Transportation Research Part C 127 (2021) 103131

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Inferring trip purpose by clustering sequences of smart card records

Hamed Faroqi [a,*], Mahmoud Mesbah [a,b]

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

ARTICLE INFO

Keywords:
Trip purpose
Public transport
String similarity
Machine learning
Data mining
Big data

ABSTRACT

Smart card transactions are known as a rich and continuous source of public transit data, but they miss some important attributes about trips and passengers. One of these missing attributes is the trip purpose attribute. This paper proposes a novel method to infer the trip purpose attribute from the sequences of trips of passengers instead of separate trips. The proposed method infers the trip purpose attribute (a missing attribute in the smart card data) from the temporal attributes (available attributes in the smart card data). First, the relation between the temporal attributes and the trip purpose attribute is learnt by discovering clusters of passengers in the Household Travel Survey dataset while each passenger is represented by one sequence of trips. Then, the discovered clusters are utilized to infer the trip purpose of smart card transactions by allocating each passenger to the closest clusters. The proposed method is implemented on the smart card and HTS datasets from southeast Queensland, Australia. The evaluation results showed a considerable improvement in inferring the trip purpose compared to the results published in the literature. Notably, the effect of considering the trip sequence was more significant than considering land use variables.

1. Introduction

Smart card records are a valuable source of urban mobility data. Smart card records are transactions of passengers in the public transit network. These datasets have attracted the attention of transport researchers and planners during the last two decades. While smart card datasets are rich in passively and continuously collected attributes in the public transit network (such as location and time of trips), they lack some important attributes such as trip purpose or passengers' demographic (Faroqi et al., 2020). Another source of urban mobility data is the Household Travel Survey (HTS) which is known as a traditional pillar for transport planning. HTS data are rich in the trip, individual, and households' attributes, but they suffer from the sample size and covering the whole population (Faroqi et al., 2018a). Despite the different size and available attributes in the smart card and HTS datasets, these two mobility datasets usually have some trip attributes in common such as the start time of the trips, which could help to fuse the datasets. Therefore, smart card and HTS are two valuable sources of mobility data that can complement each other.

Clustering techniques can discover patterns and help to fuse two datasets. Clustering objects in a dataset is usually based on distance (or similarity) between the objects and separating them into different groups. In general, members in each group (or cluster) are closer (more similar) to each other than objects in other clusters. Also, two datasets can be merged based on common attributes in both

[*] Corresponding author.
E-mail addresses: h.faroqi@uq.edu.au, faroqi.hamed@gmail.com (H. Faroqi), mahmoud.mesbah@uq.edu.au, mmesbah@aut.ac.ir (M. Mesbah).

https://doi.org/10.1016/j.trc.2021.103131
Received 21 November 2020; Received in revised form 30 March 2021; Accepted 6 April 2021
Available online 21 April 2021

datasets. Discovered clusters (learnt relations) based on available attributes in one dataset can be transferred to other datasets using the common attributes in both datasets. As abovementioned, there are some common attributes between the HTS and smart card datasets that make it possible to transfer clusters from one of the datasets to the other one. Hence, clusters of passengers with similar behaviour in the HTS dataset can be discovered and transferred into the smart card dataset to enrich its records with the trip purpose attribute.

Common attributes between the HTS and smart card datasets belong to the trip records of individuals. Temporal attributes such as the start and end time of the trips are generally well recorded while spatial attributes of the trips (e.g. origin or destination location) are sometimes eliminated from the HTS due to privacy issues. Even if the spatial attributes are available in both datasets, extracting useful information (such as the land use or point of interest) from the locational attributes would still be a potential source of inconsistency because these attributes are not (in most of the cases) available at the same level of details (e.g. locational attributes might be as detailed as stop coordination in smart card datasets, but traffic zones in HTS and 'mesh blocks' in land use data). Therefore, in this study, we only focus on the temporal attributes. Clusters of individuals with similar temporal trip attributes can be discovered from the HTS and labelled with the relevant trip purposes.

The trip purpose is one of the most important missing attributes from the smart card datasets. Enriching smart card datasets with trip purpose would extend the potentials of the datasets for planning and research purposes. For example, acknowledging the importance of estimating an O-D matrix separated by trip purpose in the planning of transport networks, smart card data could be complemented by other data sources to reduce the need for expensive data collection methods for inferring trip purpose (Alsger et al., 2018). Also, recent innovative applications have been recommended for smart card datasets. Targeting passengers for advertising purposes is an example of those recent applications (Faroqi et al., 2019) that mainly relies on inferring the trip purpose attribute. Therefore, augmenting the smart card data with the trip purpose attribute would extend its applications in both traditional and innovative ways.

The existing literature on trip purpose inference can be divided into two main categories as rule-based and model-based methods. Rule-based methods specify rules and thresholds to heuristically determine the trip purpose (or activity type) (Hasan et al., 2013; Alexander et al., 2015; Zou et al., 2018). Rule-based methods require domain knowledge for designing the heuristic rules without providing an estimation of uncertainty (Zhao et al., 2020). Model-based methods use learning models and data fusion techniques to infer the activity type (Lee and Hickman, 2014; Kusakabe and Asakura, 2014; Alsger et al., 2018). Furthermore, most of the previous studies only focused on inferring work and home trip purpose types (Chakirov and Erath, 2012; Jun and Dongyuan, 2013; Zhou et al., 2014; Zhao et al., 2020) and education trips (Devillaine et al., 2012; Lee and Hickman, 2014). Alsger et al. (2018) extended the inferring goal into five categories of work, home, education, shopping, and recreational trip purposes. Regardless of the methods and activity types, the previous studies are common in one aspect that is inferring the trip purpose attribute per trip and not taking into account the 'sequence of trips' for individuals.

The proposed method in this study follows an individual-based perspective. Each trip purpose is not only inferred based on one trip's attributes but also based on attributes of other trips of the individual during the day. In other words, preceding and following trips of an individual during the day can be informative about his or her current trip. E.g. knowing that a person has already been on a work trip make it less likely to have another work trip right after that compared to a situation that we don't know about his or her previous trips. Another example is that knowing that a person has already completed an education activity make it less likely that his or her next activity be a work activity compared to a person who just started the day (no previous activity). The two abovementioned examples are only to elaborate on the context behind the innovation of the proposed method, and they may not always be true. Considering the trip sequences instead of inferring a solo trip not only takes into account each trip's attributes, but also the relation between trips. In other words, the proposed method in this research adds an extra informative attribute (the relation between previous and next trips) to the previous studies that were built on separate attributes of each trip.

This paper proposes an unsupervised learning (clustering) method to infer the trip purpose of sequences of trips of individuals. The proposed clustering method discovers the relation between the trip purpose and temporal attributes of sequences of trips from an HTS dataset to infer the trip purpose attribute in a smart card dataset. A similarity measure for measuring similarity between sequences of trips is developed from the Jaccard similarity index. Groups of individuals with a similar sequence of trips are discovered and labelled based on the developed similarity measure in the HTS dataset. The labelled groups are then transferred to the smart card dataset according to the temporal attributes of the individuals' trips. The proposed method is run on datasets from southeast Queensland, Australia. The outcome of the method is evaluated using accuracy, sensitivity, and informedness measures and compared with results from Alsger et al. (2018).

The rest of the paper is organized as follows. The methodology section explains the proposed method step by step including developing the similarity measure, the clustering method, the labelling function, and the evaluation indices. The results section

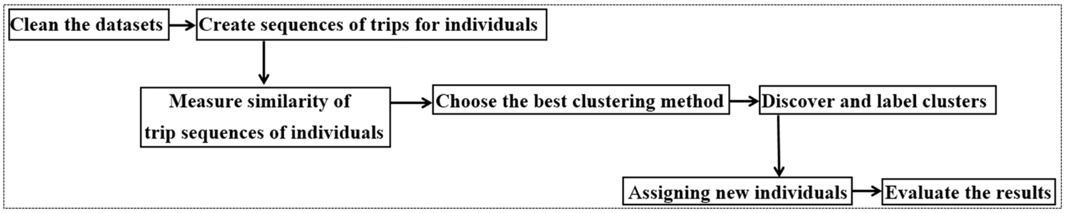
Clean the datasets
Create sequences of trips for individuals
Measure similarity of trip sequences of individuals
Choose the best clustering method
Discover and label clusters
Assigning new individuals
Evaluate the results

Fig. 1. Methodology.

describes the HTS and smart card datasets, compares the temporal attributes in both datasets, evaluates the method on the HTS, presents the outcome of running the method on the smart card data, and compares the results with the existing literature. In the end, the conclusions section discusses and summarizes the finding and implications of the proposed method.

2. Methodology

Fig. 1 summarizes the main steps of the methodology. Firstly, both HTS and smart card datasets need to be cleaned. Secondly, the temporal attributes of trips are discretised into time intervals to create the trip sequences per individual in the dataset. Next, the similarity between individuals' trip sequences is computed based on the Jaccard similarity measure. Fourthly, an Agglomerative Hierarchical Clustering (AHC) method is implemented on the HTS dataset for five different clustering distances of AHC, and the results of the clustering methods are compared using the Silhouette coefficient. After choosing the best clustering distance method, clusters of individuals with similar trip sequences are discovered and labelled. Labelled clusters can determine the trip purposes of new individuals. In the end, the HTS dataset is divided into two partitions for evaluation purposes using three evaluation indices. The proposed methodology then will be implemented on smart card data for showing its application in the next section.

Faulty or missing transactions must be removed from both datasets. These transactions can include top-up transactions (to increase the card balance), transactions with missing location or time, and transactions with the same boarding and alighting locations (Robinson et al., 2014; Tavassoli et al., 2016; Alsger et al., 2016). It should be mentioned that this study follows the literature (Pelletier et al., 2011; Alsger et al., 2018; Faroqi et al., 2018a) in focusing on individuals with more than one transaction per day. Also, transactions in the smart card dataset consist of the time and location of boarding or alighting. Each pair of boarding and alighting transactions is called a trip leg. Trip legs can be linked according to the time gap between two subsequent trip legs. In general, two trip legs are considered as part of a trip if the time gap between the two trip legs is less than a certain threshold (Alsger et al., 2018).

In this research, both datasets contain individuals who are represented by their trips during the day; a dataset $D = \{I_1, I_2, \cdots, I_N\}$ contains N individuals. An individual is represented by the temporal attributes of his/her trips during the day. Two important attributes for each trip in this study are the end time of the trip and the time gap until the next trip. The end time is the time when an individual alights from a public transit vehicle, and the time gap is the time between when an individual alights from and boards again on public transit vehicles. We use upper case characters for the end time and lower case for the time gap. For instance, "Aa" is a trip that ended at the time interval "A" with the gap of "a" hours before the next boarding. Therefore, an individual is represented by a sequence of his/her trips during the day such as AaCfZp; this particular individual had three trips during the day: the first trip ended at time interval "A" with the gap of "a" hours before his second trip; the second trip ended at time interval "C" and the gap of "f" hours; the last trip of his day ended at time interval "Z" and the gap of "p" hours.

In general, clustering methods put an object I_n from the dataset D with N objects into a set of clusters $C = \{C_1, C_2, \cdots, C_K\}$ with K clusters satisfying three conditions in Eqs. (1)–(3). Eq. (1) states that each cluster must be a subset of dataset D. Eq. (2) describes that the intersection of any two clusters must be empty. Eq. (3) satisfies a condition that the union of all clusters must be equal to the original dataset.

$$\forall_i C_i \subseteq D \tag{1}$$

$$C_i \cap C_j = \varnothing \tag{2}$$

$$\cup C_i = D \tag{3}$$

Agglomerative hierarchical clustering is known as a practical clustering method because of its flexibility in admitting user-defined distance metrics. It generates a sequence of M nested partitions $C_1, C_2, \cdots, C_M$ from the trivial clustering of $C_1 = \{\{I_1\}, \{I_2\}, \cdots \{I_n\}\}$ where each object is in a separate cluster, to the other trivial clustering of $C_M = \{\{I_1, I_2, \cdots, I_n\}\}$. The hierarchical clustering method uses a visualization tool called Dendrogram that is a binary tree capturing the nesting structure with edges between the clusters. AHC begins with each of the N objects in a separate cluster. Then, it repeatedly merges the two closest clusters until all objects are members of the same cluster; or reach K clusters if K is pre-determined (Zaki and Meira, 2014).

The main step of the AHC is determining the closest pairs of clusters at each level (of the dendrogram). Distance between two clusters is based on the distance between objects in the clusters. Distance between any two objects can be Euclidean, L2-norm, or any other user-specified measure (Zaki and Meira, 2014). Flexibility in using a user-specified distance measure (between objects) is the main privilege of the AHC. A variety of distance measures (between clusters) can be used in AHC: single link, complete link, group average, mean distance, and minimum variance (Ward's method). A concise formulation based on these distance measures is given here, and further details could be found in Zaki and Meira (2014). It might be mentioned here that after performing sensitivity analysis (presented in the Results section), Ward's method is preferred for this research.

Single Link: Given two clusters C_i and C_j, the distance between them, denoted by $\delta(C_i, C_j)$ is defined as the minimum distance between a point (object) in C_i and a point in C_j as presented in Eq. (4):

$$\delta(c_i, c_j) = \min\{\delta(x, y) | x \in c_i, y \in c_j\} \tag{4}$$

Complete Link: The distance between two clusters is defined as the maximum distance between a point in C_i and a point in C_j as presented in Eq. (5):

$$\delta(c_i, c_j) = \max\{\delta(x, y) | x \in c_i, y \in c_j\} \tag{5}$$

Group Average: The distance between two clusters is defined as the average pairwise distance between points in C_i and C_j as presented in Eq. (6) where $n_i = |C_i|$ denotes the number of points in cluster C_i:

$$\delta(c_i, c_j) = \frac{\sum_{x\in c_i}\sum_{y\in c_j}\delta(\mathrm{x}, \mathrm{y})}{n_i.n_j} \tag{6}$$

Mean Distance: The distance between two clusters is defined as the distance between the means or centroids (μ) of the two clusters as presented in Eq. (7):

$$\delta(c_i, c_j) = \delta(\mu_i, \mu_j) \tag{7}$$

Minimum Variance (Ward's Method): The distance between two clusters is defined as the increase in the Sum of Squared Errors (SSE) when the two clusters are merged. In other words, two clusters with the minimum increase in SSE are merged as presented in Eq. (8):

$$\delta(c_i, c_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j = \frac{n_i n_j}{n_i + n_j}||\mu_i\text{-}\mu_j||^2 \tag{8}$$

The Silhouette coefficient can help to choose the best option among the abovementioned distance measures. It is a method of interpretation and validation of consistency within clusters of data. It is a measure of both cohesion and separation of clusters and is based on the difference between the average distance to points in the closest cluster and points in the same cluster. For each data point $x_{i,}$ the Silhouette S_i is presented in Eq. (9) (Aranganayagi and Thangavel 2007; Zhu et al., 2010; Zaki and Meira, 2014):

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \tag{9}$$

where $\mu_{in}(x_i)$ is the mean distance from x_i to points in its cluster, and $\mu_{out}^{min}(x_i)$ is the mean of the distances from x_i to points in the closest cluster. A value close to +1 indicates that x_i is much closer to points in its cluster and is far from other clusters. A value close to zero indicates that x_i is close to the boundary between two clusters. A value close to −1 indicates that x_i is much closer to another cluster than its cluster, and therefore, the point may be mis-clustered. Finally, the Silhouette coefficient is defined as the average of S_i for all objects in the dataset (Zaki and Meira, 2014).

The proposed distance function (δ) between two individuals (objects) uses the Jaccard similarity index to measure the similarity between the individuals. In general, the Jaccard similarity index measures the similarity between two sets E and F as $\frac{E\cap F}{E\cup F}$. In this research, each trip and its preceding/following activity are represented by a set of double characters; consequently, an individual is represented by a string of double characters for his/her sequence of trips/activities, e.g. an individual with AaBbCc equals to a set of double characters of {Aa, aB, Bb, bC, Cc}. The main innovation of this research is in linking characteristics of preceding and following trips of the individuals. The proposed method implies that the preceding and following trips of an individual is informative about the individual's current trip. In simple words, the proposed method infers the purposes of all trips of an individual in one step, and not as separate trips. Therefore, the distance between two individuals is defined as the following in Eqs. (10) and (11), where $|I_i|$ stands for the number of following double characters in I_i.

$$\delta(I_i, I_j) = \frac{1}{Similarity(I_i, I_j)} \tag{10}$$

$$Similarity(I_i, I_j) = \frac{I_i \cap I_j}{I_i \cup I_j} = \frac{Number\ of\ common\ double\ characters\ in\ I_i\ and\ I_j}{|I_i| + |I_j| - Number\ of\ common\ double\ characters\ in\ I_i\ and\ I_j} \tag{11}$$

Three examples for the defined distance function are described in this paragraph. The first example assumes that there is an individual with two trips during the day as AaBb and another individual with AaCc (I_1 = {Aa, aB, Bb} and I_2 = {Aa, aC, Cc}). In example 1, the number of common double characters is 1 and the number of double characters in the union of these two individuals is 5; the distance between individuals 1 and 2 is 5. The second example assumes a third individual with AbCa trips (I_3 = {Ab, bC, Ca}). As there are no common double characters between I_3 with I_1 and I_2 then the similarity between I_3 with I_2 and I_1 is zero. The third example assumes a fourth person with one trip during the day as Aa (I_4 = {Aa}). The similarity between I_4 with I_1 and I_2 is 0.33; the similarity between I_4 with I_3 is zero.

The visual outcome of the AHC is a dendrogram presenting how individuals are merged level by level from having each individual in its cluster to having all individuals in one cluster. The dendrogram helps to decide on the number of clusters. After cutting the dendrogram tree into the desired number of clusters, labels of trip purposes can be assigned to each cluster. Each sequence of trips is represented by a set of specific Trip Purposes (TPs). The label of a cluster can be inferred based on Mode of trip purposes of members of that cluster as presented in Eq. (12) where X is a set of categorical values. Each label can consist of multiple trip purpose types relevant to the most common trip sequences in the cluster.

$$Label(C_i) = Mode(\forall_{p\in i} TPs_p)\ where\ Mode(X) = Most\ frequently\ occuring\ value\ in\ set\ X \tag{12}$$

The next step of the proposed method addresses assigning new individuals to the discovered and labelled clusters (transferring discovered and labelled clusters from the HTS to the smart card records). A new individual, represented by his/her trips during the day, is assigned to the closest cluster. The closest cluster is defined as the one with the shortest distance between its centroid (most repeated trip sequences in the cluster) and the new individual. The label of the selected cluster is considered for the new individual. I_q stands for the new individual, and μ_i stands for the centroid of cluster i as presented in Eq. (13). It should be noted that it is possible (but unlikely) that the number of trip purposes in the assigned label for a new individual (or similarly for the label of a cluster) does not match with the individual's actual number of trips during the day; e.g. an individual with five trips during the day might be assigned to a cluster with a label of three trips because that cluster is the closest among all discovered clusters.

$$Label\left(I_q\right) = Label\left(C_{\min_i \delta\left(I_q, \mu_i\right)}\right) \tag{13}$$

The outcome of the clustering method can be summarized in a confusion table that is presented by four values including True Positive (TP) (number of truly predicted objects in the class), True Negative (TN) (number of truly predicted objects in the other class (es)), False Positive (FP) (number of falsely predicted objects in the other class(es)), and False Negative (FN) (number of falsely predicted objects in the class). In the case of having more than two classes, each of these values is calculated by considering one class against all other classes (Powers, 2011; Faroqi et al., 2018b). It should be mentioned that the HTS dataset needs to be divided into two parts for learning and testing purposes. Clusters are discovered and labelled from the learning partition of the HTS data; then the testing partition of the HTS dataset is used to calculate the evaluation indices.

Three common indices to evaluate the performance of the learning process are sensitivity, accuracy and informedness. Sensitivity evaluates the number of truly predicted objects in one class in relation to the population of that class. Accuracy evaluates the number of truly predicted objects in all classes in relation to the population. Informedness considers all true and false positive and negative values; and is considered as a balanced measure. Informedness can be used even in the case of having unbalanced classes with very different sizes. Each of the mentioned measures is formulated in the following Eqs. (14)–(16). Sensitivity and Accuracy vary between 0 and 1. Informedness varies between −1 and +1. In all measures, a value closer to +1 means a better outcome (Powers, 2011; Faroqi et al., 2018b).

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Informedness = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{16}$$

3. Results

This section describes the datasets used in the study, compares the temporal attributes of the datasets, compares the distance measures and selects the best one, discovers clusters of individuals from the HTS dataset, runs the proposed method on the smart card dataset, and finally evaluates the proposed method on the HTS dataset. It should be mentioned that all the analyses are completed by R version 4.0.0 "Abor Day".

The two datasets used in this study are the household travel survey and smart card data. Both datasets are from southeast Queensland, Australia. The HTS was undertaken across southeast Queensland from 2009 through 2012. The sampling unit was the household. Every household was asked on a particular day. The HTS dataset includes sociodemographic characteristics and trip attributes of the individuals. The HTS dataset includes around 38,000 individuals with 110,000 trips across all modes of transport, among which 2,523 trips were by public transit and made by 1,233 individuals. Fig. 2 presents the histogram for transit trip purposes in the HTS dataset; trip purposes include Home, Work, Education, Shopping, and Recreational. Also, the smart card dataset in this study includes records of passengers with more than one transaction during the day on Thursday 21 March 2013. More than 130,000 passengers with approximately 600,000 transactions are in the smart card dataset. According to outcomes from Alsger et al. (2016), a

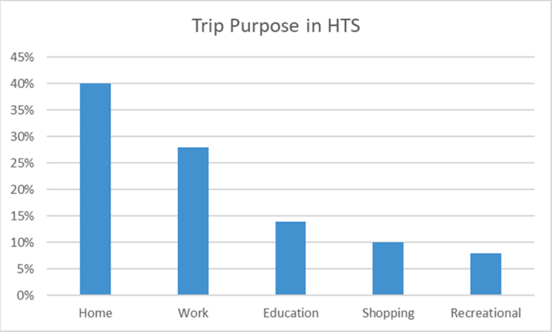

Fig. 2. Trip purposes in HTS.

30-minute time gap is considered for linking the trip legs. After cleaning the smart card dataset, 128,977 passengers with 282,453 trips remained in the dataset.

After cleaning the datasets, the trip sequences for passengers in the datasets were created. To do so, it was necessary to first determine a set of time intervals, which may change according to the case study and application. Two temporal attributes that are considered in this study for modelling the trips of individuals are the time gap between subsequent trips and the end time of the trip. In this study, the main goal is to show the applicability of the proposed method. Therefore, according to the literature (Alsger et al., 2018) and without loss of generality, three time intervals were picked for the end time of the trip, and six time intervals for the time gap between trips as follows:

End time of the trip

A: before 9
B: between 9 and 15
C: after 15

Time gap between trips (hours)

$a < 3$
$3 \leq b < 5$
$5 \leq c < 7$
$7 \leq d < 9$
$9 \leq e < 12$ (the longest time gap between two trips in the dataset is <12 h)
f: last trip of the day

Fig. 3 compares the temporal attributes between the two datasets. Only marginal differences between the bars are observable in this Figure. In Fig. 3, SCD stands for Smart Card Dataset. The biggest difference between the HTS and SCD bars belong to the time gap between trips for the last trip of the day, which hit as high as 6%. Also, the difference between the end time of the trips hits its highest (6%) for trips ended after 3 pm. It could be concluded that the temporal attributes of public transit trips in the HTS dataset follow a similar pattern as the temporal attributes in the smart card dataset.

There are five common methods (Eqs. (4)–(8)) to measure the distance between two clusters, among which Ward's method is chosen based on the following analysis. The sensitivity analysis here computes the Silhouette coefficient for each of five distance measures across a range of cluster numbers and put it in the developed distance function of Eqs. (10) and (11). In general, the Silhouette coefficient is expected to increase by increasing the number of clusters, but it should be mentioned that the number of clusters should also be set according to the context of the application and expert knowledge. Fig. 4 presents the diagram for values of Silhouette coefficients implemented on the HTS dataset for 5 to 20 clusters (which are chosen based on authors' expectation that there

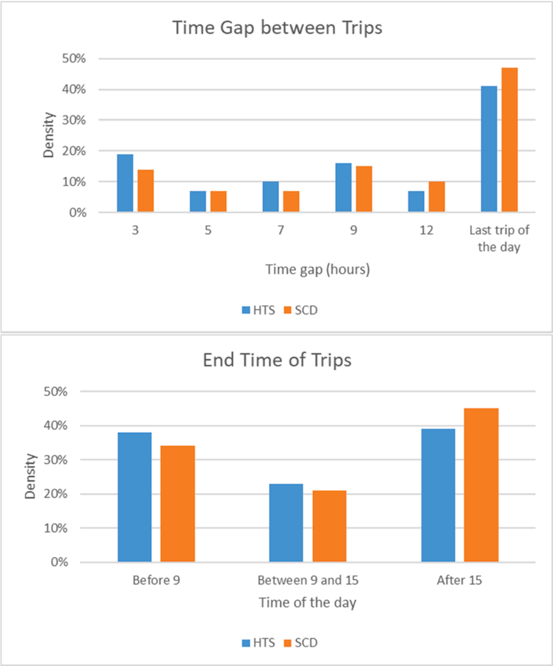

Fig. 3. Temporal attributes in HTS and smart card datasets.

would be 10 to 15 meaningful clusters in the datasets). Ward's method had the highest value of the coefficient for all number of clusters. Also, Ward's values are close to +1 which shows clusters well-separated individuals into groups with similar trip sequences.

Fig. 5 demonstrates the dendrogram of the clustering approach (AHC with Ward's method) on the HTS dataset. Without loss of generality, the dendrogram is divided into 12 clusters by the red line in Fig. 5. The individuals in the HTS dataset are clustered into 12 separate groups of individuals with similar sequences of trips during the day.

Table 1 presents the sequences of trips for the discovered groups besides their labels resulted from Eq. (12). E.g. cluster 1 contains 99 individuals with the trip sequence of BaBf (these individuals had two trips during the day: the first one ended at time interval "B" with the gap of "a" hours before their second trip of the day that ended at time interval "B" and the gap of "f" hour), who are labelled as shopping-home passengers (with 198 assigned trips). A work trip purpose type can be observed in five clusters with a total of 727 individuals. An education activity is in four clusters with a total population of 273 individuals. Four clusters are labelled with shopping activity and a total population of 272 individuals. A recreational trip purpose type is seen in only one cluster that includes 61 individuals. A home trip purpose type is chosen for the last trip of the day of all discovered clusters. Two of the labelled clusters have three activities during the day namely Work-Shopping-Home and Education-Shopping-Home sequences; however, most of the individuals (91%) are put in groups with two labels (activities) during the day. The last row of Table 1 presents the total number of individuals and labelled trips. There is a marginal difference (1.6% of the total labelled trips) between the total number of the actual trips in the HTS dataset and the labelled trips of the discovered clusters. As it was mentioned before the Eq. (13), this difference is unlikely and it is because of the mismatch between the number of trip purposes in the label and the actual trip purposes of the individuals. It should be noted that while the total number of trips in the assigned label of a cluster for an individual might not match with his/her actual number of trips, there could still be some correct inference; e.g. an individual with actual recreational-shopping-home trips might be assigned to a cluster with the label of recreational-home, which has correctly inferred the first trip of the individual (not his/her second and third trips).

Fig. 6 presents the proportions of trip purpose types in the 12 clusters that were discovered from the HTS dataset. The label for each discovered cluster is derived from Eq. (12), based on the Mode function. In 4 of 12 groups (groups 3, 4, 6, and 9) the share of the most repeated trip purposes is more than 75%. In 5 of 12 discovered clusters (clusters 1, 2, 7, 8, and 12) the Mode value is between 50% and 75%. The remaining three groups include at least 42% for the Mode. It should be mentioned that among all the discovered groups there is at least (this minimum amount happens in group 10) a 20% difference between the highest portion and the second highest one. Also, the maximum portion of 'others' trip purposes (all the available combinations of trip purposes in the HTS dataset but the six most common that are used as labels in Table 1) is 12% in group 7. Therefore, according to the gap between the highest proportion of the trip purposes and the rest of trip purposes in each cluster, using the Mode function can perform as a reliable method to label the discovered clusters.

After discovering and labelling the groups of passengers from the HTS dataset, trip purposes for passengers in the smart car dataset can be inferred as described in Eq. (13). Each passenger in the smart card datasets was assigned to the closest cluster discovered from the HTS dataset based on the distance between the passenger's trip sequence and the discovered clusters. Fig. 7 presents portions of the inferred trip purpose across all day and three different time intervals. During the day (covering all three time intervals), home trips had the highest portion followed respectively by work, shopping, education, and recreational trips. During time interval A (trips ended before 9 am), the highest portion belonged to work trips followed by education, shopping, home and recreational trips. During time interval B (between 9 am and 3 pm), the maximum portions were home, work and shopping trips followed by education and recreational trips. During the latest time interval C (after 3 pm), home trips formed the biggest slice of the pie chart followed by work, shopping, recreational and education trips. While there is no relevant source of information to validate the inferred trip purposes for the smart card data, the inferred portions follow the expected patterns.

Three public transit stops are chosen to further present the outcomes of inferring trip purpose for the smart card data. Fig. 8 presents the location of these three stops (this map is trimmed from GoogleMap). Stop number 1 is the central train station located in the Central

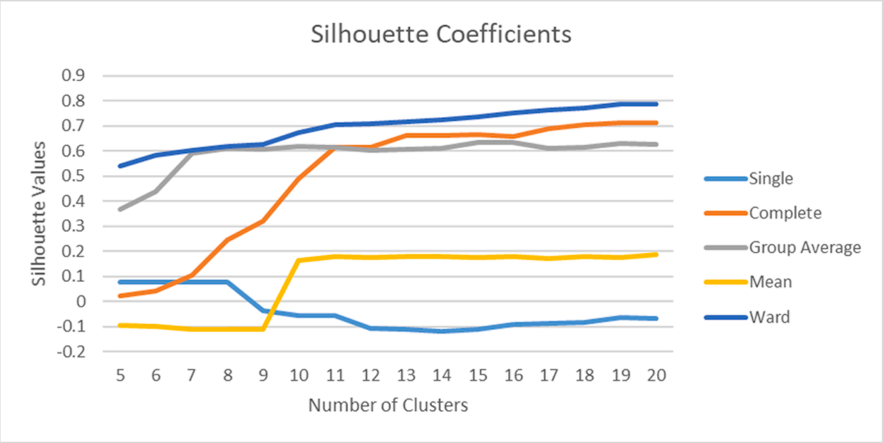

Fig. 4. Silhouette coefficients to choose the best distance measures.

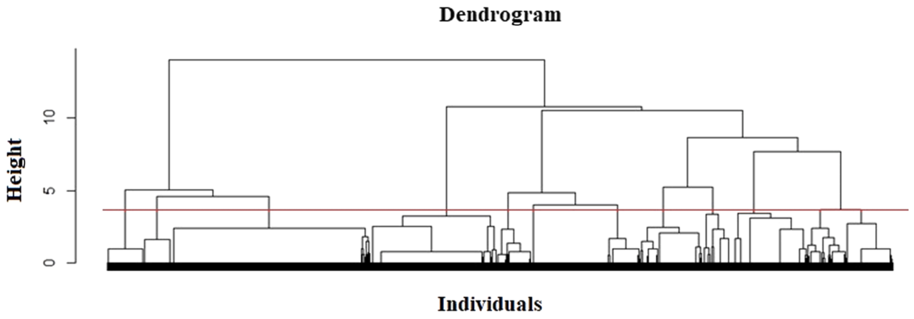

Fig. 5. Dendrogram for Ward's method.

Table 1
Discovered groups from HTS.

No.	Population	Trips sequence	Label	Number of labelled trips
1	99	BaBf	Shopping-Home	198
2	45	AdCaCf	Work-Shopping-Home	135
3	312	AdCf	Work-Home	624
4	197	AeCf	Work-Home	394
5	61	CaCf	Recreational-Home	122
6	119	AcCf	Education-Home	238
7	55	AcCaCf	Education-Shopping-Home	165
8	115	BbCf	Work-Home	230
9	58	BdCf	Work-Home	116
10	73	BaCf	Shopping-Home	146
11	50	BcCf	Education-Home	100
12	49	AbBf	Education-Home	98
Total	1233	–	–	2566

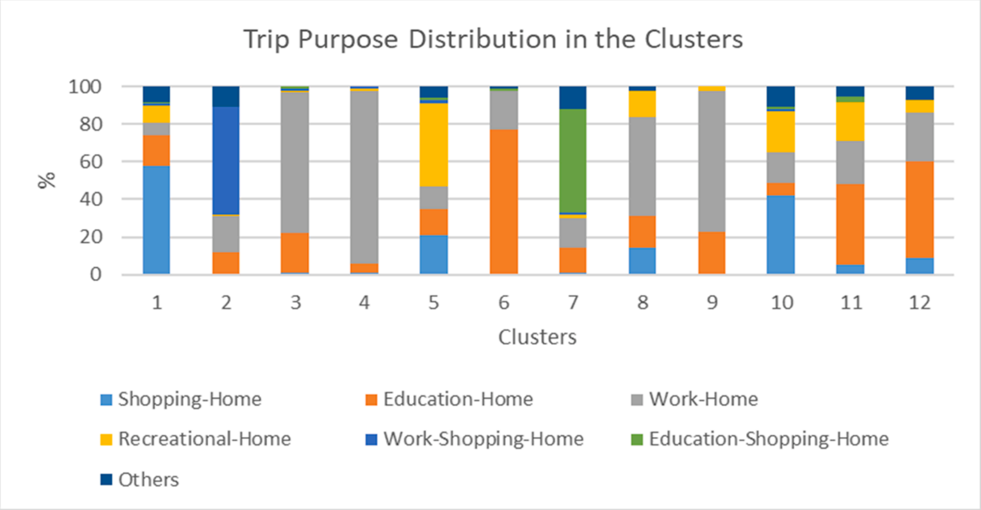

Fig. 6. Distribution of trip purposes in the discovered HTS groups.

Business District, stop number 2 is located on the campus of the University of Queensland, and stop number 3 is located next to a shopping centre (Garden City shopping centre). Proportions of the inferred trip purposes for the trips that ended at these three stops during the three time intervals (A, B, and C) are presented in Fig. 8. These pie charts can be compared across both time intervals and locations of stops.

For stop 1, work trips are in the majority during the early morning time interval (interval A); during time interval B, the work trips are still in the majority but with a smaller share; at the end of the day, the number of home trips raises and covers most of the pie chart area. For stop 2, work and education trips have close shares during time interval A; during time interval B, the portions of work trips shrink and education trips increase; at the end of the day at stop 2, education trips are still in majority (it should be mentioned that the

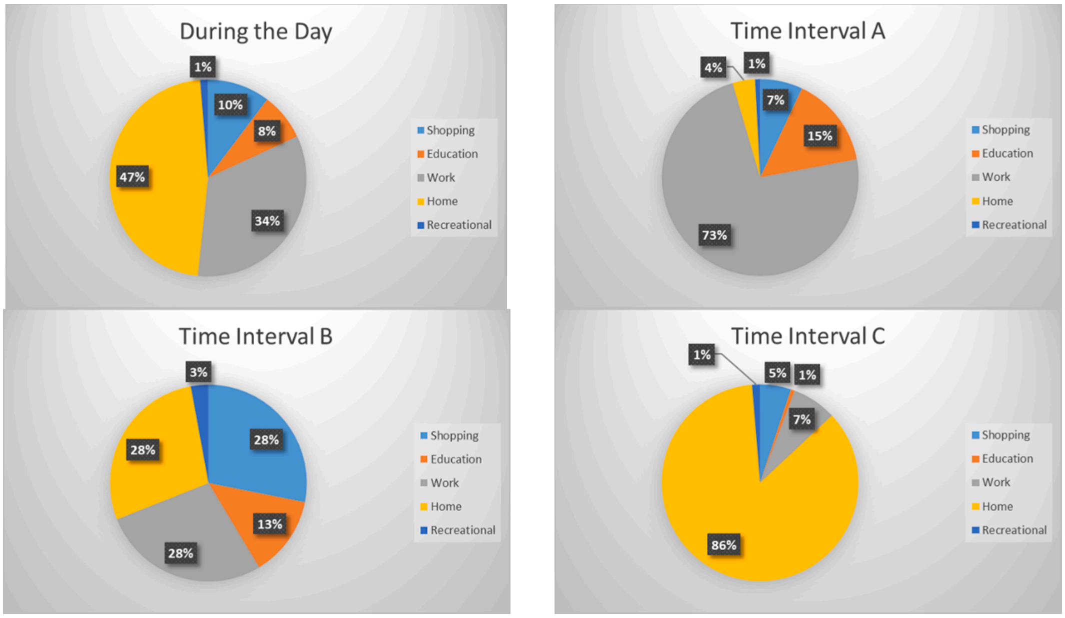

Fig. 7. Inferred trip purpose for the smart card data.

number of trips ended at stop 2 during the time interval C is very few). According to the pie chart of stop 3, work and shopping trips are in the majority during the morning peak, and it changes to the majority of home and shopping trips in the interval B; at the end of the day, the majority of the pie chart is covered by home trips. Also, it is observable that patterns of the inferred trip purposes vary across the location of these three public transit stops. Stop 1 is located at the Central Business District with a mixture of business, shopping, residential and education land use. Stop 2 is located on a university campus with the potential for work and education trips. Stop 3 is close to a shopping centre with business, shopping, and residential land uses. It is observed that the inferred trip purposes vary in the public transit network across both locational and temporal dimensions. In other words, while the spatial attributes are not directly considered in the proposed method of this paper, the effects of the spatial attributes are observed because these effects are actually hidden in the temporal behaviour of passengers at the public transit stops.

As described in Eqs. (14)–(16), three evaluation measures for the proposed method are sensitivity, accuracy, and informedness. To evaluate the proposed method, the HTS dataset was randomly partitioned into two sets: one for learning (80% of the HTS records) and one for testing (20% of the HTS records). To keep the generality of the evaluation outcome, the evaluation procedure was repeated for 10 randomly chosen learning and testing partitions. Table 2 presents average values for 10 runs of the evaluation process. Work and home trips were inferred with very high values (more than 90%) for all three evaluation measures. Also, education and shopping trips were inferred with more than 50% sensitivity and informedness. The recreational label had the lowest values for sensitivity and informedness evaluation measures.

In each run of the evaluation process, 20% of the HTS records are randomly selected for the inferring purpose. Each record in the testing partition of the HTS has an actual and an inferred trip purpose type. Table 3 presents how frequently an actual trip purpose was inferred as another trip purposes. It should be mentioned that all the numbers in Table 3 are the average of 10 runs of the evaluation step. For example, 23% of actual shopping trips were inferred as educational trip. As it was observed in Table 2, most of the work and home trips were correctly inferred. A total of 39% of the education trips were incorrectly inferred as other trips: 21% of them were inferred as work trips and 16% as shopping trips. One possible reason could be that education activities cover a diverse period that make the temporal attributes of some of the education trips similar to work or shopping trips; they can be as short as 2 h (attending only one lecture at a university, which might be similar to a shopping trip) or as long as 9 h (attending several lectures, which might be similar to a work trip).

A total of 46% of shopping trips were incorrectly inferred as other trips: 30% of which as education and 12% as work trips. Similar to the education activities, shopping activities can cover from short to long durations, which can be one of the reasons for the incorrect inferences. Recreational trips have the lowest sensitivity and informedness (according to Fig. 2, they also have the lowest share among other trip purposes in the HTS dataset). Most of the recreational trips were incorrectly inferred as education and shopping trips since education and shopping activities usually have a shorter duration than work activities.

One might argue that adding the spatial attributes (particularly, land use data) to the model could increase the sensitivity and informedness of the model. However, in addition to the explained reasons in the Introduction section and the outcomes of our model in the Results section, there is one more reason against this argument: public transit networks are in urban areas, where land use types are mixed (a university is located next to a business area (or a university is a place that can be utilized for both education and work purposes), or parkland is located in a residential area). It is challenging to allocate trips of a public transit stop to a specific land use type because there are many factors involved (such as distance from the surrounding facilities, or mixed use in multi-story facilities, or

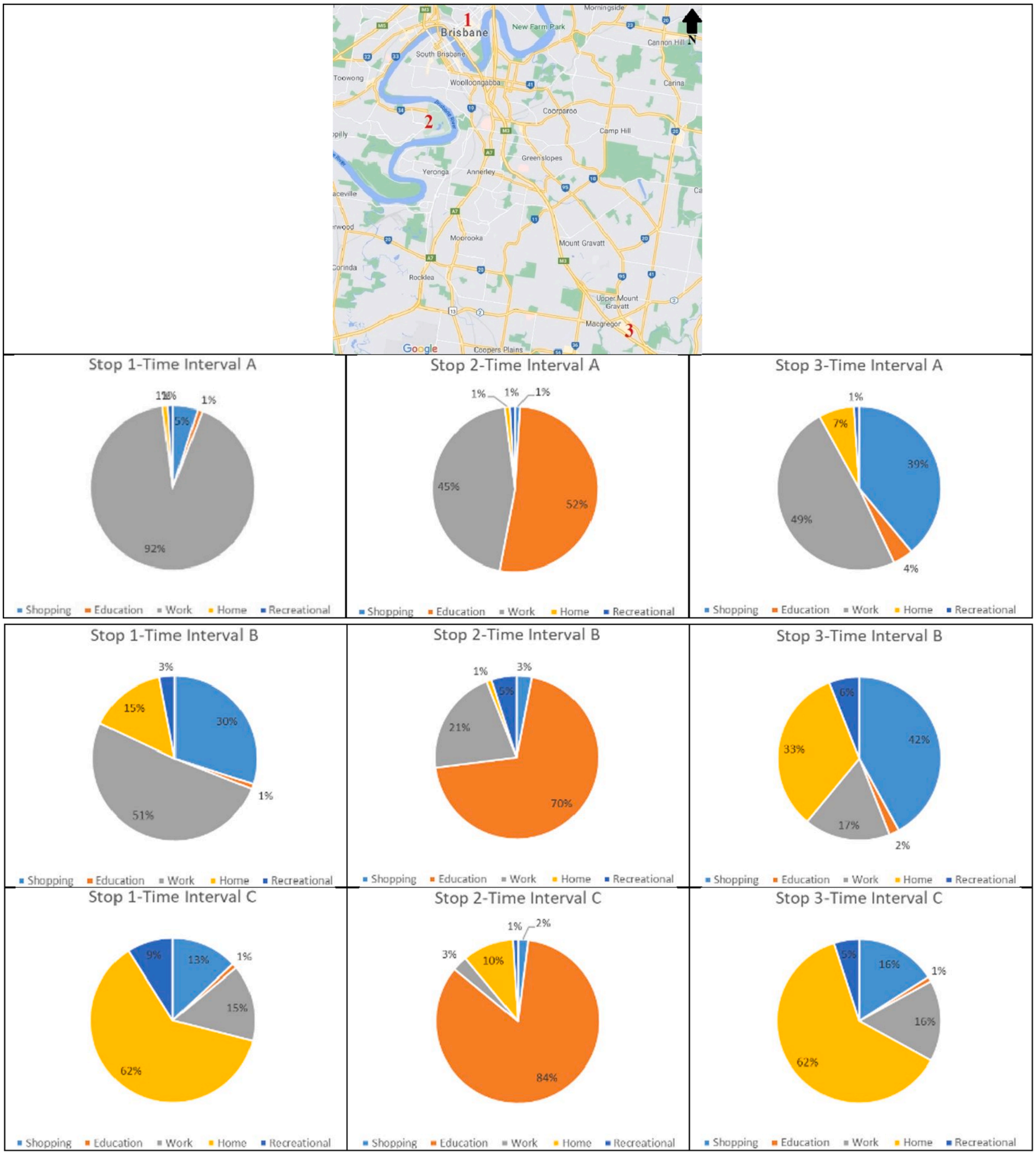

Fig. 8. Results for three public transit stops.

Table 2
Evaluation results.

Trip Purpose	Sensitivity	Accuracy	Informedness
Work	99%	93%	90%
Home	99%	99%	99%
Education	61%	93%	58%
Shopping	54%	94%	51%
Recreational	7%	95%	5%

Table 3
Error analysis.

Actual trip purpose	Inferred trip purpose	Actual trip purpose	Inferred trip purpose
Work	Work 99%	Education	Work 21%
	Home 1%		Home 1%
	Education 0.1%		Education 61%
	Shopping 0.1%		Shopping 16%
	Recreational 0%		Recreational 1%
Actual Trip Purpose	Inferred Trip Purpose	Actual Trip Purpose	Inferred Trip Purpose
Home	Work 1%	Shopping	Work 12%
	Home 99%		Home 0.1%
	Education 0.1%		Education 30%
	Shopping 0.1%		Shopping 54%
	Recreational 0%		Recreational 4%
Actual Trip Purpose		Inferred Trip Purpose	
Recreational		Work 9%	
		Home 0.1%	
		Education 39%	
		Shopping 45%	
		Recreational 7%	

capacity of facilities). One recent research (Faroqi et al., 2021) focused on including some of these effective factors and formulating the possibility of choosing a facility in the catchment area of a public transit stop, however, such examples still need to be improved to include further effective factors. Also, in the next paragraph for a clearer comparison, the results of our model are compared with a model that considers both spatial and temporal attributes in the inferring method.

To compare the improvement of the proposed method in comparison with the existing literature, the results are compared with Alsger et al. (2018) that has been one of the recent and most cited papers on this topic during the last two years. While Alsger et al. (2018) used spatial and temporal attributes for inferring the trip purposes, our study and Alsger et al. (2018) both use similar time intervals, HTS dataset, and trip purpose types. The outcome of Alsger et al. (2018) is compared with the outcome of this research. Alsger et al. (2018) reported the values only for the sensitivity evaluation measure in their paper, which are presented in Table 4. Numbers in parentheses in Table 4 show the difference between Alsger et al. (2018) and our work. It should be mentioned that the numbers in the parentheses are not from Alsger et al. (2018) and just show the improvement compared to that study. All five trip purpose types had higher sensitivity values in the proposed method of this research. The biggest improvement is in education trip purpose with 39%, which could imply the fact that it is easier to separate a student from an employee by considering their chain of trips. However, the recreational trip purpose stays at a very low sensitivity value in both methods.

4. Conclusions

This study proposes a clustering method to infer the trip purpose in smart card datasets. The method clusters trip sequences of individuals based on their temporal attributes and labels them with relevant trip purposes. In brief, the proposed methodology fuses the HTS and smart card datasets based on the common temporal attributes to enrich the smart card dataset with the trip purpose attribute. The similarity between the trip sequences is measured by the similarity between the end time and the time gap between the trips. The relation between the temporal and trip purpose attributes is learnt from the HTS dataset of southeast Queensland, Australia. The labelled clusters were run on one day of smart card data of the same case study area. It is notable that the effect of considering the trip sequence was more significant than considering the land use variables. Furthermore, the proposed model is applicable in other areas, subject to data availability and adjustments to the characteristics of the public transit network and datasets.

The major scientific contribution of this study is developing a novel approach to using trip sequences (instead of separate trips) of individuals to learn their behaviour and infer the trip purpose attribute. Also, a generic similarity concept (Jaccard) is translated into the context of trip sequences. Besides, the results show an improvement in the "sensitivity" evaluation measure when compared with the existing literature. The main practical contribution of inferring the trip purpose attribute can be considered as a step forward towards enriching passively collected mobility datasets, which eventually would be useful for comprehensive transport planning purposes. Also, enriching the smart card dataset can initiate a new wave of innovative applications in the public transit network such as targeted advertising.

One limitation of this study relates to the available HTS dataset that only covers trips of individuals during one specific day. Upon availability of datasets that cover longer periods, it should be possible to extend the trip sequences for periods longer than one day (such as one week or one month) or to add an extra temporal attribute emphasizing the day of the week. Besides, one possible source of error in the proposed method can be discretising the time dimension. While discretising the time dimension (using time windows) is a common approach in the literature, it might be a source of errors: when two trips end within a short time gap but in two different time windows. E.g. in the proposed time windows in this study, one trip could end at 8:59 and another one at 9:01; these two trips would be put into two different time windows despite having an only two-minute difference. Possible effects of this inherited disadvantage of using time windows should be further investigated in the future studies. Also, future studies can investigate developing probabilistic methods instead of choosing the most repeated pattern as the label of the discovered clusters. Furthermore, one possible source of

Table 4
Evaluation measure from Alsger et al. (2018).

Trip purpose	Sensitivity
Work	92% (−7%)
Home	96% (−3%)
Education	22% (−39%)
Shopping	46% (−8%)
Recreational	6% (−1%)

errors in our proposed method could be assigning a trip purpose to an area with no suitable facility for that type of activities (assigning a work trip to a purely residential area). Future studies may further investigate the possibility of filtering the final outcomes of our proposed method by considering such cases.

CRediT authorship contribution statement

Hamed Faroqi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Mahmoud Mesbah:** Conceptualization, Data curation, Methodology, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alexander, L., Jiang, S., Murga, M., Gonzalez, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C: Emerging Technol. 58, 240–250.

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

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., Hickman, M., 2018. Public transport trip purpose inference using smart card fare data. Transp. Res. Part C: Emerging Technol. 87, 123–137.

Aranganayagi, S., Thangavel, K., 2007. Clustering categorical data using silhouette coefficient as a relocating measure. In: International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) (Vol. 2, pp. 13-17). IEEE.

Chakirov, A., Erath, A., 2012. Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport. Institute for Transport Planning and Systems, Eidgenössische Technische Hochschule Zürich, IVT.

Devillaine, F., Munizaga, M., Trépanier, M., 2012. Detection of activities of public transport users by analyzing smart card data. Transport. Res. Record: J. Transport. Res. Board 48–55.

Faroqi, H., Mesbah, M., Kim, J., 2018a. Applications of transit smart cards beyond a fare collection tool: a literature review. Adv. Transp. Stud. 45.

Faroqi, H., Mesbah, M., Kim, J., 2018b. Inferring socioeconomic attributes of public transit passengers using classifiers. Proceedings of the 40th Australian Transport Research Forum (ATRF).

Faroqi, H., Mesbah, M., Kim, J., 2019. Behavioural advertising in the public transit network. Res. Transp. Business Manage. 32, 100421.

Faroqi, H., Mesbah, M., Kim, J., 2020. Modelling socioeconomic attributes of public transit passengers. J. Geogr. Syst. 22 (4), 519–543.

Faroqi, H., Moeckel, R., Mesbah, M., 2021. Temporal distribution of sociodemographic characteristics at transit stops. Transp. Planning Technol. 1–14.

Hasan, S., Schneider, C.M., Ukkusuri, S.V., Gonzalez, M.C., 2013. Spatiotemporal patterns of urban human mobility. J. Stat. Phys. 151, 304–318.

Jun, C., Dongyuan, Y., 2013. Estimating smart card commuters origin-destination distribution based on APTS data. J. Transport. Syst. Eng. Inform. Technol. 13, 47–53.

Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. Transp. Res. Part C: Emerging Technol. 46, 179–191.

Lee, S.G., Hickman, M., 2014. Trip purpose inference using automated fare collection data. Public Transport 6, 1–20.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transp. Res. Part C: Emerging Technol. 19 (4), 557–568.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Robinson, S., Narayanan, B., Toh, N., Pereira, F., 2014. Methods for pre-processing smartcard data to improve data quality. Transp. Res. Part C: Emerging Technol. 49, 43–58.

Tavassoli, A., Alsger, A., Hickman, M., Mesbah, M., 2016. How close the models are to the reality? Comparison of transit origin-destination estimates with automatic fare collection data. Australasian Transport Research Forum. Melbourne, Australia.

Zaki, M.J., Meira, W., 2014. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2020. Discovering latent activity patterns from transit smart card data: A spatiotemporal topic model. Transp. Res. Part C: Emerging Technol. 116, 102627.

Zhou, J., Murphy, E., Long, Y., 2014. Commuting efficiency in the beijing metropolitan area: an exploration combining smartcard and travel survey data. J. Transport Geography. 41, 175–183.

Zhu, L., Ma, B., Zhao, X., 2010. Clustering validity analysis based on silhouette coefficient. J. Comput. Appl. 30 (2), 139–141.

Zou, Q., Yao, X., Zhao, P., Wei, H., Ren, H., 2018. Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. Transportation 45, 919–944.