Les Cahiers du GERAD ISSN: 0711–2440

Challenges in spatial-temporal data analysis targeting public transport

M.S. Ghaemi, B. Agard,
V. Partovi Nia, M. Trépanier

G–2015–13

February 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Challenges in spatial-temporal data analysis targeting public transport

Mohammad Sajjad Ghaemi[a]

Bruno Agard[b]

Vahid Partovi Nia[a]

Martin Trépanier[b]

[a] *GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

[b] *CIRRELT & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

m.s.ghaemi@gmail.com
bruno.agard@polymtl.ca
vahid.partovinia@polymtl.ca
mtrepanier@polymtl.ca

February 2015

Les Cahiers du GERAD

G–2015–13

Abstract: Intelligent transportation has been emerged as one of the data mining and machine learning applications. The smart card data nowadays are continuously gathering in the public transport systems. Such data, usually convey two viable distinct information to investigate how users prefer to behave in a public transport system. The first component of the data, provides the spatial feature, which indicates the geographical coordinates of the bus stops or subway stations. The second component of the data, deals with the temporal feature that is the time of the trips that the public transport has been 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 out of the data. Due to massive data storage and diversity of the data analysis methods, various challenges are arisen in the process of exploring and exploiting the hidden patterns of the data. We review a couple of scenarios and suggest a solution to overcome the raised challenges.

Key Words: Clustering, public transport, smart card, spatial-temporal data.

Acknowledgments: My special thanks go to Thalés and NSERC whose financial supports make this project possible, and also STO for providing the data.

1 Introduction

Everyday, thousands of people are utilizing public transport systems. This means, huge amount of information is getting collected over a long time. Exploiting the hidden patterns of the stored data enables infrastructure development of the public transport system. This makes the usage of this network affordable, especially in large metropolitan cities. In this regard, several diverse researchers from different disciplines including urban computing, civil engineering, industrial engineering, data mining, etc. try to model this network. Describing behavioural patterns of users in the public transport network is the major problem that can be revealed via the smart cards data. However, most of these cases are developed based on certain postulations. Accordingly, finding a measure to evaluate behavioural 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' behaviour 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 behaviour versus change of users behaviour 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 travellers in (Lathia and Capra, 2011). Therefore, human mobility can be modelled according to the smart card data as one of the big data sources from human activity.

Smart card data, contain worthwhile digital information of daily locations visited at certain period of a large number of individuals. Other sources of digital 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 exist. The best promising source of users digital information is the smart card data. Thus this helpful information can be utilized to 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, can be 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 modelled as a spatial-temporal pattern of human mobility in (Hasan et al., 2012). A data mining approach then is used to understand passenger's temporal behaviour 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 to maintain the services and tools and 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 travellers, and finally visitor travellers for tourism and business affairs, are investigated in this work.

Smart card data gathered from Brisbane, Australia is another source of information being 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 can 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).

2 Methodology

A typical public transport network, containing subway stations, bus stops and users, is shown in Figure 1. This network, usually consists of connected bus and subway lines at few strategic locations of the city. In the modern public transport systems, 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. Spatial data includes coordinates of the bus stop or stations, e.g. latitude and longitude that can be GPS data or relative location values. Temporal data includes the time of each trip. This information is encoded in a $0-1$ vector, where 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 patterns, 2) temporal patterns and 3) spatial-temporal patterns.

2.1 Spatial data

Spatial data contains worthwhile information about the geographical details of each bus stop and are stored sequentially following the order of temporal usage. Although, enough information about the coordinates of bus stops are available, defining a measure of similarity of behaviours 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 questions. First, are two users similar according to the similar bus stops they usually use every day? Second, are they 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.

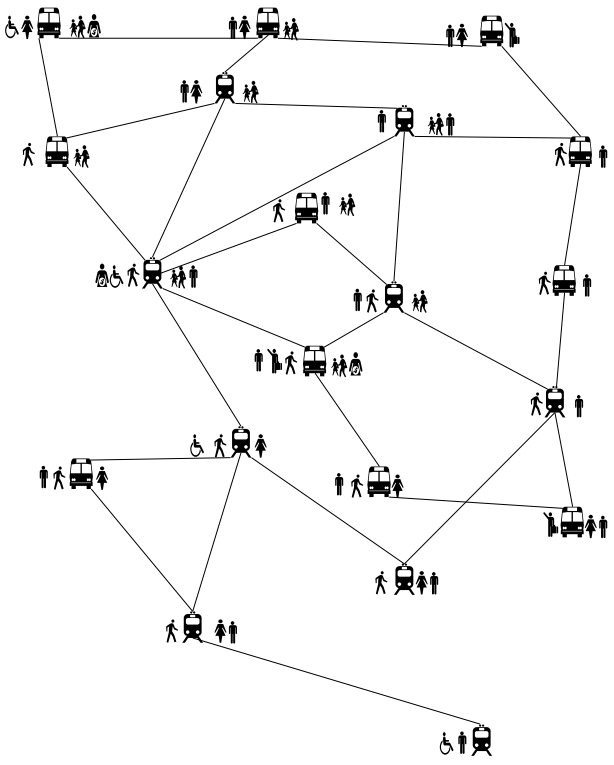

Figure 1: A typical network of public transport

Figure 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 different number of trips in various directions. Hence, their resultant traversed distance is quit identical while each uses a different path. This example shows that how the answer to the two asked question can quit change the measure of user similarity in the spatial data analysis task.

Figure 3, points out two red and blue users start and end their trips using the same bus stops but in the opposite directions. In contrary to the Figure 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 in different time period.

In Figure 4, it turns out that it is possible to ask even the third question. 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. Besides, that it can be a consequence of taking different buses from variant inceptions to the terminations, taking the same bus stops in the same route but in different time intervals would be the other reason. The former instance, is happening in the spatial-temporal data analysis.

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

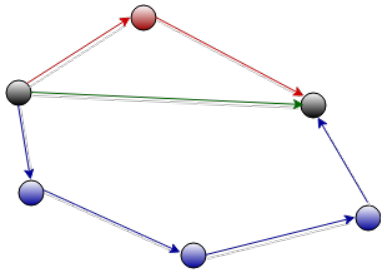

Figure 2: Three users with the same start point and end point

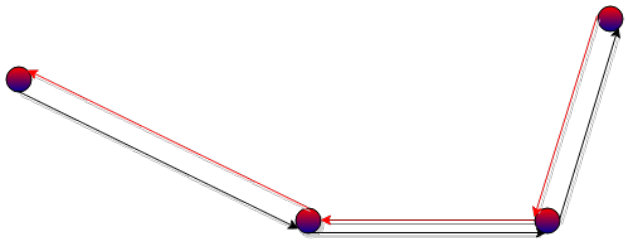

Figure 3: Two users taking the same buses in opposits directions

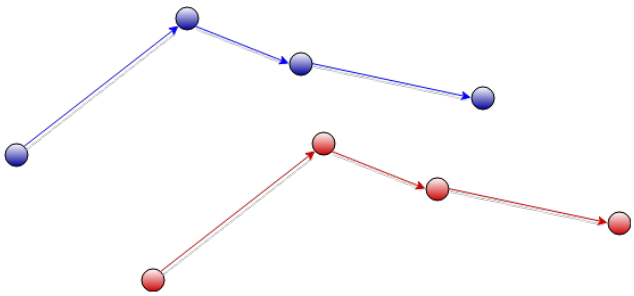

Figure 4: Two users with the same directional pattern

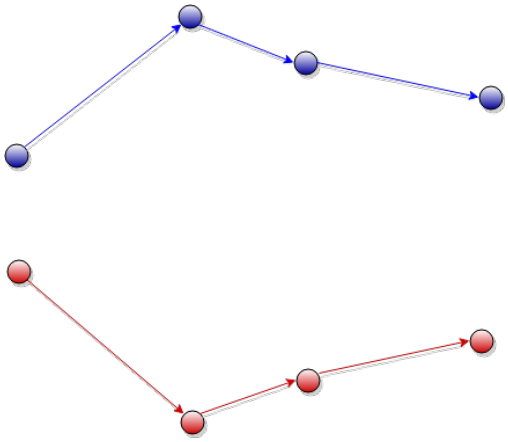

Figure 5: Two users with the same symmetric directional pattern

Considering a case where two users are following almost the same sequence of bus stops order except one. Figure 6 shows this situation, this behavior can also belong to the schedule of one user in two different days. This anomaly would probably occur often too when frequent bus stops are used by similar users. Defining this type of usage pattern as an outlier or might be a noise, because of fault in storing or capturing devices, quite depends on the definition of user similarity criterion.

In Figure 7, two users are shown, the total trip and bus stops taken by user blue, is a subset of the used bus stops by user red. In this circumstance, two users are utilizing the public transport roughly alike in a particular part of their schedule, nevertheless they behave differently beyond that interval. Hence, it turns

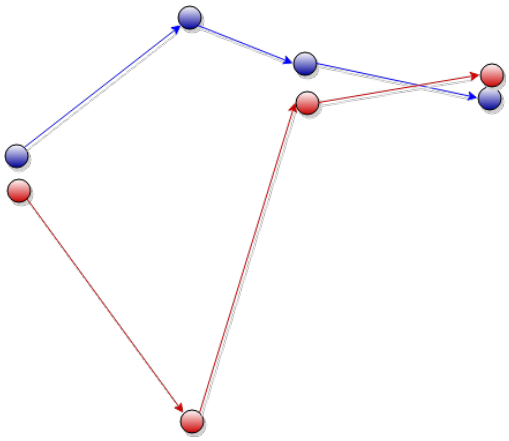

Figure 6: Two users with the same pattern of usage except one

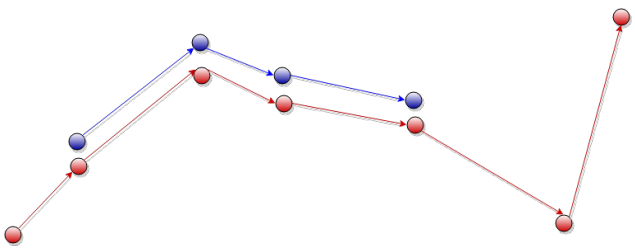

Figure 7: Two users with partial similarity pattern

out, the number of the taken bus stations is another important factor in defining the user similarity in the spatial domain.

Figure 8, shows the other scenario, where the two users differ in the number of trips. Like Figure 7, the blue one's used bus stops, is a subset of 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 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 Figure 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.

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

In the former circumstances, Euclidean distance between bus stops, was assumed in the definition of the user similarity. This presumption can be violated, if the utilization of the bus stops does not conform the uniform distribution. Despite, the utilization of the bus stops usually comes from a mixture of normal distributions, for the sake of simplicity, we can assume that bus stops are sampled from just a normal

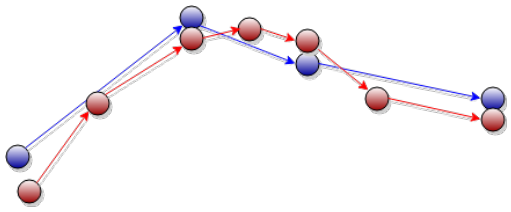

Figure 8: The same resultant traversed distance with different bus stops

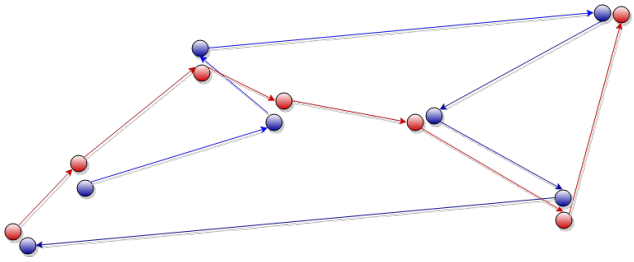

Figure 9: Two users taking the same buses with different order

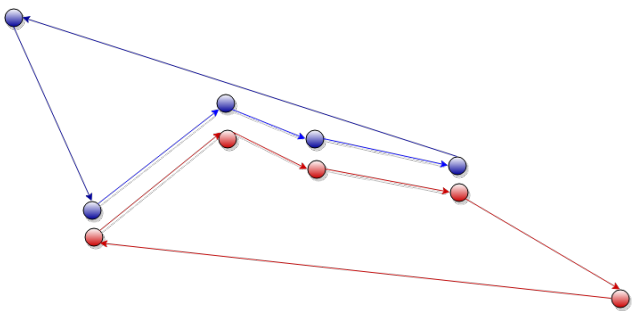

Figure 10: The same pattern of two users living in the different places

distribution. Figure 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 they are relatively close to each other, in contrary to the Euclidean distance. Accordingly, in Figure 11, the red user is following the same pattern as the user blue do (at each time point, the identical bus stops are taken from the same orbit).

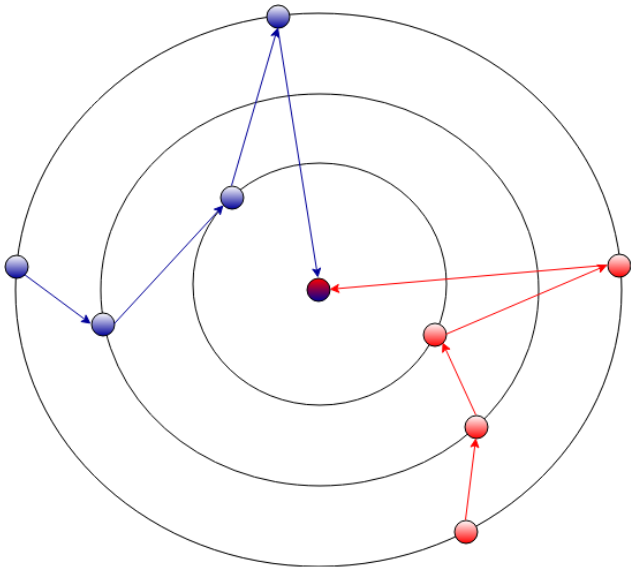

Figure 11: User similarity based on circular grid representation of bus stops

So far, a number of possible use cases are introduced about the spatial public transport usages, comparing two given users. In the real world datasets, where milions of users usually take the public transport for their daily journeys, the combination of these patterns can happen in the whole picture. Moreover, taking the temporal behaviour into account, certainly affects the complexity of the scheduling and methods of data analysis.

Two aforementioned questions, address how the user similarity criterion can be defined under few assumptions. As the first one, we assume two users are comparable if they take the same number of bus stops in their daily trip. For the second assumption, two users are similar 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 the all distances between pair of bus stops from an origin user, similarity of a user can be computed. One suggestion for the origin user, is the mean geographical coordinates of the used bus stops, at each time point. These few hypotheses preserve the defined constraints such that, resultant traversed distance of two users is similar if they take similar bus stops at each time step. Figure 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 blue and red circles (green lines) identifies the similarity of user blue and red. Similarly, the similarity of users blue and orange can be computed.

Formalizing this definition mathematically, suppose these two sequences are given as S_1 and 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,

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

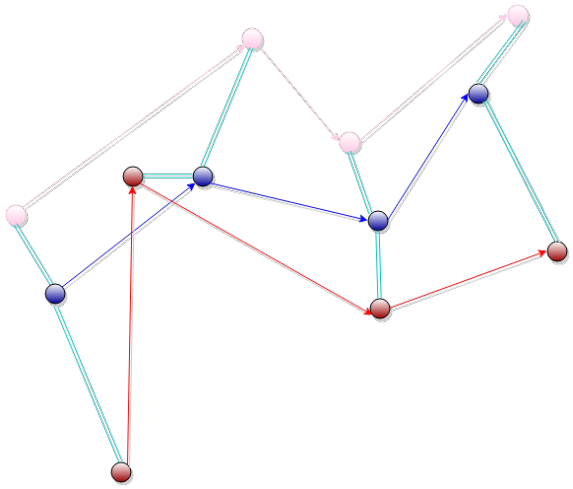

Figure 12: Pairwise bus stop difference criterion for measure of user similarity

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

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

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

2.2 Temporal data

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

Table 1: Sequence 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}	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 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,

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

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 different times.

As it is shown in Figure 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 distances in terms of time differences. Consequently, the users who take the public transport two times a day, are located on the second arc. Similar to the first arc, users with close Euclidean distances, are located proximate. 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 that preserves the given Euclidean relations, also provides a visual guide corresponding to the time schedule that makes a better interpretation for the experts.

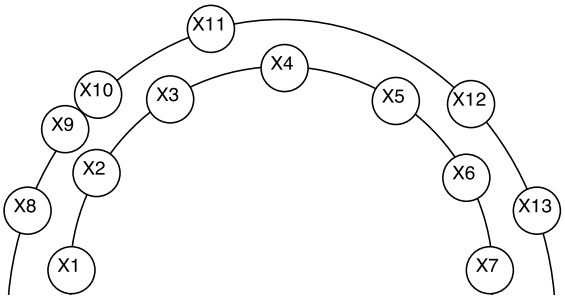

Figure 13: Mapping the temporal data into a half-circle

3 Conclusion

In this article, we reviewed a number of different use cases that can be possibly existed in analysing the smart card data. Each behavioural 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 combination of few of them as the measure to fit the data. In addition, mixing spatial and temporal data together creates even more complicated cases that we do not go through them 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, 1) At each step, similar bus stops should be taken by similar users, this is called local property. 2) If the overall resultant traversed distance are relatively nearby, this implies two users are following the similar pattern in their daily trip. This captures the global characteristic of the usage schedule in the public transport network. Moreover, it has to be stated that the proposed solutions do not cover all the possible similarity metrics. In the spatial case, several cases require another solutions. Thus, we come to the conclusion, that finding an algorithm which generalizes the similarity metric to all sort of user proximity does not exist. However, it can be designed under certain assumptions with associated interpretations to fulfill a set of constraints that are desirable for a given dataset.

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.

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.

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. Washington, D.C.

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, New York, NY, USA.

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, New York, NY, USA.

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

Morency, C., Trépanier, M., and Agard, B. (2006). Analysing the variability of transit users behaviour with smart card data. In Intelligent Transportation Systems Conference, ITSC '06, 44–49.

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.