Analysis of factors influencing regional economic expansion based on OOB coefficients under RF algorithm

Yue Xu[1, #], Zhi Cao[2,#], Muyuan Wang[1, #,*]

[1]College of Economics and Management, Nanjing University of Aeronautics and Astronautics, 210000, Nanjing, China

[2]College of Economics and Management, Northeast Agricultural University, 150000, Harbin, China

[*]Corresponding author: wang_mu_yuan@163.com

[#] equally contributed to this works

Abstract. In the context of China's economic growth, the economic situation of national new areas is particularly important, and it is of great significance to study its economic expansion factors. In this paper, the regression random forest algorithm (RF) is used and the economic panel data of Shanghai, China from 2011-2022 is selected for algorithm learning, which obtained algorithm with low error and high accuracy. After the algorithm was learned, the OOB coefficients of the influencing factors were obtained and analyzed to study their impact on regional economic inflation. The results show that the per capita disposable income and the ratio of the added value of the tertiary industry to the GDP play an important role in economic expansion. Finally, policy recommendations for regional economic development are offered, which can help address regional economic risks and contribute to regional economic improvement and growth. The research in this paper analyses the importance of factors influencing regional economic growth by using machine learning methods and quarterly and annual provincial panel data, making the conclusions drawn more innovative and robust.

Keywords: random forest; regional economy; economic expansion; influencing factors; OOB.

1. Introduction

In recent years, China's economic and social growth and development, and industrial structure have been adjusted, regional economic coordination has been strengthened, and the overall economic quality has been greatly improved. At the same time, as a regional economic policy with Chinese characteristics, the national new areas shoulder the important strategic mission of radiating and driving regional economic development. Affected by trade frictions between China and the United States and the new crown epidemic, China's GDP growth in the first quarter of 2020 dropped by 6.8% year on year, and China's domestic economic situation has a serious downward trend. Therefore, it is of great practical significance to study the influencing factors of regional economic growth to stabilize the basic economic disk and resolve the operational risks of the regional economy.

Focusing on the issue of regional economic growth in China, Chinese scholars have made rich research. At the capital factor level, LI & CHEN[1] found that there is a widening "gap" in physical capital accumulation between coastal areas and inland areas, which is an important factor leading to the regional economic growth gap. YUAN & ZHU[2] Through measuring human capital and its structure, it is found that investment in human capital stock can improve the efficiency of capital economic growth and has a lagging effect on economic growth. From the perspective of industrial structure and urbanization, XU & FANG[3] shows that both industrialization and urbanization have significant effects on economic growth in the central and western regions. WANG[4] Through the dual difference method, it is examined that the construction of free trade zone is the key factor to drive regional economic growth, and investment and trade is the important driving factor to promote regional economic growth. ZHANG & ZHANG[5] show that house price fluctuation will make regional economic growth positive change in a short time, but the long-term positive impact tends to disappear. LI & WANG[6] Research shows that house prices have a significant impact on the quality of economic growth, and there is an inverted "U" relationship between the two. WANG & XU[7] Based on provincial-level panel data, fiscal capital expenditure contributes to economic growth in the

short term, while fiscal welfare expenditure significantly promotes economic growth and contributes to our short-term and long-term economic convergence.

Previous studies have done a lot of research based on economic expansion, but in the analysis method, the use of the Solow Residual Method (SRA) or Stochastic Frontier Analysis (SFA) is required to assume the specific form of the production function, the results are inevitably affected by the production function set. In this paper, the stochastic forest is used to calculate the importance of each influencing factor by constructing an RF model. The research of this paper is beneficial to the development of the following aspects. Firstly, the importance of different factors affecting regional economic growth is analyzed by machine learning, which is more innovative than traditional econometric models. Secondly, this paper selects the provincial quarterly panel data, and compared with the annual panel data, the conclusion is more robust.

2. Constructing index system of regional economic expansion

2.1 Regional economic expansion

Regional economic expansion means that when a regional economy develops at a certain scale after a long period of rapid development, its future development situation may appear to "bubble burst". If macro-control is not timely, it will lead to an "economic bubble" more serious, and eventually burst into depression or be trapped by external and internal factors. Therefore, it is of great significance to study the influencing factors of regional economic growth for timely regulation of economic policies and prevention of economic expansion. Referring to JIA[8], the CPI (consumer price index) is used as an indicator of regional economic expansion.

2.2 Construction of influencing factor index system

Based on the existing literature, this paper from the capital factors, labor factors, urbanization factors, foreign trade factors, industrial structure factors, institutional factors, and real estate market factors. Capital factors are measured by reference to the ZHANG & ZHANG[11] method, which is mainly measured by the amount of investment in fixed assets; labor factors are generally measured and analyzed by reference to the number of employees, which is measured by reference to the YUAN & ZHU[2] method, which is measured by reference to the registered unemployment rate and per capita disposable income; urbanization factors are measured by reference to the practice of WANG & ZHAO[9], which is measured by reference to the urbanization rate (the proportion of the urban population to the total population at the end of the year) and the contrast between urban and rural consumption levels; foreign trade factors are measured by reference to the practice of WANG[4], which is replaced by the export volume of the trade zone; industrial structure factors are measured by reference to the practice of XU & FANG[3], which is measured by the ratio of the added value of the secondary and tertiary industries to the GDP. Institutional factors refer to WANG & XU's[7] methodology, using general public budget expenditure as a measure. Real estate market factors refer to the LV[10] method, the price index of newly-built commodity housing in large and medium cities is used to replace. To sum up, the index system set up in this paper is shown in Table 1.

Table 1 System of Influencing Factors

Grade I indicators	Grade II indicators	Unit	Ref
Capital factors (X1)	Investment in fixed assets (X1)	100 million yuan	[1]
Labor factors (X2)	Registered unemployment rate (X21)	%	[2]
	Per capita disposable income (X22)	yuan	
Urbanization factors (X3)	Urbanization rate (proportion of the urban population to total population at the (X31)	%	[9]
	Comparison between urban and rural consumption levels (rural consumption	/	
	Foreign trade factors (X32)		

Foreign trade factors (X4)	Value of exports (X4)	Dollar	[4]
Industrial structure factors (X5)	The ratio of the added value of secondary industry (X51)	%	[3]
	The ratio of the added value of tertiary industry (X52)	%	
Institutional factors (X6)	General public budget expenditure (X6)	100 million yuan	[7]
Real estate market factors (X7)	New Commercial Housing Price Index for Large and Med (X7)		[10]

3. Random Forest algorithm —— OOB index construction

3.1 Random Forest algorithm

Random Forest (RF) is an integrated algorithm based on Classification and Regression Tree (CART) proposed by Breiman[12] based on the original Bagging algorithm, which can be used to deal with regression, classification, clustering, prediction, and survival analysis[13]. Fang et al. showed that the RF algorithm has high accuracy and robustness, good tolerance for outliers and noise and it can handle high-dimensional data sets and is not prone to overfitting and underfitting[14]. Ouyang and Chen's study showed that the RF algorithm can obtain the weights of each variable and assess the importance and role of the variables in the model concisely and efficiently with good generalizability[15]. RF is multiple classifiers or multiple regressors that are a combination of multiple decision tree models. First, it generates a new sample by Bootstrap resampling method by randomly selecting m records from the original sample in a put-back manner, where a is the sample capacity of the original sample, and repeated k times to obtain k samples, which is θk. A decision tree model h(X, θk) is built for each sample, where a of the b feature variables (b>a) are randomly selected at the branch nodes, and the optimal feature variables are segmented according to the least informative principle based on the average reduction of the impurity of each child node. Then the tree is fully grown using the CART method to finally obtain k classification or regression results. If the original problem is a classification problem, the final result is the plurality of all tree results and if the original problem is a regression problem, the final result is the average of all tree results[14]. In this paper, we use a regression random forest as shown in Figure 1.

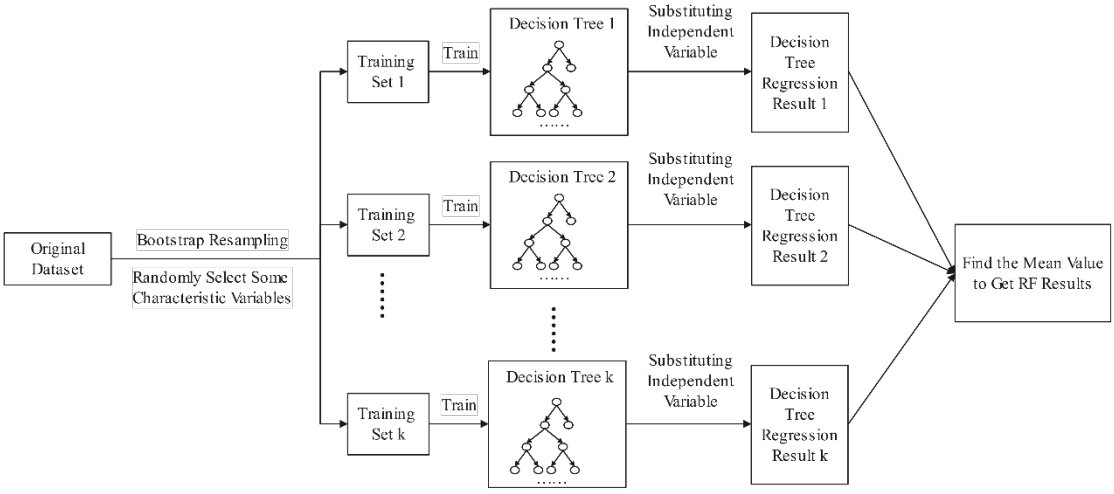

Figure 1 RF schematic

The RF algorithm is a combination of multiple binary decision trees (CART) and algorithm learning is the learning of multiple binary decision trees. When training a binary decision tree model, it is necessary to consider how to select splitting variables and splitting points and how to measure

the goodness of a splitting variable and splitting point. For the selection of splitting variables and splitting points, this paper uses the exhaustive method to find the best splitting variables and splitting points from them. The goodness of the splitting variables and splitting points is measured by the impurity of the nodes after the cut, i.e., the weighted sum of the impurity of each sub-node G(xi,vij), which is calculated as follows:

$$G(x_i, v_{ij}) = \frac{n_l}{N_s} H(X_l) + \frac{n_r}{N_s} H(X_r) \tag{1}$$

In the formula, xi is a certain splitting variable. vij is a splitting value of the splitting variable. nl, nr and Ns are the number of training samples of the left child node, the number of training samples of the right child node, and the number of training samples of the current node, respectively. Xl and Xr are the training sample sets of left and right child nodes respectively. H(X) is a function measuring the impurity of the node. In this paper, Mean Square Error (MSE) is chosen for H(X), i.e., for a given score point:

$$H(X) = \frac{1}{N_s}\left(\sum_{y \in X_l} (y_i - \bar{y}_l)^2 + \sum_{y \in X_r} (y_i - \bar{y}_r)^2\right) \tag{2}$$

In the formula, yi is the output value in the corresponding training sample set. $\bar{y}_l$ is the average of the output values in the training sample set of the left child node. $\bar{y}_r$ is the average of the output values in the training sample set of the right child node. The learning process for a node in a decision tree is mathematically equivalent to finding the cut variables and cut points that minimize G.

$$(x^*, v^*) = \arg\min_{x,v} G(x_i, v_{ij}) \tag{3}$$

The final regression result of the whole RF is the average of the regression results of all decision trees.

$$R(x) = \left(\sum_{i=1}^{k} h_i(x)\right) / k \tag{4}$$

3.2 OOB index calculation

Bootstrap resampling means that the original sample set *D* is resampled *N* times with put-back repetitions to obtain the training sample set D_b, where *N* is generally equal to *D*, such that some samples in *D* will be repeated in D_b several times and some other samples will not be drawn. In the Bootstrap sampling of *D*, the probability of each sample being selected is $\frac{1}{N}$ for each sampling and the probability of each sample being selected after *N* repeated sampling with put-back is：

$$(1 - (1 - \frac{1}{N})^N) \Rightarrow (1 - \lim_{N \to \infty}(1 - \frac{1}{N})^N) = (1 - \frac{1}{e}) \approx 0.633 \tag{5}$$

Therefore, each time the training sample set is generated, about one-third of the sample data is not selected and these unselected data are called out-of-bag (OOB) data. The RF model can use the OOB data at each node to calculate the weight of each feature variable for the dependent variable in a single decision tree for MSE analysis and the importance of a feature in the whole RF model is the average of the importance of that feature in all internal decision trees, which allows measuring the importance of each influencing factor for the dependent variable[16].

The importance of a particular node m is:

$$n_m = w_m \cdot G_m - w_l \cdot G_l - w_r \cdot G_r \tag{6}$$

In the formula, wm, wl and wr are the ratios of the number of training samples to the total number of training samples in node m and its left and right child nodes, respectively. Gm, Gl and Gr are the impurities of node m and its left and right child nodes calculated using OOB data, respectively. After calculating the importance of each node, the importance of a variable can be derived by the following formula:

$$f_i = \frac{\sum_{j \in nodes\ split\ on\ feature} n_j}{\sum_{k \in all\ nodes} n_k} \tag{7}$$

In order to make the importance of all variables add up to 1, the importance of each variable needs to be normalized:

$$f_{ni} = \frac{f_i}{\sum_{j \in all\ features} f_j} \tag{8}$$

4. The setting of test indicators

To make the model more credible, after the RF algorithm learning is completed, the goodness of fit of the model is evaluated by calculating the test metric through an unbiased estimation of the data. In this paper, two metrics, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are selected to measure the accuracy of the algorithm.

The MAE is the average of the absolute values of the deviations of all individual true values from the predicted values, which avoids the problem of errors canceling each other out and thus accurately reflects the magnitude of the actual prediction errors. A smaller MAE indicates a higher accuracy of the algorithm. Its calculation formula is:

$$MAE(X,h) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - h(x_i)\right| \tag{9}$$

RMSE is the sample standard deviation of the difference (i.e., residual) between the true and predicted values and can measure the deviation between the observed and true values. It is very sensitive to extra-large or extra-small errors in a set of data so it can reflect the precision of the prediction well. A smaller RMSE indicates a higher accuracy of the algorithm. Its calculation formula is:

$$RMSE(X,h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2} \tag{10}$$

5. Empirical analysis

5.1 Object selection and data collection

In recent years, the economy of China ' s first-tier cities has developed rapidly, and they occupy most of China ' s resources. As the representative of China ' s first-tier cities, Shanghai is the international economic, financial, trade, shipping, science and technology innovation center, and its economic development is far ahead of many cities. From 2016 to 2021, Shanghai ' s GDP grew year by year. In 2021, Shanghai ' s GDP ranked first in China ' s major cities, reaching 432.1485 billion yuan. In addition, the growth rate of fixed asset investment in Shanghai in 2021 was higher than the average of all cities in China and the average of first-tier cities in China ; retail sales of consumer goods in Shanghai rebounded in 2021 in a downturn in 2020, far higher than the average of first-tier cities in China ; shanghai also ranks first among China ' s top four cities in taxes. The rapid development of Shanghai ' s economy has also led to its price rise to a certain extent. The data show that the fluctuation of its consumer price index is more obvious than other first-tier cities in China. In summary, Shanghai ' s rapid economic development, while the consumer price index fluctuations significantly, as the object of this study, has a certain practical significance, can be used as other cities or regions of the case.

This paper collects 45 groups of relevant data in the first quarter of 2011 – 2022 in Shanghai through the databases of CSMAR, RESSET and Statistical Yearbook. Descriptive statistics are shown in Table 2.

Table 2 Descriptive statistics

Variable	Max	Min	Mean
Y	105.69	100.10	102.46
X1	2543.62	869.03	1699.36
X21	4.14	2.73	3.70
X22	7554.00	2920.00	4694.00
X31	89.60	87.60	88.70
X32	2.40	1.96	2.13
X4	72502935.00	41619417.00	51505211.00
X51	41.50	23.70	32.10
X52	76.20	58.10	67.70
X6	2910.39	632.40	1636.67
X7	103.27	98.73	100.55

5.2 Random forest training

In this paper, the data of consumer price index and fixed asset investment in Shanghai from the first quarter of 2011 to the first quarter of 2022 were collected through the databases of CSMAR, RESSET and Statistical Yearbook. RF training was carried out with SPSSPRO software. By adjusting the number of decision trees in was adjusted to 100,200,500. The accuracy of the model was measured by MAE and RMSE, and the following iterative accuracy was obtained. The results are shown in Table 3.

Table 3 Training Results

Parameter	MAE	RMSE
100	0.815	0.969
200	0.688	0.881
500	0.448	0.676
800	0.397	0.484
1000	0.298	0.434

It can be seen from Table 3 that when the number of model construction is set to 1000, MAE = 0.298, RSME = 0.434, and the algorithm accuracy reaches the optimal value, so the confidence of the result is the strongest. The model under this parameter is selected as the final analysis model.

5.3 Analysis and Discussion of OOB

Based on the above RF training, this paper obtains the final analysis model and the OOB coefficient corresponding to each explanatory variable, which is showed in Table 4:

Table 4 OOB coefficient corresponding

Explanatory variable	OOB coefficient corresponding
The ratio of the added value of tertiary industry	28.00
The ratio of the added value of secondary industry	21.70
Comparison between urban and rural consumption levels	4.30
Urbanization rate	1.70
Registered unemployment rate	2.80
Per capita disposable income	27.50
General public budget expenditure	3.50
New commercial housing price index	1.50
Export volume	4.70
Investment in fixed assets	4.20

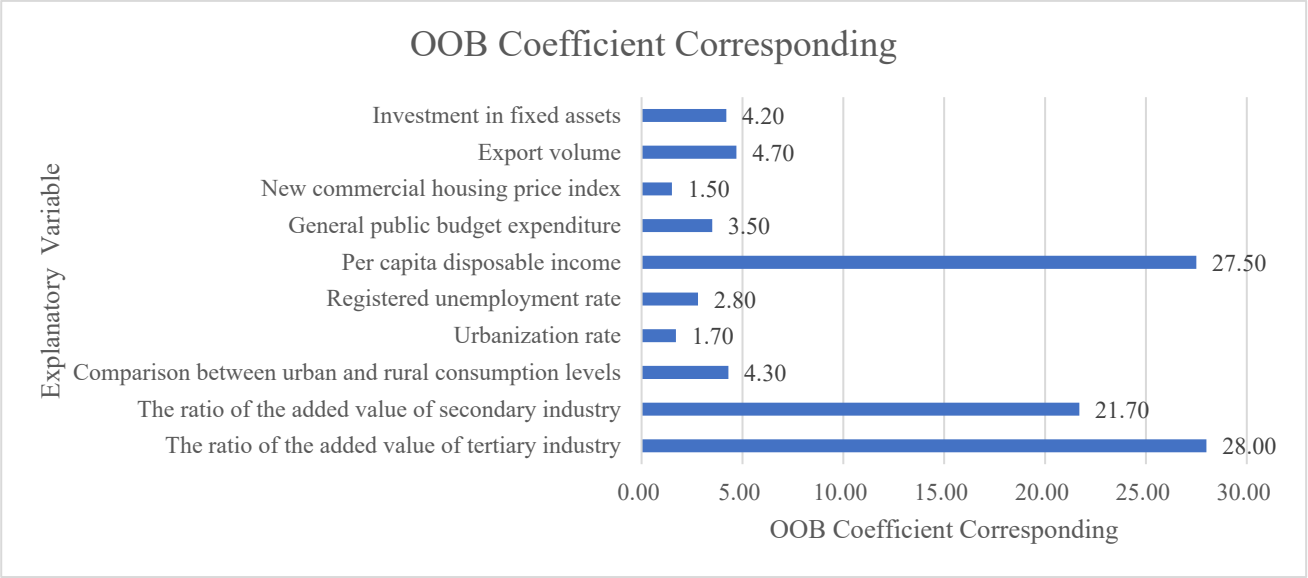

Figure 2 OOB Coefficient Corresponding

It can be seen from Figure 2 that the proportion of added value of tertiary industry to GDP (0.280) and per capita disposable income of urban residents (0.275) are more important to the consumer price index than other variables. The reason is that for the proportion of the added value of the tertiary industry in GDP, this paper believes that the proportion of the tertiary industry can reflect the vitality of the local economy. The higher the proportion of the tertiary industry is, the more dynamic the local economy is. The GDP is one of the important indicators of economic vitality, and the GDP can reflect the local consumption potential to a certain extent. Therefore, the proportion of the tertiary industry will also have a great impact on the local consumer price index. For the per capita disposable income of urban residents, the per capita disposable income can reflect the consumption ability of local residents and the level of regional economy to a certain extent. The higher the per capita disposable income is, the stronger the local residents ' consumption ability is, and the more market vitality the local economy is. Therefore, the pricing of local commodities will also rise accordingly, which leads to a higher price index of local residents ' consumption.

Secondly, the proportion of the added value of the secondary industry in GDP (21.70) also has an important impact on the consumer price index. This paper believes that this is because the proportion of the secondary industry can reflect the degree of local industrialization to some extent.

Other explanatory variables, such as the comparison of urban and rural consumption levels and export volume, have less influence on the explained variables than the above two variables. The reasons are as follows. (1) The explanatory variable does not have explanatory power on the explained variable itself. (2) The multiple collinearity between variables leads to the fact that the influence of the variable on the explained variable is covered by other variables that have collinearity with it, resulting in its OOB coefficient other variables.

6. Conclusion

In this paper, the consumer price index is used as the explained variable, and fixed asset investment and export volume are used as explanatory variables. Based on the random forest model, SPSSPRO software is used for calculation and solution. The conclusions are as follows :

In the construction of the RF model, when the number of model construction is 1000, MAE = 0.298, RSME = 0.434, the accuracy of the algorithm is optimal. The results show that the proportion of added value of the tertiary industry in GDP (0.280) and per capita disposable income of urban residents (0.275) in Shanghai have a greater impact on the consumer price index than other explanatory variables, followed by the proportion of added value of the secondary industry in GDP (0.217), whose influence is relatively smaller.

The research in this paper contributes to addressing regional economic risks and radiates to the improvement and growth of the regional economy. To make the analysis more accurate and comprehensive, subsequent studies could adopt new algorithms to further optimize our results or consider additional dimensions of influence. Further analysis of the factors influencing regional economic expansion can also be carried out to obtain more in-depth results.

References

[1] Li Shujuan, Chen Qiuhui, Xu Xianxiang. The experience of economic growth recovery under adverse shocks--based on China's economic target management practice[J]. Economic Research,2021,56(07):59-77.

[2] Yuan Hang,Zhu Chengliang. The impact of government R&D subsidies on the transformation and upgrading of China's industrial structure: a push or a drag? [J]. Finance and Economics Research,2020,46(09):63-77.DOI:10.16538/j.cnki.jfe.20191217.301.

[3] Xu Qiuyan,Fang Shengfei,Ma Linlin. New urbanization, industrial structure upgrading and economic growth in China--an empirical study based on spatial spillover and threshold effect[J]. Systems Engineering Theory and Practice,2019,39(06):1407-1418.

[4] Wang Aijian,Fang Yunlong,Yu Bo. Construction of free trade pilot zones and regional economic growth in China: comparison of transmission paths and dynamic mechanisms[J]. Finance and Trade Economics,2020,41(08):127-144.DOI:10.19795/j.cnki.cn11-1166/f.2020.08.009.

[5] Zhang Xiekui,Zhang Lian. The impact of housing price fluctuations on the regional economy--an analysis based on dynamic panel data of 35 large and medium-sized cities[J]. Urban Issues,2017(06):90-95+103.DOI:10.13239/j.bjsshkxy.cswt.170611.

[6] Li GB,Wang J. Research on the impact of housing price on the quality of China's economic growth--an empirical study based on panel data of 286 prefecture-level and above cities[J]. Price Monthly,2018(05):1-6. doi:10.14076/j.issn.1006-2025.2018.05.01.

[7] Wang Baoshun,Xu Qishuang. Fiscal spending, regional economic disparity and dynamic growth convergence[J]. Journal of Zhongnan University of Economics and Law,2021(03):48-57+90.DOI:10.19639/j.cnki.issn1003-5230.2021.0029.

[8] Jia, Kaiwei. Government public expenditure and economic growth:A re-test of Wagner's law based on MTAR model[J]. Statistics and Decision Making,2015(13):143-146.DOI:10.13546/j.cnki.tjyjc.2015.13.039.

[9] Wang J,Zhao K. A study on urbanization, aging, urban-rural gap and economic development in China - based on a moderated mediating effect model[J]. Contemporary Economic Management,2020,42(07):49-58.DOI:10.13253/j.cnki.djjgl.2020.07.007.

[10] Lv Fengyong. Study on the inflationary effect of real estate price fluctuations[J]. Price Theory and Practice,2016(12):28-32.DOI:10.19851/j.cnki.cn11-1010/f.2016.12.007.

[11] Zhang Xiaohui,Zhang Chuanna. Research on the relationship between local government debt, fixed asset investment and economic growth--an analysis based on data from 111 counties (cities) in three northeastern provinces[J]. Economic Vertical,2020(08):100-107.DOI:10.16528/j.cnki.22-1054/f.202008100.

[12] Breiman. Random forests[J]. MACH LEARN, 2001, 2001,45(1)(-):5-32.

[13] Wang Yishen, Xia Shutao. A review of random forest algorithms for integrated learning[J]. Information Communication Technology,2018,12(01):49-55.

[14] Fang Kuang-Nan, Wu See-Bin, Zhu Jian-Ping, Xie Bang-Chang. A review of random forest methods[J]. Statistics and Information Forum,2011,26(03):32-38.

[15] Ouyang C.G.,Chen P. Factor endowment, local industrial sector development and industry choice[J]. Economic Research,2020,55(01):82-98.

[16] Xiong JINGHUA,Ru JING. Research on the combined model of exchange rate forecasting based on random forest algorithm and fuzzy information granulation[J]. Quantitative Economic and Technical Economics Research,2021,38(01):135-156.DOI:10.13653/j.cnki.jqte.2021.01.008.