ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/274009746

Passenger Segmentation Using Smart Card Data

Article *in* IEEE Transactions on Intelligent Transportation Systems · December 2013

DOI: 10.1109/TITS.2014.2368998

CITATIONS
6

READS
255

3 authors:

Le Minh Kieu
Queensland University of Technology
13 PUBLICATIONS 28 CITATIONS
SEE PROFILE

Ashish Bhaskar
Queensland University of Technology
67 PUBLICATIONS 263 CITATIONS
SEE PROFILE

Edward Chung
Queensland University of Technology
151 PUBLICATIONS 762 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Arterial Travel Time Estimation View project

Project Transit Capacity and Quality of Service Research View project

All content following this page was uploaded by Le Minh Kieu on 27 April 2015.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Passenger Segmentation Using Smart Card Data

Le Minh Kieu, Ashish Bhaskar, and Edward Chung

***Abstract*—Transit passenger market segmentation enables transit operators to target different classes of transit users for targeted surveys and various operational and strategic planning improvements. However, the existing market segmentation studies in the literature have been generally done using passenger surveys, which have various limitations. The smart card (SC) data from an automated fare collection system facilitate the understanding of the multiday travel pattern of transit passengers and can be used to segment them into identifiable types of similar behaviors and needs. This paper proposes a comprehensive methodology for passenger segmentation solely using SC data. After reconstructing the travel itineraries from SC transactions, this paper adopts the density-based spatial clustering of application with noise (DBSCAN) algorithm to mine the travel pattern of each SC user. An *a priori* market segmentation approach then segments transit passengers into four identifiable types. The methodology proposed in this paper assists transit operators to understand their passengers and provides them oriented information and services.**

***Index Terms*—Automated fare collection (AFC) system, market segmentation, public transport, smart cards (SCs), transit passenger.**

I. Introduction

BETTER understanding of passengers is essential for transit authorities to satisfy customer needs and preferences. The Transportation Research Board has published a handbook on using market segmentation to increase patronage [1]. Most transit operators have defined classes of customers but not market segments. For instance, operators in South East Queensland (SEQ), Australia, classify passengers into six types (adult, senior, child, pension, secondary school student, and student) according to age and occupation. Although this classification is still useful for fare collection, whether these types differently respond to alternative services and whether new policies benefit them is unknown.

Despite the high exposure to transit passengers, transit providers have limited knowledge about their customers due to reasons such as the anonymity of passengers, the stochasticity of their behaviors, and the complexity in analyzing the disaggregated information of a massive population. Future impacts of the proposed initiatives to existing and latent passengers are obscure due to the lack of knowledge on their mobility requirements and travel behaviors. Existing service improvement projects are limited to the impacts on generic transit customers, neglecting the differences between the types or segments of passengers with different needs and behaviors. Paradoxically, the majority of studies in public transport solely focus on improving a vehicle's performance, such as the travel time or schedule adherence, without a profound understanding of passenger segments and behaviors, notwithstanding the fact that different segments of passengers would behave differently in a transit system. For instance, an irregular transit passenger would be more concerned with the service coverage, i.e., if s/he would be able to travel by public transport to the desired destination, whereas a commuter transit rider user would be more concerned with the on-time performance and the easiness of transfers.

One of the dominant factors that dictate the level of passenger characterization is the availability of data. Traditional studies on passenger travel patterns and passenger segmentation solely focus on the use of transit user surveys [1]–[3]. These surveys are generally expensive to perform, are limited in sample size, and are only valid within the study period. Transit agencies are at a critical transition in data collection technology from manual data collection toward automated data collection systems (ADCSs). Manual data collection systems with a low capital cost but a high marginal cost, small sample sizes, and sometimes unreliable accuracy are being replaced by a low marginal cost, large sample sizes, and disaggregated ADCSs. ADCSs such as automatic vehicle location, automatic passenger counter, automatic vehicle identification, and, particularly, smart card (SC) automated fare collection (AFC) systems have only recently become widely popular for collection and analysis. The proliferation of these modern technologies provides a tremendous opportunity to analyze the existing condition of transit quality of service and facilitates agencies to enhance the service quality. Public transport agencies that are able to take advantage of this massive amount of data to anticipate and actively react to the changes in the transport environment and passenger behaviors would earn an utmost advantage to attract customers. The Brisbane City Council Transport Plan for Brisbane, Australia [4], emphasizes the use of data to ensure that the strategies are effective to augment the transit experience and provides a demand-responsive transit system.

This paper augments the transit passenger characterization by passenger segmentation using the dynamic SC data. The segmentation aims to group passengers of similar travel pattern, i.e., with the same level of transit journeys at regular times and places. The market segmentation of transit passengers brings various benefits to transit authorities to better cater to their customers. A targeted survey could aim for the passengers of

Manuscript received May 24, 2014; revised September 9, 2014; accepted October 26, 2014. This work was supported by Queensland University of Technology. The data were provided by Translink, the transit operator of South East Queensland, Australia. The Associate Editor for this paper was M. M. Dessouky.

The authors are with Smart Transport Research Centre, Queensland University of Technology, Brisbane, Qld. 4000, Australia (e-mail: leminh.kieu@qut.edu.au; ashish.bhaskar@qut.edu.au; edward.chung@qut.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2014.2368998

1524-9050 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

low transit usage to understand the disutility that limits the level of ridership. Before–after studies could observe the changes in the passenger market to understand the evolution of passenger demand. Incentives and personalized service can be given to passengers of regular usage to encourage passengers to use public transport. The observation of the travel pattern also benefits operational strategies such as transfer coordination and origin–destination (OD) demand management by monitoring and inferring passenger movements through their travel habits.

The contribution of this paper is a systematic approach to mine the travel pattern and segment transit passengers solely using SC data. After the literature review in Section II, this paper reconstructs the completed "journey" of SC users from individual SC "transactions" in Section III-B. Each "transaction" includes both the boarding and alighting times and stop IDs of a transit journey between a touch on and a touch off to the ticketing device. Each "journey" is defined as the public transport travel from an origin to a destination, including the transfers, which might include one or several transactions. In Section III-C and D, a density-based clustering algorithm is adopted to mine the travel pattern from each SC user's historical itineraries and to identify the spatial OD that the cardholder usually travels as "regular OD" and the time of regular travels as "habitual time." SC users are segmented into different classes using the mined travel pattern by the *a priori* market segmentation approach in Section III-E. The analysis of segmentation results in Section IV reveals interesting travel behaviors from each segment. Finally, after the discussion of potential practical applications in Section V, the conclusion sums up this paper.

II. Related Studies

The intelligent transportation system AFC system using SCs collects large volumes of individual travel data and facilitates a large-scale, economical, and continuous method to explore the multiday behaviors of transit passengers. An emerging number of recent studies have been published using SC data, where the authors have connected individual SC boarding/alighting records to reconstruct user itineraries [5]–[7]. Most of the studies also added a multiday dimension to explore the travel pattern or repeated travel patterns of each SC user [5], [6], [8]. The literature of travel pattern mining using SC data has evolved from an aggregated to a disaggregated level of passenger analysis. Existing studies have been looking at a general transit passenger (the whole data set) [6], [9] to a group of passengers (passengers of similar characteristics) [10]–[12] and, finally, to individual passengers [5]. The proliferation of computing power is probably the reason for this trend. Although aggregated travel pattern analysis provides insights into the travel pattern of a general user, it fails to capture the individuality of travel behavior. Moreover, the typologies of trips and passengers are predefined, which might not reflect the similarity of passengers between the same class and the difference between classes.

Another trend in travel pattern analysis is the development of pattern discretization. Spatial travel pattern analysis often breaks down to stop-to-stop repeated trips [6], [9], [12]. The limitation of this method has been identified by several authors [13]. A transit stop is usually only linked with a single direction or route, whereas transit passengers normally have several route choice options within their OD locations. Any stops within the immediate vicinity that provide the same access should be considered in the same travel pattern because transit passengers might pick randomly or might pick the first arriving vehicle. Different aggregation approaches have been recently proposed to group spatially close stops into the same travel pattern. Chu and Chapleau [5] aggregated stops within 50 m of each other to form a new node. Lee *et al.* [14] and Lee and Hickman [13] proposed a stop aggregation model to group stops according to their proximity, descriptions, and catchment areas.

The problem of temporal travel pattern analysis has not received much attention as a spatial travel pattern. The existing discretization of time usually breaks down to either a number of predefined time windows (e.g., the 1-h period in [12]) or the time of the day (e.g., A.M. peak, midday, and P.M. peak in [5] and [11]). A temporal pattern is defined if the passenger repeatedly made multiple trips within a time period. It is strenuous to discretize the temporal pattern for individual passengers because different people would have different habitual behaviors. For instance, a 1-h time window may segregate the journeys at 9:59 A.M. to the journeys at 10:01 A.M., although these journeys come from the same temporal behavior.

A decent amount of research has also segmented transit passengers based on travel behaviors for fare elasticity [2] or increasing transit patronage [1], [3]. Elmore-Yalch [1] outlined three major approaches for transit passenger segmentation: 1) physical segmentation based on basic information such as demography, geography, and geodemographics; 2) product usage segmentation based on ridership such as the frequency of use; 3) physiological segmentation based on the characteristic of individual passengers; and 4) benefit segmentation based on passenger requisite. Hensher [2] segmented the transit customer market into four classes of nonconcession and concession passengers traveling long or short trips to estimate fare elasticity. Shiftan *et al.* [3] proposed a structural equation modeling approach to segment transit passengers according to the sensitivity to time, the need for a fixed schedule, and the willingness to use transit.

III. Methodology

This section introduces the data set used for the case study (see Section III-A), as well as the methods for the reconstruction of travel itineraries (see Section III-B), travel pattern analysis (see Section III-C and D), and passenger segmentation (see Section III-E).

A. Data Set

The SC data used in this paper come from Translink, which is the transit authority of SEQ, Australia. The data set is a compilation of around 34.8 million transactions made by a million SCs over 15 000 transit stops of the bus, city train, and ferry networks in SEQ from March 1, 2012 to June 30, 2012. Each transaction contains the following fields.

1) *CardID*: The unique SC ID, which has been hashed into a unique number to maintain the privacy of the cardholder.

2) *T_on*: The timestamp for touch on.
3) *T_off*: The time stamp for touch off.
4) *S_on*: The station ID at touch on.
5) *S_off*: The station ID at touch off.
6) *ValidIndicator*: A binary indicator for differentiating a valid or invalid transaction. It has been used by the operator for ticketing purposes. A valid transaction is the combination of a touch on and a touch off from the same transit line within a 2-h limit [15]. Any cases other than that, e.g., no touch off, touch off at a different line, etc., are indicated as invalid transactions. Only around 3% of the transactions are invalid.
7) *RouteUsed*: The transit line that the passenger has used.
8) *Direction*: The direction of travel (inbound/outbound).
9) *Fare*: The fare paid for the transaction in Australian dollars.

For the current analysis, the study is only performed on working days (weekdays, excluding public holidays and school holidays) because the travel behavior on working weekdays can be significantly different than that on weekends and holidays.

B. Reconstruction of Travel Itineraries

The first step to mine the travel pattern is through reconstructing the travel trip from individual transactions. The flowchart in Fig. 1 illustrates the algorithm to connect the individual transactions from each SC user on each working day into completed trips. The algorithm is built on a binary *ReconstructingIndicator* to identify the ongoing/new trip status and on a *TripID* to differentiate the completed trips.

A fixed threshold of 60 min is then used to decide if the two transactions are connected. This threshold has been differently chosen in literature [7]. Sixty minutes is chosen in accordance with Brisbane's public transport threshold for transferring trips [15].

Here, the first boarding stop and the last alighting stop of a completed trip are defined as the "origin stop" and the "destination stop," respectively. The gap between the alighting time of a transaction and the boarding time of the next transaction of the same trip is defined as the transferring time. The following four steps describe the trip construction process.

1) STEP 1: A binary *ReconstructingIndicator* is defined and assigned as 0.
2) STEP 2: The *ValidIndicator* is checked. If the indicator is equal to 0 (which denotes an invalid transaction), the corresponding trips will be discarded.
3) STEP 3: If the *ReconstructingIndicator* is 0, a variable *OriginLocation* is defined and set as equal to the current *T_on*. We also assign a new unique *TripID*, change the *ReconstructingIndicator* to 1, save the current transaction, and move to the next transaction.

   If the *ReconstructingIndicator* is 1 and the time gap between the current *T_on* and the last *T_off* is less than 60 min, we move to Step 4.

   If the time gap is more than 60 min, the transaction with the previous *TripID* is connected into a completed trip. A new *TripID* and a new *OriginLocation* are assigned, in which the current transaction is identified as the first leg from the origin stop. The *ReconstructingIndicator* is set as 1.
4) STEP 4: If the current *S_off* is different to the *OriginLocation*, the transaction is connected to the trip as a continuation journey. If it is also the last transaction of the day, the trip reconstruction process for the study passenger is finished; otherwise, we move to the next transaction.

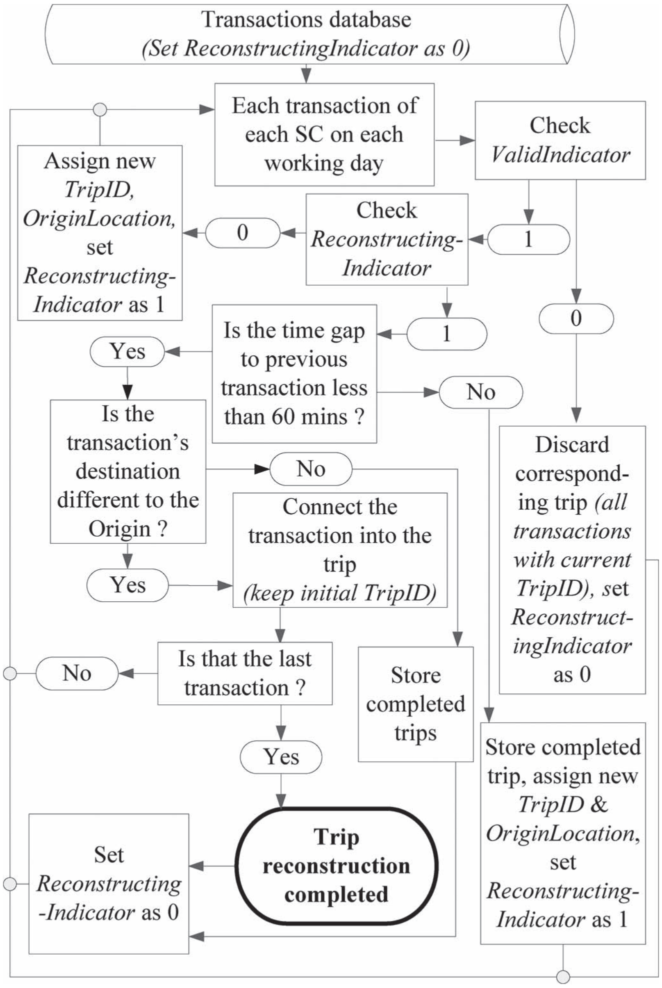

Transactions database
(Set ReconstructingIndicator as 0)
Each transaction of each SC on each working day
Check ValidIndicator
Assign new TripID, OriginLocation, set Reconstructing-Indicator as 1
0
Check Reconstructing-Indicator
1
0
Yes
Is the time gap to previous transaction less than 60 mins ?
1
No
Is the transaction's destination different to the Origin ?
No
Connect the transaction into the trip (keep initial TripID)
Yes
Discard correspond-ing trip (all transactions with current TripID), set Reconstruct-ingIndicator as 0
No
Is that the last transaction ?
Store completed trips
Yes
Store completed trip, assign new TripID & OriginLocation, set Reconstructing-Indicator as 1
Set Reconstructing -Indicator as 0
Trip reconstruction completed

Fig. 1. Trip reconstruction flowchart.

C. Mining Spatial and Temporal Pattern From Travel Itineraries

This section presents the method to mine the spatial and temporal travel pattern from the historical trip database. The spatial OD stops are represented as geographical coordinates (geographical position), whereas the temporal boarding and alighting times are represented as timestamps. We adopt a density-based clustering algorithm because of the following reasons.

1) Density-based algorithms identify clusters of high density and noise of low density. In travel pattern analysis, noise is an anomaly travel pattern that does not follow any regular travel pattern or, in other words, trips that are

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

randomly made. Our goal is to find the clusters (regular pattern) and differentiate it with the anomaly pattern.

2) Density-based algorithms can identify a cluster of any shape and size. A travel pattern could also form any shape and size due to its nature of human behavior pattern.
3) Density-based algorithms do not require the predetermination of initial cores or the number of clusters. This feature is also essential for travel pattern analysis because the number of patterns from an individual passenger is unknown.
4) Discretization is a major concern in travel pattern analysis. The existing literature has showed that there is a need for a systematic and flexible solution to spatial and temporal pattern analysis without limiting to stop-to-stop repeated trips and time-window discretization. Density-based scanning algorithms systematically produce a flexible range of high density for each passenger's spatial and temporal travel pattern.

A decent number of density-based clustering algorithms such as the density-based spatial clustering of application with noise (DBSCAN) [16] and more complex methods such as ordering points to identify the clustering structure (OPTICS) [17] and density-based clustering (DENCLUE) [18] can be found in literature. DBSCAN is then chosen as the algorithm to use in this paper because of its high computing performance to handle a large data set with over a million SC users and because it has all of the four aforementioned features of a density-based clustering algorithm.

1) DBSCAN Algorithm: The DBSCAN algorithm defines clusters as dense regions, which are separated by regions of a lower point density. The algorithm has two global parameters: the maximum density reach distance ε and the minimum number of points MinPts. A point can be considered a "core point" i_c if it has at least MinPts (density) within a radius ε, as expressed in

$$\left|N_{\varepsilon(i_c)}\right| \geq \text{MinPts} \tag{1}$$

where $N_{\varepsilon(i_c)} : \{i \text{ points in the data set} \mid d(i_c, i) \leq \varepsilon\}$.

$N_{\varepsilon(i_c)}$ is the number of points i in the data set that has a distance to i_c that is $d(i_c, i)$ less than ε. The most common distance metric used is the Euclidean distance.

A point can be considered a "border point" i_b if it has fewer points than MinPts within ε but lies within the range ε of a core point. A point is considered a "noise point" i_n if it is neither a core nor a border point. A cluster is defined by combining the core points i_c that are not more than ε distance apart, along with their associated border points i_b.

For a more detailed description of DBSCAN, see [16]. The algorithm is separately applied for mining the spatial and temporal patterns, in which the regular ODs are derived by a two-level DBSCAN application: first on the historical alighting stops and second on the boarding stops. The order of the two levels is interchangeable without changing the results. The separate application of DBSCAN increases the robustness of the overall clustering algorithm, and the outcomes of each level are useful for later passenger segmentation.

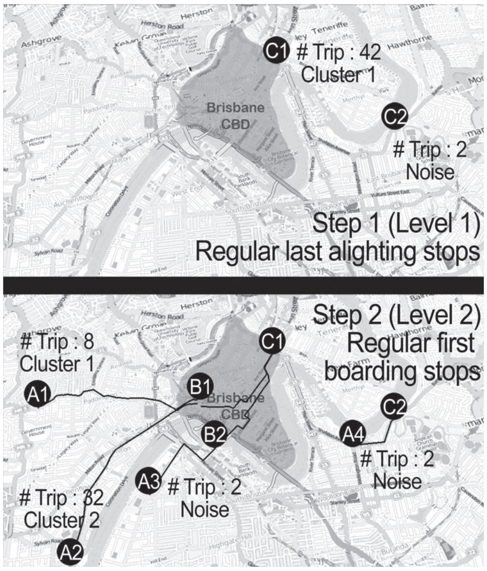


Fig. 2. Two-step DBSCAN application for regular OD mining.

2) Mining Spatial Travel Pattern (Regular OD): This section describes the mining process for regular ODs. A two-step procedure is applied to separately mine the regular last alighting and first boarding stops of each SC user. Fig. 2 illustrates the process of mining the travel pattern of an SC user on morning trips as an example for explaining the clustering method. Here, the A points represent the first boarding stops; the C points, the last alighting stops; and the B points, the transfer stops in the SC user's historical itineraries. The two levels of DBSCAN application are described in the following steps.

Level 1: The first level of DBSCAN only groups the last alighting stops (the C points). It is important to notice that, for underlying the recurring patterns, each trip's last alighting stop is considered a point in the database. In Fig. 2, the 42 trips made at the same stop C1 form 42 points at the same coordinates. The DBSCAN algorithm with ε equals 1000 m and MinPts equals 8 is applied to define Cluster 1 at stop C1 and the other two points as the anomaly pattern.

Level 2: Now, if we locate the origin stops (Stop A) and the transfer stops (Stop B), the travel pattern can be identified. The second level of the DBSCAN algorithm only groups the origin stops (the A points). The same algorithm is used to identify two clusters of boarding stops at A1 and A2.

If both the origin stop and the destination stop are not anomaly patterns, the corresponding OD is identified as a regular OD. In our example, the OD pairs A1–C1 and A2–C1 are regular ODs.

3) Mining Temporal Travel Pattern (Habitual Time): This section presents the application of DBSCAN to mine the habitual time, i.e., the time an SC user habitually boards a transit vehicle. Each journey has been stored as a timestamp record, i.e., minutes from midnight (0:00), e.g., a timestamp of 480 min is 8 A.M. The existing studies in the temporal travel pattern analysis would break down the time axis to either a number of predefined time windows (e.g., 1 h) or the time of the day (e.g., A.M. peak, midday, and P.M. peak). These time windows cannot suit everyone because each passenger has

TABLE I
EXAMPLE OF TRAVEL PATTERN: (a) REGULAR OD AND (b) HABITUAL TIME

(a)

SC ID	Regular OD ID	% Regular Trip	Origin Stop ID	Destination Stop ID	Number of Trips	Route ID sequence
X1	1	43.02	5198	5210	37	420
	2	31.40	5873	5198	27	458
X2	1	13.56	4364	1882	8	999→370
	2	54.24	8	1882	32	999→370
	3	20.34	1882	8	12	370→999
X3	1	21.28	1878	2890	10	543→141 →139
	2	17.02	1878	2888	8	542→141 →169

(b)

SC ID	Habitual time ID	% Habitual Trip	Habitual time pattern (min from 0)	Number of Trips
1	1	48.84	486.42	42
	2	45.35	1062.32	39
X2	1	37.29	398.13	22
	2	30.51	519.06	18
	3	18.64	956.6	11
X3	1	36.17	551.85	17

different travel behaviors. More importantly, a temporal pattern could be separated at the border of two time windows.

DBSCAN offers a systematic and flexible method to find the area of high density in a 1-D time axis. For this paper, we applied DBSCAN in the temporal travel pattern analysis, with ε equals 5 min and MinPts equals 6. We would see, for instance, a passenger with two journeys at 8:00 A.M., three journeys at 8:04 A.M., and one journey at 7:56 A.M. that were grouped into a temporal pattern. Given that the reliability of the travel time during peak periods can have an impact on the alighting time, the boarding time is chosen instead of the alighting time for the DBSCAN application because the SC users can actively choose the boarding time but the time when they arrive at the destination.

Table I presents an example of three SC users' travel regularity, in which passenger X2 was chosen as the example in Fig. 2.

D. Sensitivity Analysis of MinPts and ε

The application of DBSCAN requires two important parameters: the minimum number of boardings MinPts and density reach distance ε.

For the spatial travel pattern analysis, the maximum density reach distance ε denotes the walking distance of the passenger from one to another stop of the same boarding pattern. ε can be found by a travel survey and varies from case to case. Burke and Brown [19] found that people in Brisbane and Perth, Australia, significantly walk longer than people in U.S. cities, where a rule of thumb of 500 m has been usually used to measure the preferable walking distance to transit stops [20]. If ε increases, the algorithm would define more stops as regular. ε should not be too large since the OD of the transit trip might be clustered into the same boarding pattern if ε is larger than the travel distance.

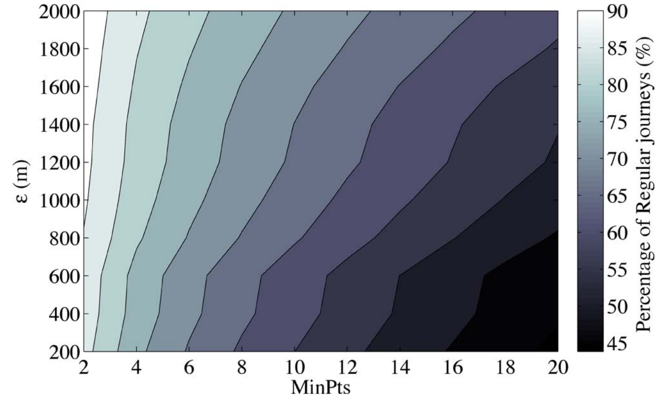


Fig. 3. MinPts and ε sensitivity analysis for regular OD mining.

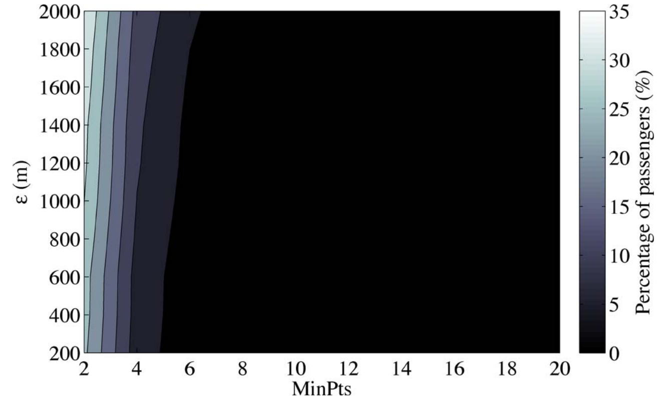


Fig. 4. Percentage of passengers with a regular OD but who traveled less than ten times.

The examination of MinPts can be broken down into how transit operators define the travel pattern. Given the number of boardings over a study period, MinPts is equal to the minimum boarding made to be considered "regular." For instance, a value of MinPts equal to 2 means that any repeated boarding will be considered regular. Fig. 3 illustrates the MinPts and ε sensitivity analysis results.

The percentage of regular journeys noticeably increases when ε increases from 400 to 600 m because most passengers would prefer walking within these distances. Another significant increase could be seen when ε exceeds 1200 m, where the OD is grouped into the same pattern. The value ε that is chosen for this paper is 1000 m.

We choose the value of MinPts to maximize the proportion of the regular travel pattern, but conversely, the algorithm should minimize the proportion of passengers who rarely travel but are still being assigned with a regular pattern because these behaviors could be unreliable. Fig. 4 demonstrates the number of passengers with a regular OD but who traveled less than ten times during the four-month study period. Given that ε has been chosen as 1000 m, MinPts has been chosen as 8 for the case study.

The parameters for the temporal travel pattern analysis have been chosen by a similar approach. Fig. 5 shows the percentage of habitual journeys for different values of MinPts and ε.

For the temporal travel pattern analysis, the maximum density reach distance ε denotes the variability of boarding times

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

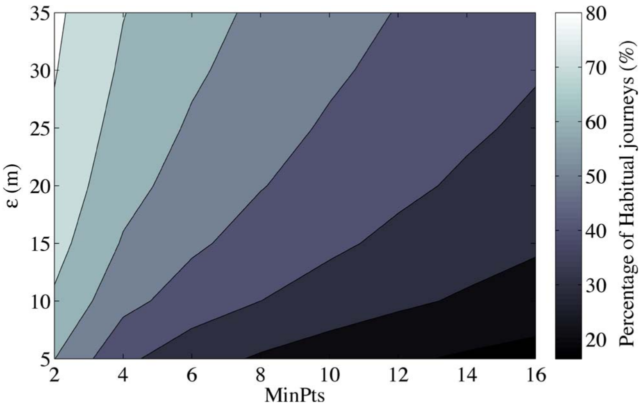


Fig. 5. MinPts and ε sensitivity analysis for habitual time mining.

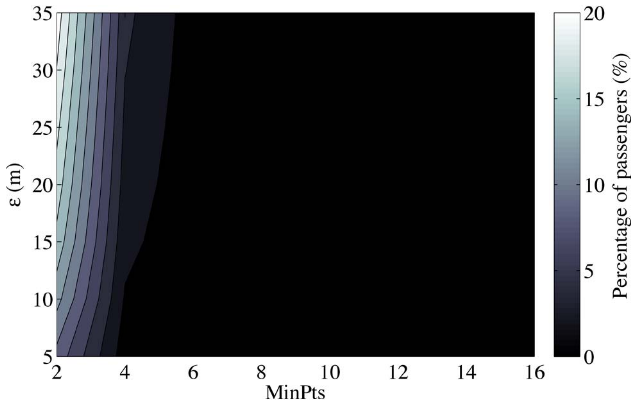


Fig. 6. Percentage of passengers with a habitual time but who traveled less than ten times.

within the same travel pattern. ε has been chosen as 5 min to allow some variability in vehicle arriving times. MinPts has been similarly chosen by the same logic as the spatial pattern analysis. Fig. 6 demonstrates the number of passengers with a habitual time but who traveled less than ten times during the four-month study period.

E. A Priori Market Segmentation Analysis for Transit Passenger Segmentation

The market segmentation analysis follows *a priori* segmentation where identifiable passenger classes are selected from the SC user population based on the proportion of regular OD/habitual time trips in the total transit usage. In the *a priori* market segmentation, the cluster-defining descriptions are selected in advance by the researcher, and conducting the study will not influence the definitions of these predefined segments [1]. The *a priori* segmentation is based on the assumption that there are stereotypes about different classes. The segmentation approach in this paper could be classified as a physiological segmentation [1], where passenger travel characteristics, i.e., spatial and temporal travel patterns, define the type of passenger. Four segments of passengers can be identified as follows.

1) Passengers with a regular OD but without a habitual time are hereafter called *regular OD passengers*. They have regular places to travel but are flexible in terms of the traveling time.

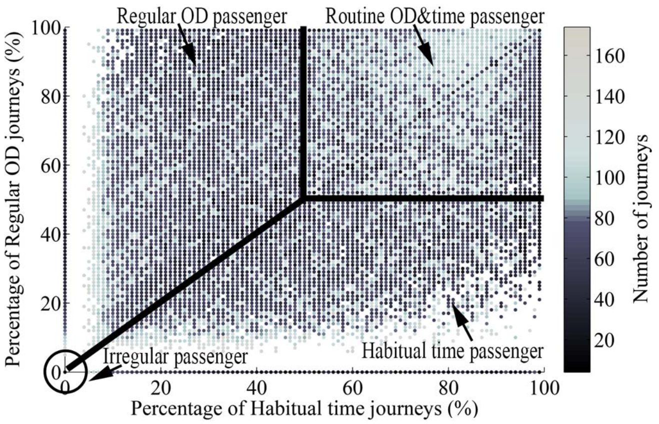


Fig. 7. *A priori* rule for transit passenger segmentation.

2) Passengers with habitual times but without a regular OD are hereafter called *habitual time passengers*. These passengers use public transport at fixed times of the day but travel between multiple ODs.
3) Passengers with both regular ODs and habitual times are hereafter called *transit commuters*. They are commuters who usually use public transport at habitual times for trips between regular ODs.
4) Passengers without neither a regular OD nor a habitual time are hereafter called *irregular passengers*. They do not follow a particular regular transit travel pattern, which means that they probably have other main travel modes.

Each SC user itinerary is revisited during the passenger segmentation process. Fig. 7 illustrated the heuristic rule to segment transit passengers according to the proportion of regular OD and habitual time journeys. Passengers during the study period traveled for a certain number of journeys following a regular OD, a habitual time pattern, or not following any pattern. Each of them is represented as a point in Fig. 7. The color of the point represents the number of journeys made within the study period.

Only passengers with no recognizable pattern are segmented into the irregular passenger type. The other passengers could be grouped into three identifiable types. Transit commuters followed spatial and temporal patterns in most of their journeys. Regular OD passengers made more spatially regular than habitual temporal journeys, and vice versa for habitual time passengers. These heuristic rules are translated to *a priori* rules for market segmentation as follows.

1) *Rule 1*: If no temporal or spatial travel pattern is identified, the passenger is classified as an irregular passenger.
2) *Rule 2*: If more than 50% of the journeys were made within habitual times and between regular ODs, the SC user is classified as a transit commuter.
3) *Rule 3*: The remaining passengers are segmented into regular OD passengers if the proportion of the regular OD journeys is more than the habitual time journeys, and vice versa for the habitual time passengers.

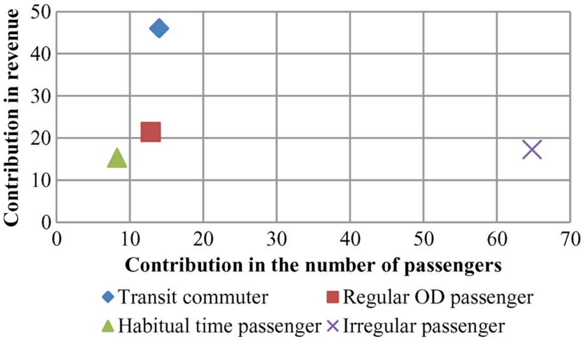


Fig. 8. Passenger segmentation result.

IV. Passenger Segmentation Analysis

This section augments the transit passenger characterization by analyzing the market segmentation results. This section further disaggregates each passenger type into subsegmentations of different transfer behaviors, modes and routes used, card types, and usage frequency. In other words, this section aims to exploit the coarse-grained information of passenger types toward a more fine-grained understanding of market segments, their needs, and the capabilities required to serve them.

Fig. 8 illustrates each market segment contribution in the number of passengers and the fare revenue. The dominance of the irregular passengers (64%) denotes that most of the SC users do not have regular travel patterns. The transit commuters only account for 14%, whereas the regular OD passengers and the habitual time passengers account for 13% and 8%, respectively.

However, the transit commuters made the largest contribution (46%) to the ticket revenue. This fact designates that those who regularly use public transport as the main travel mode are still the major income contributor for transit operators. Conversely, 64% of their customers (irregular passengers) contributed for only 17% of the revenue.

For further analysis, this section analyzes these four passenger segments in terms of the spatial and temporal daily usage (see Section IV-A), the total travel and transfer time (see Section IV-B), the modes and routes used (see Section IV-C), the card type (see Section IV-D), and the frequency of use (see Section IV-E).

A. Spatial and Temporal Pattern of Daily Usage

This section exploits the usage pattern of each passenger type to understand the daily usage of the transit network. Fig. 9 illustrates that the average number of journeys is made per passenger at different times of the day.

Fig. 9 shows that the transit commuters mainly traveled during peak periods, whereas the irregular passengers traveled any time but generally started later in the morning (from 8:00 A.M.) and finished earlier (6:00 P.M.) than any other class. One trip purpose assumption can be made that the transit commuters mostly travel for school- and work-based trips and that the irregular passengers mostly travel for less tightly scheduled trips, such as for leisure or shopping activities. The regular OD passengers such as those who are flexible in time made the most number of journeys during off-peak periods. Conversely, the habitual time passengers such as those who follow a temporal travel pattern mainly traveled during peak periods, similar to the transit commuters.

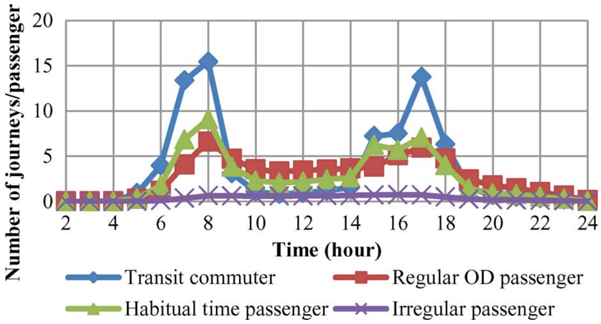


Fig. 9. Average journeys made by each type of passenger.

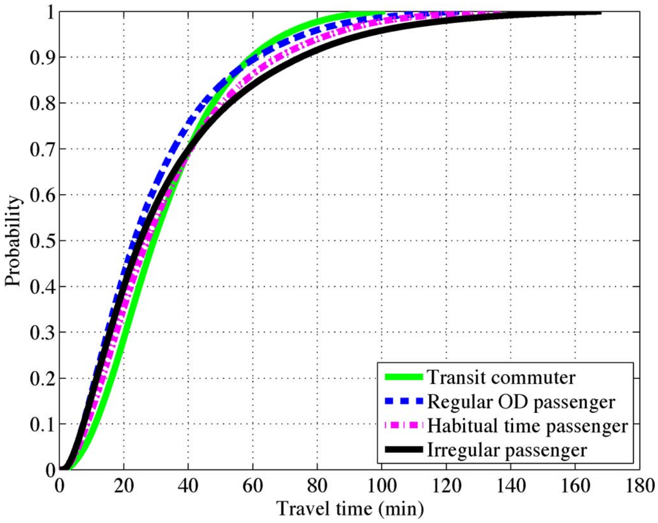


Fig. 10. CDF of the total travel time.

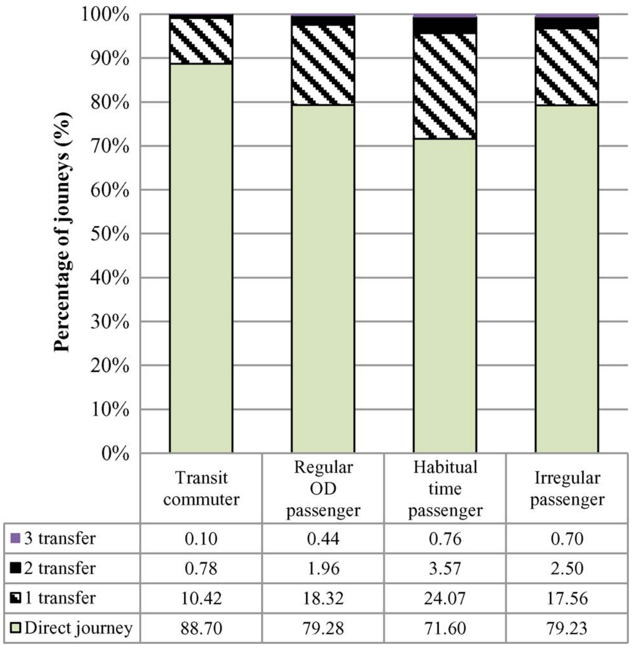
B. Total Travel and Transfer Time

This section exploits the differences between the passenger types in terms of the total travel and transfer time to augment the understanding of passenger behaviors. Fig. 10 shows the empirical cumulative density function (cdf) of the total journey travel time made by different types of passengers.

The total time spent on traveling was relatively similar among different passenger types. The transit commuters spent slightly less time for traveling than the irregular passengers and other types. This difference might come from the difference in the transfer time, which is illustrated in Fig. 11.

Fig. 11 shows that the majority of passengers make journeys with no transfer. The transit commuters made significantly less transfer than those of irregular and habitual time passengers. Nearly 90% of the transit commuters made no transfer during their journeys, leaving only over 10% of the journey having a single transfer and an insignificant number of journeys having more than one transfer. Conversely, over 20% of the irregular passenger journeys required at least a transfer. This fact implies that transfer is one of the most important disutilities of the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



	Transit commuter	Regular OD passenger	Habitual time passenger	Irregular passenger
■ 3 transfer	0.10	0.44	0.76	0.70
■ 2 transfer	0.78	1.96	3.57	2.50
▨ 1 transfer	10.42	18.32	24.07	17.56
□ Direct journey	88.70	79.28	71.60	79.23

Fig. 11. Proportion of journeys with none, one, two, and three transfers.

TABLE II
Mode Choice Decision of Different Passenger Types

	Modes used (%)					
Passenger Type	Train only	Bus only	Ferry only	Bus-Train	Bus-Ferry	Train-Ferry
Transit commuter	42.78	50.02	2.25	4.60	0.29	0.05
Regular OD passenger	24.54	63.94	4.32	6.27	0.71	0.22
Habitual time passenger	17.05	73.33	1.40	7.46	0.64	0.11
Irregular passenger	24.57	64.42	4.04	6.05	0.68	0.25

transit system, which discourages passengers to commute on a daily basis. It is consistent with findings in literature, where existing studies believed that transfer could be the decisive factor to the transit quality of service [21]. The habitual time passengers show high mobility needs, with more transfers than any other passenger types. A journey by a habitual time passenger on average would consist of 1.33 legs compared with only 1.12 legs in that by a transit commuter.

C. Modes and Routes Used

This section investigates the passenger mode and the route choice over the bus, city train, and ferry systems in SEQ, Australia. Table II shows how different passenger types used the transit system. The "train only," "bus only," and "ferry only" modes represent the journeys made by a single transit mode, whereas "bus–train," "bus–ferry," and "train–ferry" represent the corresponding two modes that were used for traveling. There is no journey that used all the three modes of transit.

The number of bus journeys exceeded that of train journeys in all passenger types, whereas the ridership share for ferry journeys was insignificant compared with the other two modes. The service coverage could be the reason for this figure. Although the bus network in SEQ consists of over 24 000 stops and the rail system consists of 146 stations, there are only 24 ferry terminals of 11 ferry lines. The ferry system with a limited commercial speed and service coverage did not attract passengers with a tight schedule. Only 2.25% of the *transit commuters'* journeys and 1.4% of the *habitual time passengers'* journeys were ferry journeys. The insignificant numbers of bus–ferry and train–ferry journeys also suggest the poor connectivity of the ferry to rail and bus networks.

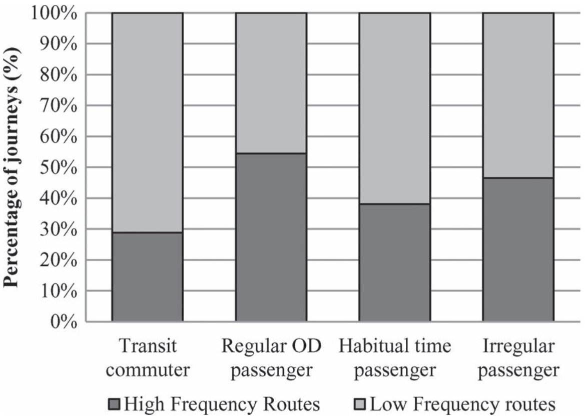


Fig. 12. High-frequency and low-frequency route choice decisions of different passenger types.

The transit commuters noticeably traveled more train itineraries than any other type. The city railway network in SEQ has a centripetal structure, which facilitates the going-to-school/work activities to the Brisbane central business district (CBD) of the transit commuters. Conversely, the bus network structure is more centrifugal and has a widespread coverage. Passengers with mobility needs to multiple destinations such as the habitual time passengers consequently used more buses than any other passenger type.

Fig. 12 illustrates the route choice decisions of bus riders during the study period. The bus routes are classified into high-frequency lines (equal or less than 15 min per vehicle) and low-frequency lines (larger than 15 min per vehicle). The SEQ network consists of 38 high-frequency lines over the total of 446 bus lines.

Despite the limited number of high-frequency bus lines (approximately 8.5% of the total number of lines), Fig. 12 clearly shows that up to 54% of the regular OD passengers' journeys were from those high-frequency lines. This figure was also high in irregular passengers, habitual time passengers, and transit commuters, with 46%, 38%, and 29%, respectively. The results show that high-frequency services were desired by transit passengers of any type. These lines promote the preferable "turn up and go" behavior, where passengers randomly arrive to transit stops without checking a schedule [22]. However, passengers of a high temporal travel pattern such as the transit commuters and the habitual time passengers were less dependent on high-frequency bus lines than the regular OD and irregular passengers. It means that passengers on a time habit are more willing to check the timetable and take the less frequent bus lines. This finding is consistent with the study where Farag and Lyons [22] asserted that people would only

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

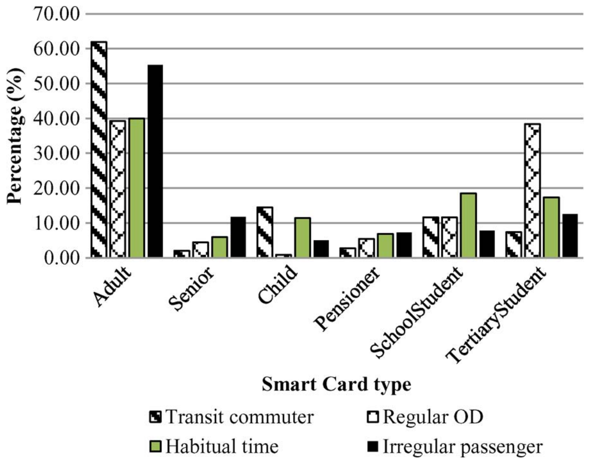


Fig. 13. Proportion of each passenger type over different SC card class.

turn up and go if there are no time constraints and the service is frequent.

D. Card Type

This section analyzes the proportion of each passenger type under the six classes of SC cards in Queensland, Australia (adult, senior, child, pensioner, secondary school student, and student) to augment the understanding of these types.

Fig. 13 shows that adults are the largest contributor in all passenger types. Most of the transit commuters are adults, who are currently charged with the highest ticket fare. Their ticket fares, along with their high number of journeys, explain why the transit commuters are the main contributors of ticket revenue. However, a large proportion of adult cards are irregular passengers, which indicates that public transport is not the main mode of transport for those people.

Most of child cards (inclusive of children of 5–14 years) are also transit commuters due to their tight schedule and lack of travel activities. It is essential that the transit system is safe and reliable so that parents allow their young children to travel by public transport; otherwise, there would be more drop-off/pick-up cars on the roads.

School students still have a tight daily schedule similar to the child class. However, they have more travel activities and require more mobility than the child class, which makes habitual time passengers as the biggest contributor for this class.

The majority of tertiary students are regular OD passengers. Their flexible study schedule is probably the reason for this trend.

The major contributor in the senior and pensioner classes is irregular passengers because passengers from these classes are flexible in time and have no mobility needs for work or study.

E. Frequency of Transit Usage

This section investigates the transit usage frequency of different passenger types by data mining techniques. A data set containing the number of boarding and travel days has been constructed for every passenger. Table III shows an example of the data set.

TABLE III
DATA SET USED IN THE FREQUENCY OF TRANSIT USAGE ANALYSIS

SC ID	Number of traveled day(s)	Number of journey(s) made
X1	48	86
X2	36	59
X3	23	47
X4	3	5
X5	1	2

The number of travel days and journeys made represent the frequency of use from each passenger. This section aims to find a threshold to differentiate between frequent and infrequent transit passengers. For that purpose, a k-mean algorithm is used to classify transit passengers into two clusters. The k-mean algorithm seeks to minimize the sum of all points to the centroid of each cluster [12]. The objective function of the algorithm is expressed as

$$\text{Minimize}: \; J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \tag{2}$$

where

j $= 1, \ldots, k$, where k is the number of predetermined clusters, which, in this case, is $k = 2$;

i $= 1, \ldots, n$, where n is the number of data points, which, in this case, is the passengers, i.e., $n = 1,010,158$;

$\|x_i^{(j)} - c_j\|^2$ distance measure between a data point $x_i^{(j)}$ and cluster center c_j. Points $x_i^{(j)}$ and c_j are located in a 2-D space of the number of travel days and the number of trips made.

The algorithm is composed of the following steps.

a) Place all the points (passengers) into the 2-D space.
b) Assign each point into the cluster of the closest centroid.
c) Recalculate the positions of the two centroids.
d) Repeat steps b and c until the centroids are stationed.

Fig. 14 illustrates the classification result, whereas Fig. 15 shows the proportion of the frequent and infrequent passengers in each passenger type.

The majority of the transit commuters frequently used the transit system during the study period, whereas almost all the irregular passengers did not travel frequently. It means that the passengers using public transport on a spatial and temporal travel pattern would also travel more than those having no travel pattern.

F. Summary of Passenger Travel Pattern Analysis

This section sums up the knowledge gained from analyzing the passenger segments. Table IV shows the descriptive statistic of each passenger type.

The understanding of each passenger type augments the passenger characterization. The following understanding about

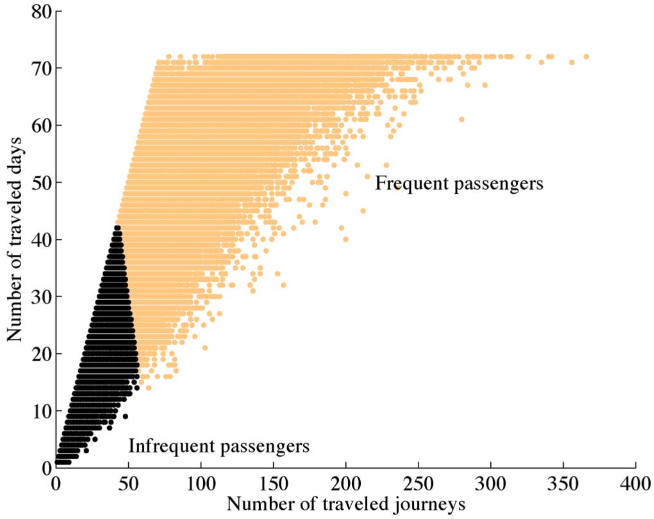


Fig. 14. Classification of the frequent and infrequent passengers.

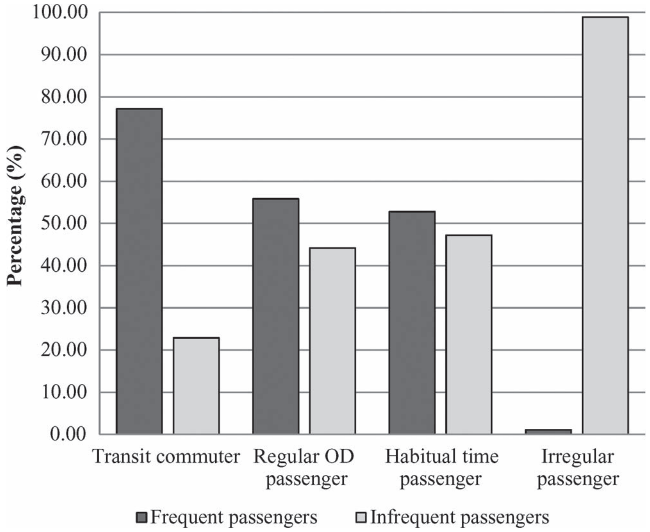


Fig. 15. Frequency of the transit usage in each passenger type.

transit riders could be gained from the passenger segmentation analysis.

1) The majority (64%) of the operated SCs are irregular passengers, who do not follow any travel pattern. These passengers rarely travel by public transport (99% of them are infrequent users of the transit system) and in total contributes to only 17% of the total ticket revenue. It means that selling more SCs would not earn much profit to both the transit authority and the society, but passengers should be encouraged to make more journeys.
2) Most of the transit commuters would travel during peak periods and travel from outside to inside the CBD in the morning peak, and vice versa for the afternoon peak. The transit commuters only proportioned for around 14% of the SC population but contributed to 46% of the revenue because their majority are frequent transit users (77%). It means that encouraging passengers to more regularly travel would also increase their travel frequency and, eventually, the ticket revenue contributions. However, these passengers are directly affected by transfers, which could be one of the decision factors limiting their patronage.
3) Regular OD passengers represent people who are flexible in time but who regularly travel between OD pairs, such as tertiary students. Regular OD passengers travel more in the off-peak period than any other type.
4) Habitual time passengers represent people who usually travel within a regular time period but travel to different destinations. Due to the need of traveling to multiple destinations, this type of passengers requires more transit mobility. Consequently, they take more transfers and more bus journeys than any other type. Examples of them are school students.

V. Discussion on Potential Applications of the Proposed Passenger Segmentation

To transit authorities, all passengers deserve the utmost attention. The understanding of each passenger type, behaviors, and needs facilitates the development of the transit system and service provision to better serve each individual passenger. The characterization of passenger demand by the proportion of each type shows the overall service requirements for each region. For instance, an area of mainly regular OD passengers would not require as much timely service as an area of transit commuters, whereas too many irregular passengers suggests a problem in the transit service provided. Transit authorities could aim to raise the number of high-value customers such as transit commuters for revenue- and transit-oriented developments. Nevertheless, they could also fulfill the needs of other passenger types to maintain the customer equity and the overall attractiveness of the system. This section suggests several practical service improvements using the knowledge of passenger segmentation.

TABLE IV
Descriptive Statistic of the Transit Passenger Types

	% Card	% Revenue	Mean legs/journey	% Frequent	% Infrequent
Transit commuter	14.04	46.04	1.12	77.14	22.86
Regular OD passenger	12.86	21.38	1.24	55.84	44.16
Habitual time passenger	8.28	15.32	1.33	52.78	47.22
Irregular passenger	64.82	17.26	1.25	1.10	98.90

The understanding of each passenger type helps transit authorities in transit strategic planning. Transit-on-demand services that serve people who need regular travel where standard routes are not available can be developed. Transfer coordination may be developed for major transfer stops used by large numbers of transit commuters and habitual time passengers.

A targeted survey could aim for the irregular passengers to understand the disutility that limits the level of ridership. The segmentation results indicate that, before thinking about attracting new customers, transit authorities should first encourage the majority of their customers to choose public transport as the main travel mode. The enormous number of operated

SCs does not tell much about the passenger transit usage because 64% of them rarely travel. Transit authorities could pay special interest to passengers who were not irregular passengers before but recently became irregular passengers. These are the potential customers whose behaviors have changed due to certain reasons. Understanding these reasons would benefit transit authorities to prevent the reduction of patronage. Conversely, the reasons for an irregular passenger to transform to another passenger type would be interesting successful stories to learn to further improve the transit system.

Transit authorities could also observe the impacts of recent transit policies to their customers. For instance, a policy on reducing the transit fare during off-peak periods would cause passengers, particularly people who have flexible daily schedules such as regular OD passengers, to travel more during off-peak periods. Transit authorities could foresee the number of affected passengers for this policy by looking at those who usually travel at the end of the peak period and are flexible in terms of time. The number of passengers at each type before and after a policy implementation is an important evaluation of different fares, marketing, and servicing strategies. For instance, more transit commuters and less irregular passengers mean that more passengers become daily users of public transport.

Finally, incentives and personalized services can be given to transit commuters, regular OD passengers, and habitual time passengers to encourage passengers to use public transport for commuting. The characterization of regular behaviors provides a tremendous opportunity for transit authorities to provide personalized information and incentives to each passenger. Once a passenger enrolled in the system, real-time information on their regular journeys could be given to individual passengers. Special incentives can be given to promote the commuting behavior, e.g., by reducing the ticket type on regular journeys. Although many SC systems are not associated with user contact information, the travel pattern can be stored and provided to each individual passenger through SC IDs. The information and the service can remain customized for each individual, and at the same time, maintain privacy.

VI. Conclusion

This paper has proposed a systematic approach to mine the travel pattern and to segment transit passengers only using SC data.

The individual transactions of each SC user on each working day were combined to reconstruct travel itineraries. DBSCAN algorithms were separately applied to mine the regular OD and the habitual time from the travel itineraries. The passengers were finally segmented into transit commuters, regular OD passengers, habitual time passengers, and irregular passengers by an *a priori* passenger market segmentation approach. Analyses on SC user types indicate interesting patterns of transit usage from each type. Practical applications of the method were also discussed, showing benefits to both SC users and transit operators.

The passenger segmentation methodology presented in this paper has enabled transit operators to segment their customers and provide them with well-suited information and services. Further extensions of this paper are in progress, in which the SC data of other time periods are used to evaluate the impacts of different policies to transit passengers. Investigations of the coming-back-home behavior and the segmentation of the trip distance further augment the understanding of passenger types. In the meantime, the findings of this paper have been useful to augment the passenger characterization and to better cater to individual transit passengers.

Acknowledgment

The authors would like to thank the Smart Transport Research Centre members, and also S. McEvoy and D. Ng of Translink, for their constructive comments. The conclusions in this paper reflect the understandings of the authors, who are responsible for the accuracy of the data.

References

[1] R. Elmore-Yalch, *A Handbook: Using Market Segmentation to Increase Transit Ridership*, vol. 36. Washington, DC, USA: Transportation Research Board, 1998.

[2] D. A. Hensher, "Establishing a fare elasticity regime for urban passenger transport," *J. Transp. Econ. Policy*, vol. 32, no. 2, pp. 221–246, 1998.

[3] Y. Shiftan, M. L. Outwater, and Y. Zhou, "Transit market research using structural equation modeling and attitudinal market segmentation," *Transp. Policy*, vol. 15, no. 3, pp. 186–195, May 2008.

[4] *Transport Plan for Brisbane 2008–2026*, Brisbane City Council, Brisbane, U.K., 2008.

[5] K. K. A. Chu and R. Chapleau, "Augmenting transit trip characterization and travel behavior comprehension," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2183, pp. 29–40, 2010.

[6] W. Jang, "Travel time and transfer analysis using transit smart card data," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2144, pp. 142–149, 2010.

[7] C. Seaborn, J. Attanucci, and N. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2121, pp. 55–62, 2009.

[8] J. M. Farzin, "Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2072, pp. 30–37, 2008.

[9] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 1971, pp. 119–126, 2006.

[10] K. K. A. Chu, R. Chapleau, and M. Trepanier, "Driver-assisted bus interview," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2105, pp. 1–10, 2009.

[11] S. Lee and M. Hickman, "Trip purpose inference using automated fare collection data," *Public Transp.*, vol. 6, no. 1/2, pp. 1–20, Apr. 2014.

[12] C. Morency, M. Trepanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transp. Policy*, vol. 14, no. 3, pp. 193–203, May 2007.

[13] S. Lee and M. Hickman, "Are transit trips symmetrical in time and space?" *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2382, pp. 173–180, Dec. 1, 2013.

[14] S. G. Lee, M. Hickman, and D. Tong, "Stop aggregation model," *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2276, pp. 38–47, 2012.

[15] *How to Use Your Go Card on the TransLink Network: TransLink Go Card User Guide (Part 1 of 2)*, Translink, Vancouver, BC, Canada, 2007.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

[17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999.

[18] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Proc. KDD*, 1998, pp. 58–65.

[19] M. Burke and A. Brown, "Distances people walk for transport," *Road Transp. Res.—J. Australian New Zealand Res. Pract.*, vol. 16, no. 3, pp. 16–29, Sep. 2007.

[20] *Transit Capacity and Quality of Service Manual*, Transportation Research Board (TRB), Washintgon, DC, USA, 2013.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[21] H. Mohring, J. Schroeter, and P. Wiboonchutikula, “The values of waiting time, travel time, and a seat on a bus,” *Rand J. Econ.*, vol. 18, no. 1, pp. 40–56, 1987.

[22] S. Farag and G. Lyons, “What affects use of pretrip public transport information?: Empirical results of a qualitative study,” *Transp. Res. Rec.—J. Transp. Res. Board*, vol. 2069, pp. 85–92, 2008.

Le Minh Kieu received the Bachelor’s degree from University of Transport and Communications, Hanoi, Vietnam, and the Master’s degree in transport engineering from Linköping University, Linköping, Sweden. He is currently working toward the Ph.D. degree at the Smart Transport Research Centre, Queensland University of Technology, Brisbane, Australia.

Ashish Bhaskar received the Bachelor’s degree in civil engineering from Indian Institute of Technology, Kanpur, India; the Master’s degree from The University of Tokyo, Tokyo, Japan; and the Ph.D. degree from Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland.

He is a Lecturer with Queensland University of Technology, Brisbane, Australia, where he is also a Traveler Information Research Domain Leader with the Smart Transport Research Centre.

Edward Chung received the B.Sc. (Hons.) and Ph.D. degrees from Monash University, Melbourne, Australia.

He is a Professor with Queensland University of Technology, Brisbane, Australia, where he is also the Director of the Smart Transport Research Centre.

View publication stats