Research and Prediction of Health Expenditure Factors in China Based on Machine Learning Methods

Xiaoqin Zhang[1], Xiaowen Wan[1, *]

[1]*School of Economics and Management, Jiangxi University of Traditional Chinese Medicine, Nanchang, 330004, Jiangxi Province, China*
Xiaoqin Zhang Email 17718319635@163.com, Xiaowen Wan Email 645479278@qq.com.
Correspond author: Xiaowen Wan Email 645479278@qq.com. telephone:13732971307.

Abstract:
Objective: To analyze the influencing factors of China's total health expenditure and predict the development trend of China's total health expenditure in the next five years based on historical statistics and provide some theoretical basis for formulating relevant policies. Methods: The main factors that may influence the total health expenditure in China were selected by the theory of health demand-supply relationship, and the degree of influence of the factors influencing the total health expenditure in China was studied by using gray correlation and principal component analysis (PCA). Based on the analysis results of the two methods and the comparative analysis of the fit of the various models, the machine learning model was finally applied to forecast the total health expenditure in China in the next five years. Results: The total health expenditure from 2021 to 2025 are 78663.60, 83950.43, 88748.01, 93397.45, and 97974.29 billion yuan, respectively. The forecast results indicate that the total health expenditure in China will continue to maintain the growth trend in the next five years, but the growth rate will gradually level off. Conclusion: Both gray correlation and PCA analysis show that the influencing factors of China's total health expenditure are mainly reflected in economic income and medical services. Therefore, it is necessary to ensure the coordinated development of total health expenditure and economy, and improve the financing structure of total health expenditure. Improve the medical service system and optimize the allocation of health resources.

Keywords: *machine learning, Elman Neural Network, Principal component analysis, prediction of health expenditure*

1 INTRODUCTION

National health expenditure (NHE) is the result of health expenditure accounting, which is used as a comprehensive measure in monetary terms to comprehensively reflect the total amount of money consumed by society as a whole in integrated services in a certain period of time in a country or region [9]. The continuous rise in total health expenditure puts enormous pressure on the sustainability of health demand and financing. Therefore, accurately grasping and predicting the development trend of health expenditure can provide an early insight into the trend of health inputs and utilization, enabling timely policy adjustments and further improvement of the health care system reform. At present, the research on health expenditure is mainly divided into two aspects: the first is the study on the influencing factors of health expenditure; the second is the forecast of health expenditure.

Domestic and foreign scholars are also more focused on the research methods on the influence of health expenditure factors, for example, Getzen et al. correlated the cross-sectional data of 20 OECD countries from 1960-1988 and concluded that the increase of health expenditure is largely a policy and expenditure management issue rather than a demographic factor [2]. Reimers et al. took advantage of the cross-sectional data to correlate 21 The analysis of the association between health expenditure and GDP in OECD countries concluded that health expenditure not only determined by income, another driver is medical progress, which is represented by different variables, such as life expectancy, infant mortality and the proportion of elderly people [1].

At present, scholars at home and abroad have a wide variety of forecasting methods for health costs, for example, Wuet al. based on the FAGM (1,1, t^a) model to

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 737-744, 2023.
https://doi.org/10.2991/978-94-6463-040-4_112

forecast health expenditures and government health expenditures, and compared with other traditional gray models to conclude that the univariate fractional gray model FAGM (1,1, t^a) has more accurate forecasting accuracy [6]. Jiang Yan et al. used a logistic regression model to predict the total health costs of Beijing from 2000-2017 and concluded that: 2000-2014 is the period of incremental increase in costs, 2014-2033 is the period of rapid growth, and after 2033 tends to be stable [4].

Traditional health cost forecasting is mostly based on selecting historical data of total health costs in a certain period and then building statistical models for univariate time series forecasting, or building regression forecasting models with many factors as independent variables over a certain period of time, which tend to reduce the forecasting performance of the models by considering many factors.

Based on this, machine learning algorithms are used in this paper to study the total health expenditure in China. Machine learning can be optimized by programming algorithms based on prior data, and the trained algorithms can provide an effective inferential prediction and can also be used to identify and extract the significant degree of relationship between inputs and outputs. This paper draws on relevant literature studies on health expenditure impact indicators and selects a system of 14 impact factor indicators in 6 areas, and analyzes the 14 impact factor indicators using gray correlation and principal component analysis, respectively, and then trains Elman neural network models with 1980-2010 principal component dimensionality reduction data as input variables and total health expenditure as output variables, and the trained models predict Health expenditure from 2011-2020, while the model parameter settings will be used, and then the 4 influencing indicators with high gray correlation and all influencing indicators will be used as input variables, and then the prediction results of the 3 machine learning models will be compared. Finally, the machine learning model will then be compared with the traditional model for predicting total health expenditure, and then use the machine learning model predicting the total health expenditure in China for the next 5 years, and relevant policy recommendations will be proposed for the research results.

2 DATA SOURCES AND SEIECTION OF IMPACTS INDICATORS

2.1 Source Analysis

The data used in this paper are obtained from statistical yearbooks, mainly from the China Statistical Yearbook published by the National Bureau of Statistics and the China Health and Wellness Statistical Yearbook compiled by the national health department.

2.2 Selection of indicators for total health expenditure factors

Health expenditure, as one of the indicators to measure the effectiveness of the health system, is itself affected by many factors. In this paper, based on the reference of related literature, 14 indicators related to 6 factors, namely, economic income, socio-demographic, health resources, medical services, health financing, and education level, are taken from the theory of health demand supply relationship. This can be seen in Figure 1 and Table 1.

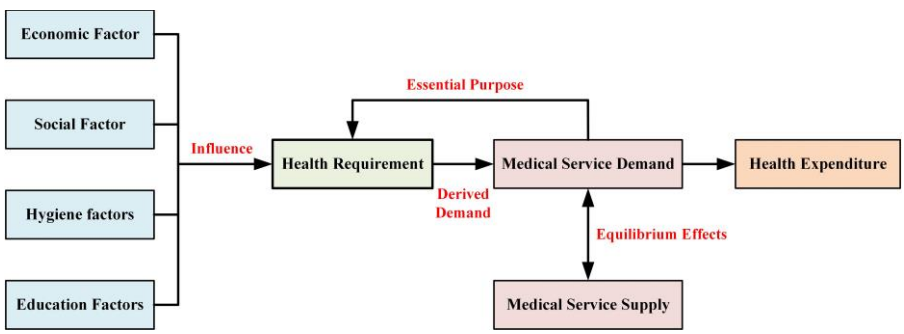

Fig.1. Health expenditure impact mechanism

Table 1 Indicator system of factors influencing total health expenditure

Type	Expression	Indicator	Units
—	Y	Total Health Expenditure	100 million yuan
Economic	X1	GDP	100 million yuan
Income	X2	Household Consumption	Yuan
Social Demography	X3	Year-end Population	Ten Thousand Person

	X4	Proportion of Population Aged Over 65	%
	X5	Human Mortality	%
Health Resources	X6	Number of Employed Persons	Unit Person
	X7	Number of Health Institutions	Pieces
	X8	Number of Beds in Health Institutions	Million Pieces
Medical Services	X9	Number of Clinical Visits	100Million Times
	X10	Number of Admissions	Ten Thousand Person
Health Financing	X11	Personal Health Expenditure	100 million yuan
	X12	Proportion of Government Health Expenditure in Fiscal Expenditure	%
Education Level	X13	Gross College Enrollment Rate	%
	X14	Junior High School Graduation Promotion Rate	%

3 ANALYSIS OF THE INFLUENCING FACTORS OF TOTAL HEALTH EXPENDITURE

In order to fully study the influencing factors of total health expenditure in China, this paper uses 2 methods to analyze the influencing factors of health expenditure in China from 1980 to 2020.

3.1 Gray correlation analysis

In the development process of a system, two factors are said to be highly correlated if they have a consistent trend of change, and the reverse is lower. The gray correlation analysis method is based on the degree of similarity or dissimilarity of the development trends between factors.

According to the calculation principle of gray correlation, this paper uses MATLAB2020b software to calculate the correlation between the total health expenditure and each influencing factor in China from 1980 to 2020 as shown in Table 2.

From Table 2, it can be seen that the correlation between the total population and the number of health care institutions at the end of the year among the 14 influencing indicators is lower than 0.6, indicating that these two indicators have a low correlation relative to the original series (total health expenditure), while all the remaining indicators are higher than 0.6, indicating that these indicators have a strong correlation on total health expenditure. As a whole, economic level, health services and health funding allocation are important factors on health spending, while differences in education level also have an indirect and important impact on health expenditure.

Table 2. Correlation degree of influencing factors of total health expenses

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
Relevance	0.91	0.93	0.47	0.81	0.66	0.88	0.46	0.86	0.87	0.9	0.92	0.72	0.85	0.67
IDX	3	1	13	10	12	5	14	7	6	4	2	10	8	11

3.2 Principal Component Analysis

Principal component analysis (PCA) is essentially finding the variance of some projection directions that are orthogonal to each other. The larger the eigenvalue of the covariance matrix of the original data, the greater the amount of information projected on the corresponding eigenvectors, which is the principal component [5] A small eigenvalue means that the data project very little information on these eigenvectors, then this data has little influence in the overall, and the data in the direction corresponding to the small eigenvalue can be removed.

Based on the impact indicator system established in the previous paper, the original data from 1980-2020 were subjected to a principal-form analysis, and firstly, KMO and Bartlett's sphericity tests were conducted using

MATLAB software, and the two indicators were used to determine the strength of correlation between variables and whether each variable was independent of the other, respectively. The test result was a KMO value of 0.863 and the significance of Bartlett's test was 0. For the KMO value: 0.8 on very suitable for principal component analysis, for Bartlett's test, if the significance is less than 0.05 or 0.01, it means that the principal component analysis can be done. From the results of the KMO test and Bartlett's spherical test, a principal component analysis can be performed.

The principal component analysis of the original data from 1980 to 2020, the contribution of the principal components obtained can be seen in Figure 2, the cumulative contribution of the first four principal components has reached 93.06%, according to the principle of principal component selection, the accumulation of the first three principal components has reached 85%, so according to the first three principal components can already replace the information of the original 15 indicators. The factor loading coefficients of the three principal components are given in Figure 3, and the importance of the hidden variables in each principal component can be analyzed from the factor loading coefficients. Principal component 3 mainly reflects the information of total population and population mortality rate at the end of the year, and principal component 2 mainly reflects the number of health care institutions, while principal component 1 can synthesize the information of the original indicators. As can be seen in Figure 3, the loading coefficients of factors such as gross domestic product, hospital admissions, number of beds in health care institutions, and absolute number of residents' consumption level are large, mainly reflecting on economic income and medical services. And it is basically consistent with the analysis method of gray correlation in the previous section. [3] [8]

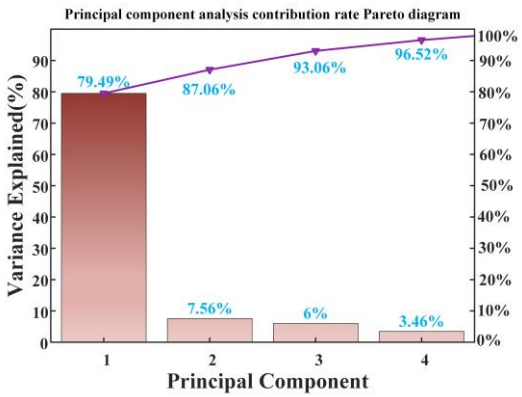

Fig.2. Cumulative contribution rate of principal component

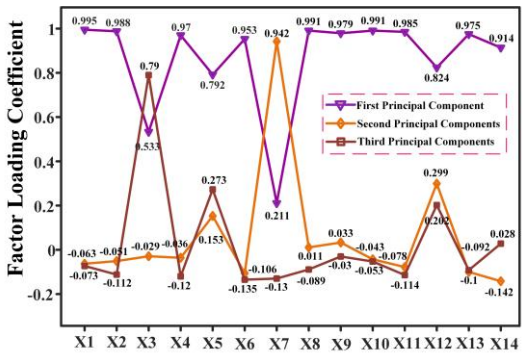

Fig.3. Principal component factor load coefficient

4 TOTAL HEAITH EXPENDITURE PROJECTIONS

The data of the impact indicators of China's total health expenditure for 41 years from 1980 to 2020 were analyzed by gray correlation analysis and principal component analysis above to analyze the impact degree of 14 indicators. In this section, based on the above analysis, the impact indicators will be used as input variables to predict China's total health expenditure by using machine learning method. The indicators with high correlation degree are selected as input variables and the data that have been dimensioned down by PCA are selected as input variables, respectively. The predictions of the data without any processing are also compared.

4.1 Elman Neural Network

Elman neural network was proposed by Professor Jeffrey Elman in 1990 as a feedback neural network model, which adds a takeover layer to the implicit layer of the feedforward neural network model to be used as a delayed operator, thus being able to achieve memory of the output value of the previous moment of the implicit layer, thus giving the system the ability to adapt to time changes and thus being able to increase the accuracy of prediction [7].

Based on the above analysis of the main influencing indicators of China's total health expenditure, this section first extracts 41 groups containing 4 principal component scores each as the input variables of the model, and the output variable is China's total health expenditure from 1980 to 2020, and the established data set will be used to train the model. Therefore, the structure of the model is the number of input layers is 4 and the number of output layers is 1.

The sample data from 1980-2010 were selected as the training set, and the model was trained by parameter adjustment, and the data samples from 2011-2020 were used as the test set to test and analyze the error of the trained model. The Elman neural network was constructed with MATLAB2020b software, the

activation function of the model was a sigmiod-type function, the output function was a linear purelin function, and the training function was a gradient descent function with momentum backpropagation and dynamic adaptive learning rate (traingdx). The number of nodes in the hidden layer has a great influence on the prediction effect of the model, so this paper uses the trial-and-error method to repeatedly train the model, and finally determines that the model fits best when the number of hidden layers is 25.

4.2 Prediction results of Elman Neural Network

In order to evaluate the prediction effect of the model, the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) are selected as prediction error evaluation indicators in this paper, and the three evaluation indicators are shown in the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - \hat{y}_i | \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

where y_i denotes the actual total health expenditure of the i sample, $\hat{y}_i$ denotes the predicted total health expenditure of the i sample, and n denotes the number of samples.

The best model was obtained by training the data after dimensionality reduction using PCA as the input of the model. Now, based on the parameter settings of this model, the input data are replaced with the original impact indicator data without dimensionality reduction (unprocessed) and the four impact indicator data with the highest selected gray correlation (gray processed), and the output data remain the national total health expenditure. Again, the data from 1980-2010 were used as the training set and the data from 2010-2020 were used as the test set. Table 3 gives a comparison of the prediction error evaluation metrics of the three methods on the training and test sets, from the table it can be seen that the MAE, RMSE and MAPE of the PCA dimensionality reduction treatment on the training set are: 14.29, 19.91 and 1.18%, respectively, and the test set are: 9.63, 12.06 and 0.02%, which are much smaller compared with the gray treatment and the method without any treatment. Therefore, the method after PCA dimensionality reduction is more superior to the other two methods.

4.3 Comparison of traditional prediction methods

In the forecasting studies of total health expenditure, more studies have used the GM (1,1) gray forecasting model and polynomial trend curve forecasting method. In order to compare the forecasting models established in this paper, this section uses MATLAB 2020b to construct the GM (1,1) model and the polynomial fitting model, and compares the results of both methods with the PCA-Elman results.

Since the gray forecasting model is less effective for long-term forecasting, only the data from 2010-2020 were selected for fitting analysis, and the fitted forecasting results for 10 years from 2011-2020 were compared with the forecasting results of the previous PCA treatment. The polynomial curve fitting model was also fitted and analyzed according to the same data, and the quadratic, cubic, quadratic, and quintuple polynomials were fitted using the MATLAB fitting toolbox, and the quadratic polynomial with the smallest RMSE value was finally selected for fitting. Equations 4 and 5 give the equations for the GM (1,1) and quadratic polynomials, respectively.

$\hat{X}^{(1)}(t+1) = 198967.7e^{0.1193t} - 178987.3$
$y_t = -5.5t^4 - 4.4\times10^4 t^3 - 1.3\times10^8 t^2 + 1.8\times10^{11} t - 9.1\times10^{13}$

Table 4 gives the comparison of the prediction results of the PCA-Elman model and the traditional model. The MAE, RMSE and MAPE values for each of the grey forecasts are: 719.17, 817.66 and 1.63% respectively, while the MAE, RMSE and MAPE values for the quadratic polynomial trend curve model are: 141, 198.1 and 0.39% respectively. Based on the comparison of the results obtained above, we can see that the indicators of the PCA-Elman model are smaller than those of the two traditional models, and the prediction accuracy is also much greater than that of the traditional model. In summary, it is more scientific and reasonable to use the PCA-Elman model to predict the change trend of total health expenditure in China.

Table 3 Comparison of prediction error evaluation indexes of the three methods in training set and test set

	Training Set			Test Set		
Model	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
PCA	19.91	14.29	1.18	12.06	9.63	0.02
Gray Processing	277.41	190.84	10.19	199.79	161.77	0.43
Unprocessed	380.02	281.17	34.13	467.66	384.40	1.09

Table 4: Comparison of prediction results of different models

Year	Actual	PCA-Elman	GM (1,1)	Polynomial Curve	Year	Actual	PCA-Elman	GM (1,1)	Polynomial Curve
2011	24345.91	24344.37	19980.39	24435.60	2016	46344.88	46343.22	45753.23126	46302.90
2012	28119	28116.63	25202.55916	27909.40	2017	52598.28	52578.83	51548.63819	52545.80
2013	31668.95	31670.5	28394.88201	31579.20	2018	59121.91	59142.35	58078.1297	59193.30
2014	35312.4	35328.64	31991.5656	35800.70	2019	65841.39	65855.78	65434.68979	65887.60
2015	40974.64	40965.65	36043.82892	40689.90	2020	72175	72165.39	73723.08045	72139.00

4.4 Total health expenditure forecast for the next 5 years

A comparison of the three machine learning models and the traditional prediction model shows that the PCA-Elman neural network model has the best prediction effect, and it can be concluded from the results of the error analysis that the model can make effective predictions for future data. Since the predictions of health expenditure above all require the principal component score values of the current year's impact indicator system, this paper still uses the Elman neural network model to predict the principal component scores of the impact factors on total health expenditure. The principal component scores of the previous three years were selected as the input to the network, and the principal component scores of the current year were used as the output, i.e., the input nodes of the network were 12 and the output nodes were 4. In this way, a total of 38 sets of data were constructed, and the data of sets 1-35 were used as the training set, and the last three sets of samples were used as the test set to test the model machine and perform error analysis. The parameters of the model are still set according to the previous paper, and the number of nodes in the hidden layer is determined to be 21 when the best effect is achieved through repeated trial and error training. Fig.4, 5 give the absolute error stacking plots of the four principal components in the training set and the absolute error values of the four principal components in the test set, respectively, from Fig. 4, the prediction error of principal component 2 is the largest compared with the other three, and the error distribution of principal components 1 and 4 is more similar. MAE= 0.04, RMSE=0.05 and MAPE = 9.6%, indicating that the training effect of the model is not bad. The prediction decision error results of the three test sets are given in Figure 5, from which it can be seen that the 1st principal component has the best prediction, the 2nd principal component has the largest error compared to the other three principal components, and the actual values of the four principal component scores in the test set are compared to the predicted values with MAE=0.03, RMSE=0.04, and MAPE=6.8%. This indicates that the model has high accuracy in the prediction of principal component scores.

Based on the above analysis, the model can be used to predict the subsequent principal component scores, then the principal component scores for 2018-2020 are used as input data and imported into the prediction model above to obtain the predicted principal component scores for 2021, then the real data for 2019-2020 and the predicted data for 2021 are used as input data and imported into the model to predict the principal component scores for 2022, and so on, until the predicted scores for 2025 are predicted.

The predicted principal component scores of the latter 5 years are imported into the PCA-Elman model, and the predicted output is the national total health expenditure in 2021-2025. Table 5 gives the predicted results of China's health expenditure in 2021-2025, and the predicted results indicate that China's total health expenditure is still growing in the next 5 years, but the annual growth rate is decreasing year by year, which indicates that the growth of health expenditure will gradually become stable, which is in line with the national policy of controlling the future health expenditure.

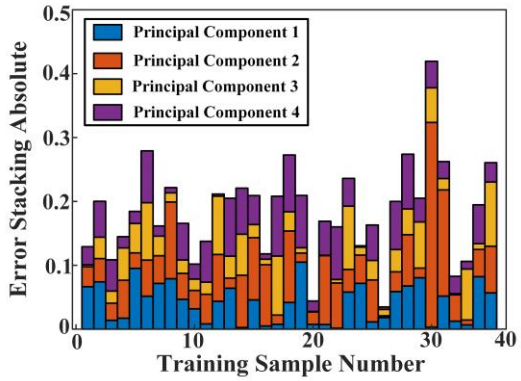

Fig.4. Stacking diagram of absolute error of training set

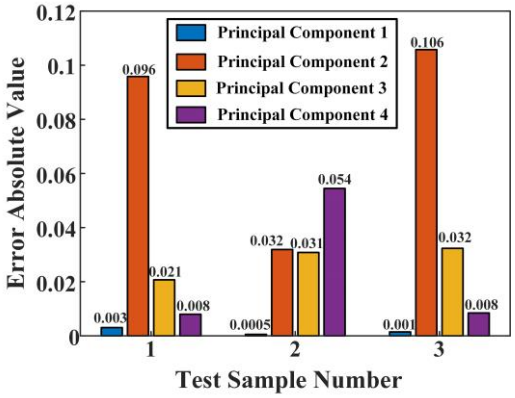

Fig.5. Absolute error of test set

Table 5: Projected total health expenditure in China over the next five years

Year	Predicted value of total health expenses (100 million yuan)	Annual growth rate (%)
2021	78663.60	8.99%
2022	83950.43	6.72%
2023	88748.01	5.71%
2024	93397.45	5.24%
2025	97974.29	4.90%

5 CONCLUSIONS

This paper takes China's total health expenditure as the research object, establishes a system of influential indicators that may affect China's health expenditure based on the theory of health demand-supply relationship and analysis of relevant literature, studies the degree of influence of influencing factors using gray correlation and principal component analysis (PCA), and uses machine learning models to forecast China's total health expenditure in the next five years, and draws the following conclusions:

(1) Both gray correlation and PCA analysis show that the influencing factors of total health expenditure in China are mainly reflected in economic income and medical services, specifically in GDP, hospital admissions, number of beds in medical and health institutions, and the absolute number of residents' consumption levels.

(2) The principal component score, the four impact indicators with high gray correlation and the 14 impact indicators without any treatment were used as input data and the total health expenditure was imported into the Elman neural network model as output data, respectively, and comparing the three machine learning models it was concluded that the principal component score model (PCA-Elman) had the best prediction effect.

(3) By comparing the PCA-Elman model and the traditional total health expenditure forecasting model, it is concluded that the effect of using machine learning method is better than the traditional method.

(4) The PCA-Elman model is used to forecast the total health expenditure in China in the next five years, which is 78663.60, 83950.43, 88,748.01, 93397.45, and 97974.29 billion yuan from 2021 to 2025, respectively. The forecast results show that the total expenditure of health in China in the next five years will continue to maintain the growth trend, but the growth rate will gradually stabilize.

ACKNOWLEDGMENT

Research on Humanities and Social Sciences in Jiangxi Universities (GL19136).

REFERNCES

[1] Dreger C, Reimers H E. Health Care Expenditures in OECD Countries: A Panel Unit Root and Cointegration Analysis[J]. IZA Discussion Papers, 2005, 2(2): 5-20.

[2] Getzen. Population aging and the growth of health expenditures[J]. Journal of Gerontology, 1992.

[3] He Sizhang. Research on Structural changes and forecast of total health expenditure in Sichuan [J] Modern Preventive Medicine, 2021, 48(23): 6.

[4] Jiang Yan. Research on the Development Stage of Total Health Expenditure of Beijing Based on Logistic Regression Model J]. Chinese Journal of Social Medicine 2021, 38(5): 3.

[5] Tang Shuyi. Prediction of Total Health Expenditure in China Based on PCA-BP Neural Network Model [D] Beijing Jiaotong University,2021 (Master Thesis)

[6] Wu W, Xinma, Zhang Y, et al. Analysis of novel FAGM (1, 1, t^α) model to forecast health

expenditure of China[J]. Grey systems: theory and application, 2019, 9(2): 232-250.

[7] Xu Rongfei. Tendency and prediction study of Total expenditure on health in China based on principal component analysis and neural network model [D] Shandong University,2013(Master Thesis)

[8] Zhang Fangfang. Trend prediction and composition analysis of total health expenditure in Guangdong based on ARMIMA model [J] Modern Preventive Medicine, 2019, 46(2): 5.

[9] Zhang Zhengzhong. China Health Cost Accounting Study [M] China Health Cost Accounting Research Report, 2009.