International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n3.47
Journal homepage: https://wepub.org/index.php/IJCSIT/index

An Empirical Analysis of Anhui Province's Quarterly GDP based on the LSTM and SARIMA Combined Model

Kai Chen

Anhui University of Finance and Economics School of Business Administration, Bengbu, China

ABSTRACT

This paper uses the data of Anhui Province's quarterly GDP from 2005 to 2022 to build different models for prediction. Two single models, LSTM neural network model and SARIMA model, are established for prediction respectively. On this basis, the combination weights are determined by particle swarm optimization algorithm. It is found that the mean square error (MSE) of the combination using PSO particle swarm optimization algorithm is 0.2236 and 7.0159 lower than that of the two single prediction models, respectively. That is, the proposed combination model based on natural selection particle swarm algorithm to determine the weights of LSTM neural network model and SARIMA model has better prediction effect and higher prediction accuracy. This combination prediction model can be considered to predict the future GDP data of Anhui Province.

KEYWORDS

LSTM neural network; SARIMA model; Combination prediction; Natural selection particle swarm algorithm

1. INTRODUCTION

China's GDP is composed of the GDP of 34 provincial administrative regions. Anhui Province is located in the hinterland of central China, with a large population. It borders Jiangsu, Zhejiang and Shanghai on the right. It is an important bridge and link for the coordinated development of eastern, central and western China. Its economic trend prediction research can reflect the development trend of China's total GDP from a certain perspective. Correctly predicting Anhui Province's GDP is not only conducive to judging Anhui Province's economic strength, but also provides an important basis for proposing constructive policies to promote Anhui Province's economic development, and even provides a certain reference value for the economic development of other provinces.

There are many theoretical studies and empirical analyses on GDP series, among which forecast analysis has a higher versatility. Forecasting is widely used in actual production activities, and a small increase in forecast accuracy will save considerable costs. GDP data comes from social life and usually contains potential characteristics. This paper constructs a model to construct GDP data forecasts, uses the theoretical foundations learned in mathematical statistics, machine learning methods, etc., uses charts to display and analyze the hidden laws in the studied data, and uses evaluation criteria such as mean square error to measure the effect of the forecast method. Artificial neural networks can independently extract valuable information for learning, simulate the development trend of the sequence, and obtain more accurate forecast results. The quarterly regional gross domestic product has obvious seasonality, and the seasonal sum autoregressive moving average model (SARIMA) is used to predict Anhui Province's quarterly GDP. Based on these considerations, this paper selects the LSTM neural network model and the SARIMA model in the forecast model of Anhui Province's GDP, and uses the weight allocation method based on the natural selection particle swarm algorithm to integrate the advantages of the two types of models to determine the combined forecast model to predict Anhui Province's GDP.

2. LITERATURE REVIEW

Different scholars adopt different methods to predict GDP. According to the number of prediction methods adopted, they can be roughly divided into single models and combined prediction models.

In the study of single model, Zhang Shuhong [1] et al. used the GDP data of Henan Province to establish an AR model. The results showed that the model fitting effect was good and could be used to predict the future GDP value of Henan Province. Hong Yan [2] et al. used the interpolation function of MATLAB's built-in toolbox to fit and predict GDP data and predicted the gross domestic product from 2015 to 2017. Zhang Jing [3] et al. used 58 GDP sample data of Gansu Province to establish a time series model. Based on the prior information of the model, two different Bayesian models and algorithms were derived. On this basis, the GDP values of Gansu Province from 2011 to 2015 were predicted. The study found that this type of Bayesian model has better prediction accuracy than traditional prediction methods. Tian Zichen [4] [17] et al. added an improved Lagrange interpolation function to optimize the GM (1,1) model based on the traditional grey model, solved the Runge phenomenon of the Lagrange difference, and used this optimized grey model to predict the GDP value of Xinjiang.

In the combined prediction model, Wang Shasha [5] et al. integrated the advantages of the mixed time series model, ARIMA model and GM (1,1) model, established a combined prediction model and applied it to the prediction of China's GDP data. Zhang Zhichao [6] et al. combined the advantages of three single prediction models and used the optimal linear combination prediction model to organically combine the single models. They found that this method can effectively predict the per capita GDP of Anhui Province. Xiong Zhibin [7] et al. combined the advantages of the ARIMA model in linear space and the advantages of the NN model in nonlinear space to form a combined model to predict GDP values. Long Huiden [8] et al. integrated the ARIMA model, GM (1,1) model and BP neural network model. The results showed that the prediction accuracy of the combination of the three single prediction models was significantly better. Mo Songjuan [9] et al. established a combined prediction model based on the IOWGA operator to combine multiple single models. The study found that the prediction effect of the combined model using the IOWGA operator was significantly higher than that of the three single prediction methods. Li Yanfei [10] et al. innovatively proposed a combined prediction model based on proximity and a new method to determine the combination coefficient to predict the GDP of Anhui Province in the next few years. Li Jia [11] et al. used the independent recurrent neural network (lndRNN) in deep learning to improve the accuracy of model prediction.

At present, in the field of GDP combination prediction research, the weight assignment method has the disadvantages of large computational complexity and poor combination prediction accuracy [16]. The swarm intelligence algorithm has become a popular algorithm for solving optimization problems due to its advantages such as high accuracy and wide application. However, the algorithm also has the disadvantage of being easily trapped in solving local optimal solutions. Therefore, this paper improves the weight and learning factor of the particle swarm algorithm based on natural selection, and uses the improved particle swarm algorithm based on natural selection to solve the weights of the LSTM neural network model and the SARIMA model to predict the GDP of Anhui Province, and conducts a combined prediction of the GDP of Anhui Province.

3. METHOD SELECTION, MODEL CONSTRUCTION AND DATA SOURCE

3.1. Anhui Province GDP Prediction Method and Model Construction

(1) LSTM neural network model

Like ordinary artificial neural network models, LSTM neural network models also capture data information through multiple layers of linear weighting and nonlinear activation function nesting. Its uniqueness lies in that LSTM neural network models add a gating mechanism based on the improved RNN model, enabling the model to better learn long-term information. However, the farther the time node in the past is from the present, the smaller the impact may be [14]. Therefore, the gate control mechanism of LSTM neural network can fully play its role, filter redundant information, and select effective prediction information that is helpful for prediction from historical data. Among them, is the parameter information newly introduced into the training process, such as the GDP of Anhui Province in each year. represents the interim result of the previous iteration process, the forget gate filters the information entering the activation function in, and is the learning result state of the previous period and the learning result state of this period. The function of the update gate is to update the state, and finally the output gate controls the effective output information. Through multiple iterations to correct the error, a better prediction model can be obtained. In the specific operation, the LSTM neural network model is used to predict the scrolling pane of Anhui Province's GDP from 1992 to 2022, and finally the estimated value of Anhui Province's GDP from 1993 to 2022 is inferred.

(2) SARIMA model

ARIMA model, namely, the autoregressive moving average model. It is a well-known random occurrence time series forecasting method, jointly proposed by American researcher Box and British statistician Jenkins in 1976. The ARIMA model regards the forecast object as a set of probability distributions with its own correlation structure characteristics. It uses the unknown time t to predict the relative stability of the future of the forecast object, makes full use of the past and present values to predict the future value of a variable of the object, and selects this statistical model to describe the autocorrelation of the time series. The function of the ARIMA model is defined as ARIMA(p,d,q)(P,D,Q)S, with relevant parameters p, P and q, Q represents the order of autoregression and moving average, d and D represent the difference time interval, and s represents the actual length of the continuous cycle. The specific steps of constructing the seasonal ARIMA model [18]: (1) Stationarity test: With the help of the difference and seasonal difference methods, the original sequence or the processed sequence must meet the stationarity requirements of ARIMA. (2) Determine model parameters: refer to the autocorrelation and partial autocorrelation diagrams of the time series that have passed the stability test, preliminarily determine the relevant parameters d, D, s, and then continue to combine the BIC related information in the basic principle to determine the p and q values corresponding to the minimum values of BIC. (3) Model diagnosis: mainly refer to the relevant parameters selected in (2) to build the model, judge whether the relevant parameters selected by the SARIMA model can meet the basic statistical indicators, and judge whether the autocorrelation and partial autocorrelation diagrams of the residuals are the characteristics of white noise sequences. (4) Prediction evaluation: use the model of the selected ideal solution to evaluate future numerical and time series trends, and compare the predicted values with the actual values to evaluate whether the effect of the SARIMA model is good.

(3) Weight allocation method based on natural selection particle swarm algorithm

In the basic particle swarm algorithm (PSO), it is assumed that the position and speed of the particle in the iteration are and respectively. The particle updates the position and speed by supervising the individual and population extreme values to further approach the optimal solution [13]. The update formula is:

$$v_{i,t} = w \times v_{i,t} + c_1 \times rand \times (pbest_i - x_{i,t}) + c_2 \times rand \times (gbest_i - x_{i,t}) \tag{1}$$

$$x_{i,t+1} = x_{i,t} + \lambda \times v_{i,t+1} \tag{2}$$

Where: w——weight, 0~1;

c_1 and c_2 ——learning factor, both in the range of 0~4;

$rand$ ——generate random numbers between 0~1;

$pbest$——individual extreme value;

$gbest$——group extreme value;

λ——speed coefficient, generally 1.

The particle swarm algorithm based on natural selection is based on the basic particle swarm algorithm. Its basic idea is to replace the worse performing half of the particles with the better half of the particles according to the calculated fitness value, and retain the historical best value of the particle memory.

(1) Predict the GDP of Anhui Province based on the selected single prediction model.

(2) The fitness function established based on the prediction results of Anhui Province's GDP. The fitness function is determined with the best combined prediction accuracy as the goal, that is, the absolute value of the relative prediction error of all prediction years is the smallest. Therefore, the fitness function is defined as:

$$f(w) = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|X_t - \sum_{i=1}^{m} q_i w_i\right|}{X_t} \tag{3}$$

Where: n——Forecast total year; X_t——Actual GDP value of period t.

(3) Initialize population particle parameters. The initial parameter values should be set according to the value range of particle parameters.

(4) Calculate particle fitness.

(5) Compare the fitness values of all particles, find the current best particle fitness value and its position, and update the search position and speed of the particle.

(6) Sort the particles according to the principle of natural selection, and replace the worst half of the particles with the best half of the particles in the group.

(7) Determine whether the termination condition of the iteration is met. If it is met, the algorithm ends; if it is not met, return to step 4.

(8) Output the optimal solution.

3.2. Data Source and Processing

The data in this article comes from the Anhui Statistical Yearbook published by the Anhui Provincial Bureau of Statistics, which contains the quarterly GDP data of Anhui Province from 2005 to 2022. When training the LSTM neural network model, the data is normalized before training, and then the Adam optimizer is selected for training. After training, the inverse process is performed, and the final result is outputted by denormalization.

4. EMPIRICAL RESULTS AND ANALYSIS

4.1. LSTM Neural Network Model Prediction Analysis

First, the GDP time series data is normalized and mapped to the [0, 1] interval. The LSTM neural network model processes the time series by inputting the data of the previous periods as feature variables into the model training to predict the number of future periods. We should transform the collected data to transform unsupervised learning into supervised learning, and then solve the problem. The data used in this paper are the GDP of Anhui Province in four quarters of each year, so the step size is 1, that is, 1x, 2x, 3x and 4x are used to predict 5x, and 2x, 3x, 4x and 5x are used to predict 6x, ..., until the GDP value of 2022 is predicted [15]. The LSTM model in this paper sets a single hidden layer, the number of hidden layer nodes is set to 50, tanh is used as the activation function, Adam is used as the optimizer, the batch size is 4, the number of iterations is 500, and the mean absolute error (MAE) is used as the loss function to measure the prediction accuracy of the model, so as to train the training set. The loss rate after training is shown in Figure 1. The training loss eventually stabilizes at about 0.2, and the effective loss eventually stabilizes at about 0. After training, the values are denormalized to obtain the final prediction sequence. The actual value is compared with the fitted value to obtain the fitting situation shown in Figure 2. As shown in Figure 2, the true value and the predicted value have the same trend and are not much different. Table 3 lists the MAE, MSE, RMSE, MAPE and MSPE of the LSTM neural network model in the GDP time series. The mean absolute percentage error (MAPE) of the final training model is as low as 0.2587%, which shows that the LSTM model has a good fitting effect.

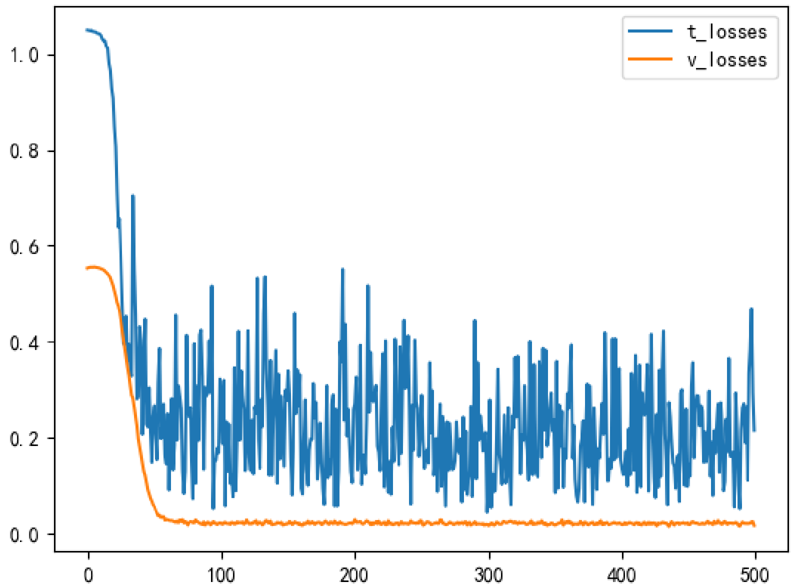

Figure 1. Loss rate of LSTM neural network model

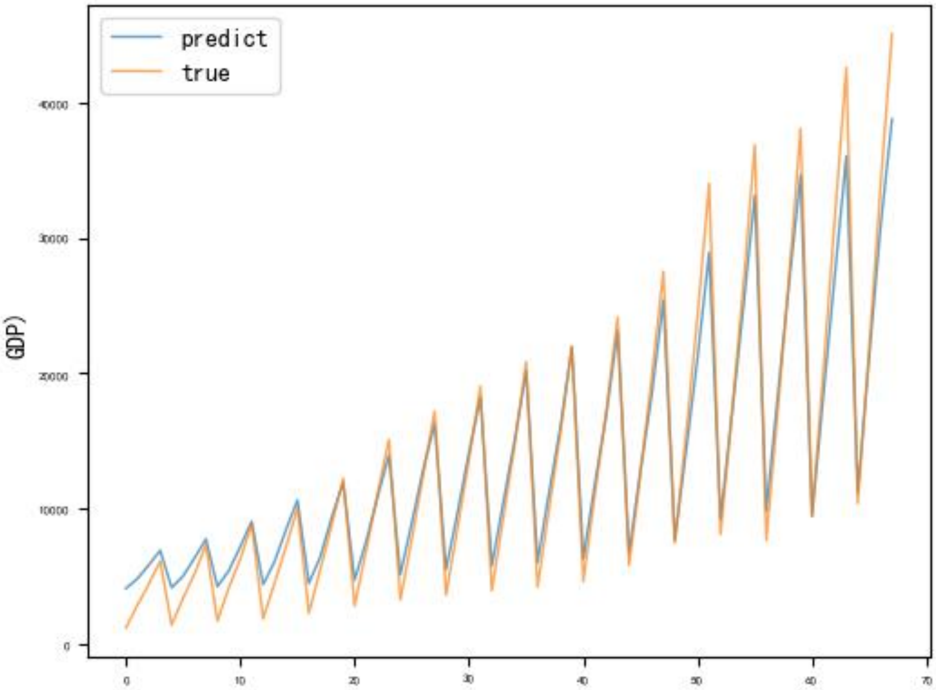

Figure 2. LSTM neural network model prediction of Anhui Province’s GDP

The model fitted by the training set was verified using the test set. The results are shown in Table 1 below:

Table 1. Prediction effect

Mean absolute error (MAE)	Mean square error (MSE)	Root mean square error (RMSE)	Mean absolute percentage error (MAPE)	Mean Square Percent Error (MSPE)
1515.5168	4074843.0	2018.6240	0.2587	25.8001

4.2. SARIMA Model Prediction Analysis

First, the ADF test is performed on the time series data of Anhui Province’s quarterly GDP. The P values of the original sequence and the sequence after first difference are both greater than 0.1, indicating that the sequence at this time is still an unstable sequence. For the second-order difference sequence based on the variable GDP, the significance P value is 0.000***, showing significance at the level, rejecting the null hypothesis, and the sequence is a stationary time series.

Table 2. ADF inspection table

variable	sequence	t	P	AIC	Critical value		
					1%	5%	10%
GDP	Original sequence	-0.068	0.953	953.96	-3.546	-2.912	-2.594
	1st order difference	-1.612	0.477	937.257	-3.546	-2.912	-2.594
	2nd order difference	-6.046	0.000***	938.265	-3.546	-2.912	-2.594

Note: ***, **, and * represent the significance levels of 1%, 5%, and 10% respectively.

After performing the second-order difference on Anhui Province's quarterly GDP, we observed the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF). The autocorrelation coefficient has tailing properties, and the partial autocorrelation coefficients both show third-order censoring. Therefore, it is assumed that the ARIMA (3,2,0) model is more suitable for fitting and prediction.

This article selects the ARIMA (3,2,0) × (0,1,1)4 seasonal additive model to model the quarterly GDP data of Anhui Province and perform significance testing, as shown in Table 3.

As can be seen from Table 3, the P values of the ARIMA (3,2,0) × (0,1,1)4 model are all greater than 0.05, the significance passes, and the fitting effect is good.

Table 3. Parameters of the ARIMA (3,2,0) × (0,1,1)4 model

		Estimate	Standard error	t	Significance
constant		-0.884	6.968	-0.127	0.899
AR	Delay 1	-1.006	0.017	-59.336	0.000
	Delay 2	-1.005	0.020	-49.509	0.000
	Delay 3	-0.995	0.020	-49.580	0.000
MA, Seasonality	Delay 1	0.830	0.157	5.289	0.000

The ARIMA (3,2,0) × (0,1,1)4 model is used to predict the quarterly GDP data of Anhui Province. Figure 3 shows that the data results have a relatively small error and the true value is within the confidence interval.

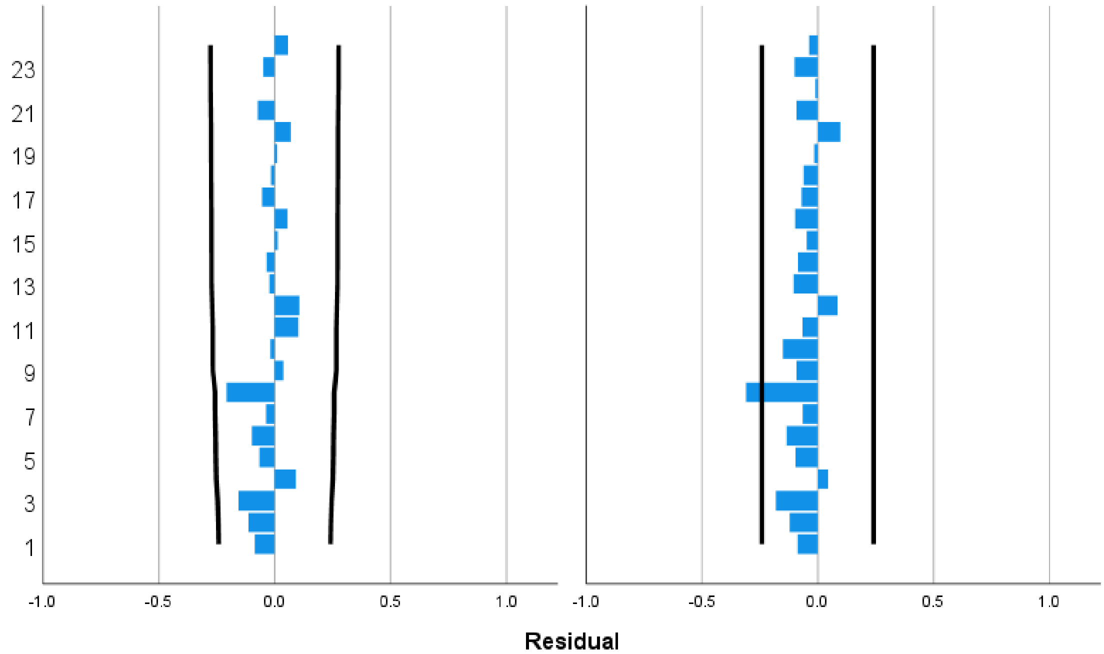

Figure 3. Residual autocorrelation and non-autocorrelation plots

The errors of the SARIMA model are shown in Table 4.

Table 4. Prediction results

Mean absolute error (MAE)	Root mean square error (RMSE)	Mean absolute percentage error (MAPE)
463.015	867.129	7.051

4.3. Combined Forecast Based on Natural Selection Particle Swarm (PSO)

The LSTM neural network and ARIMA (3,2,0) × (0,1,1)4 are selected to make a combined forecast of Anhui Province's quarterly GDP. The combined forecast formula is:

$$Q_z = w_1 q_{lstm} + w_2 q_{gm} \tag{4}$$

Where: w_1, w_2——represents the prediction weights of LSTM neural network and SARIMA model respectively;

q_{lstm}, q_{qm}——represents the prediction values of LSTM neural network and SARIMA model respectively;

Q_z——represents the combined prediction value of LSTM neural network and SARIMA model.

Based on the weight constraint $w_1 + w_2 = 1$, formula (9) can be simplified as:

$$Q_z = w_1 q_{lstm} + (1 - w_1) q_{gm}$$

Therefore, the particle swarm algorithm based on natural selection only needs to solve. The algorithm experiment in this section is carried out in the python3.10 environment. The particle swarm algorithm code based on natural selection is written to find the weight with the smallest absolute value of the combined prediction error in the test sample. The initialization settings of the main parameters of the PSO algorithm are as follows: the population size (sizepop) is 50; the maximum number of iterations (maxgen) is 500 times; the speed (rangespeed) search interval is -0.1 to 0.1. The final combined prediction effect is shown in Table 5.

Table 5. Prediction effect of PSO algorithm

Mean absolute error (MAE)	Mean square error (MSE)	Root mean square error (RMSE)	Mean absolute percentage error (MAPE)	Mean Square Percent Error (MSPE)
296.9946	181271.0737	425.7594	0.0351	0.3582

To further verify the combined prediction effect of the weights solved by the particle swarm algorithm based on natural selection, it is compared with the prediction effect of a single prediction model. The prediction results of the test samples of different prediction methods are shown in Table 6.

It can be seen from Table 6 that the weights solved by the improved particle swarm algorithm based on natural selection have better combined prediction effects, which are specifically reflected in: compared with the single prediction model, the combined prediction accuracy of the improved particle swarm algorithm based on natural selection is higher, with a mean absolute percentage error (MAPE) of 0.0351, while the root mean square errors of the LSTM neural network and SARIMA model are 0.2587 and 7.051 respectively. In comparison, the mean square error of the combination of the PSO particle swarm optimization algorithm is 0.2236 and 7.0159 lower than the first two single prediction models, respectively. The mean absolute error (MAE) and root mean square error (RMSE) of the PSO particle swarm optimization algorithm combination is also smaller than the values of the single prediction model.

Table 6. Comparison of prediction effects of different models

Model	Mean absolute error (MAE)	Root mean square error (RMSE)	Mean absolute percentage error (MAPE)
LSTM model	1515.5168	2018.6240	0.2587
SARIMA model	463.015	867.129	7.051
PSO particle swarm optimization algorithm combination	296.9946	425.7594	0.0351

5. CONCLUSION

This article uses the quarterly GDP data of Anhui Province from 2005 to 2022 to build different models for prediction. It is found that the prediction effects of using the LSTM neural network model and the SARIMA model are not much different. Drawing on the literature, it is found that most of the combined models have better prediction effects than a single model. Therefore, a weight determination method based on natural selection particle swarm was introduced to combine two types of models to predict GDP. It was found that this combination model can effectively extract information in time series and achieve better prediction results. The obvious shortcoming is the lack of improvements to particle swarm optimization. Future research can consider setting up better algorithms to iterate the positions of particles to obtain the optimal solution.

REFERENCES

[1] Zhang Shuhong, Yang Wancai, Wu Xingan. Forecast of per capita GDP in Henan Province during the "Twelfth Five-Year Plan" period [J]. Mathematical Statistics and Management, 2014, 33(03): 394-399. DOI: 10.13860/j.cnki.sltj. 2014.03.003.

[2] Hong Yan, Liu Bingxiang, Wan Xiang. A GDP forecasting method based on interpolation fitting [J]. China Management Informatization, 2016, 19(16): 123-124.

[3] Zhang Jing. Forecast of per capita GDP in Gansu Province based on Bayesian time series model [J]. Mathematical Statistics and Management, 2018, 37(02):205-210. DOI:10.13860/j.cnki.sltj.20170721- 001.

[4] Tian Zichen, Liu Miao. Empirical GDP forecasting based on the improved gray GM(1,1) model[J]. Statistics and Decision Making, 2018, 34(11):83-85. DOI:10.13546/j.cnki.tjyjc .2018.11.019.

[5] Wang Shasha, Chen An, Su Jing, et al. Application of combination forecast model in China's GDP forecast [J]. Journal of Shandong University (Science Edition), 2009, 44(02): 56-59.

[6] Zhang Zhichao. Application of combination forecast model in per capita GDP forecast of Anhui Province [J]. Journal of Guangxi University of Finance and Economics, 2009, 22(05): 32-35.

[7] Xiong Zhibin. Research on GDP time series forecasting based on the integration of ARIMA and neural network[J]. Mathematical Statistics and Management, 2011, 30(02):306-314. DOI:10.13860/j.cnki.sltj.2011.02.012 .

[8] Long Huidian, Yan Guangle. Research on GDP time series forecasting based on SARIMA, GM(1,1) and BP neural network integrated model [J]. Mathematical Statistics and Management, 2013, 32(05): 814-822. DOI:10.13860/j.cnki.sltj.2013.05.004.

[9] Mo Songjuan, Yang Guiyuan, Luo Yang. Combination forecast of China's total GDP based on IOWGA operator [J]. Statistics and Decision Making, 2015, No. 436(16): 4-7. DOI: 10.13546/j. cnki.tjyjc.2015.16.001.

[10] Li Yanfei, Wu Tao, Guo Haiyan. Anhui Province per capita GDP combination forecast model based on proximity [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2019, 36(04):95-100. DOI:10.16055/j .issn.1672-058X.2019.0004.016.

[11] Li Jia, Huang Zhihao, Wang Jiahui. GDP forecast based on independent recurrent neural network method [J]. Statistics and Decision Making, 2020, 36(14): 24-28. DOI: 10.13546/j.cnki.tjyjc.2020.14. 005.

[12] Liu Sifeng. The emergence and development of gray system theory [J]. Journal of Nanjing University of Aeronautics and Astronautics, 2004(02):267-272. DOI:10.16356/j.1005-2615.2004.02.027.

[13] Li Wan, Feng Fenling, Jiang Qiwei. Railway passenger volume prediction using improved particle swarm algorithm to optimize LSTM neural network [J]. Transactions of Railway Science and Engineering, 2018, 15(12): 3274-3280. DOI: 10.19713/j. cnki.43-1423/u.2018.12.033.

[14] Hu Jianbo, Luo Zhipeng, Li Feng. Forecasting China’s carbon emission intensity under the “carbon peak” goal—analysis based on LSTM and ARIMA-BP models [J]. Financial Science, 2022, No. 407(02):89- 101.

[15] Gu Xin. China's economic development forecast based on ARIMA-LSTM-XGBoost combination model [D]. Qingdao University, 2021. DOI:10.27262/d.cnki.gqdau.2021.001857.

[16] Ma Jingwen, Li Shuqing, Xia Mengyao. Review of research on the application of machine learning in GDP forecasting [J]. Science and Technology Information Research, 2022, 4(03):73-94. DOI:10.19809/j.cnki.kjqbyj.2022.03. 008.

[17] Wang Jiawei, Liu Miao, Tian Zichen. Empirical analysis and prediction of Xinjiang’s GDP based on improved gray model [J]. Practice and Understanding of Mathematics, 2019, 49(17): 306-312.

[18] Chen Ting. Comparative study on the application of ARIMA model and BP neural network model in AIDS incidence prediction [D]. Guangxi Medical University, 2015.

[19] Li Yongdi. Forecast of total GDP and growth rate based on ARIMA, deflator method and principal component regression model - taking Henan Province as an example [J]. Henan Science and Technology, 2021, 40(13): 149-155.