ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2821-2975 (Online)
CODEN: EGESCS

Economic Growth and Environment Sustainability (EGNES)

DOI: http://doi.org/10.26480/egnes.01.2022.01.05

RESEARCH ARTICLE

PREDICTION OF REAL ESTATE INVESTMENT IN ZHEJIANG PROVINCE BASED ON SARIMA MODEL

Liu Bingjie*, Ye Shanli

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China.
**Corresponding Author E-mail: 19857026394@163.com*

ARTICLE DETAILS

Article History:

Received 18 November 2021
Accepted 21 December 2021
Available online 23 December 2021

ABSTRACT

Taking the monthly flow data of real estate investment in Zhejiang Province from 2000 to 2020 as the research sample models of SARIMA, LSTM and Prophet are established based on the contributing factors such as the time trend, seasonal cycle, emergencies to predict the real estate investment in Zhejiang Province. And then, the predictability of the three different models are analyzed by comparing the values of RMSE, MAE and MAPE. The results show that the SARIMA has better performance in predictability than that of LSTM and Prophet Model.

KEYWORDS

Real estate investment, Box-cox conversion, SARIMA model, LSTM model, Prophet model

1. INTRODUCTION

Real estate investment is not only an important indicator of a country's continuous social and economic development, but also one of the main driving forces of economic growth, the amount of which is largely dependent on the local social and economic growth. Accurate prediction of real estate investment can serve the purpose of strengthening urban construction by effectively helping the government, real estate developers and other relevant departments to formulate investment measures of urban development in time. Many scholars have made in-depth research and analysis of real estate investment. At present, the main methods used for real estate investment are description analysis, factor analysis, support vector machine (SVM) learning method, gray model, traditional ARIMA model and so on.

Among those scholars, Li Ying put forward the "inverted U" phenomenon based on the situation of real estate investment in 2019, and specifically analyzed the real estate investment in the first quarter in such aspects as the current investment situation, and how real estate investment steers economy from the perspective of descriptive analysis (Li, 2020). Her analysis concludes that the government should make use of the "inverted U" relationship between real estate and macro-economy to enhance their strengths and avoid weaknesses, so as to make industrial restructuring. Researcher Lei Lei uses factor analysis to delve into the relevant data of real estate investment by two methods of regression prediction and time series prediction to predict the future real estate investment in China (Lei and Chen, 2020).

In 2009, Zhao Yong established the risk index system of the whole process of investment, and used the excellent nonlinear regression fitting characteristics of SVM to establish a prediction model to predict the risk of the whole process of real estate investment (Zhao, 2009). Another scholar Xu Tao uses the grey model to adjust the growth rate of the sample data by weakening the buffer sequence operator and proves that the grey model has a good accuracy in studying the prediction of real estate development investment in Chengdu (Xu and Zhao, 2011). It can provide a theoretical basis for the decision-making of the government, investors, consumers and so on. Zhang Qing Jun and Zhang Li through the traditional ARIMA model, this paper makes a short-term prediction of the amount of real estate investment, studies the periodic changes of real estate development investment in China, and makes an empirical analysis, which reveals the development trajectory, principle mechanism and periodic law of real estate investment in the process of real estate development in China (Jun and Li, 2009).

In 2016, Lu Shanghui applied the sparse coefficient seasonal model to analyze and forecast the real estate investment in Nanning (Lu, 2016). These studies all analyze the real estate investment, but most of the studies do not systematically predict the real estate investment, and the quantitative analysis of the prediction of real estate investment is limited. This paper selects the monthly flow data of real estate investment in Zhejiang Province from 2000 to 2020 (data source: National Bureau of Statistics), and uses SARIMA, LSTM and Prophet model to analyze and forecast the real estate investment data in Zhejiang Province. RMSE, MAE and MAPE are adopted to predict and evaluate the index, indicating the prediction performance of the three models, to select the best prediction model and a more scientific prediction method on the basis of comparative analysis, and explain its development law, all of which provide theoretical support for the future development of real estate investment in Zhejiang Province.

2. BASIC THEORY OF THE THREE MODELS

2.1 SARIMA model

ARIMA model is a prediction method of time series proposed by Box and Jenkins in the early 1970s (Peng and Zhang, 2016; Zhang, 2003). The operation of ARIMA model is essentially the transformation from ARIMA model to ARMA model through difference operation, the transforming process of non-stationary sequence to stationary by multiple differences. And then the stationary sequence can be fitted by ARMA model. Besides, ARMA model has good abilities not only in predicting the stationary time series, but also in analyzing the stationary series obtained by the difference (Wang, 2015).

Quick Response Code	Access this article online	
	Website: www.egnes.com.my	DOI: 10.26480/egnes.01.2022.01.05

Cite The Article: Liu Bingjie, Ye Shanli (2022). Prediction of Real Estate Investment in Zhejiang Province Based on Sarima Model. *Economic Growth and Environment Sustainability, 1(1): 01-05.*

In the prediction of time series data with seasonal and periodic trends, SARIMA model has higher accuracy than ARIMA model (Wang et al., 2020; Zhao et al., 2020). The SARIMA (*p, d, q*) (*P, D, Q) S* model can be expressed as the following structure:

$$\Phi(B)U(B^S)\nabla^d\nabla_S^D X_t = \Theta(B)V(B^S)\varepsilon_t \tag{1}$$

Where:

$$\begin{cases} \nabla^d = (1-B)^d, \nabla_S^D = (1-B^S)^D \\ U(B^S) = 1 - \Gamma_1 B^S - \Gamma_2 B^{2S} - \cdots - \Gamma_P B^{PS} \\ V(B^S) = 1 - H_1 B^S - H_2 B^{2S} - \cdots - H_Q B^{QS} \\ \Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_P B^P \\ \Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_P B^P \end{cases} \tag{2}$$

In formula (2): P is the auto-regressive operator of the seasonal part, Q is the moving average operator, S is the seasonal period, D is the order of seasonal difference, $\Phi(B)$ is the regression coefficient polynomial of the stationary series of ARMA(p, q) model,$\theta(B)$is the moving smoothing coefficient polynomial of the stationary series ARMA(p, q) model,$U(B^S)$ is the seasonal autoregressive polynomial, $V(B^S)$ is the seasonal moving average polynomial.

2.2 LSTM model

LSTM is a form of traditional Recurrent neural networks (RNN). The traditional cyclic neural network learns the timing of data by gaining hidden information in the hidden layer and updates the parameters through back propagation. However, the location interval of the information cannot be determined. As for the relevant information with a large distance from the current location, RNN will lose its learning ability. On the contrary, LSTM can maintain a good memory for time series data with a long span and has the ability to remove or add information to the cell state through the "gate" structure (Luo et al., 2019). In essence, the work of RNN can also be done by LSTM, moreover, LSTM demonstrates better performance in most tasks. The LSTM model contains three "gate" structures: the forget gate, the input gate and the output gate (Fu, 2020; Shen et al., 2020; Zhang et al., 2019; Hochreiter and Schmidhuber, 1997; Morana, 2001; Husken and Stagge, 2003).

The LSTM structure is shown in Figure 1.

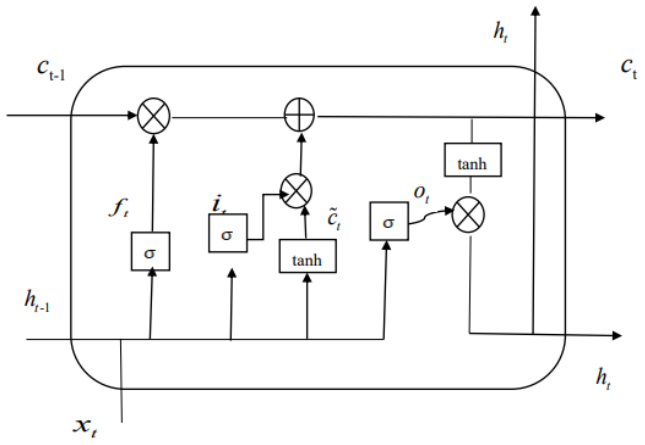


Figure 1: LSTM structure

The specific formula of the LSTM is as follows (Li and Zhang, 2020):

$$f_t = \text{sigmoid}\left(w_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$i_t = \text{sigmoid}\left(w_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{4}$$

$$o_t = \text{sigmoid}\left(w_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{5}$$

$$\tilde{c}_t = \tanh\left(w_c \cdot [h_{t-1}, x_t] + b_c\right) \tag{6}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{7}$$

In Formula (3) ~ Formula (7): $i_t$is the input gate,$o_t$is the output gate,$f_t$is the forget gate,$\tilde{c}_t$is the input value in Cell at t,$c_t$is the updated value in Cell at t, vector$h_t$stores the hidden information at t time and before, the activation function is sigmoid,tanh.w_f,w_i,w_o,$w_c$is the weight matric for f_t, i_t,o_t,$\tilde{c}_t$respectively and b_f,b_i,b_o,$b_c$is the offset of w_f,w_i;w_o,w_c respectively.

From the structure diagram and specific formula of the LSTM model, we can see that LSTM controls the flow of data information through the forget gate, which determines the amount of the previous information can be retained until now, and then it fuses the current information with the long-term stored memory information to form new storage information. The forgotten information can reduce the data redundancy of the time dimension, providing a good prediction method for data with time sequence (Zhang and Chui, 2020).

2.3 Prophet model

Prophet model is a time series prediction model proposed by Taylor et al in 2017. The model can more effectively analyze the characteristic information of the tested data and the law of social development and has a very good prediction performance (Wang et al., 2020). The core of Prophet model is to analyze the time series characteristics of data, such as periodicity, trend, holiday effect and so on.

The structure of the Prophet model[0]is as follows:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{8}$$

In formula (8): trend change function $g(t)$, to describe aperiodic changes in time series, seasonal periodic variation term $s(t)$, usually in years or weeks, holiday term $h(t)$, the changes caused by holiday factors, error term ε_t, the impact of unexpected events, usually modeled as Gaussian noise.

2.4 Evaluation index verification

The average absolute error MAE, root mean square error RMSE and average absolute percentage error MAPE are selected to test the prediction performance of SARIMA model, LSTM model and Prophet model for real estate investment (Wang et al., 2020; Yang et al., 2020). The formula is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x - x\right| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|x - x\right|^2} \tag{10}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x - x}{x}\right| \tag{11}$$

Formula (9) ~ formula (11): x is the actual value of the local real estate investment in the current month, x is the predicted value of the model, n is the total number of months predicted.

MAE, RMSE and MAPE are all important prediction and evaluation indexes. MAPE is a statistical index commonly used to measure the prediction accuracy, which is a percentage value. The value is inversely related to the prediction accuracy. The smaller the index values of MAE, RMSE and MAPE are, the better the predictive ability of the model is.

3. Empirical Analysis

3.1 Data selection and processing

3.1.1 Data selection

The data selected in this paper are from the monthly cumulative real estate investment data of the National Bureau of Statistics in Zhejiang Province from 2000 to 2020. In the process of data processing and analysis, the monthly real estate investment data is obtained by dislocation subtraction. It is necessary to preprocess the original data, because of the lack of January data in the original data will lead to low training accuracy of the model.

3.1.2 Data preprocessing

In terms of the investment data in January that are missing every year from 2000 to 2020, this paper adopts the same attribute mean substitution method, that is, calculating the monthly investment amount of other known objects with similar attributes (the investment amount from February to December of the previous year) and taking the average value to fill the missing January data. The formula for finding the mean is formula (12):

$$x = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (12)$$

After the data are preprocessed, the required experimental data are obtained. In order to analyze and forecast the real estate investment data more accurately, we train the real estate investment data from 2000 to 2019 as historical data and compare the 2020 investment data with the predicted data to get the optimal model, which is used to predict the real estate investment from February to June in 2021.

3.2 Model setting

According to the data of investment from 2000 to 2020, the trend chart of investment with time and the decomposition chart of time series are obtained by using python3.7, as shown in Figure 2:

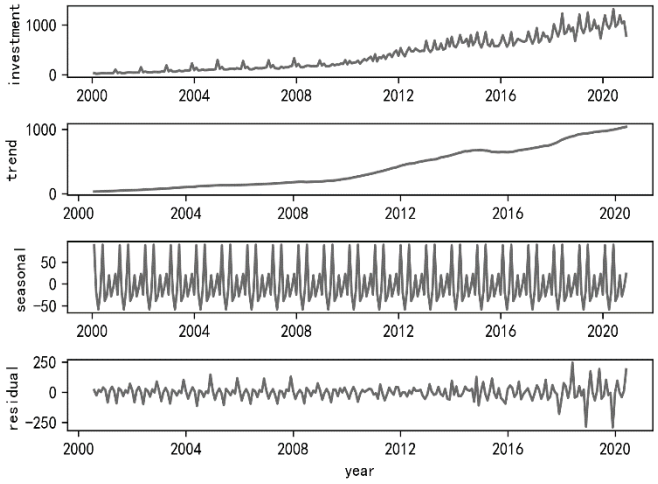


Figure 2: Data trend and decomposition chart

As can be seen from Figure 2, the overall trend of real estate investment data from 2000 to 2020 has an obvious upward or downward trend year by year, and is affected by certain annual periodicity, random factors and so on. In order to standardize the data, it is necessary to transform the experimental data by Box-cox (Li et al., 2008). Through the transformation of variables, Box-cox makes each component of the transformed error equal variance and independent of each other, and makes the residual obey the normal distribution in which the mean lag order is zero and the correlation coefficient is zero. In order to use the SARIMA model to predict the linear part, we first need to carry on the unit root test to the transformed real estate investment data. At the significance level of 1%, the transformed data and the data with the first-order difference are non-stationary. Based on the data after the first-order difference, the seasonal difference is carried out, and the *P* value of the ADF test is 0.001, less than 1%. Through the ADF test, the sequence after the difference is stable. Therefore, *d*=1, *S*=12, and the specific results are shown in Table 1.

Table 1: Test results of unit root

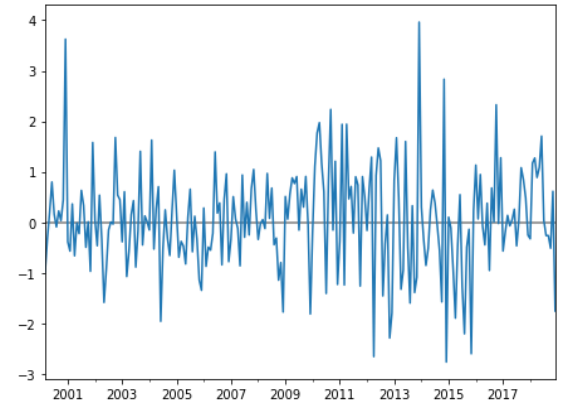
Sequence type	ADF test P value	conclusion
Sequence after Box-cox transformation	0.845	Non-stationary
First difference	0.053	Non-stationary
First difference and seasonal difference	0.001	Stationary

The order of *p, q, P and Q* is determined through the Auto-arima function according to the ACF and PACF diagram after the difference. The results are shown in Table 2.

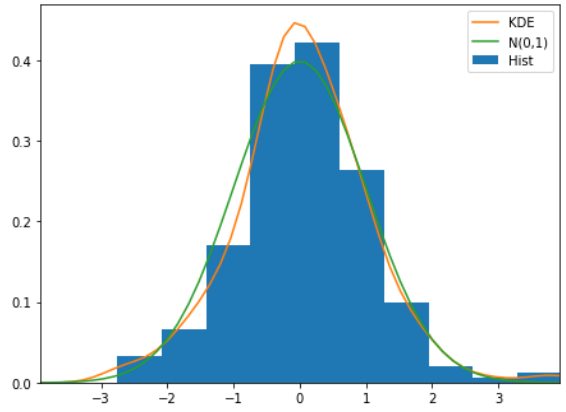
Table 2: Test results of each selection model

SARIMA (p, d, q) (P, D, Q) S	AIC
SARIMA (0, 1, 2) (2, 0, 1)12	192.74
SARIMA (1, 1, 1) (1, 0, 1)12	162.85
SARIMA (0, 1, 2) (2, 0, 2)12	189.860
SARIMA (2, 1, 1) (1, 1, 2)12	162.32

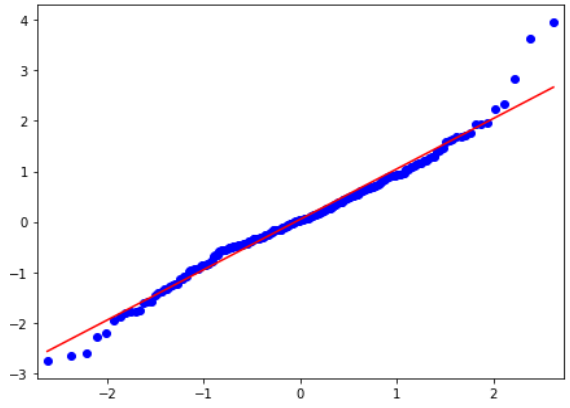
In the above table, the minimum AIC criterion is adopted, and the best model parameters are determined to be *p=2, q=1, P=1, Q=2.* Finally, SARIMA (2, 1, 1) (1, 1, 2)12 model is chosen to predict the future. The white noise test of the model shows that the corresponding *P* value of the LB test is 3.28204051e-13, less than 1%, then the differential sequence is not a white noise sequence, and there is a certain correlation between the sequences. A good SARIMA model assumes that the mean residual is 0, and the variance is constant, or the residual follows the white noise process. The sequence residual is tested, and the result is shown in Figure 3:



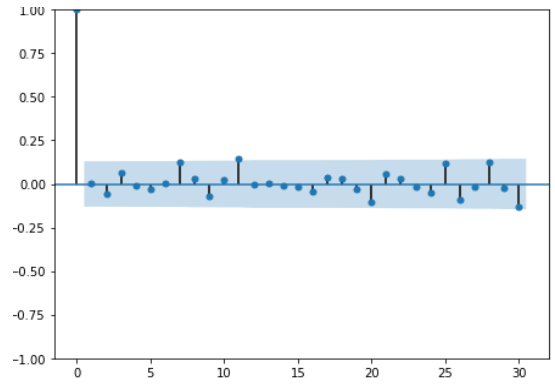
(a) Standardized residual for 'i'



(b) Histogram plus estimated density



(c)Normal Q-Q



(d) correlogram

Figure 3: Residual diagram of model test

As can be seen from Figure 3, the residual obeys a standard normal distribution. Figure 3(a) shows the residual diagram changing with time,

indicating that there is no obvious seasonality of the residual. Fig.3(b) presents that the red curve represented by the kernel density estimation (KED) roughly hinders the green curve represented by N (0,1). In the Fig.3(c), the residual diagram almost follows the linear trend except for a small number of points far away from the straight line. Fig.4(d) illustrates the auto-correlation graph between the data residual and the lagging data, from which we can see that the data is getting smaller and smaller, and the residual is white noise series, indicating a successful modeling.

3.3 Modeling of LSTM

The LSTM model mainly adopts the data information of the past few days to predict the next few days. The main processing of the prediction is as follows:

1) Data normalization processing. in order to avoid the network prediction error caused by the difference between the investment data. This paper adopts the maximum-minimum method to convert the data into a value between 0 and 1.

2) The modeling of LSTM. this paper uses the most important model of Keras-Sequential model, in which the iterative epochs of parameter model is set to 20, the number of neurons to 600, the batch size of training input into neural network to 1, the loss function to mean_squared_error, and the optimizer is set to adam.

3.4 Modeling of Prophet

Before Prophet model is adopted to make data prediction, it is necessary to adjust the preprocessed data in line with Prophet data to two fixed columns: *ds, y,* respectively referred to date and the volume of real estate investment. In order to improve the prediction performance of the Prophet model, its model parameters are further set: the interval_width is set to 0.95, the periods to 12, the prediction unit (fre) to *M*, and seasonality-mode= 'multiplicative' to model multiplicative seasonality.

3.5 Analysis of Model result

According to the buildup and parameter setting of the three models mentioned above, the seasonal ARIMA model, LSTM model and Prophet model of Box-cox transformation are used to process the data from 2000 to 2019, and the comparison between the predicted value and the actual value in 2020 is shown in Figure 4.

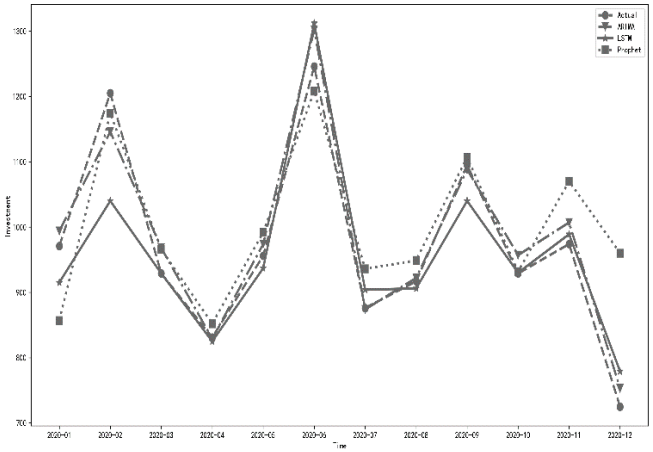


Figure 4: Comparison of the prediction performance of the three models

It can be seen from the above 12-month forecast chart of the three models in 2020 that all the three models can well predict the changing trend of real estate investment in 2020, when the investment peaked in July, higher than that of the same period. Compared with Prophet and LSTM, the overall prediction curve fluctuation of SARIMA model is closer to the actual curve fluctuation, demonstrating better overall prediction performance, thus is more suitable for the matching of investment data with trending and periodic changes. The prediction performance of SARIMA, LSTM and Prophet model is compared based on the results of three prediction evaluation indexes: MAE, RMSE and MAPE, so as to further determine the prediction performance of the three models. The comparison results are shown in Table 3.

Table 3: Comparison of model prediction performance			
MODELS	RMSE	MAE	MAPE (%)
SARIMA	25.32	24.17	2.36
LSTM	40.48	34.58	3.42
Prophet	67.53	51.55	5.80

As can be seen from Table 3, RMSE, MAE, and MAPE, the three evaluation index values of SARIMA, is 25.32, 24.17 and 2.36% respectively, all of which are lower than those of LSTM model and Prophet model. It shows that the prediction performance of ARIMA model is the best, indicating the seasonal ARIMA model has a better predictive ability in the real estate investment in Zhejiang Province. Therefore, SARIMA model has higher reliability in predicting the investment from February to June in 2021.

Based on the SARIMA model, the real estate investment in Zhejiang Province from February to June in 2021 is predicted. The results are shown in Table 4.

Table 4: Forecast of real estate investment in 2021					
Time	2021.2	2021.3	2021.4	2021.5	2021.6
Investment	1282	2384	3352	4455	5558

It can be seen from Table 4 that the real estate investment in Zhejiang Province will rise slightly from February to June in 2021, but on the whole, the fluctuation is stable.

4. Conclusion

This paper adopts the seasonal ARIMA model, LSTM model and Prophet model of Box-cox transformation to study and predict the amount of real estate investment in Zhejiang Province. By comparing and analyzing the index value of prediction evaluation, the following conclusions can be drawn:

1) In terms of the accuracy of real estate investment prediction in Zhejiang Province, the model of SARIMA performs the best, followed by LSTM model, and Prophet model is the worst. Prophet model performs better in the prediction of time series with obvious periodicity. It was originally designed to predict daily data, and most of the algorithmic optimizations are also aimed at daily forecasting. But for monthly data, the prediction effect of Prophet is not very good. The prediction results of LSTM are largely affected by the amount of data, data training and parameter adjustment.

2) The real estate investment in Zhejiang Province shows a high growth trend year by year, and can be influenced by certain seasonal factors. Besides, there is a short-term self-correlation between the monthly data of the same year, namely, the investment in the previous stage will directly affect the next stage's investment situation.

References

Fu, L., 2020. Time Series-Oriented Load Prediction Using Deep LSTM. Modern computer, (9), Pp. 25(in chinese).

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9 (8), Pp. 1735.

Hüsken, M., Stagge, P., 2003. Recurrent neural networks for time series classification. Neurocomputing, 50, Pp. 223.

Lei, L., Chen, X.N., 2020. Research on National Real Estate Investment Based on Factor Analysis and Time Series Modeling. Chinese market, (22), Pp. 5 (in chinese).

Li, R., Wang, Y., Chao, L., 2008. Short-term Wind Speed Forecasting Model for Wind Farm Based on Box-Cox Transformation. Modern Electric Power, (04), Pp. 35(in chinese).

Li, S.Y., Zhang, Y.J., 2020. Application of LSTM and Prophet Models in Predicting the Number of Tuberculosis Cases. Henan Science, 38 (2), Pp. 173(in chinese).

Li, Y., 2020. Analysis of the impact of real estate investment on economic growth in 2019. Chinese market, (1), Pp. 79 (in chinese).

Lu, S.H., 2016. Analysis and forecast of Nanning Real Estate Investment Based on Sparse Coefficient Seasonal Model. Times Finance, (12), Pp. 184 (in chinese).

Luo, Z.D., Liu, Y., Guo, W., 2019. Short-term prediction for stock price using a hybrid taylor expansion based on tracking differentiator and ARIMA Model [J]. Mathematics In Practice and Theory, 49 (23), Pp. 67 (in chinese).

Morana, C., 2001. A semiparametric approach to short-term oil price forecasting. Energy Economics, 23 (3), Pp. 325.

Peng, L.H., Zhang, X.B., 2016. Study on Prediction of Monthly rainfall Based on ARIMA-RBF Algorithm. World Sci-Tech R&D, 38 (2), Pp. 301 (in chinese).

Shen, H.J., Luo, Y., Zhao, Z.C.,etc. 2020. Prediction of summer precipitation in China based on LSTM network. Climate Change Research, 16 (3), Pp. 263(in chinese).

Wang, L.F., Nie, Y.J., Yang, X.D., 2020. Application value of SARIMA model in forecasting and analyzing inpatient cases of pediatric limb fractures. Chinese Journal of Evidence-Based Medicine, 20 (6), Pp. 651 (in chinese).

Wang, X., Chuai, J.H., Zhang, L.H., 2020. Research on Railway Passenger Flow Forecast Based on Prophet Time Series Algorithm. Computer Technology And Development, 30 (6), Pp. 130(in chinese).

Wang, X.F., Wang, B., Lu, Y.Y., 2020. Research of PM2.5 Concentration Forecasting Based on Prophet-LSTM Model. Software Guide, 19 (3), Pp. 133(in chinese).

Wang, Y., 2015. Applied Time Series Analysis. Beijing: China Renmin University Press. (in chinese).

Xu, T., Zhao, Y.M., 2011. Research on Amount of Real Estate Investment Forecast Based on GM (1,1) Model: Evidence from the City of Chengdu in Sichuan Province. Journal Of Chengdu University of Technology (Social Sciences), 19 (2), Pp. 26 (in chinese).

Yang, A.C., Wu, Y., Deng, X.S., 2020. Prediction method of electronic transformer error based on Prophet model. Automation and Instrumentation, (6), Pp. 52(in chinese).

Zhang, G., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, Pp. 159-175v (in chinese).

Zhang, K., Chui, L., 2020. Research on Multivariate Time Series Classification Algorithm Based on PCA-LSTM Model. Theoretical discussion, 36 (15), Pp. 44(in chinese).

Zhang, Q.J., Zhang, L., 2009. An empirical study and a prediction of the Cyclic Fluctuation of the real estate industry in China. Journal of Yunnan University of Finance and Economics, 25 (6), Pp. 64 (in chinese).

Zhang, Y.Y., Bai, Y.P., Hou, Y.C., 2019. Application of LSTM Model Based on Keras in Air Quality Index Prediction. Mathematics In Practice And Theory, 49 (07), Pp. 138(in chinese).

Zhao, L.W., Zhang, D.F., Zhu, Z.L., 2020. Port Traffic Volume Forecast with SARIMA-BP Model. Navigation Of China, 43 (1), Pp. 50 (in chinese).

Zhao, Y., 2009. Study on whole process risk prediction of real estate investment based on SVM of RS Pretreatment. Hebei University of Engineering, (in chinese).