M.-S. Ghaemi, B. Agard, M. Trépanier, V. Partovi Nia, Challenges of Spatial-Temporal Data Analysis in the Public Transport Domain, 2015 IFAC Symposium on Information Control in Manufacturing - INCOM, Ottawa (Ontario), Canada, May 11-13, 2015.

Challenges in Spatial-Temporal Data Analysis Targeting Public Transport ⋆

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

* *GERAD and CIRRELT Research Center, and Polytechnique Montrèal, Department of Mathematical and Industrial Engineering (e-mail: m.s.ghaemi@gmail.com).*
** *CIRRELT Research Center, and Polytechnique Montrèal, Department of Mathematical and Industrial Engineering (e-mail: bruno.agard@polymtl.ca).*
*** *GERAD Research Center, and Polytechnique Montrèal, Department of Mathematical and Industrial Engineering (e-mail: vahid.partovinia@polymtl.ca).*
**** *CIRRELT Research Center, and Polytechnique Montrèal, Department of Mathematical and Industrial Engineering (e-mail: mtrepanier@polymtl.ca).*

Abstract: Nowadays, tremendous data, are continuously gathering from the smart card in public transport domain. Such data, conveying two viable distinct information, can ensue designing intelligent transportation. More specifically, users behavior in a public transport system, can be investigated, as one of the data mining and machine learning applications. The first component of the data, provides the spatial feature, indicates the geographical coordinates of bus stops or subway stations. The second component of the data, deals with the temporal feature, being 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 from the data. Due to the massive data storage and the diversity of the data analysis methods, various challenges are arisen during the process of exploiting the hidden patterns of the data. We review a couple of scenarios and suggest a solution to overcome a number of the raised challenges. Moreover, the other aspects of this problem, are remaining as the open problems for the future research.

Keywords: Clustering, Public transport, Smart card, Spatial-Temporal data

1. INTRODUCTION

Everyday, thousands of people are using public transport systems. This means, huge amount of information is getting collected over a long period of time. Exploiting the hidden patterns of the data, enables infrastructure's development of the public transport system. This makes the usage of this network affordable, especially in the large metropolitan cities. In this regard, several researchers from different disciplines including urban computing, civil engineering, industrial engineering, data mining, etc., try to model this network.

Describing the behavioral pattern of users in the public transport network, is the major problem that can be revealed via the smart card's data. This data usually consists of two certain information about the users behavior. The first component of the data, includes the spatial attribute that provides the location of the trips. The second component of the data, comprises the temporal specification of the time-stamps, pertinent to the trips that users spend in the public transport system. Quite several different transit schemes are developed based on the variety of these two information. In most of the models, bus stops and subway stations are playing the central role regardless of the temporal features. The frequency of the used locations, is used to construct models specifying the users behavior. This knowledge can be helpful to provide special services in each station or bus stop. Nonetheless, they are incapable of clarifying the user similarity or the behavioral pattern, to discover the homogeneous groups of users who have the same manner.

Furthermore, for the sake of planning the future, it is necessary to anticipate users' future schedule ahead, based on the gathered data. In this regard, the importance of devising data analysis methods, considering the intrinsic attributes of data, i.e. spatial-temporal techniques, has emerged. Despite, spatial-temporal feature explains the significant ingredients of the users' trip, several hurdles exist in deploying all of these details to the data analysis methods. We divide the relevant data analysis approaches into two group, in order to facilitate demonstrating the

⋆ The authors appreciate the financial support of Thalès and Natural Sciences and Engineering Research Council of Canada (NSERC), and also STO for providing the data.

dilemmas. Consequently, we suggest the appropriate solutions for two cases according to the few assumptions on the closeness measure of users pattern.

Basically, the design and the development of a model imitating public transport users relies on some postulations. Accordingly, finding a measure to evaluate behavioral patterns from the history of users habit 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 the 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: first, self-report of users' feedback, and second, their real behavior versus the change of users behavior when they are encouraged by the various incentives. Finally, the authors concluded that smart card data is as valid as the human activity, extracted from the cellular phone data in order to design the future infrastructure and the travellers guidance in (Lathia and Capra, 2011). Therefore, the human mobility can be modelled according to the smart card data, producing a big source for modeling the human behavior.

Smart card data, contains worthwhile digital information of daily locations visited at the certain period of a large number of individuals. The other sources of digital information exist such as the cellular phone, credit card transactions, social network, and GPS tracker vehicle, e.g. on a bike, on a car and on a motorcycle. The best promising source of users digital information is the smart card data. Thus this helpful information can be utilized to model the urban mobility pattern (Hasan et al., 2012). The other useful information such as the travel time and the number of passengers for the sake of congestion analysis and planning improvement, can be extracted as well (Fuse et al., 2010).

Predicting the users' location according to the popular locations considering the users interaction in the city, is modelled as a spatial-temporal pattern of the human mobility in (Hasan et al., 2012). Researchers exploit the interpretable patterns, using a data mining clustering approach to understand the passenger's temporal behavior (El Mahrsi et al., 2014). Clustering approach, can help the transport operators to meet the customers' demands. The real dataset from the metropolitan area of Rennes (France) with four weeks of smart card data containing the trips of both bus and subway is tested in this approach. Furthermore, the cluster of similar temporal passengers are extracted from their boarding time, according to the generative model-based clustering approach. Then after, the effect of the distribution of socioeconomic characteristics on the passenger temporal clusters are investigated in this study.

As another example, the extensive database of the Oyster Card transactions, obtained from London's public transport users, is utilized in (Ortega-Tong, 2013). This database is deployed to classify the 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, where selecting the groups of homogeneous people is the case of interest. Four groups of users including, regular users 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 affairs, are investigated in this work.

Smart card data gathered from Brisbane Australia (Kieu et al., 2014) is another study for strategic transit planning according to the individual travel patterns. Origins and destinations that a cardholder usually travels is defined as the travel regularity. Thus, mining the travel regularity of the frequent users can be inferred to extract the travel pattern and its purposes. Reconstruction of the 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).

2. METHODOLOGY

A typical public transport network, containing subway stations, bus stops and users, is shown in Fig. 1. This network, usually consists of connected bus and subway lines at few strategic locations of the city. In the modern public transport system, instead of the old-fashion tickets, most of the people prefer to use smart cards with persuasive promotion plans and even half-price discounts for the young or old individuals. Smart card data, usually consists of two types of information; spatial and temporal. The spatial data includes coordinates of the bus stop or stations, e.g. the latitude and the longitude that can be the GPS data or the relative location values. The temporal data includes the starting time of each trip. We encode this information as a $0 - 1$ vector, where the start of the trip is identified by 1. According to these information, analysing the pattern of public transport usage based on the smart card data can be divided into three categories, 1) spatial pattern, 2) temporal pattern and 3) spatial-temporal pattern.

2.1 Spatial data

The spatial data contains worthwhile information about the geographical details of each bus stop and are stored sequentially following the order of the temporal usage. Although, enough information about the coordinates of the bus stops are available, defining a measure of similarity of behaviors in the public transport network is troublesome. The main issues about the similar trips in the spatial case can be summarized into the following two issues. Issue 1, two users are similar according to the similar bus stops they usually take every day. Issue 2, two users are categorized in the homogenous group of users, if their resultant traversed distance resembles. Moreover, it is possible to consider the following scenarios to realize how this spatial criterion is difficult to define.

Fig. 2, shows three users, red, blue, and green, who use the public transport from the same starting point and leaving the system at the same point as well, however, they use

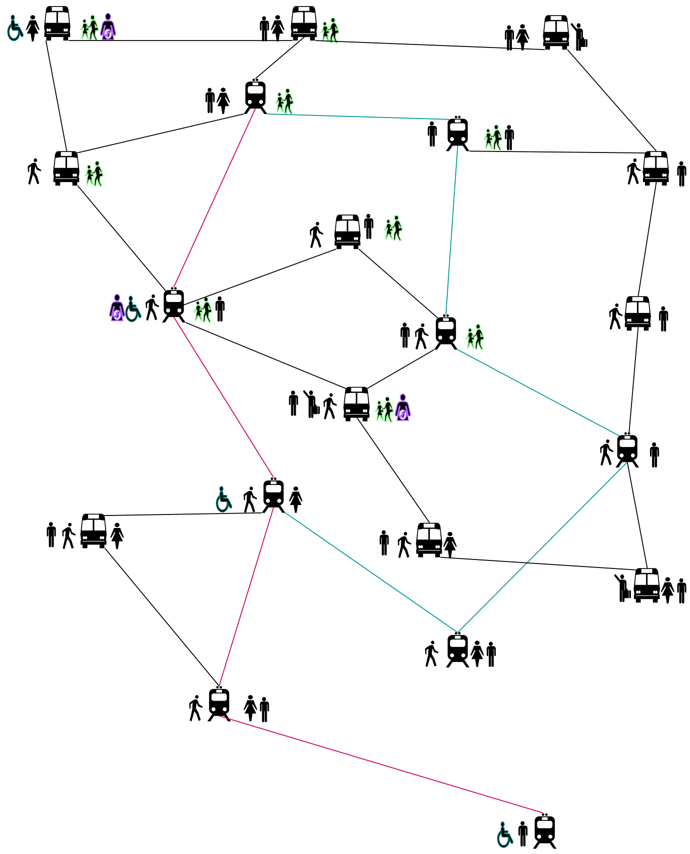

Fig. 1. A typical network of public transport system including users, buses and two lines of subway.

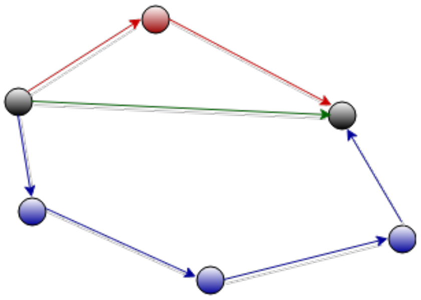

Fig. 2. Three users with the same starting point and ending point.

different number of trips in various directions. Hence, their resultant traversed distance is exactly identical while each user has a different path. This example shows how the issues can change the measure of similarity between two users in the spatial data analysis.

Fig. 3, demonstrates two users with the same pattern, but in the opposite directions. On the contrary to the Fig. 2, regardless of the resultant trips, one can define the similarity only according to the bus stops. This may reflect the trip patterns of the same user who travels between home to work and vice versa at the different time stamps.

In Fig. 4, it turns out that it is possible to raise a third issue. Despite, the starting points and the ending points are distinct for both users, and none of them use the same bus stops, still one directional routing pattern is emerged. Even they can set one of the patterns, at the different time stamps. This instance, may be happening in the spatial-temporal data analysis.

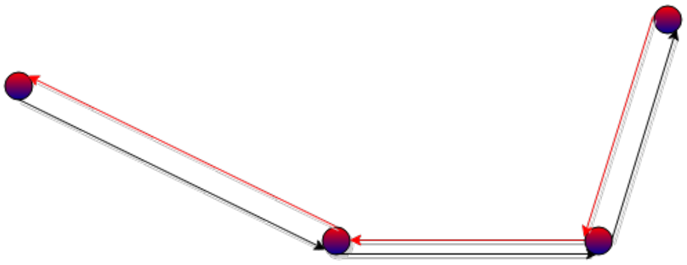

Fig. 3. Two users taking the same buses in the opposite directions.

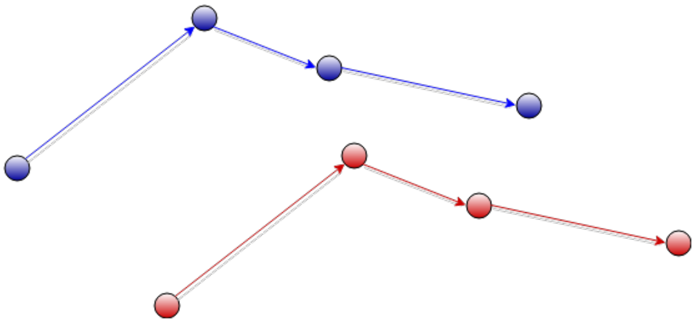

Fig. 4. Two users with the same directional pattern.

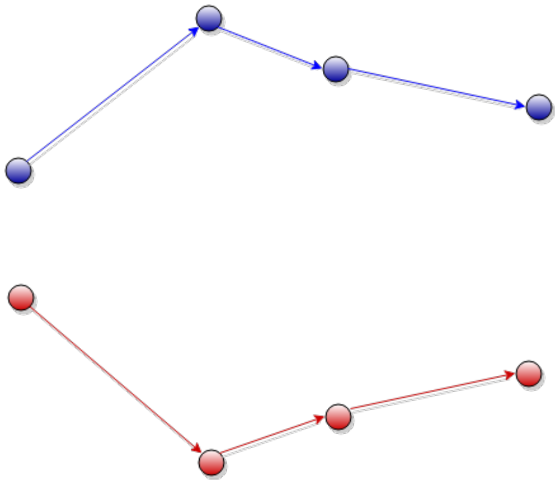

Fig. 5. Two users with the same symmetric directional pattern.

With the same argument described for Fig. 4, Fig. 5 demonstrates the fact that, this directional routing pattern, may happen in a symmetric manner as well. This symmetrical property, is holding in the horizontal orientation in Fig. 5, though the vertical orientation, $x = y$, and the $x = -y$ orientation are also presumable.

Considering a case where two users are following almost the same sequence of bus stops order except for one in their second transit, Fig. 6 shows this situation. The behavior in Fig. 6 can also belong to the schedule of one user in two different days. This anomaly would probably occur often too when the frequent bus stops are used by the similar users. Defining this type of usage pattern as an outlier or noise, affects the user similarity criterion.

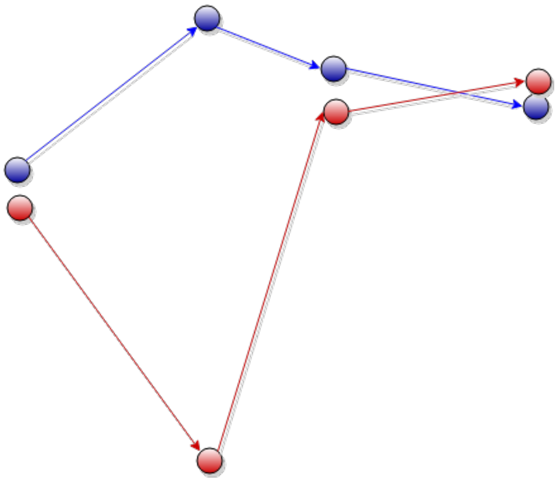

Fig. 6. The two users with the same pattern of usage except for their second transit.

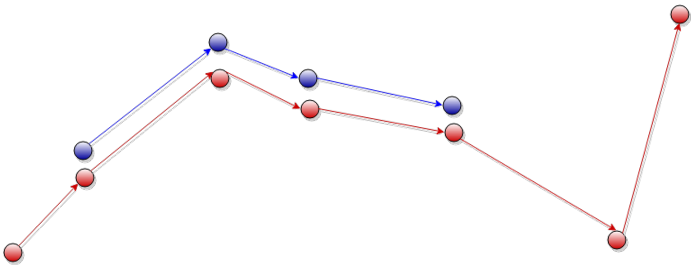

Fig. 7. The two partially similar usage pattern, but with the different number of the bus stops.

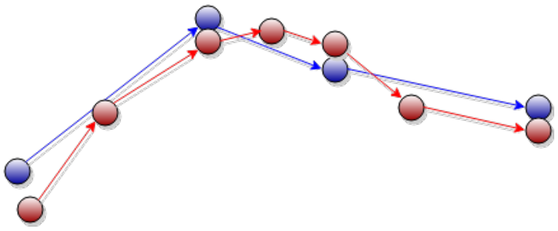

Fig. 8. The same resultant traversed distance with the different bus stops.

In Fig. 7, two users are shown, the total trip and the bus stops taken by the user blue, is a subset of the used bus stops by the user red. In this example, two users are utilizing the public transport roughly alike in the particular part of their schedule, nevertheless they behave differently beyond that interval. Hence, it turns out, the number of the taken bus stations is another important factor in defining the user similarity in the spatial domain.

Fig. 8, shows another scenario, where the two users differ in the number of trips. Like Fig. 7, the blue one's used bus stops, is a subset of the taken bus stops by the red user, meantime, the resultant traversed distance is almost the same for both users. This depicted sequence of bus stop usage trajectories, associate to the closely similar pair of users, though the number of the taken bus stops are totally different.

Suppose two users who take the same bus stops not necessarily in the same order, during their daily trip. In other words, permutations of the same bus stops can amount to the totally different resultant traversed distance. As it is shown in Fig. 9, the same bus stops are still shared between the two users without the same usage pattern. This often gets more complicated when temporal information is also got involved in this sort of data analysis dilemma.

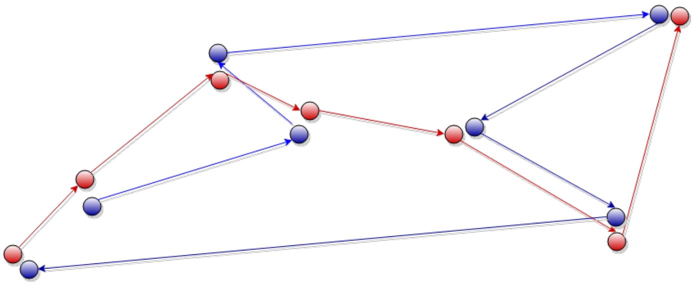

Fig. 9. The two users taking the same buses, but with the different order.

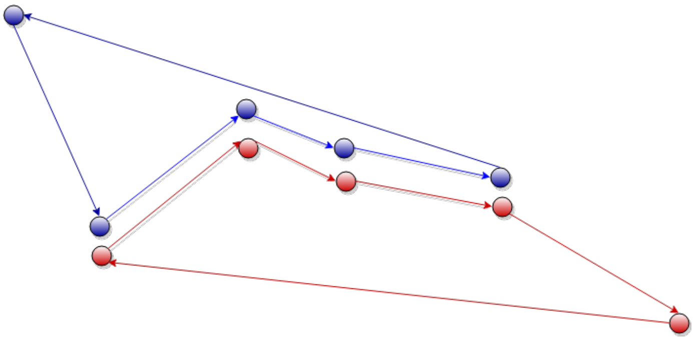

Fig. 10. The same pattern of the two users living in the different locations.

In another scenario, two users might use the public transport exactly in the same order, except the starting point and the end point. This is an ordinary pattern that appears by the users who are living in the different parts of the city, while, they take the same bus stops during their daily trip. For instance, Fig. 10, shows the two users following the same pattern in the downtown area, while living away from each other.

In all of the discussed cases (Fig. 2 to Fig. 10), Euclidean distance between bus stops, can be assumed in the definition of the user similarity. This presumption can be violated, if taking the bus stops is not a uniform distribution. Despite, the utilization of the bus stops usually comes from a mixture of normal distributions, for the sake of simplicity, we assume that bus stops are sampled from a single normal distribution. Fig. 11, illustrates a typical public transport network, where the center of the city is the mean of the spherical normal distribution, and the off-diagonal entries of the covariance matrix are zeros, because of the spherical symmetry of the density function.

This hypothesis, implies if two bus stops are taken from the same circle with the particular radius, it can be assumed that they are relatively close to each other. Accordingly, in Fig. 11, the red user follows the same pattern as the user blue do, i.e. at each time point, the identical bus stops are taken from the same orbit.

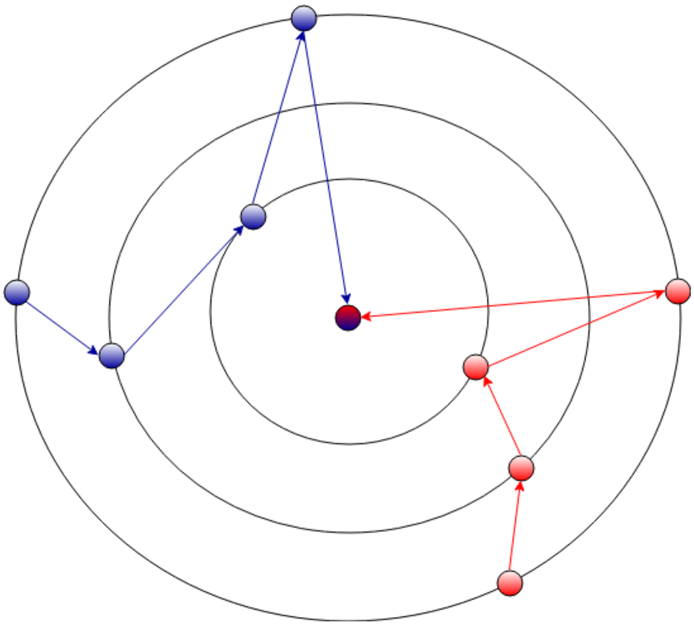

Fig. 11. User similarity based on the circular grid representation of the bus stops.

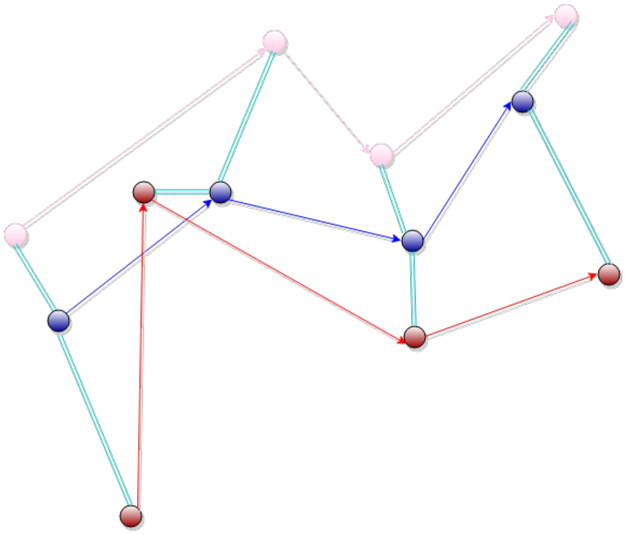

Fig. 12. Pairwise bus stop difference criterion for the measure of user similarity.

In the real world datasets, milions of users usually take the public transport for their daily journeys. Hence, the combination of the aforementioned patterns can happen in the whole picture. Moreover, taking the temporal behavior into account, certainly affects the complexity of the scheduling and the methods of data analysis.

Issue 1 and issue 2 raised in the beginning of section 2, address how the user similarity criterion can be defined under few assumptions. The first issue can be solved if they take the same number of the bus stops in their daily trip. The second issue can be solved if in the sequence of the used bus stops, each pair of the bus stops associated to the same time step, are close to each other.

Finally, by summing all distances between a pair of bus stops from a reference user, the similarity of a user can be computed. One suggestion for the reference user, is the mean geographical coordinates of the used bus stops, at each time point. These few hypotheses preserve the defined constraints such that, the resultant traversed distance of two users is similar if they take the similar bus stops at each time step. Fig. 12, shows three users, where the users red and orange are compared to the blue user. The sum of differences between all pairs of bus stop between the blue and the red circles (green lines) identifies the similarity of the user blue and the user red. Analogously, the similarity of the user blue and the user orange can be computed.

Formalizing these mathematically, suppose these two sequences are given as the S_1 and the S_2 from the same length. Each entry of the sequence, consists of (x, y) geographical coordinates of the bus stop. Hence, we define the similarity of two sequence, as the summation of Euclidean distances of the point-wise elements. Then we have,

$$\text{Similarity}(S_1, S_2) = \sum_{i=1}^{n} \text{distance}(S_{1i}, S_{2i}) \tag{1}$$

In addition, *Cosine* similarity and *Pearson* similarity are the other measurements suggested in (Li et al., 2008) as follows,

$$\text{Cosine}(S_1, S_2) = \frac{\sum_i S_{1i} S_{2i}}{\sqrt{\sum_i S_{1i}^2}\sqrt{\sum_i S_{2i}^2}} \tag{2}$$

$$\text{Pearson}(S_1, S_2) = \frac{\sum_i (S_{1i} - \overline{S_1})(S_{2i} - \overline{S_2})}{\sqrt{\sum_i (S_{1i} - \overline{S_1})^2}\sqrt{\sum_i (S_{2i} - \overline{S_2})^2}} \tag{3}$$

2.2 Temporal data

In (Agard et al., 2013), an inovative technique is introduced for grouping and characterising public transport users from the temporal data. A new distance calculation technique is proposed by the authors to apply the k-means clustering method.

Table 1. Sequence of the 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}	0	0	1	1	0	0	0
X_{12}	0	0	0	0	1	1	0
X_{13}	0	0	0	0	0	1	1

Suppose a $0-1$ vector of the temporal data is given in the input as it is shown in Table 1, to better capture the similarities between public transport users' journeys, the indices of the 1 values can be utilized as follow,

$$\text{Score}(\text{User}_X) = \sum_{i=1}^{n} i \times X(H_i)$$

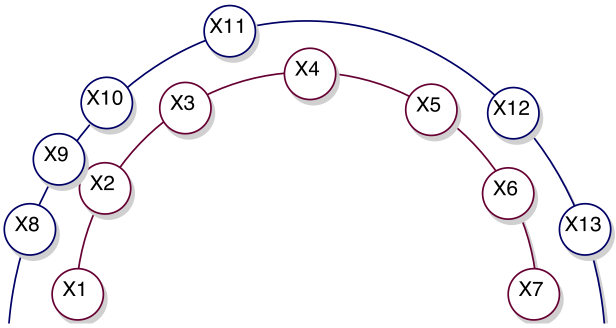

Fig. 13. Mapping the temporal data into a half-circle.

This formulation encourages the similar scores for the users who take the public transport alike while keep the users as far as possible if they use it at the different times.

As it is shown in Fig. 13, the users $X_1, \ldots, X_7$ are mapped into the firs half-circle. In this mapping, user X_1 is as far as possible from the user X_7 and as close as possible to the user X_2 on the first arc, that exactly conforms the associated Euclidean distance in terms of the time difference. Consequently, the users who take the public transport two times a day, are located on the second arc with a larger radius. Similar to the first arc, users with the close Euclidean distance, are located closely. Further properties are also held in this representation, e.g. user X_8 that takes the public transport at time H_1 and H_2 is also close to the users X_1 and X_2. This representation preserves the given Euclidean relations, also provides a better visual guide corresponding to the time schedule with meaningful interpretation for the experts.

The suggested method in (Agard et al., 2013), considers representation of the observations in the 24-hour binary vectors to model the temporal data. It provides a clock-like visual diagram with meaningful interpretation for the experts. However, the continuous time-stamps should be discretized as preprocessing step before applying this algorithm. Similar to the spatial case, this is only one scenario to define the similarity of users, and the other similarity instances require specific solutions to adapt the underlying assumptions.

3. CONCLUSION

We reviewed a number of different use cases that can be possibly existed in analysing the smart card data. Each behavioral user pattern requires a specific metric to reveal the similarity of the users according to the appropriate criterions. The expert individual is the one who is authorized to select one of these metrics or a combination of few of them as the measure of fit to the data.

In addition, mixing the spatial and the temporal data together, creates even more complicated cases that we did not consider in this study. In the spatial data analysis, we suggest a method for capturing the user similarity. This solution proposed to meet two important standards. First, at each step, the similar bus stops should be taken by the similar users, we call it local property. Second, two users are following the similar pattern in their daily trip, if the overall resultant traversed distances are close, we call it global signature.

Moreover, we emphasize that the proposed solutions do not cover all the possible similarity metrics. Thus, we conclude, finding an algorithm which generalizes the similarity metric to all sort of the user proximity is difficult to build. However, it can be designed under certain assumptions with associated interpretations to fulfill a set of constraints that is desirable for a given dataset. Investigation of the important use cases in the datasets, and finding the related metric, is the future direction of this research.

REFERENCES

Agard, B., Partovi Nia, V., and Trépanier, M. (2013). Assessing public transport travel behaviour from smart card data with advanced data mining technique. In *World Conference on Transport Research*, 13 WCTR, 15–18. Rio de Janeiro, Brazil.

El Mahrsi, M.K., Etienne, C., Johanna, B., and Oukhellou, L. (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data: A case study in rennes, france. In *ACM SIGKDD Workshop on Urban Computing*, 9p. New York City, USA.

Fuse, T., Makimura, K., and 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*, volume 2010. Orlando, Florida, USA.

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

Kieu, L.M., Bhaskar, A., and Chung, E. (2014). Transit passenger segmentation using travel regularity mined from smart card transactions data. In *Transportation Research Board 93rd Annual Meeting*, 14-4892. Washington, D.C, USA.

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

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., and Ma, W.Y. (2008). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, 34:1–34:10. ACM, Irvine, California, USA.

Morency, C., Trepanier, M., and Agard, B. (2006). Analysing the variability of transit users behaviour with smart card data. In *The 9th International IEEE Conference on Intelligent Transportation Systems*, ITSC 2006, 44–49. Toronto, Canada.

Ortega-Tong, M.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.