Identifying Potentially Problematic Bus Routes by User Behavior Transition Tendency

by Pai-Hsien Hung [a], Kenji Doi [b], Hiroto Inoi [c]

[a,b,c] *Graduate School of Engineering, Osaka University, Osaka, 565-0871, Japan*
[a] *E-mail: pshung@gmail.com*
[b] *E-mail: doi@civil.eng.osaka-u.ac.jp*
[c] *E-mail: inoi@civil.eng.osaka-u.ac.jp*

When a research focuses on the satisfaction or loyalty of the bus service, it often focuses on influence of specific variables. However, there are a wide variety of variables may influence users' decision-making. In order to prevent that the managers may ignore some unknown variables, we propose a bottom-up procedure to obtain the retention probabilities of each bus routes. Logistic regression models were calibrated which based on four behavior groups, and the significant coefficients of route variables represent the odds ratios of each route. The calibration of retention probability does not need any personal or socio-economical information but smart card transaction data. It will dramatically decrease the cost and time of data collection. We also find that there is a logarithmic relationship between number of users and retention probability. The relationship will enhance the managers estimate the retention probability effectively.

Keywords: *bus user retention, user behavior transition, smart card*

INTRODUCTION

Bus service is a common public transportation system in most of the area worldwide, and managers can use various combination of resources to implement their new service according to individual characteristics of each area. Since bus service can be designed flexibly but resource is usually limited, its goal should be thoroughly considered from the view point of finding appropriate solutions. Generally speaking, maximum income and ridership is the most common objective. However, user retention is also an important index for bus service promotion, with not so much related literature present yet. Because of these, most cities use sustainable concept to improve transportation environment, and therefore user retention becomes more and more important. Agencies started considering related issues, hoping to enhance retention and attract user, to change infrequent user into frequent user in other words.

The other similar topics like user retention are customer loyalty and satisfaction. Customer satisfaction in public transportation has been studied since mid-1960, but the loyalty in public transport is not well defined. According to Zhao et al. (2014), they suggest that loyalty has two aspects: first is a person's continuous behavior to buy a product, and the second one has to do with customer's attitudes and emotions. Although loyalty is similar to user retention, there is a critical difference between them. The loyalty is a quantification index that can measure the user's intention of purchasing the same product according to quality, satisfaction, and other causes. User retention is the result if user decide to use the same product or not. In short, loyalty is cause, and retention is result.

Moreover, most of the bus service contains numerous bus routes spread over various area. In case a manager wants to evaluate or make an improvement plan, there could be many quantification indexes to be used, e.g. income or ridership. However, such indexes cannot directly respond to the demand of user retention consideration. As a matter of fact, operation

parameters such as frequency or travel time are useful to extract performance of each bus route. However, outcomes from those parameters are rather ambiguous since they may related to users' characteristics. This paper will propose a bottom-up methodology to find potentially problematic bus routes based on user behavior transition conditions.

In the traditional planning procedure, managers often use several indexes to identify problematic parts of their bus service. For example, if income or ridership of a bus route is getting lower, that route is easily identified as a problematic one. Or, if the users' feedback is relatively negative, that route is also easily identified as problematic. However, these methods for problem identification are just based upon known causes. In reality, users will evaluate the service according to various causes and their combination. Merely using known causes to evaluate a bus service is a trap quite easily fall into.

As we know, all ridership comes from users' free will, and they respectively decide whether to use bus service or not after due consideration. It may include various causes and the weight of each cause may also variable. Therefore, we can at first identify the route which users most likely to quit using, and analyze the causes according to their characteristics. Then, other small number of potentially problematic routes will be identified based on user behavior transition condition. Needless to say, if the targeted service is limited in scope than directly applying on the whole service, the cause analysis will become more focused and reliable.

OBJECTIVE

In the past, studies of user retention required conducting questionnaire survey or household travel survey to obtain long-term user retention information. It would cost lots of budgets and also needed complicated procedure. Fortunately, as the application of smart card systems grows quickly, bus operators at present can get raw data of each transaction such as time, location and route when users board a bus. Large data sets not only present operating performance via ridership calculation but also derives users' behavior information via advanced statistical methods (Morency *et al,* 2006; Morency *et al,* 2007; Bagchi and White, 2005, Zhong *et al,* 2015). Within this research, we will first extract user behavior transition based on smart card information. Then we will cluster the transition results into several groups and therefore simplify the prediction procedure of user retention. Expectation Maximization (EM) algorithm will be used to cluster behavior, and then we derive long-term behavior transition result from cluster transition tendency of each month of a year. According to the monthly cluster results of the whole smart card users, we can derive their behavior transition between consecutive months from which we can know the decisions of users. Then, we calibrate a logistic regression model that is based on transition, routes, and other related information like ridership and number of bus stops to define users' decision making as users' preference. In that model, behavior transition is a binary dependent variable. Various independent variables will also be calibrated. Significant variables in the model will show users' tendency of specific routes. With the help of these, managers can not only easily understand users' decision tendency of some route, but also can find potential problems in their bus service.

By calibrating choice models as above, we can grasp users' decision among most of the bus routes according to various user characteristics. The coefficients of each route variables show the preference and how serious those problems are. Managers can identify where the problematic bus routes are, and grasp how serious they are. In conformity with the limited number of bus routes, managers can conduct more detailed and efficient comparison, and make the better improvement plan. To compare with traditional problem identification methods, this research simplifies the complexity of performance evaluation, and sufficiently considers the user retention at the same time. During the model calibration, various user clusters require being

grouped into a variety of groups according to the calibration in order to obtain better results.

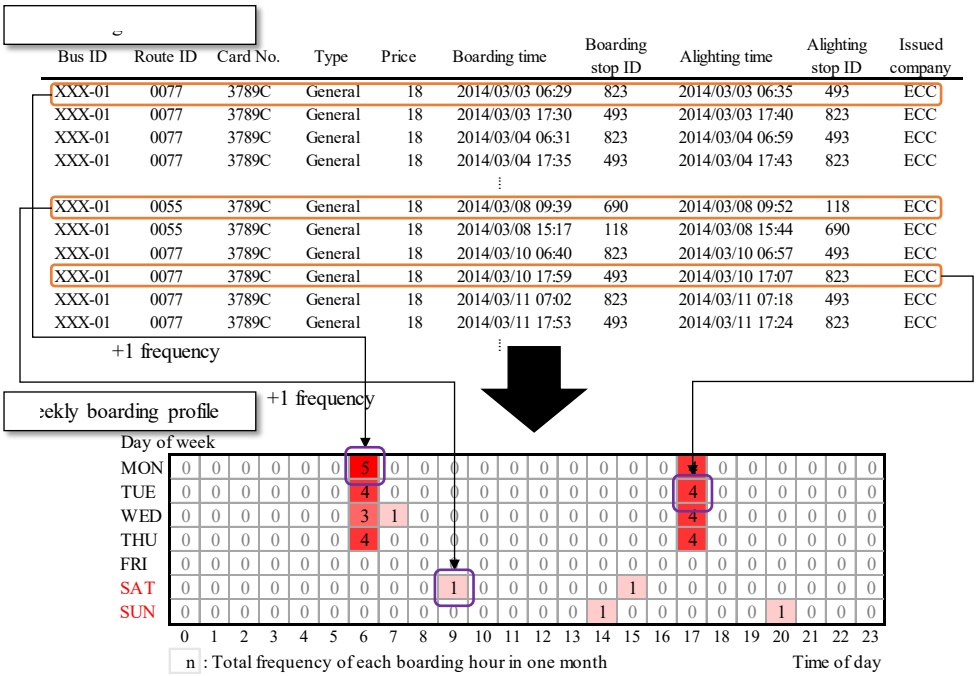
This time, a case study in Tainan, Taiwan is conducted and we propose an example to show how the bottom-up procedure works. From the calibration results of the model, we have expected to get some significant coefficients based on user behavior transition. Within this case study, we can obtain user behavior transition tendency, and long-term behavior clusters of certain bus company and bus routes; then by using that data, we can determine potentially problematic bus routes. Also, according to the routes in question, we can compare operation parameters with route characteristics of the selected bus company and routes.

CALCULATION METHODOLOGY OF BEHAVIOR TRANSITION

Behavior transition could be obtained from weekly profiles of bus users (Hung et al. In published). Weekly profiling is the key concept to conduct the clustering, and the average frequency is set to be one month. Smart card data is used in this study, and each raw data contains the boarding and alighting information of a single trip by the same card. We use each user's data in one month to conduct the clustering calculation. One single month at least includes 4 weeks of weekdays and 4 weekends. By summing each hour of one month's usage, a weekly profile can be determined. In order to prevent the rare users like one time visitors from influencing the main body of the clustering, those who use less than 4 times per month are grouped as rarely-use user cluster.

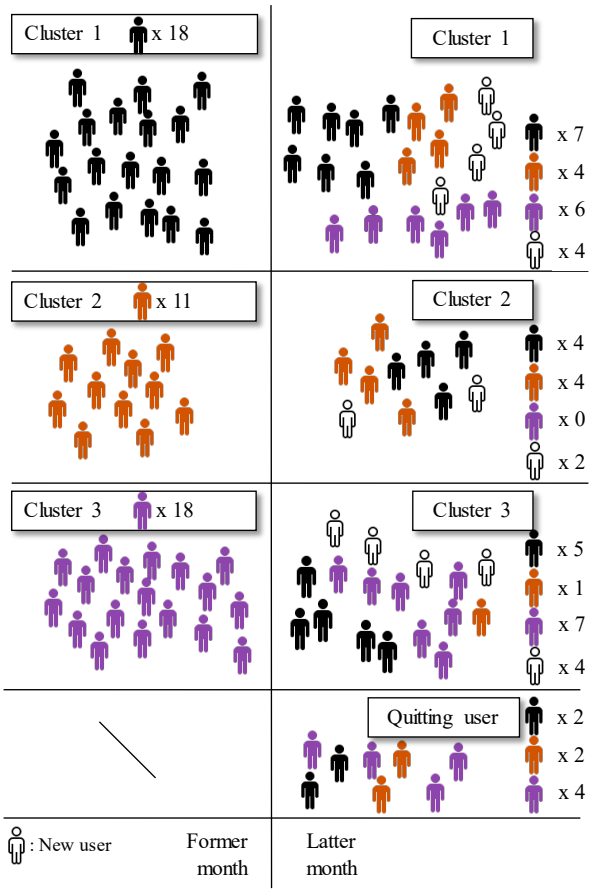
Frequency of most bus users' behavior are weekly, including weekday and weekend trip shown in weekly profile. Therefore, we consider there exists 168 (24 hours * 7 days) variables in a week and averaged the frequency in each hour (Pas, 1988; Tarigan *et al,* 2012; El Mahrsi *et al,* 2014). Figure 1 shows an example of how smart card usage raw data transfer into weekly boarding profile. For the same card number (same user), by looking at the boarding time section of smart card usage data, we accumulate each boarding records separately according to its boarding hour so that we can understand the specific frequency per hour. Peak hour characteristics as well as difference between weekday and weekend is now easy to define.

Figure 1: Example on how IC Card Usage Data Transferred into Weekly Boarding Profile



After regularity sorting, behavior transition in adjacent months is conducted by identifying the behavior clusters in the former month and the latter month. There are two special clusters in each of the former and latter clusters. One is the "New" cluster, which means that the smart cards are only used in the latter month; the other is the "Quit" cluster, which means that the cards are only used in the former month. Figure 2 is a behavior transition example, showing how to identifying the behavior cluster from a user clustering results in adjacent months.

Figure 2: Behavior Transition Example in Adjacent Months



Number of users with behavior transition

Former to latter		Latter clusters (number of users)				
		Cluster 1	Cluster 2	Cluster 3	Quit	Total
Former clusters	Cluster 1	7	4	5	2	18
	Cluster 2	4	4	1	2	11
	Cluster 3	6	0	7	4	17
	New	4	2	4	-	10

*Quitting users are who do not use bus service in latter month.

In case of computing behavior transition, clusters should be ordered for better understanding of the cluster's key characteristics. In this study, regularity is used to evaluate users' tendency to use bus service. As a regular user, more regularity means more frequency, like students or commuters. Departure time is a significant index that indicates the regularity. Besides, commuters usually take bus at peak hours, e.g. 6:00-9:00 and 16:00-19:00. Therefore, the regularity value is split into morning peak and afternoon peak, having 12:00 in the middle. Cluster regularity is the sum of standard deviation of morning and afternoon peak for all trips in the same cluster. The random cluster whose users use less than 4 times per month is forced to be the most random cluster.

CALCULATION METHODOLOGY OF RETENTION PROBABILITY

After behavior cluster computing, retention probability is conducted. This study defines the retention as the user continually uses the bus service in the latter month. We can obtain that the user is staying or quitting from the data of the same card no. also exist in the latter month. The influence from other variables will be insignificant when only consider the staying. According

to van Lierop et al. (2018), the user loyalty is a result of longer-term and trusting between user and agency. Users may be influenced by any possible or unknown variables. We use a down-up way to find the retention first, and skip to find the reasons that may influence it. Smart card data is the only one needs to collect. In addition, retention duration are varied, and it depends on the planning purpose.

Logistic model regression often used to calibrate the model with binary dependent variables (Al-Doori, 2017; Chiu Chuen, 2014; Ismail, 2011; Sun, 2013; Tao, 2017). An OR (odds ratio) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur when given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. Therefore, "Staying" or "Quitting" bus service is the binary dependent variable in this model. The user's ridership of each route are the independent variables. The OR could present the retention probability when user adds one ridership in a route in this study. Here, we bypass all personal and socio-economical information but behavior transition data. Other variables, e.g. number of routes user used, or user's ridership at peak hour, are used to enhance the calibration; and they will not obstruct the interception of OR.

The calibration of OR show in Figure 3. Sample data is the behavior clustering result from previous section, and the staying in the latter month is considered as dependent variables. Before calibration, all users split into four behavior groups, includes REG, SREG, RAN, and RARE. The behavior could be varied according to sample size or calibration result, but there is only one yearly OR for each group. Generally speaking, sample size of RAN and RARE groups are too large; users in these two groups can split into 10 or more sub-groups to obtain better calibration result. The monthly coefficients can be obtained from average of the calibration results of all sub-groups. Furthermore, the yearly coefficients of each group can be obtained from average and outliers removing of the monthly result. We can get the OR by exponentiating the yearly coefficients.

CASE STUDY

Tainan city locates at southern part of Taiwan, shown in Figure 4. From December 2010, Tainan County and Tainan City have been merged into Tainan special municipality. Its population amounts to 1,886,033 (end of 2016), its area is 2191.7 km^2, and its bus service contains 110 routes and 3 companies. The downtown area is located in the south west. From 2012 December, an smart card system has been applied to all bus services. In June of 2013, the DOT of Tainan started to operate in six main lines after re-planning the original bus services. Although the total ridership is continuously growing, the city bus mode share is still halted at 1~2%. Northern part is minor downtown. The selected bus operator provides service at northern rural and downtown area. The smart card data of the selected bus operator in 2016 (data in December is missing) is used in this study, and the amount of the operator's data is 1,459,692 rows of data.

Figure 3: Calculation Procedure of Odds Ratios

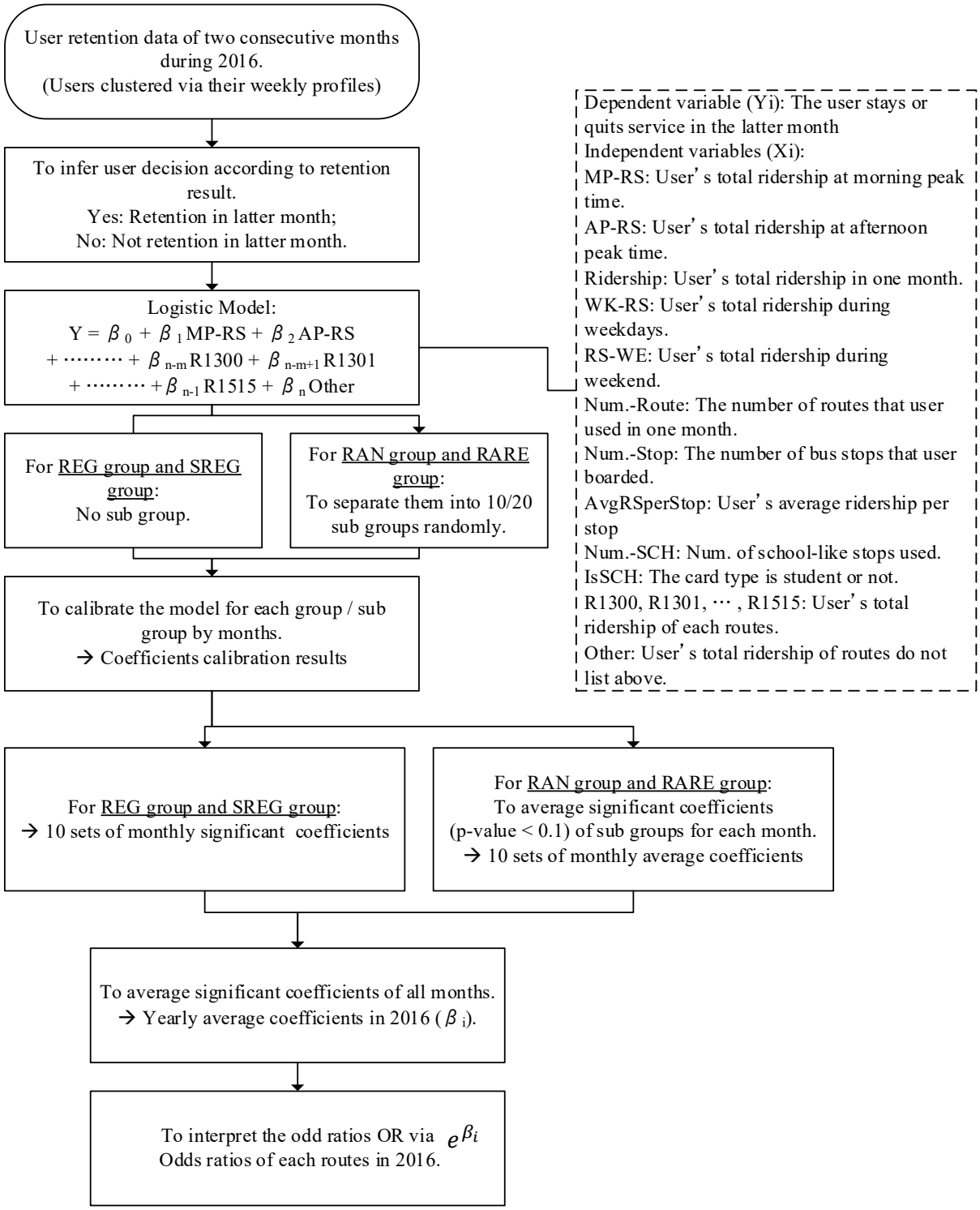
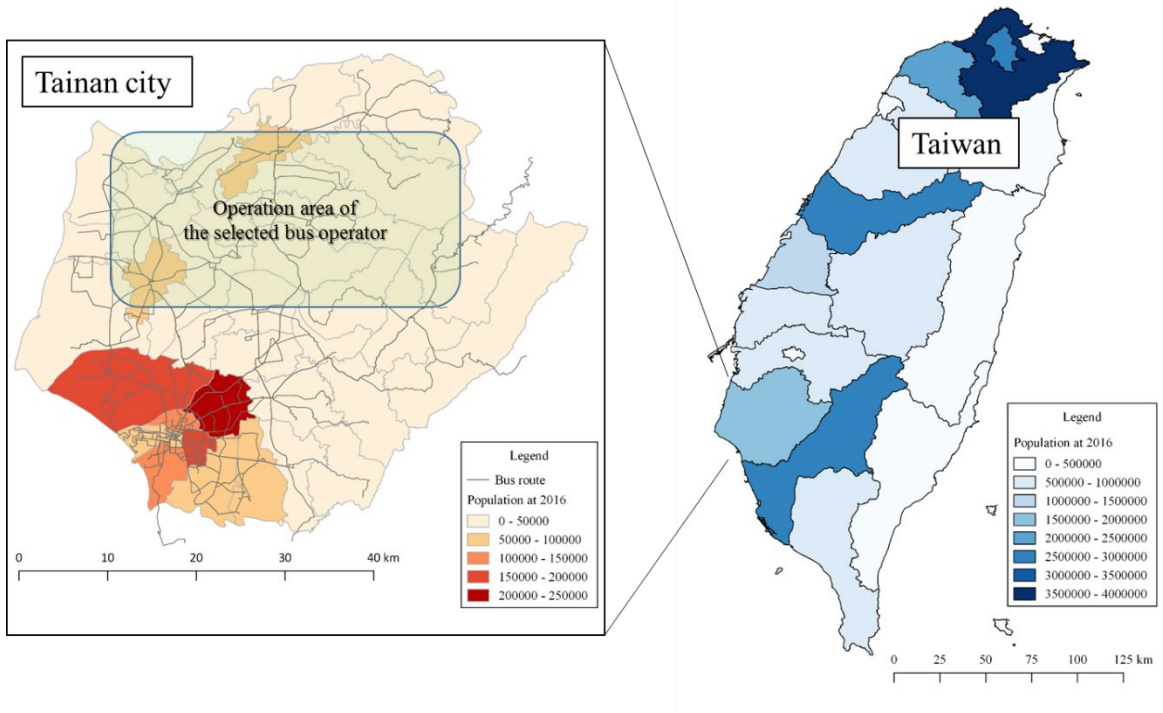


Figure 4: Population and Bus Network of Tainan City



User Behavior Clustering Result

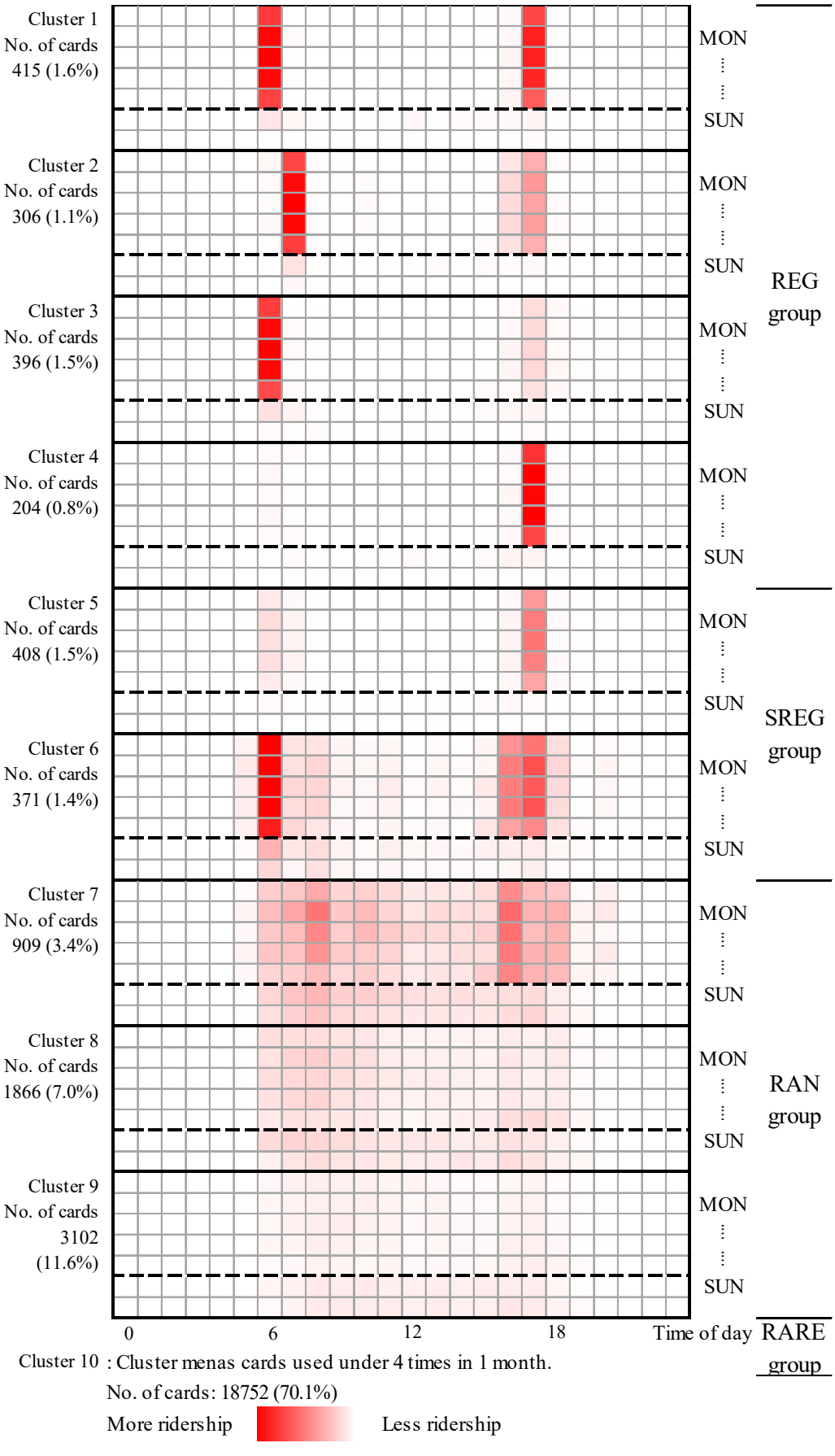
For the selected bus operator, there were 26,729 smart cards used in 2016 March, and 18,752 (70.2%) of them are random users who use less than four times in that month. The other 7,977 smart cards are users who use more than 3 times and therefore are being clustered via the EM algorithm. After clustering and regularity sorting, there comes nine usage patterns and 1 rare cluster listed in Figure 5. Cluster 1 is the most regular users', who use the bus around 6:00 and 17:00 and should be standard commuters; the average frequency shows that they use buses almost every weekday. Cluster 2 is the regular users who use the bus around 7:00, and there exists only half frequency at the afternoon peak. Cluster 3 is similar to cluster 2, but the morning peak start around 6. Cluster 4 is the users who use the bus at the afternoon peak only. Cluster 5 is similar to cluster 4, but there are about average 1.5 times in morning peak. Cluster 6 is the opposite and less regularity at afternoon peak. Cluster 7 is the users who use the bus not only at the morning and afternoon peaks but also at some adjacent several hours of both peaks, like college students or employees. Cluster 8, 9 are random users who use the bus at random departure times. The difference between these two is that cluster 8 has a higher frequency at weekends.

The results show regularity sorting clearly and closer to the real world. Because the number of users in high regularity clusters are much lower than low regularity users, and the regression model with small sample size is not easy to calibrate; we group similar behavior clusters into four groups according to regularity and frequency in the peak time. The four groups are REG group (regular), SREG group (sub-regular), RAN group (random), and RARE group (rare). Their corresponding behavior clusters and number of users show in Table 1.

Table 1: Regularity and Number of Users in Various Behavior Groups

Behavior cluster	Regularity (hour)		Number of users in cluster	Behavior group	Number of users in group
	Morning (before 12:00)	Afternoon (After 12:00)			
1	0.80	1.12	415	REG	1,321
2	0.77	1.29	306		
3	0.81	1.55	396		
4	1.59	0.92	204		
5	1.32	1.27	408	SREG	679
6	1.42	1.46	371		
7	1.63	2.02	909	RAN	5,877
8	1.63	2.12	1,866		
9	1.68	2.11	3,102		
10	1.68	2.12	18,752	RARE	18,752

Figure 5: Behavior Patterns of All Clusters in the Selected Bus Operator (March 2016)



Odds Ratios Result

There are total four sets of yearly average coefficients, includes REG, SREG, RAN, and RARE groups. REG and SREG groups have 10 sets (months) of calibration results; RAN has 100 sets (10 sub-groups * 10 months) of calibration results; and RARE group has 200 sets (20 sub-groups * 10 months) of calibration results. All McFadden R^2 values are between 0.06 and 0.25, and all VIF (Variance Inflation Factor) values of independent variables are less than 10. After insignificant (p-value > 0.1) and outliers removing, the yearly coefficients and ORs show in Table 2. In the non-route independent variables, Num.-SCH and IsSCH are both insignificant. It shows that the studentship will not significantly influence the retention decision. Other non-route independent variables in RAN group are all significant. Retention probability of AP-RS and WK-RS are both smaller than 1. The retention probability of users who add one ridership at afternoon peak or weekday will decrease 3% and 15%. Calibration results of REG and SERG groups are similar expect the OR of MP-RS in SREG group is smaller than 1 (0.9993).

Table 2: Coefficients and Odds Ratios Results of each Group

	REG group		SREG group		RAN group		RARE group	
Variables	coef.	OR	coef.	OR	coef.	OR	coef.	OR
(Intercept)	1.5316	4.6256	0.7757	2.1721	-0.0222	0.9780	-1.2928	0.2745
Num.-Route	-0.3704	0.6904	-0.4218	0.6559	-0.3188	0.7270	-	-
Ridership	-	-	-	-	0.2599	1.2968	-	-
Num.-Stop	0.5658	1.7609	0.3803	1.4628	-0.0632	0.9388	-	-
AvgRSperStop	0.2445	1.2769	0.2090	1.2325	0.4903	1.6328	-	-
MP-RS	0.0864	1.0903	-0.0007	0.9993	0.0055	1.0055	-	-
AP-RS	0.0841	1.0878	0.1104	1.1167	-0.0260	0.9744	-	-
WE-RS	-	-	-	-	0.1492	1.1609	-	-
WK-RS	-	-	-	-	-0.1614	0.8510	-	-
Num.-SCH	-	-	-	-	-	-	-	-
IsSCH	-	-	-	-	-	-	-	-
R1300	-	-	0.0672	1.0695	0.0955	1.1002	0.5261	1.6923
R1301	-	-	-	-	-0.4410	0.6434	1.3182	3.7367
R1302	-	-	-	-	-0.2652	0.7670	1.0435	2.8391
R1303	-	-	-	-	-0.2330	0.7921	1.2324	3.4295
R1310	-	-	-0.1175	0.8891	-0.3739	0.6881	1.4030	4.0675
R1311	-0.0622	0.9397	-	-	-0.4169	0.6591	-	-
R1500	0.0297	1.0301	0.0731	1.0759	-0.1127	0.8935	0.4809	1.6175
R1501	-	-	-	-	-0.2290	0.7953	0.7616	2.1418
R1502	-0.0963	0.9082	-	-	-0.1531	0.8580	1.0829	2.9531
R1503	-0.0516	0.9497	-	-	-0.2436	0.7838	1.6728	5.3269
R1504	-0.0457	0.9553	-	-	-0.1765	0.8382	1.1247	3.0794
R1505	-0.2025	0.8167	-	-	-0.4159	0.6598	1.5237	4.5891
R1506	-0.0420	0.9589	-	-	-0.3025	0.7390	0.9053	2.4727
R1507	-0.0975	0.9071	-	-	-0.1197	0.8872	0.6397	1.8959
R1509	-	-	0.1794	1.1965	-0.1175	0.8891	0.4249	1.5295
R1510	-	-	-	-	-0.2161	0.8056	1.1612	3.1939
R1511	-	-	-0.0538	0.9476	-0.2081	0.8121	1.1171	3.0560
R1512	-0.0985	0.9062	0.1404	1.1507	-0.2280	0.7961	0.4251	1.5298
R1513	-0.2714	0.7623	-	-	-0.2163	0.8055	0.7431	2.1024
R1514	-0.0904	0.9135	-0.1564	0.8553	-0.2051	0.8145	1.7514	5.7629
R1515	-0.1172	0.8894	-	-	-0.9370	0.3918	-	-
Other	-0.0560	0.9456	-	-	0.1669	1.1817	0.4889	1.6305

"-": p-value > 0.1; coef.: coefficient; OR: Odds ratio

About the route independent variables, all coefficients are significant in RAN group. Most ORs are smaller than 1 expect OR of R1300, which is 1.1002. The retention probability of users in RAN group add one ridership in routes (expect R1300) are decreasing. Expect R1311 and R1515 in RARE group, other variables are significant. All ORs of them are larger than 1, but the OR is decreasing when number of users is increasing. There are 13 significant route independent coefficients in REG group, and others are not. The retention probability of users in REG group add one ridership in most routes (expect R1500) are decreasing. There are only seven significant route independent coefficients in SREG group.

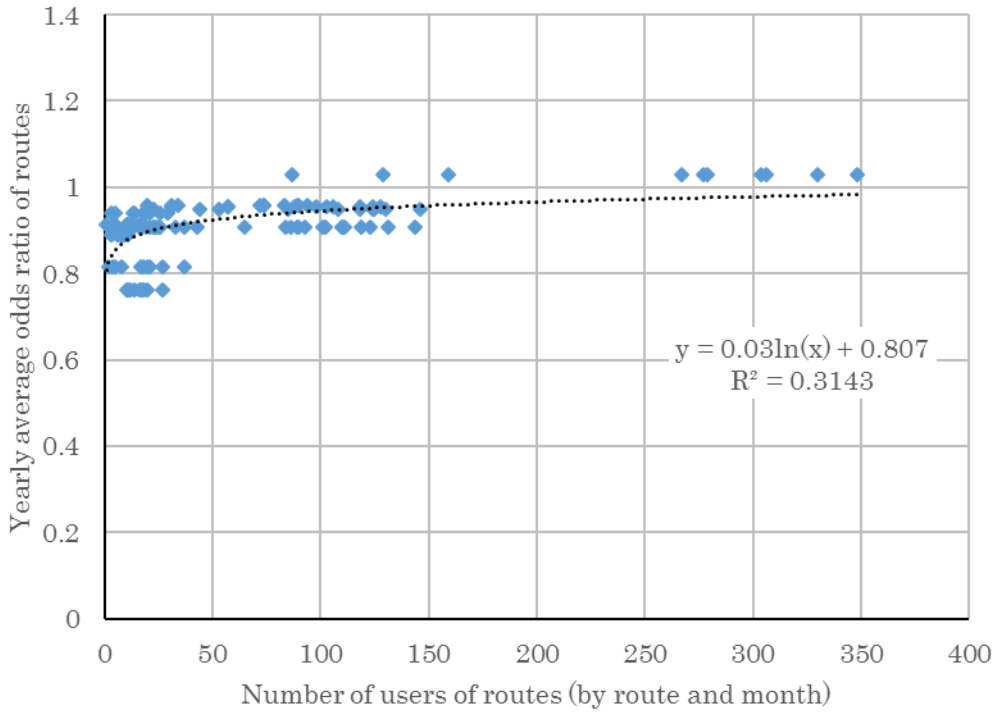
Correlation between Yearly Odds Ratio and Number of Users

Because the retention probability of each behavior cluster in each route are varied and some route variables cannot get significant coefficients, this study tries to find the relationship among the number of users and retention probability. Once we obtain the relationship, other route variables without significant coefficient or hypothetical number of users also can estimate their retention probability. The relationship between retention probability and number of users of four behavior groups (REG, SREG, RAN, and RARE) show in Figure 6 to Figure 9. Although some routes have no significant coefficients, the figures represent that there are logarithmic relationship between them. The R^2 values are between 0.3143 and 0.6155, and they show a good fit. The logarithmic relationship can derive the probability from the number of users.

The relationship of RARE group is different from other three; the probability is a decreasing curve. It means that the retention probability is decreasing when number of users is increasing. Although the relationship of REG group is an increasing curve, the OR will approach to 1. No matter how many number of users increase, the retention probability is hard to be positive. When the number of users in SREG group increases to 150 persons or more, the retention probability will increase to 10% or more (OR = 1.1). Relationship of RAN is similar to REG group, but the initial value is much lower, about 0.4. From the previous result, the relationship fits a specific curve according to the behavior, and the shape of the curve could be increasing or decreasing and initial values are varied, too.

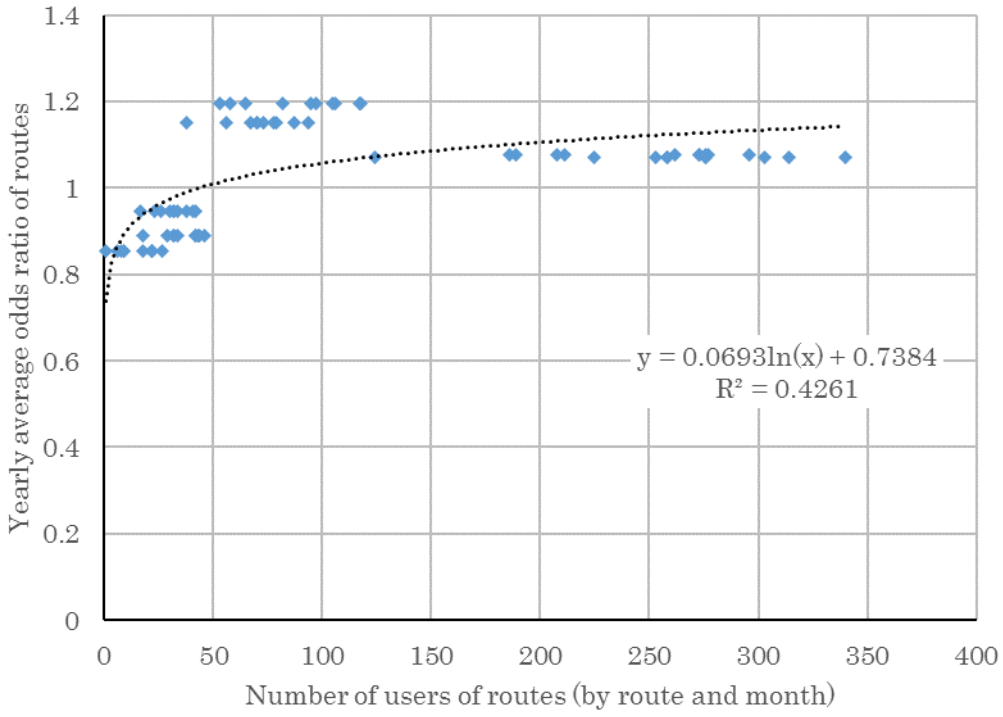
Managers can simulate the retention probability based on only number of users. Because of the personal and social-economical information are not required, we can dramatically decrease the cost and time of data collection. Moreover, we can make more calibration with available information, and obtain better result.

Figure 6: Correlation between Yearly Odds Ratio and Number of REG Users of Bus Routes



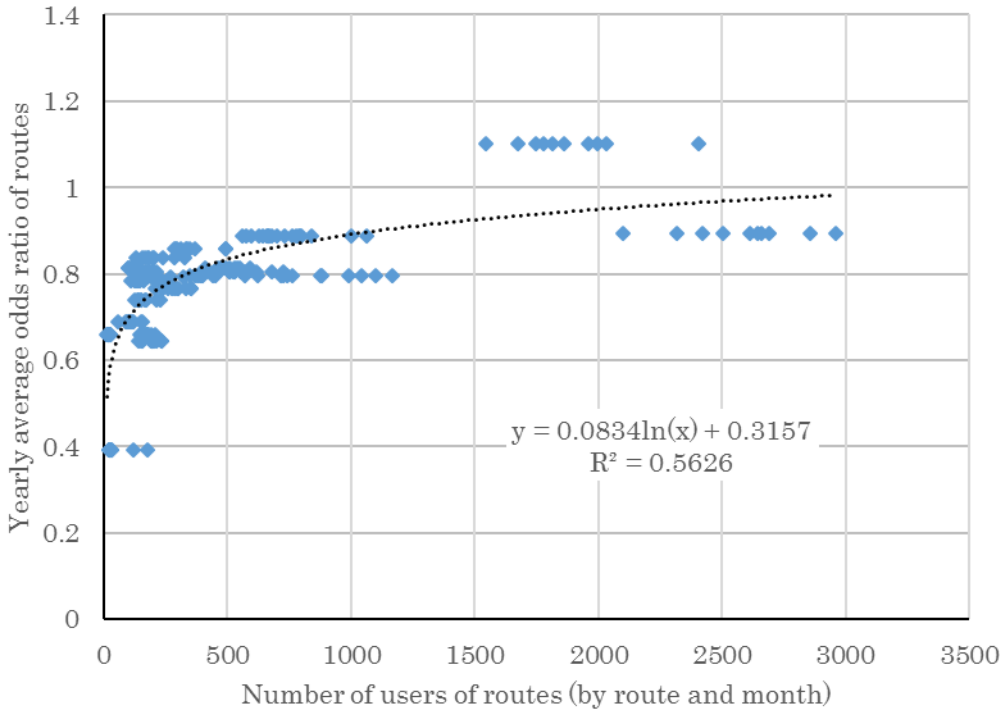
*Number of REG users in each month: 558 ~ 2,128

Figure 7: Correlation between Yearly Odds Ratio and Number of SREG Users of Bus Routes



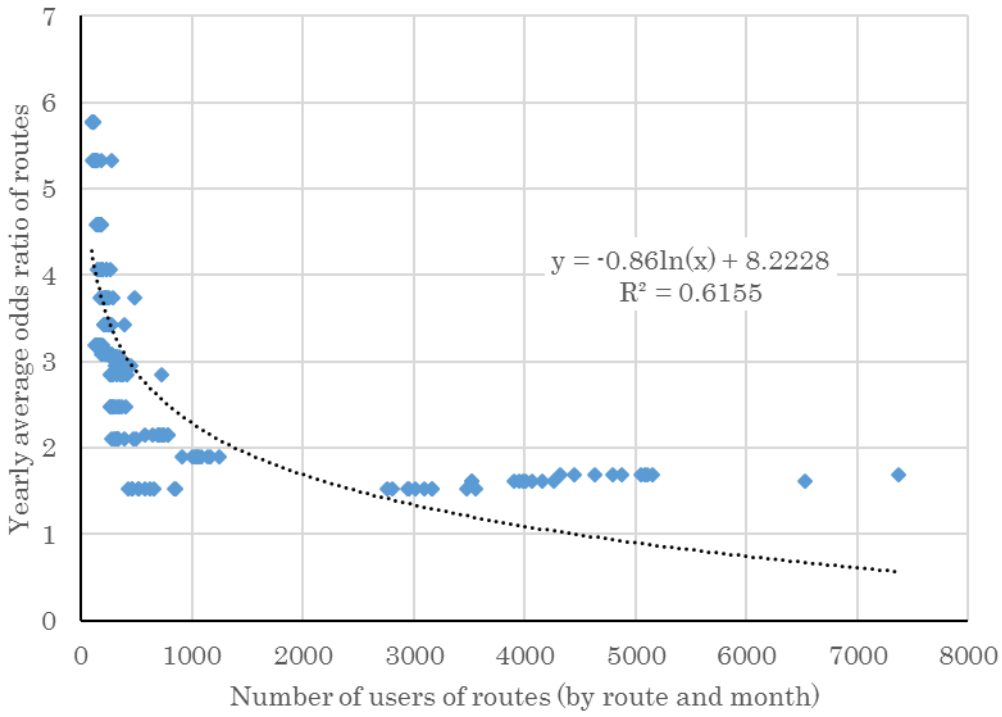
*Number of SREG users in each month: 872 ~ 1,681

Figure 8: Correlation between Yearly Odds Ratio and Number of RAN Users of Bus Routes



*Number of RAN users in each month: 8,827 ~ 12,535

Figure 9: Correlation between Yearly Odds Ratio and Number of REG Users of Bus Routes



*Number of RARE users in each month: 16,302 ~ 21,958

According to the regression formulations in Figure 6 ~ Figure 9, we give an example with 10% increasing of number of users. The number of users, increasing retention probability of original number of users, and increasing retention probability of new number of users (+10%) show in Table 3. In the traditional planning, managers preferred to allocate more resource on route with higher ridership or number of users to gain more ridership, e.g. R1300, or R1500. However, the retention probability difference represents that it might decrease the probability when the number of users increased. For example, the RARE group of R1300 and R1500 will decrease the probability because of the increasing number of users. REG group of R1505, R1512, R1513, R1514, and R1515 will decrease about 10% probability. It means that the service of these routes need to improve for REG group users.

Table 3: Retention Probability Difference after Adding 10% Number of Users

Route	Number of Users (n) in March 2016				Increasing Retention Probability of original number of users = OR(n) - 1				Increasing Retention Probability of new number of users (+10%) = OR(n * (1 + 10%)) - 1			
	REG	SREG	RAN	RARE	REG	SREG	RAN	RARE	REG	SREG	RAN	RARE
R1300	406	253	1676	4326	-0.01	0.12	-0.07	0.02	-0.01	0.13	-0.06	-0.06
R1301	86	35	140	217	-0.06	-0.02	-0.27	2.60	-0.06	-0.01	-0.26	2.51
R1302	30	22	260	276	-0.09	-0.05	-0.22	2.39	-0.09	-0.04	-0.21	2.31
R1303	103	40	135	213	-0.05	-0.01	-0.28	2.61	-0.05	0.00	-0.27	2.53
R1310	66	42	57	150	-0.07	0.00	-0.35	2.91	-0.06	0.00	-0.34	2.83
R1311	26	13	15	13	-0.10	-0.08	-0.46	5.02	-0.09	-0.08	-0.45	4.94
R1500	304	276	2659	6531	-0.02	0.13	-0.03	-0.33	-0.02	0.13	-0.02	-0.41
R1501	128	66	345	573	-0.05	0.03	-0.20	1.76	-0.04	0.04	-0.19	1.68
R1502	110	60	293	308	-0.05	0.02	-0.21	2.29	-0.05	0.03	-0.20	2.21
R1503	130	83	107	110	-0.05	0.04	-0.29	3.18	-0.04	0.05	-0.29	3.10
R1504	103	60	131	182	-0.05	0.02	-0.28	2.75	-0.05	0.03	-0.27	2.67
R1505	17	36	144	162	-0.11	-0.01	-0.27	2.85	-0.11	-0.01	-0.26	2.77
R1506	94	40	122	289	-0.06	-0.01	-0.28	2.35	-0.05	0.00	-0.28	2.27
R1507	93	90	671	908	-0.06	0.05	-0.14	1.37	-0.05	0.06	-0.13	1.28
R1509	108	117	561	2764	-0.05	0.07	-0.16	0.41	-0.05	0.08	-0.15	0.33
R1510	62	59	145	173	-0.07	0.02	-0.27	2.79	-0.07	0.03	-0.26	2.71

R1511	31	32	460	297	-0.09	-0.02	-0.17	2.33	-0.09	-0.01	-0.17	2.24
R1512	17	71	991	572	-0.11	0.03	-0.11	1.76	-0.11	0.04	-0.10	1.68
R1513	10	65	508	308	-0.12	0.03	-0.16	2.29	-0.12	0.03	-0.16	2.21
R1514	14	22	99	104	-0.11	-0.05	-0.30	3.23	-0.11	-0.04	-0.29	3.15
R1515	8	4	25	30	-0.13	-0.17	-0.42	4.30	-0.13	-0.16	-0.41	4.22

CONCLUSIONS

This study proposed a simple and efficient procedure to obtain retention probability from smart card data. By applying the EM method, bus users were clustered into several groups and their behavior transition between each adjacent month show the status of staying or quitting service. Logistic regression models were calibrated which based on four behavior groups, and the significant coefficients of route variables show the OR of each route. The calibration of retention probability does not need any personal or socio-economical information but smart card transaction data. It will dramatically decrease the cost and time of data collection.

The composition of various behavior cluster vary in different routes, and the retention probability can estimate the number of riders in latter month. Because of the probability is known, the managers may simulate the effect of various alternatives with hypothetical number of riders. Moreover, they can find the key resource allocation solution based on lower or negative retention probability. Future study should focus on the relationship among retention probability and other operational variables, e.g. frequency or stop location. More detail service planning can be evaluated according to such relationship.

The method developed in this study is different from previous ones that assumed all users in the same behavior cluster to be the same. Actually, ridership is a composition of diverse user types which changes with time. Resource allocation could be more appropriate with the understanding of who the target is and of the retention probability. This bottom-up procedure can be easily applied to other transportation modes. In addition, wider characteristics of bus service will enhance the model. Managers also can consider specific characteristics if they have sufficient data to calibrate the model. For the good of problematic bus routes and their passengers' characteristics, managers can make the more suitable improvement plans for the bus service under resource limitation in the end. Once the improvement being implemented, the service may get higher user retention in the near future. We are sure that managers will gain more user retention from the model within this research.

ACKNOWLEDGEMENTS

The authors would like to express special thanks to the Department of Transportation in Tainan, Taiwan for providing smart card data and GIS map data of Tainan.

REFERENCES

Al-Doori, Aws. "Waiting Time Factor In Public Transport By Binary Logistic Regression." Australian Journal of Basic and Applied Sciences 11.4 (2017): 72–76. Print.

Bagchi, M., and P. R. White. "The Potential of Public Transport Smart Card Data." Transport Policy 12.5 (2005): 464–474. Web.

Chiu Chuen, Onn, Mohamed Rehan Karim, and Sumiani Yusoff. "Mode Choice between Private and Public Transport in Klang Valley, Malaysia." The Scientific World Journal 2014.Figure 1 (2014): 7–9. Web.

El Mahrsi, Mohamed K et al. “Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A Case Study in Rennes, France.” The 3rd International Workshop on Urban Computing (UrbComp 2014). New York: N.p., 2014. Print.

Hung, P. H., K. Doi, and H. Inoi. “User behavior transition mapping for bus transportation planning based on time series data analysis of travel e-ticket information” Journal of the Eastern Asia Society for Transportation Studies (In published).

Ismail, Amiruddin, and Adel Ettaieb Elmloshi. “Logistic Regression Models to Forecast Travelling Behaviour in Tripoli City.” International Journal on Advanced Science, Engineering and Information Technology 1.6 (2011): 618–623. Web.

Morency, C, M Trepanier, and B Agard. “Analysing the Variability of Transit Users Behaviour with Smart Card Data.” 2006 IEEE Intelligent Transportation Systems Conference (2006): 44–49. Web.

Morency, Catherine, Martin Trépanier, and Bruno Agard. “Measuring Transit Use Variability with Smart-Card Data.” Transport Policy 14.3 (2007): 193–203. Web.

Pas, Eric I. “Weekly Travel-Activity Behavior.” Transportation 15.1–2 (1988): 89–109. Web.

Sun, Lijun et al. “Understanding Metropolitan Patterns of Daily Encounters.” Proceedings of the National Academy of Sciences 110.34 (2013): 13774–13779. Web.

Tao, Sui, Jonathan Corcoran, and Iderlina Mateo-Babiano. “Modelling Loyalty and Behavioural Change Intentions of Busway Passengers: A Case Study of Brisbane, Australia.” IATSS Research 41.3 (2017): 113–122. Web.

Tarigan, Ari K. M., Satoshi Fujii, and Ryuichi Kitamura. “Intrapersonal Variability in Leisure Activity-Travel Patterns: The Case of One-Worker and Two-Worker Households.” Transportation Letters: The International Journal of Transportation Research 4.1 (2012): 1–13. Web.

van Lierop, Dea, Madhav G. Badami, and Ahmed M. El-Geneidy. “What Influences Satisfaction and Loyalty in Public Transport? A Review of the Literature.” Transport Reviews 38.1 (2018): 52–72. Web.

Zhong, Chen et al. “Measuring Variability of Mobility Patterns from Multiday Smart-Card Data.” Journal of Computational Science 9 (2015): 125–130. Web.Zhao, J., Webb, V., & Shah, P. (2014). Customer loyalty differences between captive and choice transit riders. Transportation Research Record: *Journal of the Transportation Research Board* (2415), 80-88.