Team Control Number

202011211768

Problem Chosen

A

2020

ShuWei Cup

Summary Sheet

Summary

Over the past 30 years, China has become the world's largest infrastructure market. Under such situation, as widely used in civil engineering construction, rebar has become one of the most important steel products with the largest output. To better deepen the supply-side structural reform in the steel industry and alleviate the overcapacity in the steel industry, it's of great significance to predict and grasp the demand dynamics of rebar.

For question one, there are many factors affecting the demand of rebar, the time period and the time resolution of these variables varies. Thus, we analyzed the data and unified the time resolution and range. Then the multivariate interpolation method was employed to fill in the missing values for further modeling and analysis. Due to the uncertainty brought by the large amounts of variables, we selected 15 most important variables related to rebar demand based on random forest (RF).

For question two, we employed random forest regressor to predict the apparent demand of rebar. Firstly, we calibrated the parameters of the model. Then the samples were divided into training and testing set. Finally, random forest regressor was trained and used to predict the apparent demand of rebar of testing set. The comparison between the prediction results of random forest and multiple linear regression implied that random forest has a better performance, with a much higher R-square value and lower RMSE and MAE. The R-square value is 0.947, RMSE and MAE maintain relatively low values against actual values.

For question three, to deal with the publication time lag of several factors related to prediction, Seasonal Autoregressive Integrated Moving Average Model (SARIMA) was employed to forecast the values of some factors, which was then taken as input of random forest for further prediction. The SARIMA-RF scheme shows effectiveness in forecasting the apparent demand of rebar, with good consistency between prediction results and actual values and a R-square value of 0.92.

Key word: Random Forest, SARIMA, Multiple imputation, Rebar, Apparent demand forecast

Content

1. Introduction

1.1 Background

Rebar (short for reinforcing bar), known when massed as reinforcing steel or reinforcement steel, is a steel bar or mesh of steel wires used as a tension device in reinforced concrete and reinforced masonry structures to strengthen and aids the concrete under tension. Rebar is an indispensable structural material for infrastructure construction, which is widely used in the various building structures and civil engineering construction of houses, bridges, roads and so on. Rebar is one of the steel products with the largest output in China, which has made great contributions to national industrialization and infrastructure[1].

However, due to the obvious decline of domestic and foreign market demand along with the continuous decline of international commodity prices, the contradiction of overcapacity in the steel industry is particularly prominent[2], so it is of great significance to grasp the demand dynamics of rebar in market reasonably and effectively. From the perspective of national macro-control, predicting the demand of rebar is conducive to restructuring the new pattern of supply and enhancing the industry concentration, so as to deepen the supply-side structural reform in the steel industry, and alleviating the situation of overcapacity in the steel industry. From the perspective of current policy adjustment of environmental protection and commodities trading, the supply side constraints will be relaxed obviously in the future, thus, the investment strategy of rebar futures can be adjusted according to the forecast results of rebar demand.

1.2 Work

***Task* 1.** There are many factors affecting the demand of rebar, so variables may need to be screened or conducted in the establishment process of the prediction model. We should provide the scheme and reasons for dealing with the variables.

***Task* 2.** We should build the demand prediction model of rebar, provide the model construction ideas and schemes, and test the model performance. Different models have different interpretations of the results. We should also explore the influence path between variables and the demand of rebar.

***Task* 3.** In actual operation, the time of data release (update) and data annotation lags behind. For example, most of the monthly data is marked on the last day of each month, while the data is not published until the middle of the next month. The above factors need to be considered when using the model for prediction in practice. We also need to adjust the prediction model to make it closer to the actual application scene, and check the adjusted model.

2. Problem analysis

2.1 Data analysis

Appendix 2 provides a lot of data related to the demand of rebar, according to the variable information provided in Appendix 2 and Appendix 3, we found that the temporal resolution of most data is monthly, and only a small amount of data are provided weekly, Under such conditions, it is difficult and uncertain to predict the apparent demand of rebar every week, therefore, we attempted to aggregate the weekly apparent demand of rebar to a monthly time scale. According to the number of days in two different months in the week, the demand of one week is weighted and divided into the two months because some weeks span two months. We intended to apply the same processing method to cope with other similar situations (such as:the National turnover of construction steel). For quarterly or annual data, we directly assign quarterly/annual values to each month. Finally, we sorted out 65 variables that may affect the demand of rebar, the temporal resolution of these data is unified to monthly ranged from January, 2016 to September, 2020. According to the data of the apparent demand of rebar, we found that it has a significant periodicity (12 months), so compared with the data of the same period last year, it could be a good indicator to the forecast value of this year, thus, we took the data from the same period last year as one of the variables.

2.2 Analysis of question one

In consideration of all these factors may affect the demand of rebar. We need to establish an effective prediction and analysis model to judge the relationship between the apparent demand of rebar and factors (such as cement utilization of capacity, implementation rate of PPP projects, land premium rate, etc.) in our country. According to the analysis and prediction, we propose that the appropriate production capacity of rebar could facilitate the realization of co-ordination of supply and demand of rebar. However, in the process of building the prediction model, it is also necessary to screen a variety of influencing factors on the apparent demand of rebar. Therefore, we utilize Gini Index in Random Forest to screen and process variables.

2.3 Analysis of question two

In order to analyze the relationship between the apparent demand of rebar and the influencing factors, we take the apparent demand(nationwide) of rebar as example, analyzing the common influencing factors, constructing the prediction and analyzing model that can explore the relationship between the apparent demand of rebar and the influencing factors based on Random Forest (RF) algorithm. Furthermore, giving some reasonable suggestions for the supply of rebar.

2.4 Analysis of question three

Due to the time lag of data release (update) and data annotation, in practical situation, we need to consider the problem of time lags actually. According to the above problems, we build an auto regression model to predict the values of several variables related to rebar demand, and providing some reasonable suggestions for the rebar production.

3. Symbol and Assumptions

3.1 Symbol Description

Category	Meaning
AdR	Apparent demand of rebar
Cuc	Cement utilization of capacity
InDA_In	Declared amount of infrastructure project
InDA_Tr	Declared amount of transportation project
InNG_Fr	National government fund revenue
RE_Hc	Housing completion
Tcs	The National turnover of construction steel
nFeatures	The number of factors
nTrees	The number of decision trees
oob	The accuracy of out of bag samples

3.2 Fundamental assumptions

In order to simplify the given problem and modify it to a more suitable simulation. In reality, we make the following basic assumptions. Every assumption has a valid reason.

(1) In consideration of building a prediction and analysis model for apparent demand (nationwide) of rebar. Thus, regardless of the regional differences and major accidents.

(2) Assume that the collected data is accurate and can effectively reflect the problem.

(3) Assume that the time scale of all influencing factors in the prediction model is same: monthly.

(4) Assume that the actual situation is the same as the model.

4. Model

Random Forest (RF) is an integrated machine learning method based on decision tree algorithm. It uses the random resampling technology-bootstrap and node random splitting technology to construct multiple decision trees, and gets the final prediction results by voting. Random Forest (RF) has the ability to analyze the classification and regression characteristics of complex interactions. It also has good robustness to processing nonlinear characteristic and missing values data, and its variable importance measurement can be used as a tool for feature selection.

4.1 Estimation of missing values—Multiple imputation

There are some missing values, for example some data is not available in the given month. Though RF algorithm has good robustness in data missing problem, it is essential to ensure data integrity when training RF. At present, most of the methods to deal with data missing problems is to select the intersection of the required data, which means discard entire rows or columns of data with missing values, or just fill with mean or random values[3]. However, by doing this, it is also likely to discard some valuable data or bring uncertainty to the model. Therefore, we utilized multiple imputation missing value interpolation method to estimate missing data.

Generally, there are many variables in the given data, and the missing place and degree of each variable are different, we can utilize the regression relationship between these variables to impute missing values, this is the basic principle of multiple imputation. For instance, variables with missing values could be modeled as full variables, and the regression method could be used to determine the function relationship and estimate the missing value. Multiple imputation is executed in an iterative loop: In each step, it specifies the target column (variable) that contains the missing value as the output y, considers other columns (full variables) as input X, uses regression(we selected RF regression method in this experiment) to fit (X, y) with the known samples(Not missing) y, then predicts(Missing) y [4]. This is an iterative approach to deal with each feature, then repeat $\max\_iter$ rounds. The results of the last round in the calculation will be returned, and we would get the complete data after interpolation. Generally, this process will firstly proceed the variable with fewest missing values, and then proceed step by step, till the most serious missing variable is filled, then the whole missing value processing is finished.

4.2 Random Forest algorithm

4.2.1 Principle of Random Forest

RF is an algorithm based on the decision tree, which uses bootstrap resampling technology to generate a new training sample set by randomly and repeatedly sampling with replacement method, so as to reduce the correlation between regression models and increase the regression accuracy of regressor. According to the n decision tree generated by bootstrap sample set to compose RF[5]. The regression results of the target are determined by the synthesis results of all decision trees.

Every decision tree in RF is a binary tree, the root node contains all training bootstrap samples. According to certain principles, each node is selected from a set of randomly selected variables, after branching, the variable with the minimum "impurity" of the node is regarded as the branching variable, then we could get left node and the right node, each of which contains a subset of training data[6]. The split nodes continue to split according to the same rules till meeting the branching stop rule (Fig.1). Gini Index and Entropy can be the measurements of "impurity".

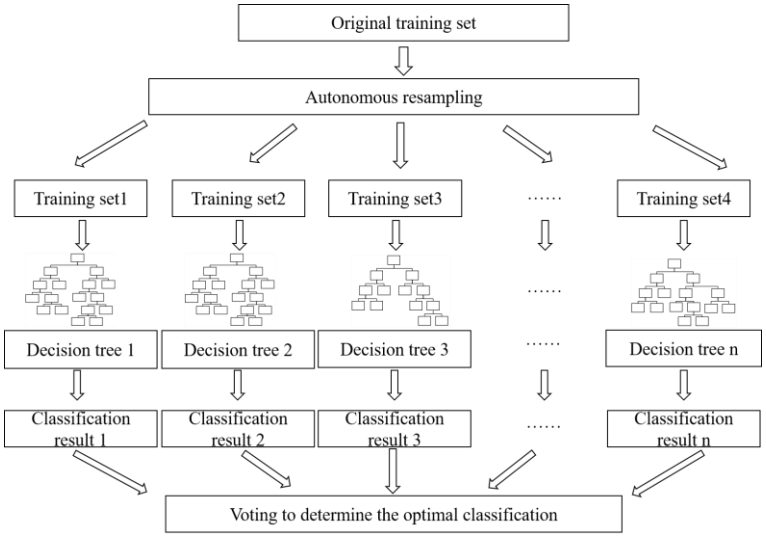

Fig 1. Principle of Random Forest

4.2.2 Feature selection

According to the feature evaluation methods, feature selection could be divided into two algorithms: Filter algorithm and Wrapper algorithm. Filter algorithm is independent of the subsequent machine learning algorithm, which could quickly eliminate a part of noncritical noise features and reduce the search range of optimal feature subset, however, it cannot indicate that the selected subset is a smaller optimal feature subset. By contrast, in the process of feature selection, Wrapper algorithm could directly use the selected feature subset to train the regressor, and easily evaluate

the pros and cons of the feature subset based on the performance of the regressor in the test set. Although this method is not as efficient as former one, the size of the selected feature subset is relatively smaller. This article utilizes RF algorithm as a basic tool to conduct Wrapper algorithm—the conventional variable importance score (VIM) of RF could be calculated by Gini Index.

The model based on the assumption that there are variables $X_1, X_2, \ldots, X_M$, and variable X_j can be obtained by eq.1 to eq.4:

In one decision tree, the Gini Index of node m is given by eq.1：

$$G_m = \sum_{k=1}^{K} P_{mk}\left(1 - P_{mk}\right) \quad (1)$$

where G_m is Gini Index of node m; K is the number of classes in the sample set; P_{mk} is the estimation of probability value that pertained to the class k at the node m.

The importance of variable X_j at the node m, which means the change of Gini Index before and after node m splitting is given by eq.2:

$$V_{jm}^{Gini} = G_m - G_{ml} - G_{mr} \quad (2)$$

where V_{jm}^{Gini} is the importance of variable X_j at the node m, G_{ml} is the Gini Index of the left node of node m, G_{mr} is the Gini Index of the right node of node m.

If X_j appears in the i-th tree for M times, the importance of variable X_j in the i-th tree could be calculated by eq.3:

$$V_{ij}^{Gini} = \sum\nolimits_{m=1}^{M} V_{jm}^{Gini} \quad (3)$$

where V_{ij}^{Gini} is the importance of variable X_j in the i-th tree, M is the frequency of occurrence of variable X_j in the i-th tree, V_{jm}^{Gini} is the importance of X_j at the node m.

If RF has N trees，then the importance of X_j in the RF could be defined as the average value of importance in all trees (eq.4):

$$V_j^{Gini} = \frac{1}{N}\sum\nolimits_{n=1}^{N} V_{ij}^{Gini} \quad (4)$$

where V_j^{Gini} is the Gini importance of X_j in RF, N is the number of decision tree in the RF, V_{ij}^{Gini} is the importance of X_j in i-th tree.

4.2.3 Accuracy assessment of Random Forest

In the bootstrap resampling process of RF, the probability that each sample in the original training set (N samples) not selected is $\left(1-1/N\right)^N$. This implies that there are approximately $1/3$ data in the training set are not selected. These unselected data are called out of bag samples (OOB), we can obtain the OOB precision estimation of each decision tree through OOB. Moreover, by averaging the OOB accuracy estimates of all decision trees in the RF, we could calculate generalization accuracy of RF.

4.3 Principal component analysis and Regression

Principal component analysis (PCA) can transform multiple correlated features into several incoherent comprehensive feature components, that is to say, it can reduce the dimension of the original factors, omit some irrelevant indicators, then the higher dimension factors could be integrated to the fewer comprehensive components which can reflect the original situation. The specific method is to carry out the feature analysis of covariance matrix, which can reduce the dimension of data and maintain the maximum contribution rate of variance[7].

Assume that there are N selected factors, and its original sample matrix X is eq.5：

$$X = (X_i) * N, i = 1, 2, \ldots n. \tag{5}$$

The correlation coefficient matrix of each index is R_{n*n} , its eigenvalue $\wedge 1 \geq \wedge n \geq 0$, and regularization eigenvector e_i , then the principal component T_i can be obtained by eq.6：

$$T_i = X_{e_i} \tag{6}$$

When the variance contribution rate of the i-th principal component is between 85% and 95%, take the first q principal component $T_1, T_2, \ldots, T_q$, so q principal component could reflect the N original factors. Its variance contribution rate could be obtained by eq.7:

$$a = \sum_{i=1}^{q} a_i \tag{7}$$

We utilize the principal components extracted from PCA to conduct multivariable linear regression. Multiple linear regression is an important method in multivariate statistical analysis. Aiming at time series data in this paper, we build a multiple linear regression prediction mode as eq.8[8].

$$Y_i = \beta_0 + \sum_{i=1}^{K} \beta_i X_{T_i} + X_T. \quad (i = 1, 2, \ldots, n) \tag{8}$$

where k is the data of explanatory variables and $\beta_i (i = 1, 2, \ldots, k)$ is called the regression coefficient.

4.4 Seasonal ARIMA Model

Actually, many variables needed by the prediction of Apparent demand of Rebar (AdR) will not be released or updated until the end of the period before which AdR should be forecast. Therefore, the prediction model established above is no longer suitable in practice. Thus, we employed Seasonal Autoregressive Integrated Moving Average Model (SARIMA) to predict the values of the variables needed by the prediction of AdR.

When it comes to time-series forecasting, SARIMA has great popularity in many

respects. Many of the variables related to AdR have random series present as a periodic change, leading to non-stationary which should be adjusted in time-series forecast. SARIMA[9] can eliminate the periodicity influence or the non-stationary in a prediction process and thus is a widely applied model for forecasting seasonal time series.

The seasonal ARIMA (SARIMA) model incorporates both non-seasonal and seasonal factors in a multiplicative model(eq.9):

$$SARIMA(p,d,q)\times(P,D,Q)s. \qquad (9)$$

where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order, and S is the time span of repeating seasonal pattern.

SARIMA is first given by eq.10 [10]:

$$\psi(B)\Phi(B^S)(1-B)^d(1-B^S)^D Y_t = \Theta_0 + \theta(B^S)\varepsilon_t. \qquad (10)$$

where $\Psi(B)=1-\psi_1 B-\psi_2 B^2-\cdots-\psi_p B^p$ is the p-order non-seasonal AR model,

$\theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_p B^p$ is the q-order non-seasonal MA model,

$\phi(B^S)=1-\phi_1 B^S-\theta_2 B^{2S}-\cdots-\phi_p B^{PS}$ is the P-order seasonal AR model,

$\Theta(B^S)=1-\Theta_1 B^S-\Theta_2 B^{2S}-\cdots-\Theta_p B^{QS}$ is the Q-order seasonal MA model.

$(1-B)^d$ represents the non-seasonal differencing of order d.

$(1-B^S)^D$ represents the seasonal differencing of order D.

ε_t is the error term $\sim \mathrm{II}(0,\delta^2)$, B is the backshift operator, S is the seasonal order.

In addition, four-step iterative cycles are needed to fit an S-ARIMA mode[11].

(1) Identify the structure of the SARIMA $(p,d,q)(P,D,Q)s.$ model;

(2) Estimate unknown parameters;

(3) Perform goodness-of-fit tests on the estimated residuals;

(4) Forecast future outcomes based on the known data.

In our study, a python statistical library named pmdarima was employed to estimate the optimal SARIMA parameters for each variable.

4.5 Prediction and evaluation Index

Tab. 1 Evaluation Metrics of Model

Metrics	Index calculation equation	Parameter description
R-square	$\frac{\sum_{i=1}^{I}(y_i-\bar{y})^2-\sum_{i=1}^{I}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{I}(y_i-\bar{y})^2}$	y represents the true value
RMSE	$\sqrt{\frac{1}{I}\sum_{i=1}^{I}(y_i-\hat{y}_i)^2}$	$\hat{y}$ represents the predicted value
Variance	$\frac{1}{I}\sum_{i=1}^{1}\left((\bar{y}-y_i)^2\right)$	$\bar{y}$ represents the Average of true values

5. Test the Models

5.1 Results and solution to Problem 1

5.1.1 The importance assessment of impact factors

There are many factors affecting the demand of rebar, such as cement price, steel price, housing loan interest rate, real estate price, etc. We selected 65 impact factors that may affect the demand of rebar from five major categories (rebar, cement, loan, infrastructure construction and real estate) of impact factors. Attached Appendix A displays the name and abbreviation of specific impact factors. However, simply taking all these factors into consideration may lead to over consumption of computing and storage. Moreover, the uncertainty of some unsuitable factors may influence the training effect of RF, which will lead to lower accuracy of prediction results. Thus，before modeling, we need to score the importance of the 65 selected impact factors, and select several important factors.

We took all the 65 features as the training set of random forest in our study. Based on the Gini Index, RF model is employed to calculate the importance of 65 factors which were then normalized by the maximum value (Fig.2). Therefore, in the feature selection, we could select the features with higher importance scores in turn, which will be used as the input factors of RF prediction.

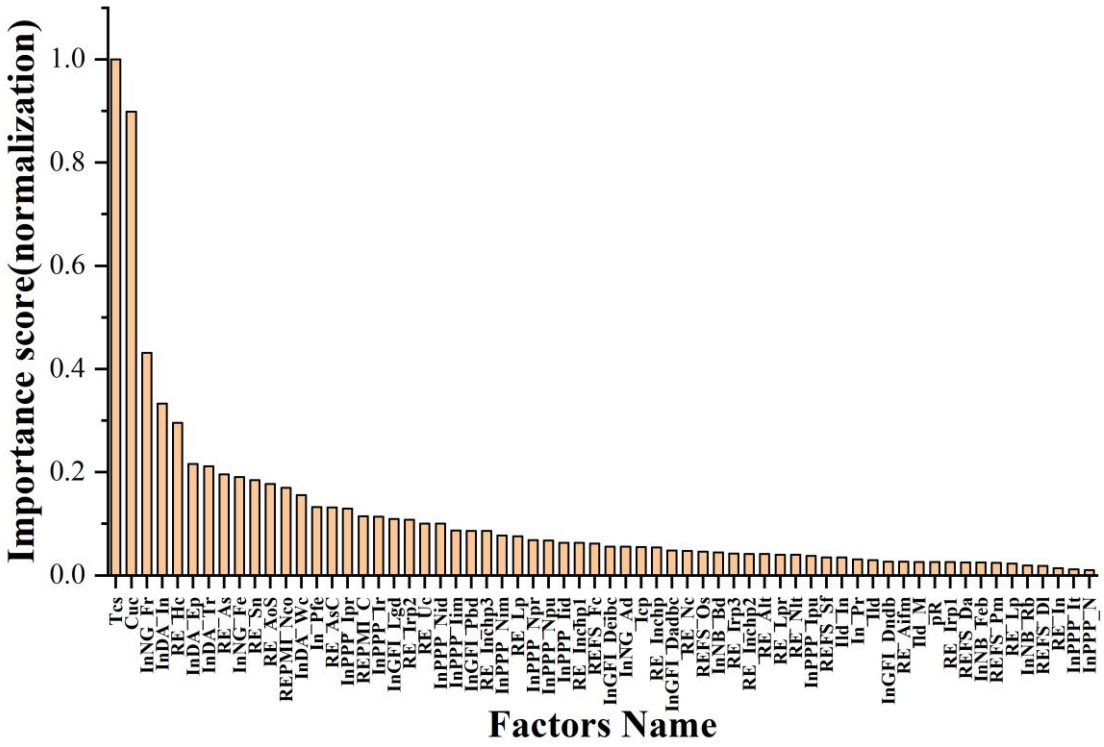

Fig. 2 Normalized importance scores of factors calculated by RF algorithm

5.1.2 Feature selection and parameter calibration

The ensemble machine learning method is generally regarded as a black box model. As one of the ensemble machine learning regressors, the parameters of RF need to be calibrated in order to obtain the best regression results. In RF regressor, we need to consider several aspects: the number of decision trees (nTrees) in RF is needed to be calibrated, the size of feature subset (nFeatures), the minimum number of samples (nSamples) when used in nodes' splitting, and the maximum depth of the tree (max_depth).

It has been concluded that, generally, when regressing: (1) The optimal size of nFeatures is the arithmetic square root of the number of features (nFeatures) in the dataset; (2) When Max_depth is not limited and nSamples = 2, the regression result is the best. When RF regressor is used to select factors, the conclusion is still applicable.

Therefore, the parameters needed to be calibrated in this study are nFeatures and nTrees. When fitting the RF regressor, all the values of apparent demand of rebar and the related factors are taken as the training set. The value of nFeatures was changed, and mean value of OOB accuracy with nTrees of 50, 100, 200 and 500 was calculated, in order to calibrate the parameter nFeatures (Fig.3).

The parameter calibration result of nFeatures shows that the regression accuracy of RF first rises with the increase of nFeatures. When nFeatures = 15, OOB accuracy reaches the maximum value, and then with the increase of nFeatures, the accuracy gradually decreases. Results show that the less important features have little influence on the regression results, and even bring extra noise and errors to the results reducing

the accuracy. Therefore, we selected 15 input factors: Tcs, Cuc, InNG_Fr, InDA_In, RE_Hc, InDA_Ep, InDA_Tr, RE_As, InNG_Fe, RE_Sn, RE_AoS, REPMI_Nco, InDA_Wc, In_Pfe, and RE_AsC. (The specific meanings of these variables are shown in Appendix A)

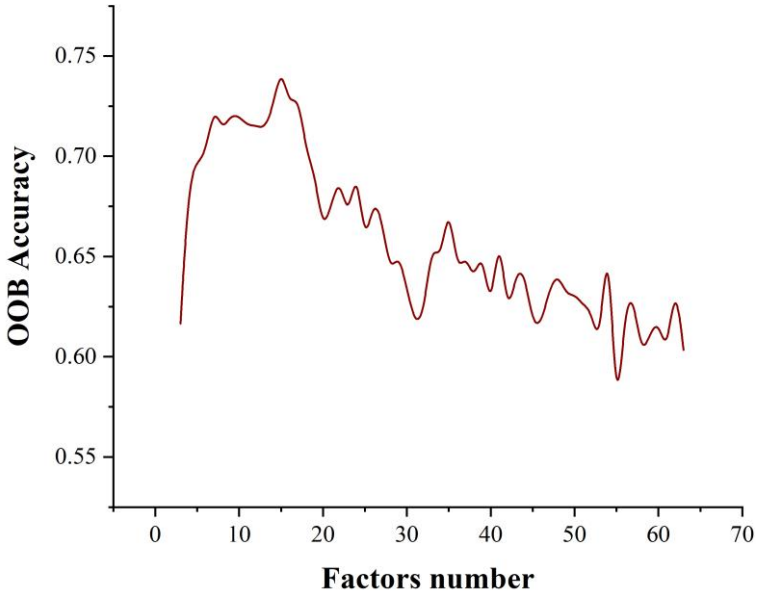

Fig.3 The variability of OOB Accuracy with the increase of factors number

According to the preliminary exploration of the data, we can see that the demand of rebar is cyclical (Fig.4), so when setting the parameters, we predict the demand of rebar this year with the demand data last year due to the periodic changes of data (as stated in section 2.1 data analysis).

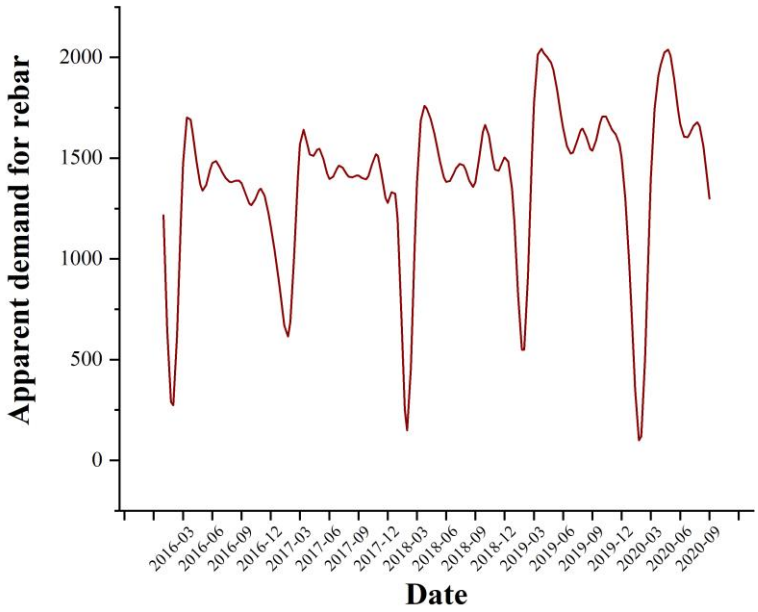

Fig. 4 Time series of apparent demand of rebar

5.2 Results and solution to Problem 2

5.2.1 Parameter calibration

In the solution to the Problem 1, we have identified the variables used to predict the apparent demand of rebar and the number of features in the random forest regressor. According to the characteristics of the RF, we need to calibrate another parameter, that is the number of decision trees in RF, so we took the factors selected in the Problem 1 as the input, with other parameters unchanged and only adjusted the number of decision trees in RF to fit the regressor, then OOB accuracy was output as the objective function. The variability of OOB accuracy and time consumption (seconds) of the prediction with the increase of the number of decision trees is demonstrated in Fig.5. Time consumption is directly proportional to the number of trees. With the increase of the number of trees, the prediction accuracy of the model first increases rapidly, then decreases slightly, and then maintains dynamic stability. So, we decided the final number of decision trees in the RF is 200 while the OOB accuracy is higher and time consumption is less. When the number of trees continues to increase, the accuracy does not increase significantly, but it will bring additional computing time.

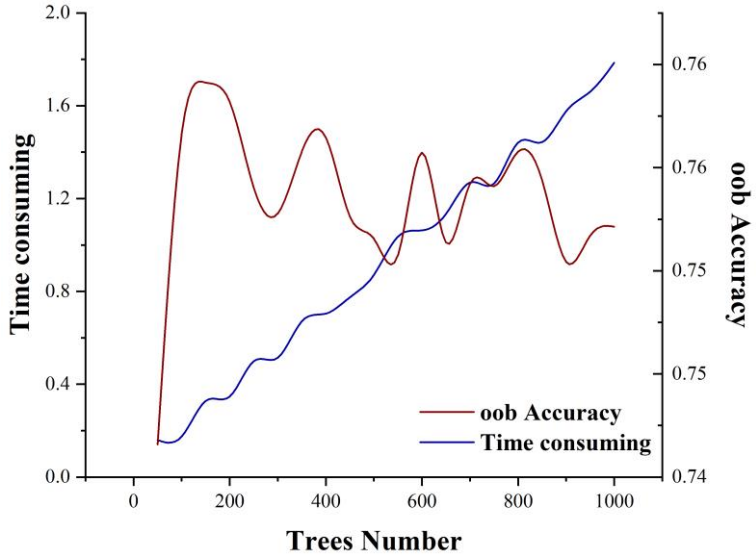

Fig. 5 The variability of OOB Accuracy and Time consuming with the increase of factors number

5.2.2 Random Forest Regressor establishment and evaluation

After data analysis, the time range of all variables is unified from January 2016 to September 2020, and the time resolution of variables is monthly. In case of the apparent demand of rebar in the same period of last year is used as the input factors

for prediction, the available data range is from January 2017 to September 2020 with a total of 45 samples. We took 28 samples (60%) from January 2016 to April 2019 as training set, the remaining 17 samples (40%) from May 2019 to September 2020 as test set to train the model and estimate the results, respectively. At the same time, we also employed multiple linear regression to predict the demand, so as to analyze the results of the model better, and compared it with the RF. Considering that multiple regression is not suitable to the high dimensional data, so we used PCA to reduce the dimensions of 15 variables, and selected five principal components with eigenvalues greater than 1 as new variables.

Similar to the experiment above, based on the 28 samples from January 2017 to April 2019, we established the regression relationship between the five principal components and the apparent demand of rebar, which was then applied to 17 samples from May 2019 to September 2020, and the predicted value was obtained by such multiple regression.

From the prediction results of RF and multiple regression (Fig.6), we respectively calculated the R-square, RMSE and MAE of the results of RF and multiple regression compared with the actual values (Tab.2). The R-square between the prediction results and the actual values of RF is 0.947, which indicated the prediction results are consistent with the actual values, RMSE and MAE remained a low level.

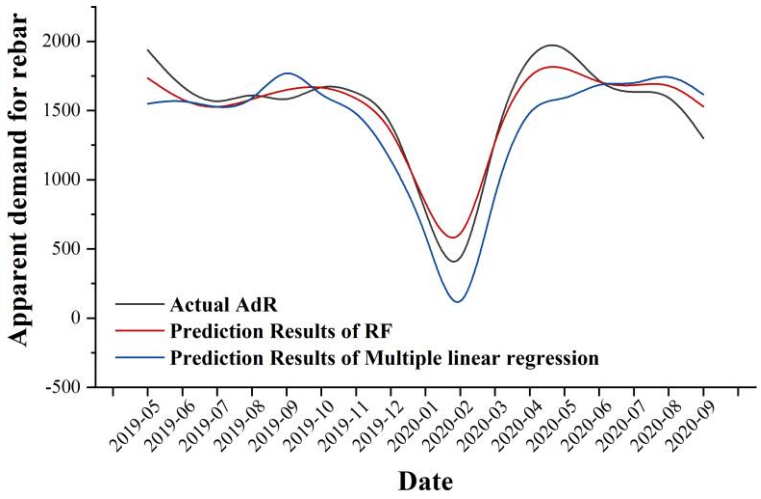

Fig. 6 Time series of prediction results of Random Forest

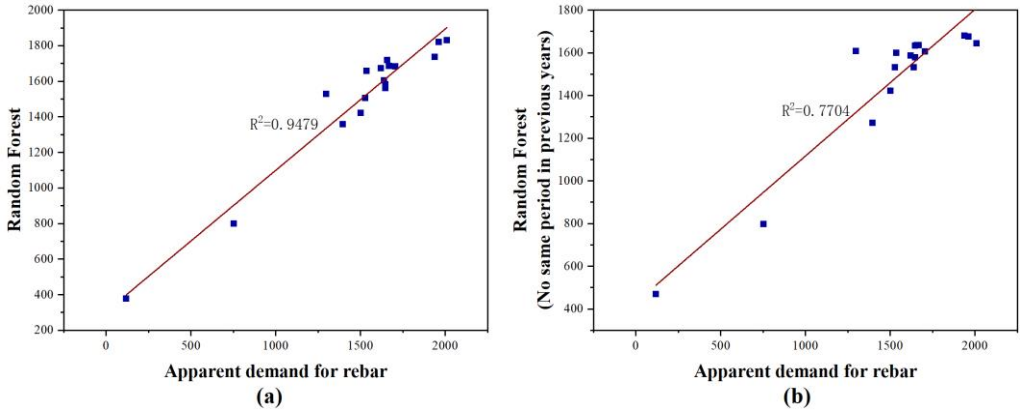

Fig. 7 Scatter plots of Random Forest prediction results against actual values ((a)With the data of the same period last year, (b)Without the data of the same period last year)

Tab. 2 Metrics of the results of random forest and multivariable linear regression

	R-square	RMSE (10kt)	MAE (10kt)
Random Forest	0.947	127.72	96.88
Multivariable linear regression	0.770	242.36	219.42

From the perspective of absolute metrics, the accuracy of RF is much higher than the multiple linear regression, and the errors are only the half. From the variabilities of the predicted values, the prediction results of RF are consistent with the actual values, and the variabilities of the apparent demand of rebar could be well predicted. Although there are still some differences between the prediction results and the actual values at some extreme value points (February, 2020 and May, 2020), compared with the multiple linear regression, the prediction results are still much closer and the variabilities is more consistent. Our results indicate that RF model has a great performance in predicting the apparent demand of rebar.

5.3 Results and solution to Problem 3

In practice, the information of some variables used to predict the apparent demand of rebar cannot be released in time, which may affect the application of RF in prediction. Therefore, we adjusted the model and employed SARIMA model, before prediction, the values of the variables in the next 6 months were forecast by SARIMA.

According to the principle of SARIMA in Section 4.4, we used the pmdarima Library of Python to calibrate the parameters of SARIMA model for each input variable, and then the values of these variables were predicted. Due to the limited space in this paper, we only displayed the forecast of the six most important variables (Fig.8). As can be seen from the figure, the variables with higher importance scores often have periodic changes. The prediction results of these variables are closer to the actual values and the variabilities are consistent. However, when the actual value has a sudden change (Fig.8 (e) (f)), the prediction result is not satisfactory. However, to tell the truth, there are nearly no time series prediction models which could resolve this problem.

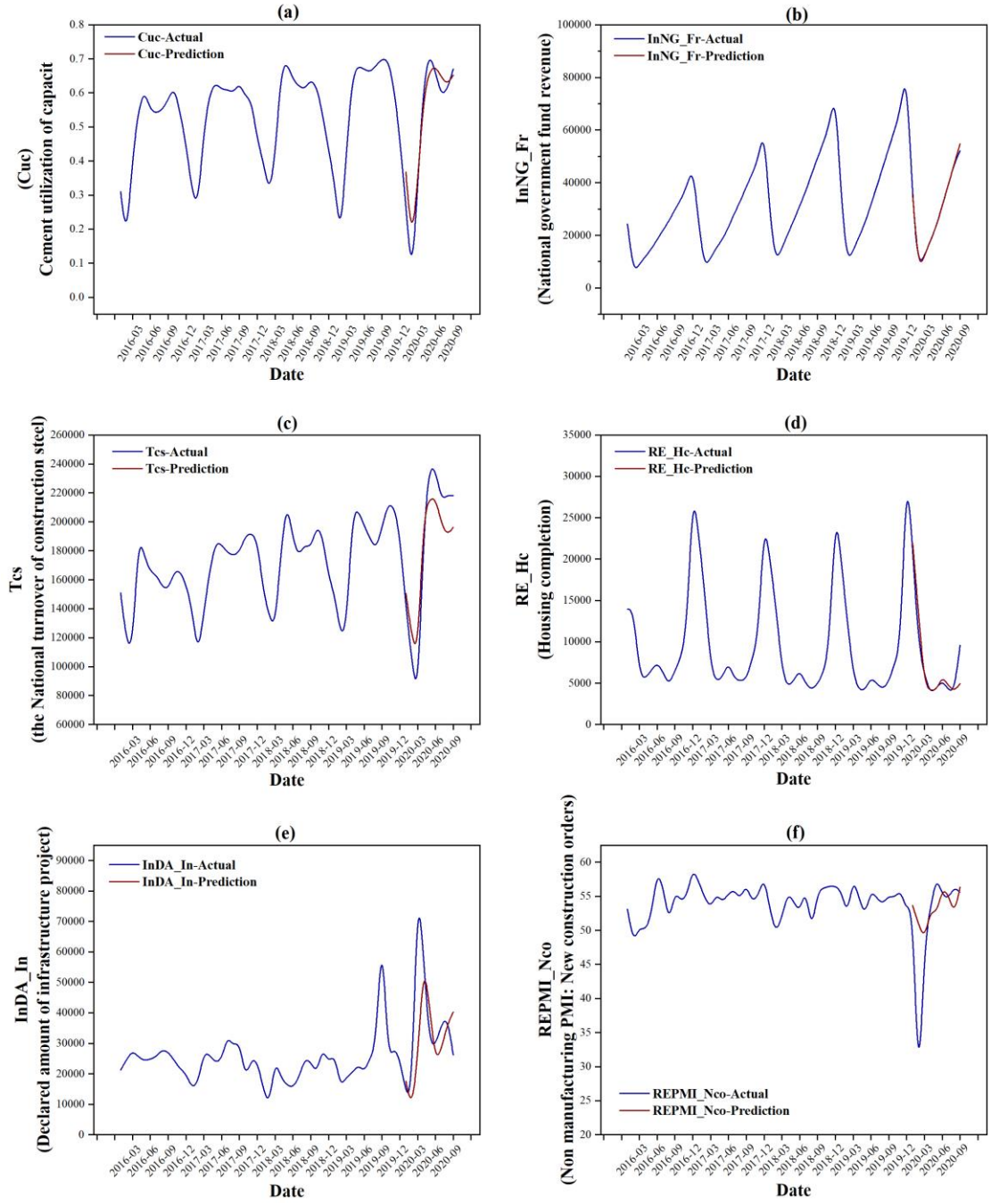

Fig. 8 Time series of several variables and their prediction using SARIMA

Taking the forecast value as the input of random forest to predict the apparent demand of rebar could solve the problem that the data is not released or updated in time, the scatter plot and variabilities between the prediction results and the actual values are shown in Fig.9. When SARIMA-forecast variables are used as input of RF, the accuracy of the prediction results is still high, which shows that the combination of SARIMA and RF model is reliable in predicting the apparent demand of rebar in the future(R-square=0.92，RMSE=133.27(10kt), MAE=103.02(10kt)).

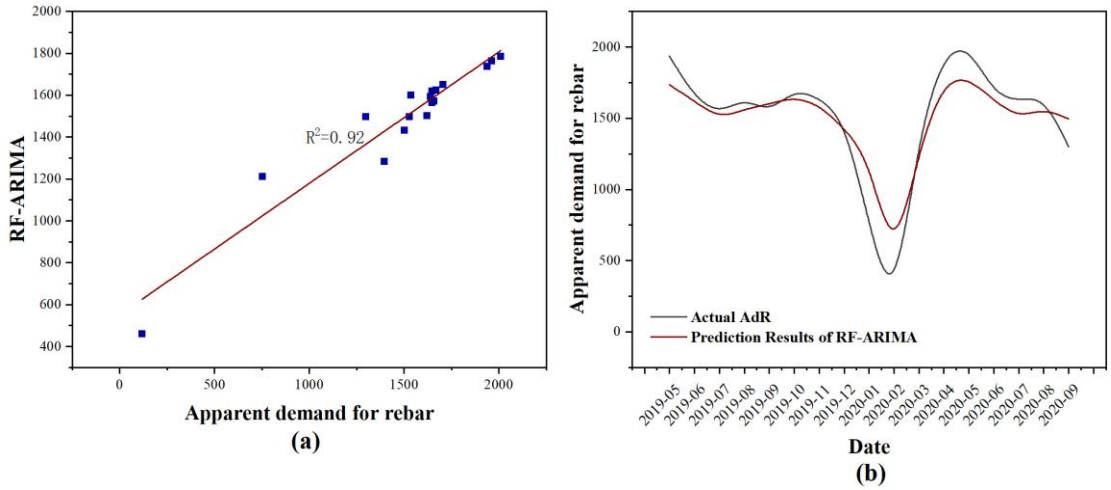

Fig. 9 Comparison between prediction results of RF-SARIMA with actual values((a)Scatter plot, (b) Time series)

6. Sensitivity Analysis

Here we test the sensitivity of the model from two perspectives: the variables and the parameters of random forest. As shown in the Fig.2, the importance scores vary from variables to variables. On the one hand, the importance score can explain, to some extent, how much AdR depends on or is related to the variable. On the other hand, this score reflects the sensitivity of the variables in the model, the greater the importance score is, the more sensitive the model is to the variable.

Take the impact of the same period of last year's data on the forecast results as an example. First, we obtained the prediction results with last year’s data as one of the input variables, and then not taking the data last year for another prediction. Metrics in the two experiments when considering/not considering the data last year are shown in the Tab.3. There are obvious differences in the metrics of the two experiments, when considering the data of same period last year, R-square significantly increased, RMSE and MAE are obviously reduced, indicating that the prediction accuracy is improved, which implies that the data of the same period last year has an important impact on the prediction results, so are the other important variables.

Thus, we concluded that the results of the model are sensitive to the variables strongly related to AdR.

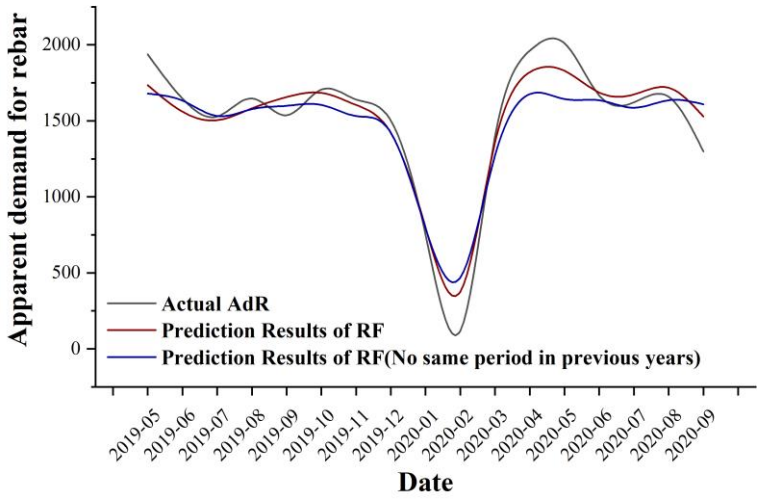

Fig. 10 Time series of prediction results of the two experiments when considering/not considering the data last year

Tab. 3 Metrics in the two experiments when considering/not considering the data last year

	R-square	RMSE (10kt)	MAE (10kt)
Results with last year's data as input	0.947	127.72	96.88
Results without last year's data	0.914	134.42	103.73

The parameters of Random Forest regressor influence the accuracy of the prediction results of the model. As described in section 5.1.2 and 5.2.1, the changes of OOB accuracy with parameters has already shown this point of view prophetically. To test the sensitivity of the model to the parameters of Random Forest, we altered the values of two important parameters (the number of the input factors (nFeatures), the number of the trees (nTrees)) in the Random Forest each time and observed the changes of model performance (Fig.11).

With the increase of nFeatures, R-square first increases and then decreases, RMSE and MAE first decrease and then increase. When nFeatures is about 15, R-square reaches the maximum value, RMSE and MAE are minimum, which is consistent with the results of parameter calibration. With the change of nFeatures, the accuracy of prediction results has a relatively greater change.

Similarly, with the change of nTrees, the changes of the three metrics are similar to that of nFeatures. The prediction results of the model are the best when nTrees is equal to 200, which is also consistent with the results of parameter calibration.

As shown in the Fig.11, the variation range of the three metrics with nFeatures is greater than the nTrees, indicating that the results are more sensitive to nFeatures due to the different importance of input variable. As the importance of input variables also affects the accuracy of the model, when nFeatures increases, the input variables have higher importance, so the accuracy of the model firstly increaces. When nFeatures is greater than 15, the importance of input variables is much lower, which may even bring more errors and uncertainties, so the accuracy of the model decreases. As for the nTrees, a smaller number of trees may be insufficient to reduce errors by voting of

multiple decision trees, thus, when the number of trees increased from 50 to 200, the accuracy of the model increases significantly, but when the number of trees exceeded a certain threshold, the accuracy of the results did not change significantly, so the results were not that sensitive to this parameter when compared with the nFeatures.

Due to the sensitivity of the model, it's of great importance to calibrate the parameters in the model.

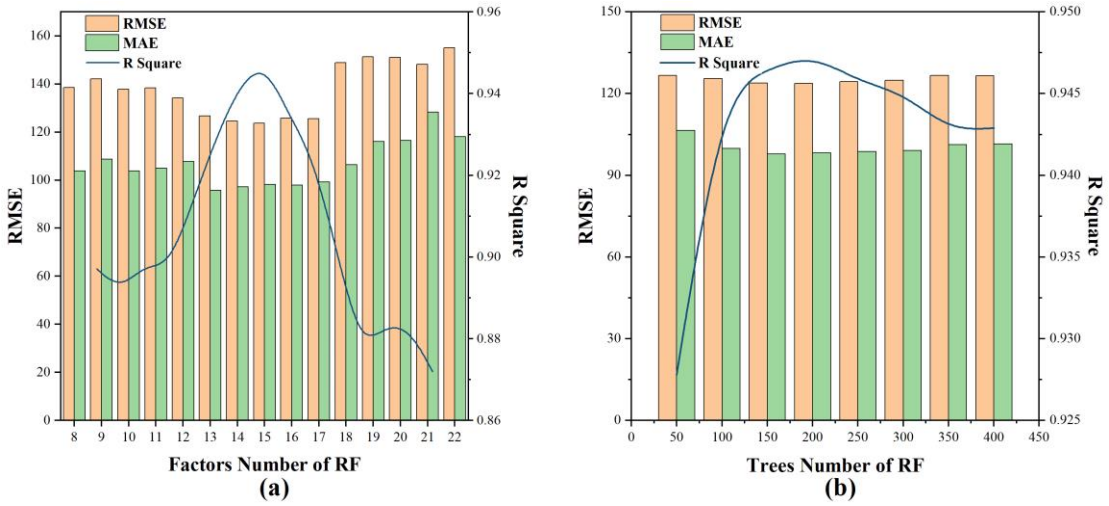

Fig. 11 The variabilities of the metrics with the changes of parameters in RF

7. Strengths and Weakness

7.1 Strengths of Model

For the current data set, the random forest algorithm has great advantages over other algorithms. According to the previous data processing and analysis, we found that rebar prediction data is a kind of high-dimensional data on time series. So, we chose to use PCA to extract the main components—Main influencing factors, and the principal components could be then fitted to predicte apparent demand of rebar by multiple linear regression. The advantage of this method is that the algorithm is simple and easy to understand, which effectively reduces the dimension of data and the interference of unimportant factors in model fitting and prediction. PCA is a comprehensive reorganization of factors, its principal components is not original, but the recombination of multiple factors. That is to say, on the one hand, it is difficult to explain the significance of a component. On the other hand, when it comes to the selection of the number of principal components, it is easy to overfit the prediction results.

As for the high-dimensional data, the RF model is widely used to analyze and predict. It does not need to reduce the dimension of data, but can effectively maintain the integrity of the high-dimensional data. In the Random Forest, the generalization error is unbiased estimated, which means a strong generalization ability. In the

process of training data, the training period is short with less time consumption, the implementation is relatively simple, and the influence of each factor in the training process can be detected. The RF has strong anti-interference ability, and it also has strong robustness for the dataset with missing values. Compared with other algorithms, RF has better anti-overfitting effect for the noise value in the data.

For the problem of time dislocation in the data, our solution is to build SARIMA model, using the same period data of previous years to fill in the missing data of the same period in the current year, and predicting the subsequent period based on the RF model.

ARIMA is one of the most widely used single variable time series data prediction methods, but it does not support time series with seasonal components. In order to support the seasonal component of the sequence, we employed the SARIMA to replace ARIMA. SARIMA is widely applied to univariate data including trend and seasonality, which is composed of a series of trend and seasonal components. SARIMA model is relatively simple, taking only its own pervious values as input and does not need other variables.

The SARIMA-RF scheme shows effectiveness in forecasting the apparent demand of rebar, with good consistency between prediction results and actual values and a R-square value of 0.92.

7.2 Weakness of Model

Though RF has many strengths, it also has many drawbacks. When the Random Forest is used to solve the regression problem, it cannot make the prediction beyond the range of the training set data, which may lead to overfitting in the modeling to some extent. RF is like a black box, which cannot control the internal operation of the model nor explain the mechanism, but only try using different parameters and random seeds. Although SARIMA could ignore the influence of seasonal changes, it can only consider the influence of its own, but not more external factors.

8. Conclusion

A prediction and analysis model of apparent demand of rebar based on RF was proposed in this paper. Firstly, the time resolution and range of 65 different factors we collected and analyzed were unified. In the light of the high-dimensional time series data with data gaps, the missing values were imputed through multivariate interpolation. Then, based on the importance scores calculated by Random Forest, the 15 input factors related to rebar demand for further prediction were selected.

While ignoring the time misalignment, Random Forest model was established to predict the apparent demand of rebar. The results of prediction achieve high accuracy, R-Square, RMSE, and MAE are 0.947, 127.72(10kt), and 96.88(10kt) against actual values, respectively. Compared with a traditional model, in which feature selection is conduct using principal component analysis and prediction is carried out with the multiple linear regression, in our model, R-Square has increased by 23%, and

prediction error has been reduced by about 50%. The prediction result of the proposed model is significantly better than the traditional method.

To deal with the problem of time lag in data publication, SARIMA was employed to forecast the values of some factors, which was then taken as input of random forest for further prediction. The SARIMA-RF scheme shows effectiveness in forecasting the apparent demand of rebar, with good consistency between prediction results and actual values and a R-square value of 0.92.

From the established models for each problem and analysis of the results, it can be inferred and discussed that:

(1) As one of the steel products, the price fluctuation of rebar is greatly affected by the macro-economic environment both foreign and domestic, especially under the influence of environmental protection policies, the demand tendency of rebar is continuous, cyclical and institutional. Environmental protection will continue to restrict the release of output, marginal effect may be weakened, and the rhythm of demand will be the key to dominate the price.

(2) From the perspective of causality and current situation analysis, the government needs infrastructure to support the economy and stabilize the economy. The manufacturing industry is expected to benefit from tax reduction and fee reduction. From the perspective of rhythm (time scale), the demand of real estate and infrastructure after the festival are the core factors to influence the price and the demand of rebar, also leading to the risk of weakening the rebar used after the middle of the year.

(3) From the perspective of industry concentration and results prediction, Although the profit center of the industry moves down, the rebound of infrastructure hitting the bottom will also hedge against the weakening of new construction in the second half of the year. On the whole, the supply and demand of the industry remains stable and weak. Land reserve and new construction at the enterprise level are important indicators to observe the future period,

References

[1] PATEL J. China Steel Rebar: Market Overview and Latest Product Developments; proceedings of the Conf Proceeding of SEAISI, F, 2018 [C].

[2] MEHMANPAZIR F, KHALILI-DAMGHANI K, HAFEZALKOTOB A J R P. Modeling steel supply and demand functions using logarithmic multiple regression analysis (case study: Steel industry in Iran) [J]. 2019, 63,101409.

[3] BUCK S F. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer [J]. Journal of the Royal Statal Society: Series B, 1960, 22

[4] BUUREN S V, GROOTHUIS-OUDSHOORN K. MICE: Multivariate Imputation by Chained Equations in R [J]. Journal of statal software, 2011, 45(3)

[5] SCORNET E. Random Forests and Kernel Methods [J]. IEEE Transactions on Information Theory, 2016, 62(3): 1485-500.

[6] ARLOT S, GENUER R. Comments on: A random forest guided tour [J]. Test, 2016, 25(2): 228-38.
[7] WOLD S, ESBENSEN K, GELADI P J C, et al. Principal component analysis [J]. 1987, 2(1-3): 37-52.
[8] LEVER J, KRZYWINSKI M, ALTMAN N. Points of significance: Principal component analysis [M]. Nature Publishing Group. 2017.
[9] ANSLEY C, SPIVEY W, WROBLESKI W J J O T R S S S C. A class of transformations for Box‐Jenkins seasonal models [J]. 1977, 26(2): 173-8.
[10] YOUNG P, SHELLSWELL S. Time series analysis, forecasting and control [J]. IEEE Transactions on Automatic Control, 1972, 17(2): 281-3.
[11] CHEN K-Y, WANG C-H. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan [J]. Expert Systems with Applications, 2007, 32(1): 254-64.

[

Appendix A

Categories	Abbreviation	Meaning	Abbreviation	Meaning
Real Estate	RE_Aifm	Average interest rate of first mortgage	RE_Irp1	the Index of residential price(T1)
	REFS_Dl	Fund Source: Domestic loan	RE_Irp2	the Index of residential price(T2)
	REFS_Fc	Fund Source: Foreign capital	RE_Irp3	the Index of residential price(T3)
	REFS_Sf	Fund Source: Self-raised funds	RE_Inchp	the Index of new commercial housing price
	REFS_Os	Fund Source: Others	RE_Inchp1	the Index of new commercial housing price (T1)
	REFS_Da	Fund Source: Deposit and advance payment	RE_Inchp2	the Index of new commercial housing price (T2)
	REFS_Pm	Fund Source: Personal mortgage loan	RE_Inchp3	the Index of new commercial housing price (T3)
	REPMI_Nco	Non manufacturing PMI: New construction orders	RE_Nlt	Number of land transactions (5 weeks average)
	RE_AoS	Area for sale of commercial housing(sum)	RE_Alt	Area of land transactions (5 weeks average)
	RE_AsC	Area sold of commercial housing (current month)	RE_Lpr	Land premium rate
	RE_As	Area sold of commercial housing	RE_Lp	Land premium (12 weeks average)
	RE_Sn	the sales number of commercial housing	RE_Lp	Land purchased (current year)
Infrastructure	InPPP_Ir	Implementation rate of PPP projects	InNB_Bd	National public budget deficit
	InPPP_N	PPP projects number(total)	In_Pr	Public revenue
	InPPP_Nid	PPP projects number(identification)	In_Pfe	Public finance expenditure
	InPPP_Npr	PPP projects number(preparation)	InDA_Ep	Declared amount of electric power project
	InPPP_Npu	PPP projects number(purchase)	InDA_In	Declared amount of infrastructure project
	InPPP_Nim	PPP projects number(implementation)	InDA_Tr	Declared amount of transportation project

	InPPP_It	PPP projects investment(total)	InDA_Wc	Declared amount of water conservancy project
	InPPP_Iid	PPP projects investment:(identification)	InNG_Ad	National government actual deficit
	InPPP_Ipr	PPP projects investment:(preparation)	InNG_Fr	National government fund revenue
	InPPP_Ipu	PPP projects investment(purchase)	InNG_Fe	National government fund expenditure
	InPPP_Iim	PPP projects investment(implementation)	InNB_Rb	National public revenue budget
	InNB_Feb	National public finance expenditure budget		
	InGFI_Lgd	Generalized financial indicators: Local government debt(current month)	InGFI_Deibc	Generalized financial indicators: the Debt of export&import bank of China(current month)
	InGFI_Pbd	Generalized financial indicators: Policy bank debt(current month)	InGFI_Dadbc	Generalized financial indicators: the Debt of Agricultural Development Bank of China(current month)
	InGFI_Dndb	Generalized financial indicators: the Debt of National development bank(current month)	**Rebar**	AdR the apparent demand of Rebar pR the price of Rebar Tcs the National turnover of construction steel
Loan	Ild	the Index of Loan demand		
	Ild_M	the Index of Loan demand (manufacturing)		
	Ild_In	the Index of Loan demand (infrastructure)		
Cement	Cuc	Cement utilization of capacity	Icp	the Index of Cement price

Appendix B

rf1-Missing value processing and feature importance score.py-Python script used for missing value processing and feature importance score.

```
import xlwt
from sklearn.ensemble import RandomForestRegressor
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
import pandas as pd
from pprint import pprint
data=pd.read_excel(r"factors.xlsx")
pprint(data)
target=data["AdR"].values
factorNames=data.columns[2:].values.tolist()
factors=data[factorNames].values
#misvalue processing
imp=IterativeImputer(n_nearest_features=10,missing_values=missValue)
imp.fit(factors)
factors=imp.transform(factors)
pprint(factors)
dataNew=data
for i in range(len(factors)):
    dataNew.iloc[i,2:]=factors[i]
pprint(dataNew)
dataNew.to_excel("factorsNew.xlsx")
#random forest fit and importance calculated
rf=RandomForestRegressor(n_estimators=200,n_jobs=4,oob_score
=True,max_depth=None,
                         max_features="sqrt",min_samples_split=2,
random_state=0,criterion="mse")
rf.fit(factors,target)
importances=rf.feature_importances_
pprint(importances)
#save results
workbook = xlwt.Workbook(encoding = 'utf-8')
worksheet = workbook.add_sheet('importance')
for i in range(len(importances)):
    worksheet.write(i,0, importances[i])
workbook.save(r'H:\project-py\appendix\importance.xls')
```

rf2-Calibration of random forest parameters.py-Python script used for the calibration of the parameters in the random forest.

```
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
import xlwt
import datetime

factorsNameData=pd.read_excel("importance.xls")
factorsNameList=factorsNameData["factor"].values.tolist()
data=pd.read_excel(r"factorsNew.xlsx")
target=data["AdR"].values.tolist()

#numbers of features
workbook = xlwt.Workbook(encoding = 'utf-8')
worksheet = workbook.add_sheet('oobScore')
for i in range(3,len(factorsNameList)):
    factorsUsed=factorsNameList[:i]
    factors=data[factorsUsed].values.tolist()
    rf = RandomForestRegressor(n_estimators=200, n_jobs=4,
oob_score=True, max_depth=None, max_features="sqrt",
                               min_samples_split=2, random_state=0,
criterion="mse")
    rf.fit(factors,target)
    rfscores = rf.oob_score_
    worksheet.write(i, 0, len(factorsUsed))
    worksheet.write(i, 1, rfscores)
workbook.save("特征选择.xls")

#numbers of trees
factors=data[factorsNameList[:15]].values.tolist()

workbook = xlwt.Workbook(encoding = 'utf-8')
worksheet = workbook.add_sheet('nTrees')
def calculateParameters(tnum):
    starttime = datetime.datetime.now()
    rf1 = RandomForestRegressor(n_estimators=tnum, n_jobs=4,
oob_score=True, max_depth=None,
                                min_samples_split=2,
random_state=0,criterion="mse")
    rf1.fit(factors, target)
    endtime = datetime.datetime.now()
    rfscores = rf1.oob_score_
    times=(endtime - starttime).microseconds
    return rfscores.mean(),times

for i in range(0,20):
```

```
    nt=50+50*i
    res=calculateParameters(nt)
    #worksheet.write(0, 0, 'nt')
    # worksheet.write( 0, 1, 'oobScore')
    # worksheet.write( 0, 2, 'time consume')
    worksheet.write(i + 1, 0, nt)
    worksheet.write(i+1, 1, res[0])
    worksheet.write(i+1, 2, res[1])
workbook.save("nTrees.xls")
```

(3)rf3-prediction.py-Python script used to predict the apparent demand of raber by random forest.

```
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
from pprint import pprint
import xlwt

#choose features
factorsNameData=pd.read_excel("importance.xls")
factorsNameList=factorsNameData["factor"].values.tolist()
factorsUsed=factorsNameList[:15]
factorsUsed.append("spl")
print(factorsUsed)

#train data
data_t=pd.read_excel(r"H:\project-py\appendix\factors-arima-train.xl
sx")
target_t=data_t["AdR"].values.tolist()
factors_t=data_t[factorsUsed].values.tolist()

#predict data
data_p=pd.read_excel(r"H:\project-py\appendix\factors-arima-predict.
xlsx")
factors_p=data_p[factorsUsed].values.tolist()
workbook = xlwt.Workbook(encoding = 'utf-8')
worksheet = workbook.add_sheet('predict')
rf=RandomForestRegressor(n_estimators=150, n_jobs=6, oob_score=True,
max_depth=None,
                                min_samples_split=2,
random_state=0,criterion="mse")
rf.fit(factors_t,target_t)
y=rf.predict(factors_p)
pprint(y)
yl=y.tolist()
```

```
for i in range(len(yl)):
    worksheet.write(i,0,yl[i])

workbook.save("results.xls")
```

(4)rf4-Multiple regression data processing.py-Python script used to extract the data needed by multiple regression.

```
import pandas as pd
factorsNameData=pd.read_excel("importance.xls")
factorsNameList=factorsNameData["factor"].values.tolist()
factorsUsed=factorsNameList[:15]
data=pd.read_excel("factorsNew.xlsx")
dataUsed=data[factorsUsed]
dataUsed.to_excel(r"H:\多元回归\factorsReg.xlsx")
```

(5)arima-parameters.py-Python script used to calibrate the parameters in ARIMA algorithm.

```
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.tsa.arima_model import ARIMA
from datetime import date
from statsmodels.graphics.api import qqplot
from pmdarima import auto_arima

# Ignore harmless warnings
import warnings
warnings.filterwarnings("ignore")
factorsNameData=pd.read_excel("importance.xls")
factorsNameList=factorsNameData["factor"].values.tolist()
factorsUsed=factorsNameList[:6]
print(factorsUsed)
# data
data=pd.read_excel(r"H:\q3.xlsx")
factors=data[factorsUsed]
factorlist=factors["REPMI_Nco"].values.tolist()
stepwise_fit = auto_arima(factorlist, start_p = 1, start_q = 1,
                          max_p = 4, max_q = 4, m = 12,
                          start_P = 0, seasonal = True,
                          d = None, D = 1, trace = True,
                          error_action ='ignore', # we don't want to know
if an order does not work
                          suppress_warnings = True, # we don't want
convergence warnings
```

```
                        stepwise = True)     # set to stepwise
# To print the summary
stepwise_fit.summary()
```

(6)arima.py-Python script used to predict the values of some factors not published when needed.

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.tsa.statespace.sarimax import SARIMAX

# Ignore harmless warnings
import warnings
warnings.filterwarnings("ignore")

#features
factorsNameData=pd.read_excel("importance.xls")
factorsNameList=factorsNameData["factor"].values.tolist()
factorsUsed=factorsNameList[:6]
print(factorsUsed)

# data
data=pd.read_excel(r"H:\q3.xlsx")
factors=data[factorsUsed]
data_index = sm.tsa.datetools.dates_from_range('2016m1','2020m9')
print(data_index)
def Arima(data:list,order,sorder):
    data=pd.Series(data)
    data.index=pd.Index(data_index)
    # setup SARIMA
    arima =
SARIMAX(data,order=order,seasonal_order=sorder,freq='M').fit()
    #prediction
    predict_y = arima.predict(start="2020-01",end="2020-09")
    for v in predict_y.values:
        print (v)

factorlist=factors["REPMI_Nco"].values.tolist()
order=(1,0,0)
sorder=(1,1,0,12)
Arima(factorlist,order,sorder)
```