The Market Segmentation on Passenger Transportation of High-speed Railway with Logistic Regression Model

Yanjin Li, Hai Zhu, Yonghong Liu*
School of Transportation and Logistics
Southwest Jiaotong University
Chengdu, P.R China
291188963@qq.com

***Abstract*—In this article, traveling passengers along Chongqing-Lichuan High-speed Railway line are regarded as research object in this paper. More than 5000 random sampling data of passenger travel and their characteristic along the Chongqing-Lichuan High-speed Railway line are obtained through SP/RP questionnaire design, and through the market segmentation by using Logistic regression to deal with these sampling data, three market segmentations are obtained: migrant workers, non-economic travel and business travel. Then three traveler characteristic indexes, age, monthly income and profession are statistically analyzed in each subordinate market segmentation, which could be helpful for the operators of Chongqing-Lichuan High-speed Railway to design differentiated products, thus realizing revenue maximization of passenger high-speed railway operation.**

Key words—Chongqing-Lichuan High-speed Railway, random sampling, Logistic regression, Market segmentation

I. INTRODUCTION

Since people came up with the concept of market segmentation, scholars at home and abroad spent a lot on studying it. In terms of overseas study, Tony pointed out that segmentation research owned 2 study directions: customer oriented and product oriented [1]. Tsai used time consumption, consumption frequency and consumption amount these three variables to subdivide and identify most valuable customers [2]. In terms of domestic study, Zhao used factor and cluster analysis method on Beijing-Shanghai High-Speed Railway passenger survey data for research [3]. Qian and Liao adopt mixed regression model to subdivide market into efficiency, economical, casual and recreational market segments [4]. The current domestic study of railway market segmentation gives priority to the method with product oriented, which reflects that there is little study on market segmentation from passenger travel characteristics. Therefore, the article based on customer oriented to subdivide market of high-speed railway by passenger travel behavior characteristics, and analysis travelers' characteristics from each market segment, which owns important research significance and practical value.

This research was supported by the Technological Research and Development Programs of China CRH sleeper Corporation (Project No.:2014X006-A).

II. SURVEY OVERVIEW

The survey is divided into two stages: preliminary investigation and formal investigation. The preliminary investigation started in Chongqing North Station waiting hall on Sep.5.2015, and respondents were high-speed rails and electric multiple units (EMU) passengers. we improved questionnaire from summarizing problems and defects in preliminary investigation stage and designed the RP/SP combination questionnaire The formal investigation started in station waiting halls along Chongqing-Lichuan railway line on Sep.12.2015.We adopt RP/SP questionnaire design to survey Chongqing-Lichuan High-speed Railway passengers.

This survey included 2 working days and a rest day, and all recycling questionnaires from 7 railway station: Chongqing North Station, Fusheng Station, Changshou North Station, Fuling North Station, Fengdu Station, Shazi Station and LIchuan Station. The amount of each station questionnaires statistic on 2 workdays and one weekend as following diagrams:

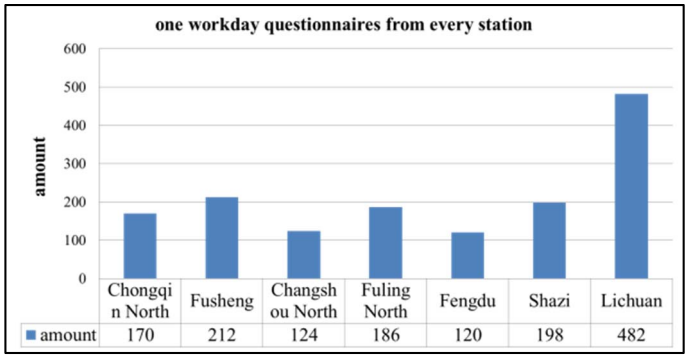

Fig. 1.One workday questionnaires from every station

We went on the preliminary investigation on a workday (Wednesday), and recycled 1492 questionnaires. From above diagram, it shows that the amount of respondents from

978-1-5090-1102-5/16/$31.00 © 2016 IEEE

Lichuan Station was the most, which would impact our effect on statistical data of survey. So, the paper improved structure of questionnaire and simplified some insignificant questions and selected another workday (Thursday) to go on the formal investigation.

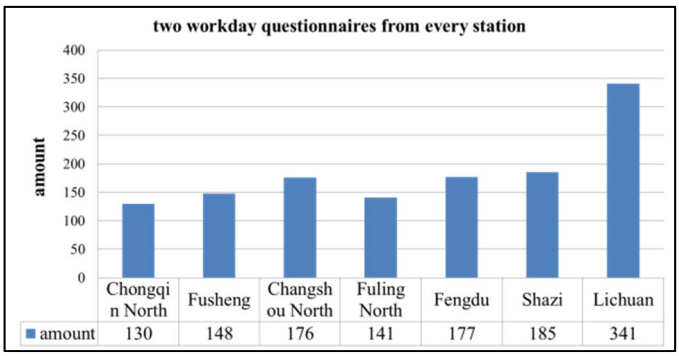

Fig. 2.Two workday questionnaires from every station

On another workday, we adopt improved questionnaire to go to on the formal investigation. From above diagram, the paper got that the amount of respondents of Lichuan Station was still the most, almost reaching the quarter of gross (1298), which hinted that there would be other factor impacting the result of survey. So as another part of survey, we should went on the formal investigation again on a weekend (Saturday) to test whether the time (workday or weekend) was an important factor on our survey.

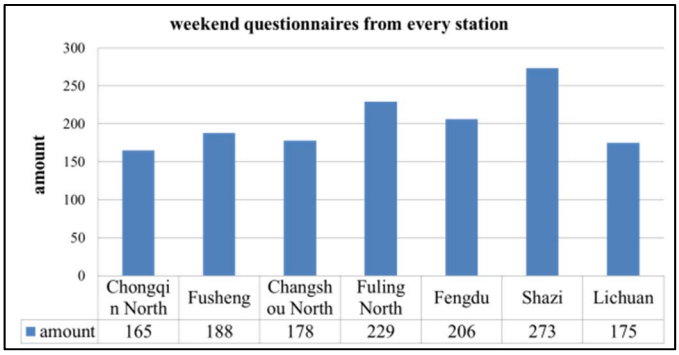

Fig. 3.Weekend questionnaires from every station

From this diagram, the amount of each station was uniformly distributed, which indicated that the time of going on survey was an important factor to impact the results. So, we selected two sets data respectively from workday (Thursday) and weekend (Saturday) as study samples. And the investigation samples used method of stratified sampling and randomly selected passengers to answer questionnaire face-to-face. Finally, the formal investigation totally recycled 5663 questionnaires and screened out incomplete information ones to get 4820 valid questionnaires.

III. BUILDING MODELS

Logistic regression is one of probabilistic classifiers, whose classification standard is to make posterior probability maximal [5]. What's more, the probabilistic classifiers algorithm can get a more significant result on sample classification problem with a variety of properties.

A. Basic Principle

Assuming that there are N passengers on the railway market, and the nth($n=1,2,3,...,4820$)passenger evaluated the ith attribute X_i $(i=1,2,3,...,6)$ was X_{ni} , and overall evaluation of product is Y_n. Assuming there are c market segments, and the occupancy of each submarket respectively is $\theta_1,\theta_2,...,\theta_c$, meeting $\theta_c \geq 0$ and $\sum_{c=1}^{c}\theta_c = 1$ .So the Logistic regression model about posterior probability $q(y|x)$ after classify is like following:

$$q(y|x;\theta) = \frac{\exp(\sum_{j=1}^{n}\theta_j^{(y)}\phi_j(x))}{\sum_{y'=1}^{c}\exp(\sum_{j=1}^{n}\theta_j^{(y')}\phi_j(x))} \quad (1)$$

B. Model Solution

In order to solute Logistic regression model, the article used maximum likelihood function, which regarded samples $\{(x_i,y_j)\}_{j=1}^{n}, i=1,2,...,6$ as a value about parameter θ, and change the function like following:

$$\begin{aligned} &likelihood \quad \prod_{j=1}^{n} q(y_j|x_i;\theta) \quad \rightarrow \\ &Log \quad likelihood: \quad \sum_{j=1}^{n}\log q(y_j|x_i;\theta) \end{aligned} \quad (2)$$

Therefore, the classification problem can be equivalent to an optimization problem like following:

$$\max_{\theta}\sum_{j=1}^{n}\log q(y_j|x_i;\theta) \quad (3).$$

The parameter θ can be differentiable, so used gradient descent method to solve this maximum likelihood function to get feasible solution $\hat{\theta}$, specific algorithm is as follows:

① giving θ proper initial value, the paper got 0.25,convergence accuracy $\eta = 0.01$;

② putting effective samples from random sampling on survey of Chongqing-Lichuan railway into

$$(x_i,y_j), \quad i=1,2,3,...,6, \quad j=1,2,3,...,4820 \quad (4)$$

③ based on the rising direction of gradient, updating parameter $\theta = (\theta^{(1)^T}, ..., \theta^{(c)^T})^T, c = 1,2,3,4$ as follows

$$\theta^{(y)} \leftarrow \theta^{(y)} + \varepsilon \nabla_y J_j(\theta), \quad y = 1,2,3,4 \tag{5}$$

Here, ε is gradient variation, the paper got 0.001. $\nabla_y J_i(\theta)$ is gradient rising direction of Log likelihood function $J_j(\theta) = \log q(y_j \mid x_j; \theta)$ about parameter $\theta^{(y)}$.

$$\nabla_y J_j(\theta) = -\frac{\exp(\theta^{(y)^T} \phi(x_i))\phi(x_i)}{\sum_{y'=1}^{c} \exp(\theta^{(y')^T} \phi(x_i))} + \begin{cases} \phi(x_i) & (y = y_j) \\ 0 & (y \neq y_j) \end{cases} \tag{6}$$

④return②,until solution $\Delta\hat{\theta} \leq \eta$ ending algorithm.

C. *Model Results*

The paper put data from 4820 effective questionnaires into ***matlab***, used Logistic regression method on market segmentation, and through gradient descent method to solve model got the figure 1 of algorithm convergence and the figure 2 of market segments as follows:

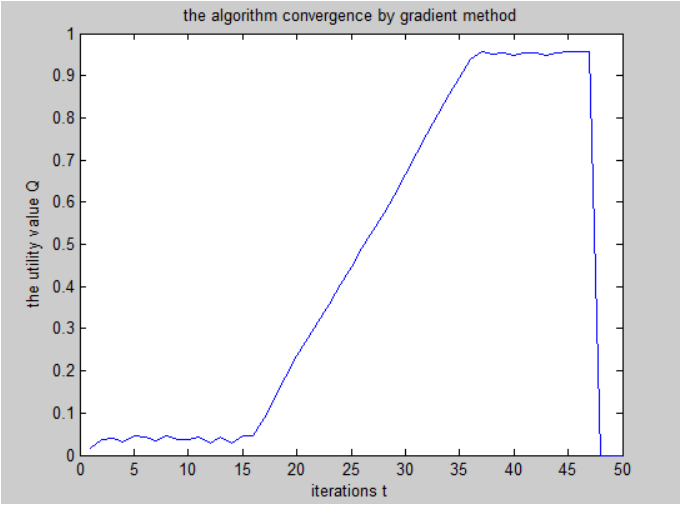

Fig. 4.The Algorithm convergence of model

The figure 1 shows that：In the initial stage of evolutionary, the solution search area is increased by reducing the choice of the elitist until algorithm iterated to 15 times. After that time, the model found feasible gradient direction and went on the convergence direction to get bigger value of utility function. When the algorithm iterated to 37 times, model found stationary solution and when it iterated to 46 times, model got final solution and the value of utility function is 0.988. Finally, the stationary solution $\hat{\theta}$ meet the convergence accuracy requirement, and finished the algorithm to make utility value zero.

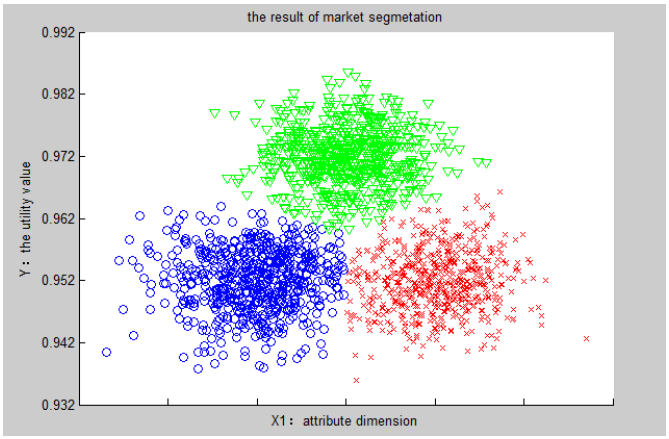

Fig. 5.Market segments by attribute dimension

The figure 2 shows that：Logistic regression successfully and observably divided samples that own 4820 questionnaires into 3 categories, and most samples were distributed in interval [0.932, 0.992], which reflected classification effect is very well. Because there were 6 attribute parameters in the model and the solution is a multidimensional vector space, this paper selected a proper dimension and went on the direction to get projecting vectors in two dimensional plane. From above diagram, the result of passenger market segmentation was well.

IV. MARKET SEGMENTATION

Extracting the data and count samples of each category, this article selected representative samples through searching category-centers to contain 55% samples, as shown below:

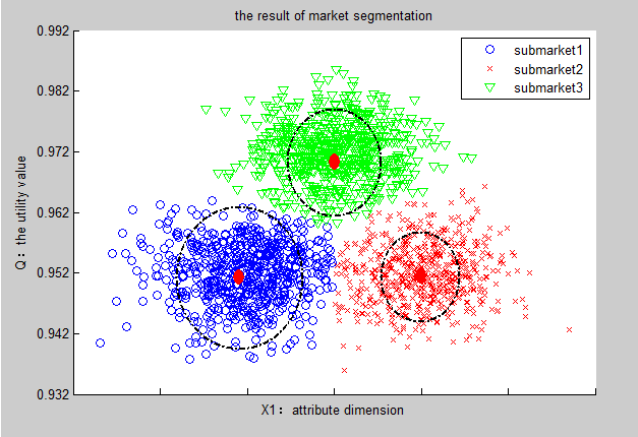

Fig. 6.Market segments by representative samples

The figure 3 shows that：To stat and analysis samples whose utility values were distributed in section [0.932, 0.992], we got sample size of these 3 submarkets. On the basis of age, month income and profession these 3 passenger characteristic indexes, the paper statistical analysis 3 submarkets: migrant workers (submarket 1), non-economic travel (submarket 2), and business travel (submarket 3), as shown below:

TABLE I. STATISTICAL RESULT OF MARKET SEGMENTATION

Segments / *Index*	*migrant workers*	*non-economic travel*	*business travel*
Capacoty	submarket 1 47.4% （1002）	submarket 2 23.2% （490）	submarket 3 29.4% （624）

(Percentages are market share, and numbers in bracket are sample sizes)

The Table Ⅰ shows that：submarket 1 (migrant workers) occupied 47.4% of the total market, and the market share of submarket 2 and submarket 3 were separately 23.2% and 29.4%.

What's more, we made statistical analysis in accordance with age, month income and profession these 3 passenger characteristic index, and got each submarket segment separately. Finally, this article gained economical and practical type, nucleus type, travel quality type these 3 subgroups. So establishing subdivision models of submarkets, as shown in the following Table Ⅱ:

TABLE II. PASSENGER CAPACITY STATISTICS IN SUBMARKETS

Passenger characteristics / *Trip Characteristics*	*Migrant workers*	*Non-economic travel*	*Business travel*
Economical and practical	submarket 11 32.1%（679）	Submarket 12 5.9% （125）	Submarket 13 2.7% （57）
Nucleus	Submarket 21 10.9%（231）	Submarket 22 20.0%（423）	Submarket 23 7.0% （148）
Travel quality	Submarket 31 2.1% （44）	Submarket 32 3.6% （76）	Submarket 33 15.7% （332）

(Percentages are market share, and numbers in bracket are sample sizes)

In order to facilitate the design of transportation products in Chongqing-Lichuan High-speed Railway, we needed to merge submarkets by certain rules, which made the results of market segmentation more practical and required. Among them, the merger rules include the following two aspects: ① the consolidation of obviously small submarkets; ② the consolidation of similar submarkets. So market segmentation after consolidation as shown below:

TABLE III. STATISTICAL RESULT OF MARKET SEGMENTATION AFTER CONSOLIDATION

Passenger characteristics / *trip characteristics*	*migrant workers*	*non-economic travel*	*business travel*
economical and practical	submarket A 43%（910）	Submarket B 8.6%（182）	
nucleus		Submarket C 27.0%（571）	
travel quality	Submarket D 5.7%（120）		Submarket E 15.7%（332）

Thus it can be seen that the result of market segmentation owned A to E, totally 5 submarkets, and each submarket has significant characteristics.

TABLE IV. CHARACTERISTICS DESCRIPTION OF SUBMARKET AFTER CONSOLIDATION

Submarket	**Market description**
A	Income: middle or lower level; Crowd: students or Middle aged Trip frequency: lower
B	Total amount: very small Crowd: Adults Income: middle level
C	Crowd: Adults Income: middle level Trip purpose: tourism or visiting relatives

D	Crowd: Adults Income: high level Trip purpose: tourism or business
E	Total amount: very small Crowd: Middle or old aged Income: middle level

V. CONCLUSIONS

For the analysis of passenger market segmentation feature, the following conclusions could be drawn：

(1)Logistic regression classification method is utilized to classify random sample data, and stationary solution can be obtained in a few iterations. The two-dimensional figure obtained by the projection of trip purpose dimension indicated that the effect of this classification method is of signification.

(2)Through the integration of the subordinate markets, in each subordinate guest market, the fact could be found that the age, monthly income and travel purpose of travelers are to be the most significant indicators of high-speed rail passenger market segmentation and market characteristics description.

(3)According to the analysis results, how to forecast the needs of all market segments, and thus how to develop scientific strategies of product pricing and seat control to achieve maximize revenue of high-speed rail operation, would be the future research directions.

REFERENCES

[1] Tony Lun, *Third Revised and Enlarged Edition*, Elsevier Science Pulishers, B.V.1986, pp. 387-423.

[2] Tsai A, "Purchase-based market segmentation methodology," E*xpert Systems with Applications*, 2004, vol. 2, pp. 265-266.

[3] J.Zhao, M.Ren, "Product differentiation and market segmentation as alternative marketing strategies," *Railway Economics Research*, 2014, vol. 6, pp. 13-17.

[4] B.Y.Qian, B.Shuai, C.S.Chen, "Study on Subdivision of DPL Passenger Market based on Mixed Regression Model," *Railway transport and economy*, 2014, vol. 1, pp. 60-65.

[5] Titter ington D M, *Statistical analysis of finite mixture distributions*, New York, Institute of Philosophy, 2005.