CIRRELT

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

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Temporal Clusters Analysis of Public Transit Passengers Using Smart Card Data

Mahnaz Moradi
Martin Trépanier

June 2018

CIRRELT-2018-28

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

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

www.cirrelt.ca

Temporal Clusters Analysis of Public Transit Passengers Using Smart Card Data

Mahnaz Moradi, Martin Trépanier*

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Mathematical and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

Abstract. Many public transit networks around the world use the smart card data which provides the information about the users. With the aim of uncovering the users' behaviours, there has been an increased interest in analyzing their temporal activities, in recent years. In this regard, several methods are developed mostly applying clustering approaches to perform data segmentation and discover the pattern of users. The next step is retrieving the information from the clusters, such as the variability of a user in cluster's membership over a period of time. This study addresses the applicability of the temporal segmented data identified in 18 clusters for measuring the stability of temporal habits of users as well as conducting descriptive analysis of clusters, fare types and the days of week to support the justification of findings. Each cluster contains users with specified time of day and number of boarding. To understanding whether the users are stable in the clusters, the sequential measurement based on the Euclidean distance between centers of clusters, as the representative of their members, is applied for each user over one month in this study. Then, the stabilities are ranked to three different levels of high and medium stable or unstable by the help of a histogram. The data and stability levels have the potential to track and visualize users, specifically, in days of week. The outcomes demonstrate the high stability of adult customers on three temporal routines, specifically, regarding the days of week. The users of the first and last working days of a week have the similar tendency in cluster's membership tracks and pretty same proportion of stability levels, having the minimum high stable and the maximum unstable users. And we discover that the most single-traveled day users commute in these two days. On the other hand, the other three working days show similar trends containing the maximum high stable users. In addition, tracking the users shows the moderate impact of long weekends as well as first and last day of month on the level of stability. Analyzing the weekends and holidays, shows no pattern and users are present on network on average once over these days. Furthermore, more than 70% of users are present on the network more than 14 days and interestingly, they are categorized in high and medium stable levels. Regarding the fare types, we recognized that regular students with the significant fewer frequency than regular adults but tending to have the same unstable frequency.

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

Acknowledgements. The authors wish to acknowledge the supporters of this study, which are the Société de Transport de l'Outaouais, the Natural Science and Engineering Research Council of Canada and Thalès Research and Technologies.

---

* Corresponding author: Martin.Trepanier@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2018

1 INTRODUCTION

The public transit agencies make decisions meet their clients' needs and satisfaction by exploring better their transit network and passengers' behaviours. For the sake of understanding how passengers behave regarding to the spatio-temporal, spatial or temporal patterns, many researches are conducted. The latter is of interest in this 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 network heavily. The temporal data have the potential to be integrated with the other sources of data such as the fare type and weather that the former was 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 analyze 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 analyzed by the help of visualization tools.

In this study, we analyze one month of bus smart card data to uncover 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 correspond 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 two commuters regular adults who have 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. Next, a descriptive analysis is conducted. 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

The most recent research focus on the analysis of spatial data to uncover the behavioural pattern of users, regardless of the temporal features of the data. Therefore, the smart card data as one of the big data sources from human mobility is used in several research works. We describe some of them conducted regarding temporal aspect, in next part.

2.1 Smart card data mining

A method is developed by Briand et al. (2017) to regroup passengers based on their temporal habits by a Gaussian mixture model. They showed the stability of users over a five-year longitudinal analysis on the dataset collected by the Société de Transport de l'Outaouais (STO) who provided the data set of this work, too.

There are some obstacles to use temporal data, such as the high-dimensional vector of hourly usage associated with each smart card. Ghaemi et al. (2017) propose a projection technique called semicircle projection (SCP) which is able to transform such high dimensional binary vector into a three-dimensional (x,y,z) feature vector that lays out the hidden temporal patterns. Then, they applied a hierarchical clustering method on the projected data to discover homogeneous groups of users who have the same temporal manner. Their novel method is described in this chapter, as their results provide the data for the present project. Another work consists of applying the k-means clustering method on the projected data with SCP in Agard et al. (2013). Their study over a one-year period of data revealed the three categories of travel behaviour among the users. However, the hierarchical method outperforms the k-means, because of producing a visual guide in the form of a binary tree, known as dendrogram. In addition it requires little prior knowledge, except for a dissimilarity measure.

2.2 The SCP methodology

In the SCP methodology, the temporal data are encoded into a 0−1 vector whereas 24 binary vector associated with the daily hours. In this vector, occurrence of 1 at a specific index represents the usage of smart cards at the corresponding hour.

Table 1 : Transformation to projected data

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
...	...	...	...	...

Then, the transformation to a three-dimensional vector for each user-day is performed by the formulas below. The reduced data in the new space are written as

Equation 1 : Reduced data formula (source: Ghaemi et al. 2017)

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

The 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 boarding for the *i*th user-day as $n_i = \sum_{j=1}^{L} X_{ij}$ 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 renormalize the half circles for long binary sequences, as well as Θ representing the angle on x axis. A schematic graph in Figure 1, visualize the projected data.

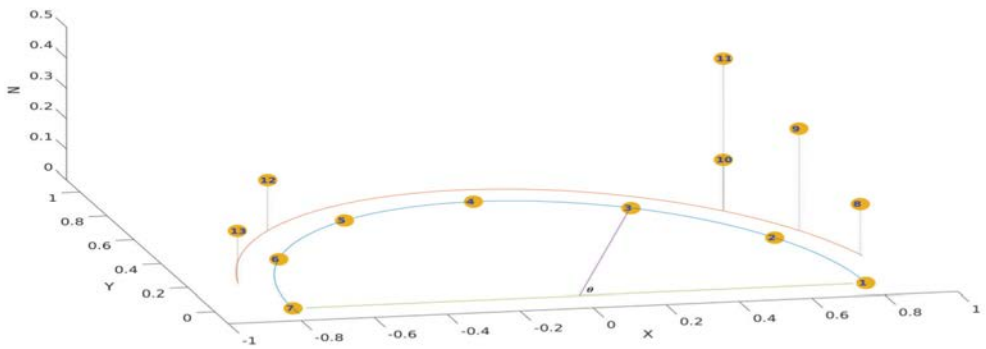

Figure 1 : 3D scatter plot of projected data (for 13 users) (source: Ghaemi et al. 2017)

They show 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. Then, they deploy a hierarchical clustering algorithm to elicit the coherent internal representation of users in terms of analogous temporal behaviour. The 18 clusters are distinguished depending on the determined similarity of clusters (See Figure 2). These clusters are categorized to the single trip, regular commuters, late commuters, long and midday, active and inactive users, showing in Figure 3.

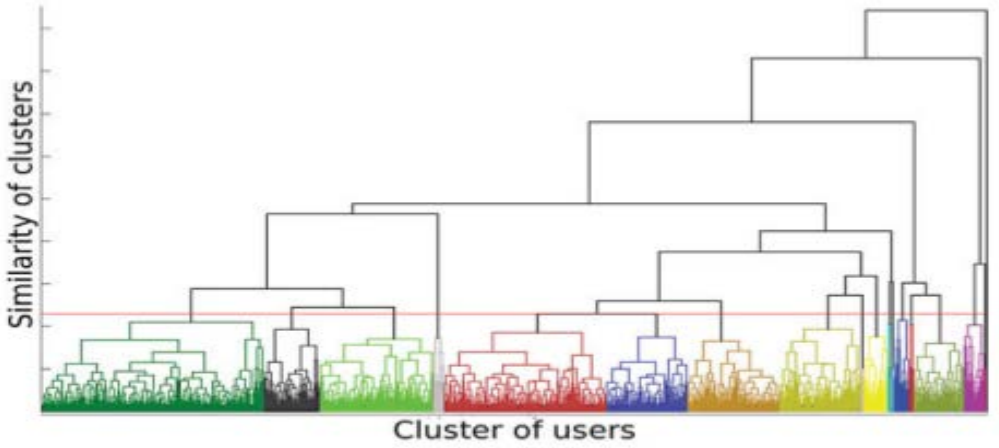

Figure 2 : Dendrogram of clusters (source: Ghaemi et al. 2017)

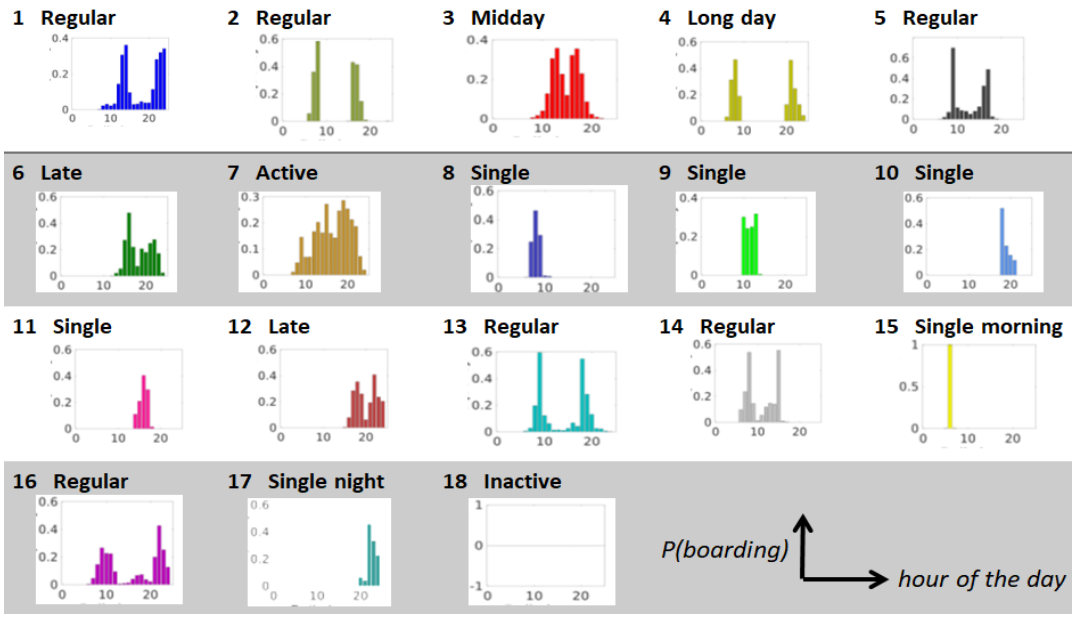

Figure 3 : daily transaction profiles of the 18 clusters (source: Ghaemi et al. 2017)

2.3 Data

The SCP method has been tested by Ghaemi et al. (2017) on the STO mid-size authority (300 buses and 220,000 inhabitants); over one-month period in April 2009 (data are gathered from 900,936 transactions, with 26,198 unique users and 416,076 card-days). For each transaction, the following attributes are present:

- Data and time of the boarding transaction;
- Card number and fare type;
- Route number and direction;
- Vehicle and driver numbers; and
- 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.

In this project, we use the clusters found by Ghaemi et al. (2017) and push further the analysis of the results.

3 DESCRIPTIVE ANALYSIS

In this chapter, the descriptive analysis is performed on the 18 clusters and the related attributes such as the distribution of clusters by projected data, fare types and days. The results of this chapter support the findings achieved by applying the methodology in next chapters.

3.1 Projected data and clusters

Visualization of the projected data taken from Equation 1, helps better understanding of clusters. The following figure is drawn by (x,y,frequency) without z to show the location of clusters according to time and boarding. The complete graphical representation (x,y,z) of projected data is shown in Figure 13 in Chapter 4.

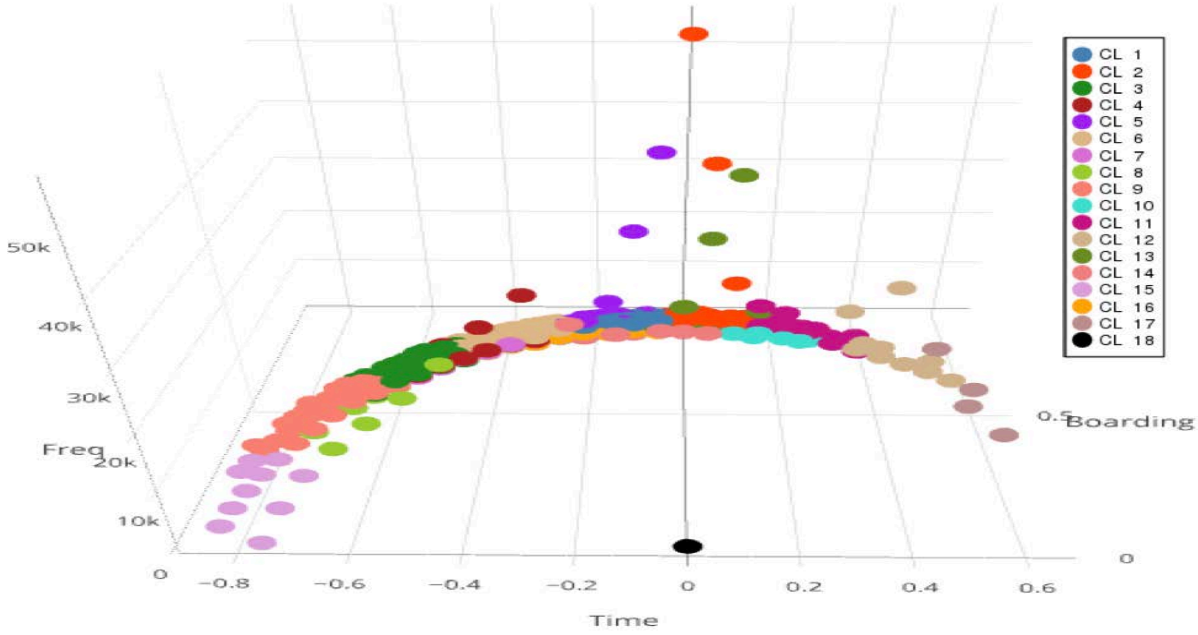

Figure 4 : Frequency distribution of 2D scatter plot of projected data

Applying the 3D scatter plot on the projection of the binary vector of timestamps onto the semicircle space, in Figure 4, turns out the user's temporal habits set on x-axis. In other words, moving from left to right covers clusters assigned from early morning to late night. And going up on y-axis, contains clusters with more boardings. The z-axis represents the frequency.

This figure and Figure 3 point out 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, frequently. While, the early birds (cluster 15) and the night persons (cluster 17) are single-trip users on the two opposite tails with different temporal usage behaviours. The pie chart in Figure 5 (Card-day cluster portions) confirms that the most frequent clusters are the ones on the pick of the 3D projected data, i.e. clusters 2, 5 and 13.

The clusters 8 and 9 are single morning users, while single afternoon and evening users are identified in clusters 11, 10.

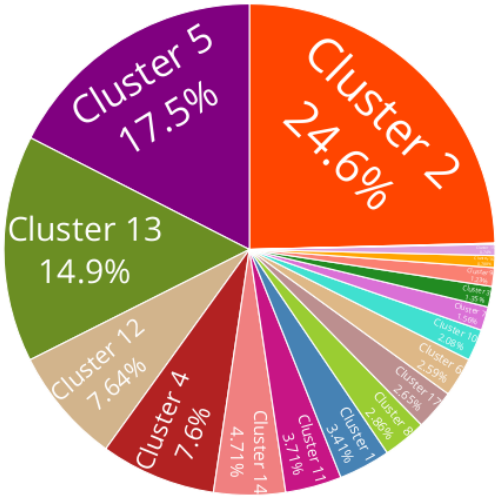

Figure 5 : Cluster Portions

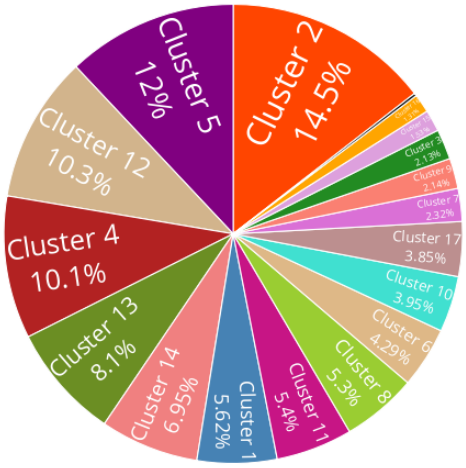

Figure 6 : Unique ID's Cluster Portions

As well, midday and long day are presented in clusters 3 and 4. Clusters 6 and 12 are the late commuters. And clusters 1, 14 and 16 contain the rest of regular users with different temporal behaviours.

Finally, the active users shown in clusters 7 and the last cluster is a singleton datum at origin (0, 0) coloured by black without any trip covered by public transport.

3.2 Fare types

In addition to the timestamp data, the fare types are taken into account to explore the type of passengers member of each cluster. The fare types are defined by the data provider (STO) in 24 different types characterized by the different group of people regarding their subdivisions regular, express and so on as shown on the legend in Figure 7.

Fare type distribution presented in Figure 7 illustrates that the highest percentage of users correspond to regular adults, regular CF and express adults, respectively. And the lowest percentage belongs to regular summer students, inter-zone campus and inter-zone UQO (*Université du Québec en Outaouais*), respectively.

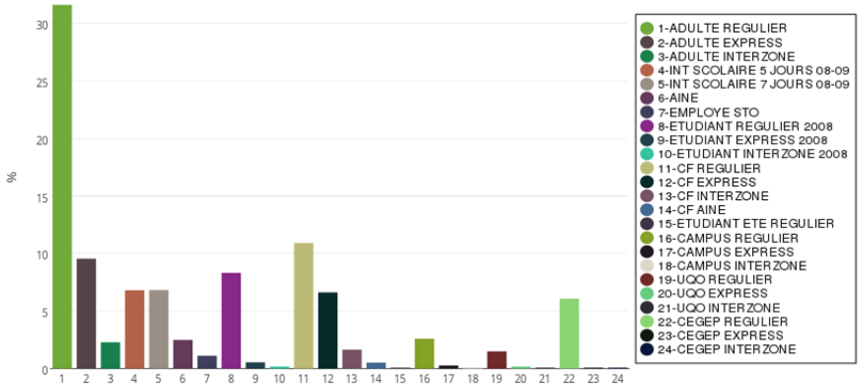

Figure 7 : Unique ID's Card Type Portions

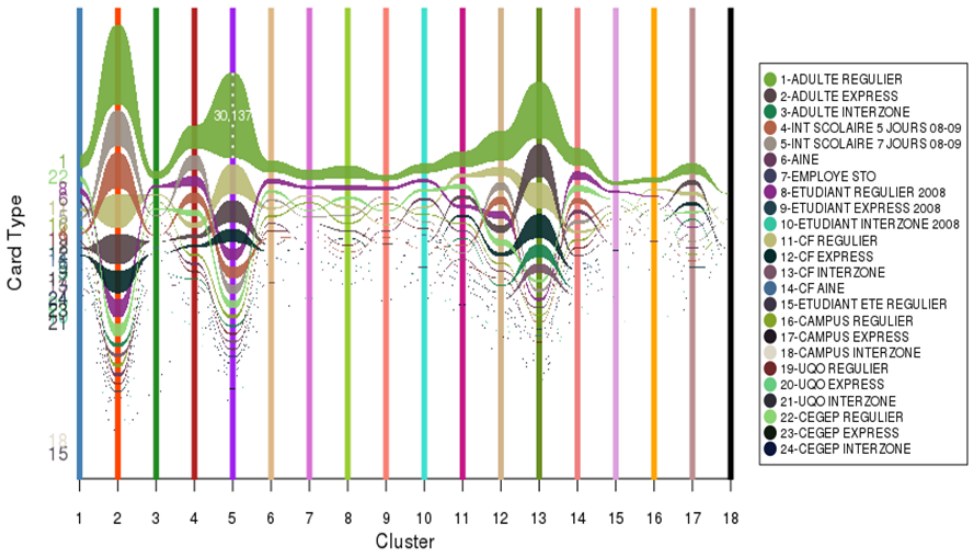

Figure 8 : Frequency Distribution of 24 Card Types vs. 18 Clusters (April 2009)

Frequency distribution of 24 card types versus 18 clusters is shown in Figure 8. Obviously, the clusters 2, 5 and 13 are the most frequent clusters and the regular adults (type 1) has the greatest portion of all clusters but other card types portions change over clusters. In cluster 2, card types 5, 4, 11 and 2 are the most seen after card type 1. In clusters 5 and 13, card types 11 and 2 are the most frequent after card type 1 but the card types 5 and 4 drop drastically.

The commuters between the morning and night (clusters 1, 3, 4, 6, 7, 14 and 16) as well as the morning single-trips (cluster8, 9 and 15) are mostly regular students after regular adults. It seems reasonable because at this time of the day, the workers and students use the public transportation to get their destination. Despite, the regular students and regular CEGEPs drop significantly in the routine scheduled users (clusters 2, 5 and 13).

Afternoon and evening users (clusters 10, 11 and 12) consist of the regular adults followed by regular students, regular CF, regular CEGEP and interscolaires. Among the night persons who are night workers or returning home late, clustered in 17, the express adults and regular CF have the highest frequency after regular adults.

Furthermore, the graphical representation in Figure 5 shows the clusters' membership hugely in clusters 2, 5, 13, 12 and 4 for all card-days which is a little different from the unique ID cards with 2, 5, 12, 4 and 13 shown in Figure 6. The clusters 2 and 5 are the most frequent in card-days as well as in cardholders, respectively. The cluster 13 (regular) is more frequent than clusters 12 (late) and 4 (long day) in card-days although it contains fewer cardholders than these two clusters. In other words, the 8.1 % of unique cardholders members of cluster 13, are present in public transit 14.9 % but the clusters 12 and 4 containing 10.3% and 10.1% of unique IDs, respectively, use public transportation 7.64% and 7.6%, respectively. It is half of the usage of cluster 13 members despite more unique IDs. The reason might be the express adults and CF as well as regular CF presence in collective transit at routine scheduled with less than maximum boardings (cluster 13) but they are absent more at before and after noon.

3.3 Clusters and fare types over days

We demonstrate the users' cluster distribution and card type distribution over all days of April 2009 in Figures 9 and 10.

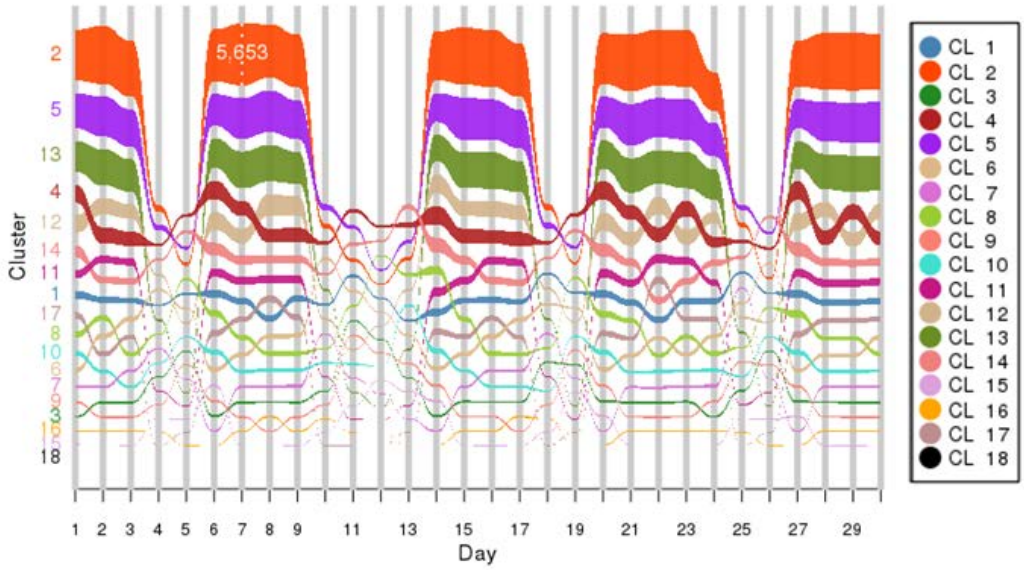

Figure 9 : Frequency Distribution of 18 Clusters (1 to 30 April 2009)

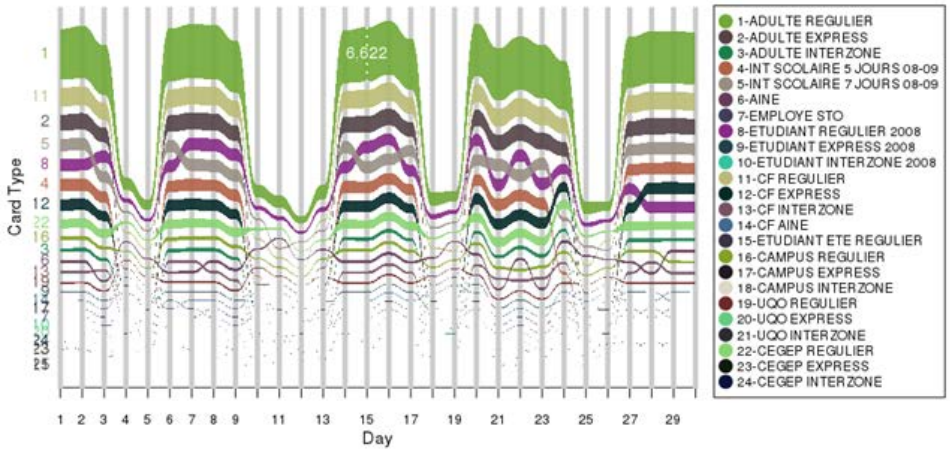

Figure 10 : Frequency Distribution of 24 Card Types (1 to 30 April 2009)

From these two figures, it turns out the largest presence in transit network on working days and the lowest on weekends as well as two holidays 10th (Good Friday) and 13th (Easter Day) April 2009. To have more interpretative results in the further sections, we separate these two kinds of days.

As expected, the clusters 2, 5, 13 (the regular commuters) as well as 4 (long day) and 12 (evening) are the dominant clusters corresponding to regular adults, regular CF, express adults, respectively,

on working days. Despite, on holidays, the cluster 13 drops drastically and is replaced by clusters 4 and 14 in which regular adults, students and CEGEPs are members of.

3.4 Day of week

It would be interesting to investigate the clusters distribution by the day of week (Figure 11). Moreover, we will apply to measure the stability methodology over each day of week in chapter 5, so this figure justifies the different clusters distribution over days of week.

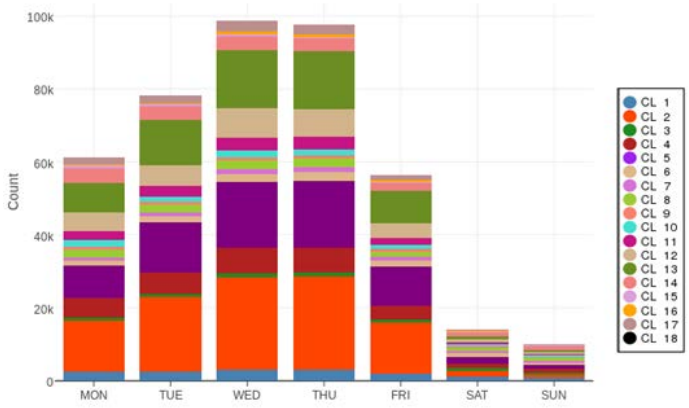

Figure 11 : Distribution of Clusters by Day of Week

Figure 11 shows that cluster 2 (regular commuters with highest boarding) has the highest portion during working days. The Wednesdays and Thursdays have the highest count because there are 5 Wednesdays and Thursdays as described in Figure 12.

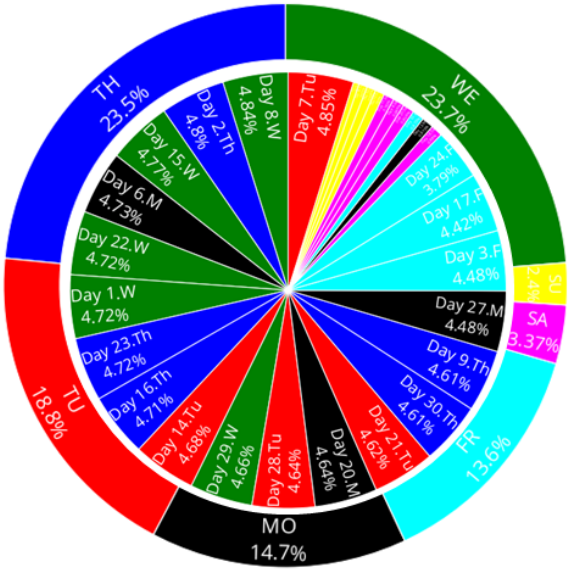

Figure 12 : Day of Week and Month Portions

It should be considered that the April 2009, starts with Wednesday and terminates with Thursday, so there are not four complete weeks which result in different total percentage of days in Figure 12 (outer). From this figure, we note that Wednesdays and Thursdays are the days more frequent than other days.

On the other hand, in Figure 12 (inner), it turns out the same proportion for working days with 4.5 % on average and the same proportion for weekends and 10th (Good Friday) and 13th (Easter Day) with 0.75 % on average.

4 METHODOLOGY

The clusters defined for each card-day associated with a cardholder have a range of 1 to 18. As these clusters have specific temporal characteristics, it would be useful for the transit agencies and the experts in public transportation to know how the users change their temporal habits.

To this end, in this section, first 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).

Next, the dominant cluster for each cardholder is defined. The methodology was performed in R statistical software.

4.1 Measuring stability of cluster's membership

To calculate the stability of users in cluster's membership, first we should find the distance between each cluster. We selected the distance based on the clusters' centroids, i.e., the average across all the points in each cluster. The primer work for clustering was done on the 3-dimensional data (projected data) (Figure 1), consequently, we have cluster's center in 3 dimensions based on equation 1.

Table 2 : Clusters' center

Card_day Data

User	Day	X	Y	Z	Cluster
X1	1	0,0576429	0,8427106	0,3478260	2
X1	2	0,0576429	0,8427106	0,3478260	2
...	...	...	...	...	...
X1	30	0,0576429	0,8810278	0,4347826	13
X2	1	0,0776429	0,8427106	0,4547826	4
...	...	...	...	...	...

→

Clusters' center Data

Cluster	X_c	Y_c	Z_c
1	-0.14892	0.840994	0.185789
2	0.032573	0.845395	0.342712
3	-0.58021	0.631664	0.334503
4	-0.39674	0.658101	0.029153
...	...	...	...
18	0	0	0

Then, based on these 18 centers showing in Table 2, the euclidean distance method is applied to obtain the distance between each cluster showing in the heat map (Table 3).

Table 3 :18 Cluster's Distances

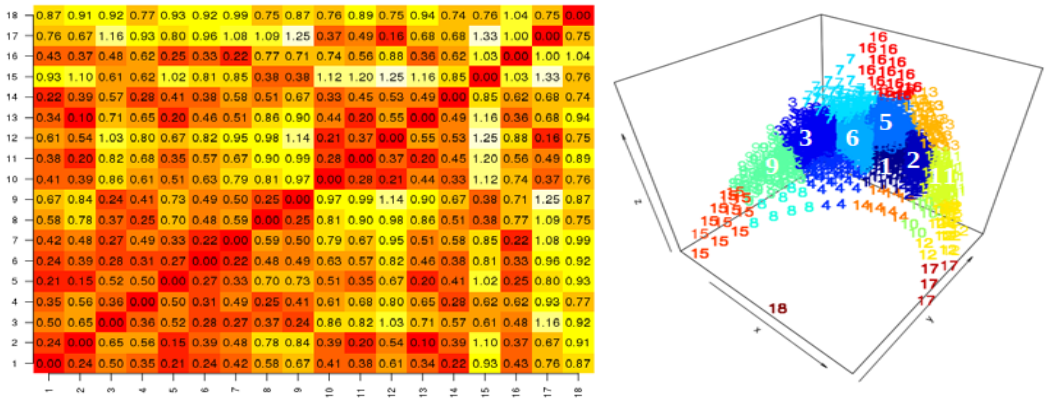

Figure 13 : 3D scatter plot of projected data

In the heat map, the more two clusters are closer; the more color is darker and vice versa. For example, in Figure 13 (based on (x,y,z) of all card-day data), center of cluster 15 is at the same distance from the centers of clusters 8 and 9 which is confirmed in the heat map dark red units of 0.38 and 0.38. In contrast, the longest distance between clusters 15 and 17 is calculated to 1.33 and shown in white on the heat map.

The closest clusters are 2, 13 and 5 who are the routine scheduled two commuters but the farthest are 15 and 17 who are the morning and night single commuters, respectively.

Based on these dissimilarity distances, we calculate the cluster's membership stability of each Card ID using its card during April 2009. Note that, as we use the dissimilarity distance, the result is the "instability" measure. Consequently, the more the measure is smaller, the more the user is stable.

It's possible to measure general or sequential stability. We selected sequential instability to be more interpretable. Sequential instability counts for cluster changing over card-days for each individual considering the distance showed on the heat map, in a sequence way. In other words, from the first day of use to the second one, second to third and so on. We sum these counts for each user.

Sequential instability for cluster appearance of a Card Id during a month is written as shown in Equation 2:

Vector ID_i: Cluster 1 → Cluster 2 → Cluster 3 → Cluster 4 → Cluster 5 →… N_i

Equation 2 : Weighted Sequential Instability

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

Taking as a vector the card-day clusters, we can move from one traveled day (j) to the next (j+1) for each ID_i , for i=1 to 26198. Then, finding the corresponding distance from the heat map.

The number of usage influences significantly 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 traveled days N_i corresponded to each user to obtain the "weighted instabilities".

For example, considering ID_1=vector(cl 1,cl 2,cl 3) on 3 traveled days. According to Table 3, WSI_1 is computed as:

$$WSI_1 = (distance[1,2] + distance[2,3])/3 = (0.24 + 0.65)/3 = 0.29$$

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 traveled days. The issue was solved by adding a constant (+1) to all the measures before being divided. This helps to distinguish the single-traveled day users from very high stable ones. Consequently, more the measure is lower; 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, by the help of a histogram and a scatter plot of instability measures, we ranked the instability to three "stability" categories defined as:

- High stable users: instability equal or fewer 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]

4.2 Dominant clusters

After measuring the stability of users, it would be 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. Note that there are some users with equal maximum cluster membership which are removed. It consists of 2745 (10%) IDs, who are assigned to medium and unstable levels more than stables.

4.3 Stability in days of week

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

5 APPLICATION

In this chapter, the methodology explained in the previous chapter is applied on a case study. The data explained in chapter 2 contains some "NA"s which represents the inactive days. To measure the stability, they are transformed to zero (See index). Furthermore, 10 days (weekends and long weekend) are removed as explained in 3.3 to apply the method on the working days as well as weekends, separately.

5.1 Measuring the stability

By applying the methodology explained in chapter 4, the weighted sequential instabilities are found and ranked, showing in Figure 14 and Figure 15. Colors represent different levels of stability. Green, blue and red covering high stable, medium stable and unstable levels, respectively.

The instabilities range between [0.05, 1.15]. More the instability is lower; 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 and regarding the number of traveled days, she or he is a stable passenger. Consequently, users with the instability equal to 0.05 are the highest stable users. In contrast, the users having instability of 1.15 are the most unstable users. They are ranked to three categories high stable, medium stable and unstable users each including 50, 40 and 10 percent of users, respectively, selected by the histogram in Figure 15.

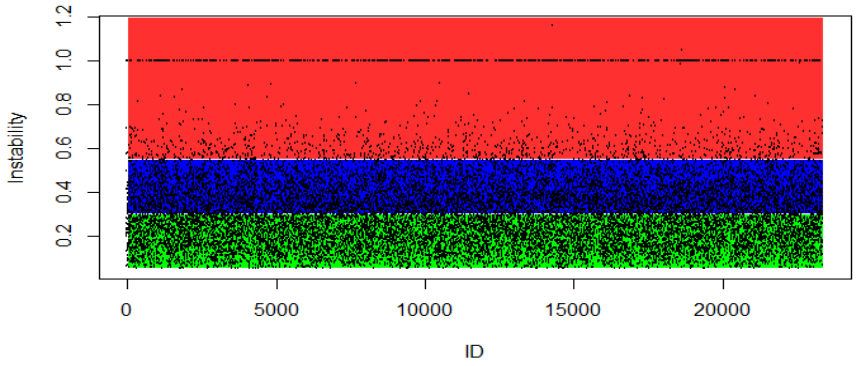

Figure 14 : Distribution of Instability

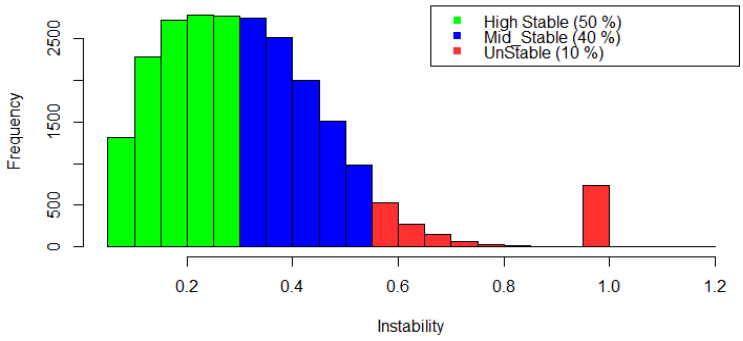

Figure 15 : Histogram of Instability

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-traveled 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 14 and single red column in Figure 15).

5.2 Dominant cluster

As described in chapter 4, the dominant cluster corresponding to each cardholder is recognized and represented in Figure 16 and Table 4.

As expected, the most frequent clusters count the most dominant clusters, i.e. clusters 2, 5, 13, 4 and 12. Obviously, the regular two-commuters (cluster 2) have the biggest portion on all levels. On the other hand, before noon (cluster 5) and after noon (cluster 13) two commuters behave differently. 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.

Dominant cluster 12 contains the pretty same unstable users as dominant cluster 2 although it has 4.68% compared to 31% users of cluster 2 (Table 4).

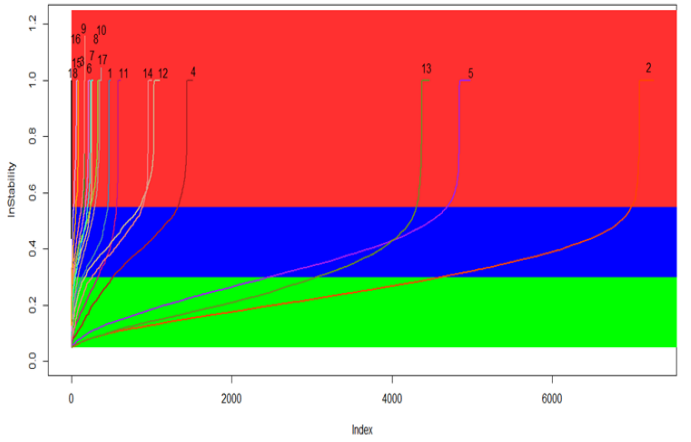

Figure 16 : Distribution of Stability Level vs. Dominant Clusters

Table 4 : Summary of dominant clusters on stability levels

	High Stable	Mid_Stable	UnStable	Sum
CL 1	133	320	34	487 (2.08%)
CL 2	4618	2368	274	7260 (31.05%)
CL 3	26	126	19	171 (0.73%)
CL 4	515	813	183	1511 (6.46%)
CL 5	2467	2220	283	4970 (21.26%)
CL 6	48	191	29	268 (1.14%)
CL 7	18	155	50	223 (0.95%)
CL 8	34	205	117	356 (1.52%)
CL 9	19	106	46	171 (0.73%)
CL 10	45	157	59	261 (1.11%)
CL 11	332	232	44	608 (2.6%)
CL 12	170	679	246	1095 (4.68%)
CL 13	3051	1252	160	4463 (19.09%)
CL 14	261	626	111	998 (4.26%)
CL 15	4	25	42	71 (0.3%)
CL 16	8	49	30	87 (0.37%)
CL 17	80	208	87	375 (1.6%)
CL 18	0	1	1	2 (0.008%)

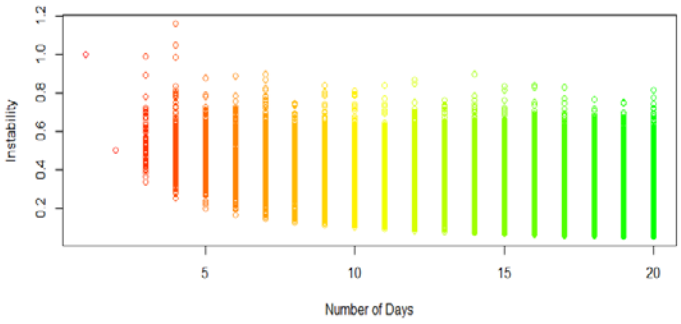

Figure 17 : Distribution of Instability over Number of Traveled Days

Furthermore, the number of traveled days plays an important role on the instability measure (stability level). More number of traveled days increases from 1 day to 20 days; more the stability improves (instability decreases), see Figure 18, showing that 73% of passengers present on transit network from 14 to 20 days have the low instability according to Figure 17.

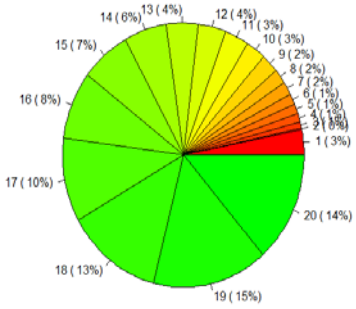

Figure 18 : Number of Traveled Days Portion

The impact of the number of traveled days regarding dominant clusters and stability levels are presented in 3D Figures 19 and 20.

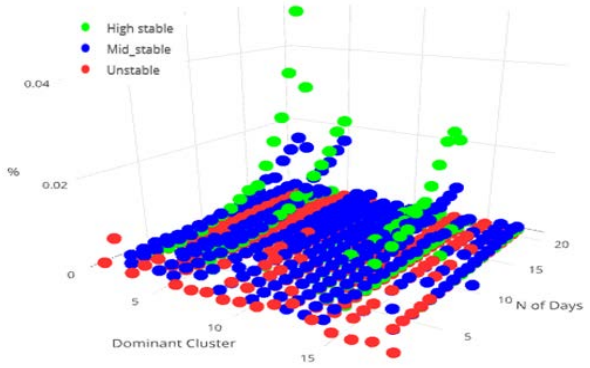

Figure 19 : Percentage of Stability levels on Dominant clusters and Number of Traveled Days

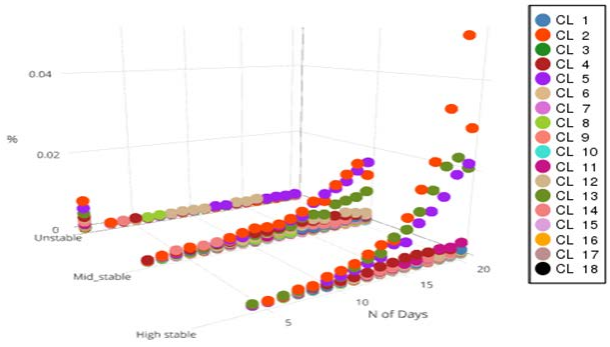

Figure 20 : Percentage of Dominant Clusters on Number of Traveled Days and Stability levels

It is evident that the number of traveled days influences high stable level dramatically. In other words, a big portion of users present on public transit more days and stay high 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-traveled day users. The single-traveled day users are the most frequent on unstable users which contain cluster 2 members more.

Another indicator was the fare types. Cardholders come from different sections of society such as adults, students, elderly people, etc. In chapter 3, the portion of cardholders according to the 24 types defined by the "STO" is shown in Figure 7. It demonstrates that the greatest clients are the regular adults followed by regular CF, express adults and so on. But what is the stability level of such clients?

Figure 21, it turns out the greatest frequency for regular adults (Type 1) on all levels. Moreover, on high level, regular CF and express adults overcome the others after regular adults. But on the second level, regular students and regular CEGEP are more unstable than regular CF and express adults. On the other hand, number of unstable regular adults is near to the number of unstable regular students although they have very different unique ID portions, 31.6% and 8.32%, respectively (Figure 7).

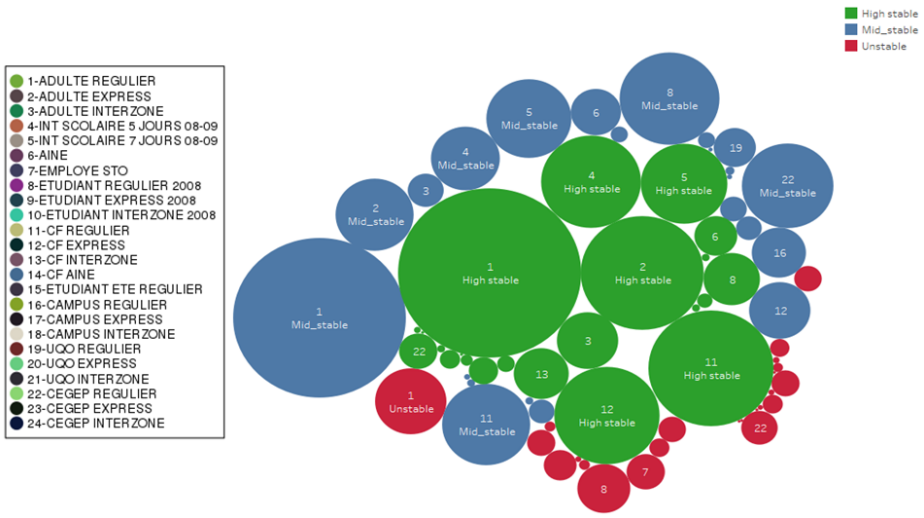

Figure 21 : Card Types Distribution on Stability levels

5.3 Users' traces over all days

In this section, we trace the flow of cluster's membership for each user over working days. In Figure 22, flow diagram of all clusters over all working days is shown on 3 stability levels. First level represents the high stable users (green color) staying in the same cluster or at least switching to very close clusters. In contrast, the red color showing the unstable users' trace who change their clusters with not close ones frequently or use their card just once, we assigned these users as unstable because they are single and have no pattern. They are not removed to the aim of exploring the clusters in this regard. 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 April, contains the most inactive users or the cluster 2, 13 and 5 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 next sections.

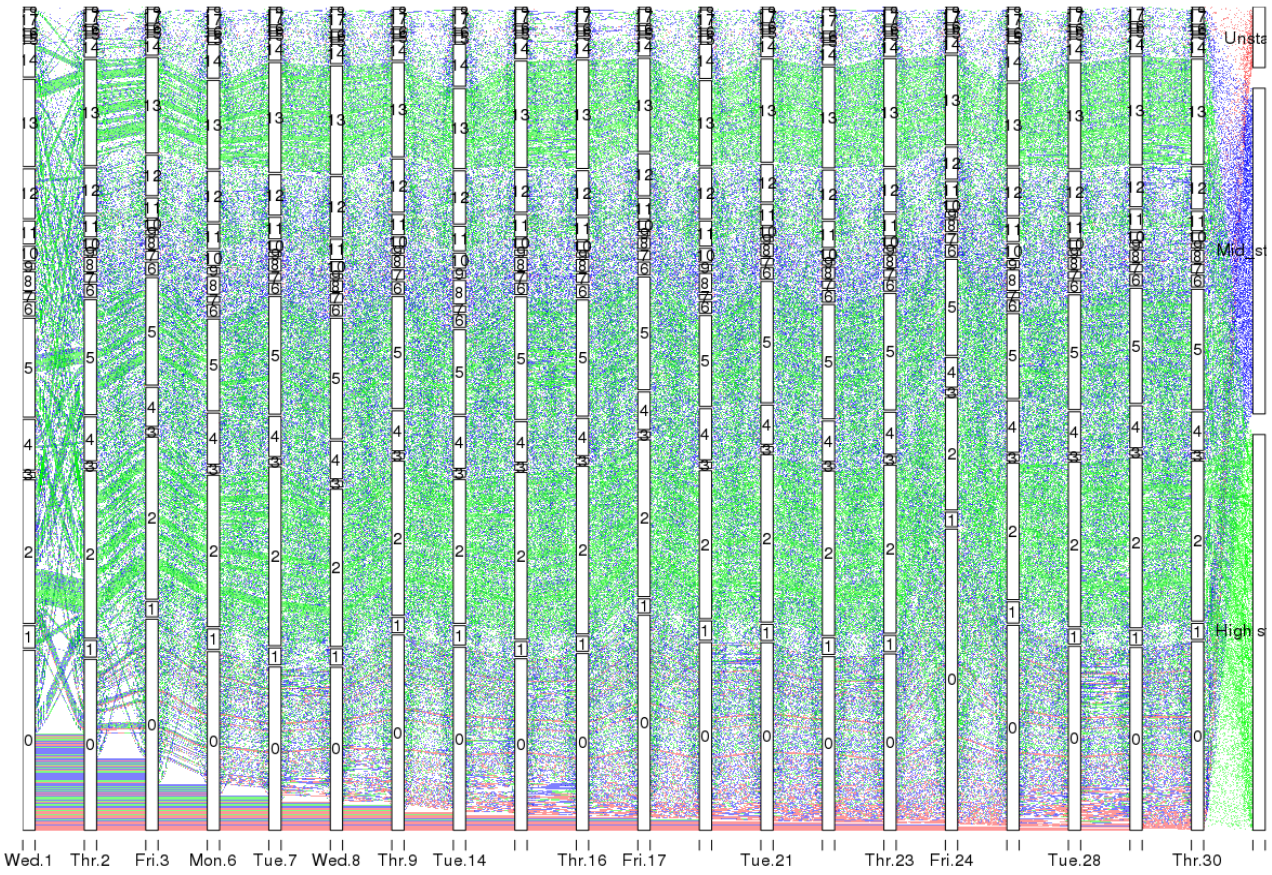

Figure 22 : Flow diagram over all 20 Working Days

5.4 User's trace over weekdays

In this part, we trace clustering membership over the weekdays: Mondays, Tuesdays, Wednesdays, Thursdays and Fridays as presented in Figures 25 to 29, respectively.

First, in Figures 23 and 24, the histogram and distribution of weighted instability over each weekday show the similar trends of Mondays and Fridays as well the other three days having similar trends with instability intervals [0.2, 1.2] resulted of about 22500 unique IDs.

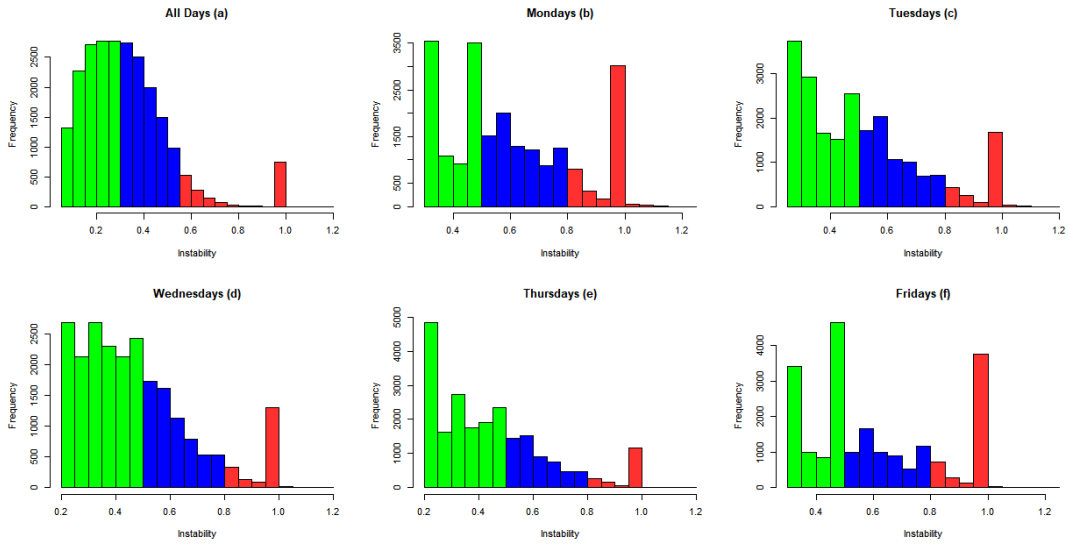

Figure 23 : Histogram of Instability on Working days

Moreover, Mondays and Fridays have the greatest unstable users (20% and 23%, respectively) and smallest high stable user's portion (42% and 47%, respectively) in comparison with the other days among them Thursdays have the highest stable users (68%) and the lowest unstable (7%). Obviously, the number of single-traveled day users on first and last working day is significantly high (the straight red line in Figure 24).

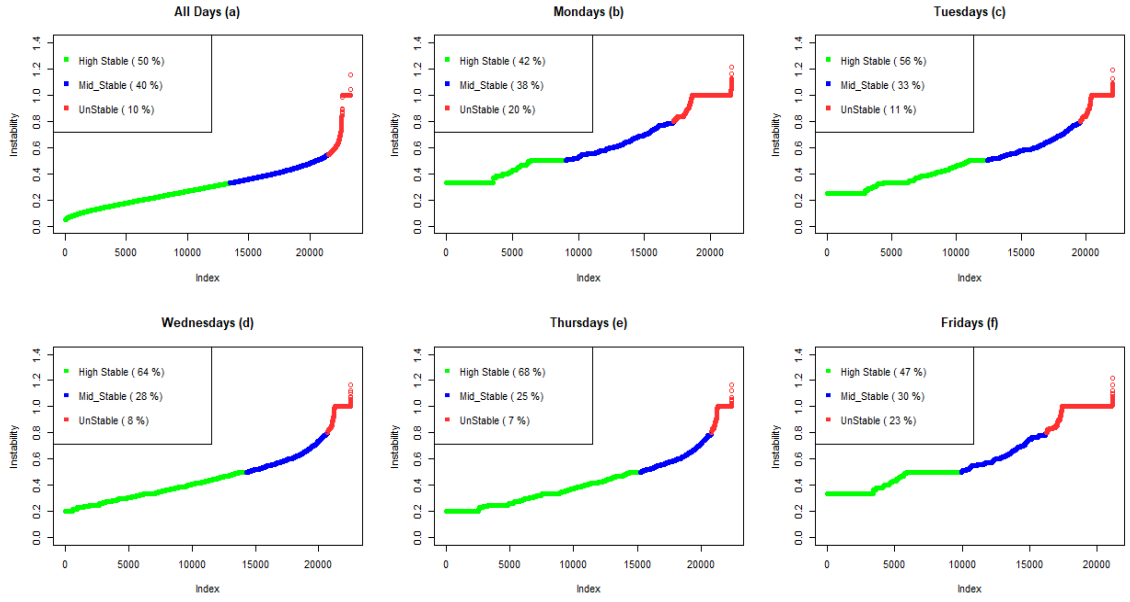

Figure 24 : Distribution of Instability on Working days

After measuring and ranking the users over each specified weekday, the flow diagrams are traced and presented in Figures 25 to 29 associated with Mondays to Fridays, respectively.

The results are described as below:

- Mondays and Fridays:

The clusters 2, 5 and 13 are the most frequent, the highest stable (green color) and they are so close to each other so that the users moving between them or even with one inactive day, stay high stable (trace 2-5-2, 5-13-13, 2-5-0,...). Furthermore, the half of high stable users members of cluster 2 on first and second Friday, are inactive on last Friday, although the same trends for clusters 5 and 13 with about a third becoming inactive on last Friday.

The users changing between clusters 2, 5 and 12 as well as 2 and 4, in addition 13 and 17 or with an inactive day are assigned to the medium stable group (blue color).

The unstable users are mostly the single-traveled day users who are active only one day (red color).

- Tuesdays:

Almost the same high stable level trends on Mondays are observed on the following day with a difference being high stable even if one or two Tuesdays are inactive. As explained in the previous sections, the number of days plays a significant role on the stability level and we have three Mondays (Monday 13 Easter day was removed) but four Tuesdays, so the algorithm decides to take Tuesday's passengers with one or two inactive days as high stables. By tracing the blue lines, it reveals that they follow the same clusters 2, 5 and 13 on all Tuesdays but the second Tuesday. This is 14 April (the day after Easter day) on which some of users members of cluster 2, change to cluster 4 or 12. And two commuter users in clusters 5 and 13 become late (cluster 12) and single evening (cluster 17) users on the day after the long weekend. As these changes happen between pretty far clusters, the corresponding users are assigned to the medium stability level.

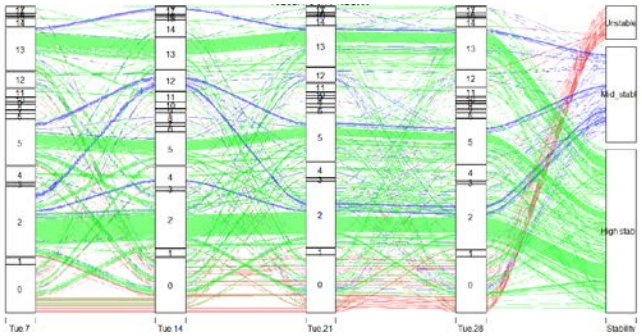

Figure 25 : User's Trace on Mondays

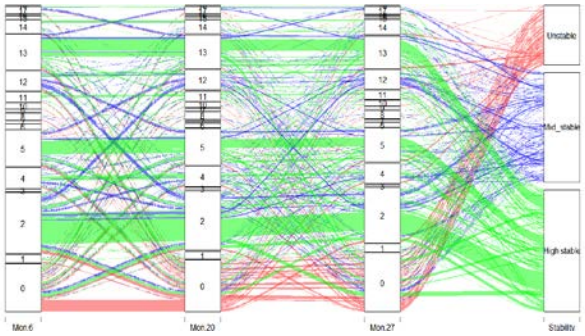

Figure 26 : User's Trace on Tuesdays

- Wednesdays and Thursdays:

Likely to other days, clusters 2, 5 and 13 are the most frequent high stable even if maximum three days are inactive (same justification as Tuesdays).

The traces on Wednesdays are not as dense as the other days. This is caused by the first Wednesday which is the first day of the month, too. We observe that the users stay on the three mentioned clusters from second to last Wednesday who come from a combination of these three clusters on first Wednesday.

By the way, as they are very close clusters switching between each other, the greatest portions of high stable level users belong to these two days.

Furthermore, there are medium stable users who are not traced on the diagram. The reason is the low frequency of the similar users because the diagram is limited to show the traces of similar users equal

or more frequent than 10 times for the best visualization. Hence, in contrast with the other days, there are the least medium stable users but with fewer than 10 similar users in this level on Wednesdays and Thursdays.

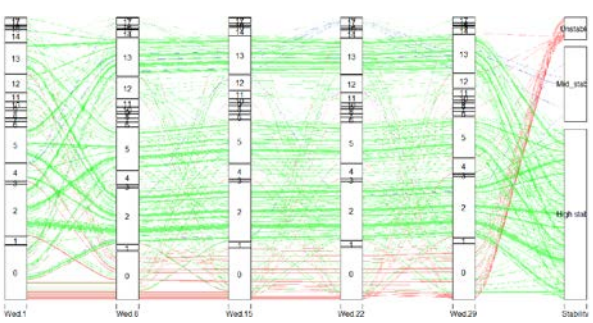

Figure 27 : Users Trace on Wednesdays

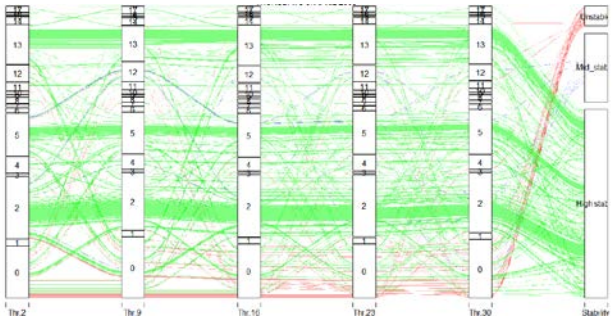

Figure 28 : User's Trace on Thursdays

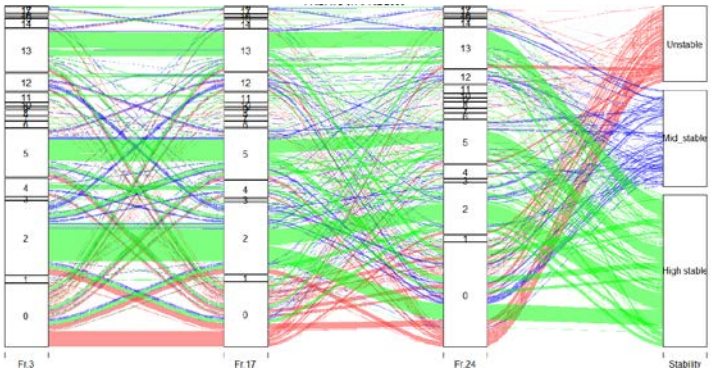

Figure 29 : User's Trace on Fridays

5.5 User's trace over weekends and holidays

In the previous sections, we removed 10 days containing weekends and holidays in order to have more homogenous days (only working days). To explore whether there are patterns on the weekends and holidays, similarly to the working days, the methodology is applied to these days, too.

Among all users, less than 0.3% use their card only weekends and holidays and 34% are present on transit network over both working days and weekends and holidays.

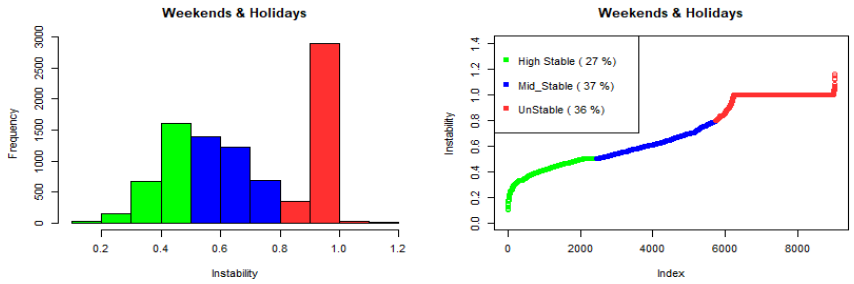

Figure 30 : Histogram and Distribution of Instability on Weekends and holidays

We observe in Figure 30 that users over these days are not very stable and the single-traveled day passengers (red column in histogram) cause the instability. Consequently, we do not expect to have similar trends corresponding to the clusters as shown in Figure 31 and described below.

- Weekends and holidays:

The long weekend has the same trend as weekends with the highest inactive users on 12th April, the day before Easter day. Furthermore, there are a little bit more users in clusters 2 and 5 on Saturday 4th, Friday 10th and Monday 13th but overall, they are single-traveled day users and spread in different clusters. Sundays are the least busy days with the highest inactive users.

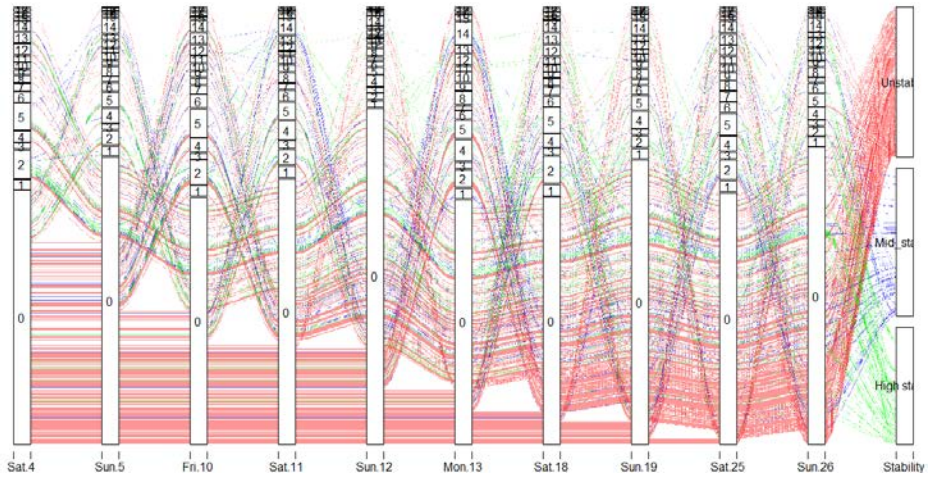

Figure 31 : User's Trace on Weekends and holidays

6 CONCLUSION

6.1 Contributions

This study proposes 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 stabilities are

computed and visualized in days of week to identify the level of stability and corresponding clusters in days of 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 removing 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 week. First and last working days contain the greatest portion of unstable users, specifically, the single-traveled day users are seen over Mondays and Fridays. In contrast, the users in the other three working days are the high stables, more. Thus, an effort to increase bus lines over these timestamps will raise customer satisfaction. Second, the first day after long weekends has a moderate difference in cluster's membership with other same working days which makes down the high stable users to the second level. Analyzing the data with several long weekends and holidays would provide the best understanding of clients' temporal variability as well as better bus service optimization. In addition, users are most inactive on Sundays and passengers use the transit network maximum once on average over all weekends and holidays.

Third, regular and express adults are the most stable passengers, while the regular students have shown the high unstable behaviours compared to the first and second groups. Finally, we found that more than 70 % of bus users are present on the transit network more than 14 days. And they are the most stables.

6.2 Limitations

Possibly the most significant limitation is the different frequencies of cluster's membership for each user. It makes the un-weighted selection of the dominant clusters from a variant interval. To improve this effect a little, we removed the users having the same maximum frequency of dominant clusters. The other limitation concerns the inactive days. As explained in Future works, with the aim of re-clustering the users, the distance between an assigned cluster-day and inactive-day is unknown. It may help considering this distance as infinitive.

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 re-clustered by the help of gene clustering method. In other words, considering each vector of cluster's membership corresponding to an individual over a month, 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 concerning its behaviour, so, to compare each user behaviour to others, aforementioned method may provide the interesting results.

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).

Briand, A. S., 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.

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.

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.

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