FINDING BEHAVIORAL RULES OF URBAN PUBLIC TRANSPORT PASSENGERS BY USING BOARDING RECORDS OF INTEGRATED STORED FARE CARD SYSTEM

Toshiyuki Okamura, Junyi Zhang, Akimasa Fujiwara
Graduate School for International Development and Cooperation, Hiroshima University
1-5-1, Kagamiyama, Higashi-Hiroshima, Hiroshima, 739-8529, Japan
e-mails: tokamura@hiroshima-u.ac.jp, zjy@hiroshima-u.ac.jp, afujiw@hiroshima-u.ac.jp

Abstract

This paper focuses on boarding records of automatic fare collection system with stored fare card that is introduced in public transport in a certain metropolis in Japan, and aims to clarify the applicability of the data by integrating all operators' record to urban city planning and operators' marketing, and to propose practical applications for data utilization such as the analysis of passengers' transfer behavior and classifying passenger based on behavioral characteristics and finding behavioral rules for marketing by using data mining methods. Although this study stands in the first step, this will be the first opened research to focus on electric fare collection system data in urban public transport system in transportation planning field and to propose the practical utilization of the data for urban transport planning and marketing.

Keywords: Stored fare card system; Public transport; Survey method
Topic Area: D5 Data Collection Methods

1. Introduction

Automatic fare collection system with stored fare card (magnetic card, IC card etc.) is widely introduced in many Japanese urban public transport operators. In the number of metropolitan areas, different public transport operators in the city adopt the same automatic fare collection system, and passengers are able to ride on multiple lines/operators within the city area including different modes (train, bus, AGT etc.) by using only one card. The most revolutionary features of the system is that complete and whole continuous boarding records (date, time, place, transfer, fare etc. for several weeks) of each 'cardholder' (passenger) inside a city can be easily collected by lower costs, as well as passengers' convenience and reducing operators' administrative costs. However, the records are not utilized for urban transport planning by local governments because each operator owns its data and no organizations handle whole data of all operators in the metropolises except revenue adjustment and allocation to each operator. Operators also do not make use of the records for their marketing nor revising their level of service. They only utilized the data for fundamental aggregation such as counting their revenue and making OD tables.

This paper focuses on boarding records of the automatic fare collection system that is introduced in public transport in the Hiroshima metropolitan area, Japan. The system is carefully designed for operators' administrative affairs such as counting and adjusting revenue between different operators, however, the interests of transport planners are the data accuracy of getting on and off place and time in day in each boarding, and the traceablity of each passenger's travel history by matching the same card issue number. This paper aims to empirically clarify the applicability of the data by integrating all operators' record to urban city planning and operators' marketing, and to propose practical applications for data

utilization such as the analysis of 1) passengers' transfer behavior (exact transfer place and waiting time are difficult to be collected by person trip survey nor interview surveys), and 2) categorizing passenger behavioral characteristics by using long term (one month) data (this is also difficult to be collected by any conventional surveys) and 3) finding behavioral rules for marketing by using data mining methods (CHAID).

2. Outlines of the fare collection system in Hiroshima

The automatic fare collection system was introduced to 6 private bus operators in the Hiroshima metropolitan area in March 1993. Currently almost all bus operators (10 operators), two railway operators (tramways and Automated Guideway Transit (AGT)) and one ferry operator in this area implement this system, which cover all public transport operators except JR (West Japan Railway: suburban and intercity railways) in Hiroshima. This is a decentralized system and each operator directly manages the system. Each operator issues and sells their own companies' cards, which can be commonly used in all operators. The boarding records are concentrated and calculated, and the revenue of selling card and deducted value by card readers are adjusted between operators.

The boarding share of the operators (modes) which implement this system is approximately 70% (number of trip base) and around half of the total passengers pay their fare with the card (Fig. 1 and 2). The number of passengers using the card amounts to more than 1.5 million per week.

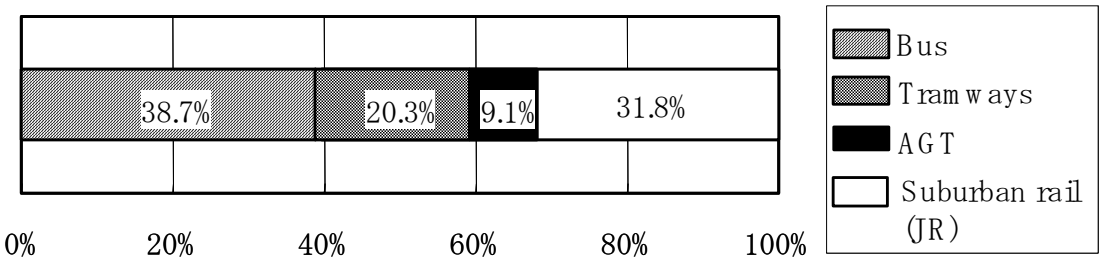


Figure 1. Share of each transit mode in Hiroshima City (1998)

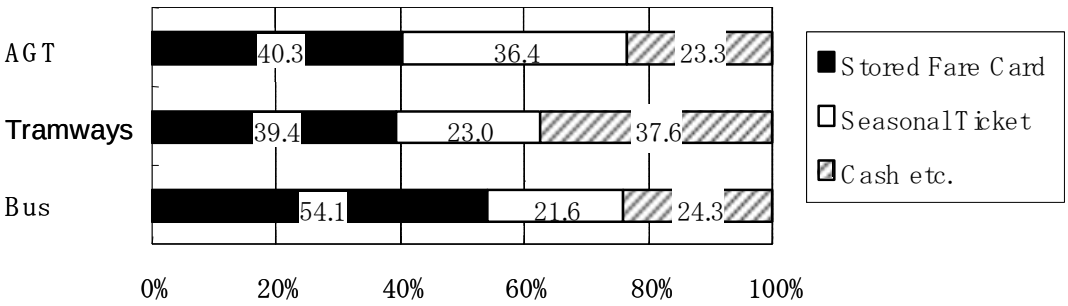


Figure 2. Ways of fare payment in different transit modes (1998)

The system in Hiroshima adopts magnetic card. There are three types of cards, including 1000 yen, 3000 yen and 5000 yen card, and these cards are pre-charged with additional 10% premium. The cards have no limited valid period, and are not rechargeable after the amounts of stored value go to zero. The Passenger is requested to insert the card into the slot of the card readers both at the entrance and at the exit of bus/tram vehicle (or into the entry and exit ticket gates at AGT stations) for each boarding. The fare is automatically deducted at the exit readers or ticket gates. If the remaining value in the card is insufficient for

the trip, passengers pay their insufficient amounts using cash or another card. Transfer rebate (20 yen) is available for within 60 minutes transfer of bus to bus and bus to tramway.

The card readers and the ticket gates record passengers' detail travel records on entry and exit date/time period and place (station or bus-stop zone), boarding line and operator, card issue number etc.. These data saved in a memory cartridge in each card reader or ticket gate are concentrated to a server in a data processing company for aggregating the amounts of adjustment to the revenue from selling card and boarding record of card users between different operators. If each operator's disaggregated records were integrated, continuous travel records (date, time, place, transfer etc.) of each 'cardholder' (passenger) inside a city could be grasped until the remaining value goes to zero by matching card issue numbers. However, passengers' disaggregate travel records have not been integrated because a private operator is unwilling to open such raw data to other operators, which are actual or potential competitors.

The outlines of the obtained data for this analysis are shown in Table 1. The authors asked all operators to show their boarding records for research purpose. Only 20% of the total records are permitted to be shown for the purpose of this study. One operator's data (AGT) was not obtained because the data format is different from others and it is impossible to integrate the disaggregate records.

Table 1. Data for analysis

Duration	Oct. 1 to Oct. 31, 2000
Operators	All public transport operators implemented this system except AGT
Major data items	Card issue number Getting on and off date, time, place (station or bus stop) Operator and line ID code Amount of remaining value in card (before and after boarding) Payment code (normal, stored value reaches down to 0, for adjustment of insufficient fare)
Sampling rate	20% (card issue number base)
Number of Samples	1,124,605 (boarding number base)

3. Practicability of data for transport planning and marketing

This data collection system is carefully designed for operators' administrative affairs such as counting and adjusting revenue between different operators, however, the interests transport of planners are the data accuracy of getting on and off place and time in day in each boarding, and the traceablity of each passenger's travel history by matching the same card issue number. This chapter focuses on the data missing concerning getting on and off place and time in day, and check the traceablity by counting passengers who hold more than one cards at the same time (because it is impossible to capture their continuous moving history). Boarding records (23,589 rides) of a certain day from the original samples are shown in Table 1 and is applied for this analysis.

3.1 Data missing concerning getting on and off place and time in day

Table 2 shows a number of data missing samples in a day categorized by missing data items. The case *fare in a trip missing* happens when remaining value of a trip is insufficient in a card and a passenger pays the rest of the fare by cash or another card at the

exit. In this case, time and place in getting on and off are recorded. *Boarding time/place missing* happens when a driver manually input the amounts of the fare for deducting in the case, for example, when a passenger fail to slot his/her card into readers in getting on a transit.

Table 2. Number of data missing samples in a day

Number of total boarding record	23,589	(100.0%)
No data missing	20,263	(85.9%)
Fare in a trip missing	1,504	(6.3%)
Boarding time/place missing	1,293	(5.5%)
Other missing (only payment recorded etc.)	524	(2.2%)

The percentage of *no data missing* and *fare in a trip missing* records, which include full information of getting on and off place and time in day, arrive up to 92.2% of the total. This shows that the data has little concern about data accuracy resulted from data missing.

3.2 Traceabily of each card holder's travel history

In this system, passengers can adjust their insufficient fare by card or cash when remaining value in a card reaches down to 0 in getting off transit. The record of this system enables to distinguish 'remaining value reaches down to 0' or 'adjust insufficient fare'. This classification is applied to estimate a percentage of passengers who hold more than one card by comparing the number of passengers whose remaining value in a card reaches down to 0 and those who adjust the balance by partially used card. The proportion of the two will be equivalent to that of passengers who hold more than one card. The numbers of those are shown in table 3.

In table 3, (B)-(C)-(D) will be the number of passenger adjusted insufficient fare by cash. (D/B) is equivalent to the percentage of passengers who hold more than one partially used card when the value of the one of the card goes to zero. The number of (D/B) is the maximum of the percentage those who hold more than one card at the same time, because passengers may purchase and use a new (unused) card when the remaining value near to zero and they may hold more than one partially used cards just before the value of another card is used up. This case has little influence on traceablity because the period having more than one card can be short. However, this number of percentage means that not a few passengers hold more than one card, and analysts have to recognize the limitation of traceablily by matching card issue number.

Table3. Percentage of passengers who hold more than one card

Number of total boarding record	(A)	23,589
Remaining value in a card reaches down to 0	(B)	1,667
Adjust insufficient fare by unused card	(C)	871
Adjust insufficient fare by partially used card	(D)	354
(D/B) (Percentage those who hold more than one card)		21.2%

4. Aggregate analysis making use of the data

4.1 on-board time distribution of transit (comparison with person trip data)

Figure 3 and Figure 4 show passengers' on-board time distribution of tramways calculated both from the card data and the Hiroshima Person Trip Survey (PT Survey) conducted in 1987. The distributions of the time in Figure 3 and Figure 4 are aggregated every one minute and five minutes respectively. These results show that while passengers percept

and answer their on-board time in multiples of 5 or 10 minutes in questionnaire surveys, the card data can obtain accurate on-board time. Smaller metropolises like Hiroshima where average trip length is not so long need more accurate on-board time data as basic statistics for transport planning.

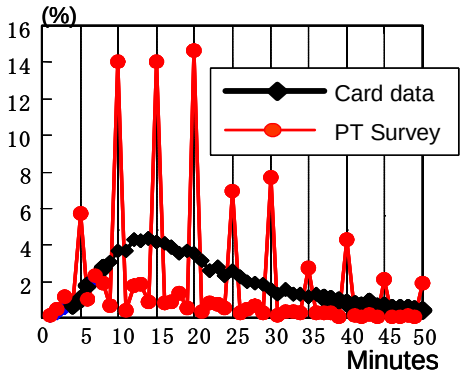


Figure 3. On-board time distribution of Tramway (aggregated by one minute)

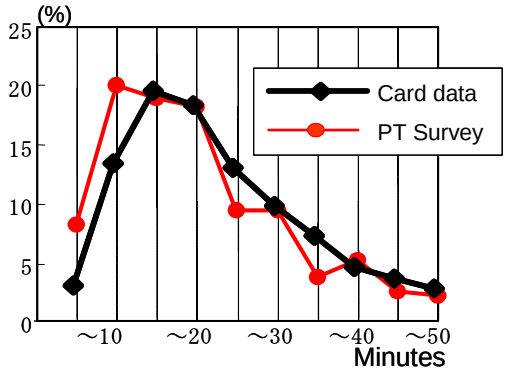


Figure 4. On-board time distribution of Tramway (aggregated by five minutes)

4.2 Number of passengers transferred and waiting time in major transfer points

Questionnaire surveys and interview surveys are difficult to acquire accurate information on transfer in public transport. For example the Person Trip Surveys in Hiroshima have inquiries for departure time at the origin, unlinked travel time, transfer place and arriving time at the final destination, however, no inquiry about waiting time for transfer. The integrated card data can obtain accurate transfer place and waiting time at the place.

Figure 5 shows the ratio of arrival passengers to transfer passengers at the major four transfer points in Hiroshima, and figure 6 shows the ratio of the total number of transfer trips in Hiroshima area to those in the major transfer points. The definition of transfer is that between public transports including different operators within 60 minutes waiting time at the same place. The card issue number is used for matching two trips before and after the transfer. The location of the four major transfer points are shown in Figure 7 and the characteristics of the four points are follows;

- Kamiyacho/Hatchobori :CBD area, commercial and business center, suburban bus terminal located
- Hiroshima Station :The central terminal of suburban and intercity rail (Shinkansen), commercial area
- Koi :West side terminal of railway, neighborhood commercial area
- Dobashi :Transfer point of tramway, business and residential area

Figure 5 and 6 shows the characteristics of the four points. Kamiyacho/Hatchobori (CBD area) has smaller ratio of transfer passengers in the point, while the ratio of those to the total transfer trips in Hiroshima area amounts to more than 40 %. These four points shares more than 80 % of the transfer passenger in the metropolis.

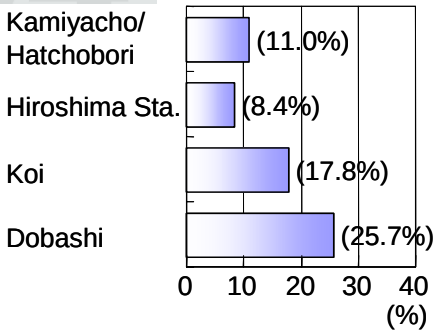


Figure 5. The ratio of arrival passengers to transfer passengers in each point

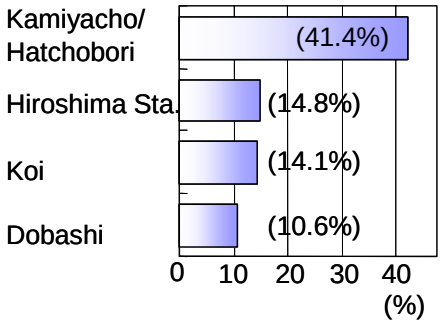


Figure 6. The ratio of the total transfer trips to transfer passengers

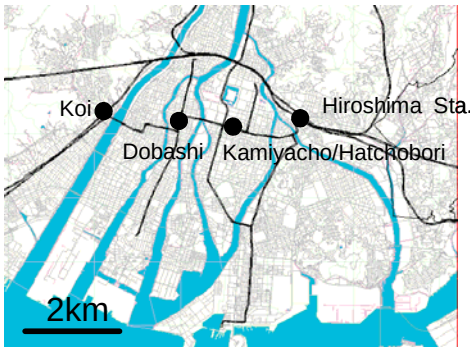


Figure7. Location of the Transfer Points

Figure 8 shows the distribution of waiting time during peak time (before 10:00) and off-peak time (10:01 to 15:00) in the two points; CBD area and small terminal. In peak time, the ratio less than 5 minutes is smaller in Kamiyacho/Hatchobori (CBD area). Considering that transit frequencies in both points are also high (a few minutes), it is suggested that the difference occurs on the longer transfer distance in Kamiyacho/Hatchobori area. Focusing on the difference between peak and off-peak, Kamiyacho/Hatchobori area has more long waiting time passengers especially more than 30 minutes. This can be a result of two reasons. First, many suburban bus routes start from Kamiyacho/Hatchobori and the bus frequency is longer than city bus. Secondly, this area has commercial complexes and many passengers go around the commercial area. The card data can apply the evaluation of easier transfer design or attractiveness of the area for activating the city center.

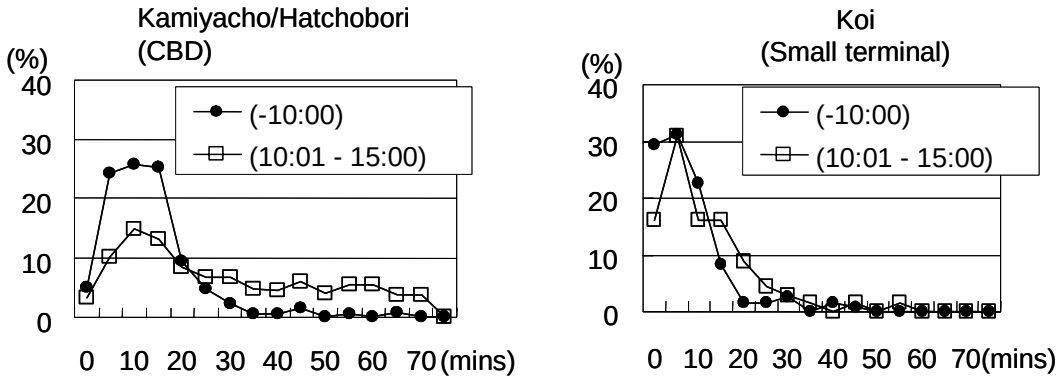


Figure 8. Distribution of Waiting Time during Peak and Off-peak time in the two points

5. Passenger classification based on behavioral characteristics

5.1 Data and classifying method

The authors classify the passengers (card holders) into groups in which passengers' boarding characteristics are similar in terms of monthly boarding frequency, boarding time period in day, average on-board time and average payment per each ride. For this analysis, boarding records of passengers are aggregated in each card holder. The data for this analysis are shown in table 4. One boarding is defined as one trip from the origin to final destination including transfer within 60 minutes.

Exhaustive CHAID (Chi-squared Automatic Interaction Detector) method, one of the most popular data mining methods for marketing segmentation, is applied for classification. This method grows multi-split decision tree and find the best split by any stopping rules. It merges values that judged to be statistically homogeneous with respect to the target valuable and maintains all other valuables that are heterogeneous. Then, select the best predictor variables to form the first branch in the decision tree, such that each node (node means a set of samples classified by decision rules on the upper levels) is made of a group of homogeneous values of the selected variable, and continues this process recursively. It works for all types of variables. For merging values, an *F* test is used for continuous variables and a chi-squared test is used for categorical variables.

5.2 Classified passenger characteristics

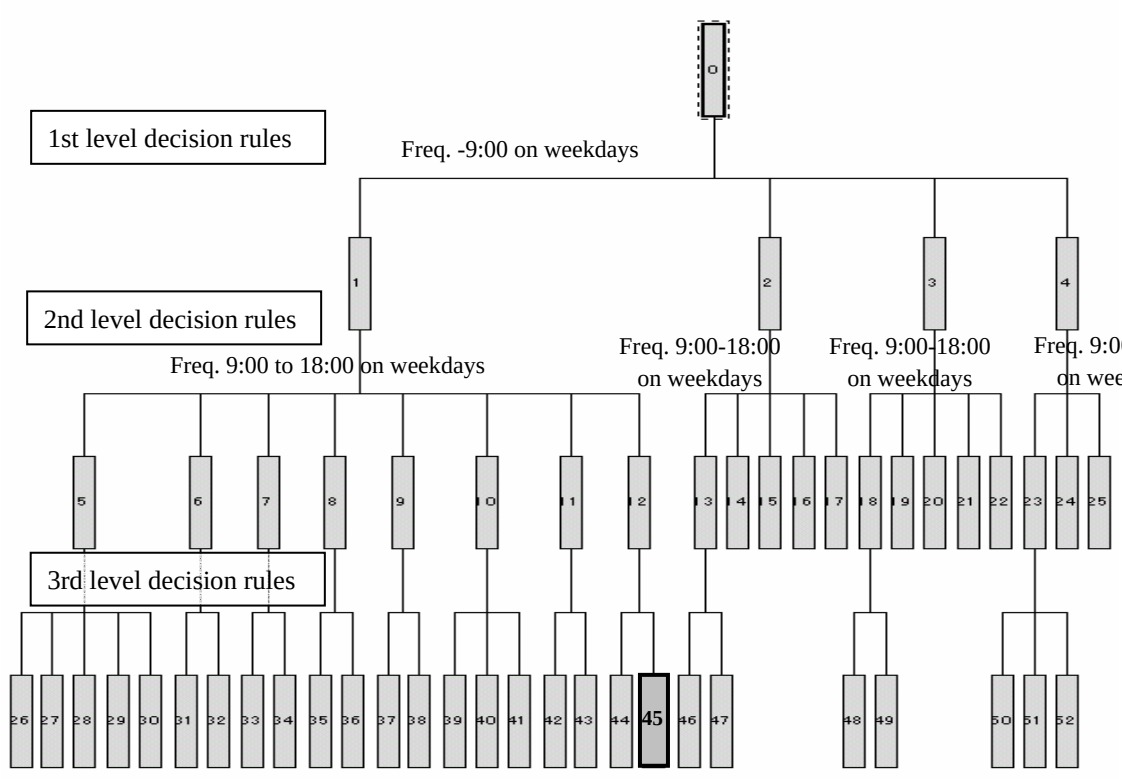
The target variable is *boarding frequency per month per passenger* and predictor variables are the data items shown in table 4. Figure 9 illustrates a decision tree that describes classification based on the data items shown in table 4, and table 5 shows the decision rules to classify into each node.

Table 4. Data for classifying passengers

Number of samples (card holder base)		8,735
Sampling rate		2.0%
Duration		Oct. 1 to Oct. 31, 2000
Data items	per each passenger	-Boarding frequency per month* before 9:00, 9:00 to 18:00 and after18:00 both on weekdays and on weekend -Average on-board time per each ride -Average payment per each ride both on weekdays and on weekend

*The data is converted to monthly boarding frequencies in cases where the stored values of cards are used up in less than one month.

In the first split, four child nodes are classified based on the variable *boarding frequency per month before 9:00 on weekdays*. The node 45, for example in Figure 9 and Table 5, is classified according to the 1st level decision rule *boarding frequency per month before 9:00 on weekdays*, has a 0 frequency value (similar to node 1), and then classified by the 2nd level rule *boarding frequency per month from 9:00 to 18:00 on weekdays* according to the frequency value of more than 12.3, and then classified again by the 3rd level rule *boarding frequency per month from 9:00 to 18:00 on weekend* having a frequency value of more than 4.



*Squares and inside numbers: nodes and node numbers

Figure 9. Decision tree overview of classifying card holders

Table 5. Decision rules of passengers based on boarding frequency

Node No. 0 (Total) Average Freq. 10.4 (/month), Variance 12.3, Number of total holders 8735 (100%)

Node No. 1 1st level decision rule : Freq. -9:00 in weekdays [0] (/month) Ave. Freq. 7.4 (/month), Variance 9.6 Number 5982 (68.

Node Number	2nd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
5	9:00-18:00 in weekdays [0]	3.6	6.0	1704	19.5%
6	9:00-18:00 in weekdays (0,1]	1.9	1.8	818	9.4%
7	9:00-18:00 in weekdays (1,1.9]	4.5	3.7	282	3.2%
8	9:00-18:00 in weekdays (1.9,2.8]	3.6	2.6	789	9.0%
9	9:00-18:00 in weekdays (2.8,4.2]	5.5	3.5	699	8.0%
10	9:00-18:00 in weekdays (4.2,6.6]	8.2	4.0	523	6.0%
11	9:00-18:00 in weekdays (6.6,12.3]	14.0	6.3	620	7.1%
12	9:00-18:00 in weekdays (>12.3]	29.0	12.7	547	6.3%

Node Number	3rd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
26	9:00-18:00 in weekends [0]	2.6	4.3	615	7.0%
27	9:00-18:00 in weekends (0,1.5]	2.0	2.1	492	5.6%
28	9:00-18:00 in weekends (1.5,2.3]	3.2	3.4	275	3.2%
29	9:00-18:00 in weekends (2.3,4.0]	4.8	3.2	194	2.2%
30	9:00-18:00 in weekends (>4.0]	14.1	14.4	128	1.5%
31	9:00-18:00 in weekends [0]	1.4	1.1	667	7.6%
32	9:00-18:00 in weekends (>0]	4.0	2.4	151	1.7%
33	18:00- in weekdays [<=1.2]	3.1	1.8	167	1.9%
34	18:00- in weekdays (>1.2]	6.5	4.7	115	1.3%
35	18:00- in weekdays [<=1.2]	3.0	1.8	688	7.9%
36	18:00- in weekdays (>1.2]	7.2	4.2	101	1.2%
37	18:00- in weekdays [<=1.2]	4.7	2.3	591	6.8%
38	18:00- in weekdays (>1.2]	9.9	5.3	108	1.2%
39	9:00-18:00 in weekends [0]	6.7	2.9	279	3.2%
40	9:00-18:00 in weekends (0,2.3]	8.0	2.6	132	1.5%
41	9:00-18:00 in weekends (>2.3]	12.4	4.6	112	1.3%
42	18:00- in weekdays [<=2.6]	11.7	3.6	476	5.5%
43	18:00- in weekdays (>2.6]	21.5	7.3	144	1.7%
44	9:00-18:00 in weekends [<=4.0]	23.7	10.0	307	3.5%
45	9:00-18:00 in weekends (>4.0]	35.7	12.5	240	2.8%

Node No. 2 1st level decision rule : Freq. -9:00 in weekdays (0, 1.9] (/month) Ave. Freq. 7.1 (/month), Variance 8.1 Number 925 (1

Node Number	2nd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
13	9:00-18:00 in weekdays [0]	3.0	3.4	319	3.7%
14	9:00-18:00 in weekdays (0,1]	2.7	1.5	125	1.4%
15	9:00-18:00 in weekdays (1,2.8]	5.2	2.7	153	1.8%
16	9:00-18:00 in weekdays (2.8,6.6]	8.3	4.2	187	2.1%
17	9:00-18:00 in weekdays (>6.6]	20.9	10.9	141	1.6%

Node Number	3rd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
46	18:00- in weekdays [0]	1.4	1.0	159	1.8%
47	18:00- in weekdays (>0]	4.5	4.2	160	1.8%

Node No.3 1st level decision rule : Freq. -9:00 in weekdays (1.9, 6.6] (/month), Ave. Freq. 12.7 (/month), Variance 10.6 Number 948

Node Number	2nd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
18	9:00-18:00 in weekdays [<=1]	6.7	4.5	379	4.3%
19	9:00-18:00 in weekdays (1,2.8]	8.3	3.9	138	1.6%
20	9:00-18:00 in weekdays (2.8,4.2]	11.0	5.5	131	1.5%
21	9:00-18:00 in weekdays (4.2,12.3]	18.0	8.4	197	2.3%
22	9:00-18:00 in weekdays (>12.3]	1.2	13.6	103	1.2%

Node Number	3rd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
48	18:00- in weekdays [<=2.6]	4.7	2.1	259	3.0%
49	18:00- in weekdays (>2.6]	11.2	5.1	120	1.4%

Node No. 4. 1st level decision rule : Freq. -9:00 in weekdays (>6.6] (/month) Ave. Freq. 31.5 (/month), Variance 12.9 Number 880 (1

Node Number	2nd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
23	9:00-18:00 in weekends [<=1.5]	28.1	11.6	569	6.5%
24	9:00-18:00 in weekends (1.5,4.0]	31.9	9.5	152	1.7%
25	9:00-18:00 in weekends (>4.0]	42.9	13.7	159	1.8%

Node Number	3rd level decision rule (monthly freq. in the following segment)	Average Freq. (/month)	Variance	Number of holders	%
50	18:00- in weekdays [0]	24.0	12.3	127	1.5%
51	18:00- in weekdays (0,6.6]	21.7	8.7	149	1.7%
52	18:00- in weekdays (>6.6]	33.2	10.1	293	3.4%

Figure 10(a) shows seven nodes of peek time frequent users such as commuting and student users whose monthly boarding frequency is higher than the total average monthly frequency (10.4 per a month) and whose boarding proportion on weekdays before 9:00 is more than 30%. Various and different characteristics are observed in these nodes of peek time frequent users. It is considered that node 20 and 50 mainly consists of students, who arrive their schools before 9:00 and usually go home before 18:00. It is also considered that node 49 and 52 mainly consists of commuting users, who start their work before 9:00 and usually do not go home before 18:00. These peak time users are also classified into frequent weekend user group and another. Users in nodes 20, 49, 24 and 25 frequently use public transport both on weekdays and weekend, while others seldom use on weekend though they frequently use on weekdays.

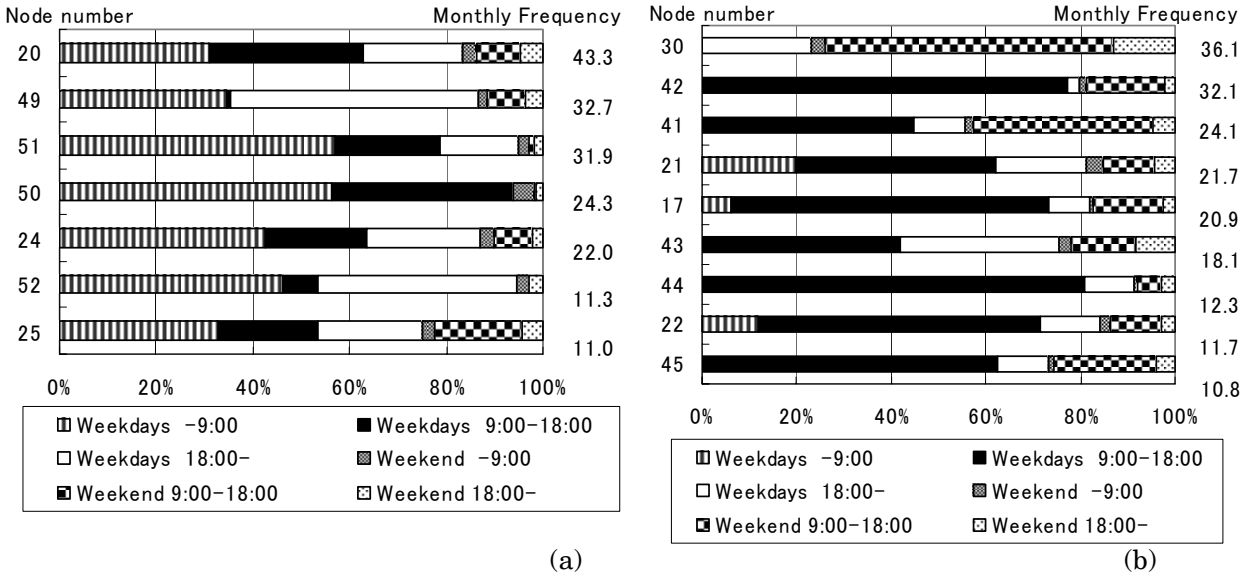


Figure 10. Average boarding proportion in a period in day :(a) peak time user group (b) off peak time user group

Figure 10(b) shows nine nodes of off-peak time frequent users whose monthly boarding frequency is higher than the total average value (10.4 per a month) and whose boarding proportion on weekdays before 9:00 is less than 20%. Various and different characteristics are also observed in off-peek time frequent users. Heavy frequent off-peak users (boarding more than 20 times per month; node 30, 42, 41, 21 and 17) amount to 10.9% of the total card holders (Table 5). In the boarding number base, these users account for around 30% of the total. Node 30 is unique as they do not ride on weekday before 18:00 by using cards. They are considered to use other transport modes or use transits by seasonal tickets for commuting on weekdays. Users in nodes 42, 17 and 44, *the boarding proportion on weekday from 9:00 to 18:00* is more than 70%, usually close their weekday activity by transit within a daytime. Node 41 is characterized by nearly same frequency of their boarding both on weekdays and weekends. Off-peak frequent users are patronages for public transport operators. Their demand elasticity is higher than that of peak time users such as commuters and students, and they are likely to use more frequently when level of services are increased.

6. Summaries

This paper focused on boarding records of the fare collection system that is introduced in public transport in the Hiroshima metropolitan area, and empirically clarified the applicability of all operators' integrated data records to urban city planning and operators' marketing. Passengers' on-board and transfer time distribution and their transfer behaviors are examined, and data missing concerning getting on and off place and time in day, and traceablity of boarding history by matching card issue number are checked and the accuracy of the data is shown by using 23,589 records. Finally, the authors classify the passengers (card holders) into groups in which passengers' boarding characteristics are similar in terms of monthly boarding frequency, boarding time period in day, average on-board time and average payment per each ride, and various and different user groups are extracted.

Suppliers of urban mass transit have regarded passengers as 'mass' and 'homogeneous'. While, retail industries, for example, have recognized that no 'average consumer' exist and that consumers are the set of countless numbers of different types of

persons, and they have succeeded by constructing relationship to each customer (Customer Relationship Management (CRM)). Transit operators now have their customers' data. Finding passengers' rule from huge size of collected boarding histories by data mining method has large potential to propose new services and fare policies targeted to a small but surely responsive passengers group in public transport. In this classification analysis, variables concerning place and line are not included and variables except concerning boarding frequency are not statistically significant. Further issues for proposing specific policies of public transport marketing are to find classification rules including above mentioned items and to applying this method to specific lines or time series data.

References

G. V. Kass, 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics, 2, 119-127

SPSS Inc: AnswerTree 3.0 User's Guide, 2001.