Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Investigating the Correlation between Activity Similarity and Trip Similarity of Public Transit Passengers Using Smart Card Data

Hamed Faroqi[a]*, Mahmoud Mesbah[b, c], Jiwon Kim[d]

[a]*PhD candidate, School of Civil Engineering, The University of Queensland, Australia*
[b]*Faculty Member, Department of Civil and Environmental Engineering, Amirkabir University of Technology, Iran*
[c]*Honorary senior lecturer, School of Civil Engineering, The University of Queensland, Australia*
[d]*Lecturer, School of Civil Engineering, The University of Queensland, Australia*

Abstract

The pattern of activities by transit passengers and the factors affecting this pattern are recently addressed in the literature. Also, the pattern of trips and the factors affecting trips are the subject of many studies. However, little is known about the correlation between the activity and trip of passengers in the public transit network. This study investigates how similar are activities of passengers if they have similar trips (or vice versa)? And what factors impact the similarity of the activity and trip of passengers? Answering these questions is useful in understanding the mobilization patterns of passengers and developing group-based transit services. Also, smart card data have provided an opportunity, which was not available before, to analyze the activity and trip of the passengers in a large scale network. In this paper, the correlation between activity similarity and trip similarity of public transit passengers is investigated. The correlation between the activity and trip of the passengers is analyzed using histograms, Pearson correlation coefficient, conditional probabilities, and hexagonal binning technique. In addition, the impact of trip length and duration on the activity and trip similarity are examined using histograms and hexagonal binning diagrams. The proposed methodology is implemented for two-day smart card data in Brisbane, Australia. Results show that it is more likely to have the activity similarity when there is a trip similarity than having the trip similarity when there is activity similarity. Also, there is a nonlinear correlation between the activity and trip similarity with the trip length and duration.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Travel behavior; big data; data mining; trip purpose; space-time

* Corresponding author. Tel.: +61-481247314 .
E-mail address: h.faroqi@uq.edu.au

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

1. Introduction

The pattern of activities by transit passengers and the factors affecting this pattern are recently addressed in the literature. Also, the pattern of trips and the factors affecting trips are the subject of many studies. However, little is known about the correlation between the activity and trip of passengers in the public transit network. This study investigates how similar are activities of passengers if they have similar trips (or vice versa)? And what factors impact the similarity of the activity and trip of passengers? Trip and activity are two inseparable parts of passenger behavior in the public transit network. Individuals use the public transit network to move and perform their desired activity. In other words, the purpose of taking a trip is doing an activity. A trip is a two dimensional (spatial and temporal) movement between an origin and destination stops in the public transit network. Both spatial and temporal dimensions are important and should be considered in the passenger trips modelling (Faroqi et al., 2017). Also, an activity has three dimensions (location, time, and type) that should be considered to have a comprehensive perspective of an activity because, for an example, two activities might be similar in location but be different at the time or type (Shen and Cheng, 2016). Consequently, both trips and activities should be considered as two fundamental elements of the passenger behavior in the public transit network.

Smart card data provides opportunities to study passenger behavior in the public transit network. It consists of boarding and/or alighting transactions of the passenger trips; and it has provided opportunities, which were not available before, to analyze the activity and trip of the passengers in large scale networks. In the past, before emerging Automated Fare Collection (AFC) systems, data for analyzing passenger trips and activities were limited to surveys that usually suffered from the size and ground truth problems (Morency et al., 2007). However, since two decades ago, AFC systems have provided valuable datasets of transactions that have enabled researchers and practitioners in understanding passenger behavior in the public transit network. These datasets, as by-product of fare gathering system in the public transit network, can reconstruct trips of the passengers. Then activities of passengers can be discovered as gaps between subsequent trips of the passengers (Faroqi et al., 2018b). Hence, smart card datasets generated by AFC systems enable researchers to reconstruct trips and activities of the passengers in the public transit network.

Investigating the correlation between the activity and trip similarity of passengers in the public transit network can emerge various applications. For instance, spatial, temporal and trip purpose patterns can be discovered; the network can be designed or upgraded according to the demand for different activities; group-based applications such as group discount or Demand Responsive Transit (DRT) can be developed. Therefore, understanding the correlation between the activity and trip similarity of the passengers through answering the aforementioned questions are useful in discovering the mobilization patterns of passengers, planning TOD models and developing group-based transit services.

This paper, for the first time (to the best of our knowledge), investigates the correlation between activity similarity and trip similarity of public transit passengers. Similarity measures for activity and trip of passengers are defined. Activity similarity considers time, location and type of the activity. Trip similarity considers spatial and temporal aspects of the trip. The correlation between the activity and trip of the passengers is analyzed using histograms, Pearson correlation coefficient, conditional probabilities, and hexagonal binning technique. In addition, impacts of trip length and duration on the activity and trip similarity are examined using histograms and hexagonal binning diagrams. The proposed methodology is implemented for two-day smart card data in Brisbane, Australia.

Rest of the paper is structured as follows. First, the existing literature of the trip and activity similarity measures and patterns is discussed. Then, similarity measures for the trip and activity of the passengers are defined in the Methodology section. Next, Results section introduces the case study and examines the correlation between the trip and activity similarity of the passengers. Finally, Conclusion section summarizes the findings and proposes the future directions.

2. Literature review

Discovering patterns of travel behavior in the public transit network has attracted attentions from the transport researchers during the recent years. Hasan et al. (2013), concentrating on 626 individuals travel records in a metro network over a three-month period, investigated the patterns in choosing their destinations. This was accomplished based on the concept of POI, utilizing a hazard function and clustering, to define the probability of staying at a place.

They found out that most people do not select the location of their non-home and non-work destinations randomly. Ma et al. (2013) determined transit passenger regularity by clustering passengers based on the location of boarding stops and then divided clusters according to the time interval of boarding transactions. They used one week's worth of AFC transactions from Beijing and compared the efficiency of three clustering algorithms (K++, C 4.5, KNN). They also showed that the regularity of a transit passenger would be a significant factor for transit market analysis. Nishiuchi et al. (2013), similar to Ma et al. (2013), studied passenger regularity based on spatial and temporal patterns using more than 500,000 transactions for 32,000 users during one month in Osaka. They also discovered a relation between the spatial and temporal patterns by drawing the frequency of use in an XY scatter plot. Finally, they found out that the trip patterns for different passenger types vary.

Kusakabe and Asakura (2016) suggested fusing the (Household Travel Survey) HTS and AFC data in order to detect trip purpose. They utilized Naïve Bayes classifier according to the jointly available attributes between the two datasets. The case study was limited to the AFC transactions in one metro station in Osaka over a 20-month period. They successfully estimated the trip purposes in 86.2 per cent of cases. Kieu et al. (2014) studied spatial and temporal aspects of travel patterns. Firstly, trips were clustered regarding the location of alighting stops, then identified groups were divided based on the location of boarding stops, and, next, according to times of the boarding transactions. They used a DBSCAN algorithm for clustering the AFC data in Brisbane for over four months. They determined 72 per cent of the studied passengers have an irregular pattern. Cats et al. (2015) identified the urban activity centers by clustering the passenger travel flows according to the spatial proximity between the stops and centers. They also considered temporal variability as a factor to divide the identified spatial clusters into the smaller ones.

Langlois et al. (2016) dealt with a big AFC dataset for four weeks in London. They investigated the heterogeneities among passengers and their longitudinal activity sequences by defining the concept of 'user areas'. They defined this concept for the stops and the stations visited by users. They were also able to reduce the dimensionality of the data using the principal component analysis (PCA) method. They detected 11 clusters of the passengers associated with a distinct sequence of user areas structures. Sun and Axhausen (2016) decomposed AFC data using a probabilistic tensor factorization model in order to investigate the interactions between the time of day, passenger type, and origin and destination zones. Han and Sohn (2016) studied the probability of choosing the next activity given that the current activity is known. This considered the activity start time, activity duration and land use around the stops. They found eight clusters of passengers with different sequences of activities. They developed a continuous hidden Markov model. More recently, Ghaemi et al. (2017) presented a new representation of the smart card dataset. This provided a visual guide to better understand temporal patterns. Seventeen clusters were identified in terms of a single trip, regular users, late commuters, long day, midday, and active and inactive groups as the temporal behavior of users by an agglomerative hierarchical clustering method. Briand et al. (2017) clustered passengers based on their temporal activities using a Gaussian mixture generative-based model. They discovered that, in addition to common temporal patterns corresponding to the regular home to work/study commute patterns, some groups make their trips earlier or are more regular than others. They also found other passengers have more diffuse activities. Faroqi et al. (2017) examined the correlation between the spatial similarity and temporal similarity of trips in the public transit network, which revealed nonlinear relations between these dimensions of the trip similarity.

The existing literature focused on different activity and trip similarity measures in order to discover activity and trip patterns in the public transit network. However, none of the studies in the literature (to the best of our knowledge) investigated the correlation between the activity similarity and trip similarity of the passengers. Hence, we examine the correlation between the activity similarity and trip similarity to move the literature of passenger behavior in the public transit network one step forward.

3. Methodology

This paper aims to investigate the correlation between activity similarity and trip similarity of public transit passengers, which are extracted from the smart card dataset including boarding and alighting transactions. Figure 1 presents steps of measuring the similarity between the activities and trips.

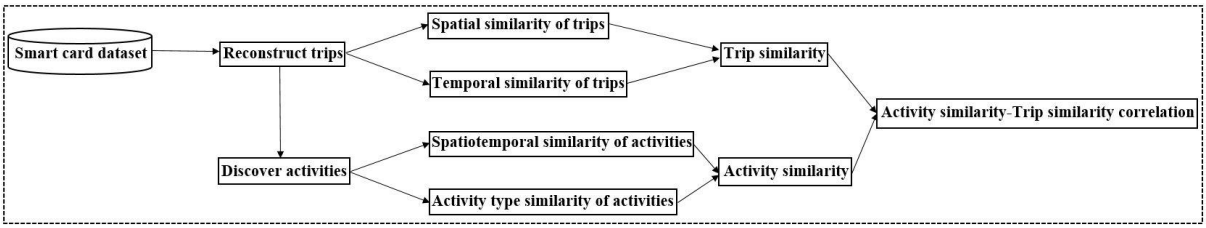

Fig. 1. Methodology

The cleaning process of the smart card datasets is adopted from Robinson et al. (2014). A trip leg happens between subsequent 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 trip legs. Various thresholds are examined for the time gap. Based on the analyses of Alsger et al. (2016), the time gap is considered as 30 minutes in this study. If the time gap between two subsequent trip legs is less than 30 minutes, then the trip legs will link together as a trip; otherwise, an activity happens between the trip legs.

3.1. Trip similarity measure

Among different available trip similarity measures in the literature, the measure of Faroqi et al. (2017) is adopted because it considers both temporal and spatial dimensions; the measures are developed specifically for smart card datasets that include boarding and alighting transactions (not like GPS trajectories that include measurements every few meters). The trip similarity is measured between the trips of two passengers by two parallel steps that are presented in Figure 2. The spatial similarity measure considers direction as well as the distance between the start and end point of trips. For two trips with close enough directions, start point, and end points, the value of spatial similarity is calculated as the ratio of the shorter trip length to the longer one. Also, the temporal similarity measure considers both boarding and alighting times in a continuous linear space. The temporal similarity value between two trips is measured as the ratio of the overlapped trip time to the longer trip time. Finally, the trip similarity matrix is calculated as the pairwise product of the spatial and temporal similarity matrices (Faroqi et al., 2017).

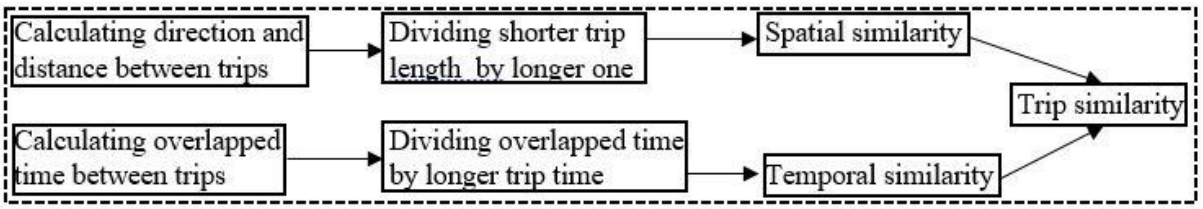

Fig. 2. Trip similarity

In brief, two trips are considered as spatially similar if the distance between the origins (destinations) is less than a threshold (A in Equation 1) and the angle between the two trips is less than a threshold (B in Equation 1). Equation 1 presents the spatial similarity measure between trips (T1, T2) that are between (O1, D1) and (O2, D2); where 'O' stands for origin and 'D' for destination; 'T' stands for trips; 'd(P1, P2)' is the distance function that measures Euclidean distance between two points; 'l(O, D)' is the length function that measures length of the trips; 'di(T1, T2)' is the direction function that measures angle between two trips; and 'SS(T1, T2)' is the spatial similarity value between two trips (Faroqi et al., 2017).

$$SS(T_1, T_2) = \begin{cases} \frac{\min(l(T_1,T_2))}{\max(l(T_1,T_2))}, & (d(O_1,O_2) \le A \mid d(D_1,D_2) \le A) \wedge (di(T_1,T_2) \le B) \\ 0, & otherwise \end{cases} \quad (1)$$

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, the spatial similarity between passenger A and 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 (Faroqi et al., 2017).

The temporal similarity should measure the joint period that two passengers use the public transit. The proposed temporal metric considers time as a linear continuous element of the movement; the first boarding transaction time and last alighting transaction time model a trip; 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 (T1, T2) that respectively are between (B1, A1) and (B2, A2); where 'B' stands for boarding time and 'A' for alighting time; 'TS (T1, T2)' stands for the temporal similarity value. The temporal similarity value between two trips is assumed as the ratio of overlapped trip time to the longer trip time of (T1, T2) (Faroqi et al., 2017).

$$TS\ (T_1, T_2)\ = \begin{cases} \frac{\min(A_1, A_2)\ -\max(B_1, B_2)}{\max((A_1\ -B_1), (A_2\ -B_2))}, & (B_1\ \geq B_2\ \wedge\ A_1 \leq A_2)\ |\ (B_2\ \geq B_1\ \wedge\ A_2 \leq A_1) \\ 0, \quad otherwise \end{cases} \tag{2}$$

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. 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 (Faroqi et al., 2017).

3.2. Activity similarity measure

Among different available activity similarity measures in the literature, the measure of Faroqi et al. (2018a) is adopted because it is a passenger-based similarity measure and simultaneously consider all three main elements (location, time, and type) of the activity; also, this measure is specifically designed for the smart card datasets; it should be noted that this activity similarity measure is validated using HTS dataset by Faroqi et al. (2018a). The activity similarity measure must be able to consider all three activity features of type, time, and location. Calculating this measure consists of two main parallel steps that are shown in Figure 3. The first step is measuring the similarity between location and time of the activities using the Space-Time-Prism (STP), and the second step is measuring the similarity of activity type. The STP determines the possible time and location span for an activity in a 3 dimensional space including two dimensions for locations (XY) and time (T) and measures the similarity by calculating the shared volume between the prisms. The activity type is based on an inferred trip purpose (Alsger et al., 2018), considering temporal (start time and duration of the activity) and spatial (land-use) attributes, and measures the similarity by calculating the time overlap of two trip identical activity types. Finally, the activity similarity matrix is calculated as the pairwise product of the STP and activity type similarity matrices (Faroqi et al., 2018a).

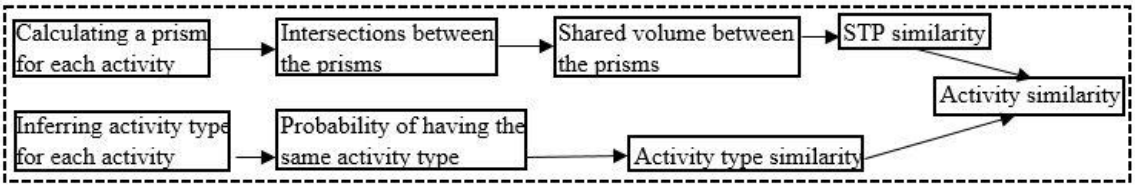

Fig. 3. Activity similarity

The main factors in forming the STP are the walking speed and the maximum walking distance. These factors can identify all locations and times that a passenger can go after alighting from the public transit vehicle and before boarding it again. A passenger can also go in different directions from the direction between the two stops. The passenger can walk in the opposite direction of the two stops for doing an activity and then walk to the next boarding stop. Therefore, the maximum allowable walking distance is related to the distance between the two alighting and next boarding stops. The STP is formulated in Equation 3 to 7 in which a passenger alights at (Xa, Ta) and then boards at (Xb, Tb), X represents the location and T represents time, Aab is the maximum walking time, S is the walking pace, and U stands for the upper line (which is equal to twice as maximum of allowable walking distance) (Faroqi et al., 2018a).

$$Future\ disc: \{X| \left||X - Xa|\right| \leq (T - Ta) * S\} \tag{3}$$

$$Past\ disc: \{X| \left||Xb - X|\right| \leq (Tb - T) * S\} \tag{4}$$

$$Potential\ path\ cylinder: \{X| \left||X - Xa|\right| + \left||Xb - X|\right| \leq (Tb - Ta - Aab) * S\} \tag{5}$$

$$Maximum\ path\ area = \{X| \left||X - Xa|\right| + \left||Xb - X|\right| - \left||Xb - Xa|\right| \leq U\} \tag{6}$$

$$STP: Future\ disc \cap Past\ disc \cap Potential\ path\ cylinder \cap Maximum\ path\ cylinder \tag{7}$$

The future disc includes all possible places that a moving object can go from the first point. The past disc consists of all points that can provide access to the second point. The future and past discs are similar to cones (Miller, 2005). The potential path cylinder inscribes all possible points at which an activity can occur at them. The potential path cylinder enlarges as the time gap between Tb and Ta increases; therefore, another surface should be considered to confine the maximum path cylinder. The maximum path cylinder indicates the border of locations, where a passenger can do his activity (Faroqi et al., 2018a).

The similarity between two STPs is defined as the shared volume between the two STPs divided by union volume of the STPs. The Jaccard index is used to measure the similarity between the STPs; it is a statistic measure for comparing the similarity and diversity of sets and is the size of the intersection divided by the size of the union of the sets (Jiang et al., 2015; Goodall, 1966). Equation 8 presents the Jaccard index value for two STPs A and B, where V stands for the volume function. For instance, if the STP of passenger A is fully inscribed within the STP of passenger B and the STP of passenger A is 50% in size of the STP of passenger B, the final similarity will be 0.5. Also, if the STP of passenger C and the STP of passenger D are next to each other and have a 50% intersection in size, the probability is 0.33. In other words, the similarity that passengers A and B do their activity in the same space-time is 50% and the same similarity for passengers C and D is 33% (Faroqi et al., 2018a).

$$Jaccard\ (A, B) = \frac{V\ (A \cap B)}{V(A) + V(B) - V(A \cap B)} \tag{8}$$

Table 1 shows the probability of the trip purpose according to the start time and activity duration, as a percentage of the total, which are derived from HTS dataset in Brisbane over a period from 2009 through 2012. These distributions are used as temporal attributes (both start time and duration) to infer the probability of different trip purposes. Five activity types are considered: work, education, shopping, recreational, and home. The temporal attributes and trip purpose inference procedure are validated using HTS by Alsger et al. (2018), in which using the HTS data of South East Queensland, origin-destination trips are categorized based on the distribution of start time and activity duration for different trip purposes. The temporal attributes, namely start time and activity duration, are extracted to identify and also confirm each trip purpose, using the HTS data of South East Queensland (SEQ). On this basis, O-D trips are categorized based on the distribution of start time and activity duration for different trip purposes. Then, the percentage of a trip for each purpose is calculated with respect to the total number of trips. It should be noted the trip purposes such as social, pick up/drop off or personal business are removed from the table.

Table 1. The probability of a trip purpose according to the start time and activity duration (Alsger et al., 2018)

Start Time		Work trips			Education trips			Shopping trips			Recreational trips			Home trips		
		Before 9 am	9 am - 3 pm	After 3 pm	Before 9 am	9 am - 3 pm	After 3 pm	Before 9 am	9 am - 3 pm	After 3 pm	Before 9 am	9 am - 3 pm	After 3 pm	Before 9 am	9 am - 3 pm	After 3 pm
Activity Duration (hour)	1-2	2.14	4.73	1.92	0.07	0.16	0.01	1.54	8.41	3.45	0.77	0.87	1.17	1.26	3.43	3.16
	3-4	1.18	1.50	0.42	0.1	0.03	0.02	0.15	0.70	0.03	0.23	0.47	0.28	0.49	2.19	0.86
	5-6	1.17	1.02	0.23	0.12	0.05	0.04	0.04	0.08	0.00	0.11	0.13	0.02	0.47	1.37	0.11
	7-8	1.84	0.51	0.05	4.13	0.31	0.00	0.01	0.01	0.00	0.08	0.05	0.01	0.39	0.35	0.01
	9-10	5.85	0.31	0.04	0.42	0.02	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.11	0.11	0.01
	11-12	1.26	0.07	0.00	0.09	0.06	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.04	0.01	0.00
	> 12	0.25	0.02	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00

The land use around the next boarding stop, as well as the land use around the alighting stop, is considered as the potential land use areas (around alighting and next boarding stops) to do the activity. Spatial and temporal rules are applied using the decision tree to infer the probability for different activity types. The highest probability among activity types is then considered as the trip purpose (Lee and Hickman, 2014; Alsger et al., 2018). For an instance, assume an activity starts at 8 am and takes 7 hours in an area with available land use of 60% work and 40% education (no shopping, recreational, nor residential); the spatial attribute (land use) indicates 60% chance for work and 40% chance for education, however, the temporal attributes (start time and duration) indicate 31% chance for work (1.84/(4.13+1.84)) and 69% chance for education (4.13/(4.13+1.84)); considering both spatial and temporal attributes, the probability of having work activity is 41% ((60% * 31%) / ((60% * 31%) + (40% * 69%))) and education activity is 59% ((40% * 69%) / ((60% * 31%) + (40% * 69%))); hence, the inferred activity type is considered as education. This method is calibrated and validated in Alsger et al. (2018).

If the activity types of two passengers are the same and the corresponding STPs do intersect, the passengers have some similarity in their activities. If the activity types are the same, the product of the activity type probabilities of two passengers is considered as their activity type similarity For instance, if passenger A has an activity type with probability of a, passenger B has the same activity type with probability of b, and the STPs are intersected, then the activity type similarity for these activities between passengers A and B is a*b. Also, if two passengers have more than one similar activity, then final activity type similarity is calculated as the sum of the activity type similarities divided by the number of intersected STPs between the two passengers. At the last step, activity similarity percentage for each pair of passengers is calculated as the pairwise product of the activity type similarity and the STP similarity values. Consequently, the activity similarity matrix is produced. This can be used to infer closeness and relations of the passengers or to cluster those (Faroqi et al., 2018a).

3.3. Correlation measure

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 the zero mention of no linear correlation and values close to +1 or -1 imply a perfect linear correlation (Sedgwick, 2012). Equation 9 presents the Pearson correlation coefficient for two vectors x and y, each of which consists of n elements; $\bar{}$x and $\bar{}$y stand for mean of x and y. In addition, a visualization technique called hexagonal binning is used to discover the nonlinear correlation between the activity similarity and trip similarity values. Conventional methods such as scatter plots cannot efficiently visualize large datasets because many of data are overlapped. The technique pairwise plot the similarity values (Lewin-Koh, 2011).

$$Pearson\ correlation\ = \frac{\Sigma_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i-\overline{y})^2}} \quad (9)$$

4. Results

The used smart card dataset is from TransLink, the public transport authority of South East Queensland (SEQ), Australia. The dataset is for two weekdays at the SEQ that include bus, train, and ferry modes. Wednesday and Thursday (20 and 21 March 2013) are chosen as the weather on those days were normal, and there were no special events. 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 20% of the whole number of transactions. Considering the analyses from Alsger et al. (2017), the sample size can appropriately represent the whole dataset. 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. Figure 4 shows the map for Brisbane, in which the City Business District (CBD) area is highlighted with a yellow circle. Also, some of the major train and bus lines representing main corridors in Brisbane are presented in the map.

Fig. 4. Map of Brisbane

Activity and trip similarity matrices are calculated based on the described methodology. Similarity values are ranged between 0 and 1. According to the matrices, on average, 86% of passengers have similar activity with at least one other passenger; and 92% of passengers have similar trips with at least one other passenger. Figure 5 presents histograms that are depicted for the similarity matrices. Histograms show frequency distribution of activity and trip similarity values for Wednesday and Thursday. Generally, smaller values, close to 0, are more frequent in all histograms. The peak for each histogram happens at the range of (0, 0.02). The frequency of the activity similarity values are higher than the trip similarity. In addition, while the number of passenger pairs having the activity and trip similarity are higher for Wednesday than Thursday, range and proportion of similarity values are the same for both days. Therefore, more number of passenger pairs have the activity similarity than the trip similarity, however, the trip similarity values vary in a wider range than the activity similarity values.

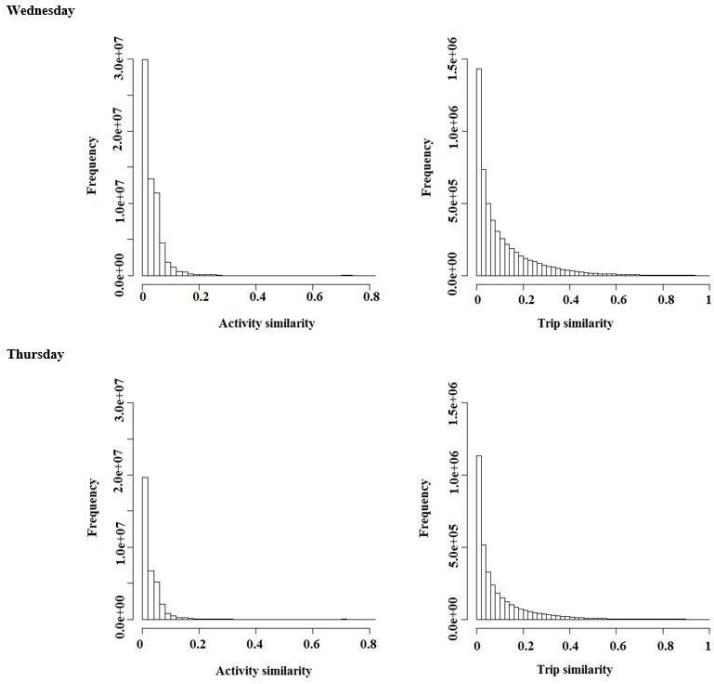

Fig. 5. Histograms for activity and trip similarity matrices

Figure 6 presents cumulative diagrams for the activity similarity at 10 different ranges. Each diagram in Figure 6 shows how each range of the activity similarity cumulated with the trip similarity. For example, the orange graph (that belongs to passengers with the activity similarity between 11 and 20%) includes 60% passengers with the trip similarity less than 10%, and 28% passengers with the trip similarity between 10 and 20%. All in all, most (more than 95%) of the activity similarity values have less than 40% of the trip similarity.

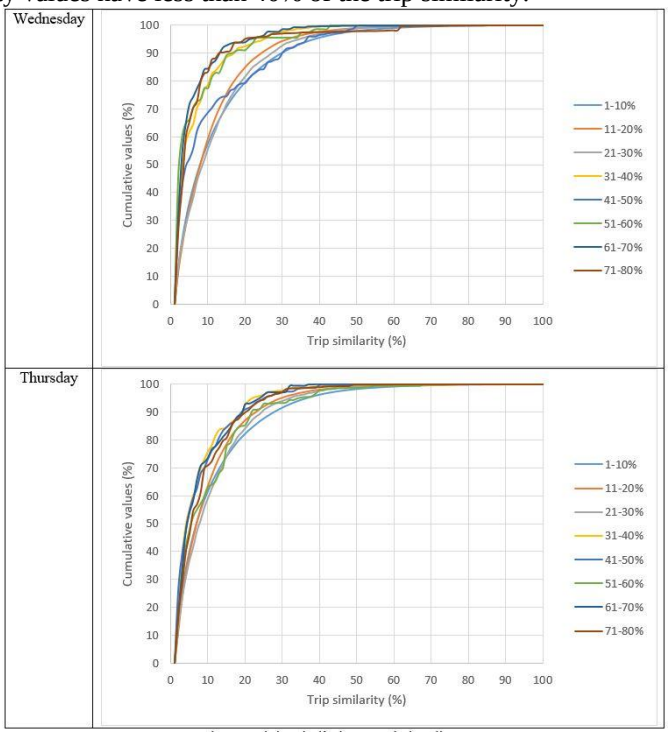

Fig. 6. Activity similarity cumulative diagram

Figure 7 presents cumulative diagrams for the trip similarity at 10 different ranges. Each diagram in Figure 7 shows how each specific range of the trip similarity cumulated with different ranges of the activity similarity. For example, the orange graph (that belongs to passengers with the trip similarity between 11 and 20%) includes 92% passengers with the activity similarity less than 10%, and 7% passengers with the activity similarity between 10 and 20%. Most (more than 95%) of the trip similarity values have less than 10% of the activity similarity. Comparing Figures 6 and 7, the trip similarity ranges for having the activity similarity are more diverse than the activity similarity ranges for having the trip similarity.

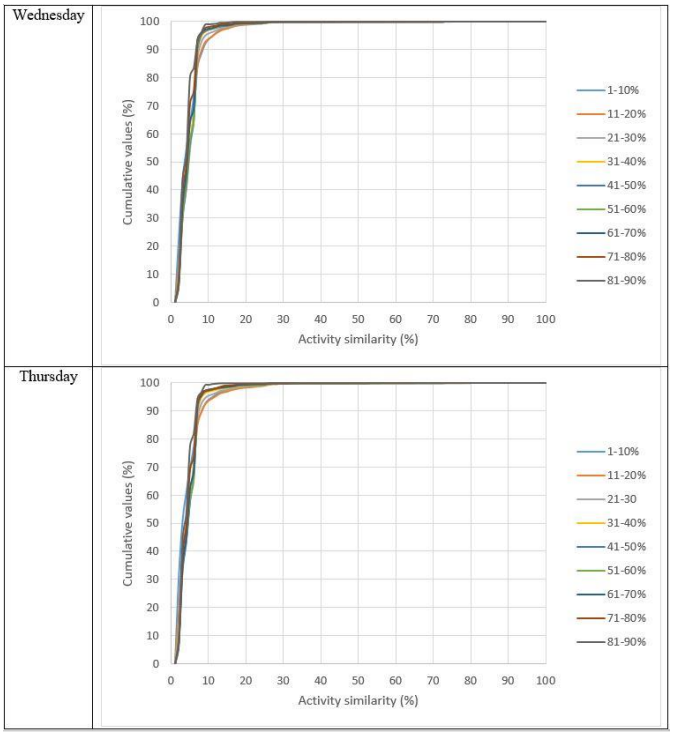

Fig. 7. Trip similarity cumulative diagram

Table 2 presents correlation and the conditional probability of the activity and trip similarity values. The correlation value is positive and close to 0.1 that means the activity and trip similarity values fluctuate positively together but with a weak linear relation. Also, values for the conditional probability for the activity similarity given the trip similarity is close to 50% that means the probability for having the activity similarity between two passengers when they have the trip similarity, is 50 %. In addition, the conditional probability for the trip similarity given the activity similarity is close to 5% that means the probability for having trip similarity between two passengers when they have activity similarity, is 5%. In other words, it is more likely to have the activity similarity when there is a trip similarity than having the trip similarity when there is an activity similarity.

Table 2. Correlation and conditional probability

	Wednesday	Thursday
Correlation (AS, TS)	0.07125783	0.0676873
P(AS|TS)	59%	48%
P(TS|AS)	4.9%	4.5%

Following diagrams, in Figure 8, present values of the correlation between the activity similarity and trip similarity at different ranges. Values of the trip similarity are presented at the left side diagram, and values of the activity similarity are presented at the right side diagram. According to the left diagram, the highest positive correlation happens at the range of (0, 0.1) of the trip similarity; range of the (0.2, 0.3) is like a boundary between positive and negative correlation values; the negative peak happens at the range of (0.9, 1). Also, according to the right side diagram, the positive peak happens at the range of (0, 0.1) of the activity similarity; range of the (0.2, 0.3) of the activity similarity works as a boundary between positive and negative correlation values; the negative peak is placed at range (0.9, 1) of the activity similarity. For Figures 8 and 9, “AS” stands for the Activity Similarity, and “TS” stands

for the Trip Similarity. The reason for declining correlation between AS and TS by increasing AS and TS values can be fewer number of passengers with high AS or TS compare to passengers with low AS or TS, which means there are more activity or trip diversity for fewer number of passengers.

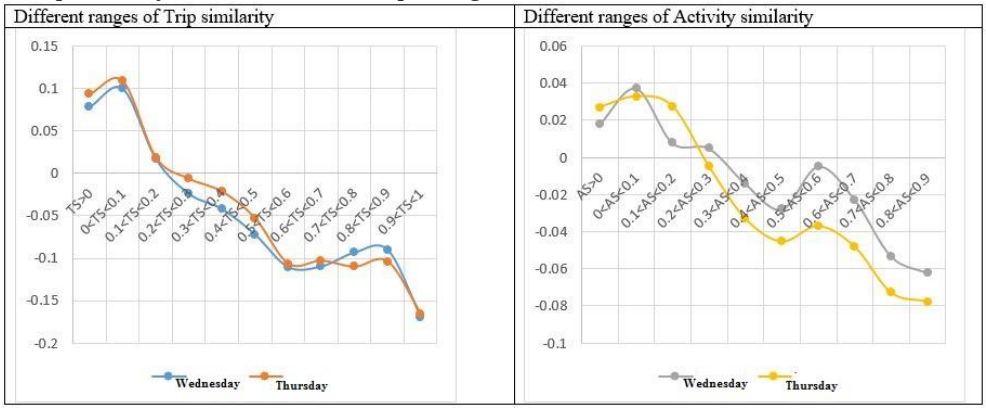

Fig. 8. Correlation between the activity similarity and trip similarity

Following diagrams, in Figure 9, present the conditional probability between the activity similarity and trip similarity at different ranges. The diagram at the left side presents P (AS|TS) and on the right side presents P (TS|AS). According to the left side diagram, the highest conditional probability for P(AS|TS), which is 66%, happens at the range of (0.2, 0.3) of the trip similarity values. After the range of (0.2, 0.3), probability of having AS decreases by increasing TS; which was seen in declining correlation between AS and TS in Figure 8. According to the right side diagram, the highest conditional probability for P(TS|AS), which is 9%, happens at the range of (0.7, 0.8) of the activity similarity values. Probability of having AS given TS is much higher than the probability of having TS given AS. Also, the probability of having TS given AS is more fluctuating than the probability of having AS given TS.

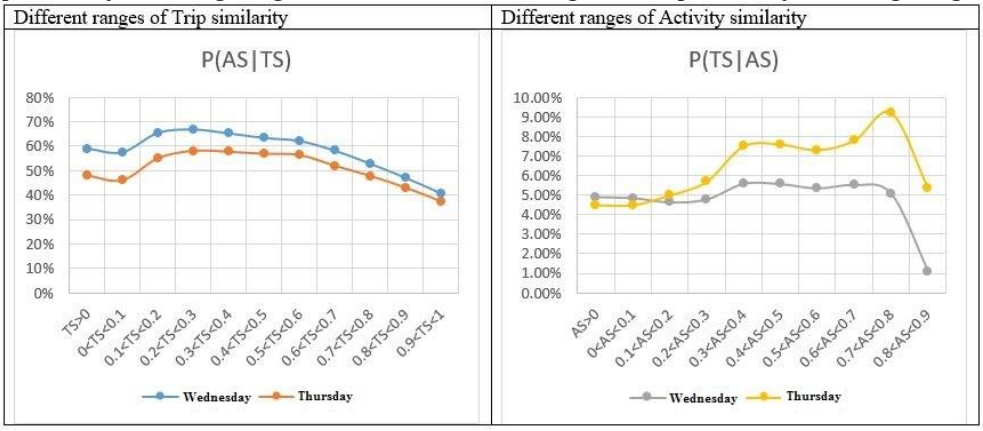

Fig. 9. Conditional probability between the activity similarity and trip similarity

Figure 10 shows the hexagonal binning diagrams for the activity similarity and trip similarity values. Color and size of each hexagonal present the number of available passenger pairs for specific values of the activity similarity and trip similarity. If a horizontal (vertical) line cut the diagrams, it is clear that probability of having the activity similarity (trip similarity) decreases for larger values of the trip similarity (activity similarity). Also, results of the correlation and conditional probability (from Figures 8 and 9) can visually be seen in these diagrams. For instance, if the rectangular of (0.8, 1) trip similarity and (0, 0.2) activity similarity is compared with the rectangular of (0, 0.2) trip similarity and (0.8, 1) activity similarity, more number of activity similarity hexagonal is seen in the former rectangular than number of trip similarity hexagonal in the latter rectangular. In other words, there are more passengers with high trip similarity and low trip similarity than passengers with low trip similarity and high activity similarity.

In addition, negative correlation between the activity similarity and trip similarity in Figure 8 can be observed in Figure 10 as well; one increases while the other one decreases. Figure 10 shows that for high values of the activity similarity (trip similarity) number of the high trip similarity (activity similarity) values are decreasing. For example,

two commuters travel on the same bus from a suburb to CBD in the morning peak, but it does not certainly mean they perform their job at the same place or time period.

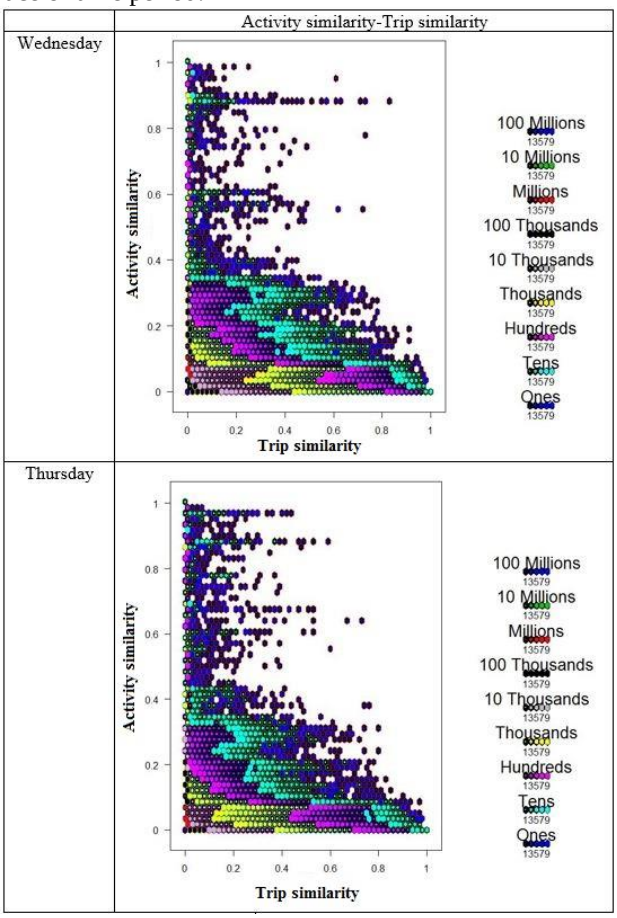

Fig. 10. Hexagonal binning diagrams for the activity similarity-trip similarity

Figure 11 presents the histograms for the trip duration and trip length. The peak for the trip duration histogram happens at the range of (30, 40) min and for the trip, length happens at the range of (6, 8) km. The trip duration histograms are ascending before the peak and descending after the peak. The trip length histograms are ascending before the peak and locally descending after the peak. It might be expected that trip duration and trip length histograms have a linear relation, but the nonlinear relation between the trip duration and trip length histograms relates to the geometry of the network and waiting time at stops or transfer trip legs.

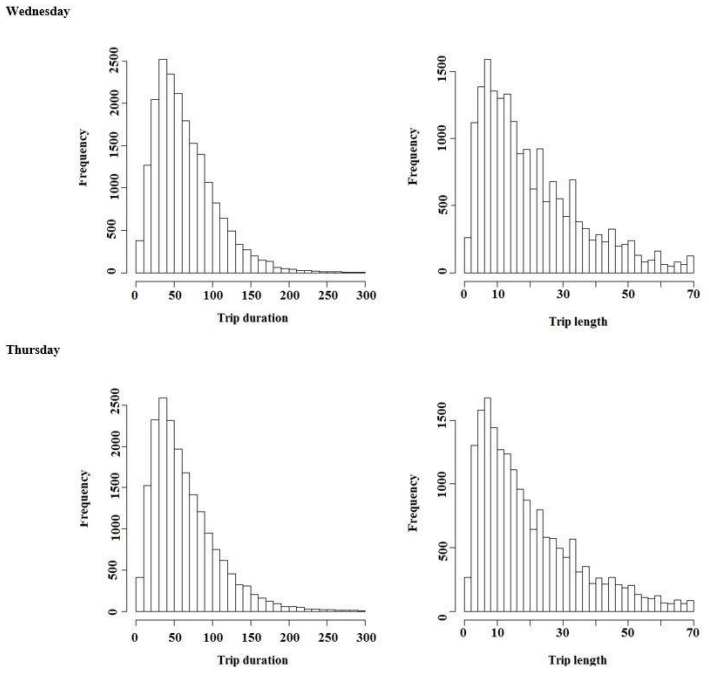

Fig. 11. Histograms for trip duration and length

Figure 12 demonstrates hexagonal diagrams for both activity similarity-trip duration and activity similarity-trip length. Hexagonal diagrams present numbers of pairs of activity similarity and trip duration or length with specific values. In trip duration diagrams, the range of (70, 90) min plays an important role in changing the trend of the diagrams. Also, in trip length diagrams, the range of (18, 22) km plays an important role in changing the trend of the diagrams. In all diagrams, if a horizontal line cut the diagram, probability of having the activity similarity decreases by increasing the trip duration or length before the mentioned ranges. In other words, passengers with shorter trip duration or length (before the mentioned ranges) are more likely to have higher values of the activity similarity than passengers with longer trip duration or length (after the mentioned ranges).

For example, passengers with 50 minute trip duration are more likely to have the activity similarity rather than passengers with 200 minutes of trip duration. Likewise, passengers with 10km trip length are more likely to have activity similarity than passengers with 40km trip length. In addition, if a vertical line cut the diagrams, at any point, the probability of having smaller values of the activity similarity is higher than larger values. For instance, the probability of having the activity similarity of 0.6 is less than the probability of having the activity similarity of 0.2 at the trip duration of 50 minute or trip length of 10 kilometers.

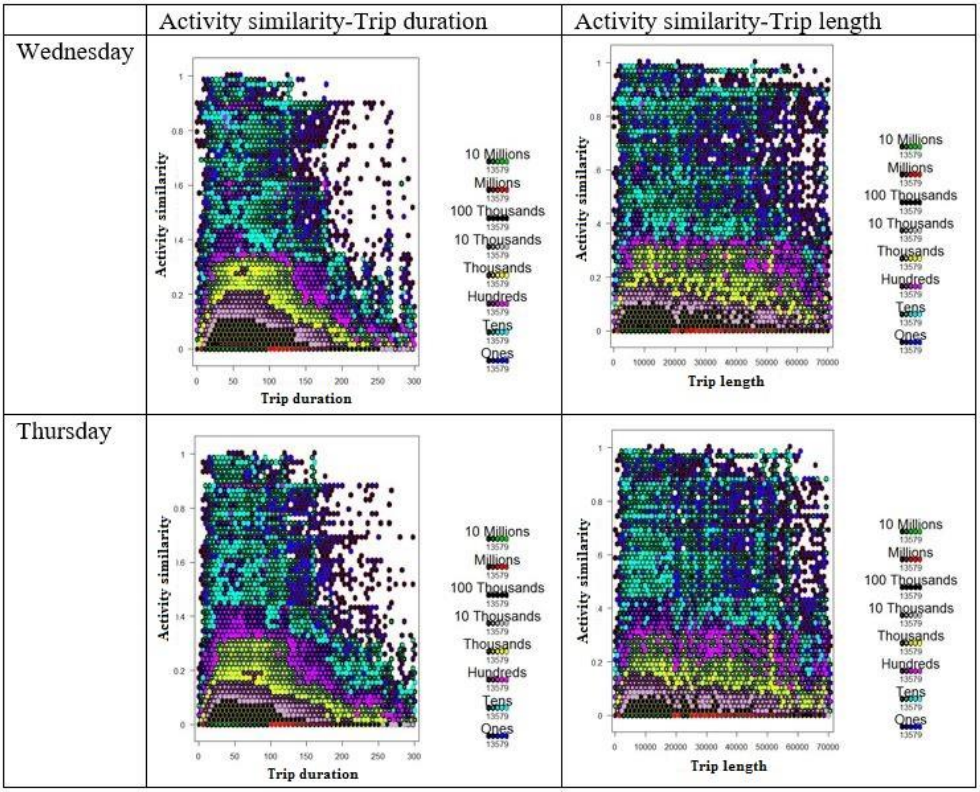

Fig. 12. Hexagonal binning diagrams for the activity similarity-trip duration and length

Figure 13 demonstrates hexagonal diagrams for both trip similarity-trip duration and trip similarity-trip length. Similar to the activity similarity hexagonal diagrams, in trip similarity-trip duration diagrams, the range of (70, 90) min plays an important role in changing the trend of the diagrams. Also, in trip similarity-trip length diagrams, the range of (18, 22) km plays an important role in changing the trend of the diagrams. In all diagrams, if a horizontal line cut the diagram, probability of having the trip similarity decreases by increasing the trip duration or length before the mentioned ranges. In other words, passengers with shorter trip duration or length (before the mentioned ranges) are more likely to have higher values of trip similarity than passengers with longer trip duration or length (after the mentioned ranges). For example, passengers with 50 minute trip duration are more likely to have trip similarity rather than passengers with 200 minutes of trip duration. Likewise, passengers with 10km trip length are more likely to have trip similarity than passengers with 40km trip length. In addition, if a vertical line cut the diagrams, at any point, the probability of having smaller values of the trip similarity is higher than larger values. For instance, the probability of having the trip similarity of 0.6 is less than the probability of having the trip similarity of 0.2 at the trip duration of 50 minute or trip length of 10 kilometer.

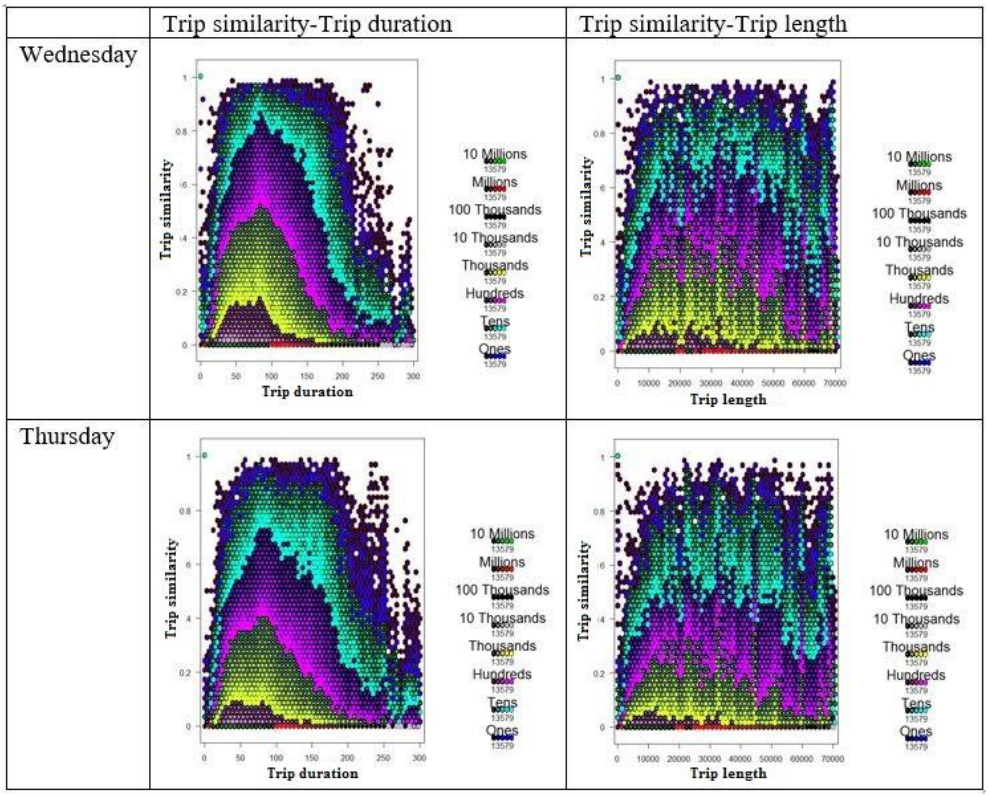

Fig. 13. Hexagonal binning diagrams for the trip similarity-trip duration and length

All the above analyses address two main questions. The research questions are: how similar are activities of passengers if they have similar trips (or vice versa)? And how trip length and duration impact the similarity of the activity and trip of passengers? Effects of having the activity similarity on the trip similarity (or vice versa) are investigated through cumulative diagrams, Pearson correlation coefficient, conditional probability, and hexagonal binning diagrams. Also, relations between the trip length and duration with the activity and trip similarity are investigated by hexagonal binning diagrams.

As results of the analyses, the correlation between the activity similarity and trip similarity of the passengers is not linear; for example, some ranges of the activity and trip similarity have a positive correlation, while there are some ranges with negative correlation values. Also, the probability of having the activity similarity given the trip similarity is much higher than the probability of having the trip similarity given the activity similarity. Moreover, the trip length and duration of some specific ranges have a positive correlation with the activity and trip similarity between the passengers. Also, there are some obvious boundaries in the trip length and duration, which change the probability of having the activity or trip similarity. The discovered ranges and boundaries of the activity similarity, trip similarity, trip length, and trip duration imply that the correlation between the activity similarity and trip similarity needs more investigation. Therefore, other parameters of the network such as geometry or accessibility should be considered in further investigation of the correlation between the activity and trip similarity of the public transit passengers.

5. Conclusions

The paper investigates the correlation between the activity similarity and trip similarity of public transit passengers. Activity similarity is measured considering time, location and type of the activity. Also, the Trip similarity is measured considering the spatial and temporal dimensions of passenger trips. The relation between the activity and trip of the passengers is examined using the histograms, Pearson correlation coefficient, conditional probabilities, and hexagonal binning technique. Furthermore, influences of the trip length and duration on the activity and trip similarity of

passengers are investigated using the histograms and hexagonal binning diagrams. The case study is two-day smart card data in Brisbane, Australia.

Results show a nonlinear correlation between the activity and trip similarity with the trip length and duration. 86% of passengers have similar activity with at least one other passenger. 92% of passengers have similar trips with at least one other passenger. While the activity similarity and trip similarity has a weak linear correlation, the probability of having the activity similarity given the trip similarity is around 50%. Furthermore, similar trends and similarity correlation for Wednesday and Thursday validate the findings. Trips with shorter duration or length are more likely to be similar. Also, activities that happen at the end of the trips with shorter duration or length are more likely to be similar. More specifically, trips with duration less than 70 minute or length less than 20 kilometers are more likely to be similar and end up in similar activities.

The findings of this paper can improve the travel behavior knowledge in the public transit network, which is necessary to design and upgrade the network. Also, group based applications such as group discount or DRT services can be developed more efficiently based on the found relation between the activity similarity and trip similarity. For example, group discounts can be designed in specific events in a way that encourage more passengers to use the public transit network. Also, TOD models that design and understand effects of the urban facilities around the transit network, can be improved by investigating the relation between trips and activities of public transit passenger.

Future directions can be led in two directions. First, enriching the smart card dataset with other mobility datasets such as mobile devices or household travel surveys in order to investigate effects of demographic attributes of the passengers on the activity and trip similarity. Second, comparing the findings from different case studies to examine the effects of urban structures on the activity and trip similarity correlation.

References

Alsger, A. 2017. Estimation of transit origin destination matrix using smartcard fare data (Doctoral thesis, University of Queensland, Queensland, Australia).

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

Alsger, A., Tavassoli, A., Mesbah, M., & Ferreira, L. 2017. Evaluation of Effects from Sample-Size Origin-Destination Estimation Using Smart Card Fare Data. Journal of Transportation Engineering, Part A: Systems, 04017003.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., and Hickman, M. 2018. Public transport trip purpose inference using smart card fare data, Transportation Research Part C: Emerging Technologies, 87,123-137.Bobadilla, J., Ortega, F., Hernando, A. & Gutiierrez, A. 2013. Recommender systems survey. Knowledge-Based Systems 46, 109–132.

Bohannon, R. W. 1997. Comfortable and maximum walking speed of adults aged 20—79 years: reference values and determinants. Age and ageing, 26(1), 15-19.

Briand, A. S., Côme, E., Trépanier, M., & Oukhellou, L. 2017. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transportation Research Part C: Emerging Technologies, 79, 274-289.

Cats, O., Wang, Q., and Zhao, Y. 2015. The identification and classification of urban centres using public transport passenger flows data. Journal of Transport Geography, 48, 10-22.

Cunningham, D. A., Rechnitzer, P. A., Pearce, M. E., & Donner, A. P. 1982. Determinants of self-selected walking pace across ages 19 to 66. Journal of Gerontology, 37(5), 560-564.

Daniels, R., & Mulley, C. 2013. Explaining walking distance to public transport: The dominance of public transport supply. Journal of Transport and Land use, 6(2), 5-20.

Faroqi, H., Mesbah, M., & Kim, J. 2017. Spatial-temporal similarity correlation between public transit passengers using smart card data. Journal of Advanced Transportation.

Faroqi, H., Mesbah, M., & Kim, J. 2018a. A model for measuring activity similarity between public transit passengers using smart card data. Journal of Travel Behaviour & Society.

Faroqi, H., Mesbah, M., & Kim, J. 2018b. Applications of transit smart cards beyond a fare collection tool: A literature review. Advances in Transportation Studies.

Ghaemi, M. S., Agard, B., Trépanier, M., & Partovi Nia, V. 2017. A visual segmentation method for temporal smart card data. Transportmetrica A: Transport Science, 13(5), 381-404.

Han, G., and Sohn, K. 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. Transportation Research Part B: Methodological, 83, 121-135.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., and González, M. C. 2013. Spatiotemporal patterns of urban human mobility. Journal of Statistical Physics, 151(1-2), 304-318.

Jiang, F., Thilakarathna, K., Kaafar, M. A., Rosenbaum, F., & Seneviratne, A. 2015. A spatio-temporal analysis of mobile internet traffic in public transportation systems: A view of web browsing from the bus. In Proceedings of the 10th ACM MobiCom Workshop on Challenged Networks (pp. 37-42). ACM.

Kieu, L. M., Bhaskar, A., and Chung, E. 2014. Transit passenger segmentation using travel regularity mined from Smart Card transactions data.

Kusakabe, T., Tsubota, T., and Bhaskar, A. 2016. Validation Study of Naïve Bayes Probabilistic Model for Transit Passengers' Trip Purpose Estimation: Case Study Exploiting Detailed Brisbane Household Travel Survey Data. In Transportation Research Board 95th Annual Meeting (No. 16-2539).

Langlois, G. G., Koutsopoulos, H. N., and Zhao, J. 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transportation Research Part C: Emerging Technologies, 64, 1-16.

Lee, S. G., & Hickman, M. 2014. Trip purpose inference using automated fare collection data. Public Transport, 6(1-2), 1-20.

Lewin-Koh, N. 2011. Hexagon binning: an overview. Online: http:// cran. r-project. org/ web/ packages/ hexbin/ vignettes/ hexagon_ binning. pdf.

Ma, X., Wu, Y. J., Wang, Y., Chen, F., and Liu, J. 2013. Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies, 36, 1-12.

Miller, H. J. 2005. A measurement theory for time geography. Geographical analysis, 37(1), 17-45.

Morency, C., Trepanier, M., and Agard, B. 2007. Measuring transit use variability with smart-card data. Transport Policy, 14(3), 193-203.

Munizaga, M., Devillaine, F., Navarrete, C., and Silva, D. 2014. Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies, 44, 70-79.

Nishiuchi, H., King, J., and Todoroki, T. 2013. Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. International Journal of Intelligent Transportation Systems Research, 11(1), 1-10.

Robinson, S., Narayanan, B., Toh, N., & Pereira, F. 2014. Methods for pre-processing smartcard data to improve data quality. Transportation Research Part C: Emerging Technologies, 49, 43-58.

Sedgwick, P. 2012. Pearson's correlation coefficient. Bmj, 345(7).

Shen, J., & Cheng, T. 2016. A framework for identifying activity groups from individual space-time profiles. International Journal of Geographical Information Science, 30(9), 1785-1805.

Sun, L., and Axhausen, K. W. 2016. Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transportation Research Part B: Methodological, 91, 511-524.