CASPT 2015

Public Transport

Identifying Temporal User Behavior through Smart Card Data

Mohammad Sajjad Ghaemi · Bruno Agard · Vahid Partovi Nia · Martin Trépanier

Abstract Nowadays, tremendous data are continuously gathering from the smart card in public transport domain. Such data, conveying two viable distinct information, can assist designing public transportation network. The first component of the data, provides the spatial feature, that indicates the geographical coordinates of bus stops or subway stations. The second component of the data deals with the temporal feature, which is the time of the trips that public transport is used. Hence, it is necessary to distill the data, in order to get the advantages of the data analysis techniques and extract the essential knowledge. More specifically, user behavior in a public transport system can be investigated as one of the data mining and machine learning applications. Extracting this information could lead analysts, engineers, managers, and strategists to excavate, design, decide, and plan more effectively.

Mohammad Sajjad Ghaemi
GERAD and CIRRELT Research Center, and Department of Mathematical and Industrial Engineering, Polytechnique Montreal
E-mail: m.s.ghaemi@gmail.com

Bruno Agard
CIRRELT Research Center, and Department of Mathematical and Industrial Engineering, Polytechnique Montreal
Tel.: 1-514-340-4711 ext 4914
Fax: 1-514-340-4173
E-mail: bruno.agard@polymtl.ca

Vahid Partovi Nia
GERAD Research Center, and Department of Mathematical and Industrial Engineering, Polytechnique Montreal
Tel.: 1-514-340-4711 ext 2349
E-mail: vahid.partovinia@polymtl.ca

Martin Trépanier
CIRRELT Research Center, and Department of Mathematical and Industrial Engineering, Polytechnique Montreal
Tel.: 1-514-340-4711 ext 4911
Fax: 1-514-340-4173
E-mail: mtrepanier@polymtl.ca

This makes the usage of the network practically efficient especially in large metropolitan cities. In this regard, we propose new methods of temporal data analysis to investigate pattern of user behavior in the public transport network.

Keywords Clustering · Public transport · Smart card · Temporal data

1 Introduction

The importance of the public transportation and its influence on the real life of many people in cities around the world, rises a new family of problems that is not confined into a particular branch of science. Hence, usage of the smart card data collected from automated payement systems, creates the opportunity for several different researchers from diverse disciplines e.g. data mining, machine learning, urban computing and planning, management, business, civil engineering, industrial engineering, statistics, mathematical engineering, geographic information system (GIS), etc. to outreach and extend their methods to analyze the data for the public transport authorities.

In most of the models, bus stops and subway stations play the central role regardless of the temporal features. The frequency of the used locations is utilized to construct the model specifying the user behavior. This knowledge can be helpful to provide particular services in each station or bus stop. Nonetheless, they are incapable of clarifying user similarity or behavioral pattern to discover homogeneous groups of users who have the same manner. In other research works, a number of measurements 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 the clustering algorithms without having any well-founded justification or explanatory translation.

Despite extensive researches have been done on public transportation domain, various obstacles have been arisen for specific purposes which require particular approaches to address them. In this study, a recent concerning problem of user clustering is introducing according to the temporal data gathered from smart cards to analyze their behavioral trip patterns in the public transport network.

Enlargement and expansion of the public transport systems which have formed independently in different cities while they are in the same regional state or country, reflects the necessity of having a strategic plan of Integrated Smart Card Fare Collection System (ISFCS). ISFCS can fill the gap of different public transport operators and also it can meet the passengers' needs and satisfactions as well. Barriers of ISFCS and their possible solutions are discussed in Yahya and Noor (2008). In Pelletier et al (2011) several other aspects of ISFCS are considered from technologies to privacy issues in three level of managements including, strategic, tactical, and operational. Moreover, discussion and comparison of planning, scheduling and survival modeling for many different purposes rather than what the smart cards are really designed for, are provided in Pelletier et al (2011).

Describing users behavior in public transport network is one of the main issues that can be revealed via the smart cards data. 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 et al (2006), by considering the variability of users behavior with smart card data, collected over a ten months period. In Lathia and Capra (2011), two viewpoints are investigated to measure the transport system's performance; self-report of users' feedback and their real behavior versus change of users behavior when they are encouraged by various incentives. Finally, authors concluded that smart card data is as important as human activity from mobile phone data for designing future infrastructure and guidance of travelers in Lathia and Capra (2011). Therefore, human mobility could be modeled according to the smart card data as one of the big data sources from human activity.

Smart card data contains worthwhile digital information of daily locations visited at certain period of a large number of individuals. Beside other sources of information such as mobile phone, GPS tracker vehicle, e.g. bike, car, motorcycle, credit card transactions, social network, and many other sources of information gathering, smart card data is the best promising source of users digital information. Thus this helpful information could be utilized to characterize and model urban mobility patterns Hasan et al (2012). 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 et al (2010).

Predicting users' location according to the popular locations as a result of users' interaction in the city, is modeled as a spatial-temporal pattern of human mobility in Hasan et al (2012). Data mining approach is used to understand passenger's temporal behavior so as to exploit the interpretable clusters in Mahrsi et al (2014). This approach can help transport operators to satisfy the customers' demands. In addition, it enables them to maintain their services and tools to meet the pleas of users 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 in this approach. Furthermore, the cluster of similar temporal passengers extracted based on their boarding time, according to the generative model-based clustering approach. Then after, the effect of distribution of socioeconomic characteristics on the passenger temporal clusters are investigated in this study.

As another example, the extensive database of Oyster Card transactions obtained from London's public transport users, is utilized in Ortega-Tong (2013). This database is deployed to classify users based on the temporal and the spatial variability, the sociodemographic characteristics, the activity patterns, and the membership. Improving the planning and the design of market research are the aim of this work, when selecting groups of homogeneous people is case 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 travelers, and finally visitor travelers for tourism and business affair are investigated in this work.

Smart card data gathered from Brisbane, Australia is another source of information that is studied in Kieu et al (2014) for strategic transit planning according to the individual travel patterns. Origins and destinations that the cardholder usually travels between is defined as travel regularity, and the definition of habitual time is the regular time of travel for each regular origins and destinations. 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, origins and destinations, number of transfers, mode and route uses, total 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 et al (2014).

The vast majority of public transport systems around the world are schedule-based. Schedules are a proper solution for the public transport user and for the public transport service provider. Most of the time, service providers operate on the same schedule for all the weekdays from Monday to Friday, and maintain distinct schedules for Saturdays and Sundays, assuming that the public transport 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. So it becomes more and more of interest of the service provider to measure or predict the amount of regularity of public transport users, using their time-stamped smart card transaction database. By applying learning methods on smart card database we aim to divide the users into several sub-populations 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).

In this paper, we propose a certain projection to satisfy the constraints between an arbitrary pair of binary temporal time series vector which can be used to find the groups of users with similar temporal behavior efficiently. Then an experimental simulation of one month record of smart card data is analyzed to extract the homogeneous cluster of users according to the temporal information.

2 Methodology

The learning methods are often divided into supervised, and unsupervised subfields, recently semi-supervised methods have attracted attention as well. All learning methods seek for dividing data into sub-populations. The difference

between supervised and unsupervised method is the existence of training data Hastie et al (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. Therefore, the supervised learning is a sub-problem of unsupervised learning. This is the reason to attack more general problem, namely the unsupervised learning.

2.1 Smart Card in Public Transport

Smart card data, usually provides 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 values. Temporal data describes the time each trip is taken, with our suggestion is encoded in a $0-1$ vector, where start of the trip is indicated by 1. According to these information, analysing users behavior is divided into three categories, 1) Spatial patterns, 2) Temporal patterns and 3) Spatial-temporal patterns.

1. In the first case, methods of spatial patterns analysing, are taking the bus stop's information into account. It turns out measure of behavioral patterns only depends on the location of bus stops taken by the users rather than having known the starting hour of their trip.
2. The second methods are seeking the information pertinent to the temporal data associated to the public transport usages. Consequently, computing user similarity score is carried out, by assuming bus stop information is unavailable. The indices of 1 occurrences in the encoded vector, are playing the central role in this approach.
3. The third scenario, is a mixture of spatial and temporal data, called spatial-temporal data analysis to investigate users behavior. It could be viewed as a combination of the last two steps or an independent approach to recognize spatial-temporal behavioral patterns in the public transport domain.

In Ghaemi et al (2015) a number of challenges to find the similarity of spatial sequence of location history of the users is enumerated, which is remained as the future direction of this research. In the rest of this section, we study the temporal aspect to exploit the users behavior according to the time they usually take the public transport. First of all, a projection is defined to transform the high-dimensional vector of temporal information into a 2D space. Next, similar group of users is discovered by deploying the hierarchical clustering approach.

2.2 Temporal Data Analysis

Our suggesting method is a simple mapping of a long binary sequence to the Cartesian coordinates. This suggestion is somehow a multi-dimensional

scaling Borg and Groenen (2005), when some equalities and inequalities are proposed for certain distance between individuals. The mapping, called *Semi-Circle Projection* (SCP) is easier to understand in the polar coordinate, i.e. in terms of radius and angle, because we suggest to map the binary sequence on a half circle. First, reserve the center of a half circle for the binary sequence of all zeros. For a binary sequence that have only one unit value, take radius equal to 1 and move the angle from 0 to π depending on the position of the unit value. For vectors with 2 unit values, we may take radius $r = 2$ to be sure these vectors do not fall over vectors with 1 unit values. Generalization for the binary sequence with n unit values is then straightforward. We choose $r = n$ and move the angle according to the average of the unit positions. However, the identity function $r_n = n$ diverges for large n. Choice of a converging r_n will help 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. After finding r and θ for each binary sequence, transformation to the Cartesian coordinates is easy through the well-known $x = r\cos(\theta)$ and $y = r\sin(\theta)$ projection. This seemingly simple transformation maps a binary sequence of any length to the Cartesian coordinates of only two dimensions (x, y). If we implement this method for the analysis of transport data, a binary vector of $d \times 24$, where d is the number of the traveled days, is compressed into only two dimensions, hugely facilitating further computation, analysis, and data visualization.

2.3 Agglomerative Hierarchical Clustering

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.

Hierarchical clustering is a breakthrough in the model-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 the 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, starts with all data in one cluster and splits the clusters until finishing with all singletons.

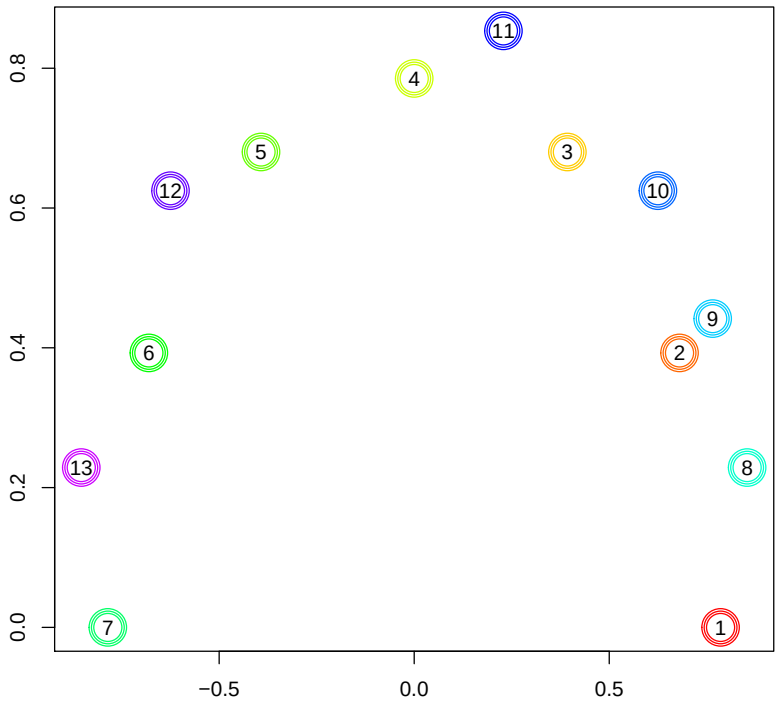


Fig. 1: Result of the SCP on the synthetic dataset

The nested groups generated using a hierarchical clustering algorithm of data, are visualized 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 dendgrogrm appeared in biology, but this term was used for the first time in Hochreiter et al (2010) 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.

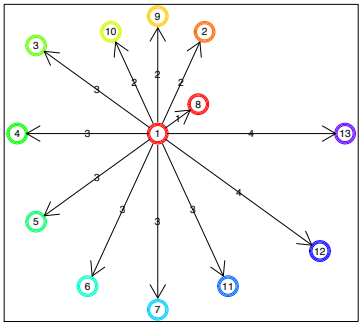
3 Experimenting the SCP method on the Gatineau dataset

After introducing the suggested ad-hoc SCP method, it should be compared with the other state-of-the-art time series distance measurements to illustrate the properties of the SCP and demonstrate how it can improve their 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 as by $D = \sqrt{\frac{(1-cc(x,y,0)^2)}{\sum(1-cc(x,y,k)^2)}}$, where $CC(x, y, k)$ is the cross-correlation between x and y at lag k, and the summatory in the denominator goes from 1 to `lag.max`. Autocorrelation-based dissimilarity, computes the dissimilarity between a pair of numeric time series based on their estimated autocorrelation coefficients that can be calculated as $D(x, y) = \sqrt{(\rho_x - \rho_y)^T \Omega (\rho_x - \rho_y)}$, where ρ_x, ρ_y are the estimated autocorrelation vectors of x and y respectively, and Ω is a matrix of weights Montero and Vilar (2014). A synthetic benchmark is considered as follows to investigate each distance measure's performance on it. The results of the three different

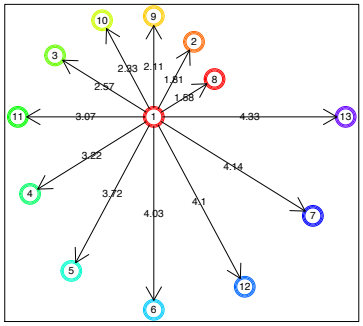
Table 1: Example of temporal data for distance calculation

User	H_1	H_2	H_3	H_4	H_5	H_6	H_7
X_1	1	0	0	0	0	0	0
X_2	0	1	0	0	0	0	0
X_3	0	0	1	0	0	0	0
X_4	0	0	0	1	0	0	0
X_5	0	0	0	0	1	0	0
X_6	0	0	0	0	0	1	0
X_7	0	0	0	0	0	0	1
X_8	1	1	0	0	0	0	0
X_9	1	0	1	0	0	0	0
X_{10}	0	1	1	0	0	0	0
X_{11}	1	0	0	1	0	0	0
X_{12}	0	0	0	0	1	1	0
X_{13}	0	0	0	0	0	1	1

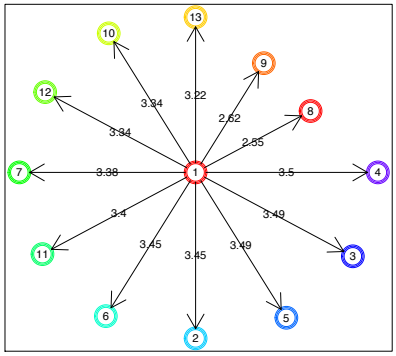
distance measures are shown in Fig 2, 3 for the users X_1, and X_8 respectively. $\{X_8, X_9, X_2\}$ could be considered as the first three nearest users to the user X_1 because of the similar time behavior. All three methods, indicate the user X_8 as the closest user to the user X_1 in Fig 2, however, X_9 is selected as the second nearest user in Fig 2c while the X_2 is selected in Fig 2a, 2b. 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 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



(a) Autocorrelation distance



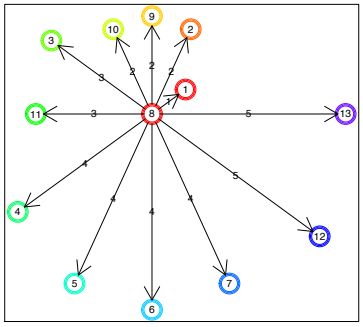
(b) SCP distance



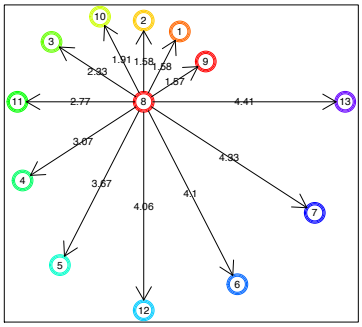
(c) Cross-correlation distance

Fig. 2: Comparison of the nearest users of X_1 with three measurements

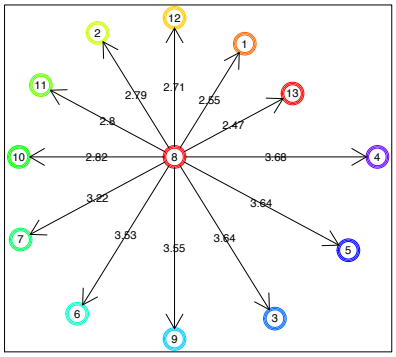
the performance of each method. $\{X_1, X_2, X_9\}$ are the first three candidates to be chosen as the nearest users to the X_8. In Fig 3, the selected users associated to the user X_8 are shown. Autocorrelation and SCP are capable of picking those users as are shown in Fig 3a, and 3b, respectively. Yet cross-correlation is able to discover only X_1 as the second closest user while X_{13} is 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 Fig 3a, 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. SCP projects each data into a lower space independently so that makes it possible to demonstrate the data in the reduced space with less computational complexity proportional to $\mathcal{O}(n)$, where n is the number of users. In Fig 1, the projected users in 2D space is shown where the aforementioned constraints are still kept.



(a) Autocorrelation distance



(b) SCP distance



(c) Cross-correlation distance

Fig. 3: Comparison of the nearest users of X_8 with three measurements

3.1 Data Analysis

This projection method is tested on the Société de transport de l'Outaouais data, over one month period (we have about 416 thousand transactions, with almost 26 thousand unique users). The first analysis in Fig 4 shows that, users usually take public transport between 15 to 20 times a month on average. Applying the 3D histogram on the projection of the binary vector of timestamps onto the semi-circle space, turns out the peak of the half-circle has the highest density which reflects the existence of a meaningful pattern depicted in Fig 5.

In Fig 6, the dendrogram of applying hierarchical clustering on the projected data, is shown which displays seven clusters that are illustrated in Fig 6. The most dominant cluster, containing regular users who usually take public transport as their routine schedule during the month frequently, is colored in red. Despite, the blue (early birds) cluster and the magenta (night persons) cluster are on the two opposite tails with different temporal usage behaviors, they are the most similar ones in terms of the number of users belonging to each cluster. The green cluster, represents users who usually prefer to com-

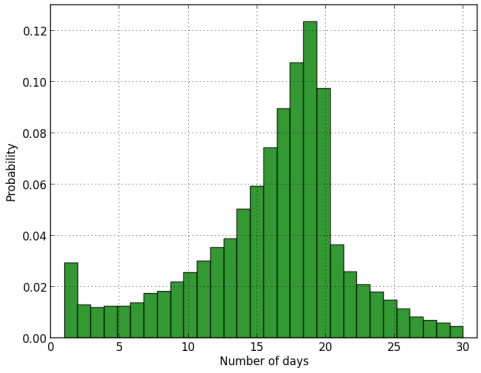


Fig. 4: Histogram of the frequency of the traveled days in one month

mute between the morning and noon rush hours. After noon and evening users are identified by the cyan and the yellow clusters respectively, where the last three clusters are the second prevalent clusters by the user proliferation. The last clusters is a singleton datum at origin $(0, 0)$ colored by black without any trip covered by public transport. The black cluster is shown as the right most leaf on the dendrogram in Fig 6.

4 Conclusion and Discussion

User's behavior modeling is crucial for predicting future financial gain, transport scheduling, traffic load, etc. Thus the main objective of the data mining on the public transport data is the discovery of peoples behavior. In this paper, we presented the analysis of the public transport smart card transactions by projecting the high-dimensional binary vector of the temporal data into a 2D semi-circle space. The new representation of the data provides a visual guide to better understand the temporal pattern. Seven clusters are identified as the temporal behavior of the users by applying the agglomerative hierarchical clustering on the transformed data, with informative demonstration. Despite a scale continuous variable carries more information, binary data carries little amount of information compared to the continuous variable. This motivates us to transform a binary sequence to one or several continuous variable to execute a computationally efficient analysis.

Furthermore, most of the data mining algorithms are developed for continuous variables that we can take advantage of them, if we properly transform binary data to continuous data. 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 from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log n)$ or even $\mathcal{O}(n^2)$ by

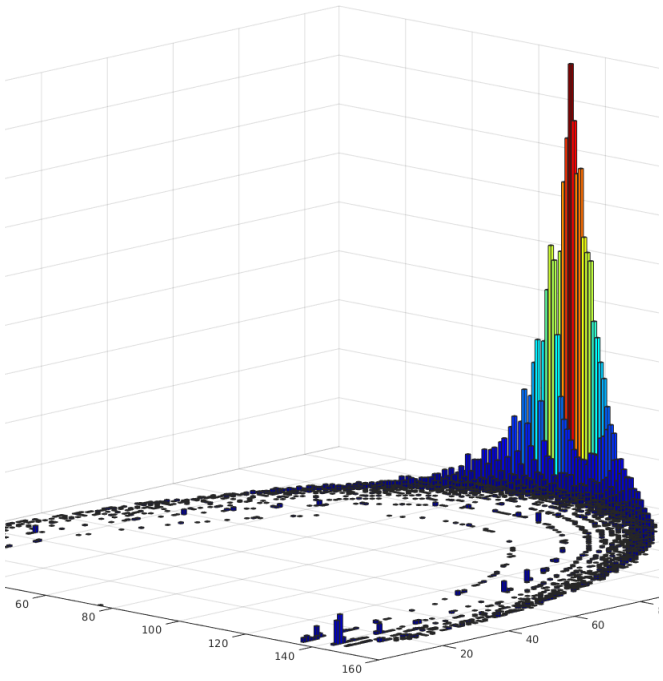


Fig. 5: 3D histogram of the projected data

certain properties of the algorithm, if n is the number of users. Remembering the binary vector of length 24×30 for each individual using the public transport in one month, if only 1000 people use the public transport, 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 aspects arise from this work. First, there is a need to provide a mathematical proof that the distance method we propose is better than the existing ones. 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 transaction times, but also the location of boarding on the territory, the route sequences, route types, etc if the data encoded as a binary vector.

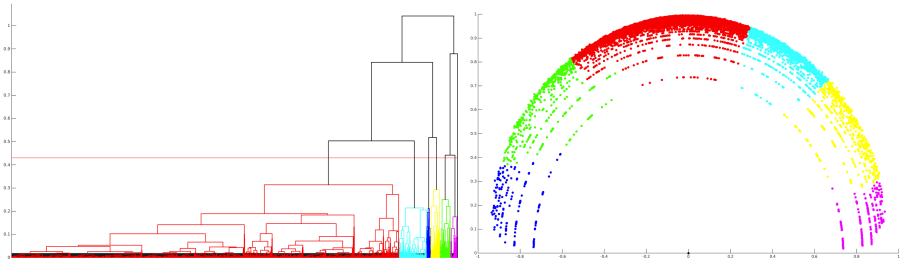


Fig. 6: Dendrogram of the hierarchical clustering with the associated seven clusters of the projected data

Acknowledgements The authors wish to acknowledge the support of the Société de transport de l'Outaouais, who provided the data for this study, and special thanks to Thalès and NSERC for supporting this project financially.

References

Borg I, Groenen P (2005) Modern Multidimensional Scaling: Theory and Applications. Springer

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

Ghaemi M, Agard B, Partovi Nia V, Trpanier M (2015) Challenges of spatial-temporal data analysis in the public transport domain. In: Proceedings of the 2015 IFAC Symposium on Information Control in Manufacturing, Ottawa, ON, Canada, INCOM'15

Hasan S, Schneider CM, Ukkusuri SV, Gonzlez MC (2012) Spatiotemporal patterns of urban human mobility. Statistical Physics 151(1-2):304–318

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York

Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Sanden SV, Lin D, Talloen W, Bijnens L, Göhlmann HW, Shkedy Z, Clevert DA (2010) FABIA: factor analysis for bicluster acquisition. Bioinformatics 26:1520–1527

Kieu LM, Bhaskar A, Chung E (2014) Transit passenger segmentation using travel regularity mined from smart card transactions data. In: Transportation Research Board 93rd Annual Meeting, Washington, D.C

Lathia N, Capra L (2011) How Smart is Your Smartcard: Measuring Travel Behaviours, Perceptions, and Incentives. In: Proceedings of the 13th International Conference on Ubiquitous Computing, ACM, New York, NY, USA, UbiComp '11, pp 291–300

Mahrsi ME, Cme E, Baro J, Oukhellou L (2014) Understanding passenger patterns in public transit through smart card and socioeconomic data. In: 3rd International Workshop on Urban Computing (SigKDD)

Montero P, Vilar JA (2014) Tsclust: An r package for time series clustering. Journal of Statistical Software 62(1):1–43, URL `http://www.jstatsoft.org/v62/i01`

Morency C, Trépanier M, Agard B (2006) Analysing the variability of transit users behaviour with smart card data

Morency C, Trpanier M, Pich D, Chapleau R (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

Ortega-Tong MA (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 MP, Trépanier M, Morency C (2011) Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19(4):557–568

Sneath PH (1957) The application of computers to taxonomy. Journal of General Microbiology 17(1):201–226, first algorithm of hierarchical clustering.

Yahya S, Noor NM (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 - Workshops, IEEE Computer Society, Washington, DC, USA, CSEWORKSHOPS '08, pp 31–36