Transportation Research Part C 36 (2013) 1–12

Contents lists available at ScienceDirect

Transportation Research Part C

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

Mining smart card data for transit riders' travel patterns

Xiaolei Ma [a], Yao-Jan Wu [b], Yinhai Wang [a,*], Feng Chen [c], Jianfeng Liu [c]

[a] *Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700, United States*
[b] *Department of Civil Engineering and Engineering Mechanics, University of Arizona, Tucson, AZ 85721, United States*
[c] *Beijing Transportation Research Center, Beijing 100073, China*

ARTICLE INFO

Article history:
Received 8 February 2013
Received in revised form 18 July 2013
Accepted 18 July 2013

Keywords:
Automatic Fare Collection System
Smart card
Transit travel pattern
K-Means algorithm
Rough set theory

ABSTRACT

To mitigate the congestion caused by the ever increasing number of privately owned automobiles, public transit is highly promoted by transportation agencies worldwide. A better understanding of travel patterns and regularity at the "magnitude" level will enable transit authorities to evaluate the services they offer, adjust marketing strategies, retain loyal customers and improve overall transit performance. However, it is fairly challenging to identify travel patterns for individual transit riders in a large dataset. This paper proposes an efficient and effective data-mining procedure that models the travel patterns of transit riders in Beijing, China. Transit riders' trip chains are identified based on the temporal and spatial characteristics of their smart card transaction data. The Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm then analyzes the identified trip chains to detect transit riders' historical travel patternsand the K-Means++ clustering algorithm and the rough-set theory are jointly applied to cluster and classify travel pattern regularities. The performance of the rough-set-based algorithm is compared with those of other prevailing classification algorithms. The results indicate that the proposed rough-set-based algorithm outperforms other commonly used data-mining algorithms in terms of accuracy and efficiency.

Published by Elsevier Ltd.

1. Introduction

Approximately 76% of those living in the United States chose privately owned vehicles for their commute to work in 2000 (ICF Consulting, 2003) and data collected for the 2009 American Community Survey indicate that 79.5% drive alone when commuting (McKenzie and Rapino, 2011). This pattern is now becoming apparent in developing countries such as China, where many rely on privately owned vehicles to commute. In 2010, for example, more than 34% of Beijing residents chose cars as their primary travel mode while only 28.2% chose transit (Beijing Transportation Research Center, 2011).

Public transit has long been considered to provide an effective way to reduce congestion, air pollution, and energy consumption (Federal Highway Administration, 2002). To improve transit services and encourage more people to use public transit, transit agencies have been striving to identify the key factors that attract transit riders through studying their travel patterns. With a better understanding of the travel patterns of transit riders, transit authorities will be able to evaluate theircurrent services to reveal how best to adjust their marketing strategies to encourage higher usage (Boyle et al., 2000). For example, knowing why some riders are especially loyal to transit can help transit agencies to determine where and when they should provide discounts to retain these loyal transit riders and potentially attract new riders (Trépanier et al., 2012). Based on identified travel patterns and transit usage regularities, transit authorities are able to evaluate the most

[*] Corresponding author. Tel.: +1 (206) 616 2696; fax: +1 (206) 543 1543.
E-mail addresses: xiaolm@uw.edu (X. Ma), yaojan@email.arizona.edu (Y.-J. Wu), yinhai@uw.edu (Y. Wang), chenf@bjtrc.org.cn (F. Chen), ljf@bjtrc.org.cn (J. Liu).

0968-090X/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.trc.2013.07.010

cost-effective fare packages for transit riders, understand how transit riders' behaviors are likely to change in response to a new fare structure, and thus select a fare policy that achieves the optimum balance between enhancing the attractiveness of the transit system and maximizing fare revenue (Taylor and Jones, 2012).

Transit planners and researches can also utilize individual travel-behavior data for activity-based trip modeling and transit travel demand analyses. For public agencies, information on the travel patterns for individual transit riders can also be utilized to quantify the effectiveness of transit-oriented development (TOD) (Dill, 2008). In particular, personal travel behavior data can reveal how TOD residents change their daily commuting behaviors and how transit use varies spatially and temporally. However, acquiring individual transit travel pattern is challenging (Tirachini, 2012). Traditional transit travel pattern analysis largely relies on rider satisfaction surveys or travel diaries (Chu and Chapleau, 2010), which is very costly and difficult to implement at a multiday level due to the low response rate andaccuracy. The use of, smart card data to track passengers' long term travel activities and patterns, such as the number of typical daily trip chains, common boarding/alighting stops and trip start/end times, offers a far more convenient and efficient data source. Smart card data records both temporal and spatial information for each rider, making it feasible to conduct individual travel pattern analysis through longitudinal analyses (Chu, 2010).

Most previous smart-card-based research into transit traveler behaviors has focused on transfer point estimation and on origin and destination inference (Munizaga and Palma, 2012). Pelletier et al. (2011) reviewed previous smart card data studies, concluding that modeling individual based trip behavior is a potentially very challenging topic. Kitamura et al. (2006) and Morency et al. (2006, 2007) utilized multiple day smart card data to analyze transit riders' travel variability and pointed out that developing a better understanding of travel variability can help reduce operational costs and manage demand. Several studies (Bagchi and White, 2004, 2005) have shown that transit agencies can encourage customer loyalty based on multiday smart card data. Utsunomiya et al. (2006) used smart card data from the Chicago Transit Authority (CTA) to extract passengers' transit usage and access distance, reporting that transit usage data can provide useful information for transit planning and market research. Webb (2010) emphasized the importance of building loyalty anddeveloped several measures (including satisfaction and quality of service) to quantify transit loyalty, although these findings were based on data gathered using a traditional customer satisfaction survey. Lee and Hickman (2011) defined regular transit users as those making two or more trips during typical weekdays, and found that travel patterns varied by card type.

Most of the aforementioned research based on smart card data extracted travel behavior information macroscopically rather than by analyzing individual transit riders' travel patterns. Chu and Chapleau (2010) applied the association rule and clustering algorithms to measure transit riders' regularity, and conducted an individual travel behavior analysis using both temporal and spatial methods. However, their analysis was based on high quality data with complete information and their methodwas not optimized for a large dataset. In reality, most transit agencies have adopted a comprehensive procedure to store smart card data, providing strict authorization and security mechanisms to protect the personal information generated from smart card data (Dinant and Keuleers, 2004). Sensitive content such as passenger age, name, boarding and alighting locations are intentionally truncated to address privacy concerns (Verykios et al., 2004), so efficient data mining approaches are needed to infer passenger travel behavior information from these incomplete smart card datasets.

Extracting transit riders' travel patterns from smart card data can be particularly challenging because the Automatic Fare Collection (AFC) system was not originally designed to support transit planning and transit performance measures (Pelletier et al., 2011). As a result, the smart card data collected by the AFC system lacks certain trip related information that affects data processing performance. To deal with this data issue, this paper proposes a robust and comprehensive data-mining procedure to extract individual transit riders' travel patterns and regularity from a large dataset with incomplete information. Specifically, two major issues are examined in this study. First, the spatial and temporal travel patterns for a particular transit rider are investigated. Here, "spatial travel pattern" means that the transit rider repeatedly visits the same or adjacent places on a multi-day basis and "temporal travel pattern" means the transit rider starts (and/or finishes) his/her daily trip during the same time period. Then we move onto determine the "regularity" of a transit rider's travel pattern, which refers to "frequency of the similar trips for this transit rider", and the frequency of the similartrips can be considered an effective measurement of travel regularity. The objectives of this study are to assist both transit agencies and transportation researchers by: (1) developing a novel data mining procedure to extract individual passengers' travel patterns and travel regularity; and (2) ensuring these data mining algorithms are capable of processing massive smart card datasets within a tolerable elapsed time.

The remainder of this paper is organized as follows. The data used in this study is first introduced, followed by an illustration of the temporal travel pattern distribution. The proposed methodology is then explained in the following order: (1) trip chain generation, (2) individual travel pattern recognition, (3) regularity level clustering, and (4) performance enhancement using the rough set theory, after which the performance of the proposed algorithm is compared with other commonly used data-mining algorithms. The paper concludes bysummarizing the research findings and suggesting directions for future research.

2. Data description

Beijing Transit Incorporated began to issue smart cards in May 10, 2006, that could be used in both the Beijing bus and subway systems. Due to the highly discounted fares (up to 60% off) provided by the smart card, more than 90% of the city's transit riders paid for their trips with their smart cards in 2010 (Beijing Transportation Research Center, 2010). There are two types of AFC systems in Beijing: flat fares and distance-based fares. Transit riders pay a fixed rate for flat fare buses by tapping their smart cards on the card reader when entering; only check-in scans are necessary. For the distance-based AFC system, transit riders need to swipe their smart cards when checking-in and checking-out.Transit riders need to hold their smart cards near the card reader device to complete transactions when entering or exiting buses. However, due to a design flaw in the smart card scan system, the AFC system on flat fare buses does not save any boarding location information, although it does store information on boarding and alighting locations, but not boarding time, on distance-based fare buses. Key information stored in the database therefore includes smart card ID, route number, driver ID, transaction time, remaining balance, transaction amount, boarding stop (for distance-based fare buses only), and alighting stop (for distance-based fare buses only). Data on more than 16 million smart card transactions are generated every day, of which 52% are from flat-rate bus riders. These characteristics of the Beijing AFC system create additional challenges for those seeking to process the data and mine useful information from it. The AFC system used in Beijing is not aunique case; the same data challenges are commonly seen in many other Chinese cities, including Chongqing (Zhou et al., 2007), Nanning (Chen, 2009), and Kunming (Gao and Wu, 2011), etc. Internationally, the AFC system in São Paulo, Brazil (Farzin, 2008) also does not record the boarding location.

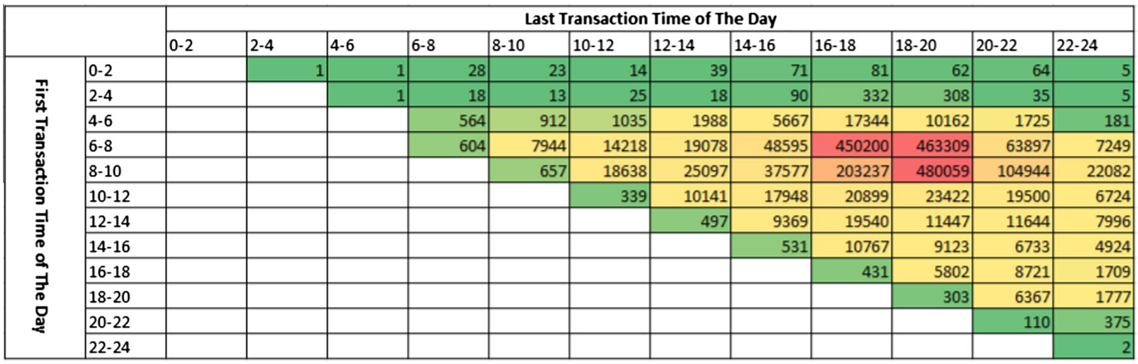
To demonstrate the temporal travel patterns and the pattern regularity for transit riders in Beijing, consider a typical travel week (in this case, the week of Monday July 5th to Friday July 9th, 2010). The transaction data from 3,845,444 smart cards was collected for that week, 58% of which (2,225,298 cards) contained two transactions for all five weekdays. Fig. 1shows the temporal frequency distribution of the "transaction pair" of the first transaction time and the last transaction time of the smart cards with two transactions per day. As shown in the red cells of Fig. 1, most of the transit riders began their first trip between 6 AM and 10 AM, and ended their travel for the day between 4:00 PM and 8:00 PM. This is likely to represent a typical commuting trip chain, where a transit rider takes a bus or subway from his or her home to their place of work in the morning and then returns home in the evening. The temporal distribution shown in Fig. 1 implies that strong temporal travel patterns exist in the multiday smart card data. However, the regular spatial travel pattern for a specific card holder remains uncertain and will be explored in the analysis describedbelow.

3. Methodology

The focus of this study is twofold: individual travel pattern recognition and travel regularity mining. A flow chart of the work performed for the study is illustrated in Fig. 2: (1) retrieve each passenger's multi-day's smart card transactions from the database; (2) generate this passenger's trip chains utilizing their spatiotemporal relationships; (3) apply a series of data mining approaches to extract this passenger's travel pattern and travel regularity based on the generated trip chains. To reduce the complexity involved in the regularity clustering algorithm, association rules were identified for the large-scale smart card data mining process.

3.1. Trip chaingeneration

Before the spatial and temporal patterns of individual transit riders can be examined, their trip chain information must be constructed. A trip chain is defined as a series of trips made by a traveler on a daily basis and is considered a useful way to demonstrate travelers' behaviors (McGuckin and Nakamoto, 2004). For flat fare buses, transit riders swipe their smart cards only during boarding and the smart card reader is not able to record either their boarding location or when and where they alight. In order to estimate transit riders' boarding stops, a Markov chain based Bayesian decision tree algorithm by (Ma et al., 2012) was therefore utilized to extract changes in the boarding volume with time between two consecutive transactions and apply this information, in conjunction with historical speed profiles retrieved from GPSdata, to calculate the



		Last Transaction Time of The Day											
		0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24
First Transaction Time of The Day	0-2		1	1	28	23	14	39	71	81	62	64	5
	2-4			1	18	13	25	18	90	332	308	35	5
	4-6				564	912	1035	1988	5667	17344	10162	1725	181
	6-8				604	7944	14218	19078	48595	450200	463309	63897	7249
	8-10					657	18638	25097	37577	203237	480059	104944	22082
	10-12						339	10141	17948	20899	23422	19500	6724
	12-14							497	9369	19540	11447	11644	7996
	14-16								531	10767	9123	6733	4924
	16-18									431	5802	8721	1709
	18-20										303	6367	1777
	20-22											110	375
	22-24												2

Fig. 1. Weekly temporal distribution for transit smart card holders with two transactions for the week of 5th–9th July 2010.

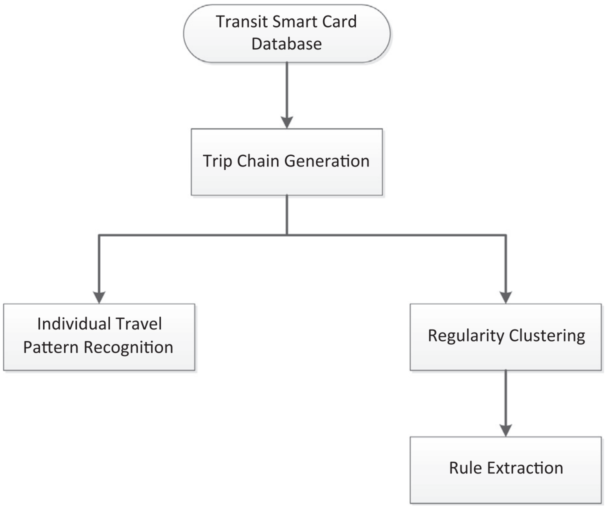

Transit Smart Card Database
Trip Chain Generation
Individual Travel Pattern Recognition
Regularity Clustering
Rule Extraction

Fig. 2. Flow chart of the study research process.

probabilities for all potential boarding stops; the stop with the maximum probability was assumed to be the boarding stop. Based on this algorithm, more than 90% of the smart card data can be accurately assigned. For distance-based fare buses, although boarding times are not recorded by the smart card reader, other information such as each passenger's boarding stop location, alighting stop location, and alighting time are known. Therefore, the missing boarding time can be approximately substituted by another passenger's alighting time at the same stop.

A fixed temporal threshold was used in our study to link several smart card transaction records into a trip chain. Fixed temporal thresholds change depending on the type of transfer activity. For instance, if a passenger transfers from a distance-based fare bus to a flat fare bus, the alighting time in the previous trip and the boarding time in the current trip are known and theappropriate 30 min time interval recorded. However, if this passenger makes a transfer from a flat fare bus to a distance-based fare bus, the alighting time is not recorded when he/she exits from the flat fare bus so a 60 min time interval was utilized to differentiate various trips in this study to take into account both in-vehicle travel time and transfer time. The determination of transfer time intervals for different transfer activities was based on the 2010 Beijing *4th Comprehensive Transport Survey* (Beijing Transportation Research Center, 2012), with the average transit transfer time and in-vehicle travel time being 25.4 min and 40 min, respectively. The same survey revealed that more than 94% of the transfer activities took less than 60 min, so if the transaction time difference between two consecutive smart cardrecords was greater than 60 min, a new trip was generated; times less than this were taken to represent a transfer activity between two routes or two transportation modes (bus and subway) (Jang, 2010).

Table 1 shows linked trip chain examples extracted from the study data. Here, *Chain ID* is a unique identifier for each trip chain sorted in ascending order by the transaction time. For each *Card ID*, the first trip's boarding time (*First Boarding Time)* and the last trip's alighting time (*Last Alighting Time*) are associated with that *Chain ID*. *Route Sequence* refers to the routes the rider took and *Stop ID Sequence* refers to the boarding andalighting stop IDs for distance-based fare buses. As previously noted, only distance-based fare buses and subways record both boarding and alighting locations, but the subway AFC system also has no check-out smart card scan reader when transit riders transfer between different lines. Take Chain ID 46388399 as an example. The transit rider boarded the distance-based fare bus on Route 635 at Stop ID 99964, and alighted at Stop ID

Table 1
Extracted trip chain information for an individual transit rider for the week of 5th–9th July, 2010.

Chain ID	Card ID	Date	First boarding time	Last alighting time	Route sequence	Stop ID sequence
46388399	1000751018309337	20100705	07:08:45	07:47:28	00635 → 10 → 13	99964,99966 → 50258,50167
46388400	1000751018309337	20100705	18:15:24	18:53:10	13 → 10 → 00635	50192,50245 → 100013,100015
46388401	1000751018309337	20100706	07:19:21	08:01:13	00350 → 10 → 13	91267,91269 → 50258,50167
46388402	1000751018309337	20100706	17:56:08	18:49:50	13 → 10 → 00635	50192,50245 → 100013,100015
46388403	1000751018309337	20100707	07:10:43	07:49:21	00635 → 10 → 13	99964,99966 → 50258,50167
46388404	1000751018309337	20100707	18:29:00	19:06:47	13 → 10 → 00350	50192,50245 → 91276,91278
46388405	1000751018309337	20100708	21:13:58	21:40:10	5 → 10	50125,50246
46388406	1000751018309337	20100709	07:16:24	08:03:46	00635 → 10 → 13	99964,99966 → 50258,50167
46388407	1000751018309337	20100709	17:25:00	18:11:59	13 → 10 → 00635	50192,50245 → 100013,100015
46388408	1000751018309337	20100709	18:30:31	NULL	00031	NULL

Note: Subway routes are denoted as one or two digits.

99966.That individual then made a transfer to subway Line 5 at Stop ID 50258, finishing his or her journey by exiting subway Line 10 at Stop ID 50167. Due to the lack of alighting location information for flat fare buses, some of the trip chains suffer from missing alighting time and stop ID sequence information, e.g. Chain ID 46388408 in Table 1. However, this does not have a huge impact on the accuracy of the individual travel pattern recognition and regularity clustering algorithms since both algorithmsare capable of handling both missing values and outliers.

3.2. Individual travel pattern recognition

Once the trip chain info has been constructed, the travel pattern for each transit rider is further investigated through clustering the trip chains. As shown by the example in Table 1, an individual transit rider is likely to show a certain travel pattern during a multi-day period. To retrieve these hidden and repeated travel patterns in an efficient manner, the density-based spatial clustering of application with noise (DBSCAN) algorithm was therefore adopted. Unlike most non-hierarchical clustering algorithms, the DBSCAN algorithm is not required to define the number of clusters (Ester et al., 1996) or identify arbitrarily shaped clusters becausehigher-density records are more likely to be grouped into a cluster. Two key parameters do, however, need to be defined in the DBSCAN algorithm: the ε distance and the minimum number of points (*MinPts*). The ε distance defines the density-reachable range; if a sample record falls within the ε distance, then this record will be included into an existing cluster. *MinPts* limits the minimum number of records in each cluster; if the number of records in each final cluster is less than *MinPts*, then these records are marked as noise. If the records are close to each other (i.e. more dense), these records are more likely to be clustered by DBSCAN. An outlier is often distinct from other dense records, so DBSCAN is able to detect these outliers.

A transit rider may begin their repeated trips in both the spatial and temporal domainsand transit riders' recurring boarding/alighting locations and times are considered simultaneously for clustering. In our application, a minimum of three records are required to form a cluster, and the ε distance is set to one. Spatially, if the frequent boarding (or alighting) stops along the recurring routes are adjacent to each other, these stops may be considered as an identical origin (or destination). Therefore, an additional algorithm was used to detect the spatial relationship between multiple routes and applied in the process of DBSCAN clustering, as follows:

Step 1: Randomly retrieve one record that is flagged as unvisited from the sorted trip chain database for an individual smart card. Flag this record as visited and form a cluster for this record.
Step 2: Check the boarding time difference between unvisited records and the last visited record. If the difference is greater than 1 h, repeat *Step 1*.
Step 3: Check the spatial relationship between unvisited records and the last visited record. If a spatial relationship exists (within 200 meters), then this record is included into the cluster formed in *Step 1* and flagged as visited.
Step 4: For each cluster, if the number of total records is less than 3, then these records in the cluster are flagged as noise; otherwise, the new cluster is confirmed.
Step 5: Continue to process those unvisited records from Step1 through Step 4 until all the records are flagged as visited.
Step 6: The number of total clusters is the number of typical trip chains per day. The recurring route, boarding/alighting stops and timings can be acquired by counting the most frequent pattern within each cluster.

Take the trip chain data from Table 1 as an example. Based on the DBSCAN clustering algorithm, several patterns can be inferred:

(1) This transit rider regularly starts his or her first trip around 7:00 AM, and ends their last trip at around 6:00 PM.
(2) Recurring routes occur for most weekdays. Although an unusual travel pattern is detected on July 8th, this is flagged as noise by the DBSCAN algorithm.
(3) As previously mentioned, transit riders may take different routes to the same location. This rider took another route, route 350, on July 6th; however, route 350 shares the same stops as route 635 along this section so the two routes are considered a "common" route, and the shared stops are grouped together.

The routes and stops frequently visited by this transit rider are depicted on the Geographic Information Systems (GIS) map shown in Fig. 3. The arrows show the weekday pattern the transit rider followed and clearly suggest that this rider takes a home-to-work trip every morning and then returns home from his or her workplace in the evening.

3.3. Regularity clustering

The historical travel pattern for a particular transit rider can be successfully extracted using the above procedure, but their individual travel pattern regularity is still unknown. As explained earlier, in this context regularity means "frequency of the similar trips for this transit rider." Identifying travel pattern regularity would help transit agencies evaluate theimpacts of transit service provision and potential network changes, enabling them to conduct more effective marketing campaigns and measure transit performance (Foote et al., 2001).

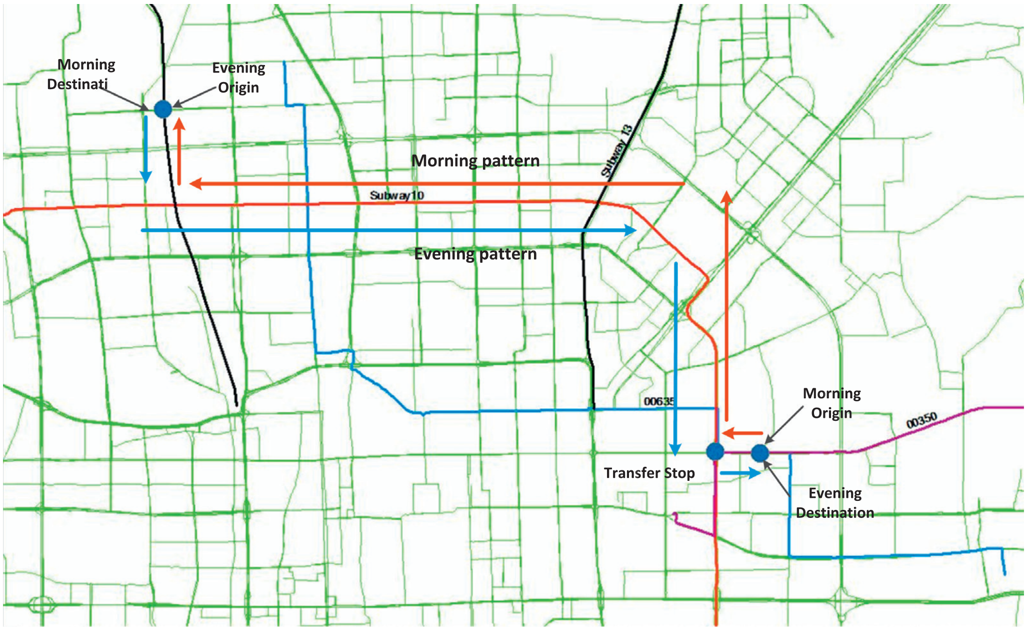


Fig. 3. Example of a transit rider's travel pattern.

Clustering algorithms have been widely used to investigate customer loyalty in the retail and on-line shopping industries (Mauri, 2003; Cheng and Chen, 2009). The same principle can be applied to cluster transit riders with similar travel patterns and place them into different regularity levels based on their temporal and spatial characteristics. Several attributes in the trip chain data were therefore selected as features for clustering as follows:

1. *Number of travel days*. The more days a transit rider travels, the more likely it is that he or she is a frequent transit rider.
2. Number of similar *First Boarding Times*. Boarding time represents a rider's temporal characteristics. If a rider begins his or her trip at a similar time of day every weekday, then this rider is more likely to be a regular transit rider.
3. Number of similar *Route Sequences*. Route sequence represents a general spatial pattern for a rider. The number of similar route sequences followed during the week may indicate a repetitive travel pattern.
4. Number of similar *Stop ID Sequences*. The Stop ID sequence may contain detailed spatial similarity information. In many cases, two different *Stop IDs* might be spatially adjacent, which can be identified by GIS buffer processing.

There may be a certain level of correlations between selected features, such as the numbers of similar *Route Sequence*s and *Stop ID sequence*s. However, these correlated features should not be eliminated since there are missing values within the Beijing transit smart card data and introducing a certain level of redundancy into the travel regularity clustering can help improve the algorithm accuracy. Redundant features (e.g. the number of similar stop IDs and the number of similar route sequences) can thus lead to more accurate clustering results.

In order to efficiently and effectively cluster regularity, a suitable clustering algorithm needs to be chosen. The K-Means algorithm is one of the well-known clustering algorithms. This algorithm tries to partition n records into k clusters by minimizing the within-cluster sum ofsquares. By continuously updating the mean of the record values, each observation is assigned into the cluster with the nearest center until no more observations can be assigned (Forgy, 1965). Although the K-Means algorithm can demonstrate a very high performance and has been applied in many fields, the algorithm suffers from two major intrinsic disadvantages. First, the K-Means algorithm relies on the random initialization of the cluster center and the solution may fall into a local optimum instead of the global optimum as a result of the selection of starting points. If the starting points are far from the true centers of the clusters, the clustering result tends to be locally optimized. Second, the algorithm could require a super-polynomial run time in the worst scenario.

K-Means++, which was proposed by Arthur and Vassilvitskii (2007), addressesthe first of these issues by enhancing the initialization process of the traditional K-Means algorithm using a randomized seeding technique to guarantees the optimal solution is obtained. An additional benefit is that the computational complexity of the K-Means++ algorithm is only $O(\log k)$, where k is the number of clusters. More details of the K-Means++ algorithm can be found in Arthur and Vassilvitskii (2007).

To equalize the magnitude and variability of the four input features, variable standardization is conducted before clustering. The range of each variable serves as the divisor to ensure each standardized variable falls between 0 and 1:

$$z = (v - \min(v))/(\max(v) - \min(v)) \tag{1}$$

K-Means++ was therefore chosen to cluster transit riders with similar travel patterns, and each standardized variable can then be incorporated during the travel pattern clustering process.

Five clusters of regularity are used here: Very High (VH), High (H), Medium(M), Low (L), Very Low (VL). The cluster centers can be expressed as:

$$\begin{aligned} c_1 &= (v_{11}, v_{12}, v_{13}, v_{14}) \\ c_2 &= (v_{21}, v_{22}, v_{23}, v_{24}) \\ &\vdots \\ c_5 &= (v_{51}, v_{52}, v_{53}, v_{54}) \end{aligned} \tag{2}$$

where v_{ij} represents the jth feature of the generated attributes from the trip chain data, and i refers to the $i$thcluster.

Then, the Euclidean distance between c_i and the zero point is calculated. This distance is defined as the cluster center distance:

$$\begin{aligned} D_1 &= \sqrt{(v_{11}-0)^2 + (v_{12}-0)^2 + (v_{13}-0)^2 + (v_{14}-0)^2} \\ D_2 &= \sqrt{(v_{21}-0)^2 + (v_{22}-0)^2 + (v_{23}-0)^2 + (v_{24}-0)^2} \\ &\vdots \\ D_5 &= \sqrt{(v_{51}-0)^2 + (v_{52}-0)^2 + (v_{53}-0)^2 + (v_{54}-0)^2} \end{aligned} \tag{3}$$

Next, D_i is sorted in a descending order. Based on the order, each regularity level is assigned to a cluster. Finally, the corresponding regularity level for each transit rider can be determined by computing and comparing the minimum distance to the center of each cluster.

A preprocessing data cleansing procedure was adopted to eliminate those smart card records with wrong transaction times; for example, a few smart card transactions were recorded as "1900/01/01". Applying the above data quality control procedure, 37,001 smart cards were randomly selected to test the proposed algorithm.

The clustered results are summarized in Table 2. If regularity levels of Very High (VH) and High (H) are considered to represent regular transit riders, approximately 41% fall into this category. The clustered results can be used to categorize different transit rider groups for various transit fare options, and provide data support for transit market analyses.

Additional individual-level daily trip and travel time information is provided in Fig. 4. Both average daily trips and average daily travel time for each passenger increased as the corresponding travel regularity became higher. On average, regular transit riders (high regularity and very high regularity) traveled more than twice per day. This is reasonable, because most regular riders take buses at some point during their daily commute. The transit authorities can focus on those transit riders with low andvery low travel regularities. The possible countermeasures include conducting customer satisfaction survey, identifying those factors influencing transit ridership and further providing additional fare discounts to attract more transit riders.

3.4. Performance enhancement using rough set theory

In Beijing, more than 16 million smart card transaction data points are generated every day. Processing and clustering such a huge amount of data is not an easy task due to the physical memory constraints of the currently available computer technology. Therefore, the K-Means++ algorithm may not be feasible for this situation without utilizing distributed computing (Cordeiro et al., 2011). To implement and execute the proposed approach in a regular personal computer, an algorithm based on therough set theory was therefore applied to improve clustering performance. The rough set theory initially proposed by Pawlak (1982) is primarily used to classify vague and uncertain data to help expert systems learn from training datasets and generate meaningful rules for classification. Unlike other commonly used data mining algorithms, rough set-based algorithms do not need any prior information about the data, such as the membership function used in the Fuzzy theory, and the Bayesian prior probability in the Naïve Bayes classifier. Rough set-based algorithms can deal with both continuous and discrete input data, and perform well under circumstances where there is missing or incomplete information. This is because rough set theories depict missing attributes using lower and upper approximations for the incomplete data, defined by probabilities (Grzymala-Busse and Grzymala-Busse,2007). Consequently, the rough set-based algorithm was deemed appropriate for dealing with the lack of boarding and alighting stop data for the flat-fare buses.

Table 2
Summary of five clusters.

Cluster center	C_1	C_2	C_3	C_4	C_5
Regularity	VL	L	M	H	VH
Cluster Center Distance	1.28	5.17	10.42	13.52	19.99
Number of Smart Cards	4809	10330	6483	9502	5877
Percentage of total	13.0%	27.9%	17.5%	25.7%	15.9%

Note: VL = Very Low; L = Low; M = Medium; H = High; VH = Very High.

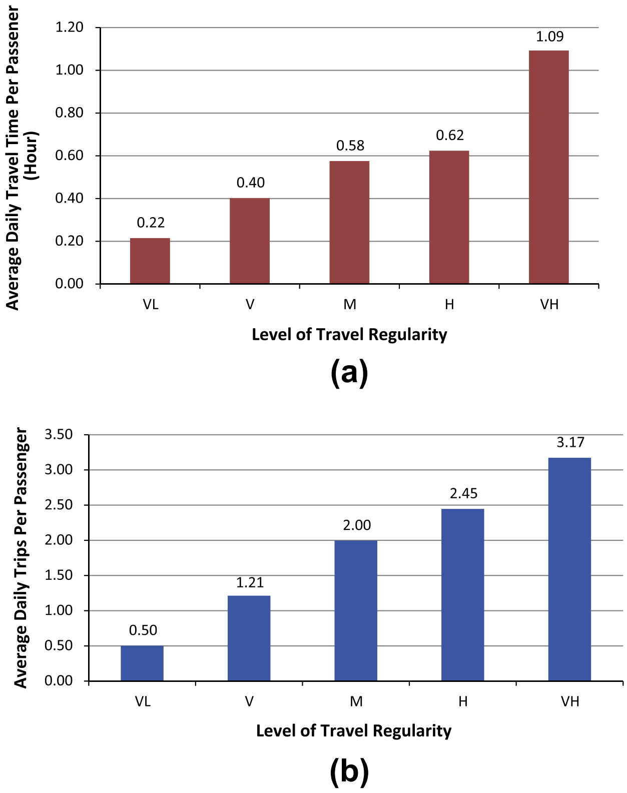


Fig. 4. (a) Individual-level average daily travel time for each cluster and (b) Individual-level average daily trips for each cluster.

The essence of the rough set-based algorithm is set approximation. Let us define any information system as: A = $(U, A \cup \{d\})$, where U represents the non-empty set of objects, or universe and A denotes the condition attributes and d the decision attributes. For our purposes, the number of travel days, the number of similar first boarding times, the number of similar route sequences and the number of similar stop ID sequences are all condition attributes, and the rider's regularity level is expressed as the decision attribute. The names of the conditionattributes and the decision attributes are considered to be the universe. The condition attributes and the decision attribute follow a many-to-one relationship: different decision attributes can be sufficiently discerned using only a subset of condition attributes. Therefore, the goal of a rough set-based algorithm is to determine the smallest number of condition attributes required to represent the decision attribute. To depict the information uncertainty and vagueness, two important concepts are described as follows. Let $B \subseteq A$ and $X \subseteq U$.

(1) $\underline{B}X = \{x|[x_B] \subseteq X\}$ is defined as the B-lower approximation of X.
(2) $\bar{B}X = \{x|[x]_B \cap X \neq \emptyset\}$ is defined as the B-upper approximation of X.
(3) $BN_B(X) = \bar{B}X - \underline{B}X$ is defined as the B-boundary region of X. If the B-boundary region is not empty, then the set X is considered "rough". (Komorowski et al., 1999).

Using the above three definitions, we can remove the superfluous attributes to leave only the equivalence classes that satisfy the minimum attributes (rules); however, finding the minimum rules is a NP-hard problem and cannot be solved in a polynomial time (Skowron and Rauszer, 1992). Fortunately, many algorithms have been proposed to obtain an optimal solution in an efficient fashion. Wróblewski (1998) developed a fast-rule induction algorithm based on a covering approach that has demonstrated both efficiency and accuracy. Its computational complexity is only $mn \log(n)$, where m is the number of universes and n is the number of attributes.

This rule induction algorithm has thereforebeen used to generate minimum decision rules in our application. Decision results classified by the K-Means++ algorithm serve as training data (as shown in Table 3), after which the rough set theory is applied to extract the hidden classification rules. Example rules can be as follows:

(1) (Number of traveling days in (5.75;7.75)) & (Number of similar boarding times in (7.25;13.25)) ⩾ (Regularity Level = High).
(2) (Number of traveling days in (17.0; Infinity)) ⩾ (Regularity Level = Very High).

Table 3
Accuracy and run time comparisons among different algorithms.

Iterations	Rough set-based algorithm		C4.5		Naïve Bayes		K-NN		Neural network	
	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)
1	98.86	59	99.31	116	87.98	64	98.57	802	97.54	120,153
2	99.01	54	99.77	119	86.34	66	99.13	798	98.12	119,868
3	99.23	69	99.60	123	87.03	65	98.96	793	96.88	123,658
4	98.92	64	99.13	118	86.99	70	99.08	826	97.65	130,147
5	98.65	59	99.00	127	85.55	68	98.97	757	98.11	121,795
6	99.19	53	98.98	133	86.78	71	99.39	788	97.96	116,583
7	99.25	60	99.13	130	87.22	69	99.12	809	97.93	130,414
8	98.97	51	99.86	121	89.53	69	98.30	825	98.21	125,478
9	99.42	56	99.75	109	88.11	65	99.35	786	97.98	123,697
10	99.26	63	99.33	111	88.02	72	99.44	779	98.14	130,186
Average	99.298	59.8	99.53	113.7	87.649	67.7	99.432	778.4	98.266	123,640.5

ms = milliseconds.

(3) (Number of traveling days in (-Infinity;2.0))&(Number of similar route sequences in (-Infinity;2.5))&(Number of the similar boarding time in (-Infinity;0.5)) ⩾ (Regularity Level = Very Low).

The rules determined by the rough set theory can then be used to classify each transit rider in terms of their level of travel pattern regularity. These rules have the added benefit that they can be easily implemented and executed in a relational database such as a Structured Query Language (SQL) database.

4. Comparison of data mining algorithms

The accuracy and efficiency of the proposed rough set-based algorithm were compared with those of several of the classification algorithms commonly used in transportation engineering research, namely the Naïve Bayes Classifier (Cestnik, 1990), C4.5 Decision Tree (Quinlan, 1993), K-Nearest Neighbor (KNN) (Cover and Hart,1967) and Three-hidden-layers Neural Network (Rumelhart and McClelland, 1986). The K-Means++ algorithm was adopted as the index algorithm for comparison, with 33% of the clustered transit riders serving as its training dataset. The rough set-based algorithm and the other four classification algorithms were applied in the training dataset to produce the corresponding classifiers. The total sample size was 37001. These classifiers were then used to process the remaining data, and the generated outputs compared to the clustered transit riders obtained using the K-Means++ algorithm to validate the accuracy of each algorithm. The entire dataset was randomly split into 33% training data and 67% test data and each algorithm executed for 10 iterations. All the algorithms were implemented in Java under an environment of a 6-core CPU and an 8 GB RAM desktop computer using the smart card data stored inMicrosoft SQL server 2008. Table 3 summarizes both the accuracy and run time (the duration taken for an algorithm to execute) statistics of all five algorithms.

The results show that the proposed rough set-based algorithm clearly outperforms the other algorithms in terms of efficiency. A t-test was conducted to evaluate the significance of the difference in accuracy between the proposed rough set-based algorithm and the other four algorithms. At a 95% confidence level, the proposed algorithm did not significantly differ from the K-NN algorithm but was 10 times faster. In addition, the proposed algorithm outperformed the Naïve Bayes and Neural Network in both algorithm accuracy and efficiency. Although the proposed algorithm slightly underperformed the C4.5 decision tree algorithm in terms of accuracy, it was twice as fast. As shown in Fig.5(a), the rough set-based algorithm demonstrated its strength in efficiency as the size of the training dataset increased. Moreover, Fig. 5(b) shows that the rough set-based algorithm would outperform the C4.5 decision tree algorithm in terms of accuracy once the size of the training dataset exceeded a certain threshold. This strongly suggests that the proposed rough-set-based algorithm is indeed suitable for handling large datasets of this type.

5. Discussion

This study opens up interesting new opportunities for leveraging smart card data to create a better understanding of transit riders' behavior and thus potentially improve public transit systems. Specifically, three major potential applications that could benefit from this study can be envisioned, asfollows:

- Travel behavior research

In the past few decades, travel demand research has shifted from a trip-based travel approach to an activity-based travel paradigm. Activity-based travel models require a substantial amount of detailed behavioral information for each traveler

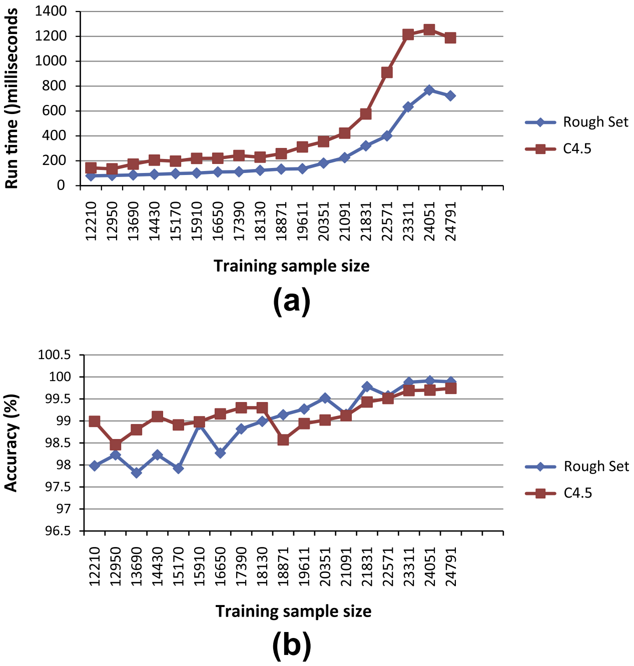


Fig. 5. Performance comparison between the C4.5 and rough set-based algorithms: (a) efficiency comparison, and (b) accuracy comparison.

extending over a relatively long period. Traditionally, this type of behavioral data has been collected using travel diaries and travel surveys, requiring an immense amount of resources to process and construct sequences of spatiotemporal activities for each traveler (Schlich and Axhausen, 2003).

The individual-level travel pattern mining algorithms developed for this study offer an alternative and novel approach for measuring the similarity and variability of transit riders through an examination of their multi-day smart card transactions. This will greatly facilitate travel behavior modeling development.

- Transit market analysis

As with other domains such as eGrocery shopping, transit agencies generally aim to develop a range of different market strategies to satisfy their passengers (Strathman et al., 2008). One typical application of transit market analysis is market segmentation (Zhou et al., 2004). Market segmentation techniques divide the entire market into several distinct segments consisting of groups of transit riders who share similar preferences and attitudes. Based on the transit rider groups identified here using the proposed travel regularity clustering algorithm, transit agencies can better allocate their limited resources to each segment to maintain and attract ridership. For example, various transit fare option can be provided that are specifically tailored for each group of transit riders. Key factors that influence transit ridership can also beidentified by integrating each market segment with transit riders' socio-demographic attributes (Krizek and El-Geneidy, 2007). For instance, most regular transit riders are commuters who do not own private cars and thus tend to be very sensitive to service reliability. In this case, improving transit service reliability (by, for example, shortening headway and providing real-time information) could be an effective measure to retain this group of transit riders.

- Transit OD estimation

Another potential use of the proposed travel pattern and travel regularity mining algorithms is to improve the accuracy of the transit OD estimation method. Each transit rider's repetitive historical routes and stops can be used as prior information for passenger alighting stop inference.

6. Conclusions

The study has proposed a series of efficient and effective data-mining approaches with which to model transit riders' travel patterns using smart card data of the type collected in Beijing, China. The DBSCAN algorithm was utilized to successfully detect each transit rider's historical travel pattern using the identified trip chains. The K-Means++ clustering algorithm and the rough set theory were then jointly applied to classify the travel pattern regularities. The performance of the resulting rough-set-based algorithmwas compared with four other classification algorithms: the Naïve Bayes Classifier, C4.5 Decision

Tree, K-Nearest Neighbor (KNN) and three-hidden-layers Neural Network. The results indicated that the proposed rough-set-based algorithm outperformed all the other data mining algorithms in terms of accuracy and efficiency.

The contribution of this study is twofold: First, a data mining approach has been proposed that is capable of identifying travel patterns for individual transit riders using a large smart card dataset. The second contribution is that the regularity levels for the data can also be successfully classified by the approach proposed here. The travel patterns and regularity levels of their customers are important information for transportation researchers seeking to understand day-to-day urban travel behavior variability and facilitate activity-based travel demand model development.

Individual travel patterns andpattern regularity also offer substantial benefits for transit agencies working to improve their transit service with the assistance of transit market analysis. Another potential application of this research is to estimate an individual transit rider's origin and destination using that rider's historical travel pattern. In terms of future work, the proposed method must now be compared with other traditional travel behavior data collection methods, such as survey studies, focus group discussions and travel diaries, in order to improve the algorithm accuracy. It would also be interesting to integrate the passenger travel pattern information obtained through this study with map-based transportation systems (Ma et al., 2011) to monitor and visualize transit performance.

Acknowledgements

The authors are grateful for the funding support from National Natural Science Foundation of China (Project No. 51138003) and China Postdoctoral Science Foundation (Project No. 2013M530018). All data used for this study were provided by the Beijing Transportation Research Center (BTRC). We are grateful to the BTRC for their data support. Special thanks to Gregory Bouche for his revision to this paper. The authors also thank the anonymous reviewers' constructive comments and suggestions.

References

Arthur, D., Vassilvitskii, S., 2007. K-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035.

Bagchi, M., White, P.R., 2004. What role for smart-card data from bus systems? In: Proceedings of the Institution of Civil Engineers: Municipal Engineer, vol. 157(1), pp. 39–46.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transport Policy 12, 464–474.

Beijing Transportation Research Center, 2010. Beijing Transportation Smart Card Usage Survey. Research Report.

Beijing Transportation Research Center, 2011. Beijing Transportation Development Annual Report, August.

Beijing Transportation Research Center, 2012. The 4th Comprehensive Transport Survey Summary Report, January.

Boyle, D.K., Foote, P.J., Karash, K.H., 2000. Public Transportation Marketing and Fare Policy. Transportation in the New Millennium, <http://onlinepubs.trb.org/onlinepubs/millennium/00093.pdf> (07.10.12).

Cestnik, B., 1990. Estimating probabilities: a crucial task in machine learning. In: Proceedings of the 9th European Conference on Artificial Intelligence, Stockholm, pp. 147–149.

Chen, J., 2009. Research on Travel Demand Analysis of Urban Public Transportation Based on Smart Card Data Information. PhD dissertation, Tongji University.

Cheng, C., Chen, Y., 2009. Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications 36, 4176–4184.

Chu, K.K., Chapleau, R., 2010. Augmenting transit trip characterization and travel behavior comprehension: Multiday location-stamped smart card transactions. Transportation Research Record: Journal of the Transportation Research Board 2183, 29–40.

Chu, K.K., 2010. Leveraging Data From a Smart Card Automatic Fare Collection System for Public Transit Planning, PhD dissertation, École Polytechnique De Montréal.

Cordeiro, R.L.F., Traina, C., Traina, A.J.M., López, J., Kang, U., Faloutsos, C., 2011. Clustering very large multi-dimensional datasets with MapReduce. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 690–698.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1), 21–27.

Dill, J., 2008. Transit use at transit-oriented developments in Portland, Oregon, Area. Transportation Research Record: Journal of the Transportation Research Board 2063, 159–167.

Dinant, J.M., Keuleers, E., 2004. Multi-application smart card schemes. Computer Law and Security Report 20 (1), 22–28.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp. 226–231.

Farzin, J.M., 2008. Constructing an automated bus origin-destination matrix using farecard and global positioning system data in São Paolo, Brazil. Transportation Research Record: Journal of the Transportation Research Board 2072, 30–37.

Federal Highway Administration, 2002. Status of the Nation's Highways, Bridges, and Transit: Conditions & Performance. <http://www.fhwa.dot.gov/policy/2002cpr/pdf/execsummary_book.pdf> (28.07.12).

Foote, P.J., Stuart, D.G., Elmore-Yalch, R., 2001. Exploring customer loyalty as a transit performance measure. Transportation Research Record: Journal of the Transportation Research Board 1753, 93–101.

Forgy, E., 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. Biometrics 21, 768–769.

Gao, L.X., Wu, J.P., 2011. An algorithm for mining passenger flow information from smart card data. Journal of Beijing University of Posts and Telecommunications 34 (3), 94–97.

Grzymala-Busse, J., Grzymala-Busse, W., 2007. An experimental comparison of three rough set approaches to missing attribute values. Transactions on Rough Sets, vol. 6. Springer, Berlin, New York, pp. 31–50.

ICF Consulting, 2003. Strategies for increasing the effectiveness of commuter benefits programs. TCRP Report 87, Transportation Research Board. Center for Urban Transportation Research, Nelson/Nygaard, ESTC.

Jang, W., 2010. Travel time and transfer analysis using transit smart card data. Transportation Research Record: Journal of the Transportation Research Board 2144, 142–149.

Kitamura, R., Yamamoto, T., Susilo, Y.O., Axhausen, K.W., 2006. How routine is a routine? An analysis of day-to-day variability in prism vertex location. Transportation Research Part A 40 (3), 259–279.

Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A., 1999. Rough sets: A tutorial. In: Pal, S.K., Skowron, A. (Eds.), Rough Fuzzy Hybridization: A New Trend in DecisionMaking. Springer, Singapore, pp. 1–98.

Krizek, K.J., El-Geneidy, A.S., 2007. Segmenting preferences and habits of transit users and non-users. Journal of Public Transportation 10 (3), 71–94.

Lee, S.G., Hickman, M., 2011. Travel pattern analysis using smart card data of regular users. Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, DC.

Ma, X., Wang, Y., Feng, C., Liu, J., 2012. Transit smart card data mining for passenger origin information extraction. Journal of Zhejiang University Science C 13 (10), 750–760.

Ma, X., Wu, Y., Wang, Y., 2011. DRIVE Net: an e-science of transportation platform for data sharing, visualization, modeling, and analysis. Transportation Research Record: Journal of the Transportation Research Board 2215, 37–49.

Mauri, C., 2003. Card loyalty. A new emerging issue in grocery retailing. Journal of Retailing and Consumer Services 10, 13–25.

McKenzie, B., Rapino, M., 2011. Commuting in the United States: 2009. American Community Survey Reports. <http://www.census.gov/prod/2011pubs/acs-15.pdf (07.10.12).

McGuckin, N., Nakamoto, Y., 2004. Trips, chains, and tours: using an operational definition. Presented at: Understanding Our Nation's Travel: National Household Travel Survey Conference, Washington DC, November 1–2.

Morency, C., Trépanier, M., Agard, B., 2006. Analysing the variability of transit users behaviour with smart card data. In: Presented at: The 9th International IEEE Conference on Intelligent Transportation Systems – ITSC 2006, Toronto, Canada, September 17–20.

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

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C 24, 9–18.

Pawlak, Z., 1982. Rough sets. Informational Journal of Computer and Information Sciences 11 (5), 341–356.

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transportation Research Part C 19 (4), 557–568.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco.

Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA, vol. 1.

Schlich, R., Axhausen, K.W., 2003. Habitual travel behavior: evidence from a six-week travel diary. Transportation 30, 13–36.

Skowron, A., Rauszer, C., 1992. The discernibility matrices and functions in information systems. In Intelligent Decision Support: Handbook of Application and Advances of the Rough Sets Theory. Kluwer, Norwell, MA, pp. 331–362.

Strathman, J.G., Kimpel, T.J., Broach, J., Wachana, P., Coffel, K., Callas, S., Elliot B., Elmore-Yalch, R., 2008. Leverage ITS Data for Transit Market Research: A Practitioner's Guidebook. TCRP Report 126, Transportation Research Board, Washington, DC.

Taylor, K.C., Jones, E.C., 2012. Fair fare policies: pricing policies that benefit transit-dependent riders. In: International Series in Operations Research & Management Science: Part 3, vol. 167, pp. 251–272.

Tirachini, A., 2012. Estimation of travel time and the benefits of upgrading the fare payment technology in urban transit service. Transportation Research Part C 30, 239–256.

Trépanier, M., Habib, K.M.N., Morency, C., 2012. Are transit users loyal? Revelations from a hazard model based on smart card data. Canadian Journal of Civil Engineering 39 (6), 610–618.

Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transportation Research Record: Journal of the Transportation Research Board 1971, 119–126.

Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y., 2004. State-of-the-art in privacy preserving data mining. ACM SIGMODRecord 33 (1), 50–57.

Webb, V., 2010. Customer Loyalty in the Public Transit Context. Masters Thesis, Massachusetts Institute of Technology.

Wróblewski, J., 1998. Covering with reducts: a fast algorithm for rule generation. In: Proceedings of RSCTC'98, Warsaw, Poland. Springer-Verlag, Berlin Heidelberg, pp. 402–407.

Zhou, T., Zhai, C., Gao, Z., 2007. Approaching bus OD matrices based on data reduced from bus IC cards. Urban Transport of China 5 (3), 48–52.

Zhou, Y., Viswanathan, K., Popuri, Y., Proussaloglou, K.E., 2004. Transit district customers in San Mateo County, California: who, why, where and how. Transportation Research Record: Journal of the Transportation ResearchBoard 1887, 183–192.