Forecast of China Railway Freight Volume by Random Forest Regression Model

Junning Gao
School of Economics and Management
Beijing Jiaotong University
Beijing, P. R. China
14125376@bjtu.edu.cn

Xiaochun Lu
School of Economics and Management
Beijing Jiaotong University
Beijing, P. R. China
xclu@bjtu.edu.cn

***Abstract*—The forecast of railway freight volume has important influence on effective allocation of railway resource. In the paper, we introduced a novel non-linear regression method: random forest regression (RFR), to quantitatively estimate China railway freight volume. Through analyzing the monthly data on railway freight volume between 2001 and 2013 by RFR model, we get a series of predicted results and the results show that Mean Absolute Error and Mean Relative Error are respectively 736.15 million tons and 3.32%. The RFR model has the characteristics of high precision of prediction, strong generalization ability, good robust performance and less adjustable parameters.**

Keywords—railway freight volume forecast; random forest regression; R language

I. INTRODUCTION

Railway freight transportation, as an important part of modern integrated transportation system, is the main transportation way of crude oil, coal, grain and other bulk materials. Therefore, it has an important influence on the growth of national economy. With the increasingly fierce competition of freight transportation market, the situation of railway freight transportation is not optimistic. According to the data of China Railway Company, the total freight volume and total turnover volume compared with the same period last year decreased respectively 4.3% and 6.1% from January to November, 2014. But considering that in 2013 the railway freight revenue accounted for 61.48 % of the total railway income, it is important for China Railway Company to pay attention to railway freight transportation. In this case, accurate prediction of railway freight volume, not only contributes to reasonable allocation of internal resources of the railway company but also provides for an important basis to evaluate economic and social benefits of railway construction projects.

II. LITERATURE REVIEW

In order to improve the accuracy and generalization ability of the railway freight volume forecasting result, many scholars have studied different methods. Early, railway freight volume forecasting is mainly by using regression analysis method or time series forecast method [1-2]. But exogenous variables which will influence future freight volume has certain fuzziness and uncertainty. Therefore, the precision will be affected at some extent. In recent years many scholars have studied this problem by using neural network model and fuzzy mathematics method. For example, Guo Dongdong and Wang Xifu have constructed the prediction model of railway freight volume based on BP network[3]; Zhang Tao and Zhang Jinlong have proposed GRNN network to eliminate the railway freight volume[4]; Song Rui and Sun Yan established hybrid RBF neural network model to forecast freight volume[5]; Xie Jianwen proposed the method of unbiased grey fuzzy Markov chain to forecast railway freight volume[6].This kind of method has made some achievement, but the disadvantage is that it has large dependence on the training samples which will conclude less robustness of the prediction model obtained from the training data. Besides, some scholars applied combined methods to the forecast. Combining GA and BP method, Li Ping proposed a new forecast model [7]; Wang Qingrong established the forecast model by the combination of network and Holt-Winter [8].

This paper introduces the random forest regression (RFR) method, which is an important application of random forest. It is a kind of statistical learning theory developed by Breiman. The algorithm has advantages of high prediction accuracy and good generalization ability, fast convergence speed and less adjustable parameters. So it has been widely applied in medicine, economics and many other fields. For example: Cui Dongwen have used the random forest regression method to predict waste water discharge[9]; Based on RFR, Qiu yiHui has established a forecast model of telecommunication customer loss[10]; Li Zhenzi et al. have studied RFR and its application in the relationship of metabolic regulation[11]. Considering that RFR method has not been applied in railway freight volume forecasting, this paper presents research on the forecast of railway freight volume based on RFR model.

III. RANDOM FOREST REGRESSION (RFR)

RFR is formed through the growth of decision tree associated with a random vector and the difference between RFR with random forest classification is the estimates of RFR are numeric variables. The output values of RFR are numeric. We assume that the training sets are extracted independently from the distribution of random vector Y and X. Therefore,

This paper is funded by Chinese Railway Corporation research project (2014F022), Beijing Planning Office of Philosophy and Social Science research project (14JDJGB034), the National Natural Science Foundation (Grant No.71132008) and "EC-China Research Network on Integrated Container Supply Chain" Project (Project No.612546) as well.

978-1-4799-1891-1/15/$31.00 ©2015 IEEE

mean squared error of any numerical prediction values $h(X)$ is:

$$E_{X,Y}[Y-h(X)]^2 \quad (1)$$

The prediction values of RFR are accomplished by taking averages of k decision trees $\{h(\theta,X_k)\}$.The theorems of RFR are similar to these of random forest classification which are as follows:

(1) Theorem 1

When $k\to\infty$, the conclusion is:

$$E_{X,Y}[Y-avh_k(X,\theta_k)]^2 \to E_{X,Y}[Y-E_\theta(X,\theta_k)]^2 \quad (2)$$

Record the right part of equation (2) as $PE(forest)$, which is also the generalization error of random forest. The average generalization error of each decision tree can be defined as follows:

$$PE(tree)=E_\theta E_{X,Y}[Y-h(X,\theta)]^2 \quad (3)$$

(2) Theorem 2

For all of θ, we assume $EY=E_X h(X,\theta)$, then there is:

$$PE(forest)\le \bar{\rho}PE(tree) \quad (4)$$

$\bar{\rho}$ in the equation (4) is weighted correlation coefficient of residual $Y-h(X,\theta)$ and residual $Y-h(X,\theta')$. θ and θ' are mutual independence.

Because

$$\begin{aligned} PE(forest) &= E_{X,Y}\{E_\theta[Y-h(X,\theta)]\}^2 \\ &= E_\theta E_{\theta'} E_{X,Y}[Y-h(X,\theta)][Y-h(X,\theta')] \end{aligned} \quad (5)$$

The right part of equation (5) is covariance. Therefore, it can be written as:

$$\bar{\rho}=E_\theta E_{\theta'}[\rho(\theta,\theta')sd(\theta)sd(\theta')]/[E_\theta sd(\theta)]^2 \quad (6)$$

Among the equation (6),

$$sd(\theta)=\sqrt{E_{X,Y}\left[Y-h(X,\theta)\right]^2}$$

Define the weighted correlation coefficient as:

$$\bar{\rho}=E_\theta E_{\theta'}\left[\rho(\theta,\theta')sd(\theta)sd(\theta')\right]/\left[E_\theta sd(\theta)\right]^2 \quad (7)$$

Then:

$$PE(forest)=\bar{\rho}\left[E_\theta sd(\theta)\right]\le \bar{\rho}PE(tree) \quad (8)$$

Theorem 2 gives the condition of precise regression forest: the low correlation between residuals and low error of decision tree. RFR can reduce the average error of decision tree through the weighted correlation coefficient $\bar{\rho}$.

Random forest prediction can be viewed as an adaptive neighboring classification and regression process. For each $X=x$, we can get the weight set $\omega_i(x), i=1,2,\cdots,n$ of the n original observations. Random forest forecast or estimation of conditional mean is equivalent to the weighted average values of dependent variable.

The steps of RFR algorithm can be summarized as: set θ as a random parameter vector, which determines the growth of decision tree (such as segmentation on which variables at each node). Corresponding decision tree writes as $T(\theta)$. Set B as X domain. Each leaf node $l=1,\cdots,L$ of decision tree corresponds to a rectangular space of B. Set rectangular space of every leaf node $l=1,\cdots,L$ as $R_l\subseteq B$. For each $x\in B$, if and only if a leaf node l meets $x\in R_l$, remember the leaf node of decision tree $T(\theta)$ as $l(x,\theta)$.

(1) Using bootstrap method to resampling to generate randomly k training set $\theta_1,\theta_2,\cdots,\theta_k$; utilizing each training set to generate a corresponding decision tree $\{T(x,\theta_1),T(x,\theta_2),\cdots,T(x,\theta_k)\}$.

(2) Assuming that feature is M dimension, we randomly extract m features from M dimension feature as the split feature set of the current node, and split this node by the best split way in these m features. (Generally speaking, the value of M is unchanged in the whole growth process of the forest).

(3) Each decision tree has a maximum growth, rather than pruning.

(4) For a new data $X=x$, the average observed value of leaf nodes $l(x,\theta)$ is the prediction value of a single decision tree $T(\theta)$. If an observed value X_i belongs to a leaf node $l(x,\theta)$ and is not 0, the weight vector $\omega_i(x,\theta)$ is:

$$\omega_i(x,\theta)=\frac{1\{X_i\in R_l(x,\theta)\}}{\#\{j:X_j(x,\theta)\}} \quad (9)$$

The sum of the weight in equation (9) is 1.

(5) Under the given independent variable $X=x$, the prediction of a single decision tree is equal to the weighted average of observed values $Y_i(i=1,2,\cdots,n)$ of dependent variable. The prediction value of a single decision tree can be obtained by equation (10):

$$\overline{\mu}(x)=\sum_{i=1}^{n}\omega_i(x,\theta)Y_i \qquad (10)$$

(6) Averaging the weight $\omega_i(x,\theta)(i=1,2,\cdots,k)$ of decision tree, we can get the weight $\omega_i(x)$ of each observed value $i\in(1,2,\cdots,n)$:

$$\omega_i(x)=k^{-1}\sum_{i=1}^{k}\omega_i(x,\theta_i)y \qquad (11)$$

(7) For all of y, the prediction of random forest can write as:

$$\overline{\mu}(x)\sum_{i=1}^{n}\omega_i(x)Y_i \qquad (12)$$

Therefore, under the given conditions $X=x$, the estimation of Y's conditional mean is equal to the weighted sum of observed values of the whole dependent variables. The weight varies with changes of the independent variable $X=x$, and the more similar the conditional distribution of Y given under $X=X_i(i\in\{1,2,\cdots,n\})$ and under $X=x$, the greater the weight.

IV. CHINA RAILWAY FREIGHT VOLUME FORECAST BASED ON RFR

A. Variables Determination of RFR Model

The input of the RFR model is 6 factors which influence the railway freight volume and the output variable, namely the explained variable, is the railway freight volume (y).

The initial RFR model is:

y=f(x1,x2,x3,x4,x5,x6)

Then, we will explain the reason why we choose these 6 factors as the input of the RFR model. According to the data of the statistics center of Railway Corporation, the railway freight transport mainly includes 28 kinds of goods. The freight quantity of coal, steel and crude oil accounted for 51.15%, 7.12% and 4.15% of general quantity. Therefore, we select the finished steel production(x1), raw coal production(x2), crude oil processing production (x3) as the influence factors of freight volume. In addition, the thermal power capacity affects directly the coal freight volume. Therefore, the thermal power capacity (x4) is selected as an influence factor of freight volume. Railway freight volume subject to the number of railway vehicle while fixed assets investment directly affect the number of railway vehicle. Therefore, fixed assets investment (x5) is selected as an influence factor of freight volume. Finally, the macro economic development will inevitably influence the freight volume of railway companies. Therefore, we select growth rate of the second industrial added value (x6) as a influence factor. To sum up, this paper selects 6 indicators to fully express affecting factors of railway freight volume, the 6 indexes include: finished steel production (x1), raw coal production (x2), crude oil processing production (x3), the thermal power capacity (x4), fixed assets investment (x5), growth rate of the second industrial added value (x6).

B. Data Processing

We choose 156 groups statistical data of railway freight volume as the subject. The data comes from the statistics center of Chinese Railway Corporation and the time horizon is from 2001 January to 2013 December. The date is divided into half. The first half of the data is training samples the function of which is to establish the model and the remaining half of the data is test samples which are used to measure the model's prediction ability. Since it isn't sensitive for random forest method to data quantities and units, there's no need to Standardize or normalize the data. The sample data are listed in table 1.

TABLE I. THE SAMPLE DATA LIST

month	freight volume (Million tons)	finished steel production (Million tons)	raw coal production (Million tons)	crude oil processing production (Million tons)	thermal power capacity (Billion kwh)	fixed assets investment (Million)	Growth rate of the second industrial added value (%)
	y	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*
2001-01	137	10.91	63.45	16.26	88.02	7.166	2.30
2001-02	131	11.31	67.97	15.75	100.43	7.186	19.00
2001-03	152	13.23	82.49	18.11	88.72	7.218	12.10
2001-04	147	12.85	79.42	18.65	95.11	7.259	11.50
2001-05	155	13.28	80.62	19.02	93.66	7.304	10.20
2001-06	149	13.61	80.93	18.32	94.65	7.347	10.10
2001-07	152	12.96	74.14	16.60	106.03	7.383	8.10
2001-08	154	13.41	78.09	16.90	100.99	7.434	8.10
2001-09	151	13.16	81.18	17.79	97.31	7.496	9.50
2001-10	158	13.54	83.89	18.50	96.62	7.564	8.80
2001-11	149	15.34	88.06	18.76	138.63	7.617	7.90
2001-12	151	13.86	100.18	16.71	76.55	7.651	8.70

C. Performance Evaluation Standard of the Model

In order to evaluate fitting degree and predict ability of the model, this research takes the following three indexes as evaluation standard. Respectively, the mean absolute error (MAE), the mean relative error (MRE), the normal mean square error (NMSE). The indexes reflect the difference between the predicted value and the true value. The smaller the indexes, the more strong the prediction ability of the model.

$$MAE = \frac{1}{n}\sum|\hat{y}_i - y_i| \qquad MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{y_i} - y_i|}{y_i}\times 100\%$$

$$NMSE = \frac{\sum(\widehat{y_i} - y_i)^2}{\sum(\bar{y_i} - y_i)^2}$$

D. Modeling Tool

The software tool which is used in this paper is open source software, R language. It includes the randomForest package which can build a random forest model easily. Because R language has more powerful function of object-oriented than other statistical or special mathematical programming language, the software is widely used in foreign countries.

Table 2 lists simply some main functions used in the random Forest package.

TABLE II. THE MAIN FUNCTIONS OF RANDOMFOREST PACKAGE

Function	Effect
randomForest	Build RFR model
Importance	List important factors
VarImpPlot	Draw important factors
Plot	Draw the curve of error
Predict	Predict the model

E. Construction and Perfection of the RFR Model

First, reading the sample data in R studio and Enter the following command to build the RFR model.

R language source program of RFR is as follows:

```
names (A)
dim (A)
n<-dim(A)[1]
m=sample(1:n,ceiling(n/2))
n<-100
NMSE<-rep(0,n)
NMSE0<-NMSE
set.seed(100)
for(i in 1:n){
        B=randomForest(y~.,data=A[-m,],ntree=i);
        y0=predict(B,A[-m,]);
        y1=predict(B,A[m,]);

NMSE0[i]<-mean((A$y[-m]-y0)^2)/mean((A$y[-m]-mean(A
$y[-m]))^2);

NMSE[i]<-mean((A$y[m]-y1)^2)/mean((A$y[m]-mean(A$y[
m]))^2)
    }
```

The initial model is:

$$y=f(x1,x2,x3,x4,x5,x6)$$

Run the RFR model and the result is obtained as follows:

Call:

Random Forest (formula = $y \sim .$, data = A[-m,], ntree = i, importance = TRUE, proximity = T)

Type of random forest: regression

Number of trees: 100

No. of variables tried at each split: 2

Mean of squared residuals: 1218032

% Var explained: 94.97

Variance explained was 94.97%. NMSE0 of the training set is 0.011243 and NMSE of the test set is 0.09003496.all of the indexes indicate the result of the model is better.

The importance of each variable MSE is as the following table:

TABLE III. THE IMPORTANCE OF EACH VARIABLE MSE

variable	%IncMSE	IncNodePurity
x1	7284302	789353113
x2	1902440	268539063
x3	6930674	742966196
x4	3547349	458508365
x5	5591733	750791314
x6	501451	52108959

From the importance of variables we can see that finished steel production (x1), raw coal production (x2), crude oil processing production (x3), the thermal power capacity (x4) and fixed assets investment (x5) are very important in RFR while the effect of growth rate of the second industrial added value (x6) in the model is too small. Accordingly, we will remove the variable x6 and establish the RFR model:

$$y=f(x1,x2,x3,x4,x5)$$

Run the RFR model and the result is obtained as follows:

Call:

randomForest(formula = $y \sim x1 + x2 + x3 + x4 + x5$, data =$A[-m,]$, ntree = i, importance = TRUE, proximity = T)

Type of random forest: regression

Number of trees: 100

No. of variables tried at each split: 1

Mean of squared residuals: 1200902

% Var explained: 95.04

Variance explained was 95.04%. NMSE0 of the training set was 0.01729747 and NMSE of the testing set was 0.07733949. All of the indexes indicate removing the variable x6 has little effect on the model. Fig.1 is NMSE curves. We can see the decreased speed of NMSE is rapid and the late of NMSE0 and NMSE curves are relatively stable, no ups and downs, which show that the model does not appear excessive fitting and the operation results is ideal.

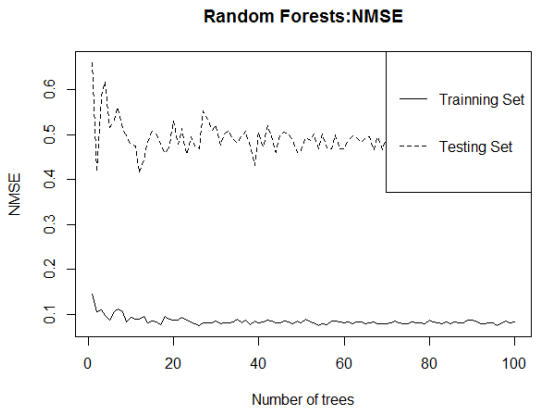

FIG I. NMSE OF RFR

Choosing different independent variables will get different RFR results. These results have listed in table 4. From table4, we can see variance proportion of the second model is higher which means the model has strong adaptability. Besides, the training set NMSE0 of the second model is small and the gap between NMSE0 and NMSE is also small respectively. Therefore, we choose the second model as the final model. In other words, the model result will be ideal if we choose finished steel production, raw coal production, crude oil processing production, the thermal power capacity and fixed assets investment as independent variables.

TABLE IV. RFR RESULTS OF DIFFERENT VARIABLES

The model number	Independence variables	Training set NMSE0	Testing set NMSE	Variance (%)
1	f=g(x1,x2,x3,x4,x5,x6)	0.011243	0.09003496	94.97
2	f=g(x1,x2,x3,x4,x5)	0.01729747	0.07733949	95.04
3	f=g(x1,x3,x4,x5)	0.02622403	0.06513233	92.87
4	f=g(x3,x5)	0.02597002	0.04209281	90.30
5	f=g(x1,x3)	0.02030649	0.1016111	92.34

Using the second model, we can get the forecasting results as shown in Fig.2. In the figure, circle represents the predicted value and curves the historical statistics. We can see the predicted value and the true value are consistent.

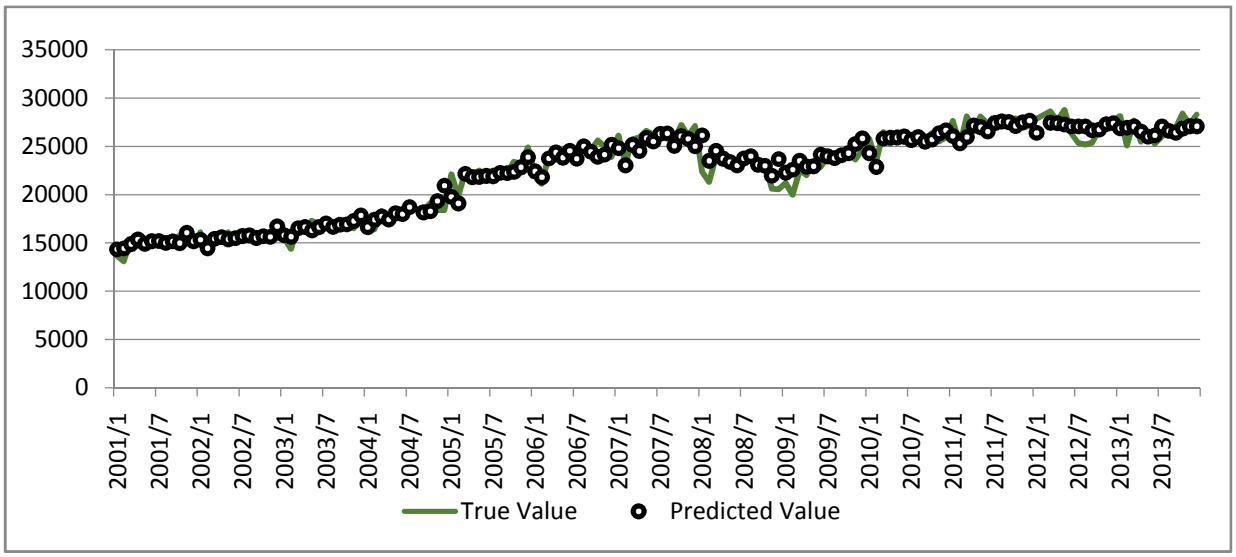

FIG II. COMPARISON BETWEEN TRUE VALUES AND PREDICTED VALUES

F. Analysis of Model Error

In order to evaluate prediction ability of the final model, we have calculated the corresponding results of RFR model based on the three indicators described in Section 4.2. Besides, in order to contrast and analyze prediction ability of different modeling methods, we also build the multiple regression model and BP neural network model by using the same training sample. The results are listed in table 5. Seen from table 5, the results of RFR model are the minimum value in three indexes which indicates that the RFR method has the following advantages relative to other modeling methods: the forecasting error is smaller and prediction and anti-disturbance ability is very strong.

TABLE V. THE PREDICTION ERROR COMPARISON OF DIFFERENT MODELING METHODS

Modeling method	MAE	MRE	NMSE
Random forest	736.15	3.32%	0.042
BP network	1008.86	4.73%	0.126
Multiple regression	1403.46	6.58%	0.227

V. CONCLUSION

It has great importance for the railway internal resources allocation to improve the forecasting accuracy of railway freight volume. This paper builds and improves the RFR model of railway freight volume based on six influence factors and contrasts the model error of three different modeling methods. The results show that the prediction of RFR model

are ideal. MAE and MRE of estimation are respectively 7.23 million tons and 3.28%. The RFR method has the following advantages relative to other modeling methods: the forecasting error is smaller and ability of prediction and anti-disturbance are very strong.

ACKNOWLEDGMENT

This paper is funded by Chinese Railway Corporation research project (2014F022), Beijing Planning Office of Philosophy and Social Science research project (14JDJGB034), the National Natural Science Foundation (Grant No.71132008) and "EC-China Research Network on Integrated Container Supply Chain" Project (Project No.612546) as well.

REFERENCES

[1] Song Guangping, "Research on Railway freight volume forecasting methods," Beijing: Beijing Jiaotong University, 2007.

[2] Wang Fang, "The short-term forecasting method of railway passenger traffic volume," Beijing: Beijing Jiaotong University, 2006.

[3] Guo Dongdong, Wang Xifu, "Prediction on railway freight volume based on BP network," railway freight transport, 2006, 1 (2), pp. 21-23.

[4] Zhang Tao, Zhang Jinlong, "Study on the forecast of Freight Volume Based on GRNN network," Logistics Sci-Tech,2014,1(10), pp.138-141.

[5] Song Rui, Sun Yan, "The forecast of Railway Freight Volume based on hybrid RBF neural network. Journal of Wuhan University of Technology," (Transportation Science & Engineering), 2014,38(6), pp.1247-1250.

[6] Xie Jianwen, Zhang Yuanbiao, Wang Zhiwei, "Prediction on railway freight volume by unbiased grey fuzzy Markov chain method," Journal of China Railway, 2009, 31 (1), pp.1-7.

[7] Li Ping, "Study on the combined forecast of railway freight volume by GA-BP model," Journal of Lanzhou Jiaotong University, 2014,33(3), pp.203-207.

[8] Wang Qingrong, "The combined forecast of railway freight volume by network and Holt-Winter model," Journal of Lanzhou Jiaotong University, 2010,29(4), pp.122-124.

[9] Cui Dongwen, "Random forest regression model and its application in waste water discharge volume prediction," Water Technology, 2014, 8 (1), pp. 31-36.

[10] Qiu yiHui, "Application of random forest in telecom customer churn prediction," Fujian: Xiamen University, 2008.

[11] Li Zhenzi, Zhang Tao, Wu Xiaoyan et al, "Analysis of random forest regression and its application in the relationship of metabolism regulation," China health statistics, 2014, 29 (2), pp. 158-160.