Research on Railway Freight Volume Prediction Based on ARIMA Model

Jianyou Zhao[1]; Jing Cai[2]; and Wenjie Zheng[3]

[1]School of Automobile, Chang'an Univ., Xi'an, Shanxi, China ST 710064. E-mail: jyzhao@chd.edu.cn
[2]School of Automobile, Chang'an Univ., Xi'an, Shanxi, China ST 710064. E-mail: caijing.chd@foxmail.com
[3]School of Automobile, Chang'an Univ., Xi'an, Shanxi, China ST 710064. E-mail: 176030320@qq.com

ABSTRACT

Railway freight transport plays an important role in transport, the accurate forecast of the freight volume is expected to guild the planning of railway business. The previous works on forecasting the freight volume have commonly used regression model, times series model, as well as grey model, however the useful information has been overlooked by using those prediction methods. The present study set out to forecast the railway freight volume by establishing an autoregressive integrated moving average model (ARIMA model), using the original data of railway freight volume in the Ningxia Hui Autonomous Region to do an empirical analysis. The results shown that the ARIMA model in forecasting railway freight volume can improve the prediction accuracy and could provide support for railway freight volume forecasting.

INTRODUCTION

Railway freight transportation, as an important mode of transportation, has the characteristics of long transportation distance, good safety, all-weather, and little influence by natural conditions, so that it has obvious advantages in the transportation of bulk goods such as coal, oil, natural gas and timber. The freight volume accounts for more than 13% of the total freight volume of the modern transportation mode. Bulk cargo transport requires higher railway capacity, and the accurate prediction of railway freight volume in the whole country has certain reference value for railway departments to arrange transportation capacity, formulate strategic planning and implement specific measures.

In previous studies, little research results has been reported at this point using time series method to forecast railway freight volume. Among the weighted moving average method, the trend moving average method, the exponential smoothing method, and the ARIMA model, the most suitable model for the prediction of the railway freight volume was the two exponential smoothing models (Song 2007). A time-series based forecasting process was proposed by Hu and Li (2014), using the expert modeling method to simplify the forecasting process and increase the modeling speed.

Several research efforts have contributed to railway freight volume forecasting, which is carried out in terms of linear regression, neural network model and grey prediction. The prediction model based on improved Gray-Markov chain model was established to predict the railway freight volumes (Guo et al. 2011). Li presented a single factor system cloud grey SCGM (1,1) model to forecast the railway freight volume (Li, 2011). Concerning the shortcomings of the methods which forecast railway freight volume, Li proposed Grey Neural Network (GNN) based on the Improved Particle Swarm Optimization algorithm (IPSO-GNN) (Li et al. 2012). In

order to improve prediction accuracy of the BP neural network model, a prediction model is presented based on combined Ada Boost algorithm and BP neural network. Li had shown that this model was effective and suitable, had higher forecasting accuracy, and was applicable to practice.

From the literature review, we know that most linear regression and grey prediction studies only take the causality of the model into consideration, but they don't fully consider the internal complexity of the railway transportation system. The neural network analysis method can obtain the railway freight volume under the comprehensive influence of multiple factors which can reflect the dynamic complexity of the railway transportation system, but the method is affected by the local minimum, thus affecting the prediction effect.

To sum up, the railway freight volume formed over time is regarded as a random sequence in this study, and an ARIMA model is used to describe this sequence approximately. By analyzing the ARIMA model, the structure and characteristics of sequence are recognized. The data sequence can be extended in the future based on the past development and change rule, and the ARIMA (2,2,2) model is established to forecast the railway freight volume in Ningxia Hui Autonomous Region. The value of this study is to propose the railway freight volume forecasting model to enhance the efficiency of the Railway Administration.

ARIMA MODEL

An autoregressive integrated moving average (ARIMA) model is fitted to time series data to better understand the data or to predict future points in the series. ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity. Non-seasonal ARIMA models are generally denoted ARIMA (p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. The model formula is:

$$\varphi(\mathrm{B})(1-B)^{d} y_t = c + \theta(\mathrm{B})\varepsilon_t \tag{1}$$

In the function: d is the dth order difference operation;

$$\theta(\mathrm{B}) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \tag{2}$$

$$\varphi(\mathrm{B}) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_q B^p \tag{3}$$

DATA ANALYSIS

This paper uses the data of railway freight volume in the Ningxia Hui Autonomous Region Statistical Yearbook in 1995–2015 years to do an empirical analysis. Ningxia is located in the geometric center of China, adjacent to Inner Mongolia, Shanxi Province and Gansu Province. As a trade center of the Silk Road, Ningxia is defined as the Inland Opening-up Pilot Economic Zone in the Belt and Road Strategy. Railway transportation is an important mode of freight transportation in Ningxia, and an efficient transport routes connecting central, southern, and western Asian countries along the Belt and Road. The accurate forecast of freight volume can make Ningxia railway department arrange transportation capacity and serve the National Freight Corridor better. The railway freight volume of the Ningxia Hui Autonomous Region in 1995–2015 years is shown in Table 1.

Table 1. Railway freight volume of the Ningxia Hui Autonomous Region

Year	Railway Freight Volume(million tons)
1995	1298
1996	1374
1997	1478
1998	1512
1999	1605
2000	2105.3
2001	2300
2002	2367.2
2003	2528.5
2004	2768.4
2005	2881
2006	3329
2007	3957
2008	4400
2009	5978
2010	6879
2011	7850
2012	8485.6
2013	8412
2014	6990
2015	5631.1

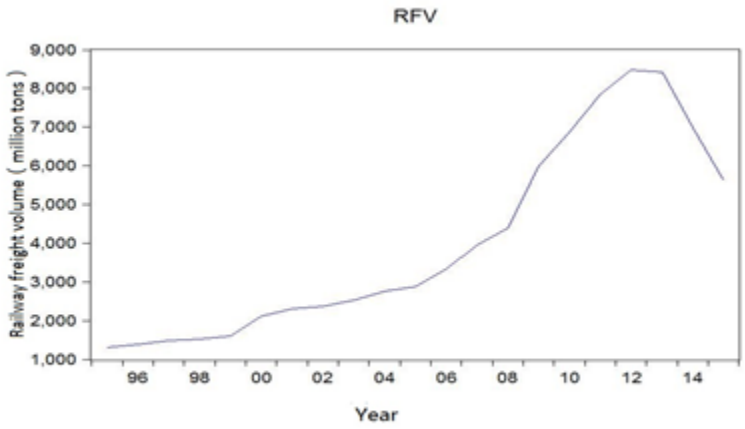


Figure 1. Sequence diagram of {RFV}

ARIMA MODEL ESTABLISHMENT

Stationary Test of Railway Freight Volume Sequence

In this paper we analyzed the sequence diagram of sequence {RFV} which represented the annual data sequence of the Ningxia Hui Autonomous Region railway freight volume by using EVIEWS 8, the sequence diagram as shown in Figure 1. As was illustrated in Figure 1, the {RFV} sequence had an obvious growth trend from 1995 to 2013 while an obvious downward

trend from 2013 to 2015 shown that the time series {RFV} has heteroscedasticity.

In order to reduce the effect of heteroscedasticity, the time series {RFV} took the natural logarithm to obtain the new time series {LNRFV}. Figure 2 was shown that there was an obvious fluctuation in {LNRFV}, which indicated that the fluctuation of {RFV} was not caused by heteroskedasticity, but caused by some of the numerical features varying with time.

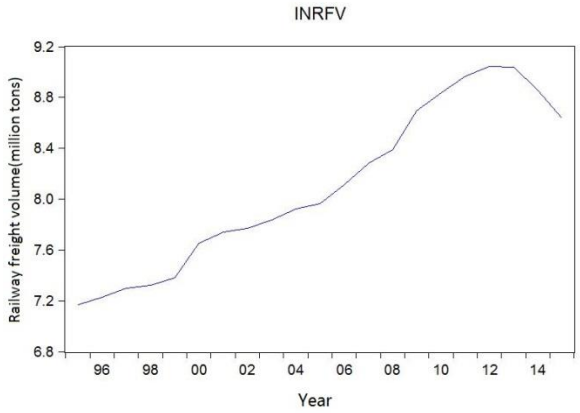


Figure 2. Sequence diagram of {LNRFV}

Table 2. Results of sequence {RFV} ADF test

		t-Statistic	Prob.*
Augmented Dickey-Fuller Test Statistic		-2.257945	0.1943
Test critical values:	1% level	-3.831511	
	5% level	-3.029970	
	10% level	-2.655194	

Augmented Dickey-Fuller Test (ADF) Test

For purpose of building the ARIMA model, Augmented Dickey-Fuller Test was used to check the stationarity of series {RFV}. ADF Test is one of the unit root test methods. It is assumed that the sequence {y} has p order sequence correlation, and the expression of whether the sequence {y} is stationary sequence is:

$$\Delta y_t = a + \beta t + \eta y_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta y_{t-1} + \mu_t \tag{4}$$

In the function: a is a constant; β is the coefficient on a time trend; p is the lag order of the autoregressive process; $\eta = \sum_{i=1}^{p} \varphi_i - 1; \beta_i = -\sum_{j=i+1}^{p} \varphi_j$.

$$\begin{cases} H_0 : \eta = o \\ H_0 : \eta < 0 \end{cases} \tag{5}$$

In the ADF test theory, the acceptance of H_0 shows that the sequence is non-stationary, while the rejection of H_0 shows that the sequence is stationary.

ADF test was performed on the sequence {RFV}.We can see from Table 2 that $t_r = -2.25794$

, $t_r > t_{0.01}$; $t_r > t_{0.05}$, the test statistic of the sample value was respectively higher than the critical value of significant level l %, 5%. It was shown that the sequence {RFV} was a non-stationary sequence.

A new stationary sequence can be obtained by differential operation for the non-stationary sequence. The mathematical model of differential operation is as follows:

$$\Delta y_t = y_t - y_{t-1} = a + u_t \qquad t = 1, 2, \cdots, T \tag{6}$$

In the function: a is a constant; u_t is a stationary sequence.

If the sequence {y} becomes a stationary sequence by dth order difference operation, and the (d-1)th order difference of sequence {y} is not stable, then the sequence {y} is dth order stationary sequence.

Respectively calculated by the first and second order difference operation, the new sequences {DRFV} and {DRFV2} were obtained from sequence {RFV}. Then ADF test was performed on the sequence {DRFV} and {DRFV2}. The results of sequence {DRFV} and {DRFV2} ADF test were listed respectively in Table 3 and Table 4. As shown in Table 3, $t_{r1} = 3.643747$, $t_{r1} > t_{0.01}$, the test statistic of the sample value was respectively higher than the critical value of significant level l %. It was shown that the sequence {DRFV} was a non-stationary sequence. As is shown in Table 4, $t_{r2} = -4.138708$, $t_{r2} < t_{0.01}$; $t_{r2} < t_{0.05}$; $t_{r2} < t_{0.1}$, the test statistic of the sample value was respectively lower than the critical value of significant level l %,5 %, and 10%. It was shown that the sequence {DRFV2} was a stationary sequence.

Table 3. Results of sequence {DRFV} ADF test

		t-Statistic	Prob.*
Augmented Dickey-Fuller Test Statistic		-3.643747	0.0180
Test critical values:	1% level	-3.959148	
	5% level	-3.081002	
	10% level	-2.681330	

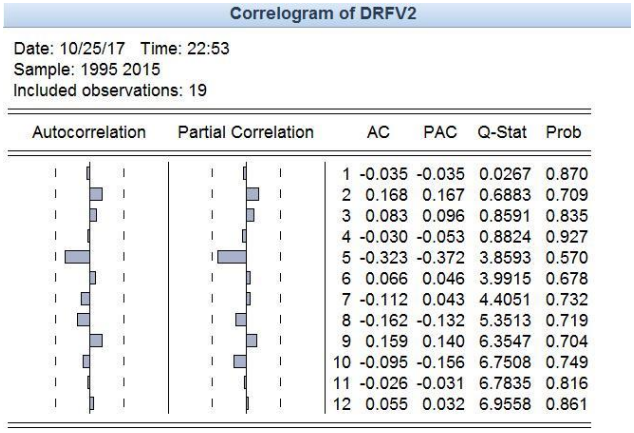
Table 4. Results of sequence {DRFV2} ADF test

		t-Statistic	Prob.*
Augmented Dickey-Fuller Test Statistic		-4.138708	0.0056
Test critical values:	1% level	-3.857386	
	5% level	-3.040391	
	10% level	-2.660551	

Model Identification and Analysis

Using the software EVIEWS 8 to export the second order difference autocorrelation and partial autocorrelation function diagram of {DRFV2}, as shown in Figure 3, we can see that the sequence {DRFV2}'s ACF function and PACF function were trailing. And sequence {RFV}

changed into a stationary sequence by calculating the second order difference operation, so we can identify ARIMA (p, 2, q) model to forecast railway freight volume.


Correlogram of DRFV2

Date: 10/25/17 Time: 22:53
Sample: 1995 2015
Included observations: 19

	Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
1			-0.035	-0.035	0.0267	0.870
2			0.168	0.167	0.6883	0.709
3			0.083	0.096	0.8591	0.835
4			-0.030	-0.053	0.8824	0.927
5			-0.323	-0.372	3.8593	0.570
6			0.066	0.046	3.9915	0.678
7			-0.112	0.043	4.4051	0.732
8			-0.162	-0.132	5.3513	0.719
9			0.159	0.140	6.3547	0.704
10			-0.095	-0.156	6.7508	0.749
11			-0.026	-0.031	6.7835	0.816
12			0.055	0.032	6.9558	0.861

Figure 3. {DRFV2}'s ACF function and PACF function diagram

The actual identification of ARMA (p, q) model requires repeated attempts, there may be more than one group (p, q) values can be identified through the test. There are two criterions types for model selection, Akaike Information Criterion (AIC) and Schwartz Criterion (SC).

In ACI theory:

$$\mathrm{AIC} = 2\mathrm{n} + \mathrm{Tln}\left(\mathrm{RSS}\right) \tag{7}$$

In the function: N is the number of parameters to be estimated; u is a stationary sequence; T is the observed value for use; RSS is the residual sum of squares.

In SC theory:

$$\mathrm{SC} = \mathrm{nln}\left(\mathrm{T}\right) + \mathrm{Tln}\left(\mathrm{RSS}\right) \tag{8}$$

In the function: N is the number of parameters to be estimated; u is a stationary sequence; T is the observed value for use; RSS is the residual sum of squares.

The established ARIMA (P, 2, q) model's different group (p,q) values were tested by AIC and SC theory. Because both AIC and SC introduce a penalty for increasing the coefficients, the smaller the value of AIC or SC is, the better choice of the variable lag order choose. The ARIMA (2, 2, 2) model or ARIMA (3, 2, 3) model was selected according to the AIC criterion, as the values of AIC listed in Table 5.

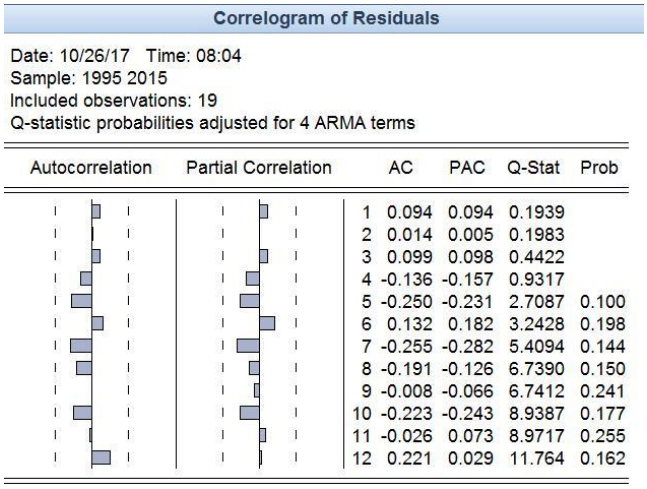
Table 5. AIC value of ARMA model

AIC value of ARMA model		MA		
		1	2	3
AR	1	15.59901	15.69038	15.40952
	2	15.43847	14.92217	15.48437
	3	15.38529	15.38981	15.10451

The SC values of ARIMA (2, 2, 2) model and ARIMA (3, 2, 3) model are calculated respectively. $SC_{2,2,2} = 15.7070 < SC_{3,2,3} = 15.78600$. The ARIMA (2,2,2) model was selected according to the SC criterion.

Model Checking and Freight Volume Forecasting

The residual of the model ARIMA (2, 2, 2) was tested by software EVIEWS 8, that was whether the residual was white noise. If the residual sequence was white noise, the model can be used to predict the freight volume of the Ningxia Hui Autonomous Region railway. The output results were shown in Figure 4, the autocorrelation and partial correlation coefficients of the residuals fall into the random range which proved the residual sequence was white noise. The ARIMA (2,2,2) model was checked and $R^2 = 0.9818169$ shown that the model was highly fitted to the railway freight volume. The fitting graph was shown in Figure 5.



Correlogram of Residuals

Date: 10/26/17 Time: 08:04
Sample: 1995 2015
Included observations: 19
Q-statistic probabilities adjusted for 4 ARMA terms

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.094	0.094	0.1939	
		2	0.014	0.005	0.1983	
		3	0.099	0.098	0.4422	
		4	-0.136	-0.157	0.9317	
		5	-0.250	-0.231	2.7087	0.100
		6	0.132	0.182	3.2428	0.198
		7	-0.255	-0.282	5.4094	0.144
		8	-0.191	-0.126	6.7390	0.150
		9	-0.008	-0.066	6.7412	0.241
		10	-0.223	-0.243	8.9387	0.177
		11	-0.026	0.073	8.9717	0.255
		12	0.221	0.029	11.764	0.162

Figure 4. Residual's ACF function and PACF function diagram

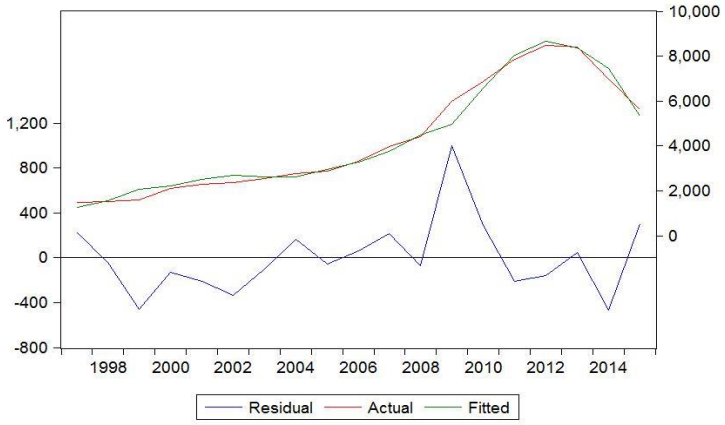


Figure 5. Fitting graph

The ARIMA (2,2,2) model parameters were calculated, and the output results were shown in Table 6. The prediction formula of the Ningxia Hui Autonomous Region railway freight volume was as follows:

$$\varphi(\mathrm{B})(1-B)^2 y_t = c + \varphi(\mathrm{B})\varepsilon_t \tag{9}$$

$$\theta(\mathrm{B}) = 1 - \theta_1 B - \theta_2 B^2 \tag{10}$$

$$\varphi(\mathrm{B}) = 1 - \varphi_1 B - \varphi_2 B^2 \tag{11}$$

$$\hat{y}_t = c + \theta_1 \cdot y_{t-1} + \theta_2 \cdot y_{t-2} - \varphi_1 \cdot \varepsilon_{t-1} - \varphi_2 \cdot \varepsilon_{t-2} \quad (12)$$

$$y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} \quad (13)$$

The prediction formula was obtained from (9) (10) (11) (12) (13)

$$\hat{Y}_t = c + \theta_1 \cdot (Y_{t-1} - 2Y_{t-2} + Y_{t-3}) + \theta_2 \cdot (Y_{t-2} - 2Y_{t-3} + Y_{t-4}) - \varphi_1 \cdot (Y_{t-1} - \hat{Y}_{t-1}) - \varphi_2 \cdot (Y_{t-2} - \hat{Y}_{t-2}) \quad (14)$$

Taking the output parameter in Table 7 into (14), the prediction model of the Ningxia Hui Autonomous Region railway freight volume was as follows:

$$\begin{aligned} \hat{Y}_t &= 3520.376 + 1.915720 \cdot (Y_{t-1} - 2Y_{t-2} + Y_{t-3}) - 1.059502 \cdot (Y_{t-2} - 2Y_{t-3} + Y_{t-4}) \\ &+ 0.641467 \cdot (Y_{t-1} - \hat{Y}_{t-1}) - 0.894061 \cdot (Y_{t-2} - \hat{Y}_{t-2}) \end{aligned} \quad (15)$$

In the function: $\hat{Y}_t$ is the forecast value in year t, Y_t is the actual value in year t.

Table 6. Parameter list of ARIMA (2,2,2) model

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
C	3520.376	836.6264	4.207824	0.0009
AR(1)	1.915720	0.141348	13.55317	0.0000
AR(2)	-1.059502	0.135628	-7.811813	0.0000
MA(1)	-0.641467	0.086385	-7.425672	0.0000
MA(2)	0.894061	0.054615	16.37037	0.0000
R-squared	0.981869			

Table 7. Forecast Railway Freight Volume of the Ningxia Hui Autonomous Region in 1995–2016 years

year	Actual railway freight volume(million tons)	Predicted railway freight volume(million tons)
1997	1478	1255.99
1998	1512	1555.78
1999	1605	2063.36
2000	2105.3	2233.81
2001	2300	2511.46
2002	2367.2	2702.50
2003	2528.5	2630.23
2004	2768.4	2607.49
2005	2881	2936.52
2006	3329	3271.71
2007	3957	3744.79
2008	4400	4474.68
2009	5978	4980.52
2010	6879	6589.92
2011	7850	8057.07
2012	8485.6	8647.55
2013	8412	8363.86
2014	6990	7455.03
2015	5631.1	5325.86
2016		5681.457

According to the forecast model of the Ningxia Hui Autonomous Region Ningxia railway freight volume, the railway freight volume value in 1995–2016 years was calculated, as shown in Table 7. The Ningxia Hui Autonomous Region's rail freight value in 2016 was 5681.457 million tons, an increase of 0.89% over 2015.

CONCLUSION

Using the Ningxia Hui Autonomous Region railway freight volume data, this paper constructs the ARIMA (2, 2, 2) model by using Time Series Forecasting model to analyze and forecast the railway freight volume of Ningxia in 2016. The model's fitting precision on railway freight volume data in the Ningxia Hui Autonomous region shows that this method overcomes the disadvantages of traditional railway freight volume forecasting models, and has a good fitting precision. The forecasting method can provide a more effective and accurate method for predicting the railway freight volume, and has certain reference value for the railway department to formulate the development plan and the industry policy, and implement the concrete measures. Extending the predicting period will reduce the prediction accuracy of the ARIMA model. In the future research if the model parameters can be improved, perhaps the accuracy of long time prediction can be improved as well.

REFERENCES

Bai, X. Y., and Lang, M. Y. (2006). "An improved BP neural network in the railway freight volume forecast." *Journal of Transportation Systems Engineering and Information Technology*, 6,158–162.

Ceselli, A., Gatto, M., Lubbecke, M., Nunkesser, M., and Schilling, H. (2008). "Optimizing the cargo express service of Swiss federal railways." *Transportation Science*, 42, 450–465.

Chen, H., Grant-Muller, S. (2001). "Use of sequential learning for short term traffic flow forecasting." *Transportation Research Part C*, 9(5), 319–336.

Dieter, W. (1997) "Short term forecasting term based on a transformation and classification of traffic volume term time series." *International Journal of forecasting*, 13(1), 63–72.

Guo, K. Q., Ma, Y. H., Wang, T. F., and Sun, L. Y. (2009). "Prediction of railway freight volume based on improved gray-markov chain." *Journal of Lanzhou Jiaotong University*, 6, 132–137.

Huang, Y., and Xu, J. H. (2010). "The railway freight volume forecasting based on grey models." *Railway Transportation and Economy*, 32(4), 86–89.

Hu, J. Q., and Li, Z. P. (2014). "Forecasting of total Social freight volume based on time series." *Journal of Logistics Technology*, 33(5), 128–130.

Li, S., Xie, Y. L., and Wang, W. X. (2012). "Application of AdaBoost-BP neural network in prediction of railway freight volumes." *Computer Engineering and Applications*, 48(6), 233–248.

Lei, B., Tao, H. L., and Xu, X. G. (2012). "Railway freight volume prediction based on grey neural network with improved particle swarm optimization." *Journal of Computer Applications*, 32(10), 2948–2962.

Li, B. (2011). "Prediction of railway freight volume based on system cloud grey SCGM(1,1) model." *Science Technology and Engineering*, 11(5), 1120–1124.

Smith, B. L., Williams, B. M. and Oswald, K. O. (2002). "Comparison of parametric and nonparametric models for traffic flow forecasting." *Transportation Research Part C*, 10(4), 303–321.

Wu, X. L., Fu, Z., Wang, X., and Zhan, H. B. (2009). "Combine-forecast method for railway freight volumes." *Journal of Railway Science and Engineering*, 6(5), 88–92.