Hindawi
Journal of Advanced Transportation
Volume 2017, Article ID 1318945, 14 pages
https://doi.org/10.1155/2017/1318945

WILEY | Hindawi

Research Article

Spatial-Temporal Similarity Correlation between Public Transit Passengers Using Smart Card Data

Hamed Faroqi, Mahmoud Mesbah, and Jiwon Kim

School of Civil Engineering, The University of Queensland, Brisbane, QLD, Australia

Correspondence should be addressed to Hamed Faroqi; h.faroqi@uq.edu.au

Received 30 April 2017; Revised 29 June 2017; Accepted 16 July 2017; Published 14 September 2017

Academic Editor: Zhi-Chun Li

The increasing availability of public transit smart card data has enabled several studies to focus on identifying passengers with similar spatial and/or temporal trip characteristics. However, this paper goes one step further by investigating the relationship between passengers' spatial and temporal characteristics. For the first time, this paper investigates the correlation of the spatial similarity with the temporal similarity between public transit passengers by developing spatial similarity and temporal similarity measures for the public transit network with a novel passenger-based perspective. The perspective considers the passengers as agents who can make multiple trips in the network. The spatial similarity measure takes into account direction as well as the distance between the trips of the passengers. The temporal similarity measure considers both the boarding and alighting time in a continuous linear space. The spatial-temporal similarity correlation between passengers is analysed using histograms, Pearson correlation coefficients, and hexagonal binning. Also, relations between the spatial and temporal similarity values with the trip time and length are examined. The proposed methodology is implemented for four-day smart card data including 80,000 passengers in Brisbane, Australia. The results show a nonlinear spatial-temporal similarity correlation among the passengers.

1. Introduction

Analysing passengers' movements in a public transit network is important in understanding passengers' travel behaviours and designing more customised public transit services. By identifying passengers with similar spatial and/or temporal trip patterns and understanding their characteristics, transit operators could design transit services that better meet different needs of different passenger groups and develop strategies to influence travellers to use the existing transit network more efficiently. For instance, if a group of passengers every day boards on a specific bus route at a specific stop and time, and they change the bus at another stop to arrive at their final destination, then a specific bus (or a minibus) can be allocated to that group of passengers between their first boarding and last alighting stop for particular periods.

With the availability of transit smart card data that provide information on boarding and alighting locations and times for each passenger trip, it is now possible to analyse spatial and temporal movement patterns for each passenger and compare them across passengers, thereby allowing a deeper understanding of individual passengers and their relationships. And each spatial and temporal dimension of the movement has its measures and units, which makes it difficult to study these dimensions simultaneously [1]. Transit authorities have developed automated fare collection (AFC) systems around the world since two decades ago. These systems not only aim to gather fares but also they turn valuable datasets out of trips as a by-product. The datasets include time and place of transactions for boarding on and/or alighting from the public transit system [2, 3]. The datasets help researchers to expand the studies and investigate nested interactions among public transit passengers [4, 5]. Previously, datasets for transport studies were gathered mostly from surveys, which were expensive to run and limited in size. Hence, smart card datasets provide opportunities to explore travel behaviours of public transit passengers in large and detailed scales.

Exploring similarities among passengers' trips, where trip similarity can be defined regarding spatial and/or temporal dimensions, can discover relationships among the passengers. By identifying how similar two passengers' trips are spatial and temporal, the "passenger similarity" can be

defined as a composite measure of trip similarity between two passengers. Such a passenger-level similarity measure can help the analysis of passenger characteristics. These measures can help the design and development of various customer-centric transit services and mobility applications. Examples of such applications include demand responsive transport (DRT) systems [6], friend recommendation systems [7, 8], and traffic flow prediction models [9].

The paper, for the first time (to the best of our knowledge), investigates the correlation of spatial similarity with the temporal similarity between public transit passengers. It measures the spatial similarity and temporal similarity of public transit passengers with a passenger-based perspective, in which one or more trips model each passenger. The spatial and temporal similarity measures are developed for the public transit network. The spatial similarity measure considers direction as well as the distance between the trips of the passengers. The temporal similarity measure considers both boarding and alighting time in a continuous linear space. The spatial-temporal similarity correlation is analysed using histograms, Pearson correlation coefficients, and hexagonal binning. In addition, the relation between the spatial and temporal similarity values with the travel time and length is examined. The proposed methodology is implemented for four-day smart card data including about 180,000 trip legs for 80,000 passengers in Brisbane, Australia.

In brief, the scientific contribution of the paper is three-fold. First, a passenger-based perspective that characterizes passengers by both boarding and alighting transactions' time and location is developed to study the travel behaviour in the public transit system. Second, specific metrics based on smart card data are developed for measuring the spatial and temporal similarities between passengers in the public transit network. Third, the correlation between the spatial and temporal similarity values is investigated.

The rest of the paper is structured as follows. The Literature Review discusses recent studies in these fields. Methodology describes the proposed methods for measuring the spatial and temporal similarities between the passengers. Case study and analysing the correlation are explained in the Results. Finally, Conclusion includes discussion, potential applications, and plans.

2. Literature Review

Studying both spatial and temporal aspects of public transit passengers' trips has been focused on just recently. Nishiuchi et al. (2013) explored spatial, daily, and hourly variations in the travel characteristics defining regularity indices for both spatial and temporal aspects. They investigated the variations on a smart card dataset extracted from 31749 passengers' trips during one month. They found out that there is no significance difference between the numbers of hourly trips during weekdays [10]. Ma et al. (2013) discovered the spatial and temporal patterns determining the regularity of transit passengers' travel patterns. They used four different clustering algorithms (K++, c4.5, KNN, 3-hidden layers NN) and compared the performance of the algorithms [11]. Tao et al. (2014) compared spatial-temporal patterns of Bus Rapid Transit (BRT) passengers with non-BRT passengers using flow-comap technique. They used Brisbane transit network as the case study and found that BRT trips involved a larger number of longer distance trips [12]. Tao et al. (2014) focused on the passengers instead of stops for discovering the patterns among the passengers. They used a smart card dataset from Brisbane and divided it into 5-time windows for further analysis of spatial patterns of the passengers' trips [13].

Kieu et al. (2015) used a modified DBSCAN algorithm to discover spatial travel patterns at stop levels from a smart card dataset [14]. Ghaemi et al. (2015) described difficulties in revealing the spatiotemporal pattern analysis in the public transit system. They proposed a method for measuring the spatial similarity between two users in the public transit [15]. Manley at al. (2016) studied variations of the regularity between the passengers' trips across the network considering different transport modes. They used DBSCAN clustering algorithm and defined three definitions for the regularity in rail and bus networks [16]. Ma et al. (2017) explored commuting patterns in the public transit network using data from the 1-month smart card at Beijing. They investigated spatial and temporal patterns at stop levels for commuters and noncommuters [17]. El Mahrsi et al. (2016) presented two different approaches for clustering smart card dataset. First one is stop-oriented, which clusters stops based on the frequency of the boarding and alighting transactions; and the second one is passenger-oriented, which clusters passengers based on the boarding times of their trips [18]. Yu and He (2017) proposed a visualisation model to present travel demand at stop level using heat maps. Also, they used a Gaussian Mixture Model (GMM) on eight-week smart card data from Guangzhou to recognize the spatial-temporal patterns of the bus travel demand [19].

The existing literature recently concentrates on the data mining techniques for discovering the spatial and temporal patterns in the public transit system using the smart card data. Most of the studies focused on discovering the spatial patterns or regularity levels at the stop level in the public transit network. Stop or route-based perspectives ignore whether the same or different passengers make the trips. Also, a few number of studies examined the temporal patterns discretising temporal dimension using time windows. However, none of those studies mentioned above explored the spatial and temporal aspects of the passengers' trips with a passenger-based perspective, which focuses on the passengers as dynamic objects characterized by both boarding and alighting transactions' time and location; the previous studies consider just boarding or alighting transactions at different levels of their study. In addition, none of the above-mentioned studies considered the direction in measuring spatial similarity nor considered time in a continuous linear space. Furthermore, no studies above investigated the correlation between the spatial similarity and temporal similarity of the passengers. Therefore, examining questions like "how the passengers' trips are correlated in the spatial and temporal dimensions" or "whether passengers with similar spatial trip patterns would be likely to have similar temporal trip patterns (i.e., passengers travelling similar places would also choose similar departure times)" or "what is the temporal (spatial)

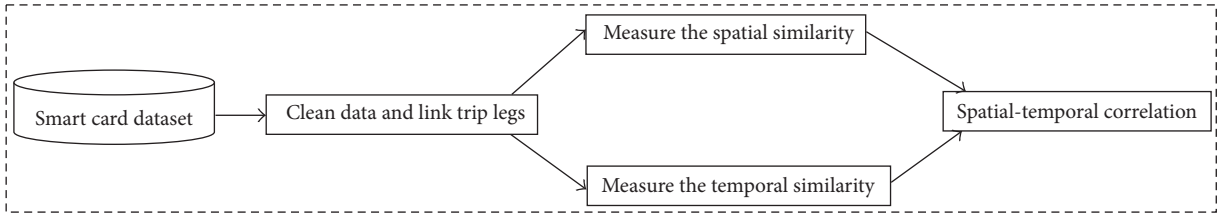

Figure 1: Methodology overview.

similarity between two passengers' trips given their spatial (temporal) similarity" or "is there any relation between spatial and temporal similarity values with trip length or time" are neglected in the literature; in other words, the paper aims to answer these questions.

Also, discovering the correlation can help to improve the Demand Responsive Transit (DRT) and friend recommendation systems in the public transit network. Knowing the correlation between the spatial and temporal similarities, performance of the DRT system can be improved in two ways. First, spatial or temporal similarity values can be predicted by knowing the correlation between them; hence, it would be just necessary to measure one of the spatial or temporal similarity values (e.g., datasets that include just spatial or temporal attributes). Second, the conditional probability models can determine the probability of having the spatial or temporal similarity at different ranges of the similarity values. The spatial and temporal similarity values have different relations in different ranges; hence, it can be used to design different DRT services according to outcome of the conditional probability models. Furthermore, the probability of encountering two passengers in the public transit network can be predicted considering the spatial-temporal similarity correlation, which leads to improving the performance of the current friend recommendation services.

3. Methodology

The proposed method aims to investigate the spatial similarity and temporal similarity between the passengers to discover the correlation between the similarities. It uses smart card data to reconstruct passenger trips. Also, it develops the spatial and temporal similarity measures in the public transit network. The measures are used to calculate the similarity matrices. The spatial and temporal similarity matrices are used to draw histograms, calculate Pearson correlation coefficients, plot hexagonal binning diagrams, and examine the relations between trip time and length with the similarities' values. Figure 1 shows the main steps of the methodology.

The smart card dataset includes time and location for the boarding and/or alighting transactions. The dataset first needs to be cleaned [20]. A trip leg is a distance and period between a consecutive boarding and alighting transactions of a passenger. A trip consists of one or more trip legs. Usually, two or more trip legs are joined as a trip based on the time gap between the consecutive alighting and subsequent boarding transactions. Various thresholds are examined for the time gap. Based on the analyses of Alsger et al. (2016) [21], the time gap is considered as 30 minutes in this study. If the time gap between two consecutive trip legs is less than 30 minutes, then the trip legs will link together as a trip.

A trip is a movement in both spatial and temporal dimensions which have different concepts and metrics to quantify. The spatial space is a 2-dimensional plane, in which objects can move back and forward in both dimensions, while the temporal space is a 1-dimension linear space, in which objects just can go forward. Also, the spatial and temporal dimensions have different units like meters and minutes. Hence, the spatial and temporal dimensions of the movement are studied in separate frameworks. In addition, a passenger moves in the network simultaneously in both spatial and temporal dimensions. Figure 2 shows an example for spatial and temporal movements by a passenger who goes from stop A to stop B and then stop B to stop C.

The proposed method for measuring the similarities is passenger-based, although most of the literatures are stops or route-based studies. Considering passengers as dynamic objects with one or more trips in the public transit network can disclose undiscovered behaviours in the network. Stop or route-based studies are usually indifferent to the passengers; they emphasize the trips disregarding whether one passenger or different passengers make the trips. However, the passenger-based perspective models the passengers with all their trips during a day. One or more trips characterize the passenger behaviour during a day. The passenger-based perspective discovers relations between the passengers. For instance, it can investigate the similarity between two passengers' behaviour in the public transit network.

3.1. Spatial Similarity. There are some trajectory similarity measures such as longest common subsequence (LCS), Fréchet distance, dynamic time warping (DTW), and TRACLUS. Briefly, LCS defines a spatial proximity threshold and two points are considered as similar or not based on the threshold; Fréchet minimizes the maximum distance between two trajectories; DTW minimizes the sum of distances at each point of the trajectories; TRACLUS is a density-based clustering algorithm that partitions the trajectories and considers closer parts as similar. The ones above are the major measures introduced, studied, and compared in the literature. Each measure has its pros and cons that make them suitable for specific applications and networks [22–28].

All the measures above emphasize the distance between the trajectories' points as the main criterion of the similarity. They are suitable for trajectory datasets that consist of location measurements every few meters. However, the smart

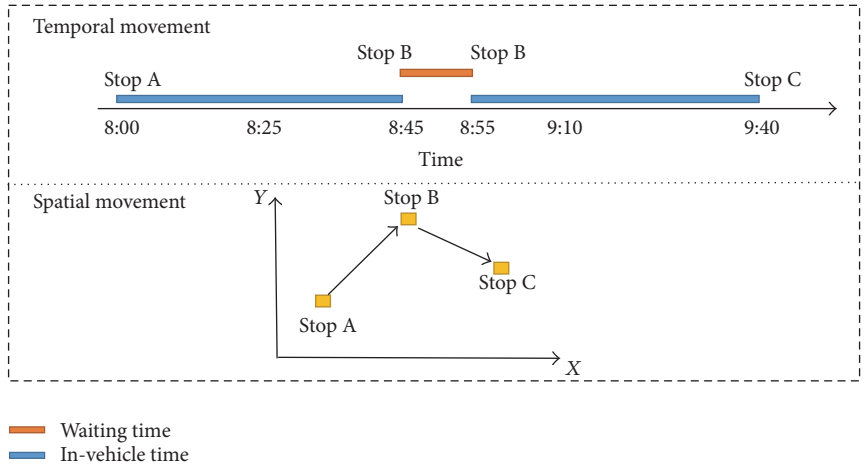

Figure 2: Spatial and temporal movements by a passenger.

card datasets include just two locations for each trip legs. Also, there are some cases in the public transit network that need to consider direction or angle between the trajectories as well as the distance. For instance, two passengers board on the same stop but travel to different stops that are located at different directions; in this case, the boarding stops are at the same location, but the trajectories are not similar. Regarding the existing algorithms, the public transit network, the structure of the smart card data, and the passenger-based perspective two criteria are considered for verifying the spatial similarity between the passengers' trips:

(1) The distance between the origins or the destinations

(2) The direction of the trips

The distance is assumed as 600 meters based on the studies in travel behaviour of public transit passengers and the walking speed [21, 22, 29]. Also, the direction of the trips is defined as the angle between the two trips; if the minimum angle between the two trips is between 0 and 6 degrees, then the two trips will be assumed in the same direction. The maximum allowable angle is considered as 6 degrees regarding the average trip length (18 km) in the case study [30]; shorter (longer) average trip length in different case studies needs bigger (smaller) maximum allowable angle. Considering the two indices together determines the similar trips in the same corridor. Therefore, two trips will be similar if the distance between the origins (destinations) is less than 600 meters and the angle between the two trips is less than 6 degrees. The value of the spatial similarity between the two trips is calculated as the ratio of the shorter trip length to the longer one. Figure 3 shows examples of the spatial similarity conditions.

Equation (1) presents the spatial similarity measure between trips (T_1, T_2) that are between (O_1, D_1) and (O_2, D_2), where "O" stands for origin and "D" for destination; "T" stands for trips; "$d(P_1, P_2)$" is the distance function that measures Euclidean distance between two points; "$l(O, D)$" is the length function that measures length of the trips; "$\mathrm{di}(T_1, T_2)$" is the direction function that measures angle between two trips; and "$\mathrm{SS}(T_1, T_2)$" is the spatial similarity value between two trips.

Spatial similarity measure between trips is as follows:

$$\mathrm{SS}(T_1, T_2) = \begin{cases} \dfrac{\min(l(T_1, T_2))}{\max(l(T_1, T_2))} & \text{if } [(d(O_1, O_2) < 600\,\mathrm{m} \parallel d(D_1, D_2) < 600\,\mathrm{m}),\ (\mathrm{di}(T_1, T_2) < 6^\circ)] \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Equation (1) is appropriate for a pair of passengers each of which has just one trip. The final spatial similarity value for a pair of passengers, who have more than one trip, is assumed as the ratio of the sum of lengths of the shorter similar trips to the greater sum of lengths of all the trips belonging to the pair of passengers. For instance, if passenger A has two trips with lengths of 3 and 6 km and passenger B has one trip with a length of 4 km that closely overlaps with passenger A's 3 km trip, then the spatial similarity between these two passengers will be $(3/(3 + 6)) * 100 = 33\%$. Also, it should be noted that similar trips for one passenger are just considered as one trip; if a passenger has some trips on the same route, then just one of them will be considered for the spatial similarity measurement. In addition, it is assumed that each trip can have similarity with just one other trip; if a trip has similarity with more than one trip, then the greatest value among the similarity values will be chosen. Moreover, the spatial similarity between passenger A and

Table 1: An example for discrete temporal similarity.

	Boarding time/time window	Alighting time/time window
Passenger 1	7:30/E	8:20/F
Passenger 2	8:01/F	8:21/F
Passenger 3	7:48/E	7:59/E

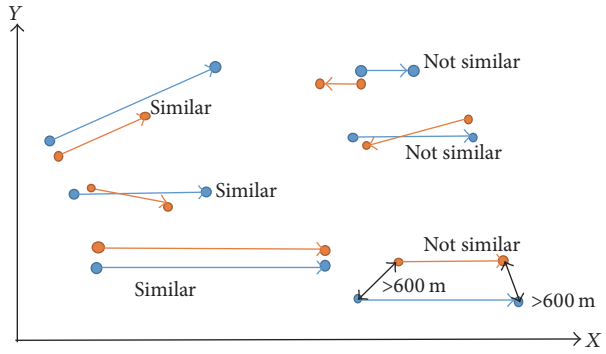

Figure 3: Spatial similarity examples between trips.

passenger B is not necessarily equal to the spatial similarity between passenger B and passenger A since the total journey length of passengers A and B can be unequal. The spatial similarity value is assumed as the minimum of the similarities between the two passengers to have a symmetric similarity value between two passengers. Based on the discussion above, Algorithm 1 presents a pseudo-code for spatial similarity between passengers (P_1, P_2) who, respectively, have m and n unique trips, where "p" stands for a passenger; "a" is defined for measuring sum of the lengths of shorter similar trips; "a_{12}" is sum of the lengths of shorter similar trip between passenger 1 and passenger 2; "a_{21}" is sum of the lengths of shorter similar trip between passenger 2 and passenger 1; "B" is a set of similar trips, in which the longest one is chosen to determine the shorter similar trip; and the other parameters are defined previously.

3.2. Temporal Similarity. The next step is measuring the temporal similarity. Previous studies mostly consider temporal dimension as a discrete variable and model it using time windows, which leads to biased results especially in times closer to the threshold or the boundary of time windows. Also, the previous studies consider boarding or alighting time as a representative of the temporal aspect of the movements. For instance, time windows are between 7 a.m. to 8 a.m. (called it "E") and 8 a.m. to 9 a.m. (called it "F"), and three passengers travel as in Table 1. Passengers 1 and 2 have the alighting time in the same time window but boarding time in the different time windows; passengers 1 and 3 have the boarding time in the same time window but alighting time in the different time windows; passengers 2 and 3 have the both boarding and alighting time in different time windows. Now, if boarding time is considered as the similarity criterion, then passengers 1 and 2 will be similar; however, if alighting time is considered as the similarity index, then passengers 1 and 3 will be similar. In addition, just one-minute boarding and alighting before or after time thresholds can affect the results, which bias the similarity measure.

Regarding the instance provided in Table 1, it is necessary to develop a similarity measure in a continuous space for the temporal dimension of the trips in the public transit network. The temporal similarity should measure the joint period in which two passengers use the public transit. The proposed temporal metric considers time as a linear continuous element of the movement; the first boarding transaction and last alighting transaction time model the trips; it is capable of considering both boarding time and alighting time in measuring the temporal similarity. Equation (2) presents the temporal similarity measure between two trips (T_1, T_2) that, respectively, are between (B_1, A_1) and (B_2, A_2), where "B" stands for boarding time and "A" for alighting time; "TS(T_1, T_2)" stands for the temporal similarity value. The temporal similarity value between two trips is assumed as the ratio of overlapped trip time to longer trip time.

Temporal similarity measure between trips is as follows:

$$\mathrm{TS}(T_1, T_2) = \begin{cases} \dfrac{[\min(A_1, A_2) - \max(B_1, B_2)]}{[\max((A_1 - B_1), (A_2 - B_2))]} & \text{if } [(B_1 > B_2,\ A_1 < A_2) \parallel (B_2 > B_1,\ A_2 < A_1)] \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Passengers can have more than one trip during a day. Trips of a passenger are temporally unique; a passenger cannot have more than one trip at the same period. The overlapped time between two trips of two passengers cannot be covered with any other trips time of these two passengers. Hence, calculating the temporal similarity between two passengers with multiple trips is simpler than the spatial similarity measuring. The temporal similarity between two passengers is assumed as the ratio of the sum of the overlapped time between the trips to the greater sum of the all trips time. Figure 4 illustrates an example for measuring the temporal similarity between two passengers.

$a = 0$;
for (i in 1: m) {$B = \varnothing$;
  for (j in 1: n) {
    if (T_i is spatially similar to T_j) {add T_j to B;}
    }
  $a = a + \min(l(T_i, \max(l(T \in B))))$;}
$\mathrm{SS}(P_1, P_2) = [\min(a_{12}, a_{21})]/[\max(\sum_{i=1}^{m} l(T_i), \sum_{j=1}^{n} l(T_j))]$;

Algorithm 1: Pseudo-code for calculating spatial similarity between passengers.

$a = 0$;
for (i in 1: m) {
  for (j in 1: n) {
    if (T_i is temporally overlapped with T_j) {$a = a + \mathrm{OT}(T_i, T_j)$;}
    }
  }
$\mathrm{TS}(P_1, P_2) = a/[\max(\sum_{i=1}^{m} l(\mathrm{TT}_i), \sum_{j=1}^{n} l(\mathrm{TT}_j))]$;

Algorithm 2: Pseudo-code for temporal similarity between passengers.

Algorithm 2 presents pseudo-code for the spatial similarity between two passengers (P_1, P_2) who, respectively, have m and n trips, where "TT" stands for trip time; "$\mathrm{OT}(T_1, T_2)$" stands for overlapped time that is calculated between two trips similar to equation (2); "a" is defined for measuring the overlapped time; and the other parameters are defined previously.

3.3. Correlation between Spatial Similarity and Temporal Similarity. Correlation discovers statistical relationships between usually two variables. The relationships can be linear or nonlinear. Pearson correlation coefficient is used to examine the linear correlation between two variables. The coefficient is measured on a scale with no units and can take a value from −1 through 0 to +1. The values close to zero mention no linear correlation and values close to +1 or −1 imply a perfect linear correlation [31]. In addition, a visualisation technique called hexagonal binning is used to discover the nonlinear correlation between the spatial similarity and temporal similarity values. Conventional methods such as scatter plots cannot efficiently visualise large datasets because many of data are overlapped. The technique pairwise plots the spatial similarity and temporal similarity values [32].

4. Results

The used smart card dataset is from TransLink, the public transport authority of South East Queensland (SEQ), Australia. The dataset for three weekdays and one weekend day from the South East Queensland SEQ bus, train, and ferry modes are selected. Wednesday to Saturday (20–23 March 2013) are chosen as the weather on all four days was normal, and there were no special events during those days. 20,000 passengers randomly are selected for each day, who approximately make 45,000 trip legs per day. The sample size for each day is almost 10% of the whole number of transactions. Considering the analysis from Alsger et al. (2017), the sample size can appropriately represent the whole dataset [33]. The dataset includes both time and location of boarding and alighting transactions, which is an important privilege of TransLink smart card dataset, while most of the automated fare collection systems around the world just include boarding or alighting transactions.

4.1. Spatial Similarity Results. Each similarity matrix consists of 400 million cells given that it is a symmetric matrix with a size of 20,000∗20,000 passenger^2. The maximum value for the similarity between two passengers is 1, and the minimum is 0. The spatial similarity matrix is calculated for each day based on (1) and Algorithm 1. 585 passengers on Wednesday, 605 on Thursday, 701 on Friday, and 572 on Saturday had no spatial similarity with others, which is equivalent to 3% of the passengers. It can be interpreted as the high similarity in the usage of the public transit network.

The maximum frequency of the spatial similarity values happens at the range of 0.05, 0.1 and then it decreases until the range of 0.45, 0.5. Two apparent irregularities occur in the diagrams: the first one at the range of (0.45, 0.5) and the second one at (0.95, 1). The former is related to passengers, who at least (approximately) half-length of their trips is in the same corridor. The reason for the first irregularity can be related to the geometry of the public transit network that is divided into two parts by a river and the location of main hubs stations. Figure 6 presents an example of having spatial similarity between 0.45 and 0.5; it shows the trips of two passengers in the network who have 48% spatial similarity. The second irregularity means it is more likely to have complete spatial similarity between two passengers rather than a spatial similarity between the range of 0.5, 0.95; it includes passengers who have the complete spatial similarity

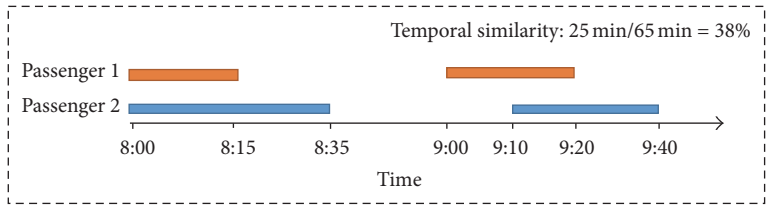

Figure 4: An example for the temporal similarity measure between passengers.

Table 2: Pearson values.

	Pearson (spatial similarity, temporal similarity)
Wednesday	0.04
Thursday	0.03
Friday	0.03
Saturday	0.01

like passengers who go from a specific suburb to city business district for work and return home using the same corridor. Histograms of Figure 5 show the distribution of similarity values in the similarity matrices.

4.2. Temporal Similarity Results. The temporal similarity matrices for the four days are generated using (2) and Algorithm 2. All the passengers have the temporal similarity with at least one other passenger. It can be interpreted as effective usage of the public transit during the day. Figure 7 represents histograms for the values of the temporal similarity for each day. All the histograms schematically are similar; the highest frequency for each histogram happens at the range of 0, 0.1 and then the frequency of the temporal similarity values decreases consistently. Lower values of the temporal similarity are more likely to happen rather than the higher values.

4.3. Results on Correlation between Spatial Similarity and Temporal Similarity. Correlation between the spatial and temporal similarity matrices can help to understand how spatial similarity and temporal similarity between two passengers are related together. Also, examining the correlation can contribute to developing a prediction model for predicting the probability of passengers encountering in the public transit. Pearson correlation coefficient is used to investigate the linear dependency between the two similarity matrices. Table 2 shows the Pearson values. Values of the coefficient for each day are between 0.01 and 0.04 which implies a weak linear dependency; a particular line cannot be allocated to scatter plots of the similarity matrices. Also, positive values of the correlations mean the spatial and temporal similarity values have a weak uphill (rather than a downhill) linear relationship.

At each point of the hexagonal binning diagrams (specific values of the spatial and temporal similarities), size and colour of the hexagonals represent the number of passenger pairs. The diagrams can be divided into five areas of different colours. In all of them, it is more likely to have some temporal similarity with zero spatial similarity rather than have some spatial similarity with zero temporal similarity, because number of the passenger pairs on the vertical axis (spatial similarity = 0) is more than the ones on the horizontal axis (temporal similarity = 0); in other words, making trips in the similar periods of a day is more probable than in the similar routes. Also, the density of the diagrams decreases by receding from the origin point. At all the charts, the spatial similarity between 0.46 and 0.52 identifies a boundary after which density changes; the range 0.46, 0.52 of the spatial similarity values is close to the range of 0.45, 0.5 at the spatial similarity histograms, which has a higher frequency among its neighbours. Moreover, there is more consistency in the temporal similarity rather than the spatial similarity. For instance, for all the spatial similarity values between 0 and 1, increasing the temporal similarity decreases the probability of having the same spatial similarity. Also, if the temporal similarity is between 0 and 0.2, the spatial similarity is between 0.52 and 1, and if the spatial similarity increases, the probability of having the same temporal similarity will increase. The results from the hexagonal diagrams can be used in developing a conditional probability model. For instance, P (spatial similarity | temporal similarity = 0.2) > P (spatial similarity | temporal similarity = 0.6) and P (temporal similarity | spatial similarity = 0.2) > P (temporal similarity | spatial similarity = 0.4). Figure 8 presents the hexagonal binning diagrams.

Figure 9 presents the relation between spatial similarity and trip length. The vertical axis is the spatial similarity, and the horizontal axes are the trip length of two passengers. Each point in the diagram is a triple (the spatial similarity (passenger i and passenger j), the trip length of passenger i, and the trip length of passenger j). All the four diagrams are schematically similar. A shape is figured at the middle of each diagram. The shape is like a wall at the $x = y$, which implies equal trip length. It means the spatial similarity for the passengers with close trip length is more likely to happen. In other words, if the difference between the lengths of trips of two passengers increases, then the probability of having the spatial similarity will decrease.

Figure 10 presents the relation between temporal similarity and trip time of passengers. Vertical axis presents the values for the temporal similarity, and the horizontal axes represent the trip time of two passengers. The diagrams include triples (the temporal similarity (passenger i and passenger j), the trip time of passenger i, and the trip time of passenger j). A similar shape is figured in all the diagrams. The shape is similar to a cone. It implies that higher values of the temporal similarity happen at a focal range, between 60 and 90

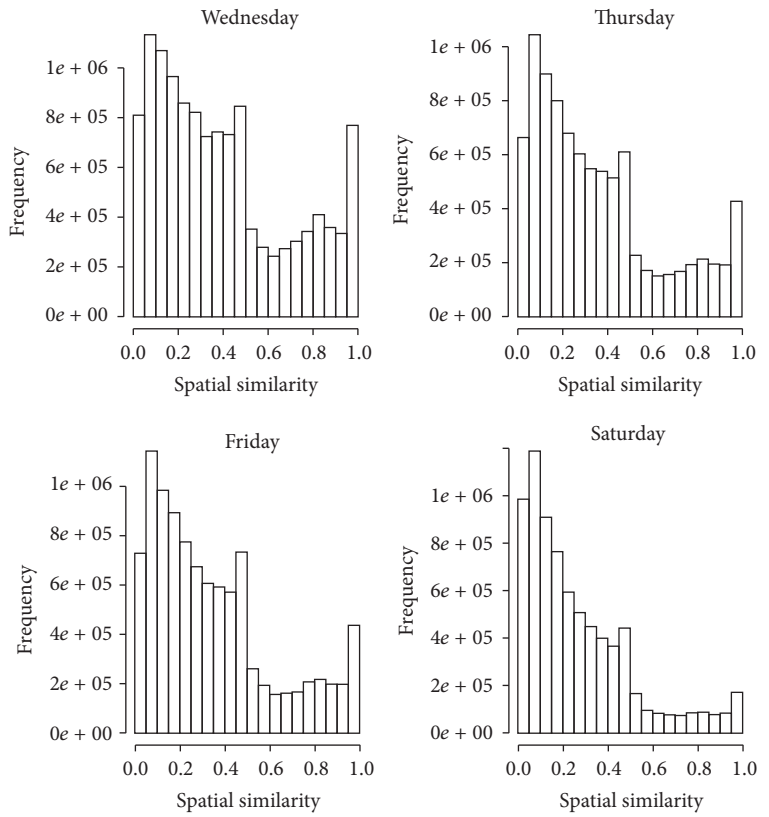

Figure 5: Spatial similarity histograms.

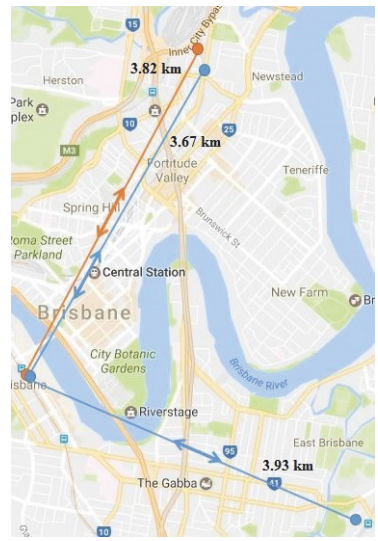

Figure 6: Example of spatial similarity range of 0.45–0.5.

minutes. Passengers with the trip time between 60 and 90 minutes are more likely to have greater values of the temporal similarity. Also, it means if the trip time is between 60 and 90 minutes, then it will be more likely to have the temporal similarity between the passengers.

All the analyses are done for the four days that include three weekdays and one weekend. Achieved results from the histograms, coefficients, and diagrams are close to each other. The similarity between the results from the different days can prove the stability of the method and analyses. The similarly

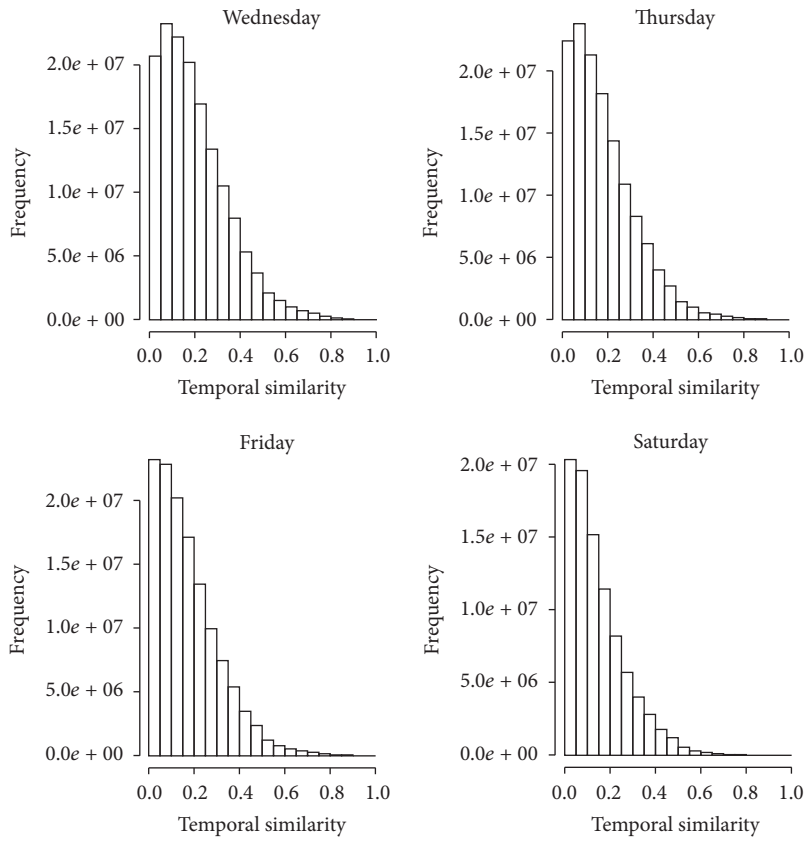

Figure 7: Temporal similarity histograms.

discovered correlations between the spatial and temporal similarity matrices of the passengers for various days validate the results.

5. Conclusion

This paper measures the spatial similarity and temporal similarity of public transit passengers with a passenger-based perspective, in which each passenger is modelled by one or more trips. The spatial and temporal similarity measures are developed for the public transit network. The spatial similarity measure considers direction as well as the distance between the trips of the passengers. The temporal similarity measure considers both boarding and alighting time in a continuous linear space. Furthermore, this paper investigates the spatial-temporal similarity correlation between passengers of the public transit system. The related similarity matrices are calculated for four-day smart card datasets including approximately 45,000 trip legs of 20,000 passengers per day. The values of similarity matrices are examined using histograms. A linear correlation between spatial and temporal similarity matrices is calculated using Pearson coefficient. The hexagonal binning technique is used to plot the frequency of correspondence values of the spatial and temporal similarity matrices. In addition, relations between the spatial similarity and the trip length of the passengers are explored by plotting 3-dimensional scatters and density diagrams. Also, 3-dimensional scatters and density diagrams of temporal similarity-trip time-trip time is plotted for investigating the relation between the temporal similarity and the trip time.

The passenger-based perspective leads to revealing the spatial and temporal relations between the passengers. 97% of the passengers have a level of spatial similarity with at least one other passenger in the dataset. Also, all passengers have a level of temporal similarity with at least one other passenger. In addition, the spatial similarity histograms show more frequency for lower values of the similarity excepting the two intervals of 0.45, 0.5 and 0.95, 1. The geometry of the network triggers the former range, and the latter shows the inclination of the passengers for having the complete spatial similarity. Likewise, values of the temporal similarity matrices are more frequent with lower values of the similarity at all the ranges; passengers are more likely to have smaller temporal similarity rather than greater values.

The spatial and temporal measures are calculated for all the pairs of passengers. The developed measures show relations with trip time and length. Passengers with closer trips length are more likely to have the spatial similarity; if the difference between the trip lengths of two passengers increases, then the probability of having the spatial similarity will decrease. Also, passengers with the trip time of 60–90 minutes are more likely to have higher temporal similarity with each other; if the trip time is between 60 and 90 minutes, then it will be more likely to have the temporal similarity between the passengers.

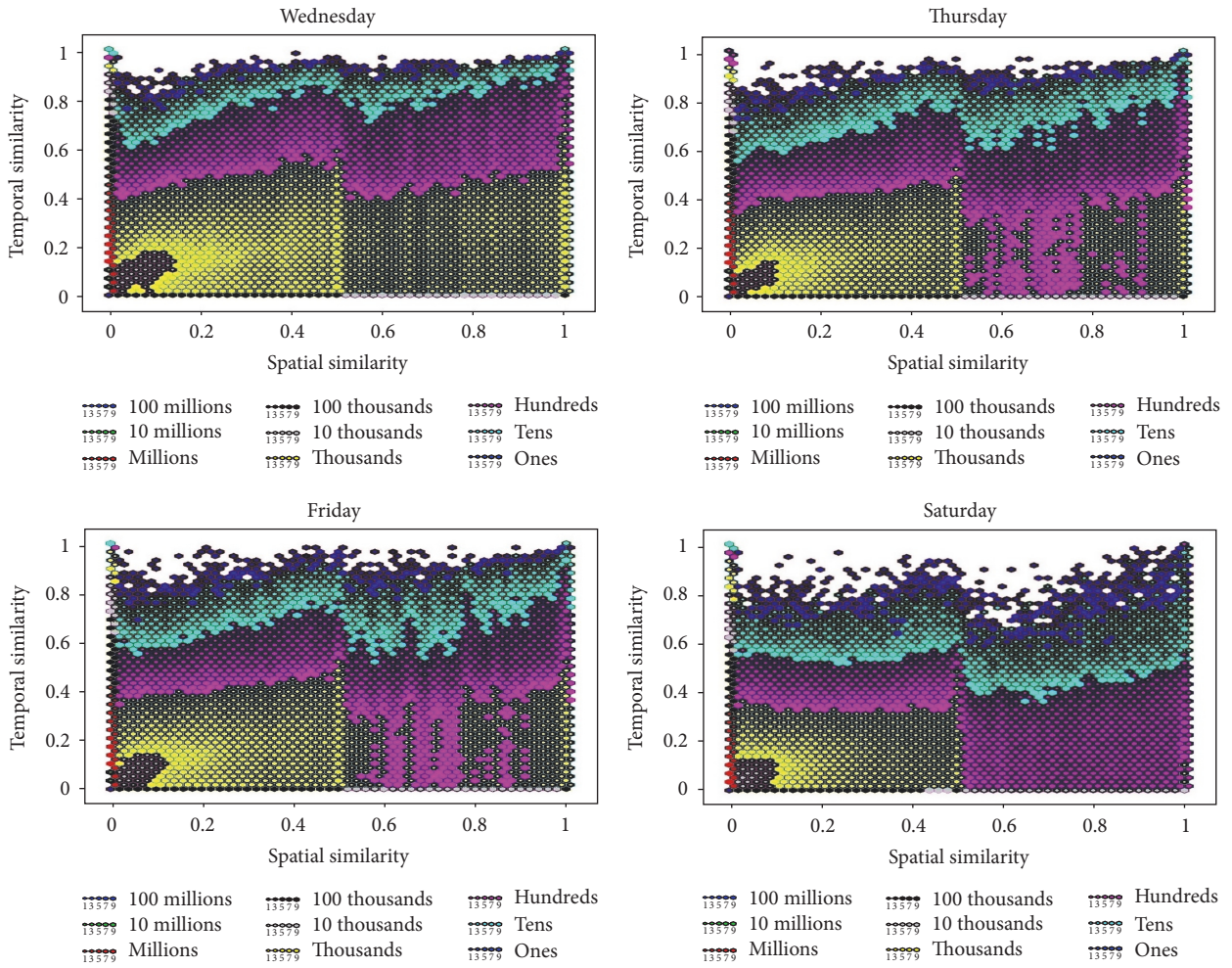

Figure 8: Hexagonal binning diagrams.

The examined correlation between the spatial and temporal similarity matrices shows a nonlinear dependency. The Pearson coefficient presents a weak linear correlation, close to zero, between the similarity matrices; positive values of the coefficient imply an uphill relationship between the spatial and temporal similarity values. Also, the hexagonal binning diagrams present nonlinear correlation with the specific patterns; the diagrams can be divided into separate sections, and specific trends can be extracted from each section, which would develop probabilistic models.

The computational complexity should be discussed. The computational complexity of the method is O ($n^2/2$) because of symmetric nature of the similarity matrix, where n is number of the passengers. Also, the structure of the similarity matrix helps to expedite the computation process because almost 90% of the cells in the matrix are zero that can be transferred to null cells for reducing the size of the matrix and computing more efficiently. Therefore, according to the structure of the similarity matrix and the computational complexity, some techniques such as cloud or parallel computing can easily handle the computations [34].

Discovering correlation between the spatial and temporal similarities of the passengers can answer the questions about the relations between the spatial and temporal similarities. Also, understanding the correlation can improve the efficiency of DRT and friend recommendation systems. DRT systems usually work based on the similar spatial and temporal patterns of the passengers' trips; if a group of passengers has similar travel patterns in both spatial and temporal dimensions, then a DRT service can be allocated to them. Spatial and temporal patterns consider passengers in groups, while the correlation can analyse relations between the passengers at individual levels and for all pair of the passengers (the patterns and correlation are two different perspectives). By knowing the correlation between the spatial and temporal similarities of two passengers, performance of the DRT system can be improved in two ways. First, spatial or temporal similarity values can be predicted by knowing the correlation between them; hence, it would be just necessary to measure one of the spatial or temporal similarity values (e.g., datasets that include just spatial or temporal attributes). Second, the conditional probability models can determine

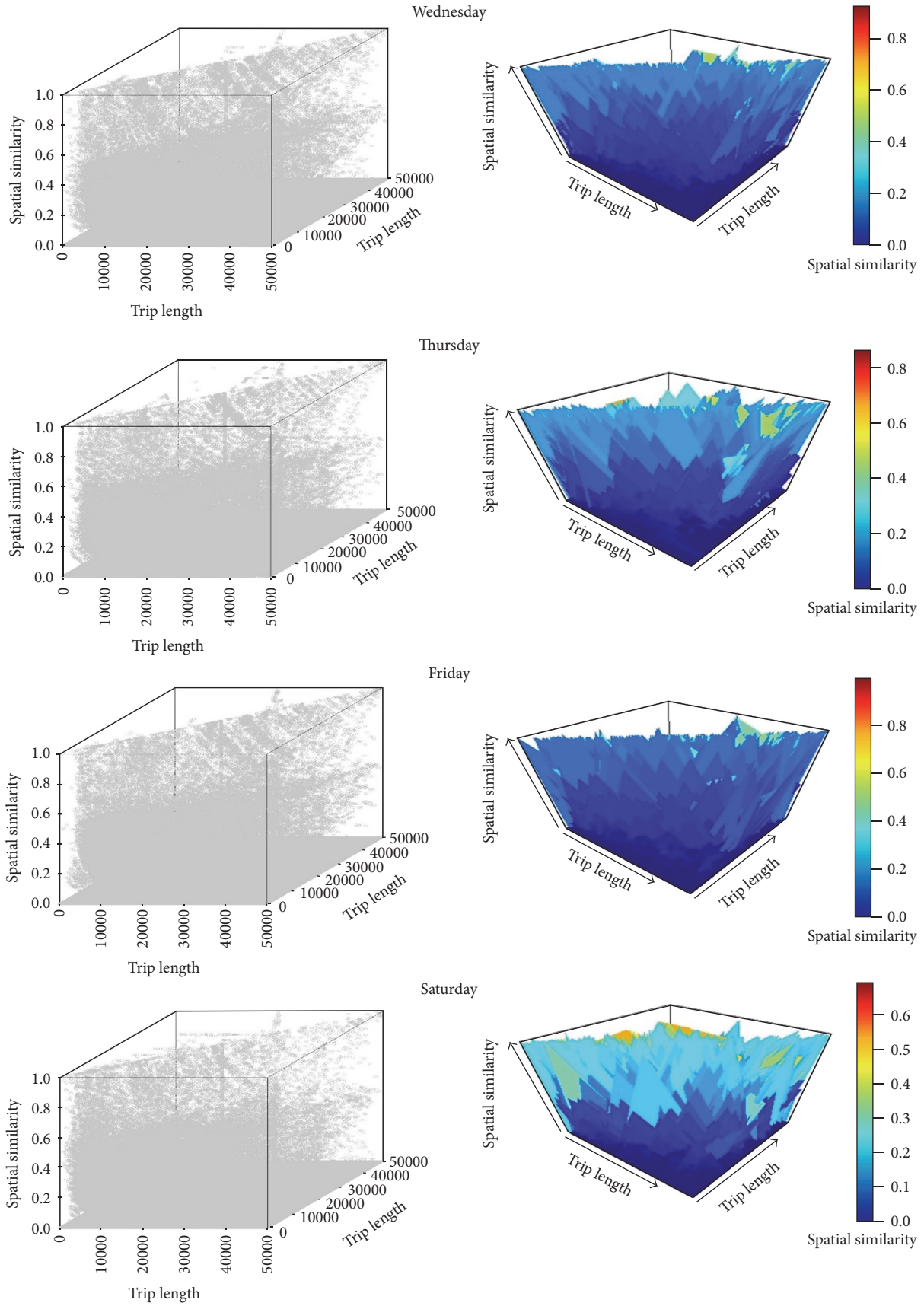

Figure 9: Spatial similarity and trip length diagram.

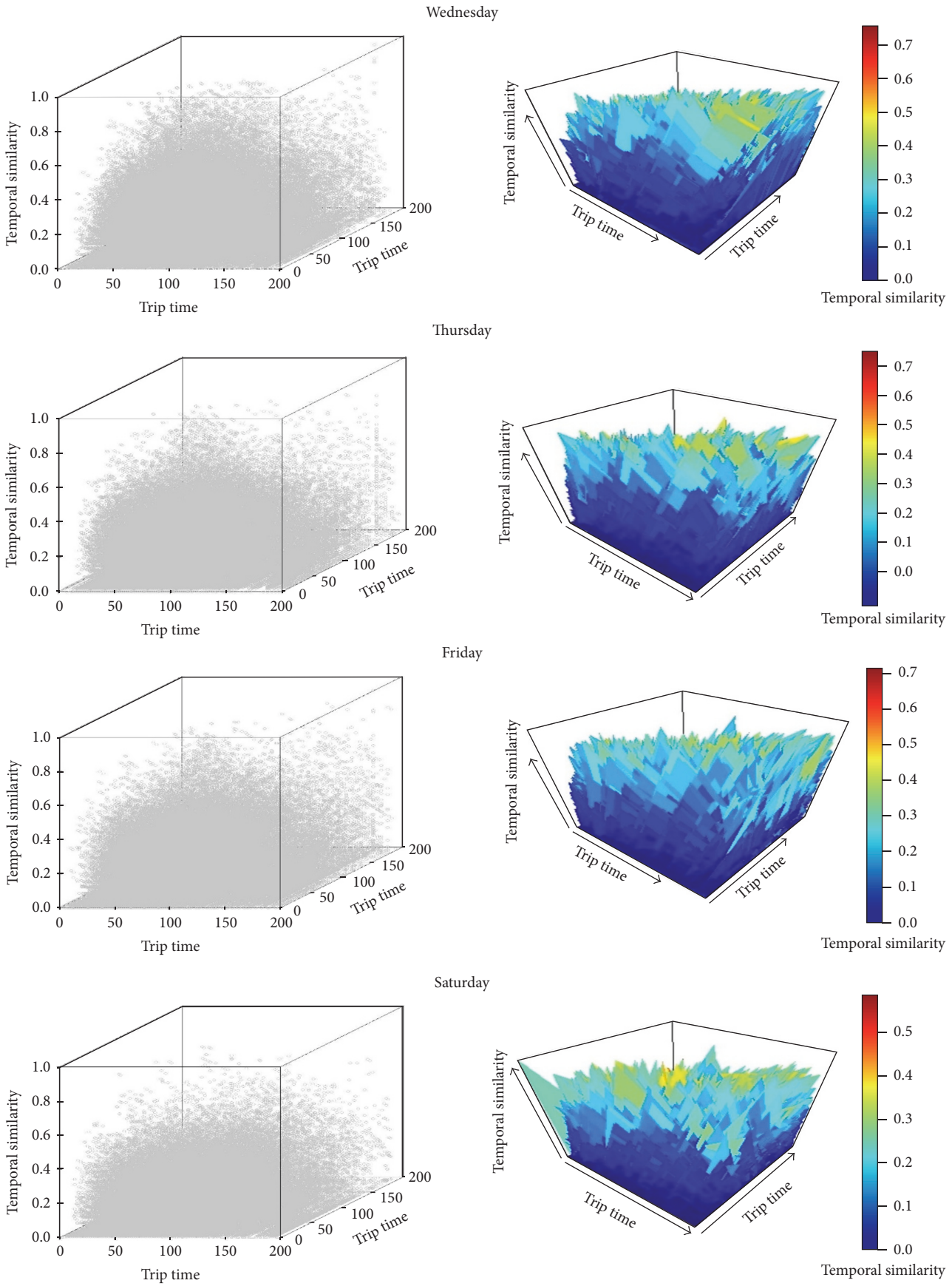

Figure 10: Temporal similarity and trip time diagram.

the probability of having the spatial or temporal similarity at different ranges of the similarity values. The spatial and temporal similarity values have different relations in different ranges; hence, it can be used to design different DRT services according to the outcome of conditional probability models. Furthermore, the probability of encountering two passengers in the public transit network can be predicted considering the spatial-temporal similarity correlation, which leads to improving the performance of the current friend recommendation services.

Additional analyses can be performed to extend this work. The proposed method may be implemented on public transit systems in other cities and the results to be compared. Also, the effects of other parameters, such as geographical location or start or end time of trips, on the spatial-temporal similarity correlation could be examined if quality data becomes available. In addition, it would be valuable to develop a local search method to do the computation just for the passengers with similarity; in other words, the proposed method discovers passengers who do not have any potential for having spatial or temporal similarities. Furthermore, the spatial-temporal similarity correlation can be reviewed with trip purpose similarity of the passengers.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. J. Miller, M. Raubal, and Y. Jaegal, "Measuring space-time prism similarity through temporal profile curves," in *Proceedings of the In Geospatial Data in a Changing World*, pp. 51–66, 2016.

[2] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: a literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[3] M. Trépanier, C. Morency, and B. Agard, "Calculation of transit performance measures using smartcard data," *Journal of Public Transportation*, vol. 12, no. 1, pp. 79–96, 2009.

[4] B. Y. Chen and W. H. K. Lam, "Special issue: smart transportation: theory and practice," *Journal of Advanced Transportation*, vol. 50, no. 2, pp. 141–144, 2015.

[5] Z.-J. Wang, X.-H. Li, and F. Chen, "Impact evaluation of a mass transit fare change on demand and revenue utilizing smart card data," *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 213–224, 2015.

[6] H. Faroqi and A. Sadeghi-Niaraki, "GIS-based ride-sharing and DRT in Tehran city," *Public Transport*, vol. 8, no. 2, pp. 243–260, 2016.

[7] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W. Y. Ma, "Mining user similarity based on location history," in *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, p. 34, 2008.

[8] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 34, pp. 13774–13779, 2013.

[9] Y. Xu, H. Chen, Q. J. Kong, X. Zhai, and Y. Liu, "Urban traffic flow prediction: a spatio-temporal variable selection-based approach," *Journal of Advanced Transportation*, vol. 50, no. 4, pp. 489–506, 2015.

[10] H. Nishiuchi, J. King, and T. Todoroki, "Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data," *International Journal of Intelligent Transportation Systems Research*, vol. 11, no. 1, pp. 1–10, 2013.

[11] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, no. Part C, pp. 1–12, 2013.

[12] S. Tao, J. Corcoran, I. Mateo-Babiano, and D. Rohde, "Exploring Bus Rapid Transit passenger travel behaviour using big data," *Applied Geography*, vol. 53, pp. 90–104, 2014.

[13] S. Tao, D. Rohde, and J. Corcoran, "Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap," *Journal of Transport Geography*, vol. 41, pp. 21–36, 2014.

[14] L.-M. Kieu, A. Bhaskar, and E. Chung, "A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 193–207, 2015.

[15] M. S. Ghaemi, B. Agard, V. P. Nia, and M. Trépanier, "Challenges in spatial-temporal data analysis targeting public transport," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 442–447, 2015.

[16] E. Manley, C. Zhong, and M. Batty, "Spatiotemporal variation in travel regularity through transit user profiling," *Transportation*, pp. 1–30, 2016.

[17] X. Ma, C. Liu, H. Wen, Y. Wang, and Y. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.

[18] M. K. El Mahrsi, E. Côme, L. Oukhellou, and M. Verleysen, "Clustering Smart Card Data for Urban Mobility Analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, 2016.

[19] C. Yu and Z.-C. He, "Analysing the spatial-temporal characteristics of bus travel demand using the heat map," *Journal of Transport Geography*, vol. 58, pp. 247–255, 2017.

[20] S. Robinson, B. Narayanan, N. Toh, and F. Pereira, "Methods for pre-processing smartcard data to improve data quality," *Transportation Research Part C: Emerging Technologies*, vol. 49, pp. 43–58, 2014.

[21] A. Alsger, B. Assemi, M. Mesbah, and L. Ferreira, "Validating and improving public transport origin-destination estimation algorithm using smart card fare data," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 490–506, 2016.

[22] K. Bringmann, "Why walking the dog takes time: Frechet distance has no strongly subquadratic algorithms unless SETH fails," in *Proceedings of the 2014 IEEE 55th Annual Symposium, In Foundations of Computer Science (FOCS)*, pp. 661–670, October, 2014.

[23] C. Hu, N. Luo, and Q. Zhao, "Fast fuzzy trajectory clustering strategy based on data summarization and rough approximation," *Cluster Computing*, vol. 19, no. 3, pp. 1411–1420, 2016.

[24] J. Kim and H. S. Mahmassani, "Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories," *Transportation Research Procedia*, vol. 9, pp. 164–184, 2015.

[25] M. Müller, "Information retrieval for music and motion," *Information Retrieval for Music and Motion*, pp. 1–313, 2007.

[26] J. Shen and T. Cheng, "A framework for identifying activity groups from individual space-time profiles," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1785–1805, 2016.

[27] H. Wang, H. Su, K. Zheng, S. Sadiq, and X. Zhou, "An effectiveness study on trajectory similarity measures," in *Proceedings of the In Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137*, pp. 13–22, 2013.

[28] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 3, article 29, 2015.

[29] D. A. Cunningham, P. A. Rechnitzer, M. E. Pearce, and A. P. Donner, "Determinants of self-selected walking pace across ages 19 to 66," *Journals of Gerontology*, vol. 37, no. 5, pp. 560–564, 1982.

[30] https://translink.com.au/about-translink/reports-and-publications.

[31] P. Sedgwick, "Pearson's correlation coefficient," *British Medical Journal*, vol. 345, article e4483, 2012.

[32] N. Lewin-Koh, Hexagon binning: an overview. 2011. https://cran.r-project.org/web/packages/hexbin/vignettes/hexagon_binning.pdf.

[33] A. Alsger, A. Tavassoli, M. Mesbah, and L. Ferreira, "Evaluation of effects from sample-size origin-destination estimation using smart card fare data," *Journal of Transportation Engineering*, vol. 143, no. 4, article 04017003, 2017.

[34] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016.