Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090301
ISSN 2371-9524 Vol. 9 Num. 3

Prediction of Taxi Quantity in Hangzhou Based on Principal Component Regression Analysis

Xinmei Wang

Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, 201804, China
xinmeiwang@126.com

Keywords: The Number of Taxis, Principal Component, Regression Prediction Model

Abstract: The congestion of a large city is closely related to the growth of the number of taxis. As an important part of urban traffic, it is particularly important to control traffic congestion to master the number of taxis. In order to assess the impact of taxis on the traffic flow in the future, a scientific method is needed to predict the number of taxis efficiently and accurately. Through qualitative analysis and quantitative correlation analysis, eight influential factors with high correlation were extracted as predictors. The multicollinearity among influencing factors was eliminated by principal component analysis. Based on the single regression prediction of principal component, the regression prediction model of taxi number is constructed. The accuracy test results show that the model has high precision and can be used to predict the number of taxis quickly. It can provide an important basis for urban traffic management departments to effectively control congestion and make accurate traffic decisions.

1. Introduction

The urban passenger transport system is an organic whole composed of various modes of transport, which jointly undertake various passenger transport services in the city. There are both competitive and complementary relations between various modes of transport in the passenger transport market, and they must be maintained in a certain proportion to ensure the stable development of urban passenger transport system. As an important part of urban passenger transportation, the proportion of taxi will inevitably affect the share rate of other transportation modes, and after the imbalance of urban bus supply and demand, it will inevitably bear part of the shortage of other transportation modes. In addition, too much taxi capacity will lead to a higher empty rate of taxi capacity waste, while too little will lead to insufficient demand and supply of urban passenger transport. Therefore, reasonable taxi capacity is of great significance for optimizing the composition of passenger transportation system and alleviating the contradiction between passenger capacity and public transport demand. Therefore, how to predict the number of taxis scientifically is an important decision-making basis for urban passenger transport management units to guide the harmonious development of urban taxi industry.

At present, researches on the number of taxis mainly focus on the estimation method of taxi input [1-3], while researches on the prediction method of taxi number are few. If the estimation method of taxi number is directly applied to forecast, multiple influencing factors need to be predicted and then the number of taxis is calculated, which will lead to the superposition of prediction errors. In literature [4], a multiple linear regression model was built to predict the taxi input volume by using the influencing factors of the taxi input volume, but the multiple linear relationships among the influencing factors were not investigated, and the extraction of the influencing factors was not thoroughly studied.

In this paper, the statistical analysis method is used to select and quantitatively analyze the influencing factors of taxi number, and the principal component analysis method is used to diagnose and eliminate the multicollinearity among the influencing factors, so as to build the principal component regression model of taxi number for the direct quantitative prediction of taxi number.

2. Extraction of influencing factors

2.1. Qualitative analysis of influencing factors

Taxi transport capacity is mainly based on the level of economic development and the actual needs of society, taking into account the social and economic benefits, and the moderate development of taxi is controlled. The main factors affecting the demand for rental cars include:

(1) Social and economic development level

(2) With the development of social economy and the improvement of people's living standards, the population flow is relatively frequent, which promotes the need for more friendship and socializing. The transportation needs associated with this will also change in quantity and quality as living standards improve. In addition, the multi-level diversity of travel needs is becoming increasingly prominent, people are no longer satisfied with simply achieving spatial displacement, but more emphasis on door-to-door to "convenient, punctual, safe, comfortable" as the standard transportation needs, taxi is the standard mode of transportation.

(3) City scale and urbanization degree

Different urban scale and urban land area and conditions determine the scale and distribution of different urban road networks. The expansion of urban scale leads to the expansion of urban space, so that the travel distance of residents increases linearly, and people's transportation demand also increases correspondingly, so it is necessary to improve the transportation speed. However, the low running speed of the existing buses cannot meet people's "punctual" requirements, forcing some people to turn to taxis, making the demand for taxis grow. The development of transportation leads to the improvement of urbanization level, and the improvement of urbanization level will in turn promote the increase of traffic demand, which leads to the corresponding increase in the demand for rental cars.

(4) Urban transportation infrastructure

Urban road traffic network, trunk road, number of parking lots and transportation hubs have a direct impact on rental cars. First, the higher the density of urban road network, the more main roads, the larger the road area, the faster the running speed of taxi, the expansion of service, the demand for taxi increases; Second, there are many winding streets and alleys in the city, which are difficult to reach by bus. Under certain conditions, these residents are likely to choose taxi as a means of transportation; Third, the layout of the transportation hub is relatively uncentralized, and the transfer is inconvenient, and the taxi may become the preferred way of transfer.

(5) The increase of urban population

With the improvement of the production level of human society and the development of the scientific and technological revolution, there has been an irreversible process of relative reduction of the rural population, relative increase of the urban population, continuous increase of the number of cities, and continuous transfer and concentration of social productive forces to cities, regardless of human will. China is no exception, the total number of non-agricultural population in cities with a population of more than 1 million has accounted for 52% of the country's non-agricultural population. Urban population growth is the most direct reason for the growth of urban traffic demand.

(6) Other factors

The government's policy system on the management of taxi, the rental price level, the management mode of taxi, the development of the tertiary industry and the influence of other modes of transportation (such as bus rapid transit) have a certain impact on the demand for taxi.

Taking the above factors into consideration, this paper selects quantifiable indicators to build a system of taxi influencing factors (Table.1)

Table 1: System of influencing factors of taxi number

Category	Influencing factor	NO
Social and economic development level	Gross domestic product of the city GDP GDP(Hundred million yuan)	R1
	Per capita consumption expenditure (RMB/person)	R2
	Total retail sales of consumer goods (Hundred million yuan)	R3
	Investment in fixed assets (Hundred million yuan)	R4
City scale and urbanization degree	Built-up area(Square kilometer)	R5
	Passenger volume(Tens of thousands of people)	R6
Urban transportation infrastructure	Urban road area(Ten thousand square meters)	R7
The increase of urban population	Permanent population at the end of the year (Thousands of people)	R8

2.2. Quantitative analysis of influencing factors

Table 2: Data table of Influencing Factors of Taxi Quantity (Hangzhou Statistical Yearbook(2005-2013))

Year	Number of taxis (vehicles)	R1 GDP(Hundred million yuan)	R2 (RMB/ person)	R3 (Hundred million yuan)	R4 (Hundred million yuan)	R5(Square kilometer)	R6(Tens of thousands of people)	R7(Ten thousand square meters)	R8 (Thousands of people)
2013	10904	8344	24833	3531	4264	462	36409	5426	884
2012	10344	7802	22800	2945	3723	453	35819	5284	880
2011	10905	7019	22642	2548	3100	433	34778	4900	874
2010	9362	5949	20219	2146	2652	413	33772	4754	871
2009	9305	5088	18594	1805	2195	393	30116	4485	833
2008	9167	4789	16719	1578	1882	367	29084	4258	820
2007	8583	4104	14896	1308	1584	345	28026	4185	807
2006	8398	3444	14472	1119	1373	327	25810	4216	789
2005	8320	2944	13438	978	1278	314	24124	3175	771

In order to analyze and verify the internal correlation between the number of taxis and the above selected influencing factors, this paper uses correlation analysis to test the degree of correlation. The correlation between two variables is measured by the correlation coefficient, and if a variable is correlated with multiple variables, it is generally measured by the partial correlation coefficient. By collecting relevant historical data of Hangzhou City (Table 2), SPSS statistical analysis software is used to conduct Person correlation analysis, and the results are shown in Table 3.

Table 3: Correlation analysis of commuting time consumption and influencing factors

	Y	R1	R2	R3	R4	R5	R6	R7	R8
Y	1.000	0.957	0.967	0.943	0.938	0.944	0.928	0.859	0.909
X1	0.957	1.000	0.989	0.990	0.989	0.990	0.982	0.938	0.964
X2	0.967	0.989	1.000	0.982	0.979	0.992	0.979	0.924	0.970
X3	0.943	0.990	0.982	1.000	0.999	0.973	0.955	0.918	0.933
X4	0.938	0.989	0.979	0.999	1.000	0.973	0.954	0.911	0.931
X5	0.944	0.990	0.992	0.973	0.973	1.000	0.990	0.936	0.983
X6	0.928	0.982	0.979	0.955	0.954	0.990	1.000	0.942	0.996
X7	0.859	0.938	0.924	0.918	0.911	0.936	0.942	1.000	0.934
X8	0.909	0.964	0.970	0.933	0.931	0.983	0.996	0.934	1.000

As can be seen from the correlation analysis results in Table 3, the Person correlation coefficients of the number of taxis (Y) and the selected influencing factors are 0.957, 0.967, 0.943, 0.938, 0.944, 0.928, 0.859, 0.909, respectively, and are significantly correlated at the level of 0.01. Therefore, it can be considered that the selected influencing factors have a high correlation with the number of taxis.

3. Construction of principal component regression forecasting model

3.1. Collinearity diagnosis

The correlation analysis table in Table 3 also shows that there is generally a good correlation among the influencing factors, that is, there may be mutual interference, resulting in the failure of the final regression coefficient to pass the significance test, and even the collinearity of the symbols carried by some regression coefficients does not match the actual economic significance. Therefore, the influence of multicollinearity among various factors must be excluded. The effect of multicollinearity on regression coefficient can be excluded by constructing regression model based on principal component analysis.

In order to verify whether there is collinearity between the simplified influencing Factor variables, the number of taxis is set as the dependent variable Y, and Variance Inflation Factor (VIF) can be found from the regression results using Backward step regression algorithm in SPSS statistical software. The results are shown in Table 4.

Table 4: The VIF value of the independent variable affecting the factor

Independent variable	R1	R2	R3	R4	R5	R6	R7	R8
VIF	896.812	162.684	545.745	453.174	242.212	2866.425	12.088	1895.73

VIF results in Table 4 show that there is collinearity among the independent variables (VIF>10), and the independent variables are not independent of each other, that is, the regression equation obtained by directly using the independent variable cannot correctly explain the meaning of commuting time consumption. Therefore, independent variables must be formed through principal component analysis to eliminate mutual interference.

3.2. Principal component analysis of influencing factors

Principal component analysis was carried out on the selected independent variables, and the statistical information of principal components was obtained, as shown in Table 5.

Table 5: Independent variable principal component statistics

Ingredient	Initial Eigenvalue		
	Sum	Variance %	Accumulate %
1	7.750	96.871	96.871
2	0.136	1.699	98.570
3	0.088	1.106	99.676
4	0.013	0.168	99.844
5	0.007	0.087	99.931
6	0.004	0.055	99.985
7	0.001	0.012	99.998
8	0.000	0.002	100.000

The above table shows that the eigenvalue of the first principal component is 7.750, which explains 96.871% of the total variance of the eight original variables (it is generally believed that effective information can be retained when the cumulative contribution rate of the principal component reaches 85% [5]). The eigenvalue of the second principal component is 0.136, much less than 1, and it explains 1.699% of the total variance of the seven primary variables. Therefore, the first principal component is selected to replace the original 8 original variables, and the score coefficients of principal components are shown in Table 6.

Table 6: Independent variable principal component coefficient

Independent variable	R1	R2	R3	R4	R5	R6	R7	R8
Principal component coefficient	0.129	0.128	0.127	0.127	0.128	0.128	0.123	0.126

Therefore, the principal component expression is:

$$Z = 0.129X_1 + 0.128X_2 + 0.127X_3 + 0.127X_4 + 0.128X_5 + 0.128X_6 + 0.123X_7 + 0.126X_8 \quad (1)$$

$$\text{among,}\begin{cases} X_1 = (\mathrm{R1} - 5498)/1913 \\ X_2 = (\mathrm{R2} - 118662)/97213 \\ X_3 = (\mathrm{R3} - 5307)/2827 \\ X_4 = (\mathrm{R4} - 443)/267 \\ X_5 = (\mathrm{R5} - 62387258)/64919187 \\ X_6 = (\mathrm{R6} - 2015)/1485 \\ X_7 = (\mathrm{R7} - 42355077)/39092689 \\ X_8 = (\mathrm{R8} - 41718)/9898 \end{cases}$$

3.3. Construction of principal component regression model

According to Equation (1) Z-value regression model, the Z-value of the sample year can be calculated by substituting the original data. It can be seen from the scatter plot of Z value and Y value (Fig. 1) that the two have a strong linear distribution relationship, so the linear regression model is used for fitting. Table 7 of the fitting model parameters shows that the model has a good goodness of fitting and passes the significance test. Therefore, the regression model of the number of taxis and principal component Z can be obtained as:

$$Y = 962Z + 9476 \tag{2}$$

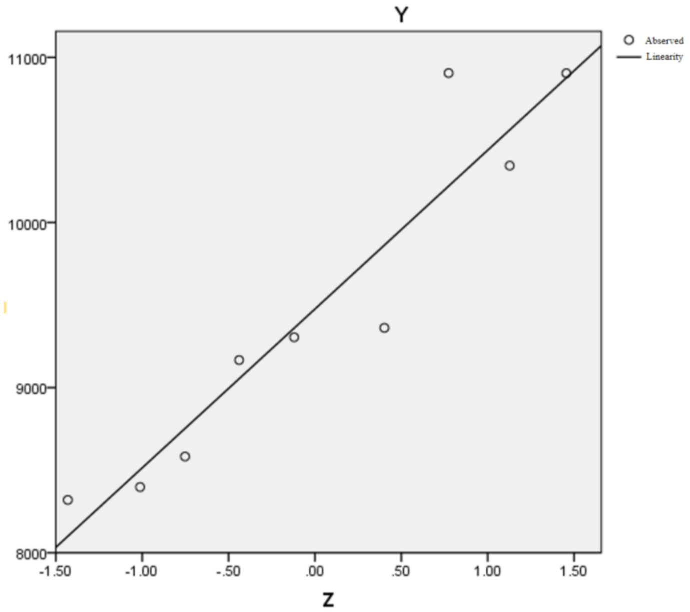

Figure 1: Scatter plot of Z and Y values

Table 7: Fit the model parameters

Equation	Model Summary					Parameter Estimates	
	R^2	F	df1	df2	Sig.	Constant	b1
Linearity	0.894646	59.44261	1	7	0.000115	9476.012	962.0058

3.4. Construction of prediction model based on Z value

In order to avoid the accumulation of errors in the prediction of many influencing factors, this paper uses the principal component (Z-value) to forecast separately and then calculates the number of taxis by substituting it into equation (2) to achieve the prediction of the number of taxis in the future. According to the scatter plot of Z-value and year (Fig. 2), it can be seen that Z-value and year have a significant linear relationship, so linear regression is used to construct the regression model of Z-value and year. As can be seen from the model parameter table in Table 8, the R square is 0.996 and passes the F test, so it can be considered that the model has a high goodness of fit. The prediction model for Z values and years is as follows:

$$Z = 0.3642T - 731.77 \tag{3}$$

Where T is the predicted year. By substituting (2), the prediction model of the number and year of taxis can be obtained as:

$$Y = 350T - 694487 \tag{4}$$

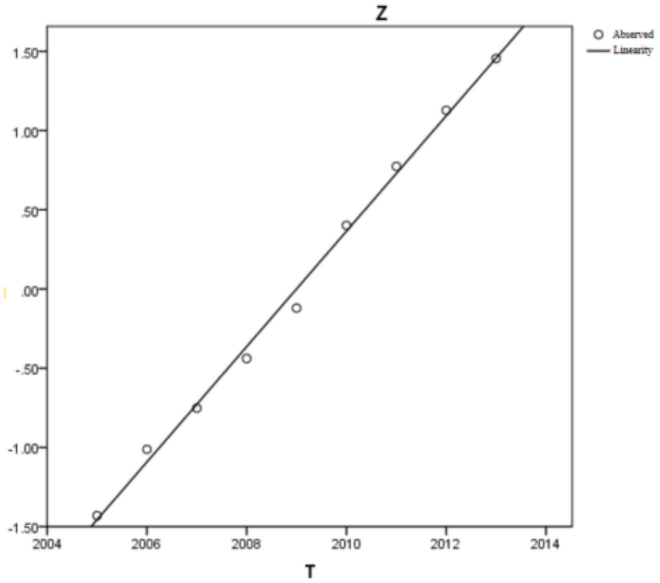

Figure 2: Scatter plot of Z value versus year

Table 8: Z-value and year fit model parameters

Equation	Model Summary					Parameter Estimates	
	R^2	F	df1	df2	Sig.	Constant	b1
Linearity	0.996	1717.30	1.00	7.00	0.00	-731.77	0.364

4. Model accuracy check

In this paper, the accuracy of the fitting model is tested by comparing the real value of the number of taxis with the model. By substituting the year into equation (4), the model value of taxi number can be obtained, and the precision analysis results are shown in Table 9.

Table 9: Model accuracy analysis

Year	Z	True value	Model value	Absolute Error	Relative Error
2013	1.45519	10904	10788	-116	1.06
2012	1.127531	10344	10438	94	0.91
2011	0.773672	10905	10087	-818	7.50
2010	0.401829	9362	9737	375	4.00
2009	-0.12045	9305	9387	82	0.88
2008	-0.43875	9167	9036	-131	1.43
2007	-0.75207	8583	8686	103	1.20
2006	-1.01227	8398	8335	-63	0.74
2005	-1.43064	8320	7985	-335	4.03
			Mean of Error	-89.90	2.42
			Error Standard Deviation	318	2

As can be seen from the precision analysis results of the model in Table 9, the average relative error is 2.42% and the standard deviation is 2%, that is, the sample prediction error is low and the error fluctuation is small. Therefore, it can be considered that the model has a high precision. Since the future changes of the influencing factors have not been thoroughly studied, the prediction model should be constantly updated according to the year data in order to maintain high model accuracy over time. Therefore, the prediction model in this paper is suitable for the short-term prediction of the number of taxis.

Formula (4) was used to predict the number of taxis in Hangzhou in the next 5 years (from 2014 to 2018) as follows: 11138, 11489, 11839, 12189 and 12540, respectively.

5. Conclusion

The number of urban taxis is an important decision-making basis for the urban passenger transport management department to formulate relevant management measures. Through correlation analysis, eight factors with high correlation with the number of taxis were extracted. Through principal component analysis, the principal components representing the original 8 influencing factors were extracted to eliminate the multicollinearity among influencing factors. Finally, based on the single regression prediction of the principal component, the regression prediction model of taxi number and forecast year is constructed. The accuracy test results show that the model has high precision and can be used to predict the number of taxis in short term. Finally, the prediction model is used to predict the number of taxis in Hangzhou from 2014 to 2018.

References

[1] Doughs, G. Price regulation and optimal service standards: The taxicab Industry [J]. Journal of Transport Economics and Policy, 1972, (20):116-127.

[2] Lu Jian, Wang Wei. Determination Method of Urban Taxi Ownership [J]. Journal of Traffic and Transportation Engineering, 2004, 4(1): 92-95.

[3] Che Lan. Development Status and Demand Forecast of urban taxi [J]. Shanxi Science and Technology, 2006, (6):104-l07.

[4] Jin Zhenyao et al. Research on taxi investment in Hangzhou [J]. Science and Technology Information, 2012, (01): 212-213.

[5] Zhang Wenlin. Operational Application of Principal Component Analysis in SPSS [J]. Market Research, 2005, (12):31-34.