Modeling Analysis and Comparison of Neural Network Simulation Based on ECM and LSTM

Lingyun Duan\textsuperscript{1,2,3,4,5}, Ziyuan Liu\textsuperscript{1,2,3,4}, Wen Yu\textsuperscript{1,2,3,4,5,*}, Wei Chen\textsuperscript{1,2,3,4}, Dongyan Jin\textsuperscript{1,2,3,4}, Denghua Li\textsuperscript{1,2,3,4}, Suhua Sun\textsuperscript{1,2,3,4,5} and Ruixi Dai\textsuperscript{1,2,3,4,5}

\textsuperscript{1}Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, P.R.China
\textsuperscript{2}Key Laboratory of Agricultural Information Service Technology, Ministry of Agriculture and Rural Affairs, Beijing 100081, P.R.China
\textsuperscript{3}Key Laboratory of Intelligent Agricultural Early Warning Technology and System, Chinese Academy of Agricultural Sciences, Beijing 100081, P.R.China
\textsuperscript{4}Beijing Engineering Research Center for Agricultural Monitoring and Early Warning, Beijing 100081, P.R.China
\textsuperscript{5}Graduate School of Chinese Academy of Agricultural Sciences, Beijing 100081, P.R.China

*Corresponding author email: yuwen@caas.cn

Abstract. Comparing the prediction effects of traditional econometric algorithm model and deep learning algorithm model, taking regional GDP as an example, two prediction models of ARMA-ECM and LSTM-SVR are established for prediction, and the prediction results of different models are compared and analyzed. The results show that there are some deviations in the prediction results of the two models, but the prediction trends are the same. The prediction accuracy of LSTM-SVR model will decrease significantly with the reduction of time series data samples, while ARMA-ECM model is not so sensitive.

Keywords: ECM, ARIMA, LSTM, SVR.

1. Introduction

Firstly, this paper summarizes the prediction research of previous scholars using ARMA model, ECM model and LSTM model in various fields, and then introduces the algorithm principle of the model, such as cointegration test, error correction model, autoregressive moving average model and LSTM model. Finally, the prediction results obtained by the two algorithms are analyzed and the advantages and disadvantages of the two prediction methods are compared.

At present, the development of modeling algorithm is divided into traditional measurement algorithm models, regression analysis models, grey system theory models and artificial neural network models. ARMA model and ECM model are classic time series forecasting methods, which can make good use of data rules for fitting and are widely used. Xia Junchang, Xie Kaigui, Cao Kan et al.\textsuperscript{[1]} (2011) established an error correction model to predict short-term electricity prices, and established a two-variable autoregressive equation. The results show that the ECM model effectively describes the short-term fluctuations of the time series and helps predict the development trend of electricity prices. Lu Wanbo and Yang Dong\textsuperscript{[2]} (2018) constructed a semi-parametric mixing data sampling error correction (SEMI-ECM-MIDAS) model for short-term prediction of China’s CPI. The results show
that the model has the best prediction accuracy when performing continuous predictions. And the prediction result is not affected by the choice of mixing dynamic cointegration relationship. Zhang Jinping, Li Hongbin and Xiao Honglin [3](2020) established error correction models to predict the annual runoff in the source area of the Yellow River. The results show that the goodness of fit of the three variable error correction model is greater than that of the two variable error correction model, which can be used for the annual runoff prediction in the source area of the Yellow River. Fan Bo and song Wenbin [4](2021) established a sales volume prediction algorithm based on ARIMA model to predict the sales volume of the product. The results show that ARIMA model is applicable to the short-term prediction of the product. Shi Tingting [5](2021) constructed ARIMA (2,1,2) model to study the short-term change law of RMB exchange rate, and predicted the RMB exchange rate in the next three months. The results show that ARIMA model is suitable for RMB exchange rate prediction, and all pass the significance test. Ni Jie, Yu Li and Jin Xiaonan [6](2021) used the improved ARIMA model to predict the passenger flow of urban rail transit. The results show that the prediction effect is good and has strong applicability.

With the development of machine learning and deep learning and the popularization of algorithms, the fitting of models to complex systems is getting better and better, and the research on future prediction and analysis using deep learning algorithms has also begun to appear widely. Zhang Yang and Zhang Fuyi [7](2021) used LSTM-ARIMA model to predict the traffic changes of mobile communication base stations with a long time span, and achieved good prediction results. Zhu Qing and Zhou Shipeng [8] (2021) used LSTM neural network in deep learning to predict and analyze GDP growth, and compared the prediction results with traditional time series ARIMA and GARCH models. The experiments show that LSTM model has high application value in macroeconomic prediction. Li Dekui, Du Shubo and Zhang Peng [9] (2021) predicted and analyzed the delayed passenger flow of urban rail transit by using LSTM model. They believe that LSTM method has smaller root mean square error than Arima method and has better prediction effect. Zhang Lei, sun Shanghong and Wang Yue [10] (2021) predicted the RMB exchange rate based on the long-term and short-term memory deep learning network model LSTM. The results show that the LSTM model has higher prediction accuracy than the traditional volatility model. Hu Yuwen [11] (2021) proposed an LSTM prediction model based on PCA and LASSO to predict the stock closing price. The results show that PCA-LSTM model can greatly reduce data redundancy and obtain better prediction accuracy.

ARMA is an econometric model based on time series data. It selects a wide range of data types without too many restrictions. Aimed at the interference of sequence order, timeliness and uncertainty of random factors, ARMA model can still show good stationarity. Therefore, ARMA model can be applied in a wide range of fields, has strong application performance, and has a good effect in the application of short-term prediction of time series. LSTM model is developed from RNN. On the basis of inheriting the advantages of RNN, it also overcomes the influence of gradient disappearance and gradient explosion. It has long-term memory function and is simpler and more accurate in sequence prediction. Therefore, based on ARMA and LSTM time series models, this paper studies and analyzes the advantages and disadvantages of the model.

2. Modeling

2.1. Data Selection and Processing

Select the GDP of each region from 1990 to 2019, per capita education years of each region from 1990 to 2017, and the fixed asset investment data of each region from 1990 to 2017 are from the Data Center of the National Bureau of Statistics. Regional GDP, regional investment in fixed assets, and per capita years of education in the region are respectively expressed as gdp, inv, and edu.
2.2. Traditional Measurement Algorithm Model

2.2.1. Modeling Ideas

![Image of modeling ideas](image)

Figure 1. Modeling ideas of traditional measurement algorithm model.

In order to eliminate the heteroscedasticity between variables, the regional GDP and regional investment in fixed assets are processed in logarithm and named lngdp and lninv respectively. Firstly, the unit root test is used to test the stationarity of variables; Then, the cointegration test is carried out to determine the long-term relationship between variables and establish the error correction model. Finally, the autoregressive moving average model is used to calculate the prediction data of independent variables, and combined with the error correction model to predict the predicted value of regional GDP from 2020 to 2050.

2.2.2. Unit Root Test. Before testing whether there is a cointegration relationship between time series variables, it is necessary to use the ADF test to test the stationarity of the time series variables. Only when the variables are all single integrals of the same order can there be a cointegration relationship. The ADF test values of lngdp, lninv, and edu in each region are neglected. Its ADF test values of the first-order difference time series of lngdp, lninv, and edu are all less than 5% critical value -2.992, so the null hypothesis that there is a unit root is rejected at the 5% significance level, that is, the first-order difference time series of lnGDP, lninv and edu are stationary.

2.2.3. Cointegration Test. According to the ADF test, lngdp, lninv, and edu are all first-order single-integration time series, and the cointegration test can be performed to determine whether there is a cointegration relationship between variables. Using the EG two-step test method:

The first step is to use the least square method to perform cointegration regression on lngdp-lninv-edu. The cointegration regression equation of the three variables is:

$$Y_t = a_0 + a_1X_t + a_2Z_t$$ (1)

In the above formula, Y is lngdp, X is lninv, and Z is edu.

The second step is to test the stationarity of the residual sequence (et) of each regression equation. Using the ADF test, if the residual sequence is stationary, the cointegration relationship between the variables exists. The critical value of ADF test for cointegration test of residual series (et) is calculated by the method of cointegration test in *Econometrics* edited by Li Zinai. The critical value of ADF test for cointegration test with 3 variables and 30 sample sizes at 5% significance level is -3.461. After testing, the ADF test values are less than -3.461, so the original hypothesis of unit root is rejected at the significance level of 5%. Therefore, the residual series of each regression equation is stationary, so there is a co-integration relationship between lngdp-lninv-edu.
2.2.4. Error Correction Model (ECM). Through cointegration test between variables, it is proved that there is a long-term equilibrium relationship. In order to avoid short-term fluctuations in the long-term equilibrium relationship and reduce errors, Error Correction Model emerges as the times require. The stable residual series of regression equation with long-term equilibrium relationship is put into the model as error correction term, and the model is established together with other explanatory variables.

Specific modeling steps:
The first step is to find the residual sequence as the error correction term (ecm) from the cointegration equation tested above. The residual formula of the three variables is:

$$\text{ecm}_{t-1} = Y_{t-1} - a_0 - a_1 X_{t-1} - a_2 Z_{t-1}$$

(2)

The second step is to establish the error correction model of lngdp-lninv-edu, and the first-order error correction model formula of three variables is:

$$\Delta Y_t = \beta_1 \Delta X_t + \gamma_1 \Delta Z_t - \lambda \text{ecm}_{t-1} + \mu_t$$

(3)

Substitute ecm as a new variable and the data corresponding to other independent variables into the above formula, and use ordinary least squares to estimate the corresponding parameters.

2.2.5. Autoregressive Moving Average Model (ARMA). The ARIMA model is a commonly used time series forecasting model at present, and it has a good fitting effect. The model first checks the stationarity of the time series, then uses the ACF and PACF functions to determine the order combined with the AIC and BIC criteria, and establishes the model prediction formula. The formula of ARMA (p, q) model is:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}$$

(4)

From the formula (3) of the Error Correction Model, the data of $\Delta X_t$, $\Delta Z_t$ and ecm$_{t-1}$ are needed to calculate the short-term fluctuation $\Delta Y_t$. Therefore, to predict the short-term fluctuations of lngdp, we need to obtain the forecast data of lninv, edu and ecm. The ARIMA model is used to obtain the predicted values of lninv and edu in each region from 2018 to 2050.

2.2.6. Forecast Results. Substitute lninv and edu into formula (4) and ecm$_{t-1}$ into formula (2) to calculate the forecast data. Substitute the forecast data corresponding to lninv, edu, ecm$_{t-1}$ into formula (3) to find the annual growth value of lngdp, and finally, calculate the annual GDP. The forecast results are shown in figure 2.

Figure 2. GDP prediction results of ARMA-ECM prediction model.

Note: In the figure 2, the yellow part is the trend of GDP development from 1990 to 2019 and the red part is the trend of GDP development from 2020 to 2050.
2.3. Deep Learning Algorithm Model

2.3.1. Modeling Ideas

![Diagram](image)

Figure 3. Modeling ideas of deep learning algorithm model.

Firstly, the long-term and short-term memory artificial neural network based on deep learning estimates the per capita years of education and fixed asset investment from 2018 to 2050; Secondly, the data of per capita years of education, fixed asset investment and GDP from 1990 to 2017 are used as the training set, and the support vector regression algorithm is used for supervised learning to establish the prediction model; Finally, the predicted value of regional GDP from 2020 to 2050 is estimated.

2.3.2. Short And Long Term Memory Artificial Neural Network (LSTM).

In 1997, Hochreiter and Schmidhuber proposed short-term and long-term memory artificial neural network (LSTM). LSTM is an improved nonlinear recurrent neural network (RNN), which avoids the problems of gradient disappearance and gradient explosion caused by too many hidden layers and large amount of calculation in traditional RNN [12]. The traditional RNN model has the problem of low information persistence. If the useful information is far from the prediction point, RNN cannot correctly predict the results according to the useful information. Therefore, when the amount of information is too large, the use effect of RNN is poor. LSTM model is improved on the basis of traditional RNN. LSTM is composed of several basic repeating units that can realize hierarchical transmission of information. Each basic unit is formed by four neural network layers interacting with each other. Unlike the traditional RNN, there is only one simple Tanh module, which is also the main reason why LSTM model can realize "persistence".

LSTM model is composed of multiple memory cells. Each LSTM network memory cell contains three gating units. The memory ability of each cell is updated and deleted by input gate, output gate and forget gate. The functions of the three gates are as follows[13]:

1. Forget gate. It determines how much information is discarded from the previous cell state. Taking the input $x_t$ of the current layer and the output $h_{t-1}$ of the previous layer as the input, the cell state output at time $t-1$ is:

$$ f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5) $$

2. Input gate. It determines how much information in the current newly entered information is saved in the memory cell and what value needs to be updated:

$$ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6) $$

$$ c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (7) $$

3. Output gate. It determines how much current information is output, and finally determines the output content:
\[ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \]  
\[ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \]  
\[ h_t = o_t \cdot \tanh(C_t) \]

\( \sigma \) is the sigmoid activation function. The value range is \([0, 1]\). 0 means to discard all, and 1 means to retain all; \( W_f, W_i, W_c \) and \( W_o \) are weight matrices; \( b_f, b_i, b_c \) and \( b_o \) are offset vectors, \( \tanh \) is hyperbolic tangent activation function, and the value range is \([-1, 1]\).

2.3.3. Support Vector Regression (SVR). SVR is a branch algorithm of support vector machine (SVM) and a linear regression model. SVM is a supervised learning algorithm, which mainly solves the classification problem. SVM determines the optimal decision surface by finding the maximum classification interval, and the sample nearest to the decision surface is included in the loss, so as to find the optimal classifier of the classification problem. SVR is the opposite of SVM. SVR aims to find the minimum classification interval and include the data outside the decision surface into the loss, so that most samples in a set are closest to the decision surface.

Firstly, the original data are normalized; Secondly, select the classifier as SVR, the data of per capita years of education, fixed asset investment and GDP from 1990 to 2017 are used as the training set to train the model and establish the prediction model.

2.3.4. Forecast Results. Taking the per capita years of education and fixed asset investment in 2020-2050 as independent variables and substituting them into the trained prediction model, the predicted value of regional GDP in 2020-2050 is estimated. The prediction results are shown in Figure 5.

**Figure 4. LSTM network model structure.**

**Figure 5. GDP prediction results of LSTM-SVR prediction model.**

Note: In the figure 5, the yellow part is the trend of GDP development from 1990 to 2019 and the red part is the trend of GDP development from 2020 to 2050.
3. Conclusions and Suggestions
Taking the regional GDP as an example, this paper uses the per capita years of education, fixed asset investment and regional GDP from 1990 to 2017, uses the traditional measurement algorithm model and the deep learning algorithm prediction model to predict, and establishes two prediction models of ARMA-ECM and LSTM-SVR. According to the empirical analysis results of the comparison of the two prediction models, the prediction results of the two models are biased, but the forecast trend is consistent. On the one hand, LSTM and SVR are data driven supervised machine learning approaches that do not take the statistical relations of data into account. In this scenario, the machine learning methods also suffer from the limited scale of training data. The prediction accuracy of LSTM-SVR model will decrease significantly with the reduction of time series data samples. On the other hand, the ECM model is a theoretically driven approach that can be helpful in the estimation of both short-term and long-term relations of time series data. ARMA-ECM model uses the long-term common random trend of basic variables to fit, and the response is not so sensitive.

Acknowledgements
This research was supported by the National Key R&D Program Project (2017YFE0104600).

Reference
[1] Xia Junchang, Xie Kaigui, Cao Kan, Li Chunyan. Error correction model for short-term electricity price forecasting in power market [J]. East China power, 2011,39 (02): 161-167
[2] Lu Wanbo, Yang Dong. Research on CPI prediction in China based on semi parametric mixing error correction model [J]. Statistical research, 2018,35 (10): 28-43
[3] Zhang Jinping, Li Honglin, Xiao Honglin, Zhang Xin. Prediction of annual runoff in the source area of the Yellow River Based on error correction model [J]. People's Yellow River, 2020,42 (07): 5-8 + 66
[4] Fan Bo, Song Wenbin. Research on product sales forecast based on ARIMA model [J]. Industrial control computer, 2021,34 (05): 128-129 + 125
[5] Shi Tingting. Prediction of the fluctuation trend of COVID-19's RMB exchange rate based on ARIMA model [J]. modern business, 2021 (14): 108-110.
[6] Ni Jie, Yu Li, Jin Xiaonan. Passenger flow prediction and research of Urban Rail Transit Based on ARIMA model [J]. Intelligent computer and application, 2021,11 (04): 135-138
[7] Zhang Yang, Zhang Fuyi. Traffic prediction method of mobile communication base station based on LSTM and ARIMA model [J]. Information and computer (theoretical Edition), 2021,33 (07): 100-102
[8] Zhu Qing, Zhou Shipeng. Research on forecasting and modeling of national economic GDP growth based on LSTM model [J]. Economic Research Guide, 2021 (19): 5-9 + 37
[9] Li Dekui, Du Shubo, Zhang Peng. Comparison of delayed passenger flow prediction methods of Urban Rail Transit Based on ARIMA and LSTM [J]. Journal of Qingdao University of technology, 2021,42 (04): 135-142
[10] Zhang Lei, Sun Shanghong, Wang Yue. Exchange rate forecasting based on deep learning LSTM model [J]. Statistics and decision making, 2021,37 (13): 158-162
[11] Hu Yuwen. Stock forecasting based on optimized LSTM model [J]. Computer science, 2021,48 (S1): 151-157
[12] Chen Wei, Qi Weiyuan, Yuan Fuxiang, Li Zhemin. Comparative analysis of rice yield prediction in Jilin Province Based on time series and cross-sectional data [J]. China agricultural information, 2018,30 (05): 91-101
[13] Duan Chuanming. Comparative analysis of traditional emotion classification methods and emotion classification methods based on deep learning [J]. Software guide, 2018,17 (01): 22-24