Transportation Research Part C 46 (2014) 179–191

Contents lists available at ScienceDirect

Transportation Research Part C

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

Behavioural data mining of transit smart card data: A data fusion approach

Takahiko Kusakabe *, Yasuo Asakura

Department of Civil Engineering, Tokyo Institute of Technology, Japan

ARTICLE INFO

Article history:
Received 8 November 2013
Received in revised form 22 May 2014
Accepted 22 May 2014

Keywords:
Smart card data
Data fusion
Data mining
Behavioural analysis
Naïve Bayes classifier

ABSTRACT

The aim of this study is to develop a data fusion methodology for estimating behavioural attributes of trips using smart card data to observe continuous long-term changes in the attributes of trips. The method is intended to enhance understanding of travellers' behaviour during monitoring the smart card data. In order to supplement absent behavioural attributes in the smart card data, this study developed a data fusion methodology of smart card data with the person trip survey data with the naïve Bayes probabilistic model. A model for estimating the trip purpose is derived from the person trip survey data. By using the model, trip purposes are estimated as supplementary behavioural attributes of the trips observed in the smart card data. The validation analysis showed that the proposed method successfully estimated the trip purposes in 86.2% of the validation data. The empirical data mining analysis showed that the proposed methodology can be applied to find and interpret the behavioural features observed in the smart card data which had been difficult to obtain from each independent dataset.

1. Introduction

Smart card systems have been installed as a method to collect fare of public transport. These systems automatically and continuously collect the records of travellers' use of the public transport because records of the fare payment can be regarded as travel records. For example, the transaction data enable us to observe volume of passengers at the point of ticket gates. That is; Smart card systems collect ID information of each traveller's smart card alongside fare collection. IDs enable us to analyse travel patterns such as each traveller's trip frequency, travel sections, and trip sequences. Travel patterns and their variability over long-term periods can thus be analysed (e.g. Bagchi and White, 2005 and Utsunomiya et al., 2006). Therefore, ID information of each traveller and long-term continuous observation are the advantages of the smart card data comparing to other conventional data.

Several studies have been attempted to develop methods to apply smart card data to analyses for transit management and planning. Chu and Chapleau (2008) presented methods to estimate the arrival times of a passenger at a bus stop and to identify linked trips using spatial–temporal concepts. Trépanier et al. (2007) presented a model to estimate the destination location for each individual boarding by using the smart card data. Seaborn et al. (2009) analysed multimodal journeys for information on transit planning using the smart card system in London by identifying the multimodal transfer combinations of bus-to-Underground, Underground-to-bus, and bus-to-bus. Kusakabe et al. (2010) and Asakura et al. (2012) estimated boarding trains of railway passengers by using smart card data and evaluated the effects of changes in train

* Corresponding author. Address: 2-12-1-M1-20, O-okayama, Meguro, Tokyo 152-8552, Japan. Tel./fax: +81 3 5734 2575.
E-mail address: t.kusakabe@plan.cv.titech.ac.jp (T. Kusakabe).

http://dx.doi.org/10.1016/j.trc.2014.05.012
0968-090X/© 2014 Elsevier Ltd. All rights reserved.

operations. They compared passengers' travel choice behaviour before and after the railway company altered the train timetable. Ma et al. (2013) developed an efficient data mining method to demonstrate the temporal travel patterns and the pattern regularity for transit riders in Beijing. Pelletier et al. (2011) categorised the previous studies on smart cards in public transport. They classified the usage of data into three purposes: strategic (long-term planning, behavioural analysis, and demand forecasting), tactical (service adjustments and network development), and operational (ridership statistics and performance indicators) purposes.

Continuous observations on travel patterns are important for transport operators to assess the current state of the effects of their efforts, such as timetable improvements and transit planning. Previously, the methods to collect the data on transit behaviour have mainly relied on behavioural surveys. The conventional behavioural survey methods are specialised in collecting data items for behavioural analysis and transit planning. For example, household surveys such as person trip surveys are used to gather data on all trips throughout a day with trip purposes, travel modes, actual origins, and destinations. In order to obtain dynamic changes in the travel behaviour, previous studies have developed advanced survey methods such as travel diary surveys (e.g. Pas and Koppelman, 1986; Axhausen et al., 2002), panel surveys (e.g. Kitamura, 1990), and travel surveys with tracking devices (e.g. Murakami and Wagner, 1999 and Asakura and Hato, 2004). They showed methods using the dynamic travel demand observation to identify temporal variation in travel behaviour, and to evaluate the impact of a change in the transportation system. However, continuous long-term observation is still difficult by using these specially designed surveys. Kitamura (1990) indicated that the disadvantages of a panel survey are possible increase in non-responses, problem of attrition (a decrease in the number of panels between the waves of the survey), possible decline in reporting accuracy due to panel fatigue, and problem of panel conditioning. Previous studies (e.g. Golob and Meurs, 1986; Kitamura and Bovy, 1987; Van Wissen and Meurs, 1989) pointed out that the number of responses by each respondent declines gradually during the survey period. These disadvantages are possibly found in travel data designed by other methods when the surveys are conducted for long-time period. To prevent such declines, long-term great efforts by survey staffs such as interviewers are needed to collect sufficient number of data (e.g. Axhausen et al., 2007). Although the information technologies enable us to automatically track the respondents where number of respondents in the most of tracking surveys remains less than a thousand, and duration of the tracking travel surveys is less than a few months (e.g. Murakami and Wagner, 1999; Asakura and Hato, 2004; Wolf et al., 2001, and Draijer et al., 2000). It is still difficult to collect day-to-day data for continuous long-term periods via the behavioural surveys because of cost, processing load, accuracy, and privacy protection of respondents. It means that survey based data are practically not available for monitoring long-term characteristics of transport demand. Behavioural surveys are more suitable for observing some specific factors to implement the planning and management policies rather than for continuous monitoring of travel demand.

The smart card data provide continuous and long-term travel information. However, they are fragmentary for behavioural analysis. For example, the data do not include the travellers' origins, destinations or trip purposes, or the data do not cover travellers' behaviour throughout the entire transport network. This is because the data are not collected with an explicit goal of behavioural analysis; rather, they are happened to be collected while fare collection. Trépanier et al. (2009) compared household travel survey data with smart card data. They showed that the large variations in transit network use on weekdays can be captured by using smart card data, although the smart card data are partially consistent with the travel survey data. Their result implies that the data is applicable to analyse transit behaviour if the insufficient parts of the data are supplemented or negligible. Several studies have attempted to develop methods to obtain the user segments and their behavioural contexts from behavioural patterns observed in smart card data. Agard et al. (2006) and Morency et al. (2007) estimated behavioural pattern groups and showed the variability of travellers' behavioural patterns. Kusakabe and Asakura (2011) proposed a method to identify within-day and day-to-day behavioural patterns of smart card users by a latent class model. However, the meanings of the segments, which are related to their activities, should be subjectively interpreted by analysts in these studies.

Data fusion is one of the approaches to integrate multiple data sources, which is applied in various fields, such as military applications, marketing, and intelligent transportation systems (e.g. Hall, 1992; Mitchell, 2007; Kamakura and Wedel, 1997; El Faouzi et al., 2011, and Shen and Stopher, 2013). For example, Shen and Stopher (2013) developed a trip purpose imputation method for Global Positioning System (GPS) data by using the National Household Travel Survey (NHTS) in the US. In their method, the trip purposes which were not directly observed by GPS data were estimated using rules obtained from the NHTS data. This study employs the data fusion methodology to derive relationships among behavioural attributes that cannot be obtained from either smart card data or survey based data alone. The survey based data directly observe detailed information on travel behaviour but being unable to continuously do so over a long-term period. In contrast, the smart card data provide only fragmentary information on travellers' behaviour though they can provide a continuous long-term period data which is difficult to achieve via a person trip survey. If the advantages of smart card data are combined with that of the person trip survey data, it would improve the effects of continuous monitoring of transport demands. Therefore, data fusion will allow us to obtain good understanding of the changes in travellers' behaviour over the continuous long-term period.

The aim of this study is to develop a data fusion methodology to estimate absent behavioural attributes in smart card data by using survey based data. The proposed method intends to enhance understanding of travellers' behaviour during monitoring of the smart card data. The method illustrates the relationship between original observed attributes of smart card data and the estimated attributes that are unobservable. In order to utilise the smart card data with survey based data, this study applies the naïve Bayes classifier method. A person trip survey data is employed as a survey based data to estimate the probability distribution of the naïve Bayes probabilistic model.

The model is applied to estimate the trip purposes of the travellers observed in smart card data. The trip purpose directly represents trip contexts and behavioural segments that cannot be observed by a smart card system though it is easily obtained by questionnaires of a behavioural survey. By using the prior information obtained by the survey based data, the proposed method possibly appends the attribute of the trip purpose to each record of the smart card data. The results of the data fusion enable us to analyse the continuous long-term features of the trip purpose of transit users which are difficult to be obtained from either the survey based data or the smart card data. In order to find changes in behavioural features of transit usage from the data fusion results, a data mining analysis with visualisation techniques is conducted. Data mining is an analysis step of the Knowledge Discovery in Databases (KDD) process to identify characteristic patterns in a huge database (Fayyad et al., 1996). The data mining analysis in this study focuses on finding the relationship of the estimated trip purpose and travel patterns. The result is expected to suggest specific behavioural segments which cause the changes of demand. Thus it will help transport operators to determine appropriate targets and timing of conducting their management practices on the basis of continuous observation.

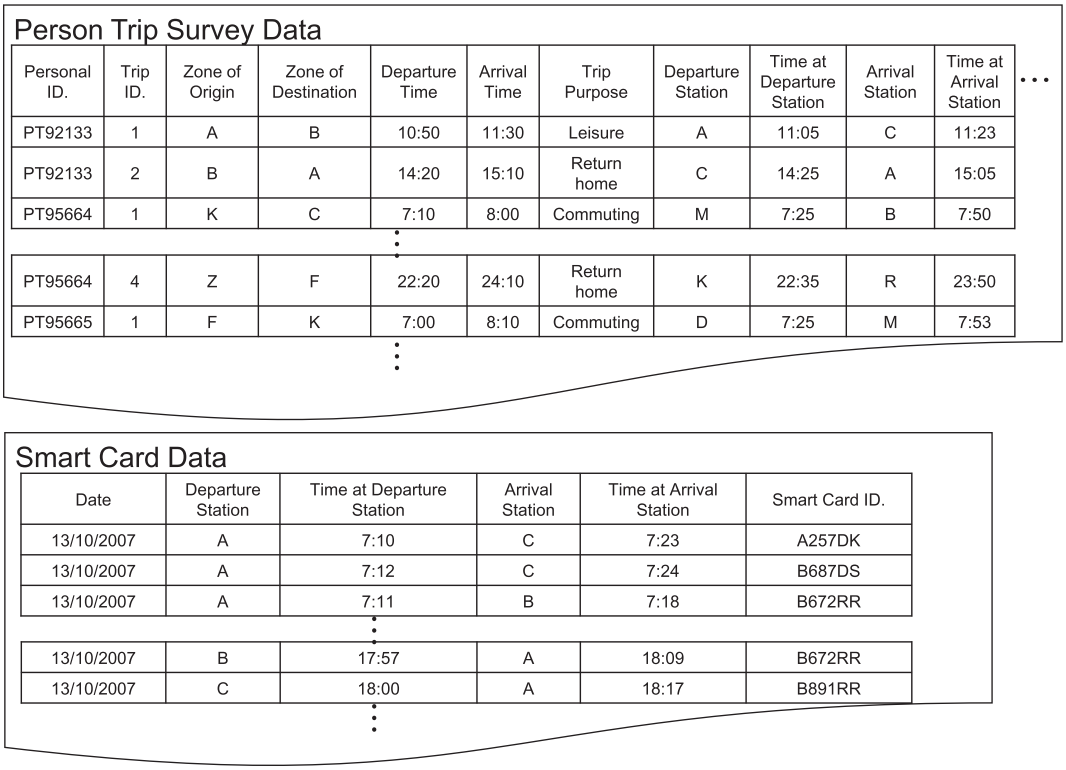
Section 2 presents the proposed data fusion methodology for smart card data and person trip survey data. Section 3 presents the empirical analysis and validation to confirm that the method can identify changes in behavioural features of transit usage observed in the smart card data. Section 4 concludes this paper.

2. Data fusion methodology

This section proposes a method for estimating the unobserved behavioural attributes by fusion of smart card data with person trip survey data. Section 2.1 shows the data structures discussed in this study. Section 2.2 describes an overview of the proposed data fusion method. Section 2.3 formulates the method using the naïve Bayes probabilistic model. Section 2.4 applies the proposed data fusion method to estimate the purpose of each trip observed in the smart card data.

2.1. Schema of smart card data and person trip survey data

Fig. 1 shows trip records stored in the smart card and person trip survey data. Records of trips using railways in the person trip survey data contain the ID of individuals, trip ID, origin and destination of the trip, departure and arrival times, and trip purpose. The boarding and alighting stations and times are also observed. The data items in the smart card data include the card ID, date, boarding station and time, and alighting station and time. The two datasets do not have common IDs. Although



Person Trip Survey Data

Personal ID.	Trip ID.	Zone of Origin	Zone of Destination	Departure Time	Arrival Time	Trip Purpose	Departure Station	Time at Departure Station	Arrival Station	Time at Arrival Station	• • •
PT92133	1	A	B	10:50	11:30	Leisure	A	11:05	C	11:23	
PT92133	2	B	A	14:20	15:10	Return home	C	14:25	A	15:05	
PT95664	1	K	C	7:10	8:00	Commuting	M	7:25	B	7:50	
				⋮							
PT95664	4	Z	F	22:20	24:10	Return home	K	22:35	R	23:50	
PT95665	1	F	K	7:00	8:10	Commuting	D	7:25	M	7:53	
				⋮							

Smart Card Data

Date	Departure Station	Time at Departure Station	Arrival Station	Time at Arrival Station	Smart Card ID.
13/10/2007	A	7:10	C	7:23	A257DK
13/10/2007	A	7:12	C	7:24	B687DS
13/10/2007	A	7:11	B	7:18	B672RR
		⋮			
13/10/2007	B	17:57	A	18:09	B672RR
13/10/2007	C	18:00	A	18:17	B891RR
		⋮			

Fig. 1. Person trip survey data and smart card data.

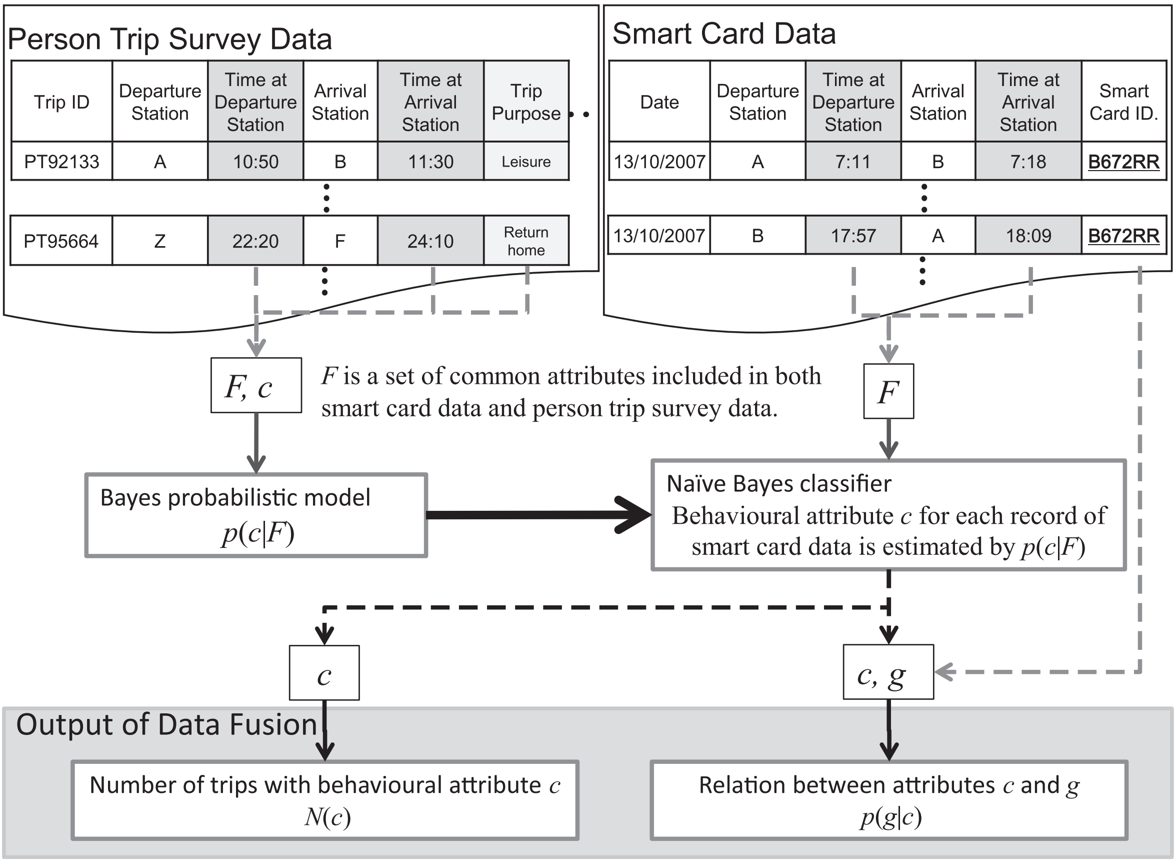
these two datasets are obtained separately, both the smart card data and the person trip survey data contain information on the boarding and alighting stations and times.

The information on the boarding and alighting stations and time is utilisable for combining the two datasets. However, these attributes are not exactly identical in the following two points. First, the precision of the information is different. In the smart card data, exact minute is recorded when a traveller passes through the gate at the station. On the other hand, boarding and alighting time in the person trip survey data is reported after the trip. The information in person trip survey is sometimes not so accurate and travellers possibly report time rounded by 5 or 10 min because the information depends on the travellers' memories. This affects the time resolution of the discrete variables that is used in the proposed method. Second, the IDs of the smart card data are not always identical to the individuals. For example, a smart card can be shared among family members or a traveller can hold more than one card. However, such usage may not be majority especially where registered monthly passes that specifies user is integrated into a smart card, which is in service in Japan.

2.2. Overview of data fusion methodology

Fig. 2 is the flow chart of the proposed data fusion method. The concept of data fusion is to estimate complementing elements of smart card data and person trip survey data. The behavioural attribute c represents the attributes only observed in the person trip survey data, such as trip purpose, origin, and destination. They directly represent behavioural contexts. These detailed behavioural attributes are comparatively easy to be obtained by questionnaires. The attributes F represents the commonly observed behavioural attributes that are included in both datasets, such as boarding stations and times. The behavioural attribute g in the figure is derived only from smart card data, such as trip frequency. A continuous collection method is needed to derive the attribute g.

The proposed method provides number of trips with attribute c in the smart card data. Also the method provides relationship between attributes c and g that cannot be obtained from either dataset alone. This study assumes that the trips observed in the smart card data have the same conditional probability distribution $p(c|F)$ as the trips in the person trip survey data. This distribution represents the probability where a traveller has behavioural attribute c at the gate of the station when he/she has attribute F. Distribution $p(c|F)$ itself is estimated from the person trip survey data. This assumption possibly causes unsuccessful estimation results when the penetration rate of the smart card is low. This is because the smart card


Person Trip Survey Data

Trip ID	Departure Station	Time at Departure Station	Arrival Station	Time at Arrival Station	Trip Purpose	• •
PT92133	A	10:50	B	11:30	Leisure	
			⋮			
PT95664	Z	22:20	F	24:10	Return home	

Smart Card Data

Date	Departure Station	Time at Departure Station	Arrival Station	Time at Arrival Station	Smart Card ID.
13/10/2007	A	7:11	B	7:18	B672RR
			⋮		
13/10/2007	B	17:57	A	18:09	B672RR

F is a set of common attributes included in both smart card data and person trip survey data.

Behavioural attribute c for each record of smart card data is estimated by $p(c|F)$

F, c
F
Bayes probabilistic model
p(c|F)
Naïve Bayes classifier
c
c, g
Output of Data Fusion
Number of trips with behavioural attribute c
N(c)
Relation between attributes c and g
p(g|c)

Fig. 2. Overview of data fusion methodology.

holders are not the representative samples of the entire travellers. The more frequent users (e.g. commuters) possibly have the higher probability of owning smart cards. For example, when attributes c and F respectively represent trip purpose and arrival time, and F perfectly explains c, attribute c of smart card users can be successfully estimated even if only commuting travellers own smart cards. When F does not perfectly explain c, it will cause under estimation of certain trip purposes because some trips are estimated as other trip purposes. This is because the trip purposes with similar features are possibly assigned to these trip purposes with higher shares. However, the effects of this bias can be ignored when the penetration rate becomes higher.

By applying the probability distribution $p(c|F)$ to the naïve Bayes classifier, attribute c is added to each trip in the smart card data. As the result of estimating attribute c of the smart card data, some relation of the attributes is derived. Time series changes in $N(c)$ that is the number of trips with behavioural attributes c can be derived by counting the trips with estimated attribute c. The relationship between attributes c and g can be summarised by $p(g|c)$, where $p(g|c)$ is the conditional probability distribution of g when c is given. This distribution is used to see whether the distribution of g is characterised by c. If the distribution of g has relationship with c, the different distributions are obtained by corresponding values of c.

2.3. Formulation of naïve Bayes probabilistic model

One of the requirements for the data fusion method of the smart card data with person trip survey data is a low calculation load because the volume of the smart card data observed over a long time period grows very large. This study employed one of the simplest data fusion methodologies with a low calculation load, namely the naïve Bayes classifier (Rish, 2001). The naïve Bayes classifier is a classification method that estimates the absent attributes of the data using the naïve Bayes probabilistic model. The naïve Bayes probabilistic model treats all the variables as discrete variables to avoid presuming the distribution form of $p(c|F)$ and to easily treat the originally discrete variables such as trip purposes. However, when the variables are discretely treated, the number of data in the person trip survey data is possibly insufficient to describe the relation in all the combinations of observed variables. To reduce the required number of the data, the naïve Bayes classifier assumes that each element of F is conditionally independent of every other element of F when c is given.

Let a vector $F = \{f_1,f_2,\ldots,f_K\}$ be a set of the behavioural attributes. Each element of F represents a common data attribute of the smart card and person trip survey data, such as boarding stations and boarding time. Let c be a variable of an absent data item in the smart card data that can be observed using the person trip survey data. In this study, we treated c and each element of F as only discrete variables. By using Bayes' theorem, $p(c|F)$ is described by

$$p(c|F) = \frac{1}{p(F)}p(c)\prod_{k=1}^{K}p(f_k|c) \tag{1}$$

where $p(c)$, $p(F)$ and $p(f_k|c)$ are probability distributions estimated from the person trip survey data. The distributions $p(c)$ and $p(F)$ are derived from the composition rate of trips having attributes c and F respectively. The conditional probability distribution $p(f_k|c)$ is derived from the proportion of trips having attribute f_k corresponding to each value of attribute c.

When F of each trip is observed by the smart card system, the behavioural attribute c of each trip is estimated by using the naïve Bayes classifier. The equation of the classifier is represented by

$$\hat{c}(F) = \arg\max_{c\in C}\, p(c|F) \tag{2}$$

where C is a set of all the possible values of c.

Note that $p(F)$ is regarded as constant when F of each trip is given by the data because it does not depend on c. By using Eq. (2), the number of trips with behavioural attribute c is described by

$$N(c) = \sum_{F\in S}\delta(c,F)N_s(F) \tag{3}$$

where

$$\delta(c,F) = \begin{cases} 1 & \mathit{if} \quad \hat{c}(F) = c \\ 0 & \mathit{otherwise} \end{cases},$$

$N_s(F)$ is the number of trips with a vector of attributes F that is observed by smart card system, S is a set of all the possible values of F.

In this equation, $\delta(c,F)$ is used to be intended to derive consistent number of trips with estimation results by Eq. (2) which is used for disaggregate estimation. This definition of number of trips possibly causes biases when behavioural attributes F does not perfectly explain c because the trip purpose with the highest probability is attached to each trip of the smart card data. By using $p(c|F)$ instead of $\delta(c,F)$, expected number of the trips with less biases will be derived though it is not consistent with the disaggregated results.

When an absent variable g of the person trip survey data is observed by the smart card data, the distribution of absent variable c of each absent variable g can be estimated by Bayesian inference. The joint probability of trips whose attributes are c and g is estimated by

$$p(c,g) = \sum_{F \in S} p(c|F)p_s(F,g) \tag{4}$$

where $p_s(F,g)$ is derived from the composition rate of trips having $\{F, g\}$, which are observed in the smart card data. Then, the distribution of g in each value of c is calculated as the posterior distribution using the person trip survey and smart card data. This is described by

$$p(g|c) = \frac{p(c,g)}{p(c)} = \frac{\sum_{F \in S} p(c|F)p_s(F,g)}{\sum_{F \in S} p(c|F)p_s(F)} \tag{5}$$

where $P_s(F)$ is composition rate of trips with a vector of attributes F that is observed by smart card system.

2.4. Estimation of trip purpose

Trip purpose is a fundamental attribute describing a behavioural context of a trip that is observed in the person trip survey data. However, it cannot be observed in the smart card data. In order to estimate the trip purpose of each trip with smart card data, the purpose of each trip is defined as the attribute c in the model proposed in Section 2.2. Trip purpose c is defined in the person trip survey data as:

$$c \in \{\text{'commuting to work'}, \text{'commuting to school'}, \text{'leisure'}, \text{'business'}, \text{'returning home'}\} \tag{6}$$

Note that the business purposes represent the trips where travellers travel between workplaces and other places except their homes such as their clients' offices.

The attribute g is defined as trip frequency that represents the trip patterns obtained from the smart card data. Trip frequency g is one of the useful factors to analyse the changes in the travel demand because the total number of trips is affected by the number of travellers as well as the number of trips conducted by each traveller.

This study assumes the following trip attributes vary among the trip purposes; time of day of trip, duration of stay at the destination, and actual destination of the trip. This assumption is set because these attributes correlate with the activities at the destination (e.g. Shen and Stopher, 2013.) Based on the assumption, the behavioural attributes F that commonly appear in smart card and person trip survey data are defined as $F = \{f_a, f_s\}$ where f_a is 'alighting time' and f_s is 'duration of stay'. The variable f_a is to represent the time of day of trip. The variable f_s is defined as an interval between the alighting and the next boarding time at the same station. This variable implicitly represents the duration of stay at the destination in addition to the travel time between the alighting station and the actual destination. This travel time includes access and egress time with different modes such as taxis and buses.

The alighting time f_a and duration of stay f_s are discrete variables. They are defined by every hour. One reason to treat the time in an hour interval, which is definitely larger than the accuracy in the smart card data, is that the time recorded in the person trip survey data is not as accurate as that of smart card data as described in Section 2.1. Other reason is to accumulate sufficient number of samples in each time interval for estimating the probability distribution $p(c|F)$.

In this study, we employed the alighting time f_a and duration of stay f_s as the simplest attributes correlating with the trip purposes. In the proposed model, any attributes can be included in F as long as they appear both in smart card and person trip survey data. For example, boarding station is one of the possible attributes. However, there might be a problem when we use this attribute because the number of samples who travelled between some stations and the target station in person trip survey seems to be insufficient for the estimation of the probability distribution $p(c|F)$. In this case, aggregation methods (e.g. clustering) of variables are required to estimate the distribution. Therefore, utilising attributes which requires supplemental processes are an issue to be solved in the future.

Fig. 3 shows the conceptual representation of the estimation target and the behavioural attributes F in the space–time dimensions. In the figure, the trip for the estimation which alights at the station A, is represented by a bold line. The trip purpose c of this trip is estimated from the alighting time f_a and the duration of stay f_s by using probability distribution $p(c|F)$. Note that the duration of stay f_s is available when the ID information of a traveller is provided since the data of this variable is defined from two consecutive trips.

3. Empirical analysis

In this section, we describe an empirical analysis using the method proposed in Section 2.4. The smart card data was obtained at a single railway station. The data are employed in order to find variation of trip purposes of passengers at this station. Although the findings in this section are station specific characteristics, it is intended to show the proposed methodology is applicable to help transit operators to find reasons for fluctuation of the number of passengers. Section 3.1 explains the datasets. Section 3.2 represents the estimation results of the probability distributions described by Eq. (1) which is derived from the person trip survey data. Section 3.3 describes the validation analyses that employ a subset of the person trip survey data. By comparing the estimated trip purpose with the actual trip purpose obtained by the person trip survey, the precision of the estimation is examined. Section 3.3 describes the similarity of the trip features between successful and unsuccessful estimation. Section 3.4 applies the proposed method to the actual smart card data to find long-term changes in traveller usage of stations.

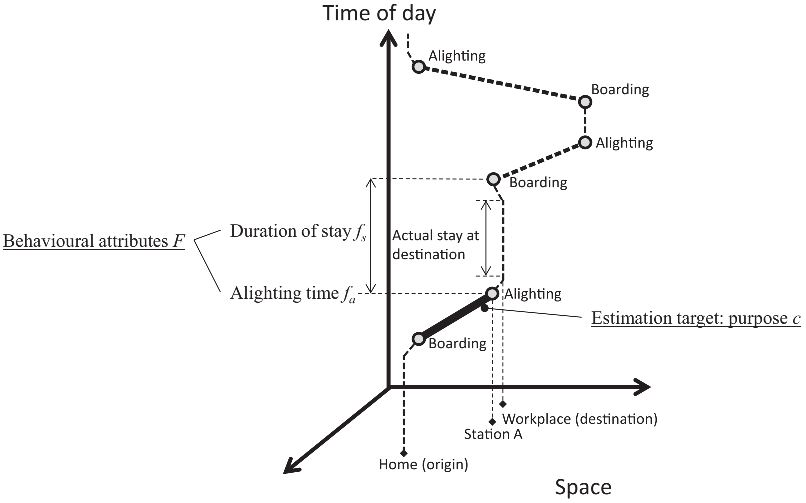


Fig. 3. Space–time representation of estimation targets and behavioural attributes *F*.

3.1. Datasets

This study employs the data obtained at a single railway station. The target railway station for the analysis is run by a private company, and the station is one of the major ones in the Osaka area that is the second largest metropolitan area in Japan. The station has many train connections to various districts that are operated by more than one railway companies. Some railway lines run in parallel; travellers can choose their boarding train from amongst several railway lines. However, the data on other railway companies were not available for this analysis. The data for estimating the probability distributions of trip purposes were person trip survey data obtained in 2002: the '4th Kei-han-shin Metropolitan Area Person Trip Survey.' The data on the trips alighting at the target station operated by the same railway company as the smart card data were used to estimate the model. The duration of stay is calculated with the alighting time and the boarding time at the target station.

The smart card transaction data to which the proposed methodology was applied were obtained from a railway operator. The contents of the transaction data records are described in Section 2.1. It was known that about 10% of the total passengers of the railway company were smart card holders. Since the railway company allowed the use of the smart card data only for research purposes, the individual ID information had been anonymised prior to the analysis. The privacy of smart card holders was strictly protected throughout this study.

The person trip survey data for travellers who alighted at the target station contained 1586 trips made by 1576 travellers. The dataset is randomly divided into the two subsets, namely estimation and validation datasets. From all the person trip data, 1,095 trips were used for estimating the probability distribution $p(c|F)$. Other 491 trips of the data were used for validation in Section 3.3.

Section 3.4 employs the actual smart card data. All the person trip survey data alighting at the target station were used for the estimation of the model. This analysis used the smart card data of travellers who alighted at the target station at least once during the data collection period. The smart card data for the analysis covered 7,074,768 trips made by 553,259 travellers observed in 20 months from October 2007 to May 2009. All the travellers made at least one trip during the period.

3.2. Estimation results of probability distributions

The naïve Bayes probabilistic model described by Eq. (1) is estimated by a subset of the person trip survey data. The estimation subset consists of 1095 trips as described in Section 3.1. The probability distribution $p(c|F)$ is the probability of observing a certain trip purpose when the trip has the attribute $F = \{f_a, f_s\}$ as explained in Section 2. The model performance depends on whether the differences in these attributes indicate the difference of the trip purposes. The two distributions $p(f_s|c)$ and $p(f_a|c)$ are estimated separately because the naïve Bayes classifier assumes that each element of F is conditionally independent of other elements of F.

Table 1 shows $p(f_s|c)$, and Table 2 shows $p(f_a|c)$ that were estimated using 1095 trips of the person trip survey data. They represent the composition of values of duration of stay and arriving time at the station at every corresponding trip purposes. According to Tables 1 and 2, the trips that commute to work and school showed relatively similar trends. Most of the travellers taking these trips arrived in the morning and had longer duration of stay than travellers on leisure and business trips. The commuting to work and school trips completed before 11 a.m. were 98.2% and 97.2% respectively. However, trips of commuting to work had longer duration of stay than those of to schools. Leisure and business trips showed identical features. They were taken during the daytime. However, the features of the duration of stay are different. One-way business

Table 1
Estimated values of $p(f_s|c)$ from person trip survey data at the target station.

f_s (h)	c				
	Commuting to work	Commuting to school	Leisure	Business	Return to home
0	0.000	0.000	0.021	0.000	0.000
1	0.002	0.000	0.075	0.088	0.000
2	0.002	0.014	0.103	0.027	0.000
3	0.000	0.000	0.133	0.083	0.000
4	0.000	0.040	0.120	0.014	0.000
5	0.010	0.051	0.073	0.018	0.000
6	0.014	0.071	0.036	0.020	0.000
7	0.014	0.125	0.038	0.000	0.000
8	0.030	0.269	0.014	0.023	0.000
9	0.132	0.151	0.035	0.000	0.000
10	0.225	0.131	0.005	0.014	0.000
11	0.165	0.050	0.015	0.076	0.000
12	0.123	0.023	0.013	0.038	0.000
13	0.093	0.013	0.000	0.025	0.000
14 and above	0.111	0.025	0.000	0.000	0.000
No return trip	0.079	0.036	0.319	0.574	1.000

Table 2
Estimated values of $p(f_a|c)$ from person trip survey data at the target station.

f_a (h)	c				
	Commuting to work	Commuting to school	Leisure	Business	Return to home
5–6	0.032	0.011	0.006	0.051	0.000
7–8	0.781	0.611	0.067	0.115	0.000
9–10	0.152	0.253	0.224	0.214	0.004
11–12	0.017	0.097	0.207	0.158	0.000
13–14	0.010	0.028	0.196	0.163	0.031
15–16	0.003	0.000	0.137	0.191	0.166
17–18	0.004	0.000	0.140	0.094	0.391
19–20	0.002	0.000	0.022	0.014	0.287
21–22	0.000	0.000	0.000	0.000	0.113
23–24	0.000	0.000	0.000	0.000	0.009

trips occupied 57.4%. In other words, these travellers did not appear again at the station on the same day, while those of the leisure trips were only 31.9%. These one-way trips represent the trips alighting at the target station and did not appear again at the station on the same day. These trips may use other stations for their return trips or stay their destination overnight. However, they were identically treated in this study in order to have consistency with the smart card data which can be observed at the station operated by the same railway company. Trips returning home increased after 3 p.m. and most of them were one-way trips. Trips returning home made up 80.0% of the total trips after 5 p.m.

3.3. Validation with person trip survey data

In order to validate the proposed data fusion method, this section examines the model by the validation subset of the person trip survey data. The validation data consists of 491 trips observed by the person trip survey as described in Section 3.1. The data includes both attributes c and F and all the data were observed in a day. By comparing the estimated trip purpose with the actual observation of the trip purpose, we validated whether the trip purpose was correctly estimated by Eq. (2).

Fig. 4 shows the estimation results of trip purposes by Eqs. (2) and (3). The number of the actual trips is the one appeared in the validation data. The number of the estimated trips is the one estimated by Eq. (3). The number of the successfully estimated trips represents the one whose actual purpose is identical to the estimated purpose by Eq. (2). The trip purposes in the figure are defined in Eq. (6). In total, 76.8% trips were correctly estimated. Among these five purposes, 82.5% of the commutes to work and 84.8% of trips returning home were correctly estimated, and they were better estimated than trips for other purposes. The correctly estimated leisure trips were 58.9%. However, only a few commutes to schools and business trips were correctly estimated; that is, 37.5% of the commutes to schools and none of the business trips were correctly estimated. One of the possible reasons for this low precision is that the numbers of these trips are quite few; no more than 2 business and 16 commutes to school trips were observed. Another reason is that some different trip purposes show similar characteristics such as duration of stay and arrival time. This means that it may be difficult to distinguish commutes to school from to work, and business trips from leisure trips in terms of these characteristics.

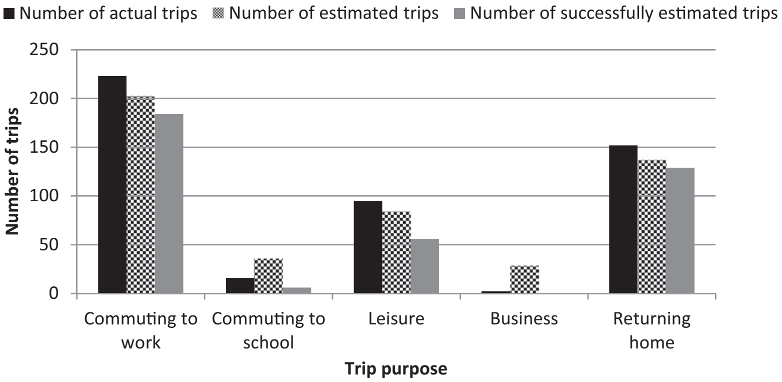


Fig. 4. Number of trips at the target station correctly estimated by trip purposes.

Trip purposes c should be characterised by the behavioural attributes F in order to obtain proper estimation results. However, the probability distributions shown in Tables 1 and 2 illustrated that 'commuting to school' and 'commuting to work' have the similar tendencies in terms of the arrival and the duration of stay. Also, the business trips have the similar tendencies to leisure ones. The similarity of the trip features seemed to cause the unsuccessful estimations shown in the above analysis. That is, the trip purposes with the similar features are undistinguishable by the behavioural attribute F. It means that the estimation results of these purposes may have some biases.

As described in Section 1, the estimated results are intended to be used for the Knowledge Discovery in Databases (KDD) process. The trip purposes which are not successfully estimated may cause wrong discoveries in the KDD process. In order to avoid such wrong discoveries, the trip purposes which are successfully characterised by the behavioural attribute F should be distinguished. Therefore, we reorganised the similar trip purposes of 'commuting', 'leisure and business', and 'returning home'. The redefined trip purposes are described as:

$$c \in \{\text{'commuting to work or school'}, \text{'leisure or business'}, \text{'returning home'}\} \quad (7)$$

Fig. 5 shows the number of trips correctly estimated when the trip purposes are redefined as Eq. (7). As shown in this figure, the number of successfully estimated trips increased. In total, 86.2% of the trips were correctly estimated. Of the total, 92.1% of the commutes, 74.2% of the leisure and business trips, and 84.5% of the trips returning home were correctly estimated. This result shows that organising trip purposes by similarity on trip characteristics improved the estimation results, although we are now unable to differentiate two purposes that might not share a distribution of trip frequency.

3.4. Application to data mining of actual smart card data

This section describes the application of the proposed model defined in Section 2.4 to actual smart card data observed for over 20 months. The intention of the analyses was to find the characteristics of day-to-day changes in the behavioural features. First, we discuss the day-to-day change in the number of trips for each trip purpose as estimated by Eq. (3). In order to find the within-day characteristics of the changes, the visualisation method which is one of the data mining techniques, was applied to the estimation results. We conclude this section by showing the month-to-month changes in the distribution of the trip frequency as derived from Eq. (5). This analysis shows the relationships between the estimated trip purposes c and trip frequency g collected from the smart card data. These relationships might illustrate the characteristics of changes in the

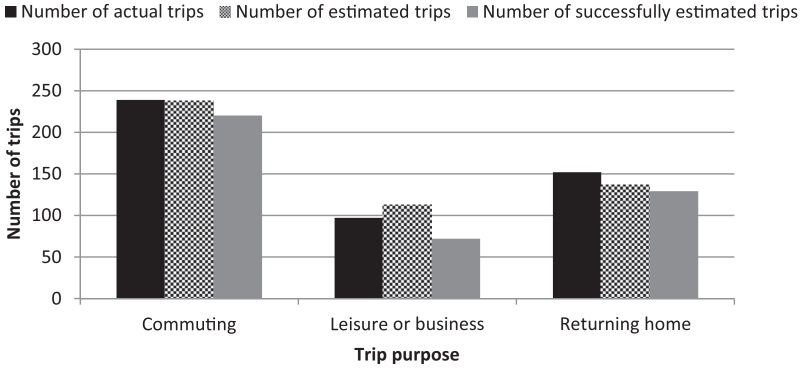


Fig. 5. Number of trips at the target station correctly estimated when trip purposes were reorganised.

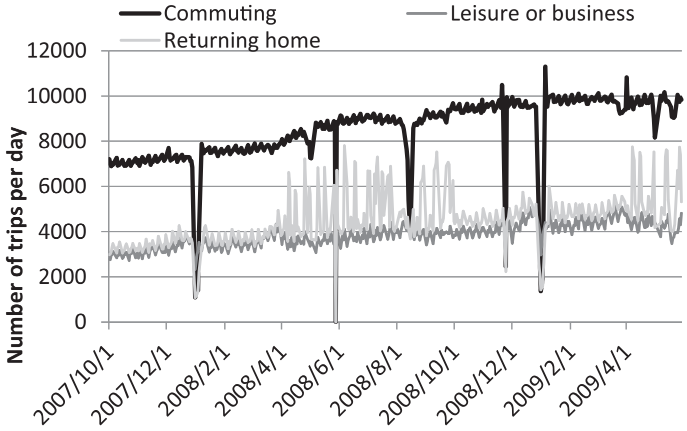
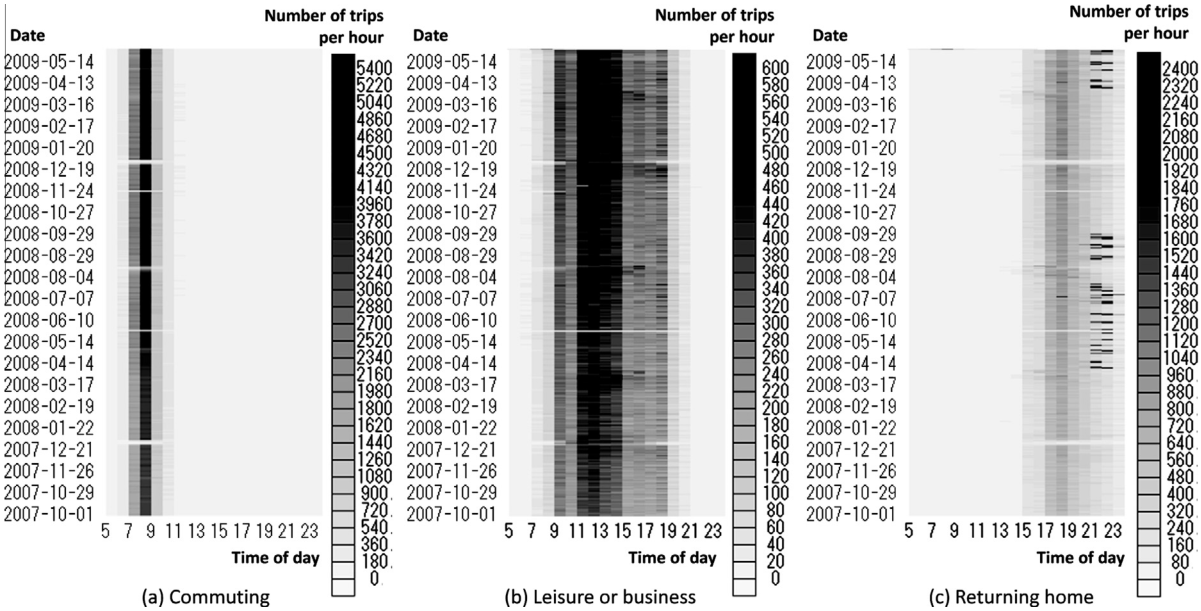


Fig. 6. Day-to-day changes in the number of trips at the target station.

demand because the changes in the number of trips are affected by the changes in the number of travellers as well as the number of trips made by each traveller.

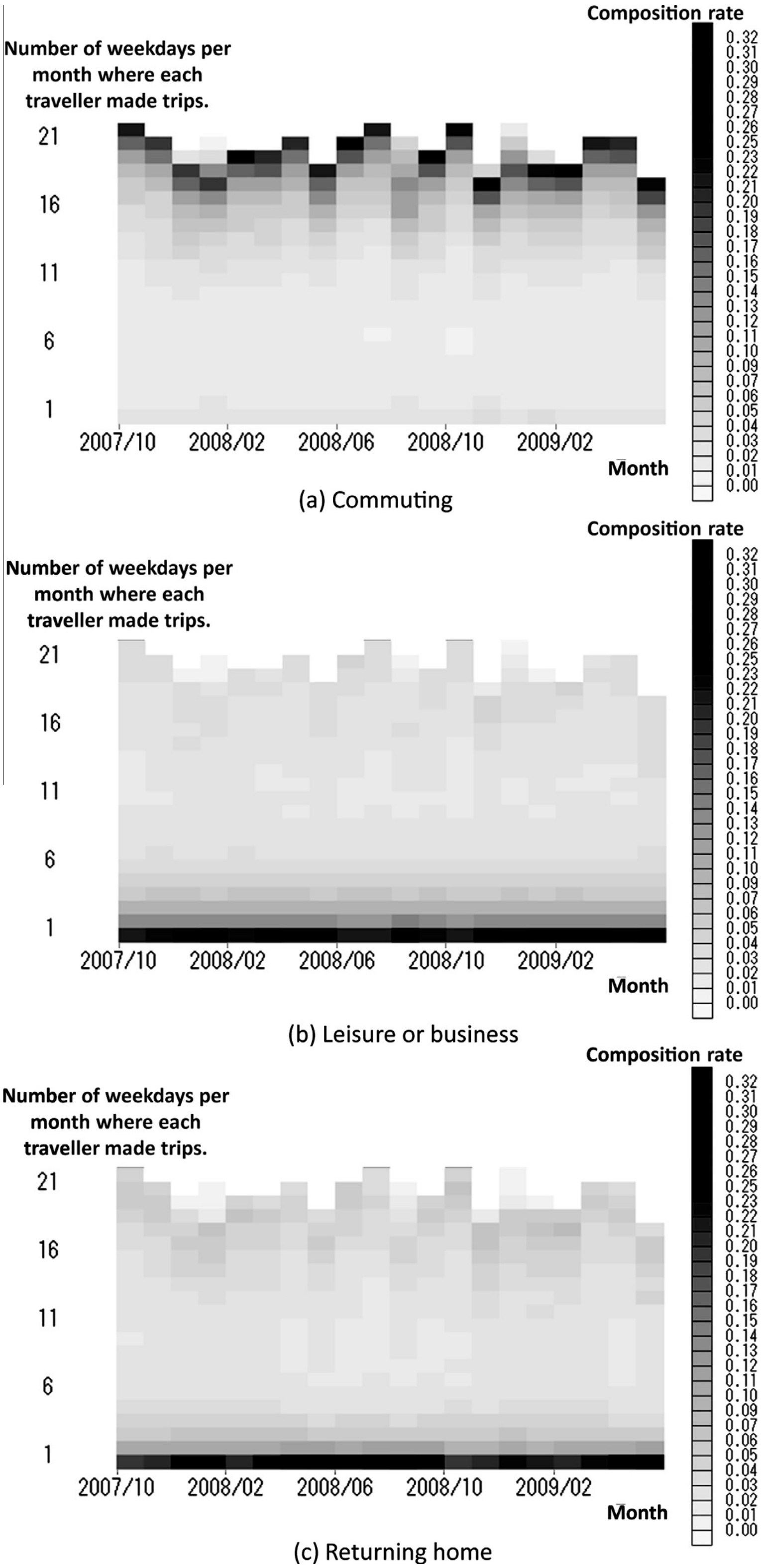
Fig. 6 shows the day-to-day changes in the number of trips of each purpose as estimated by Eq. (3). The numbers of trips of each purpose were estimated for each day for 20 months. This figure shows the number of the commuting travellers alighting at this station is twice as large as "return home" or "leisure or business" travellers because the target station is located at a central business district (CBD). The "return home" trips possibly included trips transferring to other rail line operated by different railway companies. Although we cannot distinguish whether the changes were caused solely by variations in demand or in the penetration rate of smart cards, several different characteristics of changes were found for different trip purposes. For example, the figure shows that the increase in commutes was larger than that in the other purposes. Average increase in commuting trips between October 2007 and May 2009 was 2553 trips per day. On the other hand, the increase in the leisure trips was 1193 trips per day. Commutes and leisure trips were found to be stable compared with the return home trips, although there was a large decline in commutes during the summer and New Year holiday seasons. Wide variations in returning home trips were found in summer. This might be affected by the travellers who wanted to transfer to other railway lines on the way to their homes from the stadium located at the other station in the target line. The wide variations coincided with the days when the baseball games were held.

Fig. 7 shows the visualisation of the number of trips of each estimated trip purpose according to the date and time. The horizontal axis in these figures indicates the time of day, and the vertical axis indicates the date. The grey scale represents



(a) Commuting (b) Leisure or business (c) Returning home

Fig. 7. Number of trips at the target station corresponding to time and date.



(a) Commuting

(b) Leisure or business

(c) Returning home

Fig. 8. Histograms of the number of the weekdays where each traveller made the trips in each month at the target station.

the number of trips. Several features in the number of trips at a particular time of day are shown by the figures. The number of commutes was higher before 9 a.m. throughout the period. The leisure trips increased from 11 a.m. to 3 p.m. They also increased at 5 p.m. Trips returning home drastically increased after 5 p.m. The number of trips depending on the time of day was strongly affected by the model because one input of the estimation model was the arrival time at the station. However, some characteristics may be derived from the duration of stay. Intermittent increases in trips returning home after 9 p.m. occurred from April to September and these increases corresponded to the time that baseball games ended. This characteristic was quantitatively confirmed after the finding in the data mining analysis. On 94% of the days when baseball games were held, the number of travellers between 8 p.m. and 9 p.m. increased over 20% compared to the average. On the other hand, the number of travellers in other days did not exceed 1.2 times of the average. By the Welch's t test, t-value was estimated as 14.1, that is, the number of travellers in the days of the baseball games was significantly larger than that of the usual. This result was one of the examples to find events from the data without presumption of the events. The capability of the finding of the events without the presumption will help to know the period and stations affected by events in advance of conducting the detailed surveys.

Fig. 8 shows the distributions of trip frequency for commuting and leisure trips respectively which were estimated by Eq. (5). These results show time series changes in trip purposes which are difficult to obtain from person trip survey data. The vertical axis shows the number of the days per month where each traveller made trips at the station. The horizontal axis is the month. The grey scale indicates the composition rate of each number of the days in each month when the travellers used the station. These figures show the month-to-month changes of trip frequency of each trip purpose. In most of the months, the number of the weekdays in each month corresponds to the number of the days that the most of commuting traveller made their trips. However, the frequency of commutes decreased in the holiday seasons: namely, August and December. The figures also show that the variability of the frequency tended to increase while the frequency itself decreased in these seasons. These results imply that the large decrease in the number of commutes represented in Fig. 6 was caused by the trip frequency. On the other hand, most leisure trip travellers were observed once a month. This result insists that the changes in the number of the leisure trips are caused by changes of the number in travellers.

The analysis in this section illustrated that the characteristics of the changes vary among estimated trip purposes. The results suggest that the proposed data fusion method possibly helps transport operators to interpret the cause of the changes during the monitoring of travellers. It enabled them to analyse the relationship between the original observed attributes of the smart card data and the estimated attributes that are originally unobserved. Especially, the relationship among trip frequency, number of travellers, and trip purposes are represented in the analysis. These results demonstrated the travellers' segments who contribute to the changes of the transit demand. The results also showed that the changes are caused by either trip frequency or number of the travellers. The frequency cannot be obtained from the cross sectional data such as the person trip survey data alone because continuous long-term observation is necessary to derive the frequency.

4. Conclusions

We proposed a data fusion methodology for analysing the behavioural features observed by smart card system with the person trip survey data. The method can help transport operators to monitor and data mine traveller behavioural features observed in the smart card data. The method uses the naïve Bayes classifier that is one of the simplest classification procedures. The purpose of each trip in smart card data is estimated from the arrival time at the station and time interval between the arrival and next departure at the same station—called the duration of stay. The duration of stay is available when ID information of a traveller is available and the traveller made more than one trip using the same station operated by the same railway company.

Validation using the subset of the person trip survey data, as shown in Sections 3.2 and 3.3, demonstrated that the proposed method correctly estimated 76.8% of the trip purposes. However, incorrect estimations were caused by trips of other purposes with similar characteristics in terms of duration of stay and arrival time at the station. When we redefined the trip purposes with similar features in order to distinguish the trip purposes which are successfully characterised by the behavioural attribute F, 86.2% of the trip purposes were correctly estimated.

The empirical data mining analysis in Section 3.4 showed that the proposed method was capable of helping us to find and interpret the behavioural features observed in the smart card data. The proposed method illustrated the share of trip purposes and the relationship between the trip frequency and the trip purpose which could not be obtained from either smart card data or person trip survey data alone. The method was applied to the data mining analysis on the smart card data observed for 20 months. The results showed some features in long-term changes. For example, some noticeable changes in the summer seasons were found. The proposed method showed the changes were caused by the return home trips. This interpretation provided us the supplemental information to assume the changes were affected by return trips from other station where a nearby stadium held a baseball game. The relationship between the trip frequency and the trip purpose showed how the trip purposes affected the total number of the trips. The results showed the different features in each trip purpose. Another feature was changes of the trip frequency and the number of travellers. The total number of the trips was affected by the trip frequency or number of travellers. In the holiday seasons, the commuting travellers affected the total number of trips by reducing their trip frequency. On the other hand, the changes in the number of the leisure trips are caused by changes of the number in travellers.

The proposed method can be applied to estimation of other variables such as actual origins and destinations (OD) by defining OD related variables as c whenever the behavioural attributes F are observed in both the smart card data and survey based data. In order to obtain proper estimation results, the attribute *c* that describes the estimation target should be characterised by the behavioural attribute *F* as shown in Section 3.3.

One of the advantages of the proposed methodology is that it employs a secondary data collected for other purposes by other entities; that is, smart card data are obtained for fare payment by railway operators and person trip survey data are collected for transport planning by Ministry of Land, Infrastructure, Transport and Tourism. This means that there requires no further costs to collect the data for this analysis. The method allows transport operators to continuously monitor and overview the transport demand without adding costs to obtain a survey data.

By applying the proposed method to the continuous monitoring, transport operators are able to make assumptions about the cause of behavioural changes. For example, this study represented the travellers' segments who contributed the demand changing by supplementing the trip purposes. Continuous monitoring will help them to find the magnitude of effects by implementing of policy changes as well as the spreading speed of them. It will also help to determine appropriate survey targets and timing of conducting the surveys. However, the observation period of the dataset used in this study did not include changes of transport operators' measures and policies. In the future, the proposed method will be applied to actual assessment of specific operational improvements, fare revision and transit planning.

Acknowledgement

We would like to express the deepest appreciation to Kiko YAMADA-KAWAI who gives us insightful comments and suggestions.

References

Agard, B., Morency, C., Trépanier, M., 2006. Mining public transport user behaviour from smart card data. In: 12th IFAC Symposium on Information Control Problems in Manufacturing – INCOM 2006, Saint-Etienne, France, 17–19 May 1996.

Asakura, Y., Hato, E., 2004. Tracking survey for individual travel behaviour using mobile communication instruments. Transp. Res. Part C: Emerg. Technol. 12 (3–4), 273–291.

Asakura, Y., Iryo, T., Nakajima, Y., Kusakabe, T., 2012. Estimation of behavioural change of railway passengers using smart card data. Public Transp. 4 (1), 1–16.

Axhausen, K.W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., Haupt, T., 2002. Observing the rhythms of daily life: a six-week travel diary. Transportation 29 (2), 95–124.

Axhausen, K.W., Löchl, M., Schlich, R., Buhl, T., Widmer, P., 2007. Fatigue in long-duration travel diaries. Transportation 34 (2), 143–160.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12 (5), 464–472.

Chu, K.K.A., Chapleau, R., 2008. Enriching archived smart card transaction data for transit demand modeling. Transp. Res. Rec. 2063, 63–72.

Draijer, G., Kalfs, N., Perdok, J., 2000. Global positioning system as a data collection method for travel research. Transp. Res. Rec. 1719, 147–153.

El Faouzi, N.-E., Leung, H., Kurian, A., 2011. Data fusion in intelligent transportation systems: progress and challenges – a survey. Inform. Fusion 12 (1), 4–10.

Fayyad, U., Shapiro, G.P., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag. 17 (3), 37–54.

Golob, T.T., Meurs, H., 1986. Biases in response over time in a seven-day travel diary. Transportation 13 (2), 163–181.

Hall, D.L. (Ed.), 1992. Mathematical Techniques in Multisensor Data Fusion. Artech House, Norwood, MA, USA.

Kamakura, W.A., Wedel, M., 1997. Statistical data fusion for cross-tabulation. J. Mark. Res. 35 (4), 485–498.

Kitamura, R., 1990. Panel analysis in transportation planning: an overview. Transp. Res. Part A: Gener. 24 (6), 401–415.

Kitamura, R., Bovy, P.H.L., 1987. Analysis of attrition biases and trip reporting errors for panel data. Transp. Res. Part A: Gener. 21 (4–5), 287–302.

Kusakabe, T., Asakura, Y., 2011. Behavioural data mining for railway travellers with smart card data. In: Second International Workshop on Traffic Data Collection and its Standardisation, Brisbane, Australia, 22–23 September 2011.

Kusakabe, T., Iryo, T., Asakura, Y., 2010. Estimation method for railway passengers' train choice behavior with smart card transaction data. Transportation 37 (5), 731–749.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C: Emerg. Technol. 36, 1–12.

Mitchell, H.B. (Ed.), 2007. Multi-Sensor Data Fusion – An Introduction. Springer-Verlag, Berlin, Germany.

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

Murakami, E., Wagner, D.P., 1999. Can using global positioning system (GPS) improve trip reporting? Transp. Res. Part C: Emerg. Technol. 7 (2–3), 149–165.

Pas, E.I., Koppelman, F.S., 1986. An examination of the determinants of day-to-day variability in individuals' urban travel behavior. Transportation 13 (2), 183–200.

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

Rish, I., 2001. An empirical study of the naïve Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, Washington, USA, 4–10 August 2001.

Seaborn, C., Attanucci, J., Wilson, N., 2009. Analyzing multimodal public transport journeys in London with smart card fare payment data. Transp. Res. Rec. 2121, 55–62.

Shen, L., Stopher, P.R., 2013. A process for trip purpose imputation from Global Positioning System data. Transp. Res. Part C: Emerg. Technol. 36, 261–267.

Trépanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. Syst. 11 (1), 1–14.

Trépanier, M., Morency, C., Blanchette, C., 2009. Enhancing household travel surveys using smart card data. In: Paper Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, DC, 11–15 January 2009.

Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transp. Res. Rec. 1971, 119–126.

Van Wissen, L.J.G., Meurs, H.J., 1989. The Dutch mobility panel: experiences and evaluation. Transportation 16 (2), 99–119.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. Transp. Res. Rec. 1768, 124–134.