Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 5070794, 16 pages
https://doi.org/10.1155/2019/5070794

Research Article

Comparing Sequential with Combined Spatiotemporal Clustering of Passenger Trips in the Public Transit Network Using Smart Card Data

Hamed Faroqi,[1] **Mahmoud Mesbah**,[1,2] **and Jiwon Kim**[1]

[1]*School of Civil Engineering, The University of Queensland, Australia*
[2]*Department of Civil and Environmental Engineering, Amirkabir University of Technology, Iran*

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

Received 6 December 2018; Revised 10 February 2019; Accepted 20 February 2019; Published 14 April 2019

Academic Editor: Roberta Di Pace

Smart card datasets in the public transit network provide opportunities to analyse the behaviour of passengers as individuals or as groups. Studying passenger behaviour in both spatial and temporal space is important because it helps to find the pattern of mobility in the network. Also, clustering passengers based on their trips regarding both spatial and temporal similarity measures can improve group-based transit services such as Demand-Responsive Transit (DRT). Clustering passengers based on their trips can be carried out by different methods, which are investigated in this paper. This paper sheds light on differences between sequential and combined spatial and temporal clustering alternatives in the public transit network. Firstly, the spatial and temporal similarity measures between passengers are defined. Secondly, the passengers are clustered using a hierarchical agglomerative algorithm by three different methods including sequential two-step spatial-temporal (S-T), sequential two-step temporal-spatial (T-S), and combined one-step spatiotemporal (ST) clustering. Thirdly, the characteristics of the resultant clusters are described and compared using maps, numerical and statistical values, cross correlation techniques, and temporal density plots. Furthermore, some passengers are selected to show how differently the three methods put the passengers in groups. Four days of smart card data comprising 80,000 passengers in Brisbane, Australia, are selected to compare these methods. The analyses show that while the sequential methods (S-T and T-S) discover more diverse spatial and temporal patterns in the network, the ST method entails more robust groups (higher spatial and temporal similarity values inside the groups).

1. Introduction

Automated Fare Collection (AFC) systems have been implemented in the public transit network since two decades ago. These systems not only expedite the process of fare collection but also produce valuable datasets. Smart card datasets create a great opportunity for both researchers and practitioners of the public transit network to improve the status quo. Smart card datasets usually include the location and time of boarding and/or alighting transactions of passenger trips. Contrary to a classic survey that was limited in sampling and may not have reflected the ground truth, the smart card datasets are comprehensive and reliable. The datasets can reconstruct passenger trips, which can help to understand, improve, and evaluate the network performance [1, 2]. Hence, smart card datasets attract attention of both researchers and practitioners who desire to improve the public transit network.

Studies of travel demand in the public transit network focused on understanding how passengers move in the network by modelling both the time and location of their trips [3]. Smart card datasets can help to discover the patterns of travel demand. Data mining techniques have been used to extract travel demand patterns from the smart card datasets [4–6]. In other words, data mining techniques can discover groups of passengers who have a similar travel feature based on similarity measures. For instance, passenger groups with similar travel time or length can be determined. Clustering passengers can develop various group and customer-centric transit services and mobility applications such as DRT

systems [7], friend recommendation systems [8], inferring socioeconomic attributes of passengers [9], level of access in the public transit network [10, 11], and traffic flow prediction models [12–14]. Potential implications for the spatial and temporal clustering methods are discussed later on in the discussion section. Consequently, ascertaining travel demand patterns using data mining techniques is a building block for many novel applications.

The spatial or temporal perspective in passenger clustering can form different groups of the passengers (i.e., groups of passengers with spatially similar (same routes) or temporally similar (same time) trips). Also, a passenger moves simultaneously in both spatial and temporal space. Two passengers can be similar based on the spatial similarity measure but can be dissimilar according to the temporal similarity measure. For instance, two passengers may use the same routes but in different periods of the day. In addition, a passenger can have one or more trips during a day, all of which should be considered in measuring the spatial or temporal similarity with other passengers. Spatial and/or temporal similarity measures between passengers based on their trips can be used as a measure to study the closeness or relationship between the passengers [15]. Therefore, to have a comprehensive insight of travel patterns, both spatial and temporal dimensions of the trips should be considered in the clustering of the passengers.

Spatial and temporal similarity measures should be defined separately because of fundamental differences. Spatial space is a two-dimensional space with units such as meters or inches; however, temporal space is a one-dimensional space with units such as minutes. Hence, defining a unique spatiotemporal similarity measure might be an ambiguous technique because it needs to merge these two different spaces [15]. Moreover, to have passenger clusters with both spatial and temporal similarities, it is necessary to cluster them in two steps (sequentially) or combine values of the spatial and temporal similarities (that are calculated separately) and then cluster them in one step. Also, different priorities in the sequential method (first, spatial clustering and then temporal clustering or vice versa) may reveal different passenger groups. The existing literature of the clustering passenger trips focuses on sequential spatial and then temporal clustering methods, which are examined in the next section. Consequently, passenger clusters with the spatial and temporal similarities can be discovered by different methods, which yield different outcomes.

This paper compares characteristics of passenger groups that are discovered by different methods of the spatial and temporal clustering. This paper for the first time (to the best of our knowledge) shed light on differences between the sequential and combined spatial and temporal clustering alternatives in the public transit network. Firstly, the spatial and temporal similarity measures between the passengers are defined. Secondly, the passengers are clustered using a hierarchical agglomerative algorithm with three different methods, including sequential two-step spatial-temporal (S-T), sequential two-step temporal-spatial (T-S), and combined one-step spatiotemporal (ST). Thirdly, characteristics of the discovered groups are described and compared using maps, numerical and statistical values, the cross correlation technique, and temporal density plots. Finally, some passengers are selected to show how differently passengers are clustered by the above-mentioned methods. Four-day smart card data including 80,000 passengers in Brisbane, Australia, are selected to compare these methods.

The remainder of this paper is structured as follows. Firstly, the existing literature is reviewed. Then, the spatial and temporal similarity measures and clustering algorithm are explained in the methodology section. Next, the case study and results are described in the Results section. Finally, the methodology, findings, and plans are summarized in the conclusion section.

2. Literature Review

Data mining techniques have recently been used to discover the spatial and temporal patterns in the public transit network using the smart card data. Agard et al. [4] carried out the first study using clustering algorithms to discover patterns in the smart card data; however, after 2013, the tendency to use clustering algorithms has been advanced by these datasets. Ma et al. [16] determined transit passenger regularity by clustering passengers based on the location of boarding stops and then dividing clusters according to the time interval of boarding transactions. They used one week of AFC transactions from Beijing and compared the efficiency of three clustering algorithms (K++, C 4.5, KNN). Also, they showed that the regularity of a transit passenger would be a significant factor for transit market analysis. Nishiuchi et al. [17] studied passenger regularity based on spatial and temporal patterns using more than 500,000 transactions for 32,000 users during one month in Osaka, Japan. Tao et al. [18] utilised single day transactions from Brisbane to detect the major travel paths for bus passengers at the stop level using flow-comap techniques for visualising the patterns.

Kieu et al. [5] 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 the DBSCAN algorithm for clustering the AFC data in Brisbane over 4 months. Sun and Axhausen [19] decomposed AFC data using a probabilistic tensor factorisation model to investigate the interactions between time of day, passenger type, and origin and destination zones. Manley et al. [20] analysed variation in regular and irregular travel behaviour to derive a system-wide spatial-temporal understanding of regularity in the travel behaviours. They used the DBSCAN algorithm over 49 weekdays. Also, they investigated regularity over different transit modes and found that bus mode had a higher proportion of regular travellers than others. Yu and He [21] used a 3-step methodology to discover spatial-temporal characteristics of bus travel demand using heat-map technique and the Gaussian Mixture Model (GMM). They used 8-week data from Guang Zhou. The heat-map method visually unveils the spatial-temporal travel demand patterns at a regional level. Ghaemi et al. [22] presented

Table 1: Literature review.

Clustering type	Study	Description
Temporal clustering	Agard et al. [4]	Clustering trips based on boarding time transactions.
	Ghaemi et al. [22]	Clustering passengers based on boarding time transactions.
Spatial clustering	Tao et al. [18]	Clustering trips based on locations of boarding and alighting stops.
Spatial-Temporal clustering	Ma et al. [16]	Clustering trips first based on location of boarding stops, then dividing clusters according to the time interval of boarding transactions.
	Nishiuchi et al. [17]	Clustering and investigating relations between the spatial and temporal patterns of trips.
	Kieu et al. [5]	Clustering trips first based on locations of alighting stops, then based on the location of boarding stops, and thirdly based on times of the boarding transactions.
	Sun and Axhausen [19]	Decomposing data to investigate the interactions between the time of day, passenger type, and origin and destination zones.
	Manley et al. [20]	Investigating spatial and temporal regularity over different transit modes.
	Yu and He [21]	Using heat maps to discover spatial and temporal demand of bus trips.
	Briand et al. [23]	Modelling time in a continuous space to investigate passenger exchanges between clusters over time.

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 single trip, regular users, late commuters, long day, midday, and active and inactive groups as the temporal behaviour of users by an agglomerative hierarchical clustering method. Briand et al. [23] proposed a 2-level generative model that applies the (GMM) to regroup passengers based on their temporal habits. They used 391,783 transactions by 2504 users over 4 years in Gatineau. Also, they modelled time in a continuous space. They found that clusters over time mostly exchange their cards with clusters having similar patterns. Table 1 summarizes the mentioned studies.

Consequently, the existing literature has recently focused on the spatial and temporal patterns of travel behaviour in the public transit network. While initial studies focused more on the temporal patterns of trips [4], more recent studies focused on spatial and temporal patterns of trips [20, 21, 23]; the latter studies discovered the spatial and temporal patterns by the S-T sequential clustering way, in which passengers are first clustered based on the spatial similarity and then each spatial group is clustered based on the temporal similarity [24]. However, no study considered the opposite sequential clustering of first the temporal clustering and then the spatial clustering (T-S). In other words, there is no study that has investigated differences between these two methods of sequential clustering. In addition, no study has tried to cluster passengers in the public transit network using a combined one-step way considering both spatial and temporal measures. These research gaps are addressed in this paper. The scientific contributions of this paper are twofold:

(1) Defining a one-step method for extracting the spatiotemporal patterns in the public transit network.

(2) Comparing the different methods of spatial and temporal clustering of passengers in the public transit network using the smart card dataset.

3. Methodology

This section briefly explains trip reconstruction from the smart card dataset, defining the spatial and temporal similarities, and clustering the passengers by different methods. Figure 1 presents the main steps of the methodology.

The passengers are modelled by their trips during the entire day. The trips are reconstructed from the smart card dataset that includes both boarding and alighting transactions. Each trip is made from one or more trip leg, and each trip leg comprises the space and period between the boarding and alighting transactions. Two trip legs are linked based on the time gap between the first alighting and the next boarding

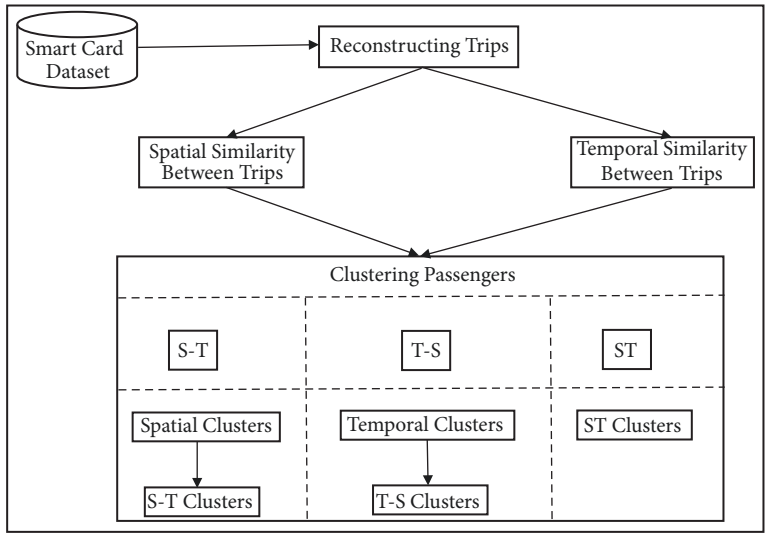


Figure 1: Methodology.

transactions. Various thresholds are examined for this time gap. Based on the analyses of Alsger et al. [25], the time gap is considered as 30 minutes in this study. Therefore, if the time gap between two trip legs is less than 30 minutes, then they will be linked as one trip.

Data mining techniques aim to discover patterns in large datasets. A clustering or unsupervised learning method as a data mining technique assembles sets of objects in the similar groups; it can assemble them in a way that increases the similarity between members of a group and/or increases the dissimilarity between members of different groups. The clustering algorithm initialises based on a similarity measure among the objects, in this case, passengers. Calculating the similarity between any pair of objects in the dataset builds a matrix that works as an input for the clustering algorithm [26].

The trip similarity is examined in two parallel steps of the spatial and temporal similarity measures. Both spatial and temporal similarity measures are adopted from Faroqi et al. [15] where more details can be found. The measures are developed specifically for smart card dataset that include boarding and alighting transactions (not like GPS trajectories that include measurements every few meters). In brief, two trips are considered as spatially similar if the distance between the origins (destinations) is less than a threshold (A in (1)) and the angle between the two trips is less than a threshold (B in (1)). Equations (1) to (3) present the spatial similarity measure and corresponding functions between trips (T1, T2) that are between (O1, D1) and (O2, D2), where "O" stands for origin and "D" for destination; each origin or destination stop is presented by coordination (x, y); "T" stands for trips; "A" stands for the maximum distance between origins or destinations; "B" stands for the maximum angle between trips; "d(O1, O2)" or "d(D1, D2)" is the distance function that measures Euclidean distance between two points; "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. Values of the spatial similarity vary between 0 and 1 [15].

Spatial Similarity Measure between Trips

$$SS(T_1, T_2) = \begin{cases} \dfrac{\min\left(d(O_1, D_1), d(O_2, D_2)\right)}{\max\left(d(O_1, D_1), d(O_2, D_2)\right)}; & \left[\left(d(O_1, O_2) < A \parallel d(D_1, D_2) < A\right) \& \left(di(T_1, T_2) < B\right)\right] \\ 0; & \text{otherwise} \end{cases} \tag{1}$$

Distance Function

$$d(O_i, D_i) = \sqrt{\left(x_{O_i} - x_{D_i}\right)^2 + \left(y_{O_i} - y_{D_i}\right)^2} \tag{2}$$

Direction Function

$$di(T_1, T_2) = \tan^{-1}\frac{m_1 - m_2}{1 + \left(m_1 * m_2\right)}; \quad m_i = \frac{y_{D_i} - y_{O_i}}{x_{D_i} - x_{O_i}} \tag{3}$$

```
a = 0;
for (i in 1:m) {B = Φ;
  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 ϵ B))));}
SS (P_1, P_2) = [min (a_12, a_21)] / [max (Σ_{i=1}^{m} l (T_i), Σ_{j=1}^{n} l (T_j)];
```

Algorithm 1: Pseudocode for the 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 + OT (T_i, T_j);}
                }
               }
TS (P_1, P_2) = a / [max (Σ_{i=1}^{m} TT_i, Σ_{j=1}^{n} TT_j];
```

Algorithm 2: Pseudocode for temporal similarity between passengers.

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 $(3/(3+6))*100 = 33$. Algorithm 1 presents the pseudocode for the spatial similarity between two passengers (P1, P2) who, respectively, have m and n unique trips, where "p" stands for the passenger, "a" is defined for measuring sum of the lengths of shorter similar trips, "a12" is sum of the lengths of shorter similar trip between passenger 1 and passenger 2, "a21" is sum of the lengths of shorter similar trip between passenger 2 and passenger 1, "B" is the set of similar trips, in which the longest one is chosen to determine the shorter similar trip, "l" is the length of the trip, and the other parameters are defined previously [15].

Two trips are considered as temporally similar if their trip time overlaps. The temporal similarity between two passengers is assumed as the ratio of sum of the overlapped time between the trips to the greater sum of the all trips time. Equation (4) 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 and "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 longer trip time. Values of the temporal similarity vary between 0 and 1 [15].

Temporal Similarity Measure between Trips

$$TS(T_1, T_2) = \begin{cases} \dfrac{\min(A_1, A_2) - \max(B_1, B_2)}{\max((A_1 - B_1), (A_2 - B_2))}; & [(B_1 > B_2 \& A_1 < A_2) \parallel (B_2 > B_1 \& A_2 < A_1)] \\ 0; & \text{otherwise} \end{cases} \tag{4}$$

Algorithm 2 presents the pseudocode for the temporal similarity between two passengers (P1, P2) who, respectively, have m and n trips, where "TT" stands for trip time, "OT(T1, T2)" stands for overlapped time that is calculated between the two trips, "a" is defined for measuring the overlapped time, and the other parameters are defined previously [15].

Separately measuring the spatial similarity and temporal similarity of the trips enables us to find similar trips in the same time interval and the same corridor (the same or opposite direction). For instance, assuming two passengers each of whom has two trips in the morning and evening and the same route (for example, a bus route between stops G and H) but opposite directions (one passenger goes from stop G to H in the morning and returns from stop H to G in the evening; another passenger goes from stop H to G in the morning and returns from stop G to H in the morning), these two passengers have temporal similarity because they both are in the public transit network in the same time period, and,

also, they have spatial similarity because each of whom have two trips traversing from stop G to stop H and from stop H to stop G.

This paper utilises the agglomerative hierarchical clustering algorithm using the Ward method that minimises the total within-cluster variance. While the agglomerative hierarchical clustering algorithm can be implemented with various methods such as Single, Average, Complete, and Ward, the Ward method is chosen according to the results of comparing these methods by Ferreira and Hitchcock [27]. It is chosen because it does not need to determine the number of clusters, and it is flexible with different similarity measures. It begins at the bottom where each object has its own cluster and merges them till all the objects form one cluster at the top. The result of the hierarchical agglomerative clustering is a dendrogram that shows how the objects are merged at each step [26]. According to the shape of the dendrograms and the Silhouette information, the dendrogram can be cut at a proper level. The Silhouette information refers to a method of interpretation and validation of consistency within clusters of data [28]. Spatial or temporal clusters of passengers are discovered after cutting the related dendrograms.

Spatial clusters include groups of passengers with similar trip routes, and temporal clusters comprise groups of passengers with similar trip times. To have groups of passengers similar in both trip routes and time, three methods are explained and compared: S-T, T-S, and ST. S-T reclusters each spatial group into several temporal groups. T-S reclusters each temporal group into several spatial groups. A potential flaw for both S-T and T-S is that, at the first step of clustering, they ignore the second similarity measure. For instance, if two passengers have high spatial similarity and low temporal similarity, then S-T would consider them in the same group, but T-S would not.

ST is a one-step clustering method that combines both spatial and temporal similarity matrices into one matrix (spatiotemporal similarity matrix). For joining the similarity matrices, the matrices are multiplied element by element. For instance, if passenger A and passenger B have 0.75 spatial similarity and 0.5 temporal similarity, then the spatiotemporal similarity between them will be 0.375. One of the premises for this method is that similarities can be taken as probabilities; in simple words, the spatial and temporal similarities are indices to ultimately measure the probability of two passengers confronting during their trips. Multiplying the spatial and temporal similarity values is calculating the probability of occurring two independent events: one is passengers travelling at the same locations and the other one is passengers travelling at the same time. In other words, if having the spatial similarity between two trips is assumed an independent event from the temporal similarity, then product of the spatial and temporal similarity values equals to having both events. Clustering passengers according to the spatiotemporal similarity matrix is the combined one-step spatiotemporal method. The combined one-step clustering method identifies the groups of passengers who are simultaneously spatially and temporally similar.

4. Results

The explained methods are applied to the smart card dataset of Translink, the public transport authority of South East Queensland (SEQ), Australia. The dataset for three weekdays and one weekend day 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 15% of the whole number of transactions. Considering the analysis from Alsger et al. [29], 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 feature of the Translink dataset as most of the AFC systems around the world just include boarding or alighting transactions. It should be mentioned that the analysis is done by R (version 3.3.2) language in RStudio framework [30]. 600 metres (value for A in (1)) and 6 degrees (value for B in (1)) are adjusted for the Brisbane public transit network [15]. In order to have a concise presentation, only the maps for a few groups for Wednesday will be presented in the paper. Figure 2 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 the direction of main corridors in Brisbane are presented in the map.

Following is an illustrative example for 12 passengers extracted from the dataset. Table 2 shows the spatial similarity values between these passengers, and Table 3 presents the temporal similarity values.

Given the similarity matrices, the hierarchical agglomerative algorithm is implemented and the outputs as dendrograms are presented in Figure 3. At the bottom of the dendrograms, the number of passengers is presented and this shows how they are merged into one cluster at each level. Also, the height of the dendrograms shows differences between the passengers; a higher height means more difference between the passenger groups. Also, values of the Silhouette information for the similarity matrices are presented in Table 4. Higher values of the Silhouette information show a better level for cutting the dendrograms.

Considering the shape of the dendrograms and values of the Silhouette information, the spatial similarity dendrogram is cut at level 4, the temporal at level 3, and the spatiotemporal at level 5. Then, each spatial group is cut into two groups considering the temporal similarity values between the members of the spatial group. Likewise, each temporal group is divided into two groups. S-T groups are $\{[1]; [3]; [2]; [6, 9]; [4, 5, 11]; [8, 10]; [7]; [12]\}$, T-S groups comprises $\{[1, 4, 5, 7, 11]; [6, 9]; [2, 3]; [8]; [10]; [12]\}$, and ST groups are $\{[1, 2, 3]; [4, 5, 7, 11]; [6, 9]; [8, 10]; [12]\}$.

According to the dendrogram and Silhouette information values in Figure 4, the spatial dendrograms are cut at 16 groups. Each spatial group is clustered into four S-T groups, which have similar spatial patterns but different temporal patterns. Members of each S-T group include the passengers who have similar boarding and alighting times of transactions with similar routes. Figure 5 shows the routes and temporal

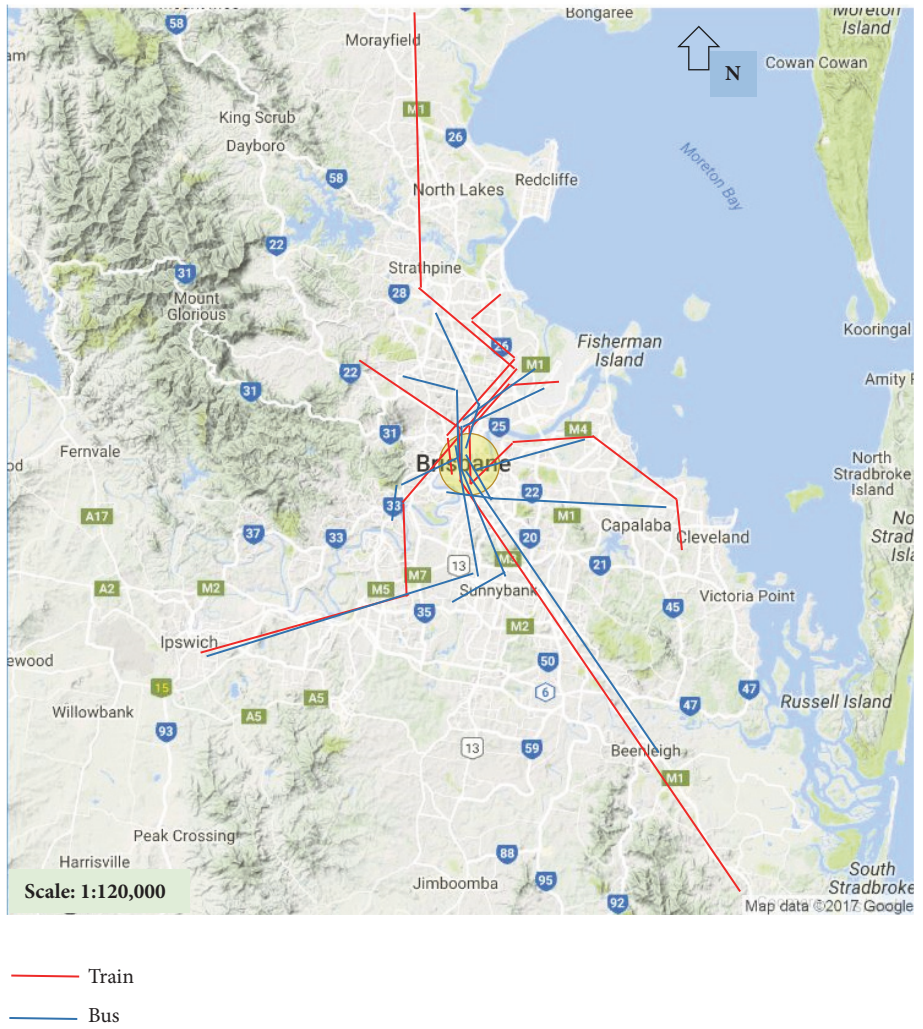


Figure 2: Brisbane map.

Table 2: Spatial similarity.

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0.33	1	0	0	0	0	0	0	0	0	0
2	0.33	1	0.33	0	0	0.66	0	0	0.66	0	0	0
3	1	0.33	1	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0.33	0	0.25	0.33	0	0.50	0.50	0
5	0	0	0	0.33	1	0	0	1	0	0.33	0.33	0
6	0	0.66	0	0	0	1	0	0	1	0	0	0
7	0	0	0	0.25	0	0	1	0	0	0	0	0
8	0	0	0	0.33	1	0	0	1	0	0.33	0.33	0
9	0	0.66	0	0	0	1	0	0	1	0	0	0
10	0	0	0	0.50	0.33	0	0	0.33	0	1	1	0
11	0	0	0	0.50	0.33	0	0	0.33	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	1

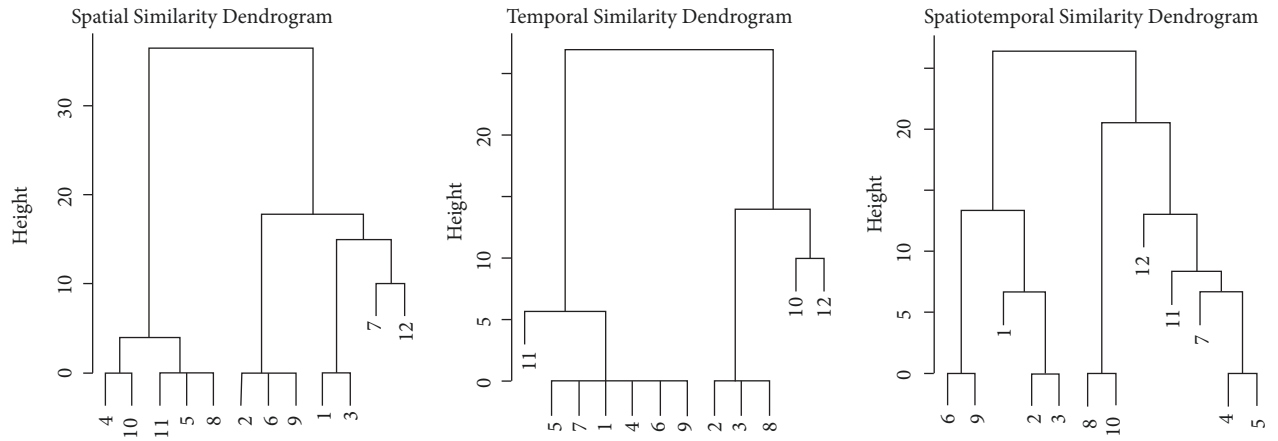


Figure 3: Dendrograms.

Table 3: Temporal similarity.

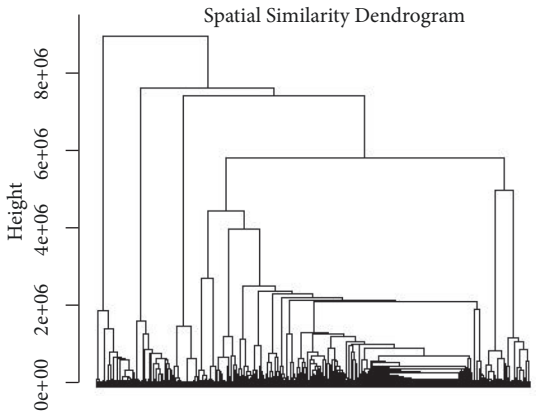
	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0.21	0	1	0.44	0.50	0.44	0	0.66	0	0	0.27
2	0.21	1	0.27	0.21	0.27	0.09	0.27	0.25	0.36	0	0.09	0
3	0	0.27	1	0	0	0	0	0.44	0	0	0	0
4	1	0.21	0	1	0.44	0.50	0.44	0	0.66	0	0	0.27
5	0.44	0.27	0	0.44	1	0.66	1	0	0.33	0	0.16	0
6	0.50	0.09	0	0.50	0.66	1	0.66	0	0.50	0	0.25	0
7	0.44	0.27	0	0.44	1	0.66	1	0	0.33	0	0.16	0
8	0	0.25	0.44	0	0	0	0	1	0	0.16	0	0
9	0.66	0.36	0	0.66	0.33	0.50	0.33	0	1	0	0.25	0
10	0	0	0	0	0	0	0	0.16	0	1	0	0
11	0	0.09	0	0	0.16	0.25	0.16	0	0.25	0	1	0
12	0.27	0	0	0.27	0	0	0	0	0	0	0	1

Table 4: Silhouette information values.

Number of groups	Spatial similarity	Temporal similarity	Spatiotemporal similarity
2	0.52	0.52	0.32
3	0.61	0.63	0.5
4	0.73	0.36	0.53
5	0.69	0.44	0.57
6	0.65	0.39	0.44
7	0.57	0.33	0.45
8	0.52	0.36	0.48

density diagrams for four S-T groups; blue points represent origin stops and red ones represent destination stops. The spatial and temporal patterns can, respectively, be presented by a map for the route and a density plot for the transactions time; hence, each S-T group is represented by a map and a density plot. Most of the spatial patterns indicate trips between suburbs and CBD of Brisbane, which can be considered as work or shopping trips. Also, most of the temporal patterns show two peaks plots representing morning and evening peak in the public transit network, which can be assumed as work-home trips. There are some one flat peak plots that can be considered as shopping-home trips, which happen all day long. Consequently, S-T groups represent passenger trips with various characteristics.

According to the dendrogram and Silhouette information values in Figure 6, the temporal dendrograms are cut at 8 groups. Each temporal group is divided into the eight T-S groups, which have similar temporal patterns but different spatial patterns. Members of T-S groups are the passengers who have similar boarding and alighting locations of transactions during specific periods and peaks. Figure 7 shows the temporal density plots and maps for four T-S groups. Temporal plots mostly present two-peak transactions during the day and some with one flat peak groups. Similar to



Number of clusters	Silhouette values	Number of clusters	Silhouette values
13	0.5089	16	0.5208
14	0.5145	17	0.5093
15	0.5199	18	0.5023

Figure 4: Dendrogram and Silhouette information values for spatial clustering.

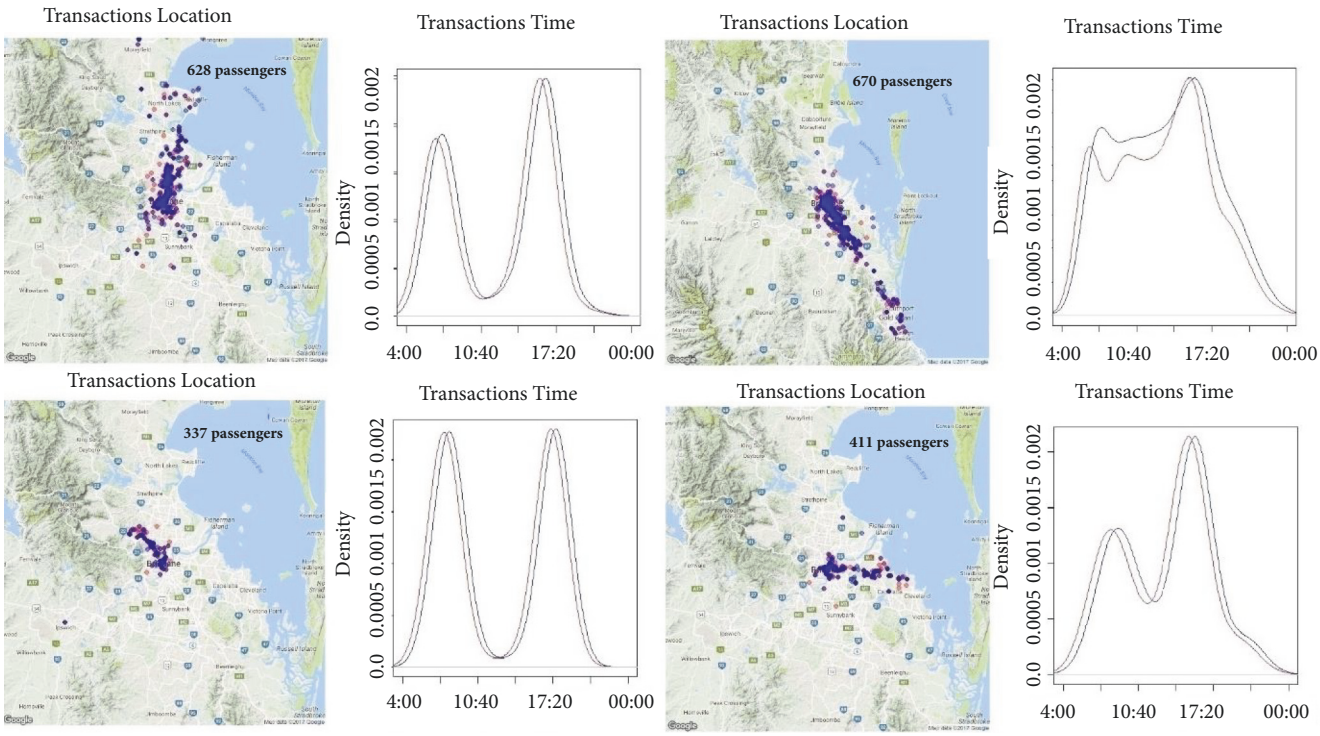
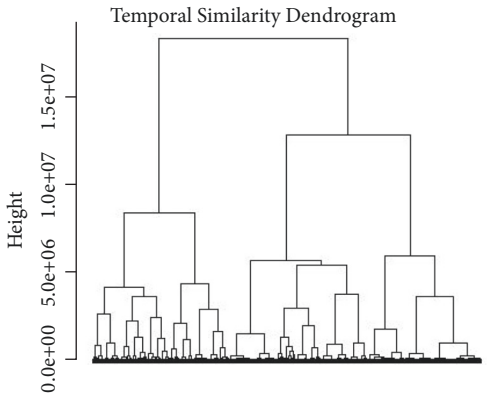


Figure 5: S-T groups.



Number of clusters	Silhouette values	Number of clusters	Silhouette values
5	0.4801	8	0.5047
6	0.4857	9	0.5036
7	0.5027	10	0.4280

Figure 6: Dendrogram and Silhouette information values for temporal clustering.

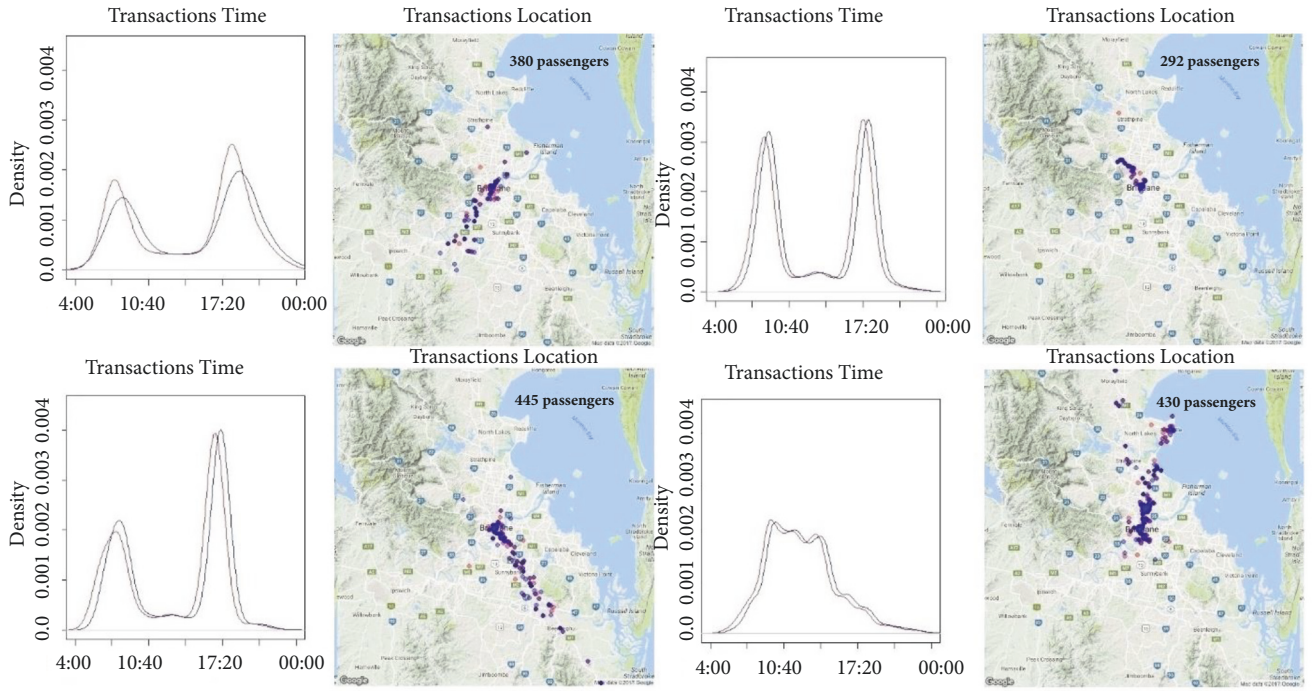
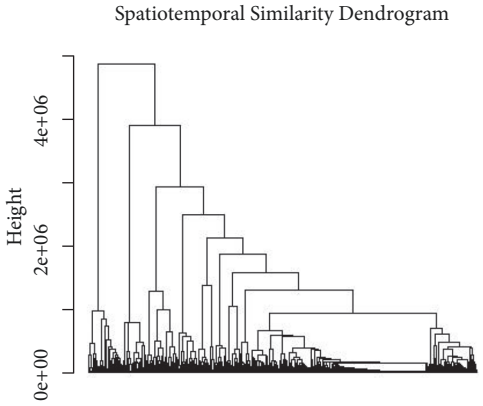


Figure 7: T-S groups.



Number of clusters	Silhouette values	Number of clusters	Silhouette values
29	0.44954	32	0.45206
30	0.4508	33	0.44873
31	0.45164	34	0.44115

Figure 8: Dendrogram and Silhouette information values for ST clustering.

S-T groups, two-peak plots can be assumed as work-home trips, and one flat peak plots as the shopping-home trips that usually happen at the middle of the day. Also, most of the spatial patterns present trips between suburbs and CBD. Therefore, T-S groups represent passenger trips with variety of temporal and spatial features.

According to the dendrogram and Silhouette information values in Figure 8, the spatiotemporal dendrograms are cut at 32 groups. Four ST groups are represented in Figure 9, in which each route and its next density plot represents a ST group. Members of ST groups are passengers who simultaneously have both similar routes and transactions times. Obvious corridors for routes from suburbs to CBD with clear peaks for the temporal density plots are observed in the ST groups. Schematically comparison, ST groups find most of the spatial and temporal patterns including trips between suburbs and CBD with peaks at the morning and evening. Hence, ST groups can represent the passenger trips with fewer numbers of groups than S-T or T-S.

Table 5 shows the mean of the spatial and temporal similarity values for the discovered groups. It should be noted that increasing the number of the groups will increase the value of similarity means; therefore, the number of groups should be considered as a factor in discussing effective values on the similarities. Also, in order to compare the different alternatives, the relative values of the spatial and temporal similarities are more important than their absolute values. As it is expected, spatial clustering has the highest (with 16 groups) mean spatial similarity, and temporal clustering leads to the highest (with 8 groups) mean temporal similarity. S-T and T-S clustering have twice the number of groups as ST clustering; however, the values for mean spatial similarity in

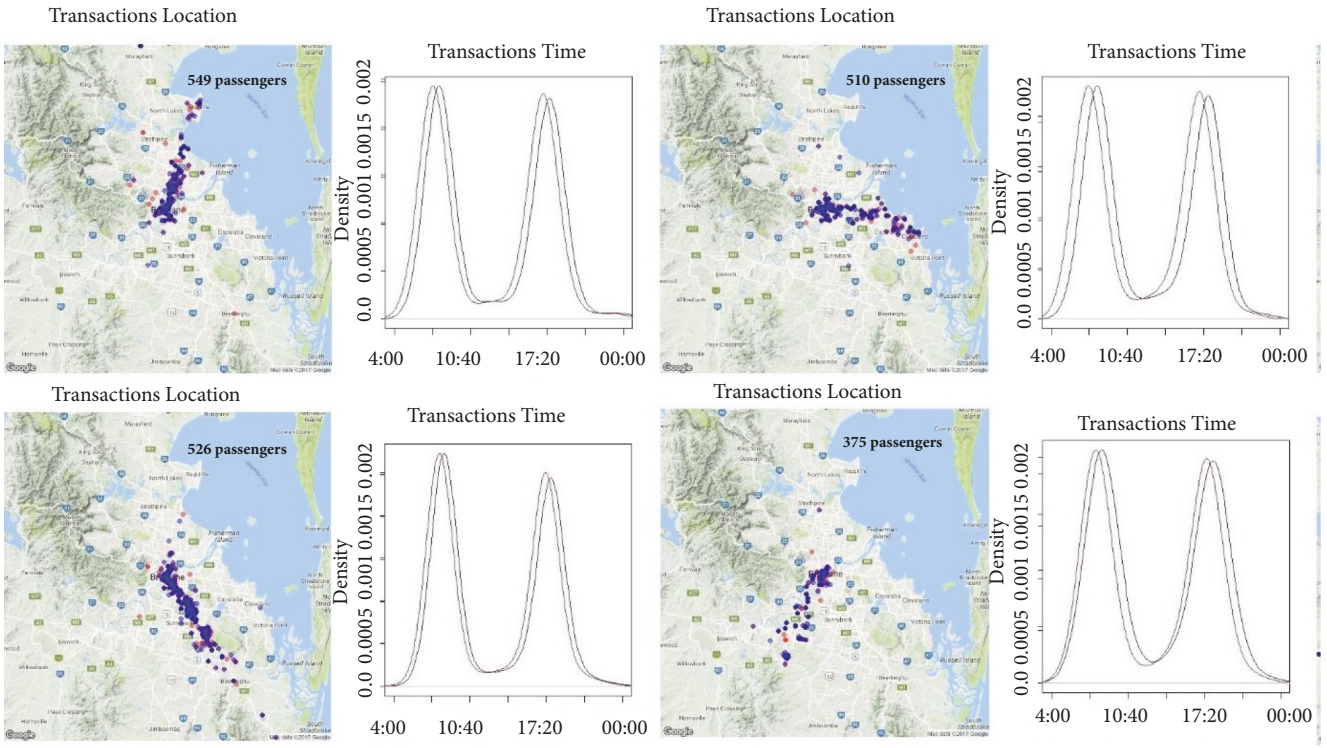


Figure 9: ST groups.

ST clustering are higher than S-T and T-S; also, the values for mean temporal similarity are close together. Furthermore, the average of means for the spatial and temporal similarity values of ST clustering are higher than the others; it basically means that members of ST groups are more likely to confront each other during their trips than S-T and T-S groups. Consequently, ST clustering leads to higher values of similarity in groups (with 32 groups) in comparison with S-T and T-S (with 64 groups) clustering methods.

Cross correlation analyses the correlation between groups of the different clustering methods. It reveals how groups from different methods are correlated. To achieve this goal, the number of passengers who are in the same group in different clustering methods are counted and then the number is divided by the size of the group. For instance, if an S-T group has 100 passengers, among which 40 remained in the same T-S group, then the cross correlation between the S-T and T-S groups is 40%. The numbers in Table 6 represent the average of all groups' correlation. For instance, 24% of S-T groups remain in the same groups of T-S, and 32% of T-S groups remain in the same groups of S-T. The fourth column (ST) has the highest value among the others, which means ST clustering covers higher proportions of S-T and T-S clustering. Also, S-T groups cover more T-S groups than the reverse situation. Considering the average of the spatial and temporal similarity values, number of groups, and the cross correlation values, the ST clustering method is a more robust method to discover the spatial and temporal patterns than S-T and T-S.

Figure 10 contains some examples to illustrate how passengers are clustered by the different methods. Figure 10 shows three examples of exchanging passengers between S-T, T-S, and ST groups. The first example shows a ST group with 274 passengers who are mostly clustered in four S-T groups; all S-T groups have similar spatial patterns, but the last one has a different temporal density plot from the ST group. The second example shows how a T-S group with 263 passengers is fit into 5 ST groups; the passengers are clustered by ST in the same spatial pattern with more diverse temporal patterns. The last example shows an S-T group with 157 passengers that are grouped in 4 T-S groups; T-S distinguishes most of the passengers in the same spatial patterns with more diverse temporal patterns.

At the end, proportions of passengers in the spatial clusters and temporal clusters are compared with proportions of passengers in the S-T, T-S, and ST clusters. According to the dendrograms and Silhouette information in Figures 4 and 6, passengers are grouped in 16 spatial clusters and 8 temporal clusters. Discovered groups by T-S and ST are compared with the spatial clusters (proportions of passengers in the S-T clusters are same as the spatial clusters); the spatial pattern of each T-S and ST group is assigned to one of the spatial clusters considering the direction and length of the discovered patterns. Also, discovered groups by S-T and ST are compared with the temporal clusters (proportions of passengers in the T-S clusters are the same as the temporal clusters); the temporal pattern of each S-T and ST group is assigned to one of the temporal clusters

Table 5: Average of the spatial and temporal similarity values in the groups.

Wed	Mean Spatial similarity	Mean Temporal similarity	Average of mean of spatial and temporal similarity values	No. groups
Spatial	0.29	0.09	0.19	16
Temporal	0.013	0.16	0.09	8
S-T	0.29	0.16	0.23	64
T-S	0.26	0.19	0.23	64
ST	0.34	0.20	0.27	32
Thu				
Spatial	0.27	0.07	0.17	16
Temporal	0.008	0.15	0.08	8
S-T	0.28	0.14	0.21	64
T-S	0.21	0.16	0.18	64
ST	0.31	0.18	0.25	32
Fri				
Spatial	0.24	0.07	0.17	16
Temporal	0.009	0.15	0.12	8
S-T	0.25	0.15	0.2	64
T-S	0.21	0.17	0.19	64
ST	0.31	0.17	0.24	32
Sat				
Spatial	0.18	0.04	0.11	16
Temporal	0.006	0.12	0.07	8
S-T	0.19	0.1	0.15	64
T-S	0.16	0.12	0.14	64
ST	0.20	0.11	0.16	32

Table 6: Cross correlation between groups from different clustering methods.

Wed	S-T	T-S	ST
S-T	100%	24%	44%
T-S	32%	100%	36%
ST	19%	13%	100%
Thu			
S-T	100%	21%	48%
T-S	26%	100%	40%
ST	18%	12%	100%
Fri			
S-T	100%	22%	48%
T-S	26%	100%	38%
ST	18%	12%	100%
Sat			
S-T	100%	19%	46%
T-S	24%	100%	34%
ST	17%	10%	100%

considering the shape, peak time, and density values of the discovered patterns. Figure 11 presents the spatial clusters and temporal clusters besides the proportions of passengers for each of the mentioned methods.

Comparing proportions of the spatial clusters with the T-S and ST clusters show that distribution of population in the spatial clusters is more similar to the T-S than ST. All the 16 spatial clusters are discovered by both T-S and ST methods. However, the proportions of population in the spatial clusters are closer to the T-S than ST method. Moreover, comparing proportions of the temporal clusters with the S-T and ST clusters shows that distribution of population in the temporal clusters is more similar to the S-T than ST. All the 8 temporal clusters are discovered by the S-T, while the ST method

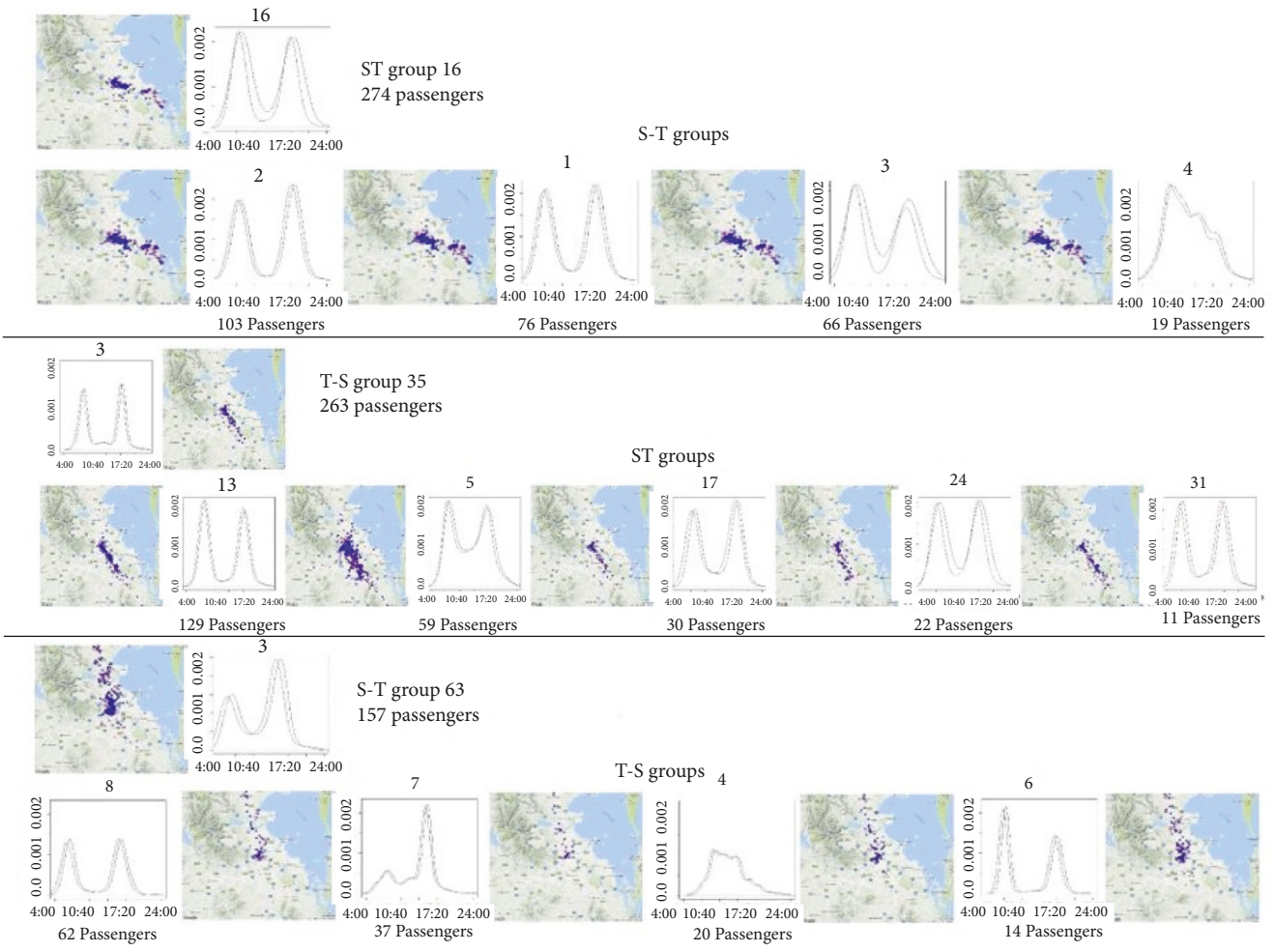


Figure 10: Passenger groups by different methods.

Table 7: Ranking of the sequential and combined methods.

	Average spatial similarity in the groups	Average temporal similarity in the groups	Spatial diversity	Temporal diversity
S-T	2	3	1	2
T-S	3	2	2	1
ST	1	1	3	3

discovered 6 temporal clusters (temporal patterns 4 and 7 are missing from the ST clusters). Also, the proportions of population in the temporal clusters are closer to the S-T than ST method. Therefore, the sequential clustering methods have a better performance in discovering diversity in the spatial and temporal patterns in the network.

Table 7 ranks the sequential and combined methods considering the results of the analyses. According to Table 5, the ST clusters have the highest spatial and temporal similarity values; average of the spatial and temporal similarity values among passengers in ST clusters is higher than the passengers in the S-T and T-S clusters. According to Figure 11, the T-S method has a better performance than the ST method in discovering the spatial diversity; the distribution of passengers in the spatial clusters is more similar to the T-S method than the ST method. Also, the S-T method has a better performance than the ST method in discovering the temporal diversity; the distribution of passengers in the temporal clusters is more similar to the S-T method than the ST method. In conclusion, while the sequential methods (S-T and T-S) discover more diverse spatial and temporal patterns in the network, the ST method entails more robust groups (higher spatial and temporal similarity values inside the groups) than the others.

5. Conclusion

The paper investigates the different clustering methods for discovering groups of passengers whose trips are spatially and

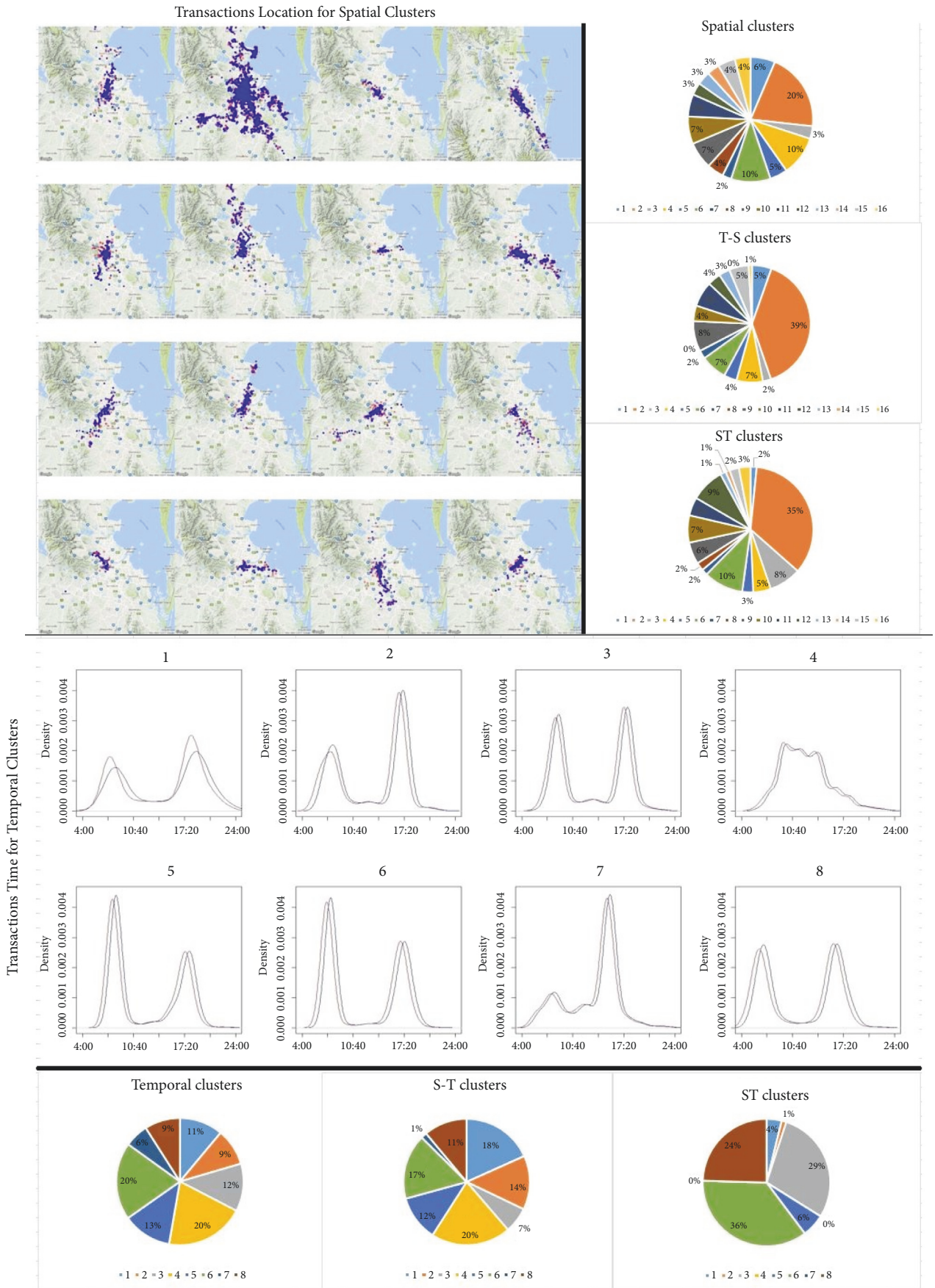


Figure 11: Proportions of passengers in the clusters.

temporally similar. First, the spatial and temporal similarity measures are defined. Then, the passengers are clustered using the hierarchical agglomerative algorithm with three different methods. The outcomes of each method are examined and compared using maps, temporal density plots, and quantitative values. Each method generates different groups with specific characteristics. The S-T method shows more diversity in the spatial dimension of the passenger trips. The T-S shows more specific temporal density plots for the groups. The ST shows a moderate combination of S-T and T-S methods with a lower number of groups and higher values for the spatial and temporal similarities in comparison with S-T and T-S. Also, ST groups cover higher proportions of S-T and T-S groups in comparison with the coverage of S-T and T-S on ST groups. In conclusion, while the sequential methods (S-T and T-S) discover more diverse spatial and temporal patterns in the network, the ST method entails more robust groups (higher spatial and temporal similarity values inside the groups) than the others.

Results from this paper are independent from the used spatial and temporal similarity measures because the alternatives of sequential or combined clustering, without loss of generality, can be implemented on other spatial and temporal similarity measures. S-T, T-S, and ST are three different methods for clustering passengers with the same spatial and temporal similarity measures. In other words, this paper investigates the effects of each method in clustering by the same similarity measures. S-T clustering can be used in those cases where the spatial diversity is the main focus, while T-S can be used in finding more specific temporal patterns. Comparing to S-T and T-S, ST is a moderate method in finding diverse spatial and temporal patterns, but it can be used to discover more robust groups of passengers, where confronting passengers during their trips is more important than diversity of the patterns. Consequently, this paper sheds light on differences between sequential and combined spatial and temporal clustering alternatives in the public transit network, which can lead to begin new trends of studies in discovering and implementing data mining techniques in the public transit network.

The main difference between S-T and T-S (as the sequential clustering methods) with ST (as the combined method) originates from ignoring one of the similarity measures at each step of clustering in the sequential methods. S-T reclusters each spatial group into several temporal groups. T-S reclusters each temporal group into several spatial groups. Both S-T and T-S at the first step of clustering ignore the second similarity measure. For instance, if two passengers have high spatial similarity and low temporal similarity, then S-T would consider them in the same group, but T-S would not. However, ST method considers both spatial and temporal similarity measures at the same step. Therefore, having more robust groups by ST is expectable.

While choosing between the S-T, T-S, and ST in practice or research might just depend on specific applications, knowing the differences between these methods can help researchers/practitioners to decide on the proper method for their desired applications. Also, focusing on the differences between the spatial and temporal clustering methods can create new trends in the public transit research area. For an instance, the clustering methods can be used in designing bus networks [31]. In simple words, designing bus networks happens in two steps (the first two steps out of four main steps in designing the public transit network [32]). First, routes of the network are designed according to the spatial movement demand by the passengers, and, then, schedules are designed according to the designed routes and temporal movement demand of the passengers. In other words, designing the bus network is similar to the S-T clustering of the movement demand of passengers. Also, it might be possible to design the network with the two other methods (T-S and ST). Designing the bus network from the T-S perspective is similar to the following: first, schedules are designed according to the temporal demand of passengers, and, then, routes are designed according to the designed schedules and spatial demand of the passengers. Considering the results from this study, it is likely to have a more reliable temporal bus network with the T-S than S-T perspectives.

Another example of implementing these clustering methods in the real world is passenger segmentation methods that discover groups of passengers, who are similar in their travel behaviour. Passenger segmentation methods are usually used in marketing applications where marketing companies are willing to target certain types of passengers. Another application of an improved clustering method is in policy making analysis where a new policy might affect passenger clusters differently. Each of the investigated methods in this study establish different sets of passengers. The ST method generates groups of passengers with more spatial and temporal similarity than S-T or T-S method. In other words, passengers in ST groups are more likely to confront each other during their trips than passengers in S-T or T-S groups, which basically means that passengers in ST groups are more likely to share a certain stop or bus route at the same time period during their trips. Higher chance of having similar passengers in a specific route or time slot is more desirable for the marketing companies than having more diversity because marketing companies are more likely to target higher number of passengers in a location or time point. Therefore, using the ST method could be a more attractive clustering method for the marketing companies. The same argument is also true for policy making analysis when the improved clustering method of the ST method is applied.

Additional analyses can be performed to extend this work. First, public transit networks and schedules can be designed according to the different methods and, then, the effectiveness of each method can be compared. Second, the effects of each method on real world applications such as demand-responsive transport should be studied. Third, trip similarity measures could be defined and compared by simultaneously considering space and time dimensions, using time-geography concepts, of the trip.

Data Availability

The smart card data used in this study has a restricted access because of privacy issues.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. K. A. Chu, R. Chapleau, and M. Trépanier, "Driver-assisted bus interview passive transit travel survey with smart card automatic fare collection system and applications," *Transportation Research Record*, no. 2105, pp. 1–10, 2009.

[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] T. Litman, *Evaluating public transit benefits and costs*, Victoria Transport Policy Institute, 2015.

[4] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *Proceedings of the IFAC Proceedings Volumes*, vol. 39, 3, pp. 399–404, 2006.

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

[6] M. Rahbar, M. Mesbah, M. Hickman, and A. Tavassoli, "Determining route-choice behaviour of public transport passengers using Bayesian statistical inference," *Road & Transport Research*, vol. 26, no. 1, pp. 64–72, 2017.

[7] 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.

[8] 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, ACM, 2008.

[9] H. Faroqi, M. Mesbah, and J. Kim, "Inferring socioeconomic attributes of public transit passengers using classifiers," in *Proceedings of the 40th Australian Transport Research Forum (ATRF)*, 2018.

[10] C. P. Tribby and P. A. Zandbergen, "High-resolution spatio-temporal modeling of public transit accessibility," *Applied Geography*, vol. 34, pp. 345–355, 2012.

[11] S. Farber, M. Z. Morang, and M. J. Widener, "Temporal variability in transit-based accessibility to supermarkets," *Applied Geography*, vol. 53, pp. 149–159, 2014.

[12] 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*, 2015.

[13] M. Saeedmanesh and N. Geroliminis, "Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks," *Transportation Research Part B: Methodological*, vol. 105, pp. 193–211, 2017.

[14] S. Seyedabrishami, M. Iranmanesh, and A. Mohades Deylami, "Short-term prediction of passenger demand in bus stations, case study: Karimkhan bridge-Jomhoori square," *Modares Civil Engineering journal*, vol. 18, no. 4, 2018.

[15] H. Faroqi, M. Mesbah, and J. Kim, "Spatial-temporal similarity correlation between public transit passengers using smart card data," *Journal of Advanced Transportation*, vol. 2017, Article ID 1318945, 14 pages, 2017.

[16] 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, pp. 1–12, 2013.

[17] 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.

[18] 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.

[19] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.

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

[21] 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.

[22] M. S. Ghaemi, B. Agard, M. Trépanier, and V. Partovi Nia, "A visual segmentation method for temporal smart card data," *Transportmetrica A: Transport Science*, vol. 13, no. 5, pp. 381–404, 2017.

[23] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 274–289, 2017.

[24] H. Faroqi, M. Mesbah, and J. Kim, "Applications of transit smart cards beyond a fare collection tool: A literature review," *Advances in Transportation Studies*, vol. 45, pp. 107–122, 2018.

[25] 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.

[26] M. J. Zaki and W. Meira Jr., *Data Mining And Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, UK, 2014.

[27] L. Ferreira and D. B. Hitchcock, "A comparison of hierarchical methods for clustering functional data," *Communications in Statistics—Simulation and Computation*, vol. 38, no. 9, pp. 1925–1949, 2009.

[28] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[29] 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.

[30] R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, 2016, https://www.R-project.org/.

[31] O. J. Ibarra-Rojas, F. Delgado, R. Giesen, and J. C. Muñoz, "Planning, operation, and control of bus transport systems: a literature review," *Transportation Research Part B: Methodological*, vol. 77, pp. 38–75, 2015.

[32] A. Ceder, *Public Transit Planning and Operation: Modeling, Practice and Behavior*, CRC Press, 2016.