during the week of the last cluster W10.

The distribution of all weeks into the 10 clusters is displayed in **Fig. 5** for each of the 51 weeks studied. We notice an increase in the share of the cluster W10 during the summer period because, as we saw before, the card users tend to travel less during this period. Furthermore, the proportion of the weeks belonging to the cluster W2 is higher in the weeks 12, 25 and 26, which coincide with public holidays on Fridays (Good Friday, National Holiday of Quebec and Canada Day). Similarly, the proportion of the weeks belonging to the cluster W3 increases in the weeks 13, 21, 36 and 41 because they correspond to public holidays on Mondays (Easter Monday, National Patriot's Day, Labour Day and Thanksgiving).

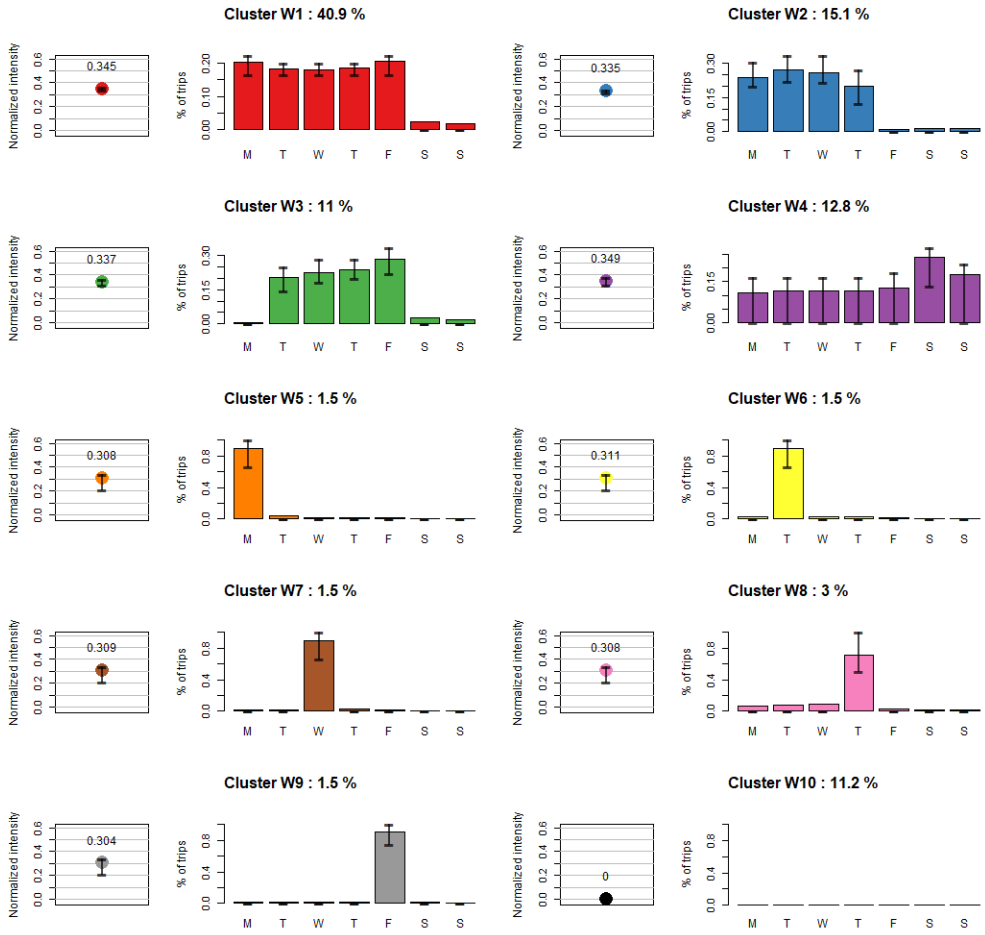



**Fig. 4** Representation of the 10 week clusters' centres with the first and third quartiles of each variable

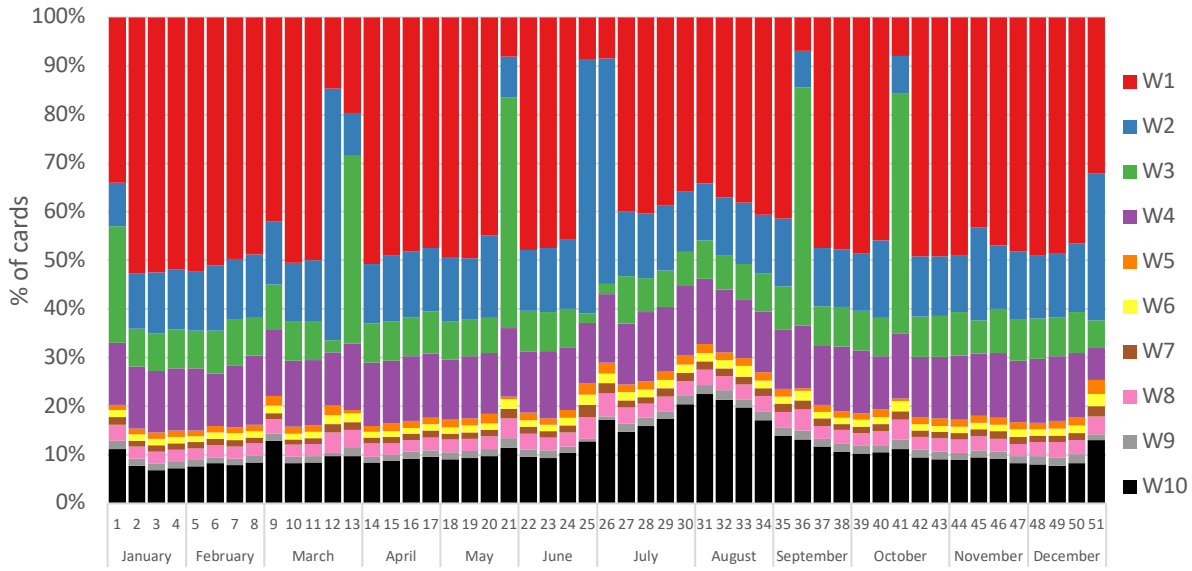
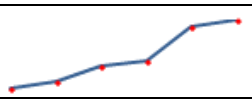
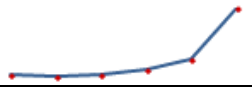
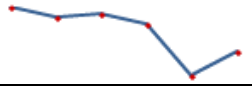
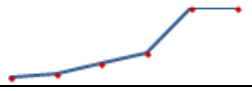
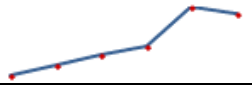



**Fig. 5** Distribution of the 51 weeks studied in the 10 clusters

In addition, **Table 7** provides some indicators to measure average intrapersonal variability within each cluster of card users (produced in section 4.1). The variations between the 6 clusters are illustrated for each indicator by the curve in the last column. Moreover, coefficients of variations are given to capture intra-cluster variability of each indicator.

First, we bring out that the average number of different week clusters (NbCL) increases from C1 to C6. This means the most frequent users, who are more regular at the monthly level according to their consistency in **Fig. 3**, are also more regular at the weekly level because all their weeks belong mainly to the same week types. The same trend is revealed by the entropy index, which is lower for the first clusters of card users. This indicates their weekly behaviours are less diversified. Moreover, the average number of weeks with no trips (W0Trips), which therefore belong to the cluster W10, is higher for the last clusters. This still echoes what we claimed previously: these users tend to travel less, especially in the middle of the year. However, the high coefficients of variation suggest that there is a lot of variability in each cluster. Similarly, the first clusters get the highest proportions of weeks which belong to the most frequent week type, and 80% of their weeks are gathered in a lower number of week clusters. This confirms their weekly transit use is more stable over the year, therefore they are more regular at the intrapersonal level according to our primary definition of intrapersonal regularity.

**Table 7** Average intrapersonal indicators per cluster of users, with coefficients of variation

| Indicator | C1 | C2 | C3 | C4 | C5 | C6 | TOTAL | Variations |
|---|---|---|---|---|---|---|---|---|
| NbCL | 4.34 | 4.65 | 5.46 | 5.80 | 7.68 | 8.00 | 6.28 |  |
| *(CV in %)* | *(35.7)* | *(25.7)* | *(22.0)* | *(25.0)* | *(21.7)* | *(21.3)* | *(29.5)* | |
| W0Trips | 1.30 | 0.90 | 1.37 | 2.89 | 5.98 | 22.63 | 5.72 |  |
| *(CV in %)* | *(204.4)* | *(192.9)* | *(128.6)* | *(80.6)* | *(83.6)* | *(53.5)* | *(152.9)* | |
| %1stCL | 0.62 | 0.59 | 0.60 | 0.57 | 0.42 | 0.49 | 0.54 |  |
| *(CV in %)* | *(24.8)* | *(21.7)* | *(22.0)* | *(22.4)* | *(34.6)* | *(41.5)* | *(29.6)* | |
| NbCL80% | 2.10 | 2.17 | 2.44 | 2.67 | 3.81 | 3.82 | 2.95 |  |
| *(CV in %)* | *(32.0)* | *(25.7)* | *(27.7)* | *(28.0)* | *(31.3)* | *(41.3)* | *(39.0)* | |
| Entropy | 0.40 | 0.44 | 0.49 | 0.52 | 0.69 | 0.66 | 0.56 |  |
| *(CV in %)* | *(35.5)* | *(24.3)* | *(22.0)* | *(21.4)* | *(21.1)* | *(30.9)* | *(28.5)* | |

4.3 Sequence analysis - User typology based on week type sequences

The results of the user segmentation based on their week type sequences are reported in this section. First, a typology in seven clusters of sequences was achieved and sorted in descending order of cluster size. The notation S*x* is used to nominate the *x*th cluster of this second user typology. **Fig. 6** shows all the sequences observed by cluster. The numbers of the 51 weeks are enumerated on the X-axis and cards are represented along the Y-axis. For each card we display its week type sequence, according to the colour of the cluster to which belongs each week. On the right of each graph all weeks of the cluster are distributed into the 10-week types.

In the first cluster S1 the regular work week W1 (in red) predominates, but we also recognize microphenomena like the public holidays already discussed in **Fig. 5** by the concentration of the types W2 and W3 in some particular weeks. Here is one of the proofs that our methodology takes into account the position of the week patterns. The cluster S2 is similar to S1 but with a higher proportion of weeks belonging to the week cluster W4, which implies more trips are made on the weekend. Moreover, the public holidays are less clear in S2 than in S1. In the cluster S3 a large proportion of weeks are in the purple week type W4, which means trips are mainly concentrated on the weekend. The cluster S4 gathers users who travel little, at least by public transit, since the cluster W10 without trips prevails. In the cluster S5 the blue week type W2 is overriding, which means the users in this group tend not to travel on Friday. The cluster S6 is composed of unpredictable users whose weeks are well distributed among the 10-week types. Finally, the cluster S7 gathers users who look like those of the S1 cluster at the beginning of the year but who travel less in the second half of the year.

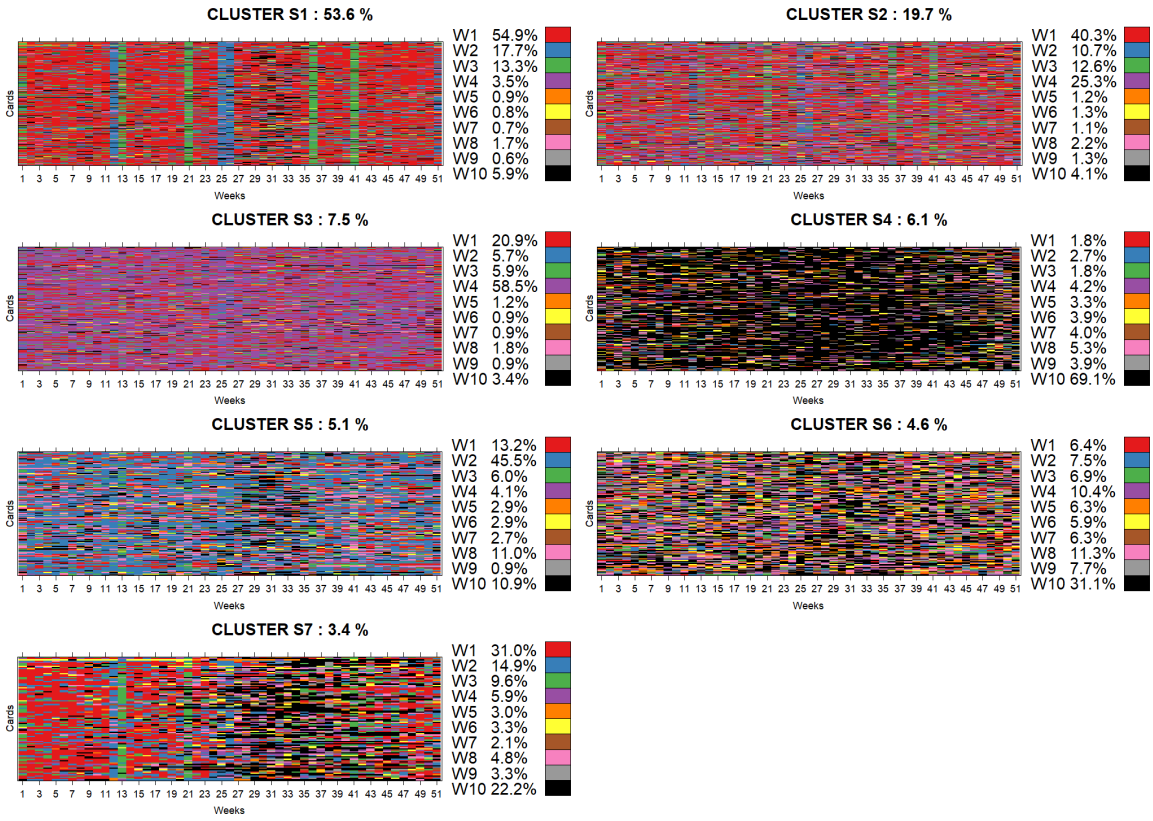



**Fig. 6** Week type sequences and week distribution in the 7 clusters of users

As the first group includes more than half of the users, we decided to force its decomposition by applying another clustering on it. The 5 subgroups obtained are presented in the following figure in descending order of their size. There are named using the code S1-$x$ where $x$ is the $x$th subgroup. This new segmentation reveals card users with a different period of summer vacation: from late July (cluster S1-3) to early August (cluster S1-4) or late August (cluster S1-2), users with dotted vacation and a decrease in trips on Friday during the summer period (cluster S1-5) or users who don't take any summer vacation (cluster S1-1), unless they continue to travel by public transport during their holidays.

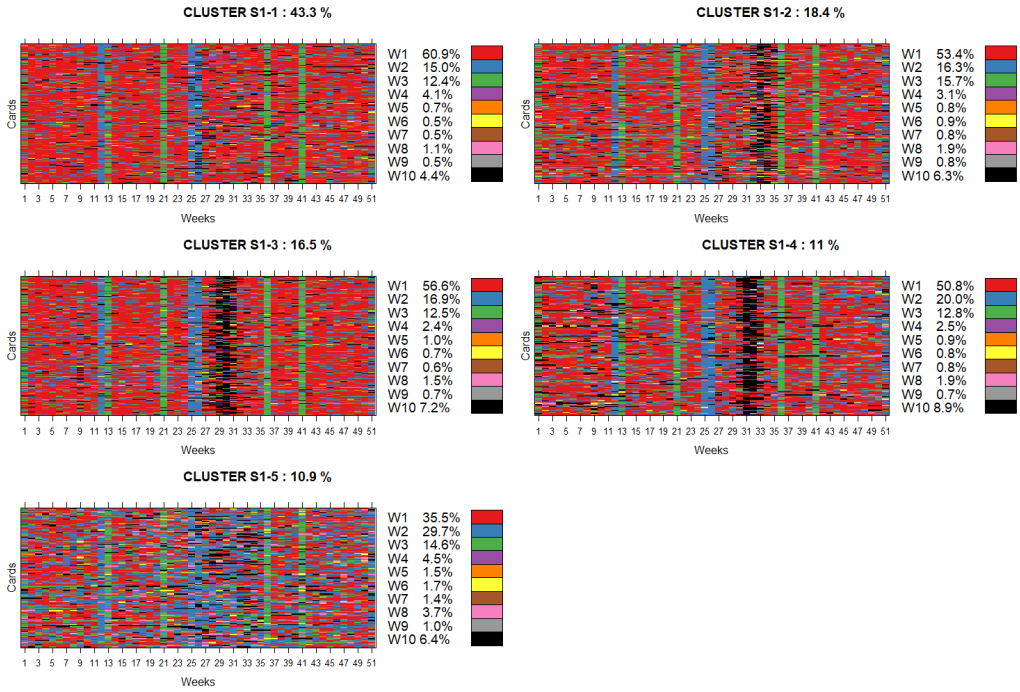



**Fig. 7** Decomposition of cluster 1: users with different holiday periods

The card proportion, average intra-cluster dissimilarity and average intrapersonal variability are given by clusters in **Table 8**. After the decomposition of the cluster S1, we end up with 11 groups more homogeneous in size, even if the clusters S1-1 and S2 stay a little bigger. As expected, the average intra-cluster dissimilarity within each cluster is lower than the one calculated with the total of the cards taken together (without clustering). However, we observe more variability between the sequences which belong to clusters S2, S1-1 (the two biggest groups) and S6 (the more motley one according to **Fig. 6**). The average intrapersonal variability is also by far higher for the cluster S6, composed of heterogeneous sequences, which is definitely the most irregular group. It is followed by the clusters S4 and S7, which gather users with longer periods of inactivity. Inversely, the smallest average intrapersonal variability is found for the cluster S1-1, composed of users with no vacation period, thus most of the time their behaviour is similar from one week to the next. Indeed, almost all their weeks belong to the same week type W1.

**Table 8** Results by cluster of sequences

| Cluster | S1-1 | S1-2 | S1-3 | S1-4 | S1-5 | S2 | S3 | S4 | S5 | S6 | S7 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Card proportion | 23.2% | 9.9% | 8.8% | 5.9% | 5.8% | 19.7% | 7.5% | 6.1% | 5.1% | 4.6% | 3.4% | 100.0% |
| Average intra-cluster dissimilarity | 2930 | 1490 | 1191 | 903 | 1233 | 4176 | 1370 | 1665 | 1428 | 2029 | 951 | 22943 |
| Average intrapersonal variability | 0.18 | 0.20 | 0.19 | 0.20 | 0.27 | 0.25 | 0.22 | 0.33 | 0.31 | 0.55 | 0.33 | 0.25 |

4.4 Comparison of the two user typologies

In this last section we will compare the user typology based on monthly use indicators (section 4.1) with the one based on week type sequences (section 4.3). First, two confusion matrixes are built, one with horizontal distribution (**Table 9**) and the other with vertical distribution (**Table 10**), in order to cross the membership of the 2850 sampled smart cards to each of the two typologies.

We begin with noting that the cards belonging to the cluster S1 (and its decomposition from S1-1 to S1-5) are mostly in the cluster C4. It is probably because S1 and C4 are respectively the two largest clusters in each typology. However, in the reverse distribution, the C4 members, along with the C3 members, are mainly represented in the cluster S1-1. They account for regular users who made the majority of their trips on business days, resulting in the observation of regular work weeks in their behaviour. Moreover, the cards of C1 and C2 are mainly found in the clusters S2 and S3, composed of quite regular users who travel on weekends too. This matching echoes the higher proportions of trips made in non-business day in these two clusters (see **Table 5**). More precisely, 56.1% of the card users in C1 (most frequent users) are in S3 (average intrapersonal variability: 0.22), whereas 51.9 % of the cards in C2 (second most frequent users) are found in S2 (average higher variability: 0.25). Therefore, it seems that the most frequent users are also the most regular ones at the intrapersonal level. The cards of S4 and S6 (the two least regular groups at the intrapersonal level according to the intrapersonal variability indicator) match with the ones in C6, so they correspond to both non-frequent and irregular users.

**Table 9** Distribution of the 11 clusters of users based on week type sequences in the 6 clusters of users based on monthly use indicators

| % of cards | | User typology based on monthly use indicators | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | C4 | C5 | C6 | Total |
| User typology based on week type sequences | S1-1 | 0.3% | 4.5% | 28.0% | 57.9% | 9.2% | 0.0% | 100.0% |
| | S1-2 | 0.4% | 1.8% | 12.5% | 64.8% | 18.9% | 1.8% | 100.0% |
| | S1-3 | 0.0% | 0.4% | 11.5% | 67.5% | 19.0% | 1.6% | 100.0% |
| | S1-4 | 0.0% | 1.2% | 8.3% | 61.3% | 26.2% | 3.0% | 100.0% |
| | S1-5 | 0.6% | 0.6% | 4.8% | 52.4% | 39.2% | 2.4% | 100.0% |
| | S2 | 2.3% | 22.1% | 34.2% | 22.3% | 14.1% | 5.0% | 100.0% |
| | S3 | 10.7% | 35.0% | 22.0% | 13.6% | 12.1% | 6.5% | 100.0% |
| | S4 | 0.0% | 0.0% | 0.0% | 1.1% | 4.0% | 94.9% | 100.0% |
| | S5 | 0.0% | 0.7% | 2.8% | 17.4% | 59.7% | 19.4% | 100.0% |
| | S6 | 0.0% | 0.0% | 0.8% | 0.0% | 15.3% | 84.0% | 100.0% |
| | S7 | 1.0% | 0.0% | 3.1% | 15.5% | 61.9% | 18.6% | 100.0% |

**Table 10** Distribution of the 6 clusters of users based on monthly use indicators in the 11 clusters of users based on week type sequences

| % of cards | | User typology based on monthly use indicators | | | | | |
|---|---|---|---|---|---|---|---|
| | | C1 | C2 | C3 | C4 | C5 | C6 |
| User typology based on week type sequences | S1-1 | 4.9% | 12.6% | 35.7% | 34.2% | 11.1% | 0.0% |
| | S1-2 | 2.4% | 2.1% | 6.8% | 16.2% | 9.7% | 1.3% |
| | S1-3 | 0.0% | 0.4% | 5.6% | 15.2% | 8.7% | 1.0% |
| | S1-4 | 0.0% | 0.8% | 2.7% | 9.2% | 8.0% | 1.3% |
| | S1-5 | 2.4% | 0.4% | 1.5% | 7.8% | 11.8% | 1.0% |
| | S2 | 31.7% | 51.9% | 37.1% | 11.2% | 14.4% | 7.3% |
| | S3 | 56.1% | 31.4% | 9.1% | 2.6% | 4.7% | 3.7% |
| | S4 | 0.0% | 0.0% | 0.0% | 0.2% | 1.3% | 43.5% |
| | S5 | 0.0% | 0.4% | 0.8% | 2.2% | 15.7% | 7.3% |
| | S6 | 0.0% | 0.0% | 0.2% | 0.0% | 3.6% | 28.8% |
| | S7 | 2.4% | 0.0% | 0.6% | 1.3% | 10.9% | 4.7% |
| | Total | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% |

Similarly, the cards of S5 and S7, characterized by a relatively long inactive summer period and a high level of intrapersonal variability too (average values of 0.31 and 0.33 respectively), are mainly in the cluster C5. This confirms that the less frequent and active card users are also less regular at the intrapersonal level and vice versa. In conclusion, with these confusion matrices we observe the same trends as in the section 4.2 which was based on several intrapersonal variability indicators. This validates our second user typology and our indicators.

Finally, two internal clustering validation measures were calculated to compare the quality of the two typologies. The Euclidean distance and the weighted Hamming distance were used respectively to evaluate these indicators. The Dunn's index gives similar results for the two user segmentations, even if the second typology based on week type sequences leads to a slightly higher value ($D_2 = 0.033$) than the first typology based on monthly use indicators ($D_1 = 0.013$). However, the Silhouette index is obviously higher for the first typology ($S_1 = 0.450\,;\, S_2 = 0.110$) and, according to the literature, this indicator often gives better results (Arbelaitz et al. 2013). This means the first typology based on monthly use indicators is better in terms of compactness and separation of the clusters. However, we argue that the second typology should not be discarded because it has a finer scale (week rather than month). What is more, it enables a more meaningful representation of the user clusters thanks to the week type sequences and the capture of several microphenomena (public holiday, vacation period).

## 5 Discussion and Conclusion

Using 51 weeks of smart card data from the Montreal transit operator, this paper provides analytical tools to study both inter and intrapersonal variability in public transit use. These tools are based on simple and comprehensive data mining methods, especially descriptive indicators, statistics and clustering algorithms. Sequential analysis is also used to consider the order of the travel patterns observed.

More specifically, the proposed analytical scheme unfolds in three major stages. First, data preprocessing was necessary to tackle the huge size of such a dataset: trip data were thus aggregated into "cards-year" and "cards-week" databases, summarizing the mobility of each card over the year or on every week of the year respectively. Secondly, interpersonal and intrapersonal variabilities were investigated. On the one hand, a typology of users was built to show differences between users in their frequency of use and activity ratio. Other temporal and spatial indicators were also calculated to support the analysis of this typology. On the other hand, a typology of weeks was produced to quantify to what extent the behaviour of each card user is similar from one week to the next. Thirdly, a sequence of week types was created for each user and served as a basis to obtain a second typology of users thanks to a weighted Hamming distance. The two user typologies obtained were finally compared and validated.

By applying this methodology to annual pass users with an amplitude of 12 months, this paper revealed that the more frequent and active users over the year tend to be more regular at the intrapersonal level too. Inversely, the occasional users have a more variable behaviour, but this type of users stand for a small minority among the annual pass users (and they are in part employees who were offered a free pass with their company). This confirms that the STM should manage to retain its users with an annual subscription since most of them are loyal and regular. Nowadays the annual pass is only the third most used product in Montreal, after the ticket book and the monthly pass, thus its popularity may still be improved. Moreover, we succeeded in detecting subtle differences between the annual pass users, for instance by identifying users who travel more on the weekend or don't travel on Friday. In an integrated pricing context, this finding could incite to create customized products, more adapted to these particular types of users.

Nevertheless, these atypical groups of users are relatively small. It is likely due to the choice of the fare studied since, as we said above, annual pass users are overwhelmingly frequent and regular passengers who use public transit to commute during business days. A greater diversity in fares may have led to a greater diversity of behaviours. For instance, it could be interesting to point out "occasional-regular" users who occasionally use public transit but in a predictable way. Further works will thus be conducted to apply the methodology with other fares. However, the research question behind this is the following: is it the transit product that influences

our behaviour or our behaviour that influences which pass we buy? The causal relationship is not clear and should be addressed.

Furthermore, in this paper we defined inter and intrapersonal variability at the monthly and weekly levels respectively, but it could be worth testing the sensibility of the proposed method with finer scales like the daily one. Finally, the spatial dimension should be more explored too, especially to measure spatial regularity.

**Acknowledgements:** This research is supported by the *Chaire de recherche du Canada sur la mobilité des personnes*. We also gratefully acknowledge the collaboration of the *Société de Transport de Montréal* for providing data and the help of Jean-Simon Bourdeau for data pre-processing.

## References

Adjengue L (2014) Méthodes statistiques: concepts, applications et exercices ; Luc Adjengue. vol Book, Whole. Presses internationales Polytechnique, Montréal

Agard B, Morency C, Trépanier M (2006) Mining public transport user behaviour from smart card data. IFAC Proceedings Volumes 39:399 - 404. doi:10.3182/20060517-3-FR-2903.00211

Agard B, Partovi Nia V, Trépanier M (2013) Assessing public transport travel behaviour from smart card data with advanced data mining techniques. Paper presented at the 13th World Conference on Transport Research, Rio de Janeiro, Brazil,

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recognition 46:243-256. doi:10.1016/j.patcog.2012.07.021

Batagelj V (1988) Generalized Ward and related clustering problems. Classification and related methods of data analysis:67-74

Briand AS, Côme E, Trépanier M, Oukhellou L (2017) Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transportation Research Part C: Emerging Technologies 79:274 - 289. doi:10.1016/j.trc.2017.03.021

Chu K, Chapleau R (2010) Augmenting Transit Trip Characterization and Travel Behavior Comprehension: Multiday Location-Stamped Smart Card Transactions. Transportation Research Record: Journal of the Transportation Research Board:29 - 40. doi:10.3141/2183-04

Cleophas TJ, Zwinderman AH (2011) Non-Parametric Tests. In: Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator. Springer Netherlands, Dordrecht, pp 9-13. doi:10.1007/978-94-007-1211-9_4

El Mahrsi MK, Come E, Oukhellou L, Verleysen M (2017) Clustering Smart Card Data for Urban Mobility Analysis. IEEE Transactions on Intelligent Transportation Systems 18:712-728. doi:10.1109/TITS.2016.2600515

Goulet-Langlois G, Koutsopoulos HN, Zhao J (2016) Inferring patterns in the multi-week activity sequences of public transport users. Transportation Research Part C: Emerging Technologies 64:1 - 16. doi:10.1016/j.trc.2015.12.012

Goulet-Langlois G, Koutsopoulos HN, Zhao Z, Zhao J (2017) Measuring Regularity of Individual Travel Patterns. IEEE Transactions on Intelligent Transportation Systems

Huang J, Xu L, Ye P (2015) Exploring Transit Use Regularity Using Smart Card Data of Students. In: Peng Q, Wang KCP, Liu X, Chen B (eds) ICTE 2015. Dalian, China, pp 617 - 625. doi:10.1061/9780784479384.080

Joh C-H, Arentze T, Hofman F, Timmermans H (2002) Activity pattern similarity: a multidimensional sequence alignment method. Transportation Research Part B 36:385-403. doi:10.1016/S0191-2615(01)00009-1

Joh CH, Timmermans H (2011) Applying Sequence Alignment Methods to Large Activity-Travel Data Sets Heuristic Approach. Transportation Research Record 2231:10-17. doi:10.3141/2231-02

Jones P, Clarke M (1988) The significance and measurement of variability in travel behaviour. Transportation 15. doi:10.1007/BF00167981

Kieu LM, Bhaskar A, Chung E (2014) Transit passenger segmentation using travel regularity mined from Smart Card transactions data. Paper presented at the 93rd Annual Meeting at the Transportation Research Board, Washington, D.C.,

Kitamura R, Yamamoto T, Susilo YO, Axhausen KW (2006) How routine is a routine? An analysis of the day-to-day variability in prism vertex location. Transportation Research Part A 40:259-279. doi:10.1016/j.tra.2005.07.002

Liu L, Hou A, Biderman A, Ratti C, Chen J (2009) Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. Paper presented at the 2009 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, Missouri,

Liu Y, Li Z, Xiong H, Gao X, Wu J Understanding of internal clustering validation measures. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on, 2010. IEEE, pp 911-916

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. doi:10.1016/j.trc.2013.07.010

Manley E, Zhong C, Batty M (2016) Spatiotemporal variation in travel regularity through transit user profiling. Transportation. doi:10.1007/s11116-016-9747-x

Morency C, Trépanier M, Agard B (2007) Measuring transit use variability with smart-card data. Transport Policy 14:193 - 203. doi:10.1016/j.tranpol.2007.01.001

Morency C, Trepanier M, Frappier A, Bourdeau JS (2017) Longitudinal Analysis of Bikesharing Usage in Montreal, Canada. Paper presented at the 96th Annual Meeting of the Transportation Research Board, Washington, D.C.,

Murtagh F, Legendre P (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification 31:274-295. doi:10.1007/s00357-014-9161-z

Nishiuchi H, King J, Todoroki T (2013) Spatial-Temporal Daily Frequent Trip Pattern of Public Transport Passengers Using Smart Card Data. International Journal of Intelligent Transportation Systems Research 11:1 - 10. doi:10.1007/s13177-012-0051-7

Ortega-Tong MA (2013) Classification of London's public transport users using smart card data. Massachusetts Institute of Technology

Pas EI, Koppelman FS (1987) An examination of the determinants of day-to-day variability in individuals' urban travel behaviour. Transportation 14:3

Raux C, Ma T-Y, Cornelis E (2016) Variability in daily activity-travel patterns: the case of a one-week travel diary. European Transport Research Review 8:1-14. doi:10.1007/s12544-016-0213-9

Saneinejad S, Roorda MJ (2009) Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. Transportation Letters-The International Journal Of Transportation Research 1:197-211. doi:10.3328/TL.2009.01.03.197-211

Schlich R, Axhausen KW (2003) Habitual travel behaviour: Evidence from a six-week travel diary. Transportation 30:13-36. doi:10.1023/A:1021230507071

Steinley D (2006) K-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology 59:1 - 34. doi:10.1348/000711005X48266

Strauss T, Michael Johan von M (2017) Generalising Ward's Method for Use with Manhattan Distances. PLoS ONE 12. doi:10.1371/journal.pone.0168288

Tao S, Rohde D, Corcoran J (2014) Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. Journal of Transport Geography 41:21-36. doi:10.1016/j.jtrangeo.2014.08.006

Ward JH (1963) Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58:236-244. doi:10.1080/01621459.1963.10500845

White P, Bagchi M, Bataille H, East SM (2010) The role of smartcard data in public transport. Paper presented at the 12th World Conference on Transport Research, Lisbon, Portugal,

Wilson WC (1998) Activity pattern analysis by means of sequence-alignment methods. Environment and Planning A 30:1017-1038. doi:10.1068/a301017

Xianyu J, Rasouli S, Timmermans H (2017) Analysis of variability in multi-day GPS imputed activity-travel diaries using multi-dimensional sequence alignment and panel effects regression models. Transportation 44:533-553. doi:10.1007/s11116-015-9666-2

Zhong C, Manley E, Arisona SM, Batty M, Schmitt G (2015) Measuring variability of mobility patterns from multiday smart-card data. Journal of Computational Science 9:125 - 130. doi:10.1016/j.jocs.2015.04.021