1-1-2019

Estimating Travellers' Trip Purposes using Public Transport Data and Land Use Information

Bo Du
University of Wollongong, bdu@uow.edu.au

Estimating Travellers' Trip Purposes using Public Transport Data and Land Use Information

Abstract

In public transport system, the equipped automated fare collection (AFC) system records travellers' spatial and temporal information and generates a mass of data daily with more than ever attraction of interest and attention from both academics and practitioners. Advances in data availability and data mining techniques provide great opportunity to investigate various researches in an efficient and effective manner. A comprehensive literature review on the application of public transport smart card data before 2011 can be referred to [1]. As some relevant studies in recent years, [2] proposed a data fusion method to infer passengers' behavioral attributes of the trips based on the naive Bayes classifier model. The proposed method was applied to a single railway station in Osaka, with boarding/alighting information recorded by smart card and validation using trip survey data. [3] applied a unsupervised machine learning method, continuous hidden Markov model, to imputing the missing activities for each trip chain with integration of both clustering and transition models. [4] conducted a comparison on OD matrices between survey data and smart card data, and showed that both trip demands showed high correlation, which implied that the latter might provide a more efficient while less expensive way to construct the OD matrices. As is well known, traditional survey serves as the major method to gather useful trip information for a long time, but it often takes high expense of manpower, time and monetary resources. Moreover, the gap between real trips and survey results can never be ignored. This study aims to investigate various travel purposes of the public transit passengers and develop a data analysis framework to estimate the trip purposes, which can be considered as an alternative or a complementarity to the traditional survey method.

Keywords

estimating, information, travellers', trip, purposes, public, transport, data, land

Disciplines

Engineering | Physical Sciences and Mathematics

Publication Details

Du, B. (2019). Estimating Travellers' Trip Purposes using Public Transport Data and Land Use Information. Tenth Triennial Symposium on Transportation Analysis (TRISTAN X) (pp. 1-4).

Estimating Travellers' Trip Purposes using Public Transport Data and Land Use Information

Bo Du

SMART Infrastructure Facility
University of Wollongong, Wollongong, NSW, 2522, Australia

Email: bdu@uow.edu.au

1 Introduction

In public transport system, the equipped automated fare collection (AFC) system records travellers' spatial and temporal information and generates a mass of data daily with more than ever attraction of interest and attention from both academics and practitioners. Advances in data availability and data mining techniques provide great opportunity to investigate various researches in an efficient and effective manner. A comprehensive literature review on the application of public transport smart card data before 2011 can be referred to [1]. As some relevant studies in recent years, [2] proposed a data fusion method to infer passengers' behavioral attributes of the trips based on the naive Bayes classifier model. The proposed method was applied to a single railway station in Osaka, with boarding/alighting information recorded by smart card and validation using trip survey data. [3] applied a unsupervised machine learning method, continuous hidden Markov model, to imputing the missing activities for each trip chain with integration of both clustering and transition models. [4] conducted a comparison on OD matrices between survey data and smart card data, and showed that both trip demands showed high correlation, which implied that the latter might provide a more efficient while less expensive way to construct the OD matrices.

As is well known, traditional survey serves as the major method to gather useful trip information for a long time, but it often takes high expense of manpower, time and monetary resources. Moreover, the gap between real trips and survey results can never be ignored. This study aims to investigate various travel purposes of the public transit passengers and develop a data analysis framework to estimate the trip purposes, which can be considered as an alternative or a complementarity to the traditional survey method.

2 Data Description

Singapore has a population of 5,399,000 and the major public transport forms consist of mass rapid transit (MRT), light rail transit and bus. Since the ez-link card launching in 2008, it remains the number one choice to pay transprot fare. To better illustrate passengers' travel purposes and departure features, the timeline of a single day is divided into six ranges according to the average daily ridership using public transport: early morning [4:00-7:00], morning peak [7:00-9:00], inter peak [9:00-17:00], evening peak [17:00-20:45], early night [20:45-23:00] and late night [23:00-4:00(+1day)].

To estimate trip purpose, the land use information of catchment areas of MRT stations plays an important role. In this study, five aggregated land use types are selected: commercial, residential,

business, education and others. In this study, commercial type represents locations open for customers like shopping mall and cinema; while business type represents the workplace, office, industrial factory, and so on. The proportion of each land use type at the catchment areas (circular coverage with station as center, 500m as radius) of the station is estimated based on "Singapore Master Plan 2014". With the proportion estimation of various land use features, they can be applied to replacing the alighting stations to reflect the characteristics of trip purpose. The illustration of the replacement procedure is shown in Table 1 as follows.

Table 1. Replacement of alighting station with its surrounding land use features

Trip Date	Borarding Time	Alighting Time	Alighting Station				
2013-08-12	7:40:42	7:57:27	HarbourFront				
			↓				
Trip Date	Borarding Time	Alighting Time	Commercial	Residential	Business	Education	Others
2013-08-12	7:40:42	7:57:27	65%	30%	5%	0%	0%

For illustration purpose, two weeks' smart card data of MRT North-East line (NEL) is adopted in this study. To avoid fluctuation, only data between Monday and Thursday is extracted, thus eight working days' data is adopted for analysis. Besides the five land use attributes illustrated in Table 1, the other three temporal attributes derived from smart card data include: average duration between trips, first trip start time range and last trip start time range. These five attributes are used to derive passengers' trip purposes based on a clustering method, which is introduced in the subsequent section.

3 K-prototypes Algorithm

The K-means algorithm is well known for its efficiency and simplicity, however, working only on numeric values prohibits it from being used to cluster real world data containing categorical. Since our sample data has six numeric attributes (average duration between trips, commercial, residential, business, education and others) and two categorical attributes (first trip start time range and last trip start time range), the K-prototypes algorithm is employed to handle data with mixed numeric and categorical characteristics. More details regarding the formulation can be referred to [5]. The procedure of K-prototypes algorithm is illustrated as follows:

[Step 1] Pre-given or randomly choose centroids;

[Step 2] Put each data point to its nearest centroid as a cluster based on the mixed measurement;

[Step 3] Re-calculate the centroid of each cluster based on its current data points;

[Step 4] Repeat 2 and 3 until the centroids no longer move or the iteration limitation is reached.

4 Experimental Results

With 16 stations spanning 20km, the NEL in Singapore plays an important role in weaving through the heart of the city, HarbourFront and heritage areas, such as Chinatown and Clarke Quay, through to the residential estates like Sengkang and Punggol. It is a typical MRT line with the coverage of miscellaneous trip purposes, like education, residential, work, entertainment and tourism. The goal of this study is to generate clusters with similar trip purposes. The clustering results in Table 2 shows that no extreme clusters exist, and most of the clusters are in similar size except Cluster 1's size is

relatively bigger. The columns of the result table refer to the clusters, and the first three rows indicate temporal features while the next five rows represent land use features.

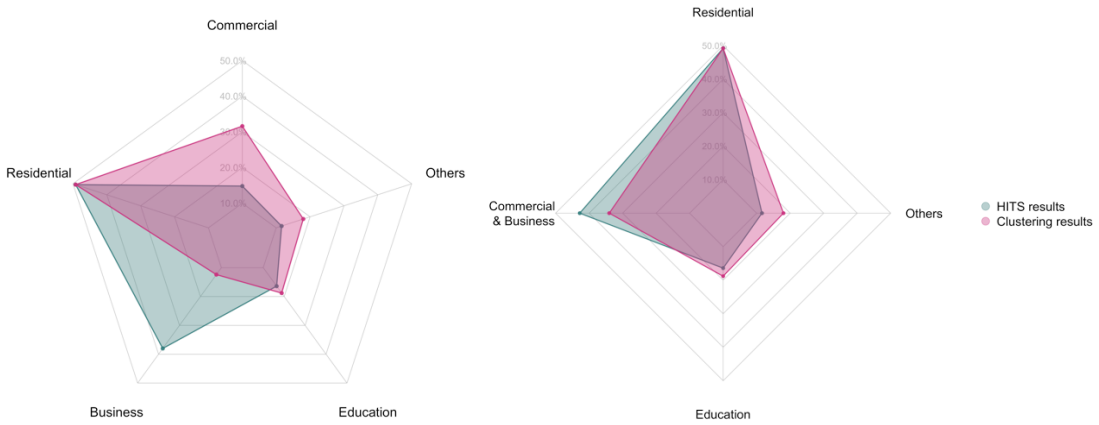
Table 2. Clustering results of trip purposes

	Cluster 1 (N= 686)	Cluster 2 (N= 299)	Cluster 3 (N= 320)	Cluster 4 (N= 339)	Cluster 5 (N= 356)
First trip start time range (average start time)	Early morning (5:30am)	Morning peak (8:00am)	Inter peak (1:00pm)	Inter Peak (1:00pm)	Inter peak (1:00pm)
Last trip start time range (average start time)	Evening peak (6:52pm)	Evening peak (6:52pm)	Inter peak (1:00pm)	Evening peak (6:52pm)	Early night (9:52pm)
Average duration (hr)	10.7	10.4	3.2	7.9	11.2
Commercial	**32.5%**	10%	**30%**	**34.7%**	**24.5%**
Residential	**47.5%**	**52.5%**	**46.3%**	**45%**	**52.5%**
Business	0%	**17.5%**	0%	0%	0%
Education	10%	**12.5%**	**11.2%**	10%	10%
Others	10%	7.5%	**12.5%**	10%	**12.5%**
Major purposes	Commercial, Residential	Residential, Business, Education	Commercial, Residential, Education, Others	Commercial, Residential	Commercial, Residential, Others

Cluster 1 shows the temporal features with first trip starting at early morning (average start time at 5:30am), last trip staring at evening peak (average start time at 6:52pm) and average duration of 10.7hr between trips, as well as land use features of residential and commercial mainly. Therefore we can infer that passengers in Cluster 1 usually travel between their residential locations and commercial areas. Similarly, Cluster 2 represents trip purposes mainly on residential, business and education. The trips of Cluster 3 are all within inter peak with short duration between trips, which indicates that the travelers in this cluster might be flexible travelers with mixed-type trip purposes and flexible schedules rather than regular commuters, and they often travel within short distance and short duration between trips. Cluster 4 represents similar trip purposes as Cluster 1, however the first trip usually happens during inter peak with shorter duration between first and last trips. Specially, tourism forms a large part of the economy (over 15 million tourists in 2014) in Singapore, therefore passengers in this cluster may include tourists. The main trip purposes in Cluster 5 include residential, commercial and others, and the trips generate later than those in the other four groups.

To validate the proposed clustering method, Household Interview Travel Survey (HITS) data with trips along NEL is used as reference. The comparison results are shown in Fig. 1. In Fig. 1(a), the HITS data was aggregated in line with the five categories in this research. However, it is worth mentioning that it is difficult to figure out the definitions of all the trip purposes in HITS data and to aggregate the trip purposes following exactly the same way in this research, hence we could notice significant difference on certain land use types, like commercial and business. In this study, commercial type represents locations open for all customers, like shopping mall. In this case, people travel to such places can either be customers or workers in those places. However, workers may belong to business type in HITS data, thus it is difficult to give a clear border between commercial and business types. In this case, commercial and business types were merged in Fig. 1(b), and we can observe similar proportions of the land use features. In general, the NEL mainly serves as a connection between residential areas and areas with business and commercial activities, which include the most famous sightseeing places in Singapore, like Sentosa, Chinatown, and so on. Similarly, many schools can be identified along this line, which explains a certain proportion

of education. All other trip purposes have been included in others type, which can be different from the definition in HITS data, as shown in the figure.



(a) Comparison based on five category (b) Comparison based on four category

Fig.1. Comparison between HITS results and clustering results

5 Conclusions

With the aid of land use information, the smart card data was analyzed to estimate passengers' trip purposes. Three temporal attributes (average duration between trips, first trip start time range and last trip start time range) and five land use attributes (commercial, residential, business, education and others) were adopted. A K-prototypes algorithm for mixed-type data was applied to obtaining the clustering results. With a MRT line in Singapore as case study, five clusters were identified to represent heterogeneous trip patterns and purposes. The proposed data analysis framework is expected to be regarded as a useful tool to impute passengers' purposes, as an alternative or a complementarity to the traditional survey approach. As a future work, numerical experiments on a large-scale public transport network will be conducted, and land use features will be adjusted to keep in line with various trip purposes in HITS data for more comprehensive and reasonable comparison.

Acknowledgements

The author thanks Land Transportation Authority of Singapore for providing relevant data support.

References

[1] M.P. Pelletier, M. Trépanier, C. Morency, "Smart card data use in public transit: A literature review", *Transportation Research Part C: Emerging Technologies* 19(4), 557-568 (2011).

[2] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach". *Transportation Research Part C: Emerging Technologies* 46, 179-191 (2014).

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

[4] C. Pineda, D. Schwarz, E. Godoy, "Comparison of passengers' behavior and aggregate demand levels on a subway system using origin-destination surveys and smartcard data". *Research in Transportation Economics* 59, 258-267 (2016).

[5] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values". *Data Mining and Knowledge Discovery* 2(3), 283-304 (1998).