IATSS Research 44 (2020) 111–124

Contents lists available at ScienceDirect

IATSS Research

Research article

User retention tendency of bus routes based on user behavior transition in an area with low mode share of public transport

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

[a] Division of Global Architecture, Graduate School of Engineering, Osaka University, 2-1, Yamadaoka, Suita, Osaka, Japan
[b] Faculty of Sustainable Design, University of Toyama, 3190 Gofuku, Toyama, Toyama, Japan

ARTICLE INFO

Article history:
Received 16 November 2018
Received in revised form 22 June 2019
Accepted 10 September 2019
Available online 26 October 2019

Keywords:
Bus user retention
Odds ratio
User behavior transition
Behavior clustering

ABSTRACT

Although both public transport and private modes have their own purposes and safety issues, most people are free to choose either way to make a trip. Previous research states that increasing the mode share of public transport is an important transportation policy to improve traffic safety, and it is a key outcome of demand management indeed. However, rather than merely focusing on increasing its ridership, a more reliable way to reach for the universality of a public transport system is through its customer retention tendency. Research on satisfaction with bus service often focuses on the influence of specific variables. However, numerous variables may influence users' decision-making. To ease their work, managers have no choice but to ignore some unknown variables. Therefore, we now propose a bottom-up procedure, which needs only smart card data, to obtain the odds ratio of usage of a specific bus route. Logistic regression models are calibrated based on four behavior groups, and the significant coefficients of route variables represent the odds ratios of the bus route usage. The calibration of odds ratio does not need any individual personal or individual socio-economical information, but only smart card transaction data. This method will dramatically decrease the cost and time for data collection. Further, the procedure proposed in this study can be encapsulated in software, which managers can then use to assist their planning.

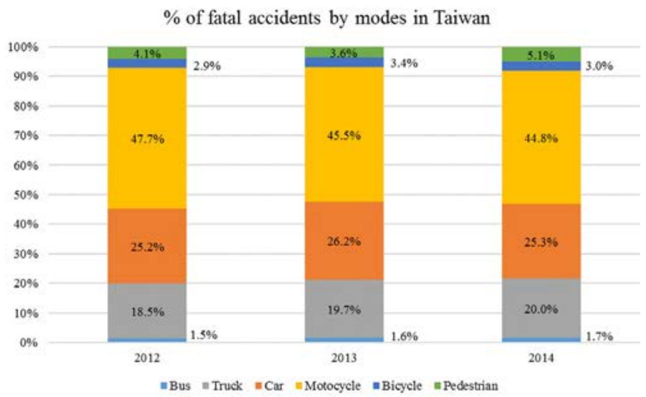
1. Introduction

In transportation planning, TDM (Transportation Demand Management) is one of three important components, Supply management, Land use management, and Demand management. It tries to explain how the users make a decision and assist users to make better decisions. Because there are various traffic problems around the world, including the mega traffic in the urban area and the inconvenience traffic environment in the rural area [1]. Moreover, sustainable transportation plays a critical role in the resource-limited environment recently; shifting users to the sustainable mode is the priority policy in the world. Those issues are the major outcome of the TDM. The current tendency is making policies for shifting users from private transportation to public one in order to reduce the traffic in the urban area and achieve the sustainable traffic environment. When users shift to public transport, the number of private modes will decrease consequently and this saves more energy and resource. In the outcome of TDM, traffic safety improvement is an indirect one and play a key role. According to the report from the WHO (World Health Organization) in 2018, there are about 1.35 million people killed by traffic accidents in the world. Most countries work hard to find the countermeasure for mitigating such a situation and try to save the economic loss made by traffic accidents. This value shows not only the severity of the traffic accident but also the influence of the national economy. Therefore, a policy can improve traffic safety efficiently will be a critical objective [2].

Traffic safety improvement – and how to achieve it – is deeply related to several major topics, e.g., driver behavior, vehicle safety, and law enforcement, and has been a critical research topic for a long time. Recent research, however, proposes that shifting users to public transport from private modes is an efficient way to improve traffic safety. Although the use of public transport cannot improve the safety of private modes directly, users traveling in public transport are much safer than those using private modes. Increasing the share of the public transport mode is, therefore, an important transportation policy for improving traffic safety. Fig. 1 shows the percentage of the number of fatal accidents by mode in Taiwan. On one hand, the percentage of fatal accidents of various private transport modes is over 70%, and the motorcycle is the highest one. On the other hand, the percentage of fatal accidents of buses is the smallest one, under 2%. According to that, decreasing the high private transport mode share – and a high percentage

[*] Corresponding author at: Division of Global Architecture, Graduate School of Engineering, Osaka University, S1-622, 2-1, Yamadaoka, Suita, Osaka, Japan.
E-mail addresses: hung.pai.hsien@civil.eng.osaka-u.ac.jp (P.-H. Hung), doi@civil.eng.osaka-u.ac.jp (K. Doi), inoi@sus.u-toyama.ac.jp (H. Inoi).
Peer review under responsibility of International Association of Traffic and Safety Sciences.

https://doi.org/10.1016/j.iatssr.2019.09.001



* National Police Agency, Taiwan, "Number of traffic fatalities and international comparison and analysis," 2015

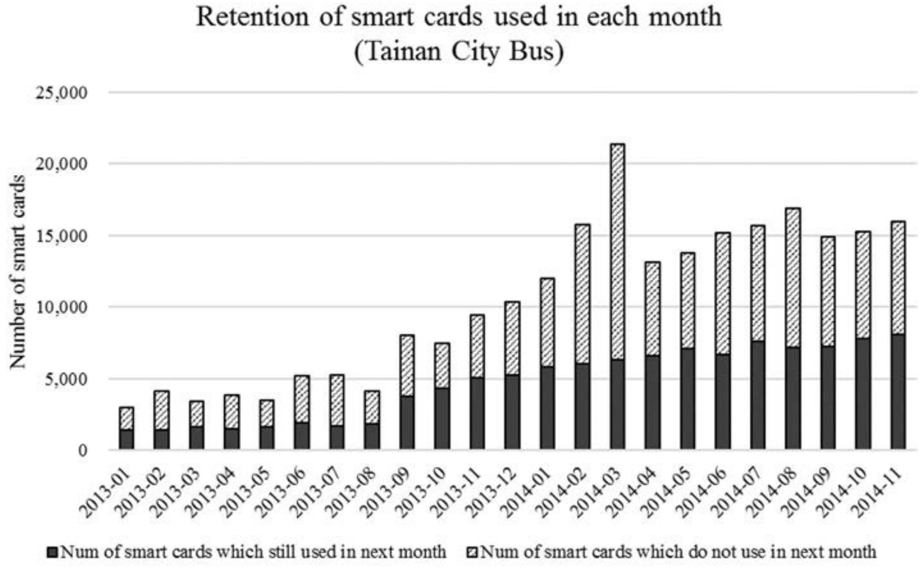
Fig. 1. The fatal traffic accident rate in Taiwan.

of fatal accidents that is a critical problem – will observably improve traffic safety in an area. Fig. 2 shows the retention of smart cards in Tainan City Bus service. In each month of 2013 and 2014, about 40–60% of smart cards will stop being used in Tainan City Bus Service in the next month. It shows that about half of users will stop using the bus and that users selected bus service as a temporary traffic mode, but did not stay in the service.

Compared to private modes, public transport provides higher capacity and safer environment; fatality rates by mode support this statement. From 2000 to 2014 in America, the average number of motorcycle deaths per billion passenger-miles was 237.57. In contrast, that of the bus was 0.2, which is obviously lower. In addition, public transport investment is the most cost-effective way to improve traffic safety for a community. According to research as above, increasing the number of users who shift to public transport and retaining them there will enhance traffic safety efficiently. [3]

Therefore, shifting users from public transport to private transport is an observably good solution for traffic safety improvement. For example, most university students in Taiwan prefer to use private transport modes, but this directly contributes to a high accident rate. The MOTC (Ministry of Transportation and Communication) in Taiwan announced a project, "Running Bus Service into University Campus in 2015," in order to decrease the high traffic accident rate of university students. That project planned several bus routes to run into the campus in order to attract students to use the bus service instead of motorcycle or other private modes. The project successfully decreased the number of traffic accidents of students by about 30%. Therefore, increasing the number of bus users is regarded as a tangible traffic safety plan.

Based on this concept of traffic safety improvement, managers prefer to use the increase in ridership or income to show the performance of service improvement. However, the ridership used for evaluating bus or other public transport usually does not consider the user composition. In reality, the usage patterns of all users are not the same; some are frequent users, and some are the fortuitous users. The general ridership simply sums up all the users' usage count without considering their various behavior characteristics, and therefore managers are only aware of overall ridership, but not the number of individual users and their characteristics. According to previous studies, users may tend to stop



* Smart card data of Tainan City Bus.

Fig. 2. Retention of smart card users of bus service in Tainan City Bus.

using service due to various reasons. Therefore, the number of users is, in fact, the number of original users plus the number of new users minus the number of users who stop using the service. The concept of user retention means that knowing the number of users who stop using the service is more important than knowing the number of users. The managers can understand their service more reliably based on the user retention; furthermore, the increase in the share of public transport users will improve the traffic safety efficiently. [4,5]

This time as described here, a case study in Tainan, Taiwan, is conducted so that we can propose an example to show how the logistic model can estimate the retention tendency. From the calibration results of the model, we have expected to get some significant coefficients based on user behavior transition, i.e., a transition meaning a change in user behavior with respect to use (or non-use) of a bus. Within this case study, we can obtain users' behavior transition tendency, and identify long-term behavior clusters with respect to a certain bus company and its bus routes. Then, by using that data, we can determine where the potentially problematic bus routes are. In addition, according to the routes in question, we can also compare operation parameters with route characteristics of the selected bus company and its routes. This case also shows that not only achieving the goal for mode shifting but also whether the users keep using the service are important. If the managers propose policies for mode shifting and keep tracing the retention result, this would result in traffic safety improvement and reduce the cost of mode shifting for repeated users.

2. Literature review

2.1. User shifting and retention

Public and private transport is a key grouping method for transport mode. In order to achieve a sustainable environment, using public transport is undeniably a needed tendency nowadays. Even so, managers must exert effort to try to overcome the high attraction of private transport, which is more convenient and has lots of incentives for personal movement. Especially in developing countries, the low quality of service of public transport is one of the major causes that result in the mode share of private transport is high. Some projects tried to use bus or bus rapid transit (BRT) services to attract users from private to public transport, and they indeed found there is potential for a modal shift with such kinds of services [6–8]. However, the more important issue is how many users will stay in the service [5].

The other topics similar to user retention are customer loyalty and satisfaction. Customer satisfaction in public transportation has been studied since the mid-1960s, but loyalty in public transport is not defined well. Loyalty has two aspects: the first one is each person's continuous behavior to buy a product, and the second one has to do with the customer's attitudes and emotions [9]. Although loyalty is similar to user retention, there is a critical difference between them. Loyalty is a quantification index that can measure the user's intention of purchasing the same product according to quality, satisfaction, and any other causes. User retention is the result of whether a user decides to use the same product or not. In short, loyalty is the cause, and retention results.

In order to increase the number of users who keep staying in the bus service, obtaining the odds of retention of the bus users is critical information that managers should understand. Although only one or just a few operators run multiple routes in the same area, users of each bus will have different feelings about the bus route that they have used. The bus route is often a basic service unit for planning or improvement. If we can find the relationship among the bus routes and their odds ratio, it can assist in making the planning easier and more efficient. In the past, studies of user retention required conducting a questionnaire survey or household travel survey to obtain long-term user retention information. Doing so costs lots in budget terms and needs complicated procedures. Fortunately, as the application of smart card systems is growing quickly, bus operators at present can easily 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 can derive users' behavior information via advanced statistical methods [10–13].

2.2. Odds ratio application for the association of user behavior and retention

When the managers design a bus service in an area, the bus route is usually the basic unit to consider, and all bus routes have their own planning targets and usage patterns. There are several properties of a bus service, including frequency, departure time, operation route, and location of stops … etc. The managers will design those properties based on the trip demand, departure time, and socio-economic characteristics of the users or potential users in the study area [14]. However, the bus service is a public transport system, and not only the planning targets will use but also the other users may use it in any scenario. All users will have different behavior patterns and feelings for each bus route. Besides, the “previous journey” is also a variable may influence the feeling of a user using the service. Besides the user's socio-economic characteristics, the experience of “previous journey” may enhance the user preference model with a variable like that [15,16]. Therefore, we pick the usage count of each route as a variable, which may have an association with the tendency of user retention.

The odds ratio can measure the association between an exposure and the outcome. In other words, the value indicates the change in the probability of an event occurs when a variable changed. In the past, it is a common methodology for medical research, and can measure the relationship between variables and outcome. Nowadays, there is some research use odds ratio to measure the relationship between the user's preference and choice. For example, one uses the odds ratio to measure the relationship between the body health index and the change of frequency of the user uses the bus [17]. Another research uses the odds ratio to measure the tendency of choosing bus service based on age and birth year [18]. In this research, we consider user retention as the dependent variable, and measure the association with behavior patterns.

3. Methodology

Within 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 certain bus route is getting lower, that route is easily identified as a problematic one. Also, 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 evaluate the service according to various causes and their combination. Merely using known causes to evaluate bus service is a trap that is quite easy to fall into. In addition, the ordinary planning procedures require collecting data concerning user characteristics from a costly sampling survey but, of course, managers do not have enough budget or time to do it frequently. Therefore, it is important to propose a procedure that can be encapsulated in a software program without considering users' socioeconomic characteristics.

3.1. Behavior clustering

Within this research, we at first extract user behavior transition, based on smart card information. Then, we cluster the transition results into several groups and, therefore, simplify the prediction procedure with respect to user retention. By an appropriate clustering method, a large number of users will split into several groups according to their behavior. The travel behavior-related models can get better calibration results based on various behavior groups [19]. An Expectation Maximization (EM) algorithm will be used to cluster behavior, and then we derive the long-term behavior transition result from the cluster transition tendency found for 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, such as ridership and number of bus stops, to define users' decision making as users' preference. Within this model, behavior transition is a binary dependent variable. Various independent variables will also be calibrated. Significant variables in the model will show users' tendency with respect to specific routes. With the help of these, managers can not only easily understand users' decision tendency of some routes, but also can find potential problems in their bus service.

As we know, behavior transition could be obtained from weekly profiles of bus users [20]. Weekly profiling is the key concept used to conduct the clustering, and the average usage frequency is set to be one month. Smart card data is used in this study, and each raw data item 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 includes at least 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 <4 times per month are grouped as a rare-use user cluster.

Usage frequency of most bus users' behavior is weekly, including both the weekday and weekend trips that appear in a weekly profile. Therefore, we considered the existence in a week of 168 variables (24 h * 7 days), and averaged the usage frequency in each hour [21–23]. Fig. 3 shows an example of how smart card usage data are transferred into a weekly boarding profile. For the same card number (same user), by looking at the boarding time section of smart card usage data, we accumulate boarding records for each card separately according to its boarding hour, so that we can understand the specific usage frequency per hour. Peak hour characteristics, as well as differences between weekday and weekend, are now easy to define.

After regularity sorting, behavior transition in adjacent months is considered 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. Fig. 4 is a behavior transition example, showing how to identify the behavior cluster from user clustering results in adjacent months.

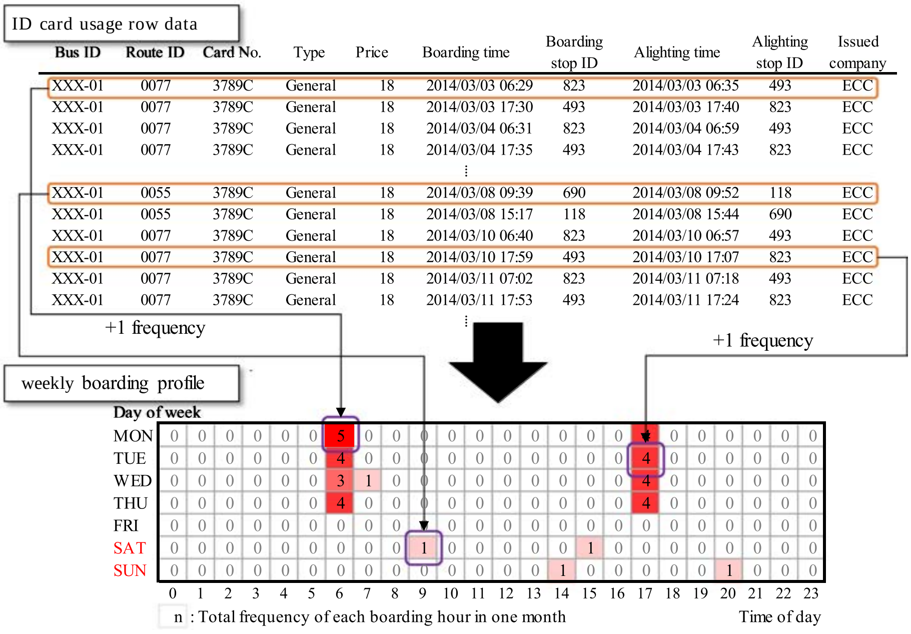
In computing behavior transition, clusters should be ordered for a better understanding of the cluster's key characteristics. In this study, regularity is used to evaluate users' tendency to use bus service. In a regular user's case, i.e., like students or commuters, more regularity means higher usage frequency. Departure time is a significant index that indicates regularity. Besides, commuters usually take a bus during 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 the standard deviation of morning and afternoon peak for all trips in the same cluster. The cluster regarded as the most random is the one whose users use a bus <4 times per month.

3.2. Binary logistic model building

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

Fig. 5 is the analysis framework of this research. In the past, there are thousands of research try to inference the association between user's preference and decision via model calibration. However, the time and financial cost are too large for practical application. Therefore, this research utilizes smart card data and big data concept to inference the relationship. Although we cannot know the reasons for the decision-making, it is possible to infer the association between user behavior and retention tendency. In Fig. 5, we calibrate the odds ratio of the measurement between attributes of the bus service users and decision of retention. The cost of this methodology will observably decrease via using the stable data source, smart card data.

By calibrating choice models as above, we can grasp users' decision with respect to most of the bus routes according to various user



ID card usage row data

Bus ID	Route ID	Card No.	Type	Price	Boarding time	Boarding stop ID	Alighting time	Alighting stop ID	Issued company
XXX-01	0077	3789C	General	18	2014/03/03 06:29	823	2014/03/03 06:35	493	ECC
XXX-01	0077	3789C	General	18	2014/03/03 17:30	493	2014/03/03 17:40	823	ECC
XXX-01	0077	3789C	General	18	2014/03/04 06:31	823	2014/03/04 06:59	493	ECC
XXX-01	0077	3789C	General	18	2014/03/04 17:35	493	2014/03/04 17:43	823	ECC
					⋮				
XXX-01	0055	3789C	General	18	2014/03/08 09:39	690	2014/03/08 09:52	118	ECC
XXX-01	0055	3789C	General	18	2014/03/08 15:17	118	2014/03/08 15:44	690	ECC
XXX-01	0077	3789C	General	18	2014/03/10 06:40	823	2014/03/10 06:57	493	ECC
XXX-01	0077	3789C	General	18	2014/03/10 17:59	493	2014/03/10 17:07	823	ECC
XXX-01	0077	3789C	General	18	2014/03/11 07:02	823	2014/03/11 07:18	493	ECC
XXX-01	0077	3789C	General	18	2014/03/11 17:53	493	2014/03/11 17:24	823	ECC
					⋮				

+1 frequency

+1 frequency

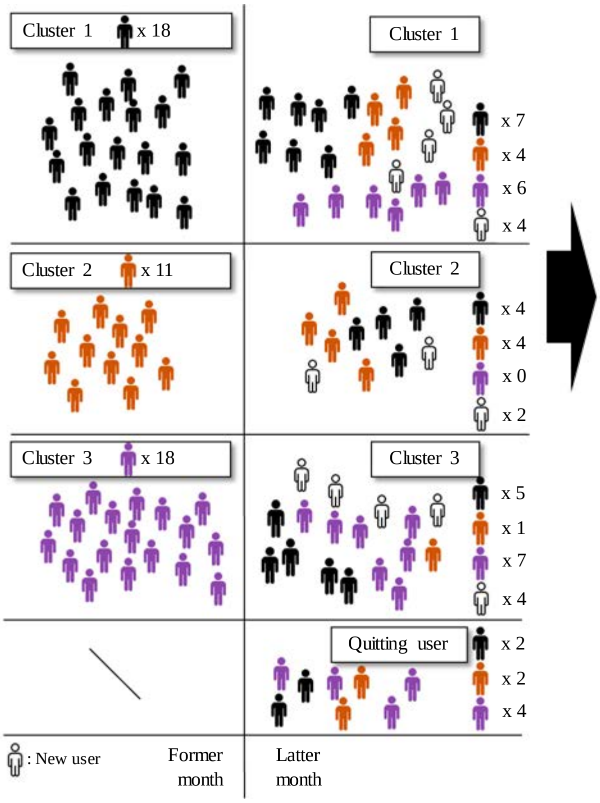
weekly boarding profile

Day of week	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
MON	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
TUE	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
WED	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
THU	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
FRI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SAT	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
SUN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0

Time of day

n : Total frequency of each boarding hour in one month

Fig. 3. Example of how smart card usage data transferred into the weekly boarding profile.



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.

Fig. 4. Behavior transition example in adjacent months.

characteristics. The procedure of building and calibrating regression model is shown in the Appendix. The coefficients of each route's variables show the preference and how serious those problems are. Managers can identify where the problematic bus routes are, and grasp how serious the problems are. In conformity with the limited number of bus routes, managers can conduct a more detailed and efficient comparison, and make a better improvement plan. Compared with traditional problem identification methods, this research simplifies the complexity of performance evaluation and, at the same time, sufficiently considers user retention. During the model calibration, various user clusters require being sorted into a range of groups in order to obtain better results.

After behavior cluster computing, retention probability is to be considered. Because the usage data of all smart cards of all months are known, the simplest way to get retention probability is via calculating the number of users using the service in former and latter months. However, it is difficult to express the influence on the usage count of the bus route. Therefore, we build a logistic model and assume that the usage count of the bus routes is independent variables and that a retention decision is a dependent variable. Fig. 6 shows the concept of inference of retention tendency. Once we obtain the significant coefficients of the model, it is possible to estimate the change in the odds ratio of bus route usage (ORBRU), i.e., the odds of customer retention, via usage count of each user, and the associated behavior characteristics of that user, e.g. total usage count, or usage count at weekend.

This study defines the retention as the user in a former month continually uses the bus service also in the latter month. We can obtain the retention figure from the data if it shows the same card number (Card No.) exists also in the latter month. Influence from other variables will be insignificant when only consider the staying ratio. According to

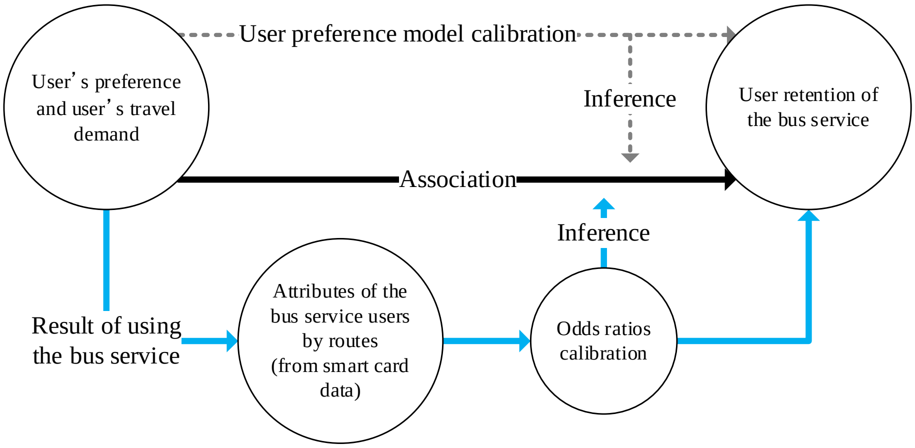
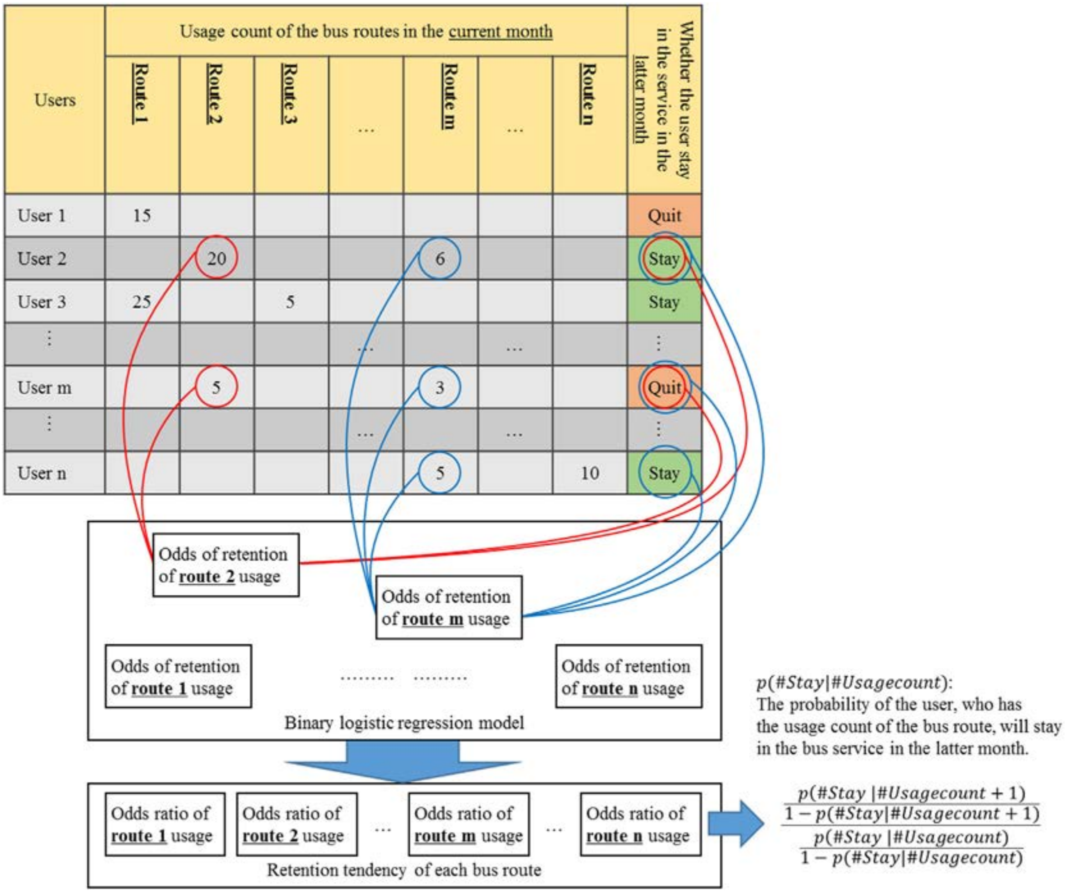

User preference model calibration
User' s preference and user' s travel demand
Inference
User retention of the bus service
Association
Inference
Result of using the bus service
Attributes of the bus service users by routes (from smart card data)
Odds ratios calibration

Fig. 5. Analysis framework.



Users	Usage count of the bus routes in the current month							Whether the user stay in the service in the latter month
	Route 1	Route 2	Route 3	…	Route m	…	Route n	
User 1	15							Quit
User 2		20			6			Stay
User 3	25		5					Stay
⋮				…		…		⋮
User m		5			3			Quit
⋮				…		…		⋮
User n					5		10	Stay

Odds of retention of route 2 usage
Odds of retention of route m usage
Odds of retention of route 1 usage
……… ………
Odds of retention of route n usage
Binary logistic regression model
Odds ratio of route 1 usage
Odds ratio of route 2 usage
…
Odds ratio of route m usage
…
Odds ratio of route n usage
Retention tendency of each bus route

$p(\#Stay|\#Usagecount)$: The probability of the user, who has the usage count of the bus route, will stay in the bus service in the latter month.

$$\frac{\dfrac{p(\#Stay\ |\#Usagecount+1)}{1-p(\#Stay|\#Usagecount+1)}}{\dfrac{p(\#Stay\ |\#Usagecount)}{1-p(\#Stay|\#Usagecount)}}$$

Fig. 6. Diagram of inference of retention tendency of using the bus service.

van Lierop et al. [24], user loyalty is a result of longer-term utilization and users' trust in the operating agency, but users may be influenced by any other possible or unknown variables. Onto the loyalty of the bus user, users may increase or decrease the usage according to various reasons. Generally speaking, the bus service satisfies the user's minimum requirement, when the user starts to use the bus service. If the user keeps using the service, it means the user has loyalty to the service. On the contrary, a user loses loyalty when a user discontinues the service. Therefore, we could regard the user retention as the index of user loyalty. Therefore, we use a bottom-up way to find the odds ratio of bus route usage first, and then proceed to find the reasons that may influence it. To calculate the odds ratio of each bus route, smart card data is the only type of data that needs collecting. Additionally, retention duration can be varied, depending on the planning purpose.

Logistic model regression is often used to calibrate a model with binary dependent variables [25–29]. An OR (odds ratio) is a measure of association between exposure and outcome. The OR represents the odds whose outcome will occur when being given some particular exposure, compared to the odds of the outcome occurring in the absence of that exposure. Therefore, “Staying” or “Quitting” of the bus service is the binary dependent variable in this model, and user's ridership of each route constitutes the independent variables. In this study, the ORBRU could present the change in the odds of retention when a user adds one ridership in a route. Here, we bypass all personal and socio-economical information but behavior transition data. Other variables, e.g., number of routes a user has used, or the user's ridership at peak hour, are used to enhance the calibration; and they will not obstruct the interception of ORBRU.

According to the above idea, we build a logistic regression model to calibrate the coefficients (log odds ratio) of the bus route usage. In order to enhance the calibration result, all samples (users) are split into four behavior groups, including REG, SREG, RAN, and RARE. REG is the users who use the service regularly, and most of them use the bus service in both morning and afternoon peak. SREG is the users who use the service sub-regularly, and their behavior is similar to REG group except the usage is lower than REG group. RAN is the users who use the service randomly. RARE is the users who use the service rarely (usage count is less than four per month). The key independent variables of the model are individual user's usage count on each bus route. Users will decide whether to use the bus service continuously or not, according to the bus route characteristics. Therefore, we assume that there is a relationship between the usage count and the odds of retention.

After the logistic model calibration, we can obtain some significant ORBRU, but others are not significant mainly due to the lower number of users. In order to be able to estimate the ORBRU and understand what factors may influence it, we build a multiple linear regression to estimate it. The diagram of building regression model is shown in Fig. 6. The dependent variable is the ORBRU that is calibrated via the logistic model, and the independent variables are monthly number of users, bus route characteristics, and socioeconomic data.

4. Case study

Tainan City is located in the southern part of Taiwan. In December 2010, Tainan County and Tainan City were 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 December 2012, a smart card system has been applied to all of the bus services. In June of 2013, the DOT of Tainan started to operate six main lines after re-planning the original bus services. Although the total ridership is continuously growing, city bus' mode share is still halted at 1–2%. The northern part of the municipality contains Tainan's lesser downtown area. Only one selected bus operator provides service for the northern rural and downtown area. Smart card data of that selected bus operator in 2016 (data in December is missing) is used in this study, and the total amount is 1,459,692 rows of data.

4.1. User behavior clustering results

For the selected bus operator, there were 26,729 smart cards used in March 2016, and 18,752 (70.2%) of them were random users who used that bus service less than four times within that month. The other 7977 smart cards were presented by users who used that bus service more than three times and, therefore, were clustered via the EM algorithm. After clustering and regularity sorting, nine usage patterns appeared plus one rare cluster, as listed in Fig. 7. Cluster 1 is of the most regular users who use the bus around 6:00 and 17:00, and should be regarded as standard commuters. The average usage frequency shows that they use buses almost every weekday. Cluster 2 is of the regular users who use the bus around 7:00, and there exists only half of its morning usage frequency at the afternoon peak. Cluster 3 is similar to cluster 2, but the morning peak starts around 6:00. Cluster 4 is of the users who use the bus at the afternoon peak only. Cluster 5 is similar to cluster 4, but their usage in morning peak is only 1.5 times on average. Cluster 6 is the opposite of Cluster 5 with less regularity at afternoon peak. Cluster 7 is of the users who use the bus not only at the morning and afternoon peaks but also in some instances at several hours adjacent to both peaks, such as college students or employees. Clusters 8 and 9 are of random users who use the bus at random departure times. The difference between these two is that Cluster 8 has higher usage frequency at weekends.

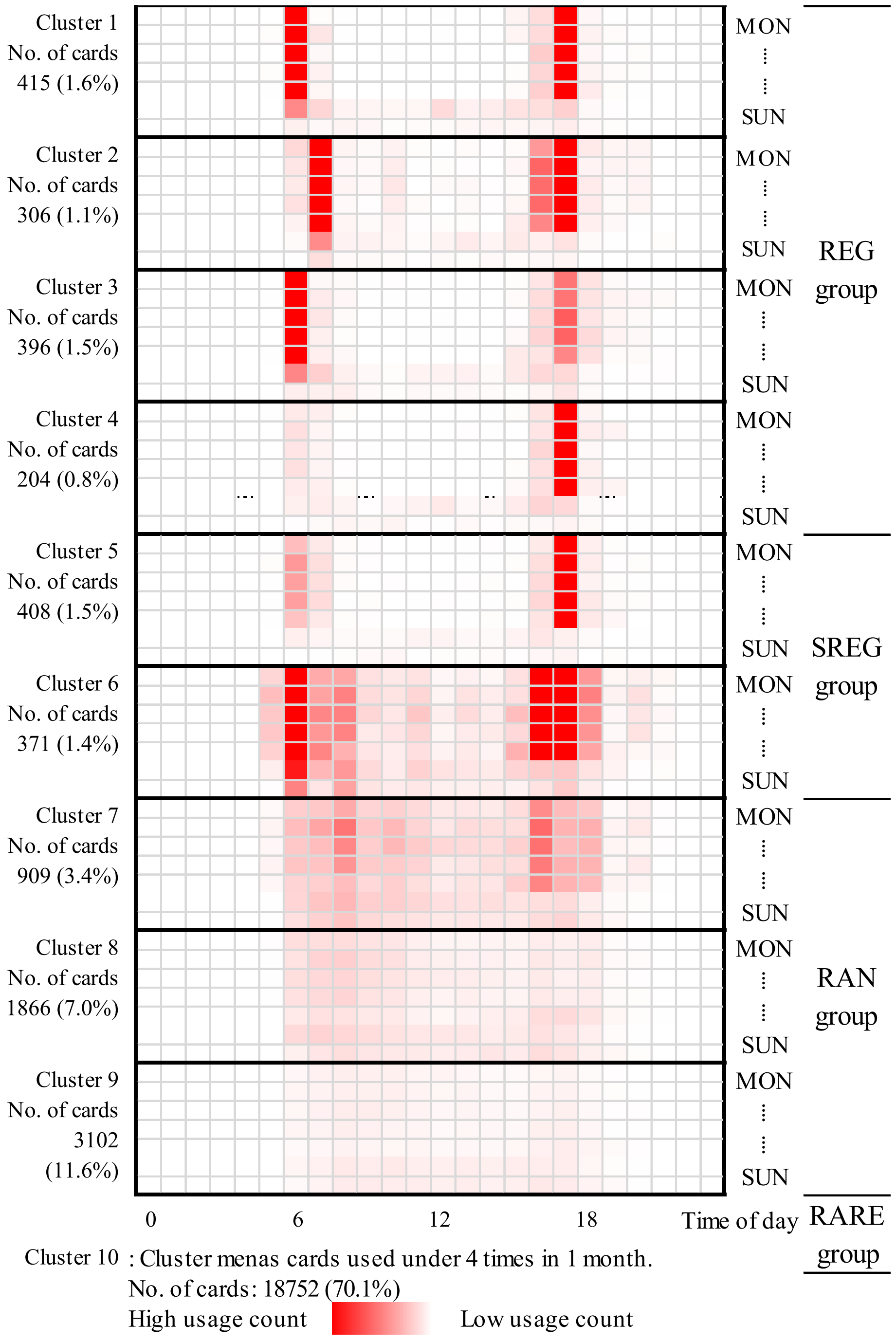


Fig. 7. Behavior patterns of all clusters in the selected bus operator (March 2016).

The results clearly show regularity sorting and are close to the situation in the real world. Since the number of users in high-regularity clusters is much lower than the number of low-regularity users, and since the regression model with small sample size is not easy to calibrate, we put similar behavior clusters together and make four groups according to regularity and frequency during the peak time. The four groups are REG group (regular), SREG group (sub-regular), RAN group (random), and RARE group (rare).

4.2. ORBRU calibration results

In this case, we select the data from one bus company in Tainan City. There are 21 city bus routes of that company, therefore, there are 21 independent variables for individual bus route usage, and plus one independent variable for other non-city bus routes. In order to enhance the explanatory power of the model, there are some independent variables are added into the model, i.e. usage count in peak hour, the number of routes and bus stops the user boarded. The dependent variables and independent variables of the logistic model are listed in Table 1. Those data could be summarized from the smart card data without extra cost. Therefore, each raw data includes one user's non-route data and summary of usage count of all bus routes. The logistic regression model in this research as shown in Eq. (1).

$$\begin{aligned} log\left(\frac{p}{1-p}\right) = {} & \beta_0 + \beta_1 MP\_UC + \beta_2 AP\_UC + \beta_3 Usagecount \\ & + \beta_4 UCinWE + \beta_5 UCinWK + \beta_6 NumofRoute \\ & + \beta_7 NumofStop + \beta_8 AvgUCStop + \beta_9 AvgUCStop \\ & + \beta_{10} NumofSCH + \beta_{11} IsSTU + \cdots + \beta_{21} NumofDay \\ & + \beta_{22} R1300 + \beta_{23} R1301 + \cdots + \beta_{28} R1500 + \cdots \\ & + \beta_{42} R1515 + \beta_{43} ROther \end{aligned} \tag{1}$$

Because all the users' behavior patterns are varied, it cannot consider which variables may influence the retention. Therefore, the calibration will include several variables with similar meaning, e.g. MP_UC, AP_UC, UCinWE… etc. However, if such variables exist in the model at the same time, it might lead to a multicollinearity problem. We will remove the unreasonable variables compositions and the variables with VIF (Variance Inflation Factors) which is larger than 10, and recalibrate again until all variables satisfy the foregoing conditions. Following are the unreasonable variables compositions.

- MP_UC, AP_UC, and Usagecount, cannot exist in the model at the same time. (Usagecount = MP_UC + AP_UC)
- UCinWE, UCinWK, and Usagecount, cannot exist in the model at the same time. (Usagecount = UCinWE + UCinWK)
- AvgUCRoute and NumofRoute, or AvgUCRoute and Usagecount, cannot exist in the model at the same time. (AvgUCRoute = Usagecount / NumofRoute)
- AvgUCStop and NumofStop, or AvgUCStop and Usagecount, cannot exist in the model at the same time. (AvgUCStop = Usagecount / NumofStop)

In this study, we use McFadden's pseudo-R^2 to assess the goodness of fit of the model. The McFadden's pseudo-R^2 statistic represents the percentage of the variance of the dependent variable that can be explained by the selected independent variables, and it shows as Eq. (2). The log likelihood of the intercept model is treated as a total sum of squares, and the log likelihood of the full model is treated as the sum of squared errors. Unlike the R^2 of linear regression, McFadden's pseudo-R^2 will not be a high value, especially in social science. The ridership of each bus route is the ridership summation of all users in each bus route, and the ridership of a single user is the summation of all

Table 1
Variables list of the binary logistic regression model.

Variables	Meaning
Dependent variable:	
$\log\left(\frac{p}{1-p}\right)$	p is the probability that Y for cases equals 1. Y is that whether the user stays or quits service in the latter month. Y is a binary variable. "Stay": 1; "Quit": 0.
Independent variables:	
Non-route variables	
MP_UC	User's total usage count at morning peak time (06:00–09:00).
AP_UC	User's total usage count at afternoon peak time (16:00–19:00).
Usagecount	User's total usage count in one month.
UCinWE	User's total usage count during weekdays.
UCinWK	User's total usage count during the weekend.
NumofRoute	The number of distinct routes that user used in one month.
NumofStop	The number of distinct bus stops that user boarded.
AvgUCRoute	User's average usage count of the bus routes that user boarded. (= Usagecount/NumofRoute)
AvgUCStop	User's average usage count per stop. (= Usagecount/NumofStop)
NumofSCH	The number of distinct bus stops whose names are actual school names that user boarded, e.g. Liouying junior high school or Baihe junior high school.
IsSTU	The card type: whether student card or not. "Yes": 1; "No": 0.
IsCharity	The card type: whether charity card or not. "Yes": 1; "No": 0.
FreqofRoute	Summation of the frequency of the distinct routes that the user boarded
StopsofRoute	Summation of the number of bus stops of the distinct routes that the user boarded
Population (persons)	Summation of the population of the distinct districts where the bus route destination, which the user boarded, located.
PDensity (persons/km^2)	Summation of the population density of the distinct districts where the bus route destination, which the user boarded, located.
RLength (km)	Summation of the length of the distinct bus routes that the user boarded.
ToSY	Summation of the linear distance from the bus route destinations, which the user boarded, to SinYing bus terminal. SinYing bus terminal is the largest terminal in the study area.
ToBH	Summation of the linear distance from the bus route destinations, which the user boarded, to BeiHe bus terminal.
TourRoute	The ratio of the number of routes, which its destination is near a sightseeing place and the user boarded, to the number of bus routes that the user boarded.
NumofDay	Number of days that the user uses the bus service in one month.
Route variables	
R1300, R1301, …, R1515	User's total usage count of each bus route. ("1300, 1301, …, and 1515" are the bus route codes, and there are 21 bus route codes in this case.)
ROther	User's total usage count of routes not in the list above.

decisions of that user. Therefore, the relationship between users' decisions and ridership of the bus route is indirect and includes various factors; consequently, we cannot expect a high value of McFadden pseudo-R^2. According to McFadden [30], values of 0.2 to 0.4 for ρ^2 (McFadden pseudo-R^2) represent an excellent fit, and the values in this research can represent an acceptable result [31,32]. After removing insignificant figures (p-value ≥0.1) and outliers, the yearly average ORs are shown in Tables 2 to 5.

$$\text{McFadden's pseudo} - R^2 = 1 - \frac{Log - likelihood(M_{Full})}{Log - likelihood(M_{null})} \tag{2}$$

M_{Full} : The model with independent variables.

M_{null} : The model without independent variables.

There are a total of four sets of yearly average coefficients, including REG, SREG, RAN, and RARE groups. REG and SREG groups have 10 sets (months) of calibration results; RAN group has 100 calibrations (since there are 10 subgroups in the RAN group, we made a calibration of each subgroup for the 10 months); and RARE group has 300 calibrations (since there are 30 subgroups in the RARE group, we made a calibration of each subgroup for the 10 months) of calibration results. All

Table 2
ORs calibrated results of REG users (use the service regularly) in each month.

Variables	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Avg.
Intercept	🟧 0.272 (-1.995)	-	🟨 0.171 (-3.153)	🟩 0.032 (-4.093)	🟧 0.216 (-2.569)	🟩 0.146 (-3.540)	🟧 5.246 (2.296)	-	🟩 0.074 (-3.589)	🟧 0.260 (-2.006)	0.802
UCinWE	🟨 0.826 (-2.680)	-	-	🟨 0.793 (-2.779)	-	-	-	-	-	-	0.810
NumofSCH	-	-	-	-	-	🟧 0.974 (-2.060)	-	-	-	-	0.974
FreqofRoute	-	-	-	-	-	🟨 1.024 (3.108)	-	-	-	-	1.024
PDensity	-	-	-	-	🟧 1.000 (-2.090)	-	-	-	-	-	1.000
NumofDay	🟩 1.309 (5.739)	-	🟩 1.318 (7.954)	-	🟩 1.309 (7.937)	🟩 1.234 (6.375)	-	-	-	🟩 1.333 (6.607)	1.301
R1300	-	-	-	-	-	🟩 0.959 (-4.511)	-	-	-	-	0.959
R1301	-	-	-	🟧 0.925 (-2.396)	-	-	-	-	-	-	0.925
R1502	-	-	🟧 0.949 (-2.077)	-	-	-	-	🟧 0.927 (-2.102)	-	-	0.938
R1503	-	-	🟨 0.959 (-3.121)	-	🟩 0.947 (-3.741)	-	-	-	-	-	0.953
R1504	-	-	-	-	-	-	-	-	-	🟧 0.957 (-2.180)	0.957
R1505	-	-	-	-	-	🟩 0.823 (-3.800)	-	-	-	-	0.823
R1506	-	-	-	-	-	-	-	-	-	🟨 0.951 (-2.726)	0.951
R1507	-	-	-	-	-	-	-	🟧 0.829 (-2.560)	-	-	0.829
R1511	-	-	-	-	-	🟧 0.950 (-2.538)	-	-	-	-	0.950
R1513	-	-	-	-	-	-	-	🟨 0.684 (-3.176)	-	-	0.684
R1514	-	-	-	-	-	-	-	-	🟧 0.904 (-2.085)	-	0.904
R1515	-	-	-	-	-	🟧 0.905 (-2.274)	-	-	-	-	0.905
ROther	-	-	🟩 0.923 (-4.036)	-	-	-	-	-	-	-	0.923
Following variables have no significant estimated parameters in all months. MP_UC, AP_UC, Usagecount, UCinWK, AvgUCStop, IsSTU, IsCharity, StopsofRoute, Population, RLength, ToSY, ToBH, TourRouteR1302, R1303, R1310, R1311, R1500, R1501, R1509, R1510, R1512											
McFadden's pseudo-R^2	0.107	-	0.207	0.213	0.168	0.149	0.112	0.183	0.189	0.139	-
Sample size	930	493	1271	1150	1268	923	395	342	1185	1307	-

Within these 10 calibration results, the coefficient is picked up according to the following conditions.
1. The p-value of ORs must be smaller than 0.1.
2. The lower and upper 95% confidence interval of each OR cannot include 1.0.
3. The outliers of the ORs are to be cut off.
4. All VIFs (Variance Inflation Factors) are smaller than 10 (most are smaller than 3).
5. T- statistics are shown in () below the OR values.
🟩: p-value <0.001; 🟨: p-value <0.01; 🟧: p-value <.05.

Table 3
ORs calibrated results of SREG users (use the service sub-regularly) in each month.

Variables	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Avg.
Intercept	-	-	-	-	-	0.063 (-7.603)	-	-	0.168 (-3.656)	-	0.115
MP_UC	-	-	-	0.914 (-3.204)	-	0.925 (-4.273)	-	-	-	-	0.919
AP_UC	-	-	-	-	-	-	-	-	1.100 (2.450)	1.109 (2.035)	1.105
Usagecount	-	-	-	-	-	-	-	-	0.913 (-4.405)	-	0.913
UCinWE	-	-	1.214 (1.981)	-	-	-	-	-	-	-	1.214
NumofRoute	-	-	0.612 (-3.673)	-	-	-	-	-	-	0.364 (-3.314)	0.488
NumofStop	-	-	-	-	0.798 (-2.889)	1.236 (2.792)	-	-	-	-	1.017
AvgUCRoute	-	1.243 (3.625)	-	-	-	-	-	1.113 (2.748)	-	-	1.178
IsCharity	-	-	-	-	3.041 (1.967)	11.906 (3.19)	-	-	-	-	7.474
FreqofRoute	-	-	-	-	-	1.071 (5.221)	-	-	-	-	1.071
StopsofRoute	-	-	-	-	-	0.99 (-2.671)	-	-	-	-	0.990
PDensity	-	-	-	-	-	0.999 (-2.900)	-	-	1.001 (3.970)	-	1.000
RLength	-	-	-	-	-	-	-	-	-	1.022 (2.464)	1.022
NumofDay	1.227 (5.968)	1.226 (2.254)	-	-	1.293 (5.439)	1.301 (8.141)	-	-	-	1.21 (4.079)	1.251
R1310	-	-	-	-	-	-	-	0.911 (-2.176)	-	-	0.911
R1500	-	1.205 (1.994)	-	-	-	-	-	-	-	-	1.205
R1501	-	-	-	-	-	-	-	0.91 (-2.616)	-	-	0.910
R1505	-	-	-	-	-	0.858 (-2.514)	-	-	-	-	0.858
R1507	-	-	-	-	-	-	-	0.887 (-2.769)	-	-	0.887
R1512	-	-	-	-	-	-	-	-	1.186 (2.147)	-	1.186
Following variables have no significant estimated parameters in all months. UCinWK, AvgUCStop, NumofSCH, IsSTU, Population, ToSY, ToBH, TourRoute, R1300, R1301, R1302, R1303, R1311, R1502, R1503, R1504, R1506, R1509, R1510, R1511, R1513, R1514, R1515, ROther											
McFadden's pseudo-R^2	0.130	0.180	0.236	0.179	0.184	0.293	-	0.202	0.200	0.245	-
Sample size	648	771	748	766	862	779	425	635	762	783	-

Within these 10 calibration results, the coefficient is picked up according to the following conditions.
1. The p-value of ORs must be smaller than 0.1.
2. The lower and upper 95% confidence interval of each OR cannot include 1.0.
3. The outliers of the ORs are to be cut off.
4. All VIFs (Variance Inflation Factors) are smaller than 10 (most are smaller than 3).
5. T- statistics are shown in () below the OR values.
🟩: *p*-value <0.001; 🟨: *p*-value <0.01; 🟧: *p*-value <0.05.

McFadden's pseudo-R^2 values are between 0.024 and 0.380, and all VIF (Variance Inflation Factor) values of independent variables are <10. Some McFadden pseudo-R^2 values here are not high, but they still show the same tendency as shown with other values that are higher. Further, the significant odds ratio can represent the retention tendency of the bus route usage.

In the non-route independent variables of REG and SREG groups, NumofSCH and IsSTU are both insignificant. This means that whether the area has a school or not and whether the card is a student card or not, the odds ratio will not be affected by these two variables. The ORBRUs of REG and SREG's NumofRoute variables are 0.678 and 0.603 respectively. These ORBRUs show that, as the number of routes a

Table 4
ORs calibrated results of RAN users (use the service randomly) in each month.

Variables	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Avg.
Intercept	0.378	–	–	–	0.414	0.384	–	0.456	0.400	0.345	0.396
MP_UC	1.034	–	0.844	0.788	0.795	0.848	–	–	0.845	0.871	0.861
AP_UC	–	–	0.813	0.824	0.832	0.837	1.195	1.177	0.813	0.852	0.918
Usagecount	1.164	–	–	–	–	1.187	–	–	–	–	1.175
UCinWE	1.159	–	1.124	–	1.162	1.132	–	–	–	–	1.144
UCinWK	0.901	0.885	0.896	0.898	–	0.894	0.887	–	–	–	0.894
NumofRoute	0.725	–	0.728	–	0.654	0.696	0.608	0.707	0.681	0.757	0.694
NumofStop	–	1.357	1.418	–	1.062	1.313	–	–	1.451	–	1.320
AvgUCRoute	–	–	–	0.830	–	1.067	–	–	1.014	0.773	0.921
AvgUCStop	0.707	0.726	0.770	–	–	0.770	0.756	0.764	0.682	–	0.739
NumofSCH	0.789	–	–	1.129	0.806	0.842	–	0.869	0.761	–	0.866
IsSTU	–	0.538	–	0.466	2.070	–	2.099	0.585	–	–	1.151
IsCharity	2.961	–	1.793	2.029	1.947	2.309	2.158	2.540	2.163	2.504	2.267
FreqofRoute	–	0.960	0.982	–	–	0.973	0.973	–	0.978	–	0.973
StopsofRoute	–	–	–	–	–	–	–	0.992	–	0.995	0.993
Population	–	–	–	–	–	–	–	–	–	–	–
PDensity	–	1.001	–	1.001	–	1.000	–	1.000	1.001	–	1.001
RLength	0.990	–	–	–	0.990	–	–	–	–	–	0.990
ToSY	–	–	–	0.945	–	–	–	–	0.967	–	0.956
ToBH	0.972	0.978	–	–	0.974	–	–	–	–	0.982	0.976
TourRoute	0.172	0.413	0.387	0.118	0.253	0.263	0.381	–	–	0.445	0.304
NumofDay	1.596	–	1.750	1.749	1.503	1.477	1.491	–	1.538	1.707	1.601
R1300	0.981	1.146	–	–	–	–	1.026	–	–	1.165	1.080
R1301	–	–	–	–	–	–	–	–	–	–	–
R1302	–	0.517	–	–	–	0.761	0.601	–	–	–	0.627
R1303	–	–	–	–	–	–	0.772	–	–	0.832	0.802
R1310	0.721	–	0.709	–	0.660	0.688	–	–	–	–	0.694
R1311	0.039	–	–	–	0.296	–	–	–	–	–	0.167
R1500	–	0.885	0.878	–	0.910	0.886	0.892	0.921	–	0.891	0.895
R1501	0.805	–	0.825	–	0.687	1.310	0.791	0.820	–	–	0.873
R1502	–	–	–	0.740	–	–	1.379	–	–	–	1.059
R1503	–	–	0.423	–	–	–	0.772	–	0.479	0.721	0.599
R1504	0.806	0.649	–	0.744	–	0.813	–	0.833	–	–	0.769
R1505	–	–	0.710	–	0.595	–	–	–	–	–	0.653
R1506	0.811	0.798	–	–	0.839	–	0.824	–	–	–	0.818
R1507	–	–	–	1.433	1.265	1.553	1.330	1.212	1.353	1.401	1.364
R1509	0.761	–	0.789	–	0.843	–	0.805	–	0.827	–	0.805
R1510	–	–	0.724	–	–	–	–	–	0.741	0.771	0.745
R1511	1.486	–	1.786	–	2.064	–	1.721	1.410	–	–	1.694
R1512	0.694	0.829	–	–	0.848	–	0.838	–	0.739	0.793	0.790
R1513	0.881	–	–	–	0.785	0.866	0.789	–	0.869	–	0.838
R1514	0.813	–	–	0.821	–	–	–	–	–	–	0.817
R1515	–	–	–	–	–	0.462	0.406	–	–	–	0.434
ROther	1.531	–	1.580	–	–	–	–	–	–	–	1.555
McFadden's pseudo-R^2	0.197–0.282	0.265–0.380	0.160–0.284	0.180–0.275	0.134–0.229	0.135–0.205	0.147–0.253	0.090–0.218	0.131–0.265	0.200–0.270	–
Sample size	536–589	542–604	495–558	514–583	502–558	542–611	614–694	662–749	447–510	553–621	–

There are 10 subgroups in each month, and the ORs in each month are the average of those subgroups.
Within calibration results, the coefficient picked up according to the following conditions.
1. The *p*-value of ORs must be smaller than 0.1.
2. The lower and upper 95% confidence interval of each OR cannot include 1.0.
3. The outliers of the ORs are to be cut off.
4. All VIFs (Variance Inflation Factors) are smaller than 10 (most are smaller than 3).

user opts for grows, the lower the probability that the user will stay in the service.

There are 13 significant routes related to ORBRUs in the REG group, and the other 9 are insignificant. The ORBRUs of users in REG group adding one usage count in most routes are decreasing. Generally speaking, the user has higher usage count should have higher probability to keep using the bus service. However, those ORBRUs are all negative. Because the REG users tend to use the service at fixed time, the higher usage count means they are students or commuters. The lower ORs here show the most bus routes could not satisfy or match user's demand.

There are only six significant routes independent of coefficients in the SREG group. In the RAN group, apart from R1301, other variables are significant. The sample size in both the RAN and RARE groups is rather large; therefore, most route-related variables are significant.

For all groups (behavior patterns), NumofDay is significant in four groups, and it means the NumofDay is a more suitable and common variable to infer the user retention than usage count. NumofDay is the number of days the user using the bus service in one month. When this value gets higher, the more days the user uses the service. Compared to the five usage count related variables, NumofDay is a common one for all users. Usage count related variables might have different influence according to the user's behavior pattern.

In the results of the previous ORs estimation, we select 11 variables to understand the relationship between user behavior and retention. The relationships are shown in Fig. 8. First is the variables related to service performance evaluation, including four variables about usage count, MP_UC, AP_UC, UCinWE, and UCinWK. When the managers evaluate the performance, the ridership (summation of usage count) is the most common index used. In Fig. 8a, it shows the ORs of the usage count of various behavior groups and time. If OR is large than 1, it

Table 5
ORs calibrated results of RARE users (use the service rarely) in each month.

Variables	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Avg.
Intercept	–	–	0.117	0.161	0.157	0.162	–	0.130	0.153	0.135	0.145
MP_UC	–	1.886	–	1.719	1.718	1.718	1.808	1.848	1.805	1.904	1.801
AP_UC	1.626	–	1.623	1.458	1.459	1.613	1.597	1.497	1.503	–	1.547
Usagecount	–	0.602	–	–	0.551	–	1.872	1.515	–	–	1.135
UCinWE	1.041	–	1.330	–	1.385	–	–	1.309	1.361	0.752	1.197
UCinWK	0.769	1.008	0.726	0.780	–	0.719	1.358	–	–	–	0.893
NumofRoute	0.474	0.398	–	–	0.287	0.610	0.241	0.408	0.440	0.409	0.408
NumofStop	–	–	1.160	0.644	1.564	1.547	–	–	–	–	1.229
AvgUCRoute	–	1.519	1.735	1.902	–	1.779	1.619	1.542	1.722	–	1.688
AvgUCStop	–	–	3.143	2.094	1.377	–	2.098	–	–	2.549	2.252
NumofSCH	–	2.566	1.378	1.898	2.122	1.315	–	2.521	2.317	2.514	2.079
IsSTU	1.728	–	–	0.533	1.821	0.547	0.581	–	1.837	–	1.174
IsCharity	1.990	–	–	1.997	1.779	1.760	1.765	1.795	1.934	1.808	1.854
FreqofRoute	–	–	0.959	0.973	0.987	–	–	–	0.980	0.989	0.977
StopsofRoute	1.017	1.016	–	1.023	1.019	1.015	1.001	–	1.016	1.022	1.016
Population	–	–	–	1.000	–	–	1.000	–	–	–	1.000
PDensity	–	–	–	1.000	1.001	–	1.001	1.001	–	–	1.001
RLength	–	1.013	–	–	1.018	–	–	0.993	0.992	0.983	1.000
ToSY	0.948	0.914	–	0.940	–	0.963	0.917	–	0.949	0.946	0.940
ToBH	0.945	1.049	1.046	1.037	–	–	0.950	–	–	1.037	1.011
TourRoute	2.655	2.765	–	1.894	2.160	2.254	–	3.773	2.908	2.434	2.605
NumofDay	2.585	–	–	2.638	2.464	–	–	2.262	2.513	2.625	2.515
R1300	1.150	–	1.615	1.350	1.657	1.373	1.638	1.361	–	1.694	1.480
R1301	5.340	–	–	–	–	11.183	–	–	9.529	3.255	7.327
R1302	5.613	–	3.757	–	2.725	–	4.181	0.113	0.042	–	2.738
R1303	5.470	–	–	6.348	5.631	4.825	–	7.875	–	–	6.030
R1310	5.719	5.558	–	–	–	5.949	–	–	–	–	5.742
R1311	–	–	–	–	–	–	–	–	–	–	–
R1500	–	1.377	–	1.314	1.424	–	1.025	1.007	–	1.510	1.276
R1501	–	2.355	1.160	–	1.630	–	–	–	–	1.393	1.634
R1502	3.305	4.376	2.694	–	3.006	3.332	2.951	2.793	0.302	2.605	2.818
R1503	–	2.832	0.077	–	–	–	–	4.179	0.220	–	1.827
R1504	–	4.213	–	–	–	3.558	3.442	3.968	–	–	3.795
R1505	–	14.555	–	–	–	–	–	4.587	–	–	9.571
R1506	6.091	2.752	–	6.764	3.383	–	–	–	2.855	–	4.369
R1507	1.755	–	–	1.681	1.449	1.944	–	2.126	1.688	1.014	1.665
R1509	0.675	0.598	0.593	0.626	–	–	0.636	0.617	0.652	0.645	0.630
R1510	–	–	–	4.492	–	–	–	5.959	–	–	5.226
R1511	2.541	–	8.697	5.613	3.899	–	1.432	2.551	–	2.761	3.928
R1512	0.307	0.151	0.123	0.190	0.247	0.208	0.354	–	0.056	0.198	0.204
R1513	1.777	0.252	–	–	3.948	–	–	–	0.217	0.222	1.283
R1514	–	4.237	–	–	–	4.452	–	–	–	–	4.344
R1515	–	–	–	–	–	–	–	–	–	–	–
ROther	–	0.453	3.399	1.342	0.365	3.218	–	–	–	4.109	2.148
McFadden's pseudo-R^2	0.049–0.145	0.058–0.188	0.084–0.206	0.050–0.160	0.048–0.135	0.043–0.114	0.024–0.132	0.056–0.151	0.056–0.140	0.060–0.178	–
Sample size	478–576	596–708	511–601	493–583	465–532	497–593	474–574	500–603	439–548	485–590	–

There are 30 subgroups in each month, and the ORs in each month are the average of those subgroups.
Within calibration results, the coefficient picked up according to the following conditions.
1. The *p*-value of ORs must be smaller than 0.1.
2. The lower and upper 95% confidence interval of each OR cannot include 1.0.
3. The outliers of the ORs are to be cut off.
4. All VIFs (Variance Inflation Factors) are smaller than 10 (most are smaller than 3).

means that the user has higher usage count will have a higher probability to stay in the service. In other words, the service is more suitable for the user has higher usage count. The result shows that the service is less suitable for the REG users in weekdays, and this is warning information to the managers. Because they usually use the bus more in weekdays, the result shows the service is not suitable for the users to have higher usage count instead.

Second is the variables related to service design, including four variables about bus route characteristics, FreqofRotue, StopsofRotue, RLength, and TourRoute. Those variables are basic characteristics for service design, the result in Fig. 8b shows the retention tendency of them. In the result, REG and SREG user have no observable tendency of all of those, because most of them are students or commuters and have no other options. However, they still have a slight tendency to using bus routes with higher frequency or longer bus routes. For RARE users, they prefer to use the bus routes to reach the sighting place. For RAN users, they have higher usage count and random boarding time in one month. If they went to sightseeing more times, they are usually tourists from other areas. That is, they will stop to use bus service next month. Here, the managers can try to enhance the service of tour-related bus routes to attract RARE users to keep using the bus service.

The third is the variables related to fare design, including three variables about the summary of the users using bus service, NumofDay, NumofRoute, and NumofStop. These variables are important indexes for evaluating the fare design. In Fig. 8c, all users have a retention tendency if they use the bus service more days in one month. It also shows the number of days is a good and general index to evaluate the retention for all users. The number of routes is also a general index to evaluate all users. When the users use fewer bus routes, the routes are more useful to the users. Although these results match the common sense, the ORs values give the manager a quantified index to evaluate various behavior pattern and their retention tendency.

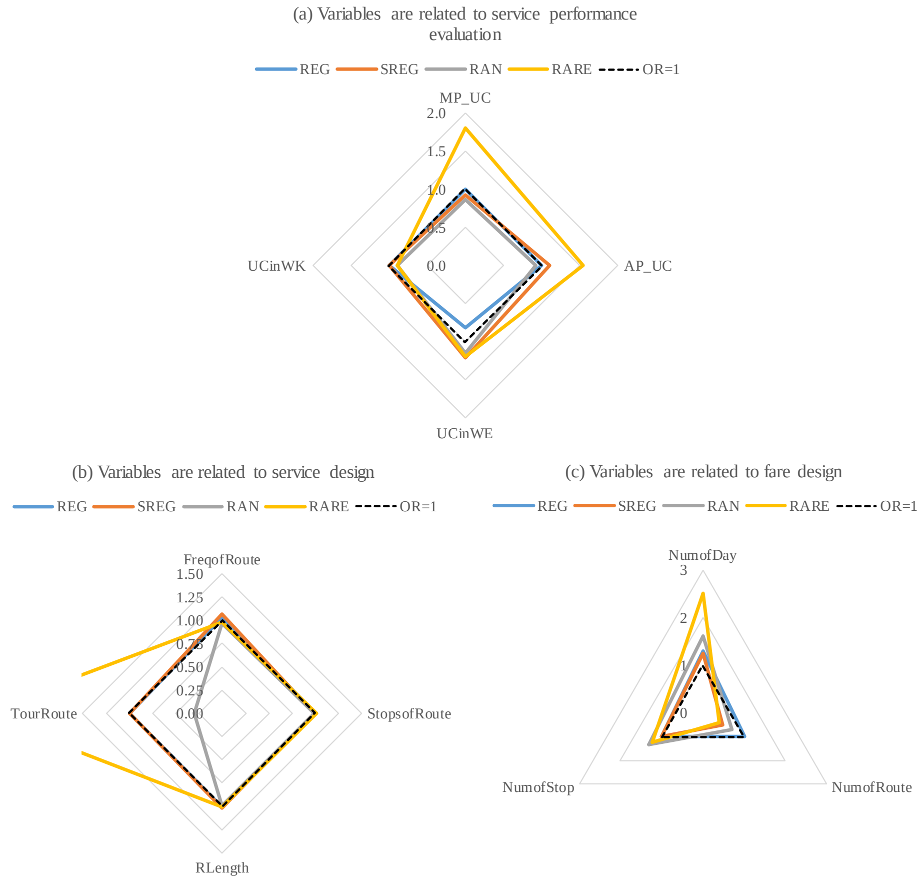


Fig. 8. Comparison of the odds ratios of three kinds of variables.

5. Conclusions

This study proposed an automatable procedure to obtain retention odds ratios for all bus routes from smart card data. The procedure can be encapsulated in a software program. Managers can then use such software to assist in their planning. Because it is automatable, managers can also have a cyclical process for review or planning of the bus service. By applying the EM method, bus users were clustered into several groups and their behavior transition from month to month showed whether they would stay or quit the service. Based on coefficients of route variables showed the ORBRU of each route. The calibration of ORBRU does not need any individual personal or individual socio-economical information but smart card transaction data. Therefore, this method dramatically decreases the cost and time for data collection.

To compare with traditional methodologies, this research uses smart card data only, a stable data resource that can be obtained from smart card system without plenty of time and budget. Managers can also obtain the ORBRU via this methodology when they put the methodology into their management system. Once the managers can understand user's reaction to the bus service, they can make quick response to user's reaction. Furthermore, the cyclic planning procedure could become faster due to the reduction of the response time. In the past, the managers needed much longer time to observe user's reaction based on a specific policy. By this research, the managers are able to reduce the observation time and use the resource more efficiently. Then limited resources could be allocated to the right position in order to increase the user retention and to use the resource more efficiently.

The composition of various behavior clusters varies in each bus route, and the odds of retention allow us to evaluate the performance of the bus routes. Once the odds of retention are known, managers can simulate the effect of various alternatives with hypothetical numbers of riders. Moreover, they can also find a better way of resource allocation based on a lower odds ratio. Future study should focus on the relationship among odds of retention and other operational variables, e.g., frequency or stop location. More detailed service planning can be proposed according to such relationships.

In Table 2 to Table 5, the managers can understand the influence of the same variables on the behavior of different types of users. In addition, the different effect of each bus route to various user behavior types is also obtained via the comparison of the odds ratios between different user behavior groups. The managers can understand the difference between the real effect and expected the effect of each bus route. Fig. 8 is the summary of the ORs for non-route variables. We split them into three groups, and use the radar graphs to show the difference. The result can assist the managers for service design and evaluation.

The method developed in this study is different from previous ones that assumed all users in the same behavior cluster to have the same tendency. Actually, ridership is a composition of diverse user types, which change from time to time. Resource allocation could be more appropriate with the understanding of who the target is and their retention probability. Needless to say, this bottom-up procedure can easily be applied to other transportation modes. In addition, it is clear that including wider characteristics of bus service will enhance the model. Therefore, managers can consider more specific characteristics if they have sufficient data to calibrate the model. In this method, once the bus routes with lower retention probability are known, their users' behavior type can also be understood. Therefore, managers can make suitable service improvement plans even with limited resources. Once the improvements are duly implemented, the service might get higher user retention in the near future. And last but not least, traffic safety will be improved if public transport systems can retain users, thereby reducing the use of private modes.

Acknowledgements

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.

Appendix A. Estimation of yearly ORBRU (odds ratio of bus route usage)

The sample data used for coefficients calibration are the behavior clustering result via the users' weekly profile, and the staying decision in the latter month is considered as the dependent variable. The estimation procedure of ORBRU is shown in Fig. A.1. Before calibration, all users are split into four behavior groups, including REG, SREG, RAN, and RARE groups. Behavior itself could be varied according to sample size or calibration result, but there is only one yearly ORBRU for each group. Generally speaking, the sample size of RAN and RARE groups are too large; users in these two groups can be split into 10 or more sub-groups to obtain a better calibration result. The monthly coefficients can be obtained from an average of the calibration results of all sub-groups. Furthermore, the yearly coefficients of each group can be obtained from the average of the monthly results and with the removal of outliers from the monthly results. We can get the ORBRU by exponentiating the yearly coefficients.

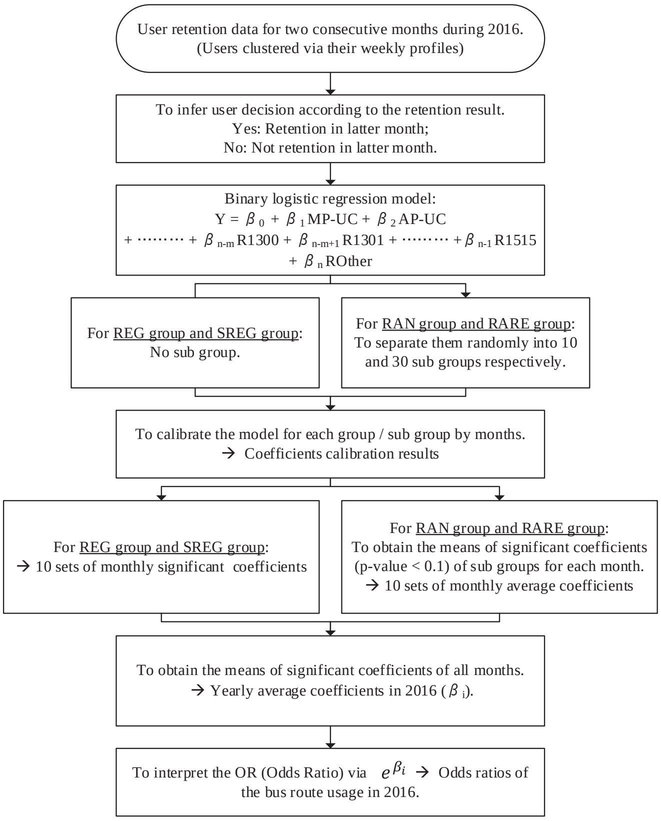

User retention data for two consecutive months during 2016.
(Users clustered via their weekly profiles)
To infer user decision according to the retention result.
Yes: Retention in latter month;
No: Not retention in latter month.
Binary logistic regression model:
$Y = \beta_0 + \beta_1 \text{MP-UC} + \beta_2 \text{AP-UC} + \cdots\cdots + \beta_{n-m} R1300 + \beta_{n-m+1} R1301 + \cdots\cdots + \beta_{n-1} R1515 + \beta_n ROther$
For REG group and SREG group:
No sub group.
For RAN group and RARE group:
To separate them randomly into 10 and 30 sub groups respectively.
To calibrate the model for each group / sub group by months.
→ Coefficients calibration results
For REG group and SREG group:
→ 10 sets of monthly significant coefficients
For RAN group and RARE group:
To obtain the means of significant coefficients (p-value < 0.1) of sub groups for each month.
→ 10 sets of monthly average coefficients
To obtain the means of significant coefficients of all months.
→ Yearly average coefficients in 2016 (β_i).
To interpret the OR (Odds Ratio) via e^{β_i} → Odds ratios of the bus route usage in 2016.

Fig. A.1. Estimation procedure of yearly ORBRU.

References

[1] M.D. Meyer, E.J. Miller, Urban Transportation Planning: A Decision-Oriented Approach, McGraw-Hill, New York, 2001.
[2] WHO (World Health Organization), Global Status Report on Road Safety 2018, 2018.
[3] J.P. Stimpson, T. Litman, The Hidden Traffic Safety Solution: Public Transportation, Am. Public Transp. Assoc., 2016 1–72.
[4] P. Adney, The Complete Guide to Passenger Retention, 2015.
[5] State of Florida Department of Transportation, Transit Ridership, Reliability, and Retention, 2008.
[6] T. Satiennam, S. Jaensirisak, W. Satiennam, S. Detdamrong, Potential for modal shift by passenger car and motorcycle users towards Bus Rapid Transit (BRT) in an Asian developing city, IATSS Res. 39 (2016) 121–129.
[7] L. Wang, L. Li, B. Wu, Y. Bai, Private car switched to public transit by commuters, in Shanghai, China, Procedia–Social and Behavioral Sciences 96 (2013) 1293–1303.
[8] L. Steg, Can public transport compete with the private car? IATSS Res. 27 (2003) 27–35.
[9] J. Zhao, V. Webb, P. Shah, Customer loyalty differences between captive and choice transit riders, Transp. Res. Rec. 2415 (2014) 80–88.
[10] M. Bagchi, P.R. White, The potential of public transport smart card data, Transp. Policy 12 (2005) 464–474.
[11] C. Morency, M. Trepanier, B. Agard, Analysing the Variability of Transit Users Behaviour with Smart Card Data, 2006 IEEE Intelligent Transportation System Conference, 2006 44–49.
[12] C. Morency, M. Trépanier, B. Agard, Measuring transit use variability with smart-card data, Transp. Policy 14 (2007) 193–203.
[13] C. Zhong, E. Manley, S. Müller Arisona, M. Batty, G. Schmitt, Measuring variability of mobility patterns from multiday smart-card data, J. Comput. Sci. 9 (2015) 125–130.
[14] TRB (Transportation Research Board), Bus Route Evaluation Standards, 1995.
[15] R. Raicu, S. Raicu, Complex aspects of transport quality, WIT Trans. Built Environ. 77 (2005) 281–290.
[16] Y.O. Susilo, T.B. Joewono, W. Santosa, An exploration of public transport users' attitudes and preferences toward various policies in Indonesia: some preliminary results, J. East. Asia Soc. Transp. Stud. 8 (2010) 1230–1244.
[17] A.A. Laverty, E. Webb, E.P. Vamos, C. Millett, Associations of increases in public transport use with physical activity and adiposity in older adults, Int. J. Behav. Nutr. Phys. Act. 15 (2018) 1–10.
[18] M. Grimsrud, A. El-Geneidy, Driving transit retention to renaissance: trends in Montreal commute public transport mode share and factors by age group and birth cohort, Public Transp. 5 (2013) 219–241.
[19] L. Ding, N. Zhang, A travel mode choice model using individual grouping based on cluster analysis, Procedia Eng. 137 (2016) 786–795.
[20] P.-H. Hung, K. Doi, H. Inoi, User behavior transition mapping for bus transportation planning based on time series data analysis of travel E-ticket information, J. East. Asia Soc. Transp. Stud. 12 (2017) 738–756.
[21] M.K. El Mahrsi, E. Côme, J. Baro, L. Oukhellou, Understanding passenger patterns in public transit through smart card and socioeconomic data: a case study in Rennes, France, 3rd International Workshop on Urban Computing (UrbComp 2014), New York, 2014.
[22] E.I. Pas, Weekly travel-activity behavior, Transportation 15 (1988) 89–109.
[23] A.K.M. Tarigan, S. Fujii, R. Kitamura, Intrapersonal variability in leisure activity-travel patterns: the case of one-worker and two-worker households, Transp. Lett. 4 (2012) 1–13.
[24] D. vanLierop, M.G. Badami, A.M. El-Geneidy, What influences satisfaction and loyalty in public transport? A review of the literature, Transp. Rev. 38 (2018) 52–72.
[25] A. Al-Doori, Waiting time factor in public transport by binary logistic regression, Aust, J. Basic Appl. Sci. 11 (2017) 72–76.
[26] O. Chiu Chuen, M.R. Karim, S. Yusoff, Mode choice between private and public transport in Klang Valley, Malaysia, Sci. World J. (2014) (2014) 7–9.
[27] A. Ismail, A.E. Elmloshi, Logistic regression models to forecast travelling behaviour in Tripoli City, Int. J. Adv. Sci. Eng. Inf. Technol. 1 (2011) 618–623.
[28] L. Sun, K.W. Axhausen, D.-H. Lee, X. Huang, Understanding metropolitan patterns of daily encounters, Proc. Natl. Acad. Sci. 110 (2013) 13774–13779.
[29] S. Tao, J. Corcoran, I. Mateo-Babiano, Modelling loyalty and behavioural change intentions of busway passengers: a case study of Brisbane, Australia, IATSS Res. 41 (2017) 113–122.
[30] D. McFadden, Quantitative methods for analyzing travel behaviour of individuals: some recent developments, Behav. Travel Model. (1978) 279–318.
[31] J. Cohen, Statistical power analysis, Curr. Dir. Psychol. Sci. 1 (1992) 98–101.
[32] B. Hu, J. Shao, M. Palta, Pseudo-R^2 in logistic regression model, Stat. Sin. 16 (2006) 847–860.