Transportmetrica A: Transport Science

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

A Visual Segmentation Method for Temporal Smart Card Data

Mohammad Sajjad Ghaemi, Bruno Agard, Martin Trépanier & Vahid Partovi Nia

To cite this article: Mohammad Sajjad Ghaemi, Bruno Agard, Martin Trépanier & Vahid Partovi Nia (2016): A Visual Segmentation Method for Temporal Smart Card Data, Transportmetrica A: Transport Science, DOI: 10.1080/23249935.2016.1273273

To link to this article: http://dx.doi.org/10.1080/23249935.2016.1273273

Publisher: Taylor & Francis & Hong Kong Society for Transportation Studies Limited
Journal: *Transportmetrica A: Transport Science*
DOI: 10.1080/23249935.2016.1273273

A Visual Segmentation Method for Temporal Smart Card Data

Mohammad Sajjad Ghaemi[a,b,c], Bruno Agard[a,b], Martin Trépanier[a,b] * and Vahid Partovi Nia[a,c]

[a]Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, 2500 ch. de Polytechnique Montréal (Québec), Canada, H3T 1J4;

[b]*Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT), Université de Montréal, P.O. Box 6128, Centre-Ville Sta., Montréal, Canada H3C 3J7*;

[c]*Groupe d'Études et de Recherche en Analyse des Décision (GERAD), GERAD-HEC Montréal 3000, Côte-Sainte-Catherine Rd., Montréal, Canada H3T 2A7*;
Received 15 Feb 2016; Revised 12 Dec 2016; Accepted 13 Dec 2016

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

In many cities, worldwide public transit companies use smart card system to manage fare collection. Analysis of this acquisitive information provides a comprehensive insight of user's influence in the interactive public transit network. In this regard, analysis of temporal data, describing the time of entering to the public transit network is considered as the most substantial component of the data gathered from the smart cards. Classical distance-based techniques are not always suitable to analyze this time series data. A novel projection with intuitive visual map from higher dimension into a three-dimensional clock-like space is suggested to reveal the underlying temporal pattern of public transit users. This projection retains the temporal distance between any arbitrary pair of time-stamped data with meaningful visualization. Consequently, this information is fed into a hierarchical clustering algorithm as a method of data segmentation to discover the pattern of users.

Keywords: clustering; public transit; smart card; temporal pattern; projection.

1. Introduction

Public transit serves the society to solve their mobility in almost every country (Gallotti and Barthelemy 2015). Thus, inter-disciplinary challenges of public transit is attended in several branch of science and engineering (Gkiotsalitis and Stathopoulos 2015). Progress of smart data and the use of automated payment system provides an explosive rich source of data, whose its analysis can promote the economy (Weisbrod and Reno 2009), reduce the air pollution, and enhance the quality of life (Ma et al. 2015). In this regard, diverse combination of tools and techniques from various disciplines, e.g. data mining, machine learning, urban computing, urban planning, management, business, civil engineering,

*Corresponding author. Email: mtrepanier@polymtl.ca

industrial engineering, statistics, mathematical engineering, geographic information system (GIS), and high-performance computing are vital to extract the meaningful piece of information from such data.

In most of public transit studies, bus stops and subway stations play the central role, regardless of the temporal features of the data. The frequency of the used locations is utilized to construct a model for identifying the user behavior. This knowledge is helpful to provide particular services in each station or bus stop. Nonetheless, such models are incapable to uncover the detailed behavioral pattern of users. In most of recent researches summary statistics such as the frequency of travel days, the count of similar starting boarding times, the number of similar transit sequences, and the repetition of similar stop/station sequences are extracted as descriptive features to be fed into clustering algorithms with few justification and explanatory translation. In recent years, user satisfaction from public transit system, quality of service and perceived quality of bus transit model are investigated based on reliability, length of journey, and driver amiability (Bordagaray et al. 2014; Del Castillo and Benitez 2013; de Oña and de Oña 2015).

Despite the extensive research that has been done on public transit domain, various obstacles are arisen for specific purposes. Such specific purposes require particularly new computationally efficient views to address them. In this study, the problem of user clustering is attacked. The ultimate aim is to uncover the temporal behavior of users in their monthly trips.

The aim of this research work is to identify group of similar users relying on the gathered data from smart cards. More specifically, groups of similar user focusing on temporal aspect of the smart card data are identified. To this end, in Section 3 we propose a projection technique which is able to transform a vector of hourly usage associated to each smart card into a three dimensional feature vector that lays out the hidden temporal patterns. Accordingly, we deploy a hierarchical clustering algorithm to elicit the coherent internal representation of users in terms of analogous temporal behavior. In Section 5 experimental results of one month record of smart card data from Gatineau (a city in western Québec, Canada) is analyzed to illustrate the effectiveness of our suggested technique.

2. State-of-the-art

2.1. *Recent research papers on the analysis of smart card data*

Public transit systems have been expanded independently in many cities regardless of their size. Thus, having a strategic plan of Integrated Smart Card Fare Collection System (ISFCS) is necessary in development and enlargement of the public transit network. ISFCS fills the gap of different public transit operators and better meets the passengers' needs and satisfaction. Barriers of ISFCS and their possible solutions are discussed in Yahya and Noor (2008). In Pelletier, Trépanier, and Morency (2011) several other aspects of ISFCS are considered from technologies to privacy issues in three levels of management including, strategic, tactical, and operational. Moreover, discussion and comparison of planning, scheduling, and survival modeling for many different purposes are provided in Pelletier, Trépanier, and Morency (2011).

Describing user behavior in public transit network is one of the main issues that can be revealed via the smart cards data (Ma et al. 2013). Accordingly, finding a measure to evaluate and disclose behavioral patterns from the history of user's habits is a crucial part of Smart Card Fare Collection System (SCFCS) analysis. Various measures are proposed in Morency, Trépanier, and Agard (2006) by considering the variability of users'

behavior over smart card data of ten months. Agard, Morency, and Trépanier (2008) applied k-means on weekly boarding; this study permitted to identify large temporal users behavior and detect important changes in users schedule (spring break for example) but the optimal number of clusters is difficult to identify and new estimation techniques for determining the number of clusters in such data seem necessary. In Lathia and Capra (2011), two viewpoints are investigated to measure the transportation system's performance; self-report of users' feedback, and real behavior while they are encouraged by various incentives. Lathia and Capra (2011) and Herrera et al. (2010) concluded that smart card data is as important as human activity on mobile phone data for designing future infrastructure and guidance of travellers. Therefore, human mobility could be modelled according to the smart card data as one of the big data sources concerning human activity.

Smart card data contain worthwhile digital information of daily locations visited at certain period by a large number of individuals. Besides, other source of information can be combined with this data such as mobile phone, GPS tracker vehicle, e.g. bike, car, motorcycle, credit card transactions, social network, (Herrera et al. 2010; Gkiotsalitis and Stathopoulos 2015). This helpful information could be utilized to characterize and model urban mobility patterns (Hasan et al. 2012; Järv, Ahas, and Witlox 2014). Other useful information such as travel time and number of passengers for the sake of congestion analysis and planning improvement could be possibly extracted as well (Fuse, Makimura, and Nakamura 2010).

Kusakabe and Asakura (2014) proposed a data fusion approach in order to estimate the trip purpose and then interpret the observed behavioral features. They are able to successfully distinguish the following different reasons: (a) commuting, (b) leisure or business and (c) returning home in 86,2% of their available trip data.

Ma et al. (2013) used a data mining technique to understand regular travel patterns in Beijing, China. First they constructed trip chains, then extracted regular patterns using clustering that leads to specific trip rules.

Ali, Kim, and Lee (2016) analyze electronic fare transactions for analyzing travel behavior of the users, in Seoul, South Korea. They used an open-source agent-based transport simulation package, `MATSim`, over smart card data to model input demand. This study permitted to generate micro-simulation travel demand models.

Amoung others, Trépanier, Tranchant, and Chapleau (2007) and Alsger et al. (2016) explore smart card data in order to estimate trip's origin and destination. Origin of trips are relatively easy to define, thanks to the first boarding check, but destinations may require prediction.

Data representation in public transit is more complicated than conventional data sets in data mining or machine learning (Nantes et al. 2015). Summary of steady sequential time model in a discrete structure is the main reason that makes it difficult to analyze the temporal behavior (Shekhar et al. 2015). The focus of this research, is to deal with the temporal datasets, that could be categorized as temporal snapshot model in spatio-temporal data as in Shekhar et al. (2015). Most of the research works in this domain perform the data mining techniques on transformed spatio-temporal attributes in a conventional way. However, because of the intrinsic structure of spatio-temporal data, independent and identically distributed (i.i.d.) observations cannot be assumed for this sort of data. Consequently, conventional data analysis algorithms often fail to capture the essential knowledge from the data. Moreover, the extracted information has no real interpretation for the experts. These are the two principal reasons that reflect the urge of why advanced techniques are required to be tailored for public transit data.

Machine learning methods are often divided into supervised, and unsupervised subfields; semi-supervised methods have attracted more attention recently. Most learning

methods seek for dividing data into sub-populations. The difference between supervised and unsupervised method is the existence of training data (Hastie, Tibshirani, and Friedman 2009). More precisely, when an indicator variable is available for sub-population allocation, the problem is called supervised learning. If dividing the whole spontaneous data into k homogeneous sub-populations is required without any guide, the problem is called unsupervised learning. Note that even the number of sub-populations, k, may be unknown.

Smart card data, may provide two distinct information: spatial and temporal. Spatial data consists of coordinates of the bus stop, e.g. latitude and longitude that could be GPS data or relative location coordinates. Temporal data describes the boarding time. According to this information, analyzing users behavior is divided into three categories, 1) Spatial patterns, 2) Temporal patterns and 3) Spatio-temporal patterns.

(1) Spatial pattern analysis focuses on location, such as the bus stop information. It turns out measuring the behavioral pattern only depends on the location of bus stops taken by the users, rather than knowing the starting hour of their trip.
(2) Temporal methods seek the information pertinent to the time associated to the public transit usages. Consequently, computing user's similarity score is carried out, by assuming that the bus stop information is unavailable. Thus taking the public transit at a specific time, plays the central role in this approach.
(3) The third scenario, is a mixture of the spatial and temporal approaches, called spatio-temporal data analysis. It could be viewed as a combination of the last two steps or an independent new approach to deal with spatio-temporal behavioral patterns.

2.2. *Extraction of users' temporal patterns in transportation*

The extraction of users temporal behaviors may be of value for planners. It may help them to plan the service more effectively. Classical approach to to discover users need includes counting passengers on board. Then, a generic demand is estimated. Smart card data permit to get more information: it is possible to extract generic behavior for all the users, or it is possible to follow a specific card. Clustering algorithms permit to subdivide the whole population of users in different groups that share certain behavior. The number of groups may vary depending on the accuracy needed by planners.

The majority of clustering algorithms can be divided into distance-based methods or model-based methods. Distance-based techniques are easy to understand and simple to implement. On the contrary, model-based approaches are flexible and adapt to complex data patters, but are counter-intuitive to implement or interpret.

Hierarchical clustering is a breakthrough in distance-based clustering context, 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. The dissimilarity measure is a positive semi-definite symmetric mapping of pairs of groups onto the set of real numbers. This measure, however, may not satisfy the triangle inequality unlike a distance. Hierarchical algorithms require a dissimilarity measure to merge clusters in order to build a nested structure of clusters. The common dissimilarities include single linkage (or nearest neighbors), complete linkage (or farthest neighbors), average linkage, and centroid linkage. There are two variants of hierarchical clustering depending on the direction of the construction of the nested groups. Agglomerative clustering starts with every observation as a singleton and consequently merges the closest clusters to end up with all data in one cluster. Divisive algorithms, on the contrary, start with all data in one cluster and split the clusters until finishing with all singletons.

The nested groups generated using a hierarchical clustering algorithm of data, are visu-

alized through a dendrogram. It provides an informative representation and visualization for different potential data structures, specifically while real hierarchical relations exist in the data. Dendrogram illustrates the nested structure or the evolutionary pattern of the members of a particular set. The idea of the dendrogram first appeared in evolutionary biology, and then applied in practice as an illustrative clustering tool in Sneath (1957). The height of the dendrogram expresses the dissimilarity between each pair of clusters. The initial groups are the leaves and every merge of clusters appears with an increasing height.

An automatic cutting point on the dendrogram has been a well-known problem for decades. Estimating a grouping, cutting a dendrogram, and model selection are closely related concepts. The ultimate estimated grouping is found by cutting the dendrogram at some height. One expects a visible gap in the height of the dendrogram for a natural grouping, but providing a universal cutting point on a dendrogram is counter-intuitive. An approximate model selection criteria such as AIC Akaike (1973) or BIC Schwarz (1978) can be applied to cut the dendrgroam if a statistical model is used to produce nested clusters (Heller and Ghahramani 2005; Heard, Holmes, and Stephens 2006). Most of the statistical models for clustering are a sort of mixture model (McLachlan, Do, and Ambroise 2004). The R package `NbClust` (Charrad et al. 2014) provides 30 different techniques to discover the optimal number of clusters in a data set. However, dendrogram itself provides a fairly well description of the clusters, so that it enables the experts in each domain to have a profound insight where to cut the dendrogram for finding the appropriate groups of data.

Data mining approach is used to understand passenger's temporal behavior to exploit the interpretable clusters (Mahrsi et al. 2014). This approach helps transportation operators to become aware of the customers' demands. In addition, it enables them to maintain their services and meet the user's requests more effectively. The real dataset from the metropolitan area of Rennes (France) with four weeks of smart card data containing trips of both bus and subway is tested. Furthermore, the cluster of similar temporal passengers extracted based on their boarding time, according to a generative model-based clustering approach. After, the effect of distribution of socioeconomic characteristics on the passenger temporal clusters are investigated in this study.

As another example Ortega-Tong (2013) studied the extensive database of Oyster Card transactions obtained from London's public transit users. This database is deployed to classify users based on the temporal variability, the spatial variability, the socio-demographic characteristics, the activity patterns, and the membership type. Improving the planning and the design of market research are the aims of this work, when selecting groups of homogeneous people is also of interest. Four groups of users including, regular users consist of workers and students commuting during the week, portion of them who make leisure journeys during the weekends, occasional users containing leisure travellers, and visitor travellers for tourism and business affair are investigated in this work.

Smart card data gathered from Brisbane, Australia is another source of information studied in Kieu, Bhaskar, and Chung (2014) for strategic transit planning according to the individual travel patterns. Origins and destinations of cardholder is defined as travel regularity, and the definition of habitual time is the regular time of trips. Thus, mining the travel regularity of the frequent users could be inferred to extract the travel pattern and its purposes. Reconstruction of user trips is made by spatial and temporal characteristics, then the frequent users are grouped by applying k-means clustering technique on the trip features including, origin and destination, number of transfers, travel mode and route uses, total travel time, and transfer time. In the last step, three level of Density Based Spatial Clustering of Application with Noise (DBSCAN) are applied to find the travel regularity Kieu, Bhaskar, and Chung (2014).

Schedules are a proper solution for the public transit user and for the public transit service provider. Most of the time, service providers operate on the same schedule in weekdays from Monday to Friday, and maintain distinct schedules for Saturdays and Sundays, assuming that the public transit user follows the same travel behavior during weekdays. It could be true for people with a regular schedule. However, society is constantly changing and more people now work only four days while other people work distantly once or twice a week. In addition, there are an increasing number of citizens with non-regular schedule such as immigrants or tourists. Hence, it becomes more of interest of the service provider to measure and predict the amount of regularity of public transit users, through their time-stamped smart card transaction database. By applying learning methods on smart card database we aim to divide the users into several subpopulations to obtain the clusters of users according to their behavior. These clusters can be put back in the context of daily mobility. Hopefully, by the analysis of these clusters we better understand the categories of the users, especially those who have a regular pattern of travel (Morency et al. 2010).

2.3. *Synthesis and justification of the needs*

The state of the art shows large interest of researchers in extracting knowledge from smart card data. Authors propose many directions, tools, and methods to explore this rich source of data. Those contributions may be classified in three main domains:

The first set of studies focuses on understanding the data, e.g. what happens on the network? this aspect is about extracting many indicators, evaluate characteristics and identify behaviors in the data. All information available from the smart cards are manipulated, formatted, and analyzed: boarding times, stops, lines and directions are the main information that are explored here.

The second set of studies deals with explaining the behaviors, e.g. why do we observe those behaviors? Here researchers explore the reasons that explain what they observe. Various sources of external data are widely used, depending on the intention of the authors. The idea is to cross, fusion, and predict from a data set. The smart card data are put in relation with the external sources of data to explain why one behavior or another is observed.

The third set of research consists of taking advantage of the extracted knowledge to help in decision making. Various objectives are considered: i) to improve the service for the user, with no supplementary cost for the transit operator, ii) to keep the same service but with minimized cost for the transit system operator.

Nevertheless, all of those topics rely on a good extraction of the user's behavior from the smart card data set. Besides, that extraction could be improved considering a better metric for the comparison of users' behavior. Traditional metrics consider Euclidean distance (which could not be used in our case), but also Dynamic Time Warping (DTW) and cross correlation. The two last metrics are powerful and widely recognized for the comparison of time series, but specific properties required for the analysis of customers' behavior could be used to improve this extraction. The main goal of this paper is to contribute in this aspect.

3. Proposed methodology

We suggest a two-stage visual method for analysis of temporal user behavior. The first stage consists of semi-circle projection to reduce the high-dimensional data into highly interpretable lower space. In the second stage a hierarchical clustering is applied on the

preprocessed data to extract the cluster structure for the expert.

We offer a simple mapping of boarding time information to the Cartesian coordinates. This suggestion is a sort of a multidimensional scaling (Borg and Groenen 2005), when some equalities and inequalities are proposed for certain distance between individuals. The mapping, that we call *Semi-Circle Projection* (SCP) is easier to understand in the polar coordinate, i.e. in terms of radius and angle.

First, reserve the center of a half circle for zero boarding time. For users with one boarding, take radius equal to $r_1 = 1$ and move the angle from 0 to π depending on the time of boarding. For vectors with 2 boardings, take radius $r_2 = 2$. Generalization for users with sequence of n boardings is then straightforward. Choose $r_n = n$ and move the angle according to the average time of boardings. However, the identity function $r_n = n$ diverges for large n. Choice of a converging r_n helps us to renormalize the half circles for long binary sequences, if needed. Our suggestion is $r_n = (1 + \frac{1}{n})^n$ having $\lim_{n\to\infty} r_n = e$, where e is the Euler constant. The third coordinate is required, as this method maps [0 1 1 0] and [1 0 0 1] fall on the same projection, because both have the same number of unit values (being two) and both have the same average of positions for unit values (being 2.5). This appeals to add another coordinate with a scale measure over the position of the unit values to distinguish these two users from each other in the projected space. We suggest the standard deviation of the position of the unit values as the z coordinate, giving a larger value to [1 0 0 1] comparing to [0 1 1 0], so they would not be mapped on the same point in three-dimensional projection. Suppose there are m user-day entities, organized in the binary matrix $X_{m\times L}$ whose the rows indicate the daily usage for specific smart card. This mapping can be formalized as follows,

$$\theta_{m\times 24} = \begin{bmatrix} 1\ 2 \cdots L \\ 1\ 2 \cdots L \\ \vdots\ \vdots\ \cdots\ \vdots \\ 1\ 2 \cdots L \end{bmatrix} \odot X$$

where $\odot$ is Hadamard (elementwise) product operator. Let r represents the number of boardings, thus for all smart cards with equal r, reduced data in the new space is written as

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

The number of boardings 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 $r_i = \left(1 + \frac{1}{n_i}\right)^{n_i}$.

The suggested simple transformation maps a binary sequence of any length to the Cartesian coordinates of only three dimensions. Implementing this method for travelled days, is compressed into only three dimensions, hugely facilitating further computation, analysis, and data visualization. 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. The result of this method on the dataset from Table 1 is shown in Figure 1.

4. Projection properties

The suggested projection includes two parameters, r_i, and θ_i for each temporal usage X_i that contribute to map the binary representation into three dimensional semi-circle space. The number of ones or frequency of usage is one of the most important factor which leads the projection through the mentioned parameters. Thus, the properties of this projection are somehow proportional to the total amount of usage determined by sum of ones. This implies the range of r_i, θ_i and $\mathrm{var}(\theta_i)$ are decreasing by increase in frequency of usage across time. Furthermore, it is evident for any pair of temporal usage encoded in binary vectors denoted by X_1, and X_2, where $X_1 \neq X_2$, and $n_1, n_2 \in \{1, 2\}$, SCP maps them onto distinct points in three dimensional reduced space, i.e. the projection is unique. However, this interesting property may not hold for $n_i > 2$. Below we separately state some properties of the projection in terms of these parameters r_i, θ_i and $\mathrm{var}(\theta_i)$. Remind that the first two axes of SCP is constructed using (r_i, θ_i) and the third axis is $\sqrt{\mathrm{var}(\theta_i)}$.

Theorem 1. *Suppose $n_i \in \mathbb{N}$, then*

$$\frac{r_{n_i+1}}{r_{n_i}} < \exp\left\{\frac{1}{n_i+1}\right\}.$$

See Appendix for the proof.

It is evident that r_{n_i} is an increasing sequence of n_i. Theorem 1 states that the rate of increase of the radius in SCP is very tight as the number of boarding n_i increases.

Theorem 2. *Suppose a pair of binary vectors X_i and $X_{i'}$ of length L, both with the same number of boarding n_i, then*

$$\max|\tilde{\theta}_i - \tilde{\theta}_{i'}| \leq 1 - \frac{n_i+1}{L}$$

where $\tilde{\theta}_i = \frac{1}{Ln_i}\sum_{j=1}^{L}\theta_{ij}$.

See Appendix for the proof.

Theorem 2 states that the angular range decreases as n_i increases. This property along Theorem 1 gives a vague idea about concentration of data after SCP for large n_i for the x-y coordinates of the SCP.

A similar result exists for the z-coordinate as well.

Theorem 3. *Suppose a set of all possible binary vectors with n_i number of boarding, and accordingly a set of all possible binary vectors with $n_i + 1$ number of boarding. Then the projection of such points over the z-coordinate of the SCP satisfy*

$$\min z_{n_i} < \min z_{n_i+1},$$

and

$$\max z_{n_i+1} < \max z_{n_i}.$$

Trivially if $\min z_L = \max z_L$ because the set of possible boarding with L number of boarding includes only one member.

Adding Theorem 3 to the other properties suggests that the data have the tendency of concentration as the number of boarding increases. So if a clustering technique is

implemented after projection, one may expect one or several clusters of data with large number of boarding raised naturally as the property of the concentration of data with large number of boarding. Subsequently SCP better maps the data with small number of boarding. Therefore, from public transport perspective, we recommend clustering after SCP if a clustering of data with small number of boarding is of interest.

5. Experimental results

Experimental design consists of two steps to analyze the data. First of all, SCP is applied on the high-dimensional binary data to project the data into the lower dimension. Next, hierarchical clustering reveals the structure of the users where similar ones grouped together. To this end, we first show the performance of SCP method on a small synthetic example by comparing our suggested method with other standard techniques. Then we use smart card data to discover similar groups of users in Gatineau transit network.

5.1. *Demonstration of Semi-Circle Projection (SCP)*

After introducing the suggested adhoc SCP method, we compare it with the other state-of-the-art time series distance measurements to illustrate the properties of the SCP. This demonstrates how one can improve the drawbacks for the temporal user behavior. Two commonly used distance measures, namely, cross-correlation distance, and autocorrelation-based dissimilarity distance are used from the `TSdist` package in `R` as the base measures for this comparison.

The cross-correlation based distance measure between two numeric time series is calculated by

$$D(x,y) = \sqrt{\frac{(1-\{\text{CrossCorr}(x,y,0)\}^2)}{\sum_{k=1}^{K}(1-\{\text{CrossCorr}(x,y,k)\}^2)}},$$

where $\text{ccorr}(x,y,k)$ is the cross-correlation between x and y at lag k, and the sum in the denominator goes from 1 to the maximum lag say K. Autocorrelation-based dissimilarity, computes the dissimilarity between a pair of numeric time series based on their estimated autocorrelation coefficients that can be calculated as $D(x,y) = \sqrt{(\rho_x - \rho_y)^\top \Omega (\rho_x - \rho_y)}$, where ρ_x, ρ_y are the estimated autocorrelation vectors of x and y respectively, Ω is a matrix of weights, and $\top$ denotes the transpose operator (Montero and Vilar 2014).

The results of the three different distance measures are shown in Figure 2, 3 for the users X_1, and X_8 respectively. $\{X_8, X_9, X_2\}$ could be considered as the first three nearest users to the user X_1 because of the similar time behavior. All three methods, indicate the user X_8 as the closest user to the user X_1 in Figure 2, however, X_9 is selected as the second nearest user in Figure 2(b) while the X_2 is selected in Figure 2(a), 2(c). Despite, the reasonable justification for the first two nearest users selected by cross-correlation distance, picking the user X_{13} as the third closest user to the X_1 violates the assumption of the temporal behavior in this dataset. Autocorrelation-based dissimilarity and the SCP measures preserve the constraints of the temporal distance for the user X_1. Next, the user X_8 is taken into account to follow up the performance of each method. The users $\{X_1, X_2, X_9\}$ are the first three candidates to be chosen as the nearest users to the X_8. In Figure 3, the selected users associated to the user X_8 are shown. Autocorrelation and the SCP are capable of picking those users as are shown in Figure 3(a) , and 3(c), respectively. Yet cross-correlation is able to discover only X_1 as the second closest user while X_{13} is

Table 1.: Synthetic example of temporal data associated to 13 users and the corresponding usage during 7 hours, e.g. user 1 entered the public transit in the very early hour of day where the related index is 1.

User	1	2	3	4	5	6	7	...	24
X_1	1	0	0	0	0	0	0	...	0
X_2	0	1	0	0	0	0	0	...	0
X_3	0	0	1	0	0	0	0	...	0
X_4	0	0	0	1	0	0	0	...	0
X_5	0	0	0	0	1	0	0	...	0
X_6	0	0	0	0	0	1	0	...	0
X_7	0	0	0	0	0	0	1	...	0
X_8	1	1	0	0	0	0	0	...	0
X_9	1	0	1	0	0	0	0	...	0
X_{10}	0	1	1	0	0	0	0	...	0
X_{11}	1	0	0	1	0	0	0	...	0
X_{12}	0	0	0	0	1	1	0	...	0
X_{13}	0	0	0	0	0	1	1	...	0

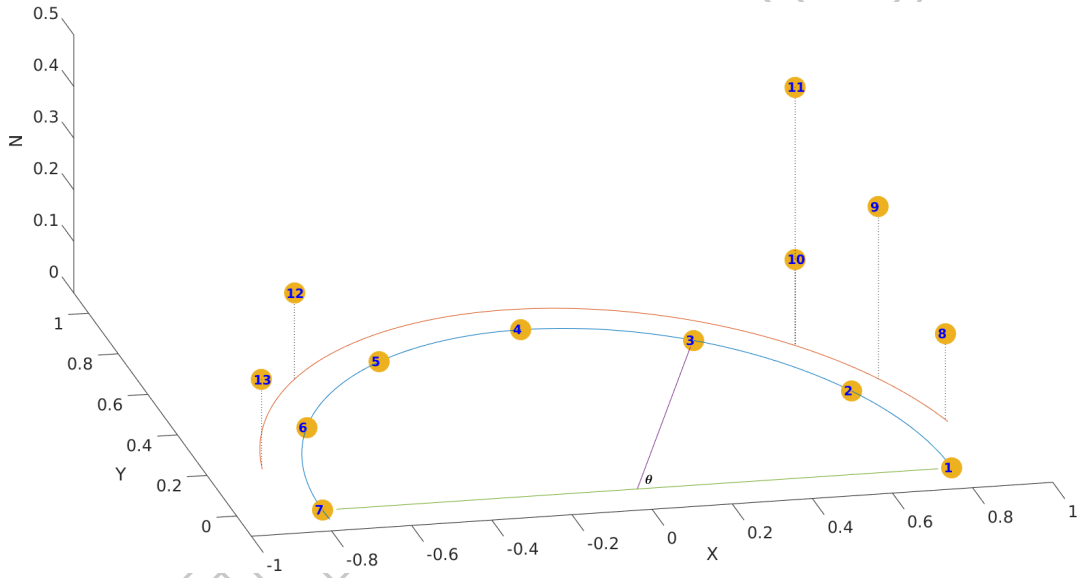
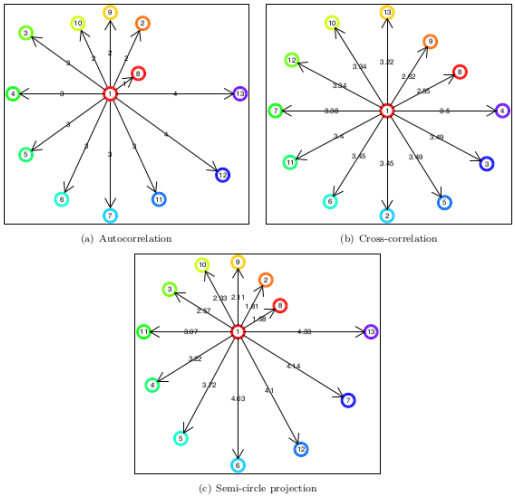


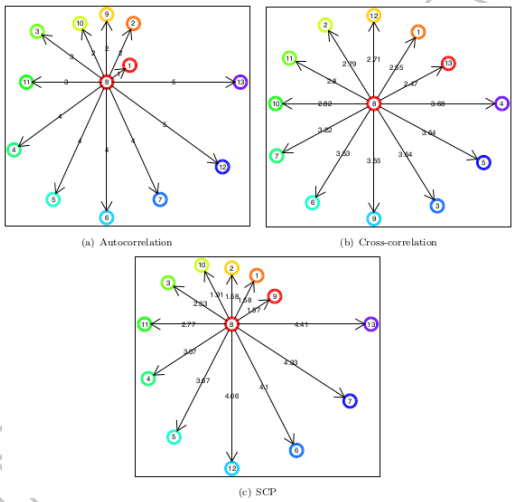
Figure 1.: Result of the Semi-Circle Projection on the synthetic dataset from Table 1 in three dimension which illustrates how similar users are located close to each other.

chosen as the first nearest similar user. Apparently, cross-correlation is not well-tailored to extract the similar users according to the temporal pattern. Regarding the discrete values of the autocorrelation distance that is redundant for couple pairs, e.g. in Figure 3(a), the same distance is assigned between four pairs, (X_8, X_4), (X_8, X_5), (X_8, X_6), and (X_8, X_7) which should not be the same. However, the correct order with associated distance is restrained by the SCP method. Moreover, the time series measurements are designed to give a value for a pair of vectors which requires $\binom{n}{2}$ flops. The SCP projects each data into a lower space independently to demonstrate the data in the reduced space with less computational complexity. The computational complexity of the SCP is of order $\mathcal{O}(n)$, where n is the number of projecting users. In Figure 1, the projected users from Table 1 into 3D space is shown where the aforementioned constraints are still kept.



(a) Autocorrelation (b) Cross-correlation (c) Semi-circle projection

Figure 2.: Comparison of the nearest users of X_1 with three similarity measurements, autocorrelation, cross-correlation, and semi-circle projection, respectively. As we expect, observations show that SCP method effectively sort out the similar users according to the temporal usage related to the user 1.



(a) Autocorrelation (b) Cross-correlation (c) SCP

Figure 3.: Comparison of the nearest users of X_8 with three measures of similarity, autocorrelation, cross-correlation, and semi-circle projection, respectively. As it could be seen, SCP is able to find out the analogous users by projecting them into three dimensions.

5.2. *Experimenting the SCP method on Gatineau dataset*

Société de transport de l'Outaouais (STO) in Gatineau, Québec, Canada, provides the data of this study. The STO authority has started to use smart card system since 2001 in

its 200-buses network. Everyday, data of every transaction is gathered from public transit users at bus stops boarding passengers. For each transaction, the following properties are present:

(1) Date and time of the boarding transaction;
(2) Card number and fare type;
(3) Route number and direction;
(4) Vehicle and driver numbers;
(5) 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. Additionally, we suggest to encode the temporal data into a $0-1$ vector whereas 24 binary vector associated to the daily hours. In this vector, occurrence of 1 at a specific index represents the usage of smart card at the corresponding hour. To deal with the binary values, discrete structure is usually suggested as the first option to entail the data for further process.

This projection method is tested on the mid-size authority (300 buses and $220,000$ inhabitants), over one month period in April 2009 (data is gathered from $753,016$ transactions, with $26,176$ unique users and $416,076$ card-days). From the first analysis of usage histogram shown in Figure 4, it turns out a large subset of users prefer to take the public transit between 15-20 days per month, on average. Figure 5 (3D histogram of

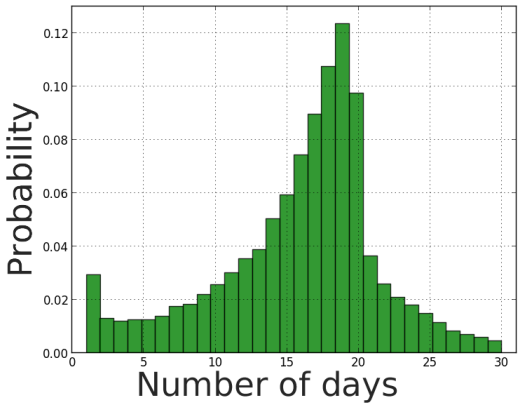


Figure 4.: Histogram of the frequency of the travelled days in one month.

the projected users on xy-plane) demonstrates how many users overlapped on the same point without the z-axis that captures the standard deviation of the timestamps. In other words, the frequency of this histogram states the proliferation of the same sum of usage indices for different behaviors. Moreover, this illustrates the peak of the half-circle has the highest density which reflects the existence of a meaningful pattern depicted in Figure 5.

The dendrogram in Figure 6(a) shows the visual aggregation of users on the projected data. In Figure 6(b) existing clusters for the cutting point are illustrated. Vertical axis represents similarity measure between clusters. Similar users are grouped in the bottom of the dendrogram; higher in the hierarchy, clusters are grouped together. The closer the groups are the more in the bottom, the more different they are bigger is the dissimilarity and higher are the steps in the dendrogram. It is then easy to identify possible cuts in the dendrogram that will stop the grouping process where too much dissimilar clusters are merged together; it is simply about translating the red line from the top to the bottom in order to identify big steps in the dendrogram.

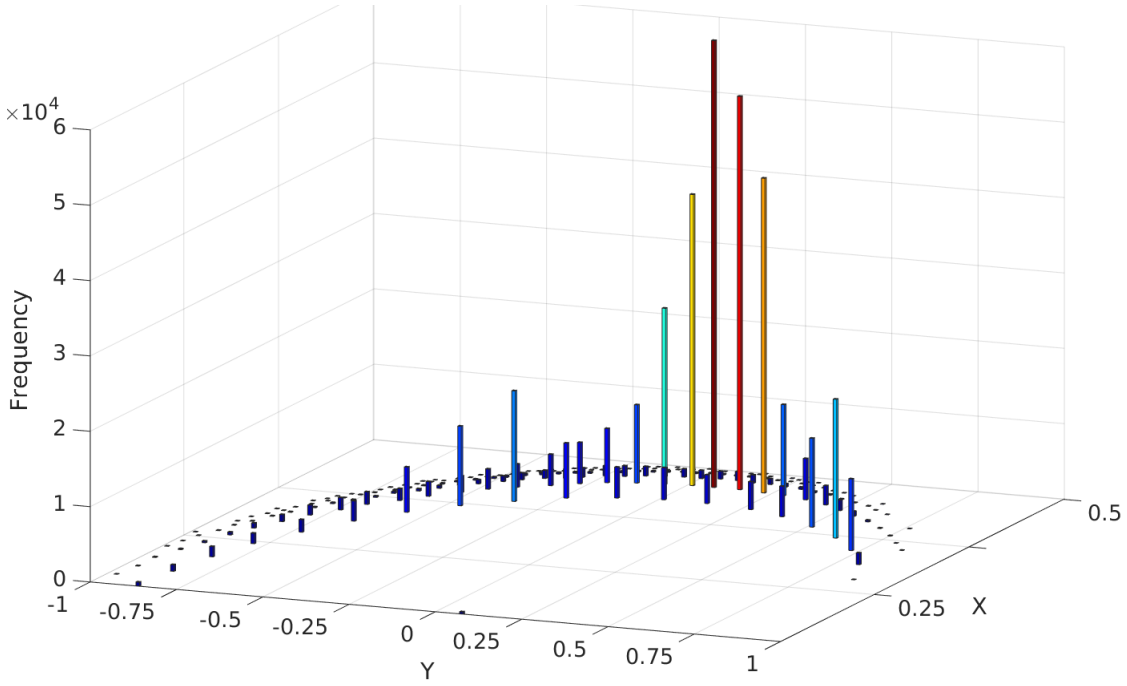


Figure 5.: 3D histogram of the overlapped projected data on xy-plane.

Different options may be possible. From Figure 6(a), cuts in $2, 3, 4, 6, 9$ or 18 clusters are to be considered. For a similarity/dissimilarity perspective they are close options, besides for an expert in the application domain it is possible to differentiate between these options. Considering domain expertise, a cut on the top, in $2, 3$ or 4 clusters, will be completely unbalanced (few customers on the right will be separated from all other users on the left in another group), this option will not be useful for the context of explaining users behaviour, we would have one conclusion that applies to a large group, which is not really useful for the practitioner. Options in 9 or 18 groups are still available, both could be processed and compared. From a methodological aspect here, we selected 18 groups for a more accurate prospect. These clusters of user behavior in public transit are described as the following categories.

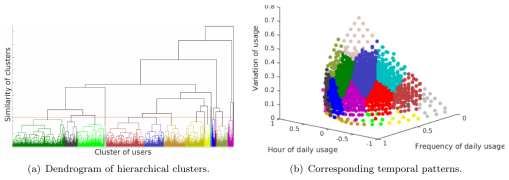
Single trip: as it is shown in Figure 7, significant number of patterns belong to this group ranging from early morning to late night, though with different number of users and distribution that is shown in Figure 6. Members of this group commute once or few times a day for one-way monadic trips at certain hours.

Regular commuters: four type of the users take the public transit regularly that could be seen in Figure 9. The pattern of these loyal users shows the frequent of usage all day long. Taking the number of users in this group into account, considerable portion of users are categorized as regular users who rely on public transit for their daily trip.

Late commuters: another category of users is determined in Figure 11, which demonstrates typical evening and late night usages for the most expeditions occurring or done on many occasions. These users usually enter the public transit network after the work for different purposes or come back to home late night.

Long day: this category is characterized by a two-peak distribution of the transactions during a typical day of travel shown in Figure 12. This is generic schedule of morning and evening peak period travel time.

Midday versus long day: the patterns of long day users and midday travellers are shown in Figure 12(a) and Figure 12(b), respectively. The former group of users usually behave



(a) Dendrogram of hierarchical clusters. (b) Corresponding temporal patterns.

Figure 6.: Dendrogram of the hierarchical clustering with the associated clusters of the projected data. Figure 6(a), shows 18 clusters form the total temporal patterns that exist for the one month period of the smart card usage. These clusters are shown on the projected data, in Figure 6(b).

as a combination of regular users and late commuters. This reflects the fact that long day users are intrinsically take the public transit for habitual commuting to work and late night circulation. In analogy, the subscriber of the latter cluster, are more similar to the late commuters whose pattern is shifter over to left. This implies the spread of transactions revolves around lunch peak time and evening rush hour.

Active versus inactive: Figure 13 shows the active users and the inactive smart cards' behaviors. Active users never miss any bus along their way as it could be seen in Figure 13(a). However, few users never used their smart card for the given one month interval with null pattern as it is shown in Figure 13(b). These two groups of user have the most extreme behavior in the public transit network in comparison to the remaining ones.

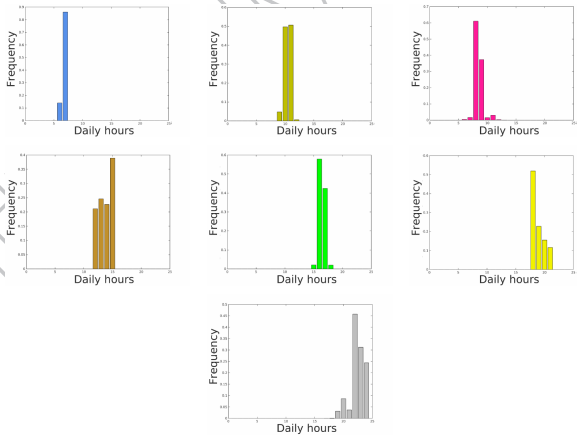


Figure 7.: Pattern of single trips ordered by early to late.

Let us look at the proportion of cards-day corresponding to each cluster in Figure 14, by working days, Saturdays and Sundays for the duration of the given month. It shows that during the working days (from Monday to Friday), the proportion of regular and single clusters (pendulum AM-PM trips) are much higher than the other ones. However, the proportion of the late commuters and the active clusters increased over the weekend, while a sharp drop of regular users is seen. This happens because people move later in

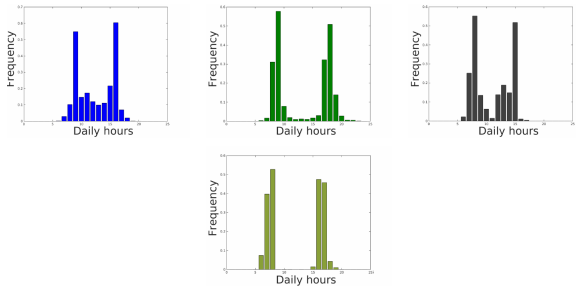


Figure 8.: Autocorrelation distance

Figure 9.: Pattern of regular users.

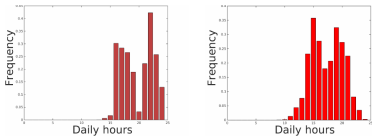
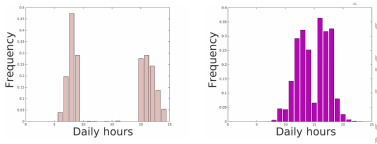


Figure 10.: Autocorrelation distance

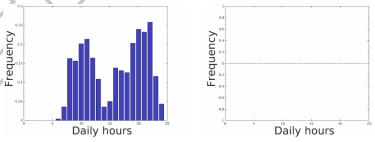
Figure 11.: Patterns of late commuters.



(a) Long day pattern (b) Midday pattern

(a) Long day pattern (b) Midday pattern

Figure 12.: Patterns of long-day trips vs midday excursion.



(a) Active users (b) Inactive smart cards

(a) Active users (b) Inactive smart cards

Figure 13.: Patterns of active users versus inactive cards.

the afternoon, and the trips are less characterized by pendulum movements like in the working days. It is also interesting to look at the distribution of the cluster by the entire days of month shown in Figure 15 where the same patterns could be found scaled by the frequency of trips per day.

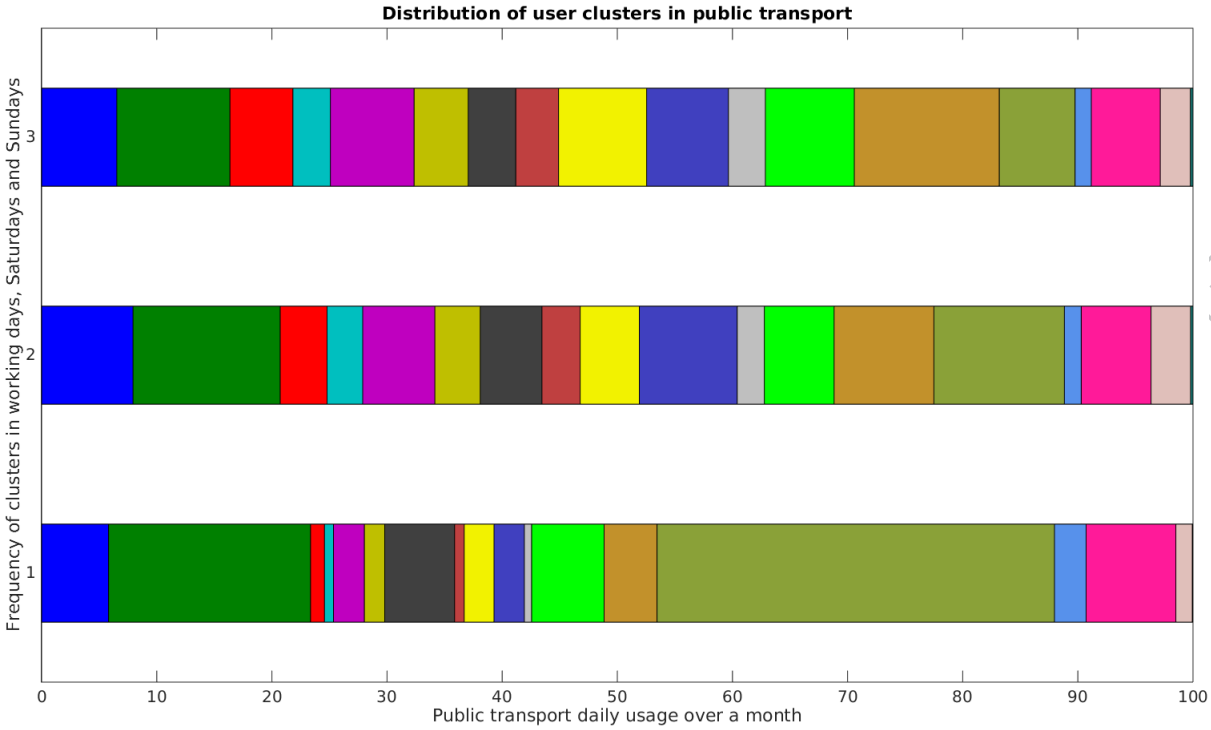


Figure 14.: Distribution of clusters shown in Figure 6 for usual working days and weekends.

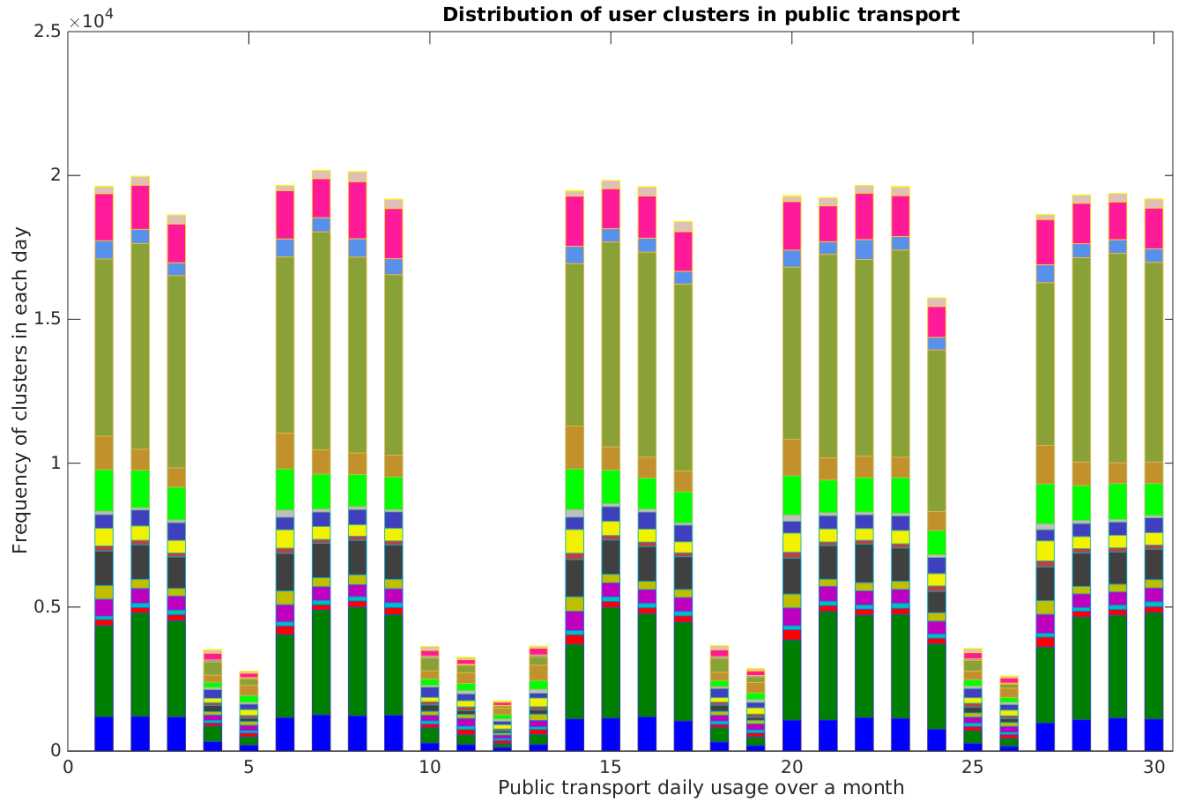


Figure 15.: Daily cluster distribution for the entire period of the month.

6. Conclusion and Discussion

User's behavior modeling is crucial for predicting future financial gain, transportation scheduling, and traffic load. Thus, the main objective of the data mining on the public transit data is uncovering people's behavior. We presented the analysis of the public transit smart card transactions by projecting the high-dimensional binary vector of the temporal data into a three dimensional semi-circle and three-dimensional space. The new representation of the data provides a visual guide to a better understanding of the temporal pattern. Seventeen clusters are identified in terms of single trip, regular users, late commuters, long day, midday, active and inactive groups as the temporal behavior of the users by applying agglomerative hierarchical clustering on the transformed data. Despite a continuous variable carries more information, binary data carries little amount of information compared to the continuous variable. This motivated us to transform a binary sequence to one or several continuous variable to execute a computationally efficient analysis. In this research study, 24 hours user-day pattern is used as the original data, however, our method is flexible to analyze even more complicated patterns such as 30 day user-day, or 365 user-day efficiently.

Most of the data mining algorithms are developed for continuous variables that we can take the advantage of them, if we properly transform binary data to continuous and informative space. Benefiting from a proper transformation we also gain computational feasibility through dimension reduction. Developing a particular data structure, one can decrease the computational time complexity of the hierarchical clustering algorithm from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log n)$ or even $\mathcal{O}(n^2)$ by certain properties of the algorithm, where n is the number of users. Remembering the binary vector of length 24×30 for each individual using the public transit in one month, if only 1000 people use the public transit, the amount of storage and computing facility required for analysis of such data with recent data mining algorithms is cumbersome, even with today computational power. The issue becomes worse if we analyze data of several years.

Several issues arise as future directions of this work. First, there is a need to define an equivalent metric on the binary space corresponding to the Euclidean measure on the projected three dimensional space. Second, the analysis of spatial data remains as the open question for our future research because of the existence of complex scenarios which require sophisticated techniques to compute the similarity of the users. Third, the technique can be applied to other sorts of vectors, not only including timestamps, but also the location of boarding on the territory, the route sequences, route types, etc. if the data are encoded in a binary vector.

7. Acknowledgements

The authors wish to acknowledge the support of Société de Transport de l'Outaouais (STO), who provided the data for this study, and special thanks to Thalès and Natural Sciences and Engineering Research Council (NSERC) of Canada RDCPJ-446107-12 for supporting this project financially. Vahid Partovi Nia is partially supported by the Canada excellence research chair in data science for real-time decision making.

8. Appendix

8.1. *Proof of Theorem 1*

The rate of radious growth is decreasing by increase in boarding.

$$\begin{aligned} \frac{1}{n_i+1} &< \log(1+\frac{1}{n_i}) \\ \frac{n_i}{n_i+1} &< n_i \log(1+\frac{1}{n_i}) \\ \frac{n_i}{n_i+1} &< \log(r_{n_i}) \\ -\log(r_{n_i}) &< -\frac{n_i}{n_i+1} \end{aligned} \tag{1}$$

$$\begin{aligned} \log(1+\frac{1}{n_i+1}) &< \frac{1}{n_i+1} \\ (n_i+1)\log(1+\frac{1}{n_i+1}) &< \frac{n_i+1}{n_i+1} \\ \log(r_{n_i+1}) &< 1 \end{aligned} \tag{2}$$

By adding the inequalities 1, 2 we have,

$$\begin{aligned} \log(r_{n_i+1}) - \log(r_{n_i}) &< 1 - \frac{n_i}{n_i+1} \\ \log(r_{n_i+1}) - \log(r_{n_i}) &< \frac{1}{(n_i+1)} \\ \frac{r_{n_i+1}}{r_{n_i}} &< \exp\left\{\frac{1}{n_i+1}\right\} \blacksquare \end{aligned}$$

8.2. *Proof of Theorem 2*

The range of angle is decreasing by increase in boarding. We start with an example where $n_i = 3$. Then $X_{\underset{i}{\operatorname{argmin}}\,\tilde{\theta}_{n_i=3}} = [1,1,1,\dots]$ and $X_{\underset{i}{\operatorname{argmax}}\,\tilde{\theta}_{n_i=3}} = [\dots,1,1,1]$. Therefore, the maximum range of angle $|\tilde{\theta}_i - \tilde{\theta}_{i'}|$ for a set of temporal usages with the same boarding varies between $\min\tilde{\theta}_i$, and $\max\tilde{\theta}_i$ that is $[\frac{n_i(n_i+1)/2}{n_iL}, \frac{n_iL-n_i(n_i+1)/2}{n_iL}]$ according to the definition, this implies the range of angle is shrinking by $(1-\frac{n_i+1}{L})$.

8.3. *Proof of Theorem 3*

The range of variance is decreasing by increase in boarding. In this section we use the alternative definition of variance that can be defined as,

$$\frac{1}{n-1}\sum_{i=1}^{n}(\theta_i-\bar{\theta}) = \sum_{i,j=1,i\neq j}^{n}\frac{(\theta_i-\theta_j)^2}{2(n-1)^2}$$

In order to prove the decrease of variance we have to show that minimum of variance is monotonically increasing by increase in boarding while the maximum of variance is monotonically decreasing.

Minimum variance of the temporal usages is monotonically increasing by increase in boarding.

Proof by induction.

First of all, we show that our claim is true for the first step.

$$\min z_{n_i=2} = .5 < \min z_{n_i=3} = 1$$

this is true according to the definition where it occurs at $X_{\underset{i}{\operatorname{argmin}}\, z_{n_i=2}} = [1, 1, \dots]$, and $X_{\underset{i}{\operatorname{argmin}}\, z_{n_i=3}} = [1, 1, 1, \dots]$.

Let $s_n = \sum_{i,j=1, i\neq j}^{L} (\theta_i - \theta_j)^2$ now, we suppose that $\frac{s_n}{2(n-1)^2} < \frac{s_{n+1}}{2n^2}$ is true, then we want to show that $\frac{s_{n+1}}{2n^2} < \frac{s_{n+2}}{2(n+1)^2}$ also holds based on the first assumption.

$$\begin{aligned} \frac{s_n}{2(n-1)^2} < \frac{s_{n+1}}{2n^2} &= \frac{s_n + \sum_{i=1}^{n} (n+1-i)^2}{2n^2} \\ \frac{(n-1)^2}{n^2} &< \frac{n^2}{(n+1)^2} \end{aligned} \tag{3}$$

Thus by multiplying two inequalities in 3 we have,

$$\frac{s_n}{2n^2} < \frac{s_n + \sum_{i=1}^{n} (n+1-i)^2}{2(n+1)^2} \tag{4}$$

Lemma 4. *For any $n \in \mathbb{N}$, we have*

$$\frac{\sum_{i=1}^{n} (n+1-i)^2}{2n^2} < \frac{\sum_{i=1}^{n+1} (n+2-i)^2}{2(n+1)^2} \tag{5}$$

Proof.

$$\begin{aligned} 2n^3 + 5n^2 + 4n + 1 &< 2n^3 + 7n^2 + 6n \\ \frac{2n^2 + 3n + 1}{2n} &< \frac{2n^2 + 7n + 6}{2(n+1)} \\ \frac{n(n+1)(2n+1)}{2n^2} &< \frac{(n+1)(n+2)(2n+3)}{2(n+1)^2} \\ \frac{\sum_{i=1}^{n} (n+1-i)^2}{2n^2} &< \frac{\sum_{i=1}^{n+1} (n+2-i)^2}{2(n+1)^2} \end{aligned}$$

Thus by adding two inequalities from 4, and 5 we have,

$$\frac{s_n + \sum_{i=1}^{n} (n+1-i)^2}{2(n)^2} < \frac{s_n + \sum_{i=1}^{n} (n+1-i)^2 + \sum_{i=1}^{n+1} (n+2-i)^2}{2(n+1)^2}$$

That is,

$$\frac{s_{n+1}}{2n^2} < \frac{s_{n+2}}{2(n+1)^2} \blacksquare$$

Maximum variance of the temporal usages is monotonically decreasing by increase in boarding.

Proof by induction.

First of all, we show that our claim is true for the first step.

$$\max z_{n_i=2} = 264.5 > \max z_{n_i=3} = 169$$

this is true according to the definition where it occurs at $X_{\underset{i}{\text{argmax}}\, z_{n_i=2}} = [1, \ldots, 1]$, and $X_{\underset{i}{\text{argmax}}\, z_{n_i=3}} = [1, 1, \ldots, 1]$.

Now, we suppose that $\frac{s_n}{2(n-1)^2} > \frac{s_{n+1}}{2n^2}$ is true, then we show that $\frac{s_{n+1}}{2n^2} > \frac{s_{n+2}}{2(n+1)^2}$ based on the first assumption.

By setting $A = S_n, B = S_{n+1} - S_n$ such that $B = \sum_{i=1}^{\frac{n}{2}} (\frac{n}{2} + 1 - i)^2 + \sum_{i=1}^{\frac{n}{2}} (L - \frac{n}{2} - i)^2$, and $C = L - \frac{n}{2} - \frac{n}{2} - 1$, suppose,

$$\frac{A}{2(n-1)^2} > \frac{A+B}{2n^2} \tag{6}$$

Now, we have to show that the following inequality is true.

$$\frac{A+B}{2n^2} > \frac{A+2B+C}{2(n+1)^2}$$

By expanding equation 6 we have,

$$\begin{aligned} n^2 A &> (n-1)^2 A + (n-1)^2 B \\ (2n-1)A &> (n-1)^2 B \\ (2n+1)A &> (n-1)^2 B + 2A \end{aligned} \tag{7}$$

Thus we have to show that, the following inequality is true.

$$\begin{aligned} (n+1)^2 A + (n+1)^2 B &> n^2 A + 2n^2 B + n^2 C \\ (2n+1)A &> (n^2 - 2n - 1)B + n^2 C \end{aligned}$$

From 7 we know that $(2n+1)A > (n-1)^2 B + 2A$, now it is sufficient to show that $(n-1)^2 B + 2A > (n^2 - 2n - 1)B + n^2 C$.

By rearranging $(n-1)^2 B + 2A > (n^2 - 2n - 1)B + n^2 C$ we should show that $A + B > \frac{n^2}{2} C$ holds.

$$A + B = 2 \sum_{j=1}^{n+1} \left[\sum_{i=1}^{\frac{j}{2}} (\frac{j}{2} + 1 - i)^2 + \sum_{i=1}^{\frac{j}{2}} (L - \frac{j}{2} - i)^2 \right]$$

$$\sum_{i=1}^{\frac{j}{2}}(\frac{j}{2}+1-i)^2 = \frac{\frac{j}{2}(\frac{j}{2}+1)(j+1)}{6} > \frac{j^3}{24}$$

$$\sum_{i=1}^{\frac{j}{2}}(L-\frac{j}{2}-i)^2 > \frac{j}{2}C' > \frac{j}{2}C$$

$$A+B > \sum_{j=1}^{n}(\frac{j^3}{12}+jC) > \frac{(n+1)^4}{24}+\frac{(n+1)^2}{2}C > \frac{n^2}{2}C\blacksquare$$

Therefore, we prove that the minimum variance of temporal usages is monotonically increasing and maximum variance of temporal usages is monotonically decreasing by increase in boarding. This implies the range of variance is also monotonically decreasing by increase in boarding and for the extreme usage where one enter the network at every single hour the minimum and maximum variance is collapsed on the same point.

References

Agard, Bruno, Catherine Morency, and Martin Trépanier. 2008. "Mining Smart Card Data from an Urban Transit Network.." In *Encyclopedia of Data Warehousing and Mining,* edited by John Wang. 1292–1302. IGI Global.

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle." In *Second International Symposium on InformationTheory,* edited by B. N. Petrov and F. Csaki. 267–281.

Ali, Atizaz, Jooyoung Kim, and Seungjae Lee. 2016. "Travel behavior analysis using smart card data." *KSCE Journal of Civil Engineering* 20 (4): 1532–1539.

Alsger, Azalden, Behrang Assemi, Mahmoud Mesbah, and Luis Ferreira. 2016. "Validating and improving public transport origindestination estimation algorithm using smart card fare data." *Transportation Research Part C: Emerging Technologies* 68: 490 – 506.

Bordagaray, Maria, Luigi dell'Olio, Angel Ibeas, and Patricia Cecín. 2014. "Modelling user perception of bus transit quality considering user and service heterogeneity." *Transportmetrica A: Transport Science* 10 (8): 705–721.

Borg, I., and P.J.F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications.* Springer.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* 61 (6): 1–36. http://www.jstatsoft.org/v61/i06/.

de Oña, Rocio, and Juan de Oña. 2015. "Analysis of transit quality of service through segmentation and classification tree techniques." *Transportmetrica A: Transport Science* 11 (5): 365–387.

Del Castillo, JM, and FG Benitez. 2013. "Determining a public transport satisfaction index from user surveys." *Transportmetrica A: Transport Science* 9 (8): 713–741.

Fuse, T., K. Makimura, and T. Nakamura. 2010. "Observation of travel behavior by ic card data and application to transportation planning." In *Special Joint Symposium of ISPRS Commission IV and AutoCarto, .*

Gallotti, Riccardo, and Marc Barthelemy. 2015. "The multilayer temporal network of public transport in Great Britain." *Scientific data* 2: 140056.

Gkiotsalitis, Konstantinos, and Antony Stathopoulos. 2015. "A utility-maximization model for retrieving users willingness to travel for participating in activities from big-data." *Transportation Research Part C: Emerging Technologies* 58, Part B: 265 – 277.

Hasan, S., C. M. Schneider, S. V. Ukkusuri, and M. C. Gonzlez. 2012. "Spatiotemporal patterns of urban human mobility." *Statistical Physics* 151 (1-2): 304–318.

Hastie, Trevor J., Robert J. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Heard, N. A., C. C. Holmes, and D. A. Stephens. 2006. "A quantitative study of gene regulation involved in the immune response of *Anopheline* mosquitoes: An application of Bayesian hierarchical clustering of curves." *Journal of the American Statistical Association* 101 (473): 18–29.

Heller, K. A., and Z. Ghahramani. 2005. "Bayesian hierarchical clustering." In *Twenty-second International Conference on Machine Learning,* 297–304.

Herrera, Juan C., Daniel B. Work, Ryan Herring, Xuegang (Jeff) Ban, Quinn Jacobson, and Alexandre M. Bayen. 2010. "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment." *Transportation Research Part C: Emerging Technologies* 18 (4): 568 – 583.

Järv, Olle, Rein Ahas, and Frank Witlox. 2014. "Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records." *Transportation Research Part C: Emerging Technologies* 38: 122 – 135.

Kieu, Le Minh, Ashish Bhaskar, and Edward Chung. 2014. "Transit passenger segmentation using travel regularity mined from Smart Card transactions data." In *Transportation Research Board 93rd Annual Meeting,* Washington, D.C. January.

Kusakabe, Takahiko, and Yasuo Asakura. 2014. "Behavioural data mining of transit smart card data: A data fusion approach." *Transportation Research Part C: Emerging Technologies* 46: 179–191.

Lathia, Neal, and Licia Capra. 2011. "How Smart is Your Smartcard: Measuring Travel Behaviours, Perceptions, and Incentives." In *Proceedings of the 13th International Conference on Ubiquitous Computing,* Beijing, China. 291–300. New York, NY, USA: ACM.

Ma, Tai-Yu, Philippe Gerber, Samuel Carpentier, and Sylvain Klein. 2015. "Mode choice with latent preference heterogeneity: a case study for employees of the EU institutions in Luxembourg." *Transportmetrica A: Transport Science* 11 (5): 441–463.

Ma, Xiaolei, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. 2013. "Mining smart card data for transit riders' travel patterns." *Transportation Research Part C: Emerging Technologies* 36: 1 – 12.

Mahrsi, M.K. El, E. Côme, J. Baro, and L. Oukhellou. 2014. "Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France." In *3rd International Workshop on Urban Computing (SigKDD),* .

McLachlan, G.J., K.-A. Do, and C. Ambroise. 2004. *Finite Mixture Models*. New York: Wiley.

Montero, Pablo, and Jos A. Vilar. 2014. "TSclust: An R Package for Time Series Clustering." *Journal of Statistical Software* 62 (1): 1–43. http://www.jstatsoft.org/v62/i01.

Morency, Catherine, Martin Trépanier, and Bruno Agard. 2006. "Analysing the Variability of Transit Users Behaviour with Smart Card Data." In *Intelligent Transportation Systems Conference, 2006. ITSC'06.*, 44–49. IEEE.

Morency, Catherine, Martin Trépanier, Daniel Piché, and Robert Chapleau. 2010. "Bridging the gap between complex data and decision-makers: an example of an innovative interactive tool." *Transportation Planning and Technology* 33 (6): 465–479.

Nantes, Alfredo, Dong Ngoduy, Ashish Bhaskar, Marc Miska, and Edward Chung. 2015. "Real-time traffic state estimation in urban corridors from heterogeneous data." *Transportation Research Part C: Emerging Technologies* .

Ortega-Tong, Meisy A. 2013. "Classification of London's public transport users using smart card data." Master's thesis. Massachusetts Institute of Technology. Department of Civil and Environmental Engineering.

Pelletier, Marie-Pier, Martin Trépanier, and Catherine Morency. 2011. "Smart card data use in public transit: A literature review." *Transportation Research Part C: Emerging Technologies* 19 (4): 557–568.

Schwarz, Gideon E. 1978. "Estimating the dimension of a model." *Annals of Statistics* 6: 461–464.

Shekhar, Shashi, Zhe Jiang, Reem Y. Ali, Emre Eftelioglu, Xun Tang, Venkata M. V. Gunturi, and Xun Zhou. 2015. "Spatiotemporal Data Mining: A Computational Perspective." *ISPRS*

International Journal of Geo-Information 4 (4): 2306–2338.

Sneath, P. H. 1957. "The application of computers to taxonomy." *Journal of General Microbiology* 17 (1): 201–226.

Trépanier, Martin, Nicolas Tranchant, and Robert Chapleau. 2007. "Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System." *Journal of Intelligent Transportation Systems* 11 (1): 1–14.

Weisbrod, Glen, and Arlee Reno. 2009. *Economic impact of public transportation investment.* American Public Transportation Association.

Yahya, Saadiah, and Noriani Mohammed Noor. 2008. "Strategic Planning of an Integrated Smart Card Fare Collection System - Challenges and Solutions." In *Proceedings of the 2008 11th IEEE International Conference on Computational Science and Engineering,* 31–36. Washington, DC, USA.