Forecast of Steel Price on ARIMA-LSTM Model

Haidi Wu[1], Mingxun Li[2], Kimhong LIM[2], Cunrong Li[3*]
whdllll@163.com, lmx@126.com, 1193858434@qq.com, 2832962632@qq.com

[1]Industrial Engineering Department, School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan, China

[2]Mechanical Engineering Department, School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan, China

[3]Technical research and development Department, Suizhou Industrial Research Institute of Wuhan University of Technology, Wuhan University of Technology, Wuhan, China

Abstract— Forecasting the price of steel is important for the manufacturing industry to make procurement plans and production plans. Price is affected by many factors, considering its time-series characteristics, this paper uses the ARIMA model to predict the linear part and LSTM to predict the non-linear residual part. Simulation results show that ARIMA-LSTM model has higher accuracy.

Keywords: ARIMA; steel price; forecasting; time series; ARIMA-LSTM

1 INTRODUCTION

Steel price is affected by many factors and fluctuates over time. Time series can reveal internal statistical characteristics of the data without totally knowing its influencing factors [1]. Commonly used time series models include AR model, MA model, and ARMA model [2]. These models can only solve stationary sequences. Non-stationary sequence has certain timing and fluctuating characteristics and needs to be solved by ARIMA. This paper uses ARIMA-LSTM to analyze and predict steel prices.

2 METHOD AND SOURCE

2.1 ARIMA model

ARIMA, also known as differential autoregressive moving average, is proposed by scholars Box and Jenkins [3]. ARIMA model, often recorded as ARIMA (p, d, q), is composed of AR model and MA model. Parameter d is the number of differences to make non-stationary series stationary.

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2326770

The basic steps to build an ARIMA prediction model are described as follows. First, get data and analyze characteristics, such as randomness, stationarity. Non-stationary series needs to be stationary before being modelled. Then, establish a suitable model and check relevant parameters. The value of parameters p and q can be attained by autocorrelation function and partial autocorrelation function plots. Finally, use the verified model to make predictions.

When the random process is related to its previous values and external random terms, its mathematical form can be expressed as follows:

$$yt=c+\alpha_1 y_{t-1}+\ldots+\alpha_p y_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\ldots+\theta_q\varepsilon_{t-q} \tag{1}$$

In formula (1), the value y_t at time t has a multivariate function relationship with y_{t-p}. θ_t and α_t are the coefficients.

2.2 Lstm model

The nonlinear characteristics in the time series data cannot be reflected in the linear model ARIMA, and the nonlinear model needs to be established. LSTM neural network can solve this problem and it has good performance in time series problems [4]. LSTM unit has a memory cell, input gate, forget gate, and output gate [5]. It is shown in figure 1.

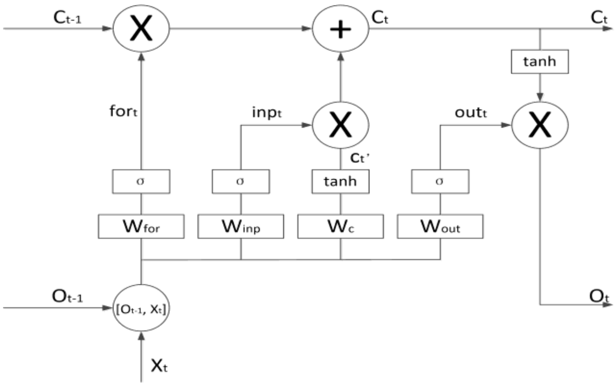


Figure 1 LSTM unit

The forward formulas of its unit are calculated as follows:

$$inp_t=\sigma(W_{inp}.[O_{t-1}, X_t]+b_{inp}) \tag{2}$$

$$out_t=\sigma(W_{out}.[O_{t-1}, X_t]+b_{out}) \tag{3}$$

$$c_t=for_t * c_{t-1}+ inp_t * c_t' \tag{4}$$

$$c_t'=tanh(W_c.[O_{t-1}, X_t]+b_c) \tag{5}$$

$$O_t=out_t * tanh(c_t) \tag{6}$$

Xt is the input vector of this unit. Ot is the final output of this unit. Winp, Wout and Wfor are the parameters of these three gates. B is the bias of gates. Tanh and σ are the arithmetic function of units. The output of the previous moment can be used as the input of the next moment to continue to participate in the calculation.

2.3 Data Source

The data comes from the 2011-2022 data of SPCC, a type of steel, in Wuhan. In this paper, the original sequence is divided into the training set and the test set according to seven to three.

3 ESTABLISH ARIMAMODEL

Before analysis, it is necessary to test the stationarity of the data. Figure 2 shows the data arranged in time. It can be seen that this sequence has an upward trend and is non-stationary.

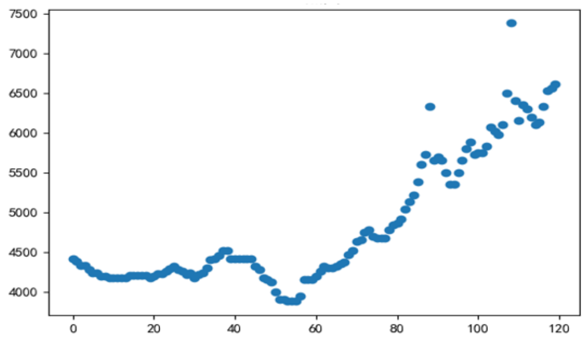


Figure 2 Raw data

3.1 Data Processing

Determine whether there are outliers and deal with those abnormal values. If there exist outliers, delete them and fill in new values. According to the 3σ principle of the normal distribution, values exceeding 3σ are outliers, but the density plot and QQ plot in the figure 3 show that the data does not conform to the normal distribution. Combined with the box plot, there exist no outliers in the series, so subsequent analysis can be performed.

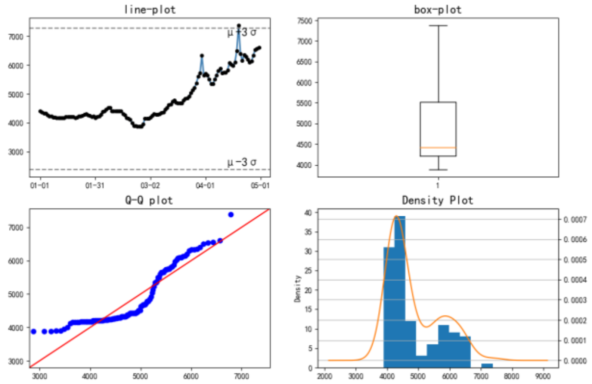
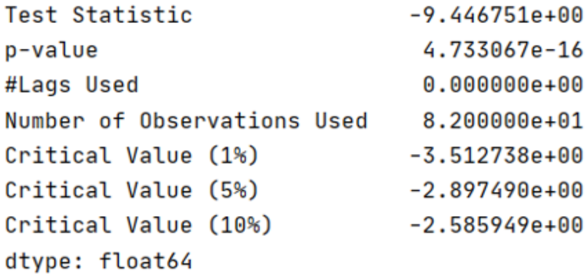


Figure 3 Outlier Identification

3.2 Stationary test and White noise test

The original sequence is not stationary, but the sequence after the first difference is stationary. The unit root test is shown in figure 4. The p-value is much less than 0.05 and the statistic T value is less than -2.89, so the null hypothesis is rejected and the first-order difference sequence is considered to be stationary at the 5% significance level.



```
Test Statistic                  -9.446751e+00
p-value                          4.733067e-16
#Lags Used                       0.000000e+00
Number of Observations Used      8.200000e+01
Critical Value (1%)             -3.512738e+00
Critical Value (5%)             -2.897490e+00
Critical Value (10%)            -2.585949e+00
dtype: float64
```

Figure 4 Unit Root Test

White noise is also stationary, but it is meaningless. To eliminate the influence of white noise, the stationary series needs to be tested. The P values are all less than 0.05, so at this significance level, the null hypothesis is rejected and the tested sequence is not a white noise sequence.

3.3 The values of p and q

Parameters p and q play an important role in the ARIMA model. Information criteria and graphs are used to determine the values. According to correlation plots, the value of parameters p and q can be determined, and then the corresponding values of BIC (Bayesian information criterion) can be calculated as follows.

$$BIC=K\ln(n)-2\ln(L) \tag{7}$$

In equation (7) K is the number of parameters, n is the number of samples, and L is the likelihood function.

After determining the model, the white noise test should be performed to ensure that there exists no useful information in the tested series. The p-value of q statistics should be over 0.05, so the residuals are white noise sequences.

4 MODEL ANALYSIS AND FORECAST

After the model is qualified, it can be used for prediction.

4.1 Model analysis

Table1 and figure 5 show that when parameters p=1 and q=0, the value of BIC is the smallest, and that the model should be ARIMA (1, 1, 0).

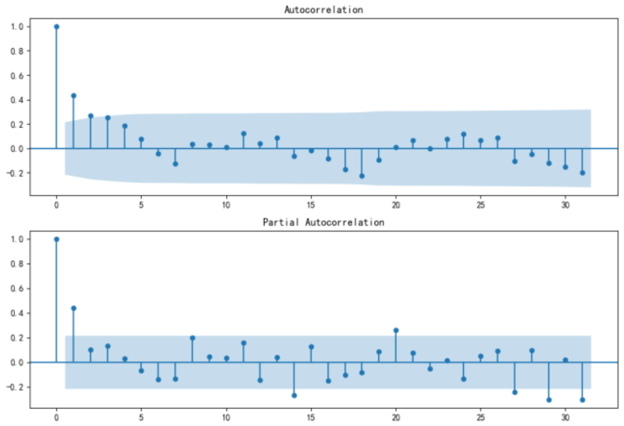
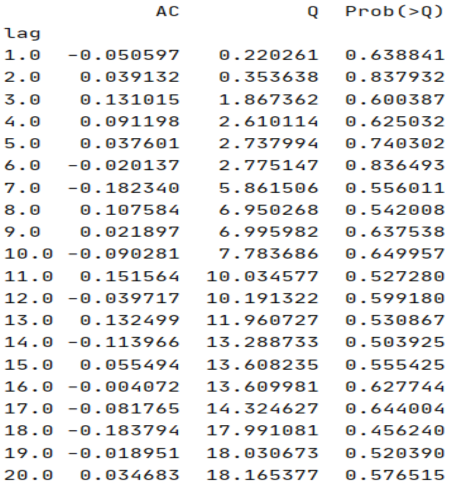


Figure 5 ACF and PACF plots

Table 1 Value of BIC

ARIMA	BIC value
ARIMA(1,1,1)	893.416
ARIMA(1,1,0)	891.137
ARIMA(0,1,2)	897.543
ARIMA(0,1,1)	895.412

Figure 6 shows the values of the white noise test and that the residuals are the random sequence. All useful information has been included in model ARIMA (1, 1, 0).



```
                AC          Q  Prob(>Q)
lag
1.0   -0.050597   0.220261  0.638841
2.0    0.039132   0.353638  0.837932
3.0    0.131015   1.867362  0.600387
4.0    0.091198   2.610114  0.625032
5.0    0.037601   2.737994  0.740302
6.0   -0.020137   2.775147  0.836493
7.0   -0.182340   5.861506  0.556011
8.0    0.107584   6.950268  0.542008
9.0    0.021897   6.995982  0.637538
10.0  -0.090281   7.783686  0.649957
11.0   0.151564  10.034577  0.527280
12.0  -0.039717  10.191322  0.599180
13.0   0.132499  11.960727  0.530867
14.0  -0.113966  13.288733  0.503925
15.0   0.055494  13.608235  0.555425
16.0  -0.004072  13.609981  0.627744
17.0  -0.081765  14.324627  0.644004
18.0  -0.183794  17.991081  0.456240
19.0  -0.018951  18.030673  0.520390
20.0   0.034683  18.165377  0.576515
```

Figure 6 White Noise Test of Residuals

The correlation test of residuals is shown in Figure 7. It can be seen that the residual is stable and fluctuates around zero. The histogram plot, density plot, and QQ plot show that the sequence obeys the normal distribution. The autocorrelation graph is all in the blue area. This means it is a purely random sequence of white noise.

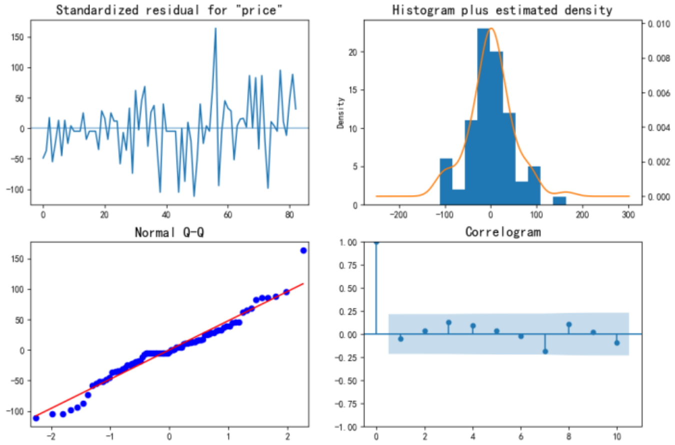


Figure 7 White Noise Test of Residuals

4.2 Evaluation index

The quality of this model needs to be evaluated, and the commonly used evaluation indicators are mean absolute error, root mean square error, and so on. Indicators RMSE and MAE are used in this paper.

$$RMSE=sqrt(sum(ytrue_i\text{-}ypre_i)^2/n) \tag{8}$$

$$MAE= sum\ (abs\ (ytrue_i\text{-}ypre_i))/n \tag{9}$$

In these formulas, $ypre$ represents the predicted value, $ytrue$ represents the true value, and n represents the number of samples.

4.3 Prediction

As the prediction length increases, the accuracy of the model will decrease; the original model may fail when the external environment changes greatly, so the rolling prediction is adopted. When predicting the next month's data, the actual value obtained in the previous month is used to update the correction coefficient of the model, and then the corrected model is used for out-of-sample forecasting. The results are shown in figure 8.

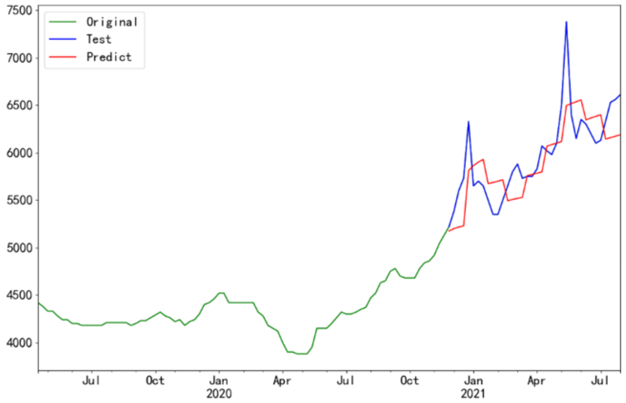


Figure 8 ARIMA forecasting

The results of forecasting are rounded to two decimal places and shown in table 2.

Table 2 Prediction of ARIMA

No	value	prediction	No	value	prediction
0	5380	5200.04	18	6070	5797.37
1	5600	5216.10	19	6020	6068.06
2	5730	5228.17	20	5980	6085.38
3	6330	5810.89	21	6100	6101.17
4	5650	5863.821	22	6500	6117.08
5	5700	5900.83	23	7380	6496.83
6	5650	5928.77	24	6400	6517.42
7	5500	5674.14	25	6150	6536.62
8	5350	5686.13	26	6350	6555.89
9	5350	5700.11	27	6300	6345.79
10	5500	57 13.76	28	6200	6365.58
11	5650	5493.80	29	6100	6382.55
12	5800	5507.27	30	6130	6399.85
13	5880	5518.25	31	6330	6143.20
14	5730	5529.55	32	6530	6158.30
15	5750	5758.38	33	6560	6173.18
16	5750	5770.42	34	6610	6188.09
17	5830	5783.97			

To verify the performance, this paper tests two models: ARIMA and ARIMA-LSTM. ARIMA-LSTM means that use ARIMA to predict the linear part and use LSTM to predict the non-linear residual part, finally these two parts are added together. Figure 9 shows the prediction results of these two models. Table 3 shows the values of evaluation indicators of these two models.

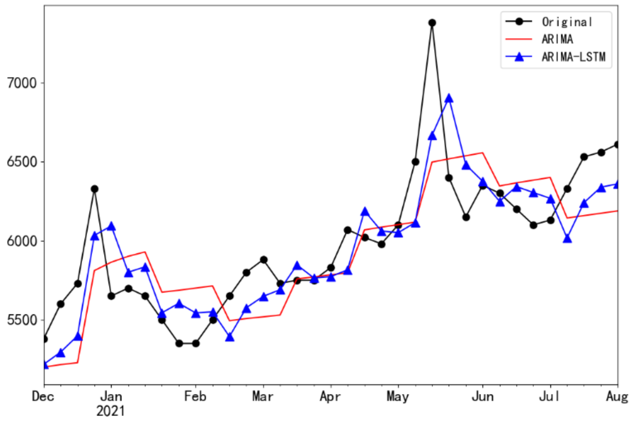


Figure 9 Forecasting results of ARIMA and ARIMA-LSTM

Table 3 Evaluation of Models

	RMSE	MAE
ARIMA	309.98	256.36
ARIMA-LSTM	259.17	211.72

It can be seen from Table 3 that the evaluation indicators of ARIMA-LSTM are lower than ARIMA model, and the prediction is better.

5 CONCLUSION

In this paper, ARIMA and LSTM are combined to predict steel price, ARIMA is used for rolling prediction on the linear part to improve accuracy, and LSTM is used to predict the nonlinear part. Finally, the two results are mixed. Simulation experiments show that the mixed model has better prediction accuracy.

Acknowledgment. This research is supported by the key research and development project of Hubei Province (Project number: 2020BAB013; Project name: Development and application of the intelligent networked system for multi-dimensional heterogeneous data special-purpose automobile industry).

REFERENCES

[1] S. Z. Yang, Y. Wu, J. P. Xuan. Engineering application of time series analysis[M]. Wuhan: Huazhong University of Science and Technology Press, 2007.

[2] H. M. Yang, Z. S. Pan, W. Bai. Overview of Time Series Forecasting Methods[J]. Computer Science, 2019, 46(01): 21-28.

[3] Y. Ding, D. Wu, W. Li, et al. ARIMA seasonal model predicts the trend of hepatitis E in China[J]. Journal of Nanjing Medical University (Natural Sciences), 2020, 40(11): 1725-1729.

[4] L. Yang, Y. X. Wu, J. L. Wang, et al. Overview of Research on Recurrent Neural Network[J]. Journal of Computer Applications, 2018, 38(S2): 1-4.

[5] Q. Yang, Z. W. Wang. Research on Global Stock Index Prediction Based on Deep Learning LSTM Neural Network[J]. Statistical Research, 2019, 36(03): 66-68.