A Forecasting Model of the Proportion of Peak Period Boardings for Urban Mass Transit System: A Case Study of Osaka Prefecture

Yan Cheng, Corresponding Author
Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University
4800 Cao'an Road, Jiading District, Shanghai, P.R. China, Post Code 201804
Tel: (86)021-69583757; Email: patty_1234@126.com

Xiafei Ye
Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University
4800 Cao'an Road, Jiading District, Shanghai, P.R. China, Post Code 201804
Tel: (86)021-69589875 Fax: (86)021-69583712; Email: yxf@tongji.edu.cn

Zhi Wang
Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University
4800 Cao'an Road, Jiading District, Shanghai, P.R. China, Post Code 201804
Tel: (86)021-69585031; Email: zhiwang@tongji.edu.cn

Word Count:

Abstract 193 words
Text 2679 words + 12 tables/figures x 250 words (each) = 5679 words

Number of References 16
Number of Figures 8
Number of Tables 4

Submission Date: 8/1/2015

ABSTRACT

At the planning and design phase of urban mass transit system, the aim is to grasp the features of spatial and temporal distributions of passenger flow during peak period. For this goal, dynamic passenger assignment model should be applied. An indispensable input parameter of this model is time-varying interstation OD matrix of peak period. To gaining this parameter, the first step is figuring out peak period boardings (excluding interchanges, PPB). Since all-day boardings can be extracted from the given all-day OD matrix, the study focuses on forecasting the proportion of PPB. Taking Osaka Prefecture as research area, this article firstly proposes a new concept of station catchment area as the border of data collection, which can be determined by two types of risks. With the help of Spearman correlation analysis, 3 factors prove to be significantly associated with the proportion of PPB. Then two regression models are conducted with socio-economic and land-use characteristics as independent variables respectively. Results show that the model with the proportion of resident population as independent variable has a better performance, of which the adjusted R^2 reaches 0.951 and the standard error of verification data is only 7.8 percent.

KEY WORDS

Urban Mass Transit, Peak Period, Proportion of Boardings, Two Types of Risks, Regression Model

INTRODUCTION

As the backbone of urban transportation, mass transit system is playing a more important role. Owing to its high construction cost and big difficulty to modification after construction, predictability and accuracy become the major aims of its planning and design. Maximum passenger flow of one-direction section in peak hour (MPSP) is an essential guidance of these complicated works. Combining the inputted time-varying interstation OD matrix of peak period with the route and departure time decisions of passengers, dynamic passenger assignment model can give the answer of MPSP. However, time-varying interstation OD matrix of peak period has been rarely studied. The reason for this situation is that most of existing studies on dynamic passenger assignment model for urban mass transit system assumed that time-varying interstation OD matrix was given (*1*-*3*) or could be estimated by using actual automated fare data (*4*). But the assume is not tenable at the planning and design stage, because the actual operational data hasn't been generated. The only available data is the estimated all-day interstation OD matrix.

To obtain time-varying interstation OD matrix of peak period, all-day boardings (excluding interchanges, similarly hereinafter) based on all-day interstation OD matrix are the first data wanted. According to socio-economic and land-use situation, the proportion of peak period boardings (PPB) can be determined and used to forecast the PPB. With passengers' specified arrival time of destination, every one's departure time become available eventually.

Some researchers have pay attention to PPB. Wang et al (*5*) analyzed the fluctuations of boardings within a day of Beijing Metro stations, but characteristics of different kinds of stations haven't been analyzed. Based on this research, Deng (*6*) categorized stations into 8 types according to the land-use around stations, and investigated floating ranges of the proportion of PPB of each type. Results showed that stations with mainly residential land around have a higher proportion than others. Nevertheless, there wasn't any quantitative analysis on influence factors and their influence degree of the proportion of PPB. This article fills the gap, with 370 stations of Osaka Prefecture as study objects. Furthermore, regression models are established to forecast the proportion of PPB.

STUDY AREA

Osaka Prefecture is the core section of Keihanshin Metropolitan Area, the second largest metropolitan area of Japan. According to the latest census in 2010, it had 8.87 million residents on an area of 1898.47km^2. The population density and employment density of Osaka Prefecture at the census tract level are shown in Figures 1 and 2. The 10km-radius circle around Osaka station has higher population and employment densities than other place. That is the circle where Osaka City is located.

This area is supported by a rail system consisting 52 lines of 14 companies, the total length of which reaches 743.45 km. 370 stations are selected from the system as study objects, each of which has complete data.

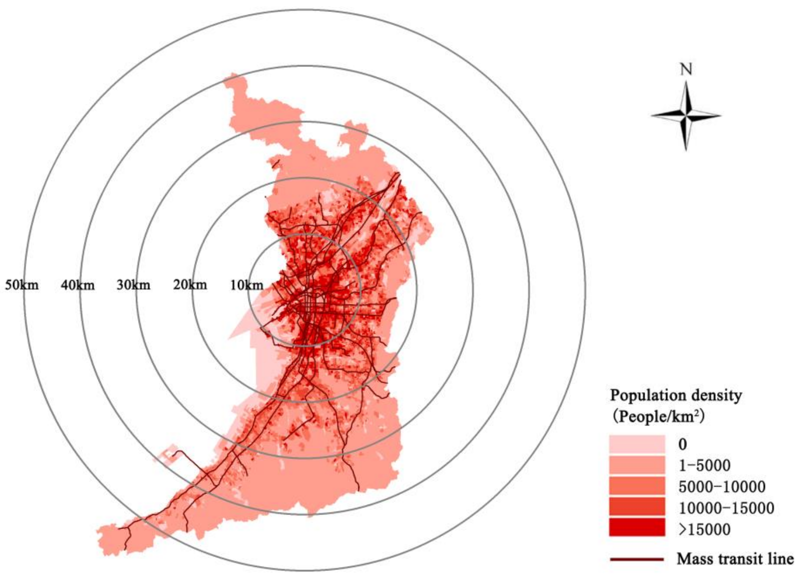


FIGURE 1 Population Density of Osaka Prefecture at Census Tract Level.

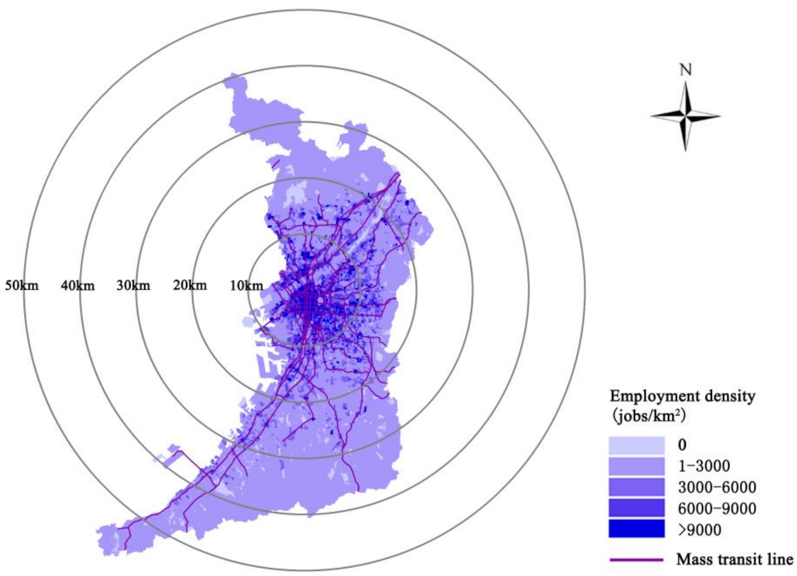


FIGURE 2 Employment Density of Osaka Prefecture at Census Tract Level.

DATA RESOURCE

Three types of data are needed in this research.

(1) Different period boardings of study objects in one day

This data is an announced result of the 5th Keihanshin Metropolitan Area Personal Travel Survey in 2010 (*7*).

(2) Actual catchment area of study objects

This data can be investigated from the results of the 11th Japan Metropolis Transportation Census in 2010 (*8*).

(3) Socio-economic and land-use data around study objects

This data which is in the format of GIS file can be downloaded from the official website of Information Policy Division, Policy Bureau, Ministry of Land, Infrastructure, Transport and Tourism (*9*).

PEAK PERIOD

Urban mass transit system has the advantages of fast speed, punctuality and large capacity. Thanks to these, it undertakes a lot of long- and moderate-distance trips, aims of which are mainly commuting and going to school. This type of passengers has rigid travel demand. Their specified arrival time of destination are fixed and relatively close, so that the peak period of urban mass transit system is usually the morning peak period. The latest Urban Traffic Annual Report indicated that the peak hour of 29 lines which are surveyed within Osaka Prefecture all appeared from 7 a.m. to 9 a.m. on weekdays, as shown in Figure 3. Therefore, morning peak period is confirmed as the time range of the study.

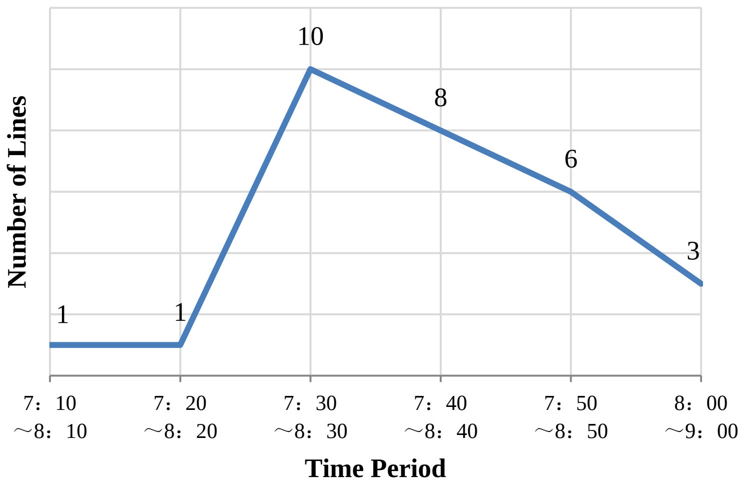


FIGURE 3 Number of Lines with Different Peak Hour within Osaka Prefecture.

Based on different period boardings of stations in Osaka Prefecture (Figure 4), it isn’t difficult to find that from 6 a.m., stations have gradually entered into the peak hour of boardings. This is mainly because passengers living in suburbs need more time to commute to downtown area. So the boarding time of them is earlier than the actual peak hour of lines. On the other hand, the work start time varies with the kind of job. For example, people who working at shopping mall and places of entertainment often begin to work at 10 a.m. That’s why the boardings between 9 a.m. to 10 a.m. is still larger than the amount of off-peak period. For the above reasons, the peak period of this study is from 6 a.m. to 10 a.m.

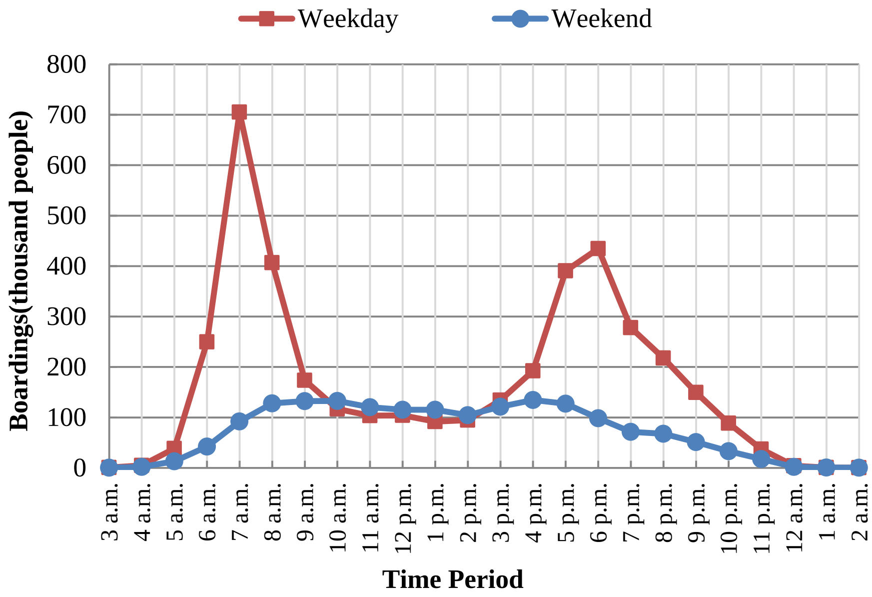


FIGURE 4 Different Period Boardings of Station within Osaka Prefecture.

REASONABLE CATCHMENT AREA (RCA)

Station catchment area is regarded as the collection boundary of all kinds of basic data. Thus, whether it is proper have a direct impact on the ability to describe reality and predict the future of the model.

It was assumed by the majority of station-level ridership researches that station catchment area was equivalent to pedestrian catchment area (PCA), which is a circle with the station as the center and a distance threshold as the radius. All stations of each study shared an identical distance threshold, ranging from 500m to 2000m (*10-16*). This practice lost sight of the effect of station density.

Few scholars explained how to justify the distance threshold for delimiting catchment area (*12*, *15*). They prepared several pre-determined distance, collected data within each catchment area and substituted the data into correlation analysis or forecasting models. By comparing results, the distance threshold was decided. Within this catchment area, correlation or the goodness of fit achieved the maximum. This method means that the distance threshold varies with independent variables, that is to say, this method establishes a correlation between them. But in fact, station's catchment area is only related to station characteristics, nothing to do with other factors (e.g.: population, land-use around stations).

It has been a desire that catchment area can be as close as possible to the actual one, the union area of all boardings' origins. However, it is unrealistic in the forecast period. This article defines the union area of 90 percent boardings' origins as station RCA, which guarantees the accuracy of model, and meanwhile avoids the meaningless expansion of catchment area due to few passengers. Though its forming method is the same as the previous research, a circle with the station as the center, its distance threshold is justified by a new method, which is called two types of risks.

Two Types of Risks

The choice of distance threshold is critical to the forming of RCA. A very low distance threshold creates very small areas, which leave out most of riders. And on the other, with a very high distance threshold, some faraway areas having no contributions to boardings may be contained into it. Either case can distort the final results of subsequent forecast. For this reason, two types of errors should be observed during the process of comparing and analyzing alternative catchment areas.

The first types of error refers to the one that alternative catchment area hasn't included the area which is a part of actual catchment area. The probability of this error is called Risk I, expressed as follows:

$$P(M_1) = \frac{A - (A \cap A')}{A} \tag{1}$$

The second types of error refers to the one that alternative catchment area includes the area which isn't a part of actual catchment area. The probability of this error is called Risk II, expressed as follows:

$$P(M_2) = \frac{A' - (A \cap A')}{A} \tag{2}$$

where *A* is the actual catchment area of station, and *A'* is the alternative catchment area of station. Since both catchment areas of study objects are divided into census tracts, the calculation of two types of risks replaces actual area by the number of tracts.

Both of risks of RCA are hoped to reach the minimum simultaneously, but it is hard to realize. The study comes up with a new idea of determining RCA, selecting the catchment area with the minimum sum of two types of risks, based on the premise that it covers no less than 90 percent of boardings.

The Distance Threshold of RCAs under Different Station Densities

The distance between origin and station is the primary issue of passengers when they are choosing which station to go. If the OD matrix unchanged, with the increase of station density in a certain area, there will be more accessible stations for passengers; correspondingly, for a station, the union area of the origins will shrink. Forasmuch, station RCAs are not uniform, but have a link with the station density of its location.

In order to understand how the station density affects the distance threshold, several 5km-wide bands are created around Osaka station, up to a maximum limit of 55km. 370 research objects are grouped by the station density of the bands which they are located in. From 0 to 1.00 stations/km^2, groups are set at intervals of 0.10 stations/km^2, and the remaining is classified as a group (if a group is void, it will be merged into the first nonempty group whose station density is less than it). Each station has 15 alternative catchment areas, the radius of which varying from 100 to 1500m at intervals of 100m. For one group, its values of two types of risks equal the average value of all stations in it.

TABLE 1 The Distance Threshold of RCAs under Different Station Densities

Group	Band	Station Density (stations/km^2)	Number of Station	Distance Threshold(m)	Risk I (percent)	Risk II (perent)
>1.00	0-5km	1.13	84	900	27.5	180.9
0.50-1.00	5-10km	0.58	102	900	26.2	116.0
0.30-0.50	10-15km	0.30	68	900	35.7	40.5
0.20-0.30	15-20km	0.25	51	1000	32.2	73.5
0.10-0.20	20-25km	0.12	28	1000	29.6	67.8
	25-30km	0.14	14			
	30-35km	0.11	8			
0-0.10	35-40km	0.09	6	1000	34.6	33.0
	40-45km	0.07	4			
	45-50km	0.05	2			
	50-55km	0.03	3			

Table 1 shows the RCA of every group as well as their risks. The distance threshold of RCA decreases with the increase of station densities. It is 1000m when the station density is less than 0.30 stations/km^2 and becomes 900m when the station density is greater or equal to 0.30 stations/km^2. Data collection will take the RCAs under different station densities as the border.

FORECASTING MODEL OF THE PROPORTION OF PPB

Factors Influencing the Proportion of PPB

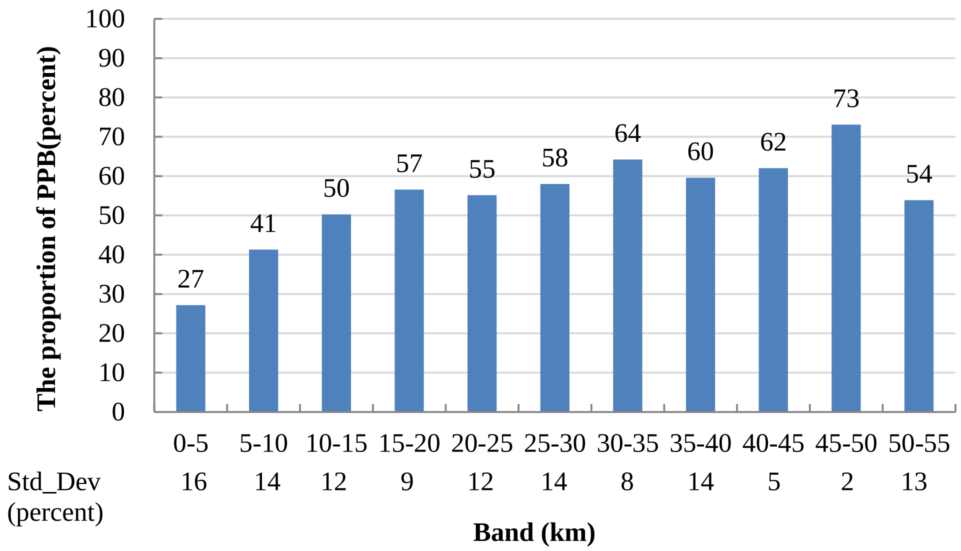


FIGURE 5 Proportion of PPB of Study Objects.

From the center of Osaka Prefecture to its periphery, the proportion of PPB has a tendency to rise, as shown in Figure 5. But the difference between stations within the same band cannot be ignored.

The proportion of PPB is a relative concept. It is not only related to the boardings during peak period, but also related to the ones of all-day. In view of the fact that there is no difference in station characteristics and its intermodal connection between passengers who enter the same station during peak period and off-peak period, it is considered that the influence of these two factors on the proportion of PPB can be neglected.

Investigations of the station boardings data of Osaka Prefecture show that the major trip purposes of passengers who boarding during peak period are going to work (60 percent) and school (17 percent). Other purposes (private matter, business, going home, etc.) account for less than 10 percent each. It is indicated that residents are the main source of PPB. But it doesn’t mean the more the residents are, the higher the proportion of PPB becomes. If the jobs within the station RCA are enough to meet the requirements of residents, then they don’t need to take mass transit to farther places. So the jobs and other land use within the station RCA have an impact on PPB. These two factors also influence the boardings of all-day, because the boardings of all-day consist of three parts. The first part is the morning peak period boardings, what this study focuses on. The second part is the off-peak boardings of people living around the station with flexible travel demand. And the last part is the evening peak period boardings of people working around the station, who travelling back to home.

In summary, two types of factors, socio-economic and land-use characteristics, are selected to be candidates. In order to eliminate the effect of absolute value as possible, factors are reflected by relative value. The socio-economic characteristic is quantified by the proportion of resident population within station RCA (the denominator is the sum of resident population and jobs), while the land-use characteristics are quantified by the proportion of residential, commercial and industrial floor area within station RCA (the denominator is the gross floor area).

TABLE 2 Bivariate Correlations between Factors and the Proportion of PPB

Variables	The proportion of PPB	The proportion of resident population	The proportion of residential floor area	The proportion of commercial floor area	The proportion of industrial floor area
Asymptotic significance of K-S test	0.008	0.000	0.001	0.000	0.000
Correlation coefficient	1.000	0.741**	0.655**	-0.635**	0.034

*Notes: ** Correlation is significant at the 0.01 level (2-tailed).*

As shown in Table 2, the result of Kolmogorov–Smirnov test shows asymptotic significances of 5 distributions are all less than 0.050, which means none of variables follows a normal distribution. Bivariate correlations among the proportion of PPB and factors are calculated through Spearman correlation analysis. It is proved that except the proportion of industrial floor area, correlation coefficients between the proportion of PPB and the other three factors (and their significance) tend to be high. The proportion of resident population and residential floor area is positive correlated and the proportion of commercial floor area is negative correlated, which is consistent with the previous qualitative analysis.

To some extent, land-use can reflect socio-economic situation. Correlation analysis verifies this point of view. Correlation coefficient between the proportion of resident population and the proportion of residential floor area and commercial floor area are both high, equaling 0.854 and -0.733 respectively. Hence, two regression models with socio-economic characteristic and land-use characteristics as independent variables are established and compared.

Regression Analysis

Figures 6 and 7 are XY scatter diagrams, used to explore whether the relationship among dependent variable and independent variable(s) is linear or not. It is clear that all the relationships are linear.

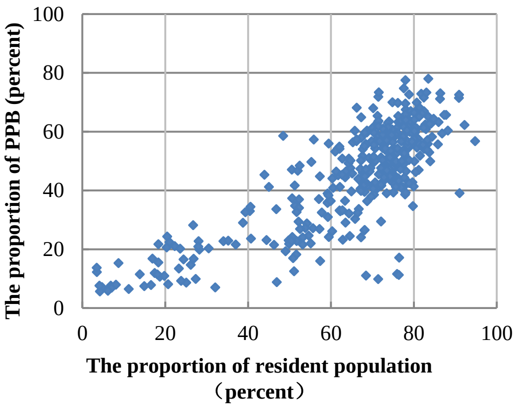


FIGURE 6 Scatter Diagram of the Socio-economic Characteristic and the Proportion of PPB.

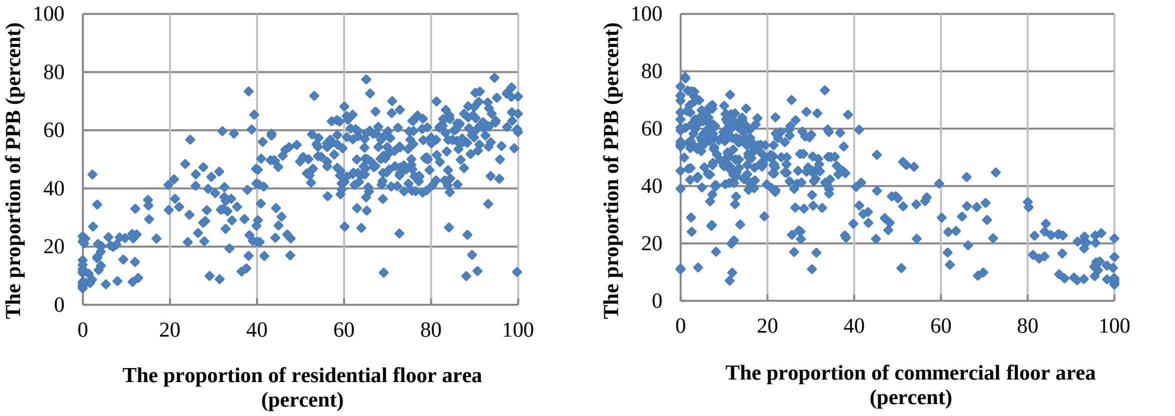


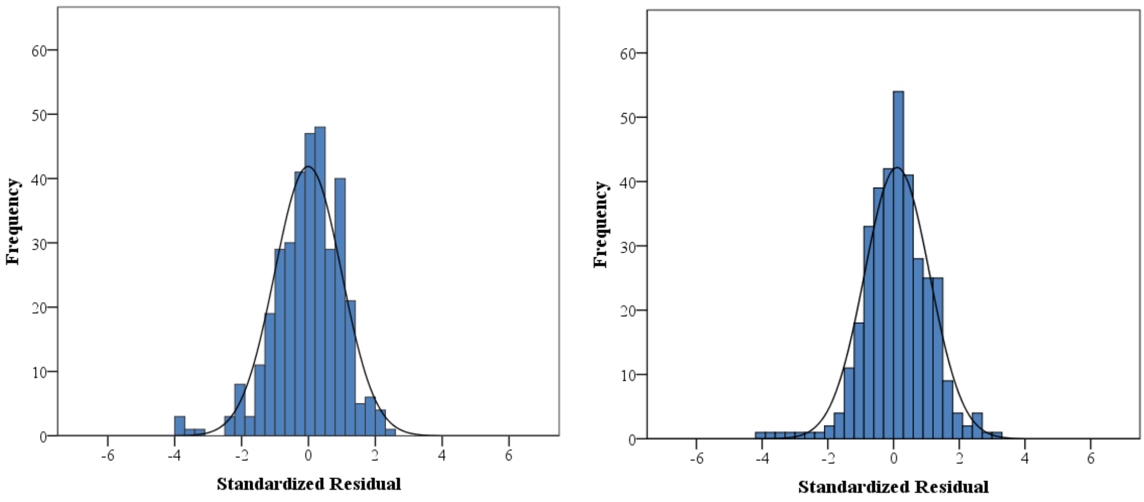
FIGURE 7 Scatter Diagram of the Land-use Characteristics and the Proportion of PPB.

370 stations are divided into two groups. Data of 350 stations is used for calibrating regression coefficients, and the rest 20 stations' data is used for validation. Intercept of regression models is equal to the value of dependent variable when all independents are zero. The variables of this study have definite meanings. The proportion of PPB is regarded to be zero when the proportion of resident population equals zero or the proportions of residential and commercial floor area equal zero at the same time. As a result, there is no intercept in the following two models. Table 3 presents results of two models. Regression model I has an R^2 of 0.975 (adjusted R^2=0.951), and an F-statistic value of 6798.60, significant at 0.000 level. Regression model II has a slightly inferior performance, of which adjusted R^2 is equivalent to 0.917, and an F-statistic value of 1922.22. Figure 8 shows standardized residual of two regression models.

TABLE 3 Two Regression Models. Dependent Variable: the Proportion of PPB

Model	Variable(s)	Coefficients	Std. error	t-statistic	Sign.	VIF
Regression model I	The proportion of resident population	0.694	0.008	82.454	0.000	NA
Regression model II	The proportion of residential floor area	0.668	0.012	54.017	0.000	1.163
	The proportion of commercial floor area	0.159	0.020	8.019	0.000	1.163

Notes: NA= not available



(*a*) Regression model I (*b*) Regression model II

FIGURE 8 Standardized Residual of Two Regression Models

TABLE 4 Validations of Two Regression Models

No.	Actual proportion of PPB (percent)	Error of regression model I (percent)	Error of regression model II (percent)
1	22.89	1.56	-2.78
2	36.42	7.59	-14.34
3	32.11	12.56	-5.36
4	50.01	1.40	-11.66
5	51.22	0.54	-0.25
6	23.11	7.75	-3.59
7	39.80	6.87	-19.28
8	46.43	-3.90	-16.82
9	40.77	8.54	11.79
10	65.32	-15.89	-33.98
11	60.43	-6.06	-6.23
12	55.06	2.18	6.10
13	65.31	-12.42	-21.88
14	56.71	-8.26	-0.55
15	61.32	-4.16	1.53
16	52.81	-0.13	-3.74
17	59.53	-6.73	1.58
18	54.01	-7.04	-8.14
19	60.34	-12.68	-17.41
20	62.22	1.79	-2.35
	Standard Error	7.8	12.8

Two models are validated by inputting the remaining 20 stations' data. The standard error of regression model I is only 7.8 percent, 5 percent less than regression model II, as shown in Table 4. Because of its better performance and higher predictive power, regression model I becomes the best choice of forecasting model of the proportion of PPB, which is expressed as follows:

$$PPB\% = 0.694 \times RP\% \quad (3)$$

where PPB% is the proportion of PPB and RP% is the proportion of resident population.

CONCLUSIONS

Time-varying interstation OD matrix of peak period has been rarely studied. With a new method proposed, this article puts emphasis on the proportion of PPB. Station RCAs are justified by two types of risks, indicating that the distance threshold of RCA decreases with the increase of station densities. To investigate the factors, socio-economic and land-use characteristics are expressed in relative form and analyze the correlation among the proportion of PPB and alternative independent variables. Result shows the proportion of resident population, residential floor area and commercial floor area are influence factors. Since the first factor is significantly correlated with the latter two, regression models with socio-economic characteristic and land-use characteristics as independent variables respectively are established and compared. With the help of regression analysis and validation, it is confirmed that regression model I with the proportion of resident population as independent variable has a better performance. It is in a simple mathematical form and the required data isn't very difficult to collect. It is convinced that this model will contribute to the forecasting of time-varying interstation OD matrix of peak period in the future.

REFERENCES

1. Poon, M.H., S.C. Wong, and C.O. Tong. A Dynamic Schedule-based Model for Congested Transit Networks. *Transportation Research Part B*, Vol. 38, 2004, pp. 343-368.
2. Liu, X.H., Research on the Dynamic Flow Assignment Model Based on Train Schedule for Urban Subway Network. Chang'an University, Xi'an, China, 2013.
3. Yang, D.C., Research on Schedule-based Rail Transit Passenger Flow Assignment. Beijing Jiaotong University, Beijing, China, 2013.
4. Chan, J., Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data. Massachusetts Institute of Technology, Cambridge, MA, 2007.
5. Wang, J., J.F. Liu, and Y.L. Ma. Temporal and Spatial Passenger Flow Distribution Characteristics at Rail Transit Stations in Beijing. *Urban Transport of China*, Vol. 11, 2013, pp. 18-27.
6. Deng, J., Interactive Relationship between Ridership Characteristics at Subway Station and Surrounding Land-use: A Case Study in Beijing. Beijing Jiaotong University, Beijing, China, 2015.
7. Results of the 5th Keihanshin Metropolitan Area Personal Travel Survey. Transportation Planning Committee of Keihanshin Metropolitan Area. www.kkr.mlit.go.jp/plan/pt/data/pt_h22/index.html. Accessed May 11, 2014.
8. Results of the 11th Japan Metropolis Transportation Census. Ministry of Land, Infrastructure, Transport and Tourism, Japan. www.mlit.go.jp/sogoseisaku/transport/sosei_transport_tk_000035.html. Accessed March 5, 2014.
9. GIS Data of Japan. Ministry of Land, Infrastructure, Transport and Tourism, Japan. nlftp.mlit.go.jp/ksj/jpgis/jpgis_datalist.html. Accessed Jan. 10, 2015.
10. Kuby, M., A. Barranda, and C. Upchruch. Factors Influencing Light-rail Station Boardings in the United States. *Transportation Research Part A*, Vol. 38, 2004, pp. 223-247.
11. Sohn, K., and H. Shim. Factors Generating Boardings at Metro Stations in the Seoul Metropolitan Area. *Cites*, Vol. 27, 2010, pp. 358-368.

12. Gutierrez, J., O.D. Cardozo, and J.C. Garcia-Palomares. Transit Ridership Forecasting at Station Level: An Approach Based on Distance-decay Weighted Regression. *Journal of Transport Geography*, Vol. 19, 2011, pp. 1081-1092.
13. Choi, J., Y.J. Lee, and T. Kim. An Analysis of Metro Ridership at the Station-to-station Level in Seoul. *Transportation*, Vol. 39, 2012, pp. 705-722.
14. Zhao, J., W. Deng, and Y. Song. Analysis of Metro Ridership at Station Level and Station-to-station level in Nanjing: An Approach Based on Direct Demand Models. *Transportation*, Vol. 41, 2014, pp.133-155.
15. Sung, H., K. Choi, and S. Lee. Exploring the Impacts of Land Use by Service Coverage and Station-level Accessibility on Rail Transit Ridership. *Journal of Transport Geography*, Vol. 36, 2014, pp. 134-140.
16. Zhou, J.Z., and D.Y. Zhang. Direct Ridership Forecast Model of Urban Rail Transit Stations Based on Spatial Weighted LS-SVM. *Journal of The China Railway Society*, Vol. 36, 2014, pp. 1-7.