World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Assessing longitudinal stability of public transport users with smart card data

Mahnaz Moradi[a], Martin Trépanier[a] *

[a]Polytechnique Montréal and Interuniversity Research Centre on Entreprise Networks, Logistics and Transportation (CIRRELT)
P.O. box 6079, station Centre-Ville, Montréal, Québec, H3C 3A7, Canada

Abstract

Many public transit networks around the world use the smart card data which provides the information about the users. In this regard, several methods are developed by mostly applying clustering approaches to perform data segmentation and discover the pattern of users. This study addresses the applicability of the temporal segmented data identified in 18 clusters for measuring the stability of users' temporal habits as well as conducting descriptive analysis of the clusters, fare types and the days of the week to support the justification of findings. Each cluster contains users with their specific time and number of boardings. To understand whether the users are stable in the clusters, the sequential measurement based on the Euclidean distance between centres of the clusters, as the representatives of their members, is applied for each user over one month in this study. We ranked calculated measures to three different levels of high and medium stable or unstable using a histogram. The outcomes demonstrate the high stability of adult customers on three temporal routines, particularly regarding the days of the week. The users of the first and last working days of the week have a similar tendency in clusters' membership tracks and pretty the same proportion of stability levels, having the minimum high stable and the maximum unstable users. Regarding the fare types, we recognized that regular students have the same unstable frequency in spite of having a significantly less frequency than regular.

Keywords: Smart card, customer segmentation analysis, cluster stability, passenger tracking.

1. Introduction

The public transit agencies make decisions to meet their clients' needs and satisfaction by conducting a better exploration of their transit network and passengers' behaviours. For the sake of understanding how passengers behave regarding the spatiotemporal, spatial or temporal patterns, many research is conducted. The latter is of interest in this

* Corresponding author. Tel.: +1-514-343-7240; fax: +1-514-340-4173.
E-mail address: mtrepanier@polymtl.ca

study because temporal patterns are related with the different types of days such as working days, weekends and holidays as well as with the time of days such as morning or evening rush hours which affect the flow of the network heavily. The temporal data have the potential to be integrated with the other sources of data such as the fare type and weather, former of which is integrated to our study. To investigate this effect, the analysis of the smart card data which is enormously used in many cities is necessary. They provide the precise temporal information including date and time of transactions as well as many characteristics about the passengers and network, and it would be very helpful to analyse and discover this kind of data.

The data mining and machine learning statistical methods provide considerable tools in this regard, such as regression models, clustering and visual representation tools. The clustering methods were applied in the prior work of Ghaemi et al. (2017) on the same data set. In this study the results of their clustering is analysed by the help of visualisation tools.

In this study, we analyse one month of bus smart card data to discover the behavioural pattern of users. Specifically, we aim to discover the cluster membership's stability of passengers. These clusters elicit the coherent internal representation of users in terms of analogous temporal behaviour for each travel day associated with the corresponding user. Measuring the variability of clients in public transit network over a period of time is very helpful in the strategic transit planning and scheduling issues. The results demonstrate the high stability of most users according to the temporal activities in different particular working days, while high instability over weekends and holidays. Moreover, frequency analysis shows the high stability of regular adult commuters who belong to two clusters with the biggest portion of users. In addition, we found that more the users are on the transit network, the more they are stable.

This paper presents some related works and describes the methodology applied by Ghaemi et al. (2017) to achieve the clusters which we use in this study, followed by the description of the data used in this work. A descriptive analysis is conducted afterwards. In the methodology section, the method of measuring stability and dominant cluster's membership on each individual cardholder are described. Then, the results of applying the methodology on a case study over one month of the STO's smart card data on working days as well as the results of tracking the passengers over weekdays regarding the level of stability are shown. Finally, contributions and recommendations are presented.

2. Related Work

Smart card fare collection systems as one of the big sources of human mobility are used in several research works to characterise user behaviour. Many works on the use of smart cards in public transport are reported in Pelletier et al. (2011). Here we will focus on two components related to this paper: clustering methods as well as loyalty and stability assessment.

2.1. Smart Card Data Mining

The huge quantity of transactions produced by smart card fare collection systems makes them a very good candidate for the application of data mining techniques. Bagchi and White (2005) analysed the consistency of users' travel behaviour over time and produced targeted marketing campaigns to retain users in specific groups. Morency et al. (2007) applied the k-means technique to daily and weekly profiles to create groups of users based on their individual behaviours. They showed that membership is quite stable for weekdays. Ma et al. (2013) used a neural network approach to identify regular passengers, looking at the time and the location of transactions. The results indicate that their proposed rough set-based algorithm outperforms other commonly used data-mining algorithms in terms of accuracy and efficiency. The DBSCAN technique has also been used by Kieu et al. (2014) to measure similarity, based on the location of the first and last boarding stop of the users during the day. This has also been preferred by Zhang et al. (2016) to cluster users based on their spatiotemporal pattern. Kusakabe et al. (2014) used a Naïve Bayesian method to add more attributes to the smart card users by fusing with household surveys. Cats et al. (2015) used a hierarchical agglomerative method to identify and classify public transport activities in urban centres.

2.2. Stability Assessment

Since the first studies on smart card data in public transit, it became clear that users were not "stable" in their behaviours all over the days, weeks, months and years. After applying k-means technique, Morency et al. (2007) tried to assess the stability of the users by looking at the belonging to groups for 11 consecutive weeks, and showed that 80% of the users will not remain in the same cluster for all week. Trépanier and Morency (2010) studied the loyalty of users over a period of 40 months and found an annual turnover rate of 40% amongst the users of a mid-size Canadian transit network.

Cui et al. (2014) assessed the overall stability of groups of smart card users with a method composed of frequency distribution and grouping. Huang et al. (2015) looked at the regularity of students by analysing the activity rate related to the smart card transactions. Lee and Hickman (2011) emphasised on the need to use multiday smart card data to better understand the variability in the use of public transport network. To do this, Zhong et al. (2015) looked at correlation matrices between temporal patterns and selection of stops by users of Singapore.

3. Methodology

This study uses the clusters produced in Ghaemi et al. (2017) as an input for the public transport user stability assessment method. This section will first describe the work done to create the clusters, and then will present the indicators used to measure the stability of users.

3.1. The SCP Methodology

Ghaemi et al. (2017) propose a projection technique called semicircle projection (SCP) to transform a high-dimensional binary vector into a three-dimensional (x,y,z) feature vector that lays out the hidden temporal patterns. This is useful to classify public smart card users based on their daily use of the network. The left part of Table 1 shows the raw data used for the analysis. In the 24-hour vector, each variable represents whether a specific user, on a specific day, performed a smart card transaction in the public transport network. This is called a card-day. A same card can have multiple card-day within a month. This configuration does not lead to a clear classification with traditional k-means technique using Euclidean distance because the method cannot differentiate between the times of the day where the transactions occurred, as shown in the source paper. They showed that the SCP method outperforms the other state-of-the-art time series distance measurements such as cross-correlation distance, and autocorrelation-based dissimilarity distance in performance and computational complexity.

Table 1. 24-hour binary vector transformed to 3D vectors.

User	Day	1	2	3	4	5	6	...	24
X1	1	1	0	0	1	0	0	...	0
X1	2	1	0	0	1	0	0	...	0
...	...	...	...	...	...	...	...	...	...
X1	30	0	0	0	0	0	0	...	1
X2	1	0	1	1	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...

User	Day	X	Y	Z
X1	1	0,0576429	0,8427106	0,3478260
X1	2	0,0576429	0,8427106	0,3478260
...	...	...	...	...
X1	30	0,0576429	0,8810278	0,4347826
X2	1	0,0776429	0,8427106	0,4547826
...	...	...	...	...

The right part of Table 1 is obtained through the transformation:

$$\left[x_i = r_i \sin\left(\frac{\pi}{Ln_i}\sum_{j=1}^{L}\theta_{ij}\right), \quad y_i = r_i \cos\left(\frac{\pi}{Ln_i}\sum_{j=1}^{L}\theta_{ij}\right), \right.$$
$$\left. z_i = \sqrt{\frac{1}{L-1}\left\{\sum_{j=1}^{L}\theta_{ij}^2 - \frac{(\sum_{j=1}^{L}\theta_{ij})^2}{L}\right\}} \right].$$

(1)

where x coordinate represents the number of trips, the y coordinate represents the average time of trips, and the z-axis shows the time variability of the trips to capture the standard deviation of the timestamps. The number of boardings for the ith user-day is $n_i = \sum_{j=1}^{L}X_{ij}$ that is the number of unit elements in the vector X_i, L = 24 denotes the number of time intervals, and converging radius $r_i = (1 + 1/n_i)^{n_i}$ to renormalise the half circles for long binary sequences, as well as Θ representing the angle on the x-axis.

Then, a hierarchical clustering algorithm is deployed to elicit the coherent internal representation of users in terms of analogous temporal behaviour. In this paper, our starting point is the set of 18 clusters obtained from the method. Please refer to Ghaemi et al. (2017) for further details.

3.2. Measuring (In)stability of Cluster's Membership

In this section, we explore the methodology of measuring the stability of users in clustering membership over the period of use. This method is based on selecting centroids and measuring the Euclidean distance between them which is extracted from Leskovec et al. (2014).

We select to calculate the sequential instability of users. Sequential instability will look at cluster changing over card-days for each individual. The changes in clusters are weighted using the distance between them (0 if no change). In other words, from the first day of use to the second one, second to third and so on. We sum these distances for each user.

Vector ID_i: 1st Cluster → 2nd Cluster → 3rd Cluster → 4th Cluster → 5th Cluster →… N_i

The sequential instability for cluster appearance of a Card ID during a month is the following equation:

$$WSI_i = \left(\sum_{j=1}^{N_i} distance[vector[ID_{i,j}], vector[ID_{i,j+1}]]\right)/N_i$$

(2)

Regarding the card-day clusters as vectors, we can move from one travelled day (j) to the next (j+1) for ID, for i=1 to 26198 (the number of cards). The amount of usage significantly influences the stability and users commute between 1 and 20 days (10 removed days are weekends and holidays). So, in the next step, the measures are divided by the number of travelling days N_i corresponding to each user to obtain the "weighted instabilities".

Note that some users never changed their clusters or were present just one day on the transit network. So their instability measure is zero and could not be divided by the number of travelling days. The issue was solved by adding a constant (+1) to all the measures before being divided. This helps to distinguish the single commute travelling day users from highly stable ones. Consequently, lower the measure, more the user is stable in clusters membership. In other words, if a passenger does not change its cluster or change with the closest ones considering the number of days present on the network, she or he is a stable passenger.

Finally, using a histogram and a scatter plot of instability measures, we rank the instability to three "stability" categories defined as:

- High stable users: instability equal or less than 0.3 [.05, .3)
- Medium stable users: instability between 0.3 and 0.55 [.3, .55]
- Unstable users: instability greater than 0.55 (.55,1.15]

After measuring the stability of users, it is interesting to find the dominant cluster for each user and explore its relationship with the stability ranks. To this end, we look for the maximum clusters membership corresponding to each user. This means that if a card is in cluster #4 40% of the time, 50% of the time in cluster #3 and the remaining

in cluster #1, the dominant cluster for this user will be cluster #3. Cards with equally balanced cluster memberships are removed from this part of the analysis.

The same methodology is applied over the days of the week to explore the stability of passengers on five days of the week. The cluster membership for its corresponding users who belong to one of the three stability categories is traced by the alluvial diagrams presented in section 5.

4. Descriptive analysis of clusters

In this section, we present the case study, the obtained clusters and their descriptive analysis.

4.1. Case study

The SCP method has been tested on the data from the *Société de transport de l'Outaouais*, a mid-size authority (300 buses and 220,000 inhabitants); over a one-month period in April 2009 (data are gathered from 900,936 transactions, with 26,198 unique cards and 416,076 card-days). For each transaction, the following attributes are present:
- Date and time of the boarding transaction;
- Card number and fare type;
- Route number and direction;
- Vehicle and driver numbers;
- Stop number at boarding.

Note that for the sake of security and privacy purposes, card numbers are encrypted so that all user information is completely anonymous.

4.2. Clusters

Figure 1 presents the average daily distribution of smart card transactions for the 18 clusters that were found in the Ghaemi et al. (2017) study. The labels are meant to be informative here. The "Regular" clusters (1, 2, 5, 13, 14, 16) regroup users that have pendular trips during the day, at different time frames. Cluster 4 represents users that could be labelled as regular, but with a much longer day of activities than the others. Cluster 3 is mostly characterised by transactions made between the peak hours, while cluster 12 regroups pendular trips made after the afternoon peak hour. Members of cluster 7 are very active all through the day, while cluster 18 regroups inactive users (very little use of the network). The other clusters regroup people that perform single trips at different times of the day. Let us remember that these profiles are linked to card-days. This means that the same card could be, for example, part of cluster 2 for 20 days of the month, and member of cluster 11 for 5 days, etc.

By applying the 3D scatter plot on the projection of the binary vector of timestamps onto the semicircle space, in Figure 2, we obtain the user's temporal habits set on x-axis. In other words, moving from left to right covers clusters associated with the early morning to late nights. And going up on y-axis, contains clusters with more boarding. The z-axis represents the frequency.

This figure points out that the peak of the half-circle has the highest frequency, which contains regular commuters clustered in 2, 5 and 13 who usually take public transport as their routine schedule during the month. While, the early birds (cluster 15) and the night people (cluster 17) are single-trip users on the two opposite ends with different temporal usage behaviours.
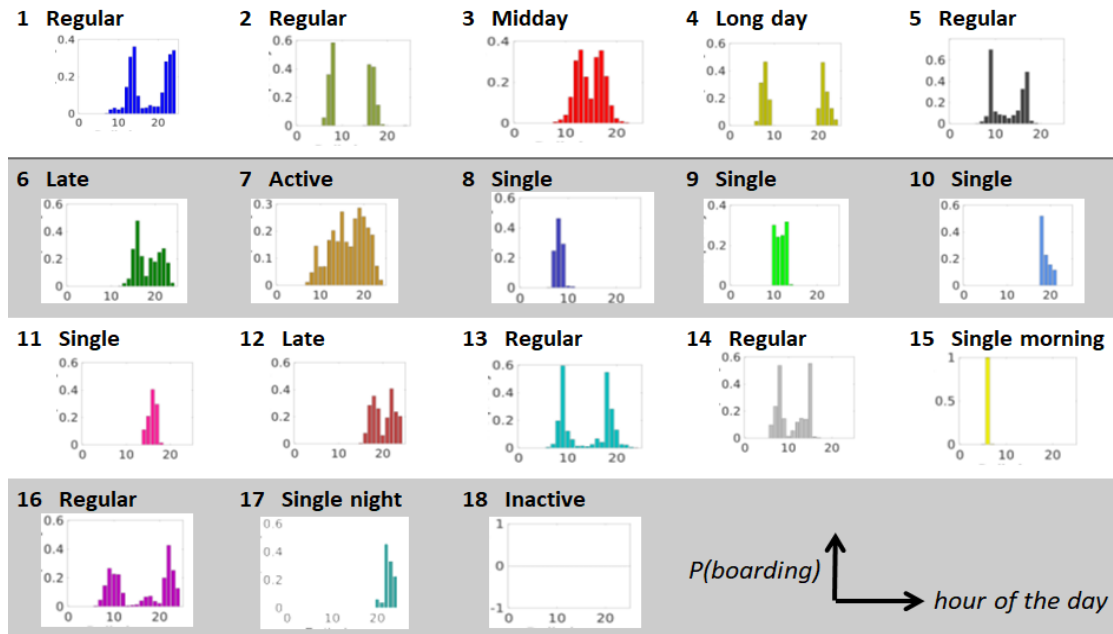
Fig. 1. Daily transaction profiles of the 18 clusters proposed by Ghaemi et al. (2017).
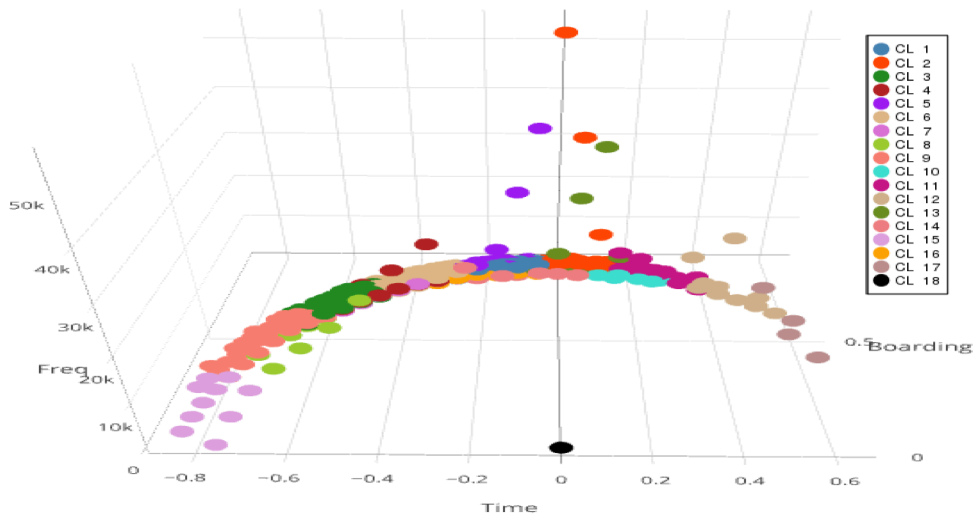

Fig. 2. Frequency distribution of 2D scatter plot of projected data.

4.3. Daily Distribution of Clusters

We demonstrate the users' cluster distribution and card type distribution over the days of April 2009 in Figure 3. It demonstrates that the largest presence in transit network is on working days and the lowest is on weekends as well as two holidays 10th (Good Friday) and 13th (Easter Day) April 2009. To have more interpretative results in the coming sections, we will work on regular weekdays.

The figure shows that regular clusters 2, 5 and 13 are the most frequent during weekdays. Interestingly, the 4th and 5th place in terms of the number of card-days switches frequently between clusters 4 (Midday) and 12 (Late).
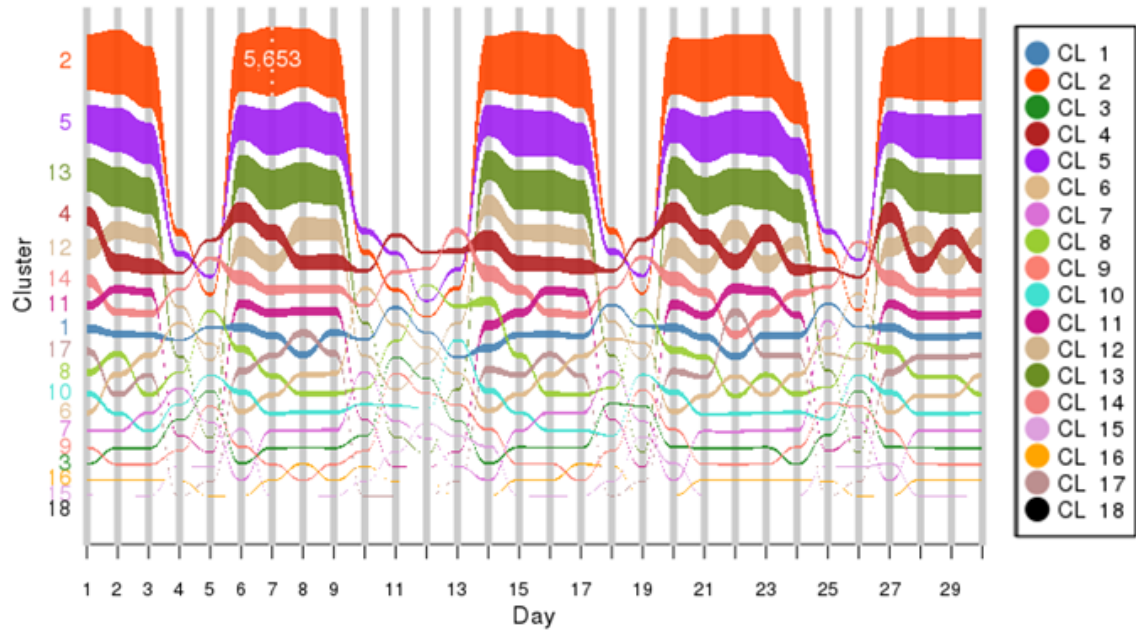
Fig. 3. Frequency distribution of the 18 clusters (1 to 30 April 2009).

This kind of figure remains anecdotal because it does not show the individual changes that happened to the card during the month. This is why we will try to assess the stability with indicators in the next section.

5. Stability assessment

In this section, we will present results on instability and dominant cluster indicators, as well as their weekly and monthly variations.

5.1. Instability

By applying the methodology explained above, the weighted sequential instabilities are found and ranked, showing in figures 4a) and 4b). Colours represent different levels of stability. Green, blue and red covering highly stable, medium stable and unstable levels, respectively.

The instabilities range between [0.05, 1.15]. Lower the instability, more the user is stable in clusters membership. In other words, if a passenger does not change her/his cluster or changes to the closest ones, considering the number of travelling days, she/he is a stable passenger. Consequently, users with the instability equal to 0.05 are the most stable users. In contrast, the users having instability of 1.15 are the most unstable users. They are classified into three categories: highly stable, medium stable and unstable users, each including 50, 40 and 10 percent of users, respectively, selected by the histogram in figure 4b).
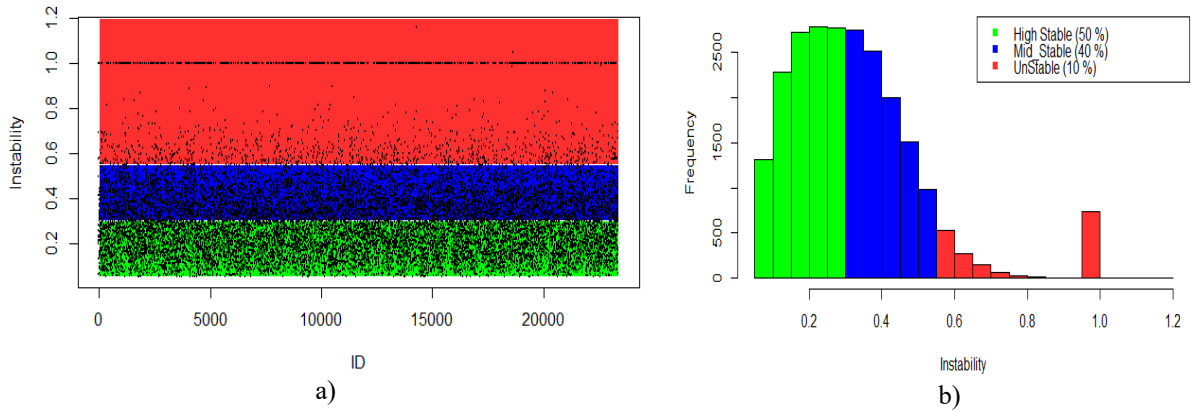
Fig. 4. Instability indicator. a) distribution amongst users, b) frequency histogram

The ranking is done by the help of histograms and scatter plots. The lowest instability (0.05) belongs to the 0.07% of cardholders who used their card more than once and never changed their cluster. On the other hand, the 3% of users are the single-travelling day users over 30 days with instability equal to "1", so there is no pattern and they are considered as unstable users (the straight line in red region in figure 4a) and single red column in figure 4b).

5.2. Dominant cluster

As described before, the dominant cluster corresponding to each cardholder is recognized and represented in figure 5.

As expected, the most frequent clusters count the most dominant clusters, i.e. clusters 2, 5, 13, 4 and 12. Obviously, the regular pendular commuters (cluster 2) have the biggest portion on all levels. On the other hand, two commuters behave differently before noon (cluster 5) and after noon (cluster 13). Cluster 13 is less unstable and more frequent on high stable level as well as half of mid stable where the cluster 5 is passed. Moreover, high intervals of instability of all clusters are around 1 except cluster 9 which contains the highest instability.
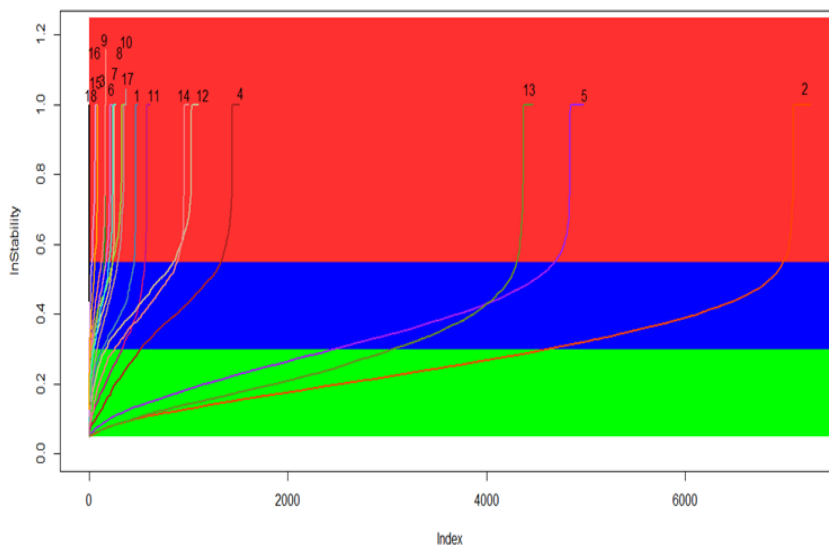

Fig. 5. Dominant clusters. Cumulative distribution vs. instability.

The impact of the number of travelling days regarding dominant clusters and stability levels is presented in Figure 6. It is evident that the number of travelling days influences highly stable level dramatically. In other words, a big portion of users present on public transit for more days and stay highly stable. They are members of clusters 2, 5 and 13. They show the same tendency but more moderate on the medium level. In contrast, the unstable level is affected by the single-travelling day users. The single-travelling day users, which mostly contain cluster 2 members, are the most frequent among unstable users
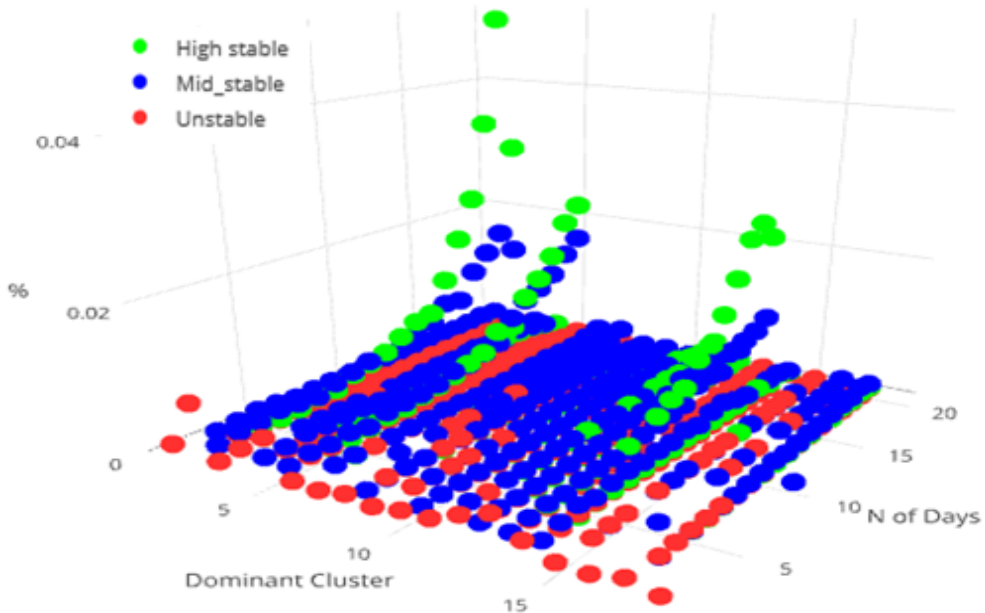

Fig. 6. Dominant clusters. Instability vs. the number of days travelled.

5.3. Trace Over workdays

In this section, we trace the flow of cluster's membership for each user over working days. In Figure 7, flow diagram of all clusters over all working days is shown on 3 stability levels. First level represents the highly stable users (green colour) who stay in the same cluster or at least switching to very close clusters. In contrast, the red colour shows the unstable users' trace who change their clusters not with close ones frequently or use their card just once. We assigned these users as unstable because they are singular and have no pattern. They are not removed in order to remain observable in the clusters. The other level, medium stability, shows the trace of users with the pretty close assigned clusters coloured in blue. The block number and size are the number and size of each cluster on a specified day. Number 0 means inactive users.

The flow diagram over 20 days does not help to describe the traces in detail. For example, Friday 24th of April, contains the most inactive users or the cluster 2, 13 and 5 which are the most dominant clusters causing the high stability but for the other levels or clusters it is not so clear. Consequently, for the reason of interpretability and exploring the stability over weekdays, the flows are traced over weekdays and weekends in the next section.
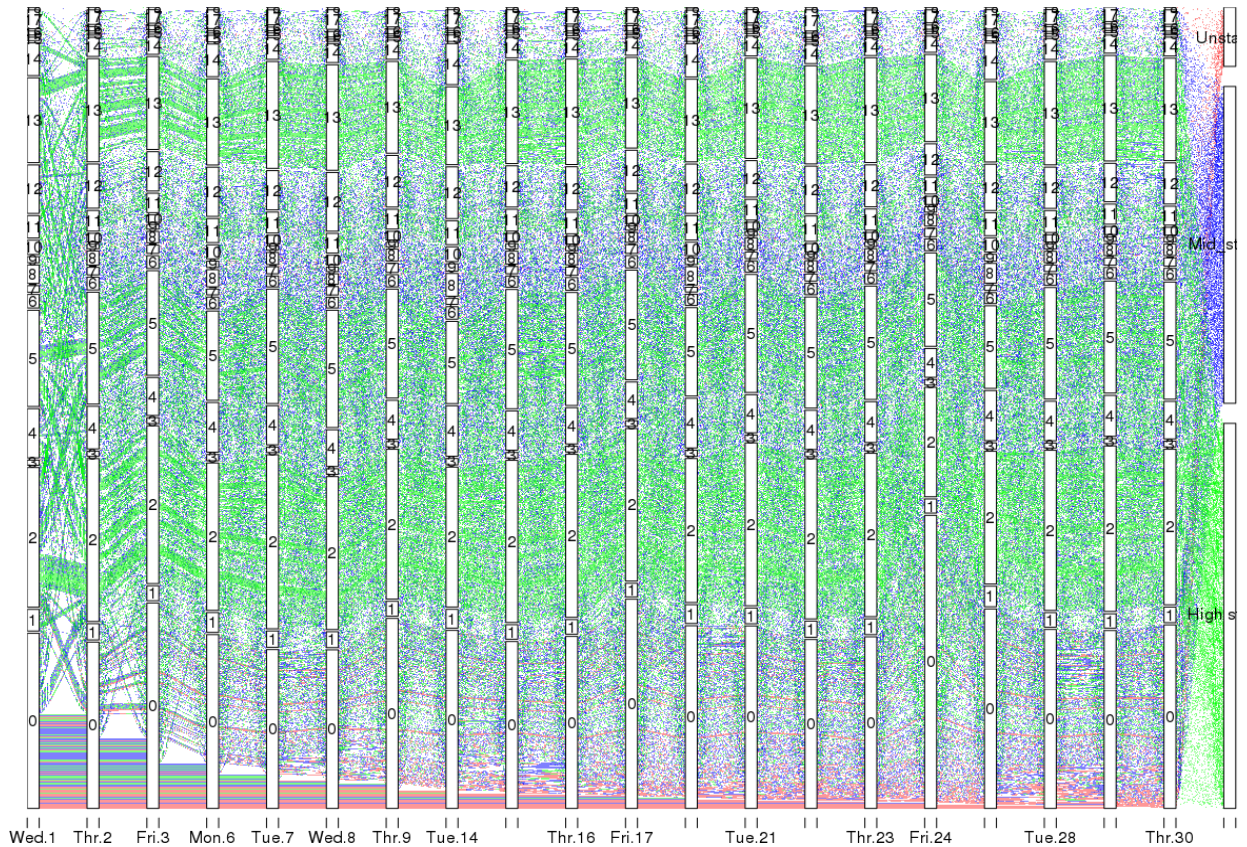
Fig. 7. Flow diagram over 20 work days.

5.4. Weekly Stability

In this part, we trace clustering membership over the weekdays: Mondays, Tuesdays, Wednesdays, Thursdays and Fridays as presented in figure 8. The cumulative distributions of weighted instability over each weekday show the similar trends of Mondays and Fridays as well. The other three days have similar trends with instability intervals [0.2, 1.2] resulted from about 22500 unique IDs.

Moreover, Mondays and Fridays have the greatest unstable users (20% and 23%, respectively) and smallest highly stable users portion (42% and 47%, respectively) in comparison with the other days. Thursdays have the highest stable users (68%) and the lowest unstable (7%). Obviously, the number of single travelling-day users on the first and the last working days is significantly high (the straight red line in figures 8b) and 8f)).
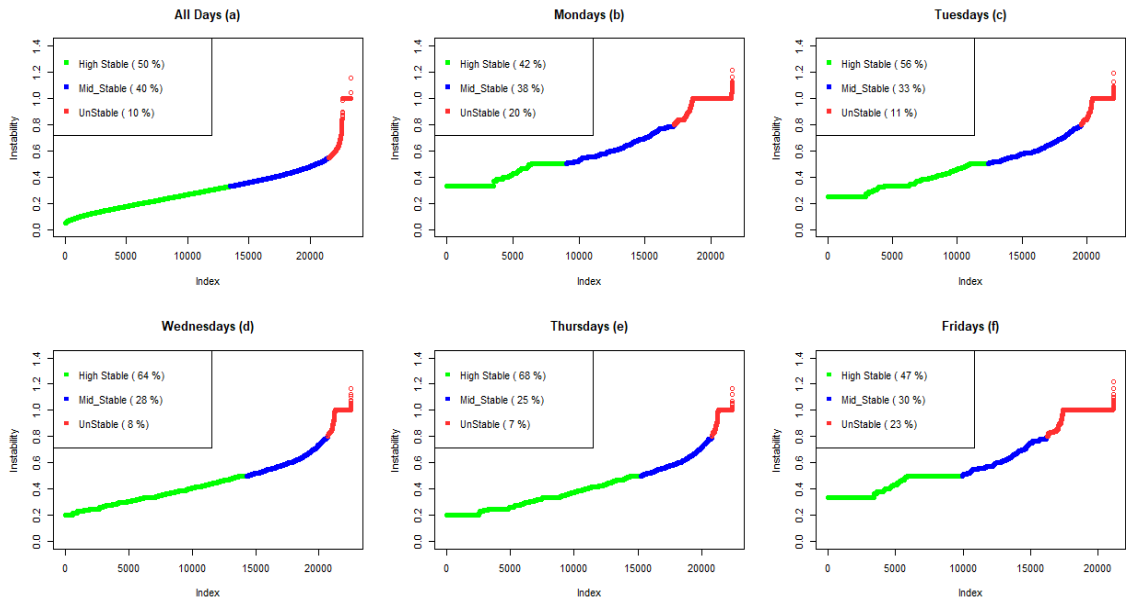
Fig. 8. Cumulative distribution of instability on working days

6. Conclusion

6.1. Contributions

This study proposed a method of measuring the stability of temporal habits of public transport users based on the smart card data. The temporal habits or temporal clusters are taken from another study on the same dataset. In this work, we measured the sequential stability of the cluster's membership for each user over one month and ranked them to three different levels. In addition, the stability indicators are computed and visualised in days of week to identify the level of stability and corresponding clusters in days of the week. It shows when the cluster's membership changes and how influences the level of stability. Moreover, a descriptive analysis is developed to support our findings such as justifying the removal of weekends and holidays.

The empirical findings in this study are critical for bus service managers. First, the findings proved that customers were highly stable on three temporal routines, specifically, regarding the days of the week. The first and the last working days contain the greatest portion of unstable users, specifically, the single commute travelling day users are witnessed over Mondays and Fridays. In contrast, the users in the other three working days are mostly the highly stable ones. Thus, an effort to increase bus lines over these timestamps will raise customer satisfaction. Secondly, the first day after long weekends has a moderate difference in cluster's membership with other similar working days which demotes the highly stable users to the second level. Analysing the data with several long weekends and holidays would provide the best understanding of clients' temporal variability as well as better bus service optimisation. In addition, users are most inactive on Sundays and passengers use the transit network at most once on average over all weekends and holidays.

6.2. Limitations

Possibly the most significant limitation is the different frequencies of cluster's membership for each user. It makes the unweighted selection of the dominant clusters from a variant interval. To slightly improve this effect, we remove the users having the same maximum frequency of dominant clusters. The other limitation concerns the inactive days.

As explained in section 6.3., with the aim of reclustering the users, the distance between an assigned cluster day and inactive day is unknown. It may help to consider this distance as infinite.

6.3. Future Works

The above findings were extracted from a single case study over one month; thus, a similar study conducted on a larger sample size, especially with complete weeks, is required for future effort to strengthen the findings of the present study. Furthermore, the users assigned with different temporal clusters have the potential to be reclustered using gene clustering method. In other words, each vector of cluster's membership corresponding to an individual over a month is considered as a chromosome and the clusters as genes. The distances between each user (chromosome) are the total distances between their clusters (genes) which would be the same distances used in this study to measure the stability of each user. But in our case, each user is measured with her/his behaviour in mind. So, to compare each user's behaviour to the others', aforementioned method may provide more interesting results.

Acknowledgements

References

Agard, B., Partovi Nia, V., & Trépanier, M. (2013). Assessing public transport travel behaviour from smart card data with advanced data mining techniques. In World Conference on Transport Research (Vol. 13, pp. 15-18).

Bagchi, M., White, P.R. (2005). The potential of public transport smart card data. Transport Policy 12, 464–474.

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

Cats, O., Wang, Q., Zhao, Y. (2015). Identification and classification of public transport activity centres in Stockholm using passenger flows data. Journal of Transport Geography, 48, 10-22

Cui Cl., Zhao Yl., Duan Zy. (2014), Research on the Stability of Public Transit Passenger Travel Behavior Based on Smart Card Data, 14th COTA International Conference of Transportation Professionals, July 4-7, Changsha, China.

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.

Huang, J., Xu, L., Ye, P. (2015). Exploring Transit Use Regularity Using Smart Card Data of Students. Fifth International Conference on Transportation Engineering, September 26–27, Dailan, China.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.

Kassambara, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Vol. 1). STHDA.

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

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., Hickman, M. (2011). Travel Pattern Analysis Using Smart Card Data of Regular Users. Transportation Research Board 90th Annual Meeting, Washington DC.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets. Cambridge university press.

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

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

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.

Trépanier, M., Morency, C., (2010). Assessing transit loyalty with smart card data. In: Presented at the 12th World Conference on Transport Research, Lisbon, Paper No. 2341.

Zhong C., Manley E., Müller Arisona S., Batty M., Schmitt G. (2015), Measuring variability of mobility patterns from multiday smart-card data, Journal of Computational Science, Volume 9, pp.125-130.

Zhang, J., Zhao, C. Tian. C. Xu, X. Liu, Rao, L. (2016), Spatiotemporal Segmentation of Metro Trips Using Smart Card Data, IEEE Transactions on Vehicular Technology, vol. 65, no. 3, pp. 1137-1149.