Short-term passenger volume forecast and model analysis of Beijing public transport

Xiangyu Li[a*], Manman Xie[b*]
[a]School of Traffic and Transportation, Beijing Jiaotong University, Haidian District, Beijing 100044, China;
[b] School of Geosciences and Info-Physics, Central South University, Changsha, Hunan 410075, China
*Contribute equally to this work
*18221163@bjtu.edu.cn; phone 18974067198

ABSTRACT

The prediction of bus passenger volume is the fundamental research content of bus transfer optimization. In order to get more accurate passenger volume data and improve the utilization efficiency of urban traffic resources, according to randomness, time-varying and uncertainty of public transport passenger volume in Beijing, combined with the current new coronavirus pneumonia epidemic, this paper collected the relevant data of Beijing in the past 40 years, and predicted and analyzed them from four dimensions of public transport, urban scale and residents' economic level, taxi and sudden health events by BP neural network and regression analysis. The results show that BP neural network has good prediction results, and BP neural network is suitable for large sample size, which needs to fit or predict complex nonlinear relationships.

Keywords: BP neural network, public transport forecast, regression analysis, public health events

1. INTRODUCTION

The scale of cities is constantly exaggerated, and traffic congestion and environmental problems are becoming more and more serious, which greatly increases the demand for transportation. Giving priority to the development of public transport is considered to be the best strategy to solve large and medium-sized urban traffic problems. The public transport system is inseparable from the prediction of passenger traffic. Only on the premise of mastering the change law of passenger traffic volume in advance, can the bus group formulate scientific and efficient operation management, route optimization and dispatching ratio, then reasonably allocate resources. Based on the accurate prediction of bus passenger volume in Beijing, this paper promotes the efficient allocation of buses, meets the basic living and travel needs of residents, and efficiently uses urban traffic resources.

Recently, researchers have researched on the prediction of the bus passenger volume. Liu C [1] et al. established a prediction model of the number of people on and off the bus line site based on the improved BP neural network. The training and testing of relevant data in Harbin proved it had high prediction accuracy. Jiang Ping [2] et al. took Hefei's passenger flow as an example to compare and analyze the Elman model and BP neural network prediction of bus passenger flow. The results show the Elman model is better than the BP neural network model. Guo [3] applied the BP neural network based on EMD to the prediction of public transport passenger flow, and collected the relevant data of public transport in Shijiazhuang. There was a problem that the small amount of data led to the short prediction time. Jia Qinglin [4] established a prediction model combined with wavelet transform theory and BP neural network. After comparison, it is found that the prediction accuracy and fitting degree of wavelet neural network prediction model are improved, and it has adaptability. Because of COVID-19, researchers have recently increased the study of urban traffic under public health emergencies. Ma [5] et al. used a genetic algorithm to construct an emergency bus route optimization method considering the requirements of epidemic prevention. This method can get an optimization scheme that meets the needs of epidemic prevention and residents ' travel from many options. Li [6] et al. proposed a public transport operation decision support system framework for the prevention and control of infectious diseases and conducted verification analysis with Xiamen.

Therefore, under the background of COVID-19, according to the relevant data of public transport in Beijing in the past 40 years, combined with the quantifiable factors affecting bus passenger traffic, this paper introduces the impact of public health emergencies on passenger traffic, and applies BP neural network and regression analysis method to predict bus

passenger traffic in the next three years, and carries out a comparative analysis, hoping to improve the efficiency of urban public transport, reduce energy consumption and travel costs.

2. RELATIVE THEORY

2.1 BP neural network

BP neural network is a feedforward network model, which includes two processes of forward propagation of signal and backward propagation of error, and three levels of input layer, middle layer and output layer. In this paper, a three-layer BP neural network structure is used. The selection of specific node number and excitation function is shown in Figure 1.

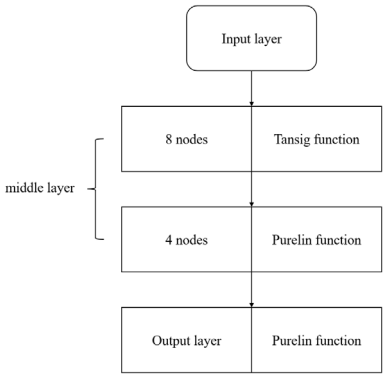


Figure 1.Neural network construction

2.2 Multivariate linear regression analysis

Multivariate linear regression is commonly used in the prediction and analysis of continuous numerical variables.

Multiple linear regression assumes that the explanatory variable y and the explanatory variable xi meet the following linear relationship.

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{1}$$

Where ε is an unobservable perturbation and $\beta_i \beta_i$ is a regression coefficient.

When multiple linear regression is used for analysis, it is required that there is no problem of complete multicollinearity, serious multicollinearity and heteroscedasticity among the explanatory variables, and the significance test of the model and coefficient should be carried out.

2.3 Factor analysis

Factor analysis is a method of comprehensive processing of multiple indicators to get a few comprehensive variables, which is suitable for the strong correlation between various indicators. The preliminary regression results show that there is a strong correlation between each variable, so factor analysis can reduce the dimension of the index in qualitative analysis.

3. EMPIRICAL ANALYSIS

3.1 Data source and pretreatment

Rail transit bears the travel pressure of some residents. After the completion of rail transit, it will compete with the bus to a certain extent. The size of cities with different populations will also be different, and the regions with more population have high urban density and large public transport passenger volume. The economic situation of residents affects people' s travel mode choice and travel ' times. For example, when the economic level decreases, people travel by public transport, but people will reduce the number of trips because of the decrease of disposable income meanwhile. Taxi and bus are competitive. Therefore, this paper takes urban public transport, urban scale and residents ' living standards as influencing actors.

This paper collects Beijing public transport data, per capita disposable income data, taxi data, and GDP growth rate influenced by health emergencies from 1980 to 2020. The data is from Beijing Statistical Yearbook.

Table 1.Evaluation index

Category	Index
Public transportation	Passenger Transport Volume (million)
	Number of bus lines (lines)
	Bus line length (km)
	Operating vehicles (vehicles)
	Rail transit passenger volume (million)
	Number of rail transit lines (lines)
	Rail transit line length (km)
Urban Scale and Economic Level of Residents	GDP (billion CNY)
	Population (million)
	Per capita disposable income (CNY)
Taxi	Number of taxis (vehicles)
	Taxi passenger volume (ten thousand people)
Other	gdp growth rate

Considering that bus passenger traffic will be affected by other emergencies and random factors, and the outbreak of COVID-19 lead to a substantial reduction in bus passenger traffic, this paper selects health emergencies as the influencing factors. Because health emergencies have a greater impact on GDP, so choose GDP growth rate to calculate the impact factor.

$$P = A/(A_{min} - A_{max}) \tag{2}$$

$$A = (A_i - A_j)/A_i \tag{3}$$

Where P is the impact factor of health emergencies, A is GDP growth rate, $A_{i,j}$ is GDP for that year and GDP for the previous year.

Because of the lack of collected data, it is necessary to preprocess it. When the missing data is in the middle year, the cubic spline interpolation method is used to complement. When the missing data are in the starting or ending year, by making the scatter plot of indicators changing with years, it can be got that the population has a strong linear relationship with time. Therefore, the population is supplemented by linear fitting. MATLAB is used to fit equations to get the final fitting results.

3.2 Establishment of BP neural network

In order to avoid overfitting to a certain extent, this paper uses the normalized mean square error as the expression of the loss function, the learning rate is set to 0.05, and the momentum gradient descent algorithm is used to calculate.

In order to test the generalization effect of the network, this paper randomly selects the data of 10 years (including 1982, 1983, 1984, 1988, 1993, 1999, 2004, 2007, 2018 and 2015) from the collected data of 41 years as the test set and the other as the training set. The training set is used to establish the model, and then the test set is used for error analysis of the model.

3.3 Multivariate linear regression analysis

This paper uses SPSS and STATA to perform regression analysis and improvement on public transport passenger volume. Since this paper is predictive regression, we pay more attention to goodness of fit R^2.

In the joint significance test, this paper conducts multiple linear regression of bus passenger volume on 12 explanatory variables, establishes multiple linear regression model, and conducts hypothesis test on the model. The p value is 0.0000, less than 0.05, showing that the established model is effective. The goodness of fit R^2 was 0.9468, greater than 0.9, indicating that the fitting effect was good.

Complete multicollinearity will lead to model failure, but rarely in practice. Approximate multicollinearity often occurs, but it only affects the significance of the coefficient of a single variable, resulting in inaccurate estimation of the regression

coefficient of a single variable, and does not affect the ability of the model to predict the explained variable. Since the regression model established in this paper is mainly used for prediction, the approximate multicollinearity problem is not considered.

Heteroscedasticity does not affect the unbiasedness and consistency of model estimation, but it will lead to the failure of hypothesis testing, so that the established model cannot be evaluated. This paper uses the white test to analyze the heteroscedasticity of the model. The p value is 0.4154, which is greater than the significance level 0.05. Therefore, the original hypothesis is accepted, that is, there is no heteroscedasticity problem in the data.

3.4 Factor analysis

In this paper, SPASS is used for factor analysis, and MATLAB is used to calculate and test the regression parameters. Firstly, KMO and Bartlett test are used to quantitatively determine whether the selected indicators are suitable for factor analysis. KMO value is 0.779, greater than 0.7, suitable for factor analysis. The gravel map showed the curve slowed down after the three components, and the total variance contribution rate of the two components reached 97.4%. Therefore, this paper finally selects three components, and uses principal component extraction, maximum variance method for rotation factor analysis.

Table 2.Postrotatory component matrix

	Component		
	1	2	3
Number of bus lines	0.96	0.14	0.12
Bus line length	0.94	0.24	0.17
Operating vehicles	0.95	0.16	-0.02
GDP	0.96	-0.28	-0.01
Population	0.98	0.01	-0.12
Per capita disposable income	0.97	-0.23	0.01
Number of taxis	0.84	0.47	-0.07
Taxi passenger volume	0.46	0.86	-0.18
Rail transit passenger volume	0.92	-0.25	-0.26
Number of rail transit lines	0.94	-0.31	-0.10
Rail transit line length	0.94	-0.32	-0.06
Public health emergency	0.42	0.02	0.90

By analyzing the above table, it can be concluded that the first component mainly reflects the influence of bus internal, social economy and rail transit on bus passenger traffic, the second component mainly reflects the influence of taxi, and the third component mainly reflects the influence of public health emergencies. Subsequently, the coefficient vectors v_1, v_2 and v_3 of these three comprehensive variables are calculated as follows.

Table 3.Comprehensive variable coefficient table (1)

	Number of bus lines	Bus line length	Operating vehicles	GDP	Population	Per capita disposable income,	Number of taxis
First variable	0.019	-0.027	0.041	0.179	0.115	0.164	-0.059
Second variable	0.115	0.166	0.144	-0.12	0.069	-0.097	0.334
Third variable	0.14	0.20	0.01	0.02	-0.09	0.03	-0.03

Table 4.Comprehensive variable coefficient table (2)

	Taxi passenger volume	Rail transit passenger volume	Number of rail transit lines	Rail transit line length	Public health emergency
First variable	-0.197	0.221	0.208	0.203	-0.158
Second variable	0.571	-0.072	-0.125	-0.142	-0.101
Third variable	-0.15	-0.23	-0.08	-0.04	0.90

Further, three comprehensive variables are got by the following calculation method.

$$x_i^* = v_1 \times x \tag{4}$$

$$\mathrm{x_i^* = v_1 \times x} \tag{5}$$

x_i^* represents the ith synthesis variable, v_i represents its coefficient, x represents 12 explanatory variables.

3.5 Regression analysis

The three comprehensive variables got by factor analysis are used for multiple linear regression of passenger volume. The p value of the test is 0.0000, through the joint significance test, but the goodness of fit is 0.6093, greater than 0.6. The fitting effect is general.

3.6 Error analysis

The data of the test set are used for error analysis by BP neural network, multiple linear regression and factor analysis, and the following conclusions are got.

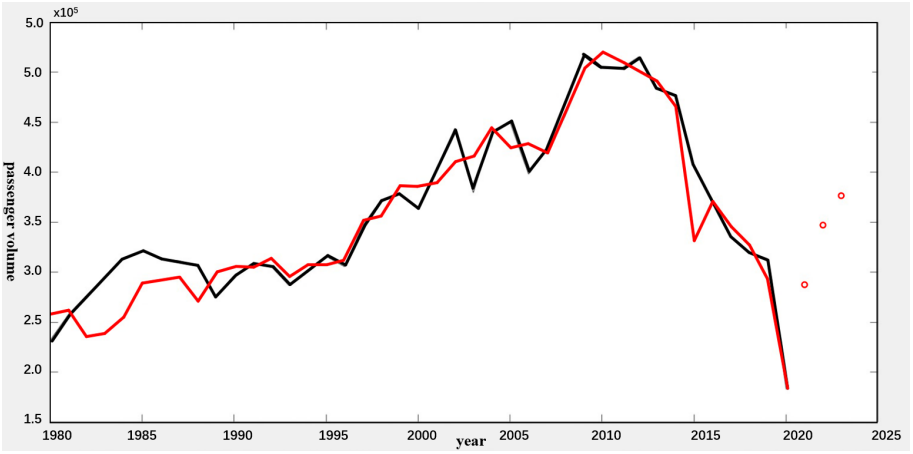
Table 5.Error comparison

	BP neural network	Multiple linear regression	Factor analysis
Maximum error	18.9%	14.0%	30.4%
Average error	9.6%	5.5%	10.6%

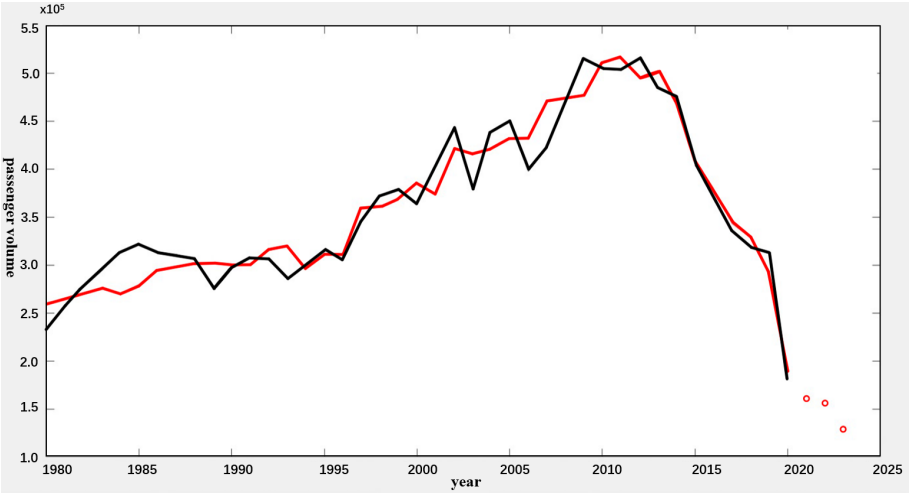
From the results, the prediction error of multiple linear regression is the smallest, followed by BP neural network, and the largest error is factor analysis combined with multiple linear regression.

3.7 Scenario prediction

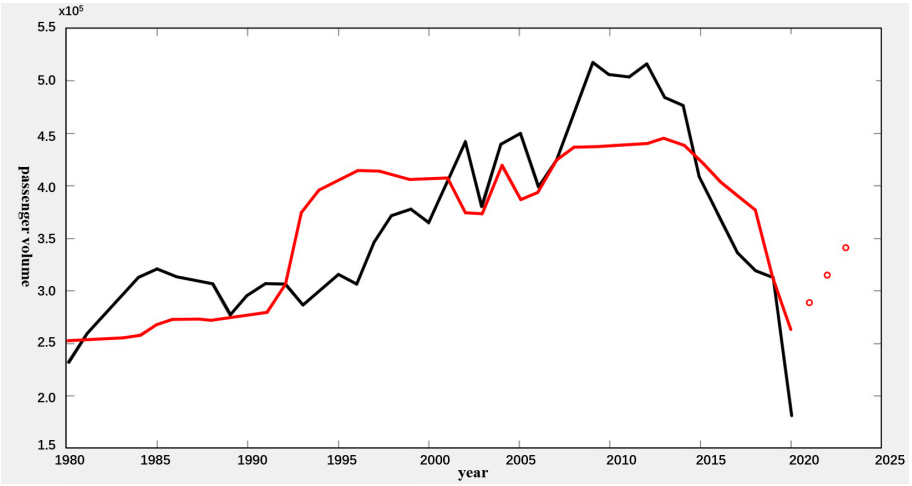
This paper predicts the bus passenger volume in the next three years according to the average growth rate. In this scenario, besides public health emergencies, the future growth rate of other indicators is the average of the previous three decades. The results of passenger volume (10,000 passengers) for the next three years predicted by the three methods are shown in Figure 2 and the specific results are shown in Table 6.



(a) neural network



(b) Multiple linear regression



(c) Factor Analysis Combined with Multiple Linear Regression

Figure 2.Forecasting results. The red circle represents the forecasts for the next three years. The red line represents the forecast passenger volume and the black line represents the actual passenger volume.

Table 6.Comparison of prediction results

	BP neural network	Multiple linear regression	Factor analysis
2021	287523	160941	288721
2022	347099	156109	314892
2023	376135	129013	340928

Based on the average growth rate scenario, BP neural network, multiple linear regression and factor analysis are used to predict the bus passenger volume of Beijing in the next three years, and the following conclusions are got.

1. On the test set, the effect of multiple linear regression is better than that of BP neural network and factor analysis, but multiple linear regression does not match the actual passenger volume in the next three years. The analysis believes that this is because of insufficient explanatory variables, that is, the impact of the epidemic on passenger volume is not fully considered, resulting in the endogeneity of the regression model. Finally, the BP neural network with better performance in the test set is taken as the final prediction result. The final prediction is shown in Table 7.

Table 7.Prediction results

Year	BP neural network prediction results
2021	287523
2022	347099
2023	376135

2. BP neural network realizes a mapping function from input to output, and has the function of realizing any complex nonlinear mapping. Therefore, it is very suitable for solving the problem of complex internal mechanism. When the relationship between the dependent variable and their respective variables is very complex or even there is no explicit expression, BP neural network can fit the relationship between the two. However, there are also the following problems when using BP neural network: unreasonable network structure will lead to slow convergence, and also affect the training and generalization ability of the network. It is easy to fall into the local optimum. There is a phenomenon of overfitting. Therefore, BP neural network is suitable for large sample size, which needs to fit or predict complex nonlinear relationship.

3. Multiple linear regression is a prediction method based on probability theory and mathematical statistics. Its accuracy depends on the choice of explanatory variables. Ideally, the disturbance term should not be related to each explanatory variable, which needs to choose as many explanatory variables as possible to reduce the impact of endogeneity. However, too many explanatory variables often bring multiple collinearity and heteroscedasticity, large deviations between prediction and practice. Therefore, multiple linear regression is suitable for long-term prediction with fewer samples and simpler relationships.

4. Factor analysis combined with multiple linear regression can be seen as an improvement of multiple linear regression. Factor analysis can reduce the dimension of multidimensional data and solve the problem of multicollinearity and heteroscedasticity to a certain extent.

4. CONCLUSION

This paper collects public transport data, per capita disposable income data and taxi data in Beijing in recent 40 years, and calculates the impact of public health emergencies by GDP growth rate. In this paper, the BP neural network is constructed based on the existing data, and the feasibility of using multiple linear regression analysis and factor analysis is verified. Then, the above three methods are used for error analysis on the test set. Finally, these three methods are used to predict the bus passenger volume in the next three years. The results show that BP neural network has good prediction effect. The research results have significance for passenger volume prediction under the influence of major public health emergencies.

In terms of public health emergencies, this paper only considers its impact on GDP growth rate to consider its impact on passenger traffic, and public health emergencies also affect passenger traffic from two aspects of traffic control and residents ' travel willingness. Future research will start from the above two aspects to further improve the accuracy of traffic passenger traffic prediction.

REFERENCES

[1] Liu C, Zhang Y Q, Chen H R. Transit Station′s Temporal Getting on/off Flow Forecasting Model Based on BP Neural Network[J]. Communications Standardization, 2008.

[2] Jiang P, Huang Z P. Forecast of Common Traffic Passenger Volume Based on Neural Network[J]. Communications Standardization, 2008.

[3] Guo Y N. EMD-based BP neural network for forecasting passenger flow of public transportation[D],Tianjing:Tianjing University, 2015.

[4] Jia Q L, JIN M J，ZHANG T，SUN F. Prediction of public transport passenger flow based on wavelet neural network model[J]. Journal of Wuhan Polytechnic University,2020,39(3):50-54.

[5] Ma C X, Wang C, Hao W, Liu J, Zhang Z L. Emergency customized bus route optimization under public health emergencies[J]. Journal of Traffic and Transportatin Engineering, 2020,20(03):89-99.

[6] Li J. Research on Decision Support for Public Transport Operations and Management for Epidemic Prevention and Control of Infectious Diseases[J]. Zhongguo Gonglu Xuebao/China Journal of Highway and Transport, 2021, 33(11):30-42.