Advances in Social Science, Education and Humanities Research, volume 427
5th International Conference on Social Sciences and Economic Development (ICSSED 2020)

Prediction of Private Car Ownership in China Based on the Improved PCA-Logistic Model

Bin Liu[1]Tianshu Zhao[1,*]Ranxia Zhang[1]

[1]*School of Maritime Economics and Management, Dalian Maritime University, Dalian Liaoning 116024, China*
[*]*Corresponding author. Email: 727848847@qq.com*

ABSTRACT
The change of private car ownership is a long-term nonlinear process with complicated influencing factors and nonlinear relationship also exists among them, which requires reasonable and accurate methods for analysis and prediction. In order to solve these problems, the traditional principal component analysis (PCA) method is improved in this study. Considering the integrity of the data, eight representative main factors influencing China's private car ownership are chosen. Firstly, the improved PCA method is used for "nonlinear" process of dimension reduction. Then, the Logistic regression model is applied to describe the relationship between private car ownership and the extracted principal components. The improved PCA-Logistic regression model is proposed finally. Compared with the results of the traditional PCA-Logistic method, it can be found that the improved PCA-Logistic model can effectively eliminate the nonlinear relationship between the data and it can change the quadratic nonlinear principle component regression curve obtained by the traditional method into a linear curve, which is more in line with the principle of PCA. The obtained nonlinear Logistic regression curve is in better agreement with the actual data, indicating that this method can evaluate the private car ownership in China more accurately. Based the model, the private car ownership in China from 2020 to 2024 have been predicted and the results show that China's private car ownership may exceed 300 million in 2020 and exceed 400 million in 2024.
Keywords: *traffic engineering, private car ownership, improved principal component analysis (PCA), Logistic regression model, prediction*

1. INTRODUCTION

With the sustainable development of China's economy, the continuous improvement of social income level, the gradual acceleration of urbanization, the personal car ownership in China has been climbing year by year [1]. According to the data of the National Bureau of Statistics, the number of personal cars in China was only about 280,000in 1985. By 2018, the number of personal cars exceeded 200 million. Though personal cars bring us convenience, they also bring a series of problems to our life, such as road congestion, frequent traffic accidents. The prediction of personal car ownership can provide data support for urban road traffic planning and provide a basis for the government's investment cost budget for the construction of traffic facilities [2].

For the prediction of car ownership, foreign researchers have done many studies and put forward many models such as comperta model based on set model [3], multiple Logit model based on non-set model and multiple Probit model [4]. Based on local demographic and social data, these models have been successfully applied to developed countries. It can be seen from the above studies that there are certain regional features in foreign prediction models. Because the automobile culture and economic level in foreign countries are quite different from that in China, it is difficult to directly apply these studies to the analysis of personal automobile ownership in China [5,6]. The research on the prediction of car ownership in China started late, but many models and methods were also put forward. Using BP neural network, Chen and Kong established a prediction model with time series to analyze and predict the ownership of private cars in China [7]. Zhu et al. used the grey system theory to establish a prediction model of private car ownership and analyzed the private car ownership in a certain region from 1996 to 2007 [8]. Zhang and Chang eliminated the repeated information among the factors that affected the car ownership by PCA and established the PCA-BP neural network prediction model to analyze the car ownership in Nanjing from 1978 to 2005and also predicted the car ownership in Nanjing in the future [6]. Based on the traditional Logistic model and the extreme difference format logistic model, Ren et al. established a Logistic combination model weighted by error and standard deviation to predict the car ownership in China, and concluded that the car ownership would reach 235 million in 2020 [9].

Through sorting out relevant literatures at home and abroad, the author improves the traditional PCA method by using the logarithmic transformation method. Combined with the Logistic model, we propose the improved PCA-Logistic model and analyze the personal car ownership in China.

2. MODELLING

2.1. Improved principal component analysis

The traditional PCA is a "linear" dimensionality reduction method. If there is a non-linear relationship between variables, the effect of dimensionality reduction will not be obvious. Therefore, it is necessary to improve the traditional PCA. Based on the analysis of the original influencing factors data, it is found that it presents a trend of similar exponential form over time, so the improved method adopts the logarithmic transformation method to achieve better dimensionality reduction effect. The specific steps are as follows [10]:

(1) Log transformation of the original data

The original data matrix is assumed to be $\mathrm{X} = (x_{\mathrm{ij}})_{n\times p}$ with $y_{\mathrm{ij}} = \ln x_{\mathrm{ij}}$. The matrix after the logarithm transformation is obtained and it can be expressed as $\mathrm{Y} = (y_{\mathrm{ij}})_{n\times p}$, $\mathrm{i} = 1,2,...n;\ \mathrm{j} = 1,2,...p$.

(2) Replace the original data with y_{ij} as the new data and standardize it to eliminate the influence of each feature on the order of magnitude, the formula is obtained as: $\mathrm{Y}^{*} = (\mathrm{Y} - \mu)/\sigma$, where μ is the mean value of each indicator data and σ is the standard deviation of each indicator data.

(3) According to the standardized matrix, the covariance matrix is calculated by $\mathrm{c} = \frac{1}{n-1}\mathrm{Y}^{*}\mathrm{Y}^{*\mathrm{T}}$, where c is the covariance matrix and n is the number of elements of the data index. Y^{*} is the standardized data and the superscript T is the transpose operation.

(4) The eigenvalues and eigenvectors are calculated according to the covariance matrix and the eigenvalues are arranged from the largest to the smallest.

(5) The cumulative contribution rate of the principal component is calculated to extract the principal component according to the calculation results. Generally, the corresponding component with cumulative contribution rate of more than 85% is selected as the principal component.

2.2. Improved PCA-Logistic Model

Combined with the above methods, an improved PCA-Logistic model is proposed. Firstly, several representative factors affecting the personal car ownership in China are selected as the evaluation indexes, and the original data are logarithmically transformed to eliminate the non-linear influence and then standardized to eliminate the order of magnitude differences. Finally, the dimension reduction factor analysis module in SPSS is used to analyze the improved principal component, determine the principal component and calculate the score of the principal component.

According to the calculation results of principal component contribution rate and cumulative contribution rate, the first principal component FAC1_1 is selected as the independent variable. The ownership of individual cars is considered as the dependent variable. And then the Logistic model regression analysis is conducted. The final model equation adopted here is:

$$y = \frac{m}{1 + \alpha_0 \alpha_1^{x}} \tag{1}$$

where α_0 and α_1 are the parameters to be determined. The appropriate maximum personal car ownership m is selected as the fixed value. The initial values of α_0 and α_1 are given for iterative solution. The nonlinear least square method is used to solve the parameters and the iterative algorithm used here is Marquardt method.

3. EMPERICAL ANALYSIS

3.1. Variable selection and data source

It can be found that there are various factors affecting the change of personal car ownership, such as economic factors, social factors, and environmental factors [11]. Considering the availability of complete long-term data, this paper selects eight influencing factors including urban population P_1, urbanization level P_2, per capita GDP P_3, resident consumption level P_4, highway mileage P_5, proportion of primary industry production P_6, proportion of secondary industry production P_7 and proportion of tertiary industry production P_8.The sample data of China's personal car ownership Y and its influencing factors are shown in Table 1. The time interval is 1985-2018. The data are all from the National Bureau of Statistics of China.

3.2. Improved Principal Component Analysis

SPSS22.0 is used to extract the principal components according to the following steps:

(1) Logarithmic transformation of raw data.

(2) After transformation, the eight influencing factors are standardized. PCA is carried out and the total variance of interpretation is shown in Table 2.

Table 1 Data of car ownership and influence factors of China

Year	t	P_1 (×10^4)	P_2 (%)	P_3 (¥)	P_4 (¥)	P_5 (km)	P_6 (%)	P_7 (%)	P_8 (%)	Y (×10^4)
1985	1	25094	23.71	866	440	942400	28.1	42.6	29.3	28.49
1986	2	26366	24.52	973	497	962800	26.8	43.4	29.8	34.71
1987	3	27674	25.32	1123	565	982200	26.5	43.2	30.3	42.29
1988	4	28661	25.81	1378	714	999600	25.4	43.4	31.2	60.42
1989	5	29540	26.21	1536	788	1014300	24.7	42.4	32.9	73.12
1990	6	30195	26.41	1663	831	1028300	26.7	40.9	32.4	81.62
1991	7	31203	26.94	1912	932	1041100	24.2	41.4	34.5	96.04
1992	8	32175	27.46	2334	1116	1056700	21.4	43.0	35.6	118.2
1993	9	33173	27.99	3027	1393	1083500	19.4	46.1	34.5	155.77
1994	10	34169	28.51	4081	1833	1117800	19.5	46.1	34.4	205.42
1995	11	35174	29.04	5091	2330	1157000	19.7	46.7	33.7	249.96
1996	12	37304	30.48	5898	2789	1185800	19.4	47.0	33.6	289.67
1997	13	39449	31.91	6481	3002	1226400	18.0	47.0	35.0	358.36
1998	14	41608	33.35	6860	3159	1278500	17.2	45.7	37.1	423.65
1999	15	43748	34.78	7229	3346	1351700	16.1	45.3	38.6	533.88
2000	16	45906	36.22	7942	3721	1679800	14.7	45.4	39.8	625.33
2001	17	48064	37.66	8717	3987	1698000	14.1	44.7	41.3	770.78
2002	18	50212	39.09	9506	4301	1765200	13.4	44.3	42.3	968.98
2003	19	52376	40.53	10666	4606	1809800	12.4	45.5	42.1	1219.23
2004	20	54283	41.76	12487	5138	1870700	13.0	45.8	41.2	1481.66
2005	21	56212	42.99	14368	5771	3345200	11.7	46.9	41.4	1848.07
2006	22	58288	44.34	16738	6416	3456999	10.7	47.4	41.9	2333.32
2007	23	60633	45.89	20505	7572	3583715	10.4	46.7	42.9	2876.22
2008	24	62403	46.99	24121	8707	3730164	10.3	46.8	42.9	3501.39
2009	25	64512	48.34	26222	9514	3860823	9.9	45.7	44.4	4574.91
2010	26	66978	49.95	30876	10918	4008229	9.6	46.2	44.2	5938.71
2011	27	69079	51.27	36403	13133	4106387	9.5	46.1	44.3	7326.79
2012	28	71182	52.57	40007	14698	4237508	9.5	45.0	45.5	8838.6
2013	29	73111	53.73	43852	16190	4356218	9.4	43.7	46.9	10501.68
2014	30	74916	54.77	47203	17777	4463913	9.2	42.7	48.1	12339.36
2015	31	77116	56.1	50251	19397	4577296	8.8	40.9	50.2	14099.1
2016	32	79298	57.35	53980	21227	4696263	8.6	39.8	51.6	16330.22
2017	33	81347	58.52	59201	22902	4773468	7.9	40.4	51.6	18515.11
2018	34	83137	59.58	64644	25002	4846531	7.2	40.7	52.2	20574.93

Table 2 Component score coefficient matrix

factor	Initial eigenvalue (%)			The sum of the squares of the weights (%)		
	total	variance	sum	total	variance	sum
1	6.865	85.812	85.812	6.865	85.812	85.812
2	1.017	12.718	98.531	1.017	12.718	98.531
3	.078	.981	99.512			
4	.029	.357	99.869			
5	.006	.073	99.942			
6	.004	.054	99.996			
7	.000	.003	99.999			
8	.000	.001	100.000			

As seen in Table 2, the characteristic value of the first principal component is 6.865, and its variance accounts for 85.812% of the total variance, which plays a major role in personal car ownership. According to the principal component selection criteria, the original eight influencing factors are replaced by this principal component.

(3) The matrix of component score coefficients is shown in Table 3. According to Table 3, the factor expression of the first principal component can be written as:

$$Y_1 = 0.146P_1 + 0.145P_2 + 0.145P_3 + 0.144P_4 + 0.142P_5 - 0.145P_6 - 0.008P_7 + 0.143P_8 \quad (2)$$

The normalized data are calculated according to the above expressions, and the data related to the first principal component FAC1_1 are shown in Table 4.

Table 3 Component score coefficient matrix

Influence factor	component	
	1	2
P_1	.146	.015
P_2	.145	-.013
P_3	.145	.063
P_4	.144	.047
P_5	.142	-.017
P_6	-.145	-.059
P_7	-.008	.981
P_8	.143	-.100

3.3. Principal component regression curves from the traditional PCA and the improved PCA

Fig. 1 and Fig. 2 show the regression curves of the first principal component FAC1_1 and the time obtained by the traditional PCA and the improved PCA, respectively. It can be seen that the regression curve of the traditional PCA method is almost quadratic and non-linear with the time, while the regression curve obtained by the improved PCA is almost linear with the time. Since the PCA is more suitable for linear data, it can be seen that the improved PCA can eliminates the non-linear relationship between the data and it is more in line with the principle of PCA, which is conducive to improving the accuracy of prediction.

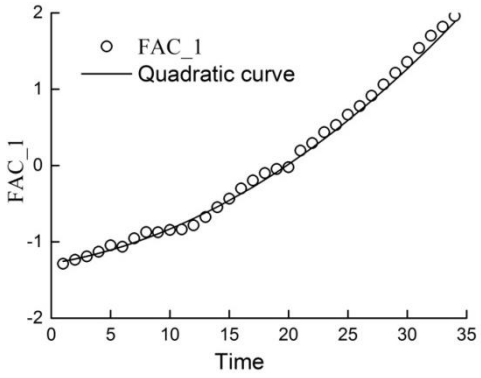

Fig.1 Regression curve of FAC1_1 and t under traditional PCA

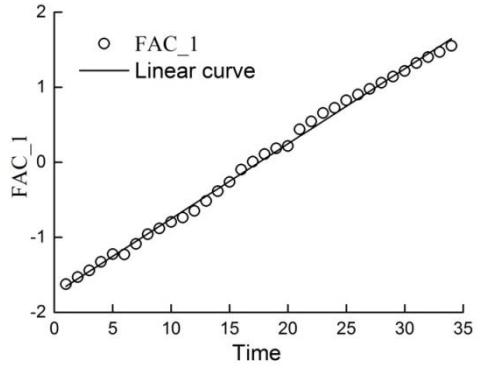

Fig.2 The regression curve of FAC1_1 and t under the improved PCA

Using linear regression model, the regression formula between the first principal component FAC1_1 and the time by the improved PCA method can be obtained as: $FAC1\_1 = -1.754 + 0.1t$.

Table 4 Calculated data of principal components FAC1_1

No.	FAC1_1	No.	FAC1_1	No.	FAC1_1
1	-1.62551	13	-0.51785	25	0.8258
2	-1.53095	14	-0.38518	26	0.90326
3	-1.44312	15	-0.26309	27	0.97853
4	-1.32814	16	-0.097	28	1.05912
5	-1.22421	17	0.00821	29	1.1414
6	-1.22822	18	0.10915	30	1.21696
7	-1.08958	19	0.18725	31	1.32068
8	-0.95964	20	0.21707	32	1.40169
9	-0.88286	21	0.43955	33	1.47019
10	-0.79524	22	0.54391	34	1.55138
11	-0.73836	23	0.65727		
12	-0.64814	24	0.72567		

3.4. Regression Analysis of Logistic Model

3.4.1. Logistic Regression Analysis of Improved PCA

Logistic model is used for regression analysis of the first principal component FAC1_1 and personal car ownership. As the maximum car ownership *m* is uncertain, it needed to be estimated in advance. There are roughly three estimation methods including direct maximum selection method, expert judgment method and pure mathematical derivation method. The author uses the expert judgment method to determine the maximum personal car ownership. According to the literature [12], the satisfaction point of car ownership in developed countries is 0.62, but for developing countries such as China with low per capita income and high population density, the literature pointed out that when the satisfaction point is 0.5, the regression result will be more suitable. Therefore, it is assumed that the satiation point of car ownership rate in China is 0.5 and the population is 1.4 billion and the proportion of personal automobiles is assumed to be 75%. It is calculated that the maximum number of personal car in China is 525 million, and Eq. (1) becomes:

$$y = \frac{52500}{1+\alpha_0\alpha_1^{x}} \tag{3}$$

The initial values of constants α_0 and α_1 are assumed to be 100 and 0.5, respectively. The tool SPSS22.0 is used to conduct Logistic nonlinear regression. The parameters are shown in Table 5.

According to Table 5, the correlation coefficient of fitting degree can be obtained as $R^2 = 0.998$. The closer R^2 is to 1, the closer the regression curve is to the data. It can be seen that the fitting degree of the logistic regression curve is very high and the final curve equation is:

$$y = \frac{52500}{1+74.903\times 0.081^{x}} \tag{4}$$

Table 5 Estimation values of parameters of the improved PCA-Logistic model

parameter	estimate	The standard error	95% confidence interval	
			Lower limit	Upper limit
a	74.903	3.488	67.798	82.009
b	.081	.003	.075	.086
		R^2=.998		

3.4.1. Comparison of logistic regression of the traditional PCA-Logisticmodel and improved PCA-Logistic model

Similarly, the traditional PCA-Logistic model is used and the parameter estimation results are shown in Table 6, $R^2 = 0.993 < 0.998$ indicates that the regression curve fitting degree of the improved PCA-Logistic model is better and the prediction result will be more accurate.

Table 6 Estimation values of parameters of the traditional PCA-Logistic model

parameter	estimate	The standard error	95% confidence interval	
			Lower limit	Upper limit
a	28.128	1.569	24.932	31.325
b	.220	.008	.204	.236
		R^2=.993		

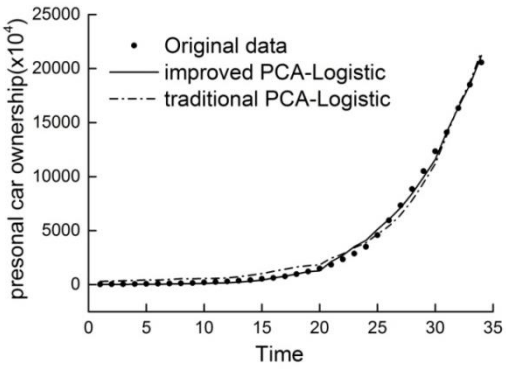

Fig.3 The comparison of the predicted personal car ownershipfrom the traditional PCA-Logistic model and improved PCA-Logistic model

Fig.3 is the comparison of the predicted personal car ownership obtained by the traditional PCA-Logistic and improved PCA-Logistic model. As seen from Fig.3, before 2007, the predicted data of personal car ownership obtained by the traditional PCA-Logistic model is higher than the actual data point of personal car ownership, while after 2007, it is lower than the actual data point. The data of personal car ownership obtained by the improved PCA-Logistic is closer to the actual data point, which once again proves that the improvement of PCA is conducive to improving the fitting degree of Logistic regression curve and improving the accuracy of prediction.

3.5. Prediction of Personal Car Ownership

In order to predict China's personal car ownership, the corresponding timeis substituted into the formula $FAC1\_1 = -1.754 + 0.1t$ to obtain the corresponding FAC1_1. Then the results are put into Eq. (4) to obtain the

predicted value of personal car ownership in 2020-2024, as shown in Table 7.

Table 7 The prediction of private car ownership from 2020 to 2024

year	private car ownership($\times 10^4$)
2020	30457.62
2021	33591.96
2022	36514.53
2023	39164.7
2024	41507.77

4. SUMMARY

The traditional PCA method is improved by the logarithmic transformation method, and the improved PCA-Logistic model is proposed in combination with Logistic model. This paper uses the improved PCA to conduct "nonlinear" dimensionality reduction treatment on eight representative main factors affecting the personal car ownership in China and extract the principal components, and uses the logistic regression model to study the relationship between the principal component and the car ownership.

By comparing the estimation results of PCA logistic model before and after improvement, it is found that the improved PCA-Logistic model can effectively eliminate the non-linear relationship between long-term data, and the obtained linear principal component regression curve is more in line with the principle of PCA, and also effectively improve the fitting degree of logistic model regression, so as to more accurately predict the future personal car ownership in China. It is predicted that the number of personal cars in China will exceed 300 million in 2020 and 400 million in 2024.

The method in this paper only analyzed the influencing factors when extracting the principal components, and did not consider the relationship between them and the car ownership. In the following work, new models will be considered comprehensively to obtain more reasonable evaluation indicatorsand more influencing factors will be considered for analysis.

REFERENCES

[1] Li Shaoyi. Is China ready for auto society? [J]. Automotive Observer, 2010(10):79.

[2] Yan Xiao. Research on the market diffusion of new energy vehicles in China under the Background of Low-carbon transportation [D]. China University of Geosciences, 2016.

[3] Dargay J., Gately D. Income's effect on car and vehicle ownership, worldwide: 1960-2015 [J]. Transportation Research Part A, 1999, 33:101-138.

[4] Ai C, Norton E C. Interaction terms in logit and probit models[J]. Economics letters, 2003, 80(1): 123-129.

[5] Shi Qiong, Wu Qunqi. Analysis for ownership and use of private car [J]. Journal of Chang'an University (Social Science Edition), 2005, 7(2):24-29.

[6] Zhang Xuewu, Chang Jinyi. Research on urban car ownership prediction based on PCA-BP neural network [J]. Computer Simulation, 2012, 29(12):376-379.

[7] Chen Yong, Kong Feng. Application of BP neural network to forecast possession of private automobile [J]. Computing Technology and Automation, 2003, 22(3):67-70.

[8] Zhu Kaiyong, Zhou Shengwu, Lou Keyuan, Sun Chengtong. Research on grey model about forecast and control for total number of private cars [J]. Journal of China University of Mining & Technology, 2008, 37(6):868-872.

[9] Ren Yulong, Chen Rong, Shi Lefeng. Prediction of civil ownership in China based on Logistic combination model[J]. Journal of Industrial Technological Economics, 2011, 30(8):90-97.

[10] Qu Shuanghong, Li Hua, Li Gang. Some common improvement methods based on principal component analysis [J]. Statistic & Decision, 2011(5):155-156.

[11] Chen Xuanxuan. Study on the influencing factors of car ownership from the perspective of population, resources and environment [D]. Capital University of Economics and Business, 2008.

[12] Wang Yini. Prediction of demand for automobile in China: an analysis based on the Gompertz model[J]. Research on Financial and Economic Issues, 2005(11):45-52.