Modeling Mode Choice Behaviors for Public Transport Commuters: A Case Study of Beijing

Weng Jiancheng[1], Tu Qiang[1], Yuan Rongliang[2], Lin Pengfei[1], Chen Zhihong[3]
（1. Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing 100124, China; 2. Beijing Municipal Institute of City Planning &Design, Beijing 100044; 3. Beijing Municipal Transportation Operations Coordination Center, Beijing 100073）

Abstract

Based on an analysis of the factors that influence the mode choice behaviors of commuters who use public transport (subways and buses), this research developed a questionnaire by using a combination of revealed preference (RP) and stated preference (SP) techniques; both online and paper surveys were conducted to gather commuters' travel choices between subways and buses. A binary logit (BL) specification was proposed to examine public transport commuters' travel choices. The regression coefficients were estimated using maximum likelihood estimation. Finally, based on the data obtained from Beijing public transport smart cards, a support vector machine (SVM) classification model was established to identify commuters' mode choices, and the accuracy was found to be as high as 94.24%. The estimated mode choice model was employed to predict the market shares of both subways and buses after a new fare scheme was implemented. The results showed that the model had high prediction accuracy: the average absolute error for predicting the market share of buses was 5.93%.

Keywords: public transport, commuting trip, Binary-Logit model, support vector machine, travel choice

1. Introduction

Faced with increasingly intensified traffic congestion, research on the priority given to public transport is of great importance. Given the continuously reducing cost of public transport combined with implementation of a series of public transport priority policies, the daily average usage of public transport smart cards was approximately 15 million instances in Beijing, which constitutes nearly 50% of the urban residents' travel demand *(1)*.

With the rapid development of urban public transport and the increasing separation between jobs and residential locations, passenger travel mode choice demonstrates an obvious imbalance in some large cities. During the morning and evening peaks in particular, travelers tend to choose rail transit, which is of high reliability and high speed and can thus lead to heavy congestion and a potential safety risk for the subway. This could also lead to a high unloaded ratio and a low utilization ratio of facilities for ground buses. With the scale expansion of the public transport system over the past five years, the proportion of rail transit travel has nearly doubled *(2)* in Beijing (Table 1). In 2014, the maximum load factors of Line 5 and Line 13 exceeded 120% in

peak hours on more than 220 days each *(1)*, indicating that the rail passenger flow in the peak section is obviously oversaturated. Research on mode choice behaviors for public transport travelers is a top priority for many big cities with similar problems; thus, it can provide support to formulate policies and measures and offer reasonable travel guidelines to public transport passengers.

Table 1 Proportion of Public Transport Travel

Travel mode	2009	2010	2011	2012	2013	2014
Ground bus	78.45% (18270 km*)	73.23% (18743 km)	69.65% (19469 km)	67.67% (19547 km)	60.18% (19697 km)	58.49% (20249 km)
Rail transit	21.55% (228 km)	26.77% (336 km)	30.35% (372 km)	32.33% (442 km)	39.82% (465 km)	41.51% (527 km)

*Note: The total mileage of the lines for ground bus or rail transit is given in the bracket.

In recent years, the research on travel mode choice has been mostly based on traveler survey data to analyze the factors that influence travel choice behavior. Logit or deformation models are used to establish a travel mode choice model and reveal the travel choice decision process and characteristics for residents of different types or in different conditions. The logit model was derived by Luce (1959), but the initial model can only be used to predict the choice of two transport modes *(3)*. Subsequently, McFadden (1974) discussed the logit model, and his system theory of the disaggregate model came to be widely used in the study of the travel mode choice *(4)*. Cervero and Radisc (1996) and Cervero and Kockelman (1997) used the binary logit model to compare the travel mode of commuting and non-commuting travel *(5-6)*. Asensio (2002) and Pinjari (2007) developed the MNL (Multinominal Logit) model to compare several travel patterns *(7-8)*. Based on travel survey data from German residents in 2003, Bohler (2006) analyzed the trip distance and travel mode quantitatively, studying holiday travel choice behavior, travelers' personal characteristics and change strategy of travel choice during holidays *(9)*. The MNL model was used to study the relationship between travel mode choice and travel activity selection by Stephan (2007), and it was found that travel mode choice changes greatly for different travel activities *(10)*.

The commuting trip is the main use of public transport services for big cities, which has stability in space and time to some extent compared with other types of trips, and the time arrangement of commuters affects the choice of other activities and trips. Commuters play an important role in public transport passenger flow in the morning and evening peaks, which causes an uneven distribution of passenger flow and high load on the subway. The previous studies conducted on commuting trip characteristics have mainly focused on the analysis of influencing factors such as commuting travel mode and trip time. Abane (1993) used a simple MNL model to study the choice behavior of travel plans in commuting trips and found that commuters' gender, income and age are the key factors *(11)*. By establishing a discrete choice model, Bhat (1997) found that the number of commute stops is closely related to commuting travel choice behavior *(12)*. Krishna Rao (1997) investigated the commuting travel behavior in the Atlanta area and found that the gender of the commuter had a significant influence on the travel behavior on the way to or

from work *(13)*. In his investigation, Turner (1998) found the existence of interaction between different family members in commuting trips *(14)*. Lu and Pas (1999) constructed a structural equation model and studied the relationship among social economic impact, activity participation and travel behavior *(15)*. Focusing on commuting activity, Golob (2000) analyzed the relationship among influence factors of trip time, participation of activities and trip generation *(16)*. The BL (Binary Logit) model based on the data of SP survey was established by Alvinsyah (2005), and the results showed that the higher the service performance of the public transport system is, the more likely it is that travelers will choose it *(17)*. By establishing a BL model, Gebeyehu and Takano (2007) found that the factors of bus trip cost, convenience and frequency of departure have an important influence on the travel choice of bus *(18)*.

In conclusion, there is much research on travel mode choice and commuting trip behavior, but few studies have focused on public transport commuters, analyzing the characteristics of commuting trip or establishing prediction models of mode choice in the public transport system to reveal the travel choice decision-making process of public transport commuters. Instead, in previous research, the verification method for the travel mode choice model had insufficient actual data support, which made it more difficult to further verify and apply the model.

This paper established a binary logit model to predict mode choice for public transport commuters based on the data of a mode choice behavior questionnaire survey. Furthermore, to verify the prediction model with actual data, a SVM (support vector machine) classification model was established to identify commuters' mode choices. The estimated mode choice model was employed to predict market shares of both subways and buses after a new fare scheme was implemented, and the results were validated with data obtained from Beijing public transport smart cards.

2. Data

2.1 Questionnaire design

Survey by questionnaire was conducted to collect information of public transport travel choice, including RP (Revealed Preference) and SP (Stated Preference) surveys. Parts of RP survey questionnaires were used to obtain personal attributes and travel characteristics as the basic input parameters of the travel choice prediction model.

Travel intention is the most important content of the survey and also the key factor that causes commuters to choose a particular public transport travel mode or transfer. Compared with travelers with other trip purposes, commuters are much more concerned about the cost of a trip, which means that they are more sensitive to the change of trip cost. Meanwhile, commuters pay more attention to the reliability and efficiency of the trip plan. For other travelers, the efficiency and reliability of travel mode is independent, which means they will think travel mode has better efficiency if running speed is higher, and they will think travel mode has better reliability if the punctuality rate is higher or the passenger waiting time is shorter*(19)*. However, for commuters, the regularity of the commuting trip makes them obtain more comprehensive information about different choices on their commuting travel routes such that they are able to fully understand the

advantages and disadvantages of each trip plan. Because of the strong time constraints of the commuting trip, commuters will comprehensively consider the reliability and efficiency of travel modes, predict the trip time for each trip plan, and evaluate and make choices based on the estimation of trip time. Therefore, the cognition of trip time and trip cost are key factors in trip plan choice and transfer from one travel mode to another for commuters. In a SP survey, the major purpose is to investigate selection tendency of commuters when the difference of trip cost and time among trip plans changes and to construct the main travel mode choice probability model of commuters in public transport.

The questionnaire is divided into two parts based on the above analysis, revealed preference (RP) and stated preference (SP) survey, and includes 26 questions. In the RP survey, the personal attributes and travel information of travelers were collected. In the SP survey, different levels of trip cost difference (2, 3, 4, 5, 6 RMB) and trip time difference (0-5, 6-15, 16-30, >30 min) between bus and subway were combined to ask commuters in which situation will they exchange the original main commuting travel mode for another. For the trip cost and trip time difference between bus and subway and based on the development orientation of bus and subway in most cities, the trip efficiency of the subway is considered higher than that of a bus, and the trip cost of the subway is higher.

2.2 Data collection and processing

An online survey was implemented over one week in December 2015, with 565 valid questionnaires collected. A paper survey was implemented in the same period, covering various locations in Beijing, particularly focusing on competitive stations where commuters can choose both bus and subway or important transfer stations. Because the investigation concerned public transport commuters, the survey period was set as early peak (07:00-09:00) and late peak (17:00-19:00), with 309 valid questionnaires collected.

In the preprocessing of survey data, using the method of conditional filtering, the rationality of the correlation between the selected results and the different variables was judged, and invalid data with obvious logical errors were excluded. Specific cleaning rules are as follows:

(1) Because the research object of this paper is public transport commuters, it is necessary to remove the survey data from those who have no fixed jobs.

(2) Because of the public welfare of public transport combined with the fare level in Beijing, the cost of each trip will not be too high for public transport commuters. Based on this, data for a single trip costing more than 10 RMB were considered invalid.

(3) If the trip distance and time or the trip distance and cost are obviously out of proportion, the data were considered invalid.

(4) In SP survey concerning trip efficiency and cost difference, there is a certain logic in the setting of the subjects and options, which means that the selection of the former question must be smaller than the latter. Thus, data that did not conform to this logic were excluded.

In total, 708 valid samples were obtained after data cleaning, including 369 commuters whose main trip mode is ground bus and 339 commuters whose main trip mode is rail transit. The description statistics of the collected data are shown in Table 2.

151 **Table 2 Descriptive Statistics of Survey Data**

Variable	Descriptive	Percent (%)	Variable	Descriptive	Percent (%)
Main trip mode	Bus	52.12	Main trip mode	Metro	47.88
Individual characteristics			**Individual characteristics**		
Gender	Male	51.72	Gender	Male	59.18
	Female	48.28		Female	40.82
Age group	⩽20	3.39	Age group	⩽20	4.64
	21-35	72.31		21-35	78.15
	⩾36	26.29		⩾36	17.21
Education level	Below high school	2.84	Education level	Below high school	1.38
	High school	16.48		High school	5.52
	College degree	30.68		College degree	11.72
	Bachelor degree	38.64		Bachelor degree	53.1
	Master degree and higher	11.36		Graduate study and higher	28.28
Average monthly income (RMB)	⩽1500	8.09	Average monthly income (RMB)	⩽1500	15.17
	1501-3000	4.62		1501-3000	10.34
	3001-5000	37.57		3001-5000	22.76
	5001-8000	30.64		5001-8000	29.66
	⩾8000	19.08		⩾8000	22.07
Household Car Ownership	Yes	64.37	Household Car Ownership	Yes	50.34
	No	35.63		No	49.66
Travel Information			**Travel Information**		
Trip distance (km)	＜5	22.67	Trip distance (km)	＜5	2.8
	5-10	41.28		5-10	11.19
	10-15	20.93		10-15	25.17
	15-20	8.72		15-20	18.18
	＞20	6.4		＞20	42.66
Trip time (min)	＜30	26.16	Trip time (min)	＜30	8.22
	30-40	25.58		30-40	13.7
	40-50	20.35		40-50	15.75
	50-60	12.21		50-60	13.01
	＞60	15.69		＞60	49.32

152 Note: Limited by length, the attributes of occupation are not included in the table, and other attributes are just partially listed or
153 group merged.

3. Modeling and discussion

To reasonably guide passenger flow, reduce subway travel pressure during the peak hour and optimize public transport travel structure, this paper mainly studied how the variables influence public transport commuters when they choose between the travel modes of bus and subway, assuming that the two travel modes are equally valuable options. Therefore, the BL model was selected to predict the public transport commuting travel choice behavior and discuss the transfer rule of subway users to ground bus.

3.1 Model foundation: binary logit model

The basis of binary logit model is the theory of "utility maximization". The default of the theory is that the trip behavior decision is always the pursuit of utility maximization; that is, the travel decision makers always choose the trip plan to obtain the most effectiveness for themselves, which provides them with demand satisfaction or pleasure, and the utility is different with the change of factors such as characteristics of trip plan and travelers' personal attributes.

Assuming that U_{in} is the utility derived by individual n to choose mode i and that it is usually divided into a random term $\varepsilon_{\mathrm{in}}$ and a fixed term V_{in}, it can be expressed as follows:

$$U_{\mathrm{in}} = V_{\mathrm{in}} + \varepsilon_{\mathrm{in}} \tag{1}$$

Assuming that X_{ink} is the k^{th} explanatory variable in mode i for individual n, θ_k is the coefficient associated with the k^{th} explanatory variable, and K is the number of variables included in the model, then the formula for V_{in} is

$$V_{\mathrm{in}} = \sum_{k=1}^{K} \theta_k X_{ink} \tag{2}$$

Based on the theory of "utility maximization", the probability P_{in} that traveler n chooses travel mode i from travel mode set A_n is:

$$P_{\mathrm{in}} = Prob\left(U_{\mathrm{in}} > U_{\mathrm{jn}},\ \mathrm{i} \neq \mathrm{j},\ \mathrm{j} \in \mathrm{A}_n\right) = P\left(V_{\mathrm{in}} + \varepsilon_{\mathrm{in}} > V_{\mathrm{jn}} + \varepsilon_{\mathrm{jn}},\ \mathrm{i} \neq \mathrm{j},\ \mathrm{j} \in \mathrm{A}_n\right)$$

If $\varepsilon_{\mathrm{in}}$ is in double exponential distribution with the parameters of (0, 1), the probability to select the first mode in the BL model is:

$$P_{1\mathrm{n}} = \frac{1}{1+e^{-\left(v_{1n}-v_{2n}\right)}} \tag{3}$$

The probability to select the second mode is:

$$P_{2\mathrm{n}} = 1 - P_{1\mathrm{n}} = 1 - \frac{1}{1+e^{-\left(v_{1n}-v_{2n}\right)}} = \frac{e^{-\left(v_{1n}-v_{2n}\right)}}{1+e^{-\left(v_{1n}-v_{2n}\right)}} \tag{4}$$

3.2 Factor analysis

Before the model is established, the independent variables to be put into it are filtered to eliminate irrelevant variables and avoid interference of the model calibration. The filtering method of the independent variables is based on the correlation test between each independent variable and the dependent variable, using the Pearson chi-square test (Table 3).

Table 3 the Result of Chi-Square Test

Attribute	Chi-Square value	Sig.	Attribute	Chi-Square value	Sig.
Age	4.363	0.225	Average monthly income	4.704	0.032
Occupation	14.556	0.1827	Car in household	1.663	0.645
Education level	11.035	0.026	Gender	6.821	0.009

According to the correlation test, gender, education level, average monthly income and the main public transport travel mode choice of commuters have a certain relationship. To reduce the number of parameters to be determined in the model, a correlation analysis of the influence factors with too many groups was conducted, including the average monthly income and education level.

The results showed that the correlation coefficient in the income groups of 0-1500 RMB, 1501-3000 RMB and 3001-5000 RMB is greater than 0.5, indicating strong correlation, and that the correlation coefficient in income groups of 5001-8000 RMB and over 8000 RMB is weak. The selection results of commuters whose education level is below high school and college have a stronger correlation (0.793), whereas the correlation with other education levels is weak, and the selection results for commuters with education levels of bachelor degree, master degree or higher have stronger correlations (0.521).

Finally, the discrete variables of personal attributes to be put into the commuter travel mode choice model are gender, average monthly income (divided into three groups: ≤5000 RMB, 5001-8000 RMB and >8000 RMB) and education level (divided into two groups: college degree and lower, bachelor degree and higher). In addition to the personal attribute variables, this research is mainly concerned about the influence of trip cost and time consumption on public transport travel choice behavior, so bus trip cost and time, subway trip cost and time were selected as the independent variables of travel attributes incorporated into the model.

3.3 Analysis results

To establish the BL model of public transport commuting travel mode choice, the model input parameters and model structure are shown in Tables 4 and 5.

Table 4 Model Input Parameters

Variable name	Attribute	Value
Gender	discrete variable	1：male 2：female
Education level	discrete variable	1：college degree and lower 2：bachelor degree and higher
Average monthly income	discrete variable	1：≤5000 RMB

		2：≥5001 RMB & ＜8000 RMB 3：≥8001 RMB
Bus trip time	continuous variable	actual value
Bus trip cost	continuous variable	actual value
Subway trip time	continuous variable	actual value
Subway trip cost	continuous variable	actual value

Table 5 Introduction of Model Structure

Mode selection	Intrinsic dumb element	Gender	Education level	Average monthly income	Bus trip time	Bus trip cost	Subway trip time	Subway trip cost
	X_{in1}	X_{in2}	X_{in3}	X_{in4}	X_{in5}	X_{in6}	X_{in7}	X_{in8}
Bus	1	1	1	1	X_{1n5}	X_{1n6}	0	0
Subway	0	0	0	0	0	0	X_{in7}	X_{in8}

Therefore, the travel utility difference when the individual *n* chooses the ground bus or subway as the main travel mode is

$$V_{\text{bus}} - V_{\text{subway}} = \theta_1 \times (1-0) + \theta_2 \times X_{in2} + \theta_3 \times X_{in3} + \theta_4 \times X_{in4} + \theta_5 \times (X_{in5}-0) - \theta_6 \times (X_{in6}-0) + \theta_7 \times (0-X_{in7}) + \theta_8 \times (0-X_{in8}) \quad (5)$$

where V_{bus} is the fixed term of the utility function when commuter *n* chooses bus as the main travel mode; V_{subway} is the fixed term of the utility function when commuter *n* chooses subway as the main travel mode; and θ_1, θ_2, θ_3, θ_4, θ_5, θ_6, θ_7 and θ_8 are coefficients of independent variables.

The parameters of the model are calibrated based on the survey data, and the results are shown in Table 6.

Table 6 Calibration Results for Main Travel Mode Choice Model of Public Transport Commuters

Parameters	θ	S.E	t
Constant term	-0.2533	0.6271	-3.7039
Gender	1.0309	0.2473	4.1689
Education level	-0.1864	0.1691	-2.9028
Average monthly income	-0.4217	0.1569	-2.6861
Trip distance	-0.1148	0.0326	-3.5209
Bus trip time consumption	-1.4302	0.2723	-5.2523
Bus trip cost	-0.1703	0.0126	-13.5184
Subway trip time consumption	-0.9084	0.0948	-9.5810
Subway trip cost	-0.2021	0.0164	-12.3295

$L(0)$ ： -2944.4892 $L(\hat{\theta})$ ： -3172.8334 $-2[L(0) - L(\hat{\theta})]$=456.6874

ρ^2 =0.2798 $\bar{\rho}^2$ =0.2682 N=4248

Therefore, the main travel mode choice model for public transport commuters is as follows:

$$V_{1n} - V_{2n} = -0.2533 + 1.0309 \times X_{n1} - 0.1864 \times X_{n2} - 0.4217 \times X_{n3} - 0.1148 \times X_{n4} - 1.4302 \times X_{n5} - 0.1703 \times X_{n6} + 0.9084 \times X_{n7} + 0.2021 \times X_{n8}$$

$$P_{1n} = \frac{1}{1 + e^{-(V_{1n} - V_{2n})}} = \frac{1}{1 + e^{-(-0.2533 + 1.0309 \times X_{n1} - 0.1864 \times X_{n2} - 0.4217 \times X_{n3} - 0.1148 \times X_{n4} - 1.4302 \times X_{n5} - 0.1703 \times X_{n6} + 0.9084 \times X_{n7} + 0.2021 \times X_{n8})}} \times 100\%$$

$$P_{2n} = 1 - P_{1n}$$

where V_{1n} is the fixed term of utility function when commuter n chooses bus as the main travel mode; V_{2n} is the fixed term of utility function when commuter n chooses subway as the main travel mode; P_{1n} is the probability that commuter n chooses bus as the main travel mode; P_{2n} is the probability that commuter n chooses subway as the main travel mode; X_{n1} is the value of gender of commuter n; X_{n2} is the value of education level of commuter n; X_{n3} is the value of average monthly income of commuter n; X_{n4} is the value of trip distance of commuter n; X_{n5} is the value of bus trip time of commuter n; X_{n6} is the value of bus trip cost of commuter n; X_{n7} is the value of subway trip time of commuter n; and X_{n8} is the value of subway trip cost of commuter n.

It can be seen that the model has high precision, whereas the value of the evaluation parameters of goodness ratio ρ^2 and the goodness ratio after adjustment of degree of freedom $\bar{\rho}^2$ are both in the range of 0.2~0.4.

From the model calibration parameters and t-test values, the absolute values of the t-values of the input parameters' attributes in the model are all greater than 1.96, consistent with the result of contingency table analysis. This result indicates that the variables of gender, age, income, trip distance, bus trip time, bus trip cost, subway trip time and subway trip cost all have an important influence on the choice of travel mode at the 95% confidence level.

Female commuters are less likely than male commuters to take the subway as their main commuting travel mode (coefficient is positive), which is different from some other researches and presumably because the subway in Beijing is too crowded for female commuters while the bus is relatively comfortable. Another distinctive feature of public transport in Beijing is the intolerable delay for bus commuters because of serious congestion in peak hours. As a result, public transport commuters with higher income levels are inclined to choose subway travel (coefficient is negative), presumably because of this group's high sensitivity to trip time while the time value is positively correlated with income *(20)*.

Concerning the influence of travel attributes, the longer the trip distance is, the more likely

it is that consumers will choose subway travel (coefficient is negative). Through a comparative analysis of influence coefficient of bus and subway trip cost, the higher the trip cost difference is between subway and bus, the more likely commuters are to choose bus as the main travel mode, and the influence of subway trip cost on travel mode choice is greater than bus trip cost. Similarly, the longer the trip time difference is between subway and bus, the more likely it is that commuters will choose subway as the main travel mode, and the influence of bus trip time on travel mode choice is significantly greater than subway trip time. Through a comparative analysis of the coefficients of trip time and cost, commuters are more concerned about trip time, which means that the influence of efficiency is much greater than cost on the travel mode choice between bus and subway.

4. Model validation with big data

Based on the public transport multi-source smart card data, the accurate classification of public transport commuters was realized. Combined with the big data analysis of public transport fare adjustment in the end of 2014 in Beijing, the prediction accuracy of the travel mode choice model for public transport commuters was validated.

4.1 Multi-source data and SVM classifier

The merging and preprocessing of the public transport multi-source data were realized (Table 7), based on bus IC card data, rail transit AFC data, bus GPS data and related static basic data, using the preprocessing method of public transport multi-source data proposed by Wang*(21)* and by following steps that included invalid data filtering and related field extraction.

Table 7 Public Transport Basic Data Integration

Card ID	START_ TIME	END_ TIME	START_ LINE	START_ STATION	END_ LINE	END_ STATION	MODE*	START_STATION NAME	END_STATION NAME
00001076	7:12:05	7:49:04	5	21	10	35	R	TianTongYuan Northern	HuJiaLou
00001074	8:46:08	9:21:10	52	16	52	1	B1	NanLiShiLu	PingLeYuan
00016445	6:19:25	6:19:25	87	5	87	9	B2	BaiShiQiao Estate	DongWuYuan
00016447	6:59:05	7:58:32	6	55	10	7	R	DaLianPo	ZhiChunLi
00016432	18:36:00	19:57:59	94	33	10	41	R	ShaHe	JinSong

*Note: In the table, “R” means a subway trip; “B1” means a flat-fare bus trip (only exists before the public transport fare adjustment in the end of 2014); “B2” means a sectional-fare bus trip.

By obtaining 978 IC card numbers of public transport commuters or non-commuters through online survey, the data can be matched with the travel information of multi-source data. According to the preliminary study of commuting trip characteristics in the research group, boarding time, boarding line, boarding station, alighting time, alighting line and alighting station were selected as the feature values to establish an SVM classifier for commuter identification.

SVM is a classification model that is suitable for small samples, with strong feasibility and universality in the machine learning field. In recent years, this method has been gradually used in

the field of transportation. Wassantachat (2009) used the online SVM classifier and a background modeling technique (OSVM-BG) to estimate traffic density based on virtual loop detector data *(22)*. Sun (2012) solved the problem of SVM in practical application and proposed parallel SVM using in the identification of traffic congestion *(23)*. Mingheng (2013) proposed a multi-step traffic flow prediction model based on SVM *(24)*. Compared with other classification models, SVM classifier is especially suitable for the case of limited samples and has the advantages of effectiveness and stability *(25)*.

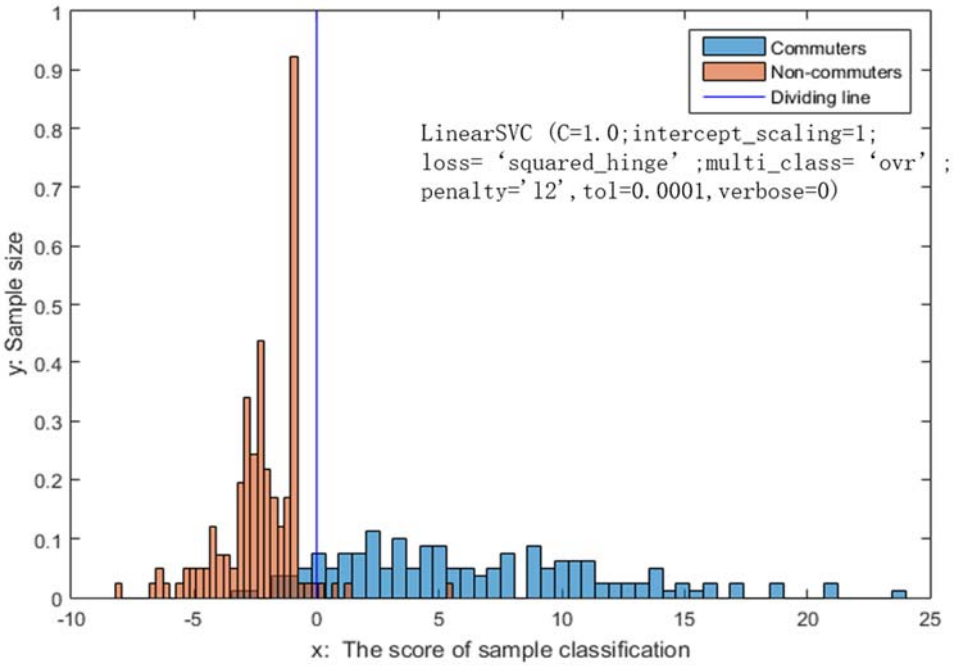

Figure 1 Distribution of Test Set Classification Result with SVM Model

All 978 samples were randomly divided into a training set and a test set at a proportion of 7:3. Using the test set of 295 samples to examine the SVM classification model, the result is shown in Figure 1. By calculating the SVM model, each sample was given a score that could be compared with the score of the dividing line to classify two travel groups; the rate of correctly identifying public transport commuters was as high as 94.24% (Figure 1). Based on the research of Wang *(21)* combined with the classification of commuters, the public transport trip chain of commuters was defined and extracted to analyze the main trip chain structure (bus or subway) and trip distance distribution. The results indicated that the SVM classification model can be used to identify public transport commuters' mode choice with high accuracy and is an effective way to validate the predicted results of the BL mode choice model.

4.2 Validation and application

Based on the travel choice model and smart card data used to calculate the proportion of travelers choosing bus or subway as their main commuting travel mode after the fare adjustment on 28 December 2014, the results were compared to validate the accuracy of the BL model. This was

also a case study for the model's application.

Validation of the travel mode choice prediction after the fare adjustment

The proportion of commuters choosing bus or subway as their main commuting mode within different trip distances after the fare adjustment was calculated by travel mode choice model and smart cart data, and the results are shown in Figure 2.

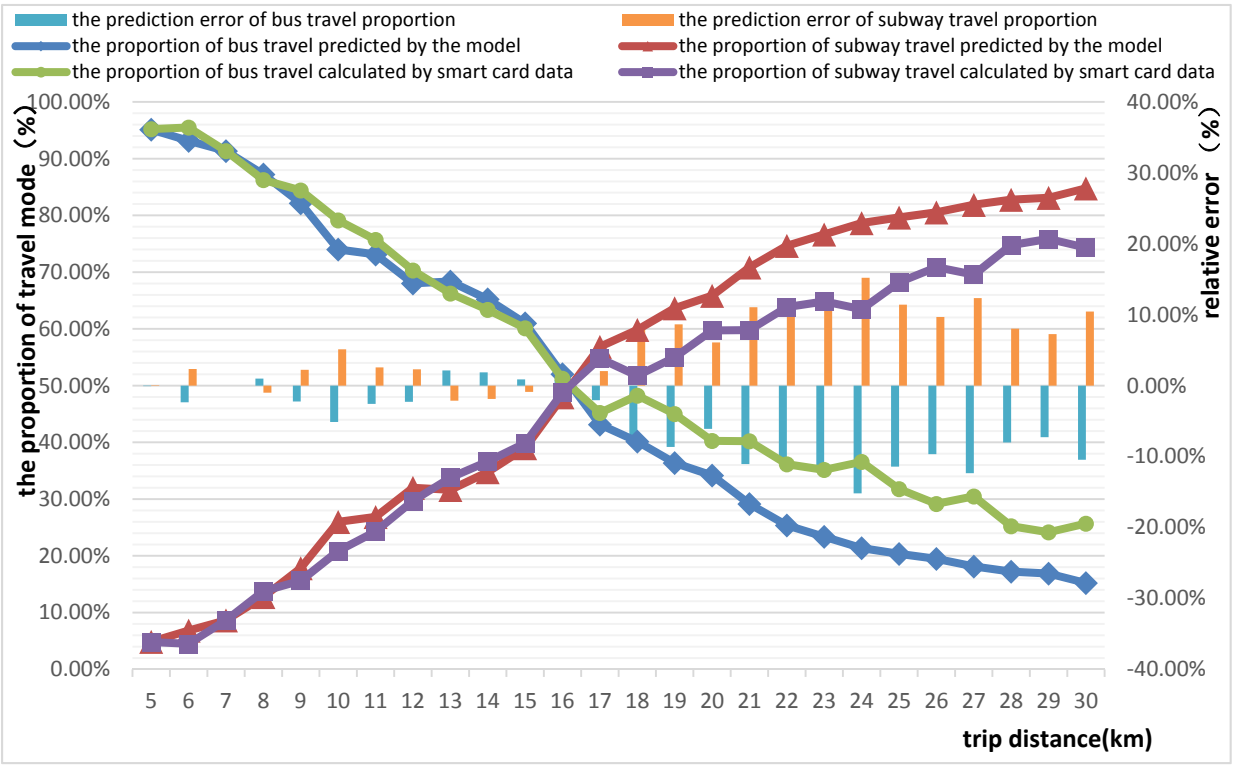

Figure 2 Results for Smart Card Data and Model Calculation after the Fare Adjustment

The results showed that the model proposed is relatively accurate in its calculation of public transport commuting travel mode choice. The absolute error in the prediction with commuting trip distance below 21 km is less than 10%, and the mean absolute error is 2.89%. When the commuting trip distance is over 21 km, the mean absolute error of prediction is 10.80%, and the predicted results showed that bus travel accounts for a smaller proportion compared with the fact of the situation. Overall, the average absolute error of predicted results in different trip distances is 5.93%, and the prediction error rises with an increase in trip distance.

Validation of the transfer ratio prediction for commuters after the fare adjustment

The proportion variation of commuters choosing bus and subway with different trip distances after the fare adjustment was calculated by the model and smart card data to validate the sensitivity of the model to the variation of trip cost difference. The results are shown in Figure 3.

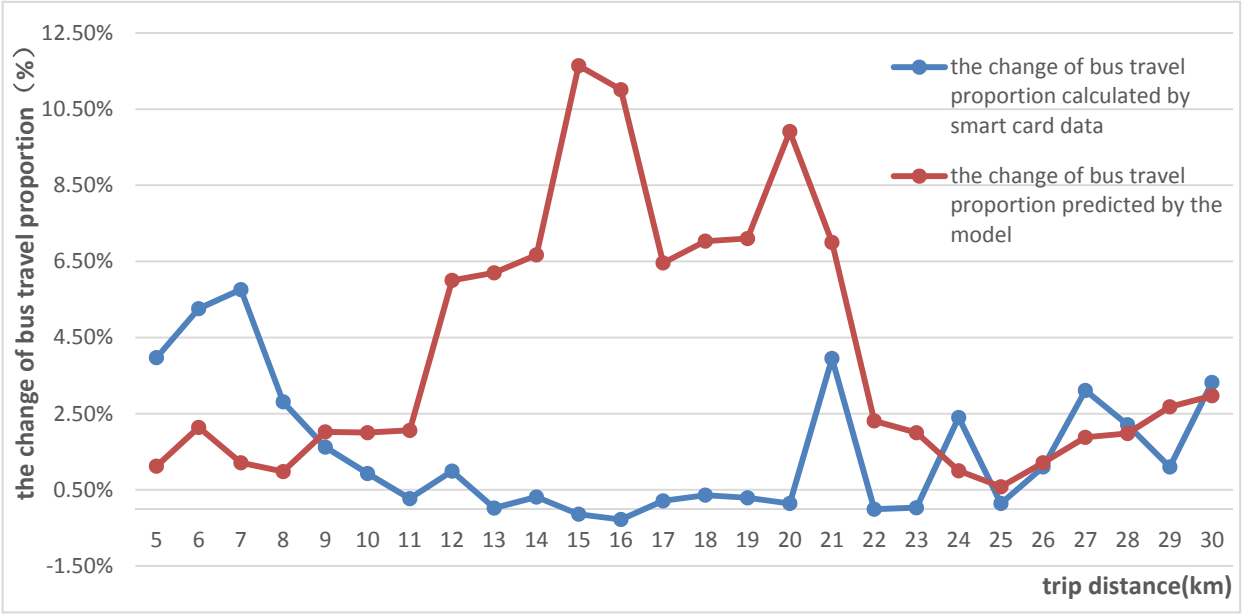

Figure 3 Calculation Result of Bus Travel Sharing Ratio Changes Based on Smart Card Data and Prediction Model

By calculating the smart card data, although the trip cost difference between bus and subway increased after the fare adjustment, the public transport travel of commuters has not changed significantly. In addition to the 4.45% increase in bus travel for trip distances shorter than 8 km, the travel proportion of bus and subway has not changed much in other trip distance ranges, consistent with the related research *(26)*.

Comparing the predicted results of the model with the calculation results of smart card data, the predicted results of the model showed that the proportion of bus travel increases from 3% to 11% within different trip distances. Except for short-distance travel, a large number of travel mode transfers occur in the trip distance range of 12 km to 21 km, where the proportion of bus travel increases on average by 7.9%, different from the actual situation calculated based on the smart card data. The prediction error is presumably caused by external factors such as new subway lines (four new subway lines were opened after 28 December 2014 in Beijing) and subway coverage. The results also indicate that commuters' concerns about the trip cost difference are less than the weight of trip cost given in the model, thus causing the model's sensitivity to trip cost to be higher than the actual.

Overall, the model can accurately predict the main travel mode choice of public transport commuters under different fare levels for subway and bus within different trip distances, and the prediction error is acceptable, with an average absolute error below 6%. The model's prediction of the proportion of bus travel is slightly less than the actual situation, and the proportion of the subway travel is slightly higher than the actual value.

In addition, through the validation, the model's sensitivity to the variation of trip cost difference is slightly higher, indicating that the weight of trip cost given in the model may be slightly higher than the actual. However, the prediction error is acceptable, with the maximum deviation to predict the travel proportion after the fare adjustment below 12%, under the conditions

that some influential external factors exist, such as the new subway lines.

5. Conclusion

Through the combination of SP and RP techniques, a survey by questionnaire was conducted to obtain the travel choices of public transport commuters. The main travel mode choice logit model of public transport commuters was established to reveal the mechanism of the mode choice of public transport commuters and determine the influential factors and their degree of importance in trip plan selection. This research can provide effective theory and data support for policy making related to public transport, line network planning and service level improvement.

In addition, based on a massive set of public transport multi-source smart card data and the machine learning theory, an SVM classification model was constructed with high accuracy to identify public transport commuters' travel mode choice. The estimated mode choice model of binary logit was employed to predict market shares of both subways and buses after a new fare scheme was implemented in Beijing, and the predicted results were compared to the actual results calculated by the SVM model based on smart card data. The average absolute error of predicted results is 5.93%, and the prediction error rises with the increase of trip distance. The results show that the model has high precision, and the error is within the acceptable range. This indicates that the model can be used to forecast the impact of policies and measures such as public transport fare adjustment to improve the scientific nature of policy making. More importantly, the present study abandons the traditional method of model validation using survey data and realizes instance verification and analysis of a large sample, with a wide range and high accuracy, and instead uses public transport smart card data and the data mining technique. This method can be used to validate other mode choice models in the future.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (NFSC) (No.51578028) and the Ministry of Transport of the People's Republic of China (No. 2015318221020). The authors would like to show great appreciation for the support.

References

1. Beijing Municipal Transportation Operations Coordination Center. The annual report of transportation operations coordination in Beijing. 2015.
2. Beijing Transport Research Center. Beijing Transport annual report. 2015
3. Luce D. Individual Choice Behavior [M]. New York: John Wiley and Sons, 1959.
4. McFadden D. Conditional logit analysis of qualitative choice behavior [M]. Frontiers in Econometrics. New York: Academic Press, 1974:105-142.
5. Cervero R, Radisch C. Travel choices in pedestrian versus automobile oriented neighborhoods

[J]. Transport Policy, 1996, 3 (3):127–141.
6. Cervero R, Kockelman K. Travel demand and the 3Ds: density, diversity, and design [J]. Transportation Research, 1997, Part D 2 (3):199–219.
7. Asensio J. Transport mode choice by commuters to Barcelona's CBD [J]. Urban Studies, 2002, 39(10): 1881–1895.
8. Pinjari A.R, Pendyala R.M., Bhat, C.R., et al. Modeling residential sorting effects to understand the impact of the built environment on commute mode choice [J]. Transportation,2007, 34 (5): 557–573
9. Susanne Bohler， Sylvie Grischkat， Sonja Haustein， Mareel Huneeke. Encouraging environmentally sustainable holiday travel[J].Transportation Research Part A，2006，40，652-670.
10. Krygsman Stephan, Arentze T, Timmermans H. Capturing tour mode and activity choice interdependencies: A co-evolutionary logit modelling approach [J]. Transportation Research Part A: Policy and Practice, 2007, 41:913–933.
11. Abane. A. Mode choice for the journey to work among formal sector employees in Accra, Ghana [J]. Journal of Transport Geography, 1993, 1(4): 219-229.
12. Bhat CR. Work travel mode choice and number of non-work commute stops[J]. Transportation Research B, 1997, 31(1):41-54.
13. Subba Rao P V, Sikdar P K, Krishna Rao K V, et al. Another insight into artificial neural networks through behavioural analysis of access mode choice [J]. Computers, Environment and Urban Systems, 1998, 22:485–496.
14. Turner T, Niemeier D. Travel to work and household responsibility: new evidence [J]. Transportation, 1997, 24(4):397-419.
15. X Lu, EI Pas. Socio-demographics, activi1y participation and travel behavior[J]. Transportation Research A, 1999, 33(1):1-18.
16. Golob T F. A simultaneous model of household activity participation and trip chain generation [J]. Transportation Research Part B: Methodological, 2000, 34:355–376.
17. Alvinsyah, Soehodho S, Nainggolan P J. Public transport user attitude based on choice model parameter characteristics(case study: Jakarta bus way system)[J]. Journal of Eastern Asia Society for Transportation Studies, 2005, 6:480-491.
18. M Gebeyehu, SE Takano. Diagnostic evaluation of public transportation mode choice in Addis Ababa[J] Journal of Public Transportation.2007, 10(4): 27-50.
19. Nurdden A, Rahruat R A, Ismail A. Effect of transportation polices on modal shift from privatecar to public transport in Malaysian [J]. Journal of Applied Sciences, 2007, 7(7): 1013-1018.
20. Tao Wang, Xiaokuan Wang, Xiaoming Liu. The Measurement Method for the Travel Time

422 Cost and the Analysis of Its Influencing Factors[J]. Road Traffic & Safety,2006,(04):19-22.

423 21. Yueyue Wang. Research on Methods of Extracting Commuting Trip Characteristic Based on
424 Public Transportation Multi-Sourced Data[D]. Beijing University of Technology, 2014.
425 22. Wassantachat.T, Zhidong Li, Jing Chen, et al. Traffic Density Estimation with On-line SVM
426 Classifier[C]// Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE
427 International Conference on. IEEE, 2009:13-18.

428 23. Sun Z Q, Feng J Q, Liu W, et al. Traffic congestion identification based on parallel SVM[C]//
429 Natural Computation (ICNC), 2012 Eighth International Conference on. IEEE, 2012:286-289.

430 24. Mingheng Z, Yaobao Z, Ganglong H, et al. Accurate Multi steps Traffic Flow Prediction Based
431 on SVM[J]. Mathematical Problems in Engineering, 2013, 32(2):544-554.

432 25. F. Debole, F. Sebastiani. An Analysis of the Relative Hardness of Reuters-21578 Subsets:
433 Research Articles.2005,56(6):584~59

434 26. http://society.people.com.cn/n/2015/0108/c136657-26350924.html