Digital Infrastructure and Regional Economic Growth: An Empirical Study based on Random Forest Regression

Tianyi Xue*

Business School, Nanjing University, Nanjing, China

*Corresponding author: xuetyttkx@163.com

Keywords: Random Forest Regression, Digital Infrastructure, Economic Growth.

Abstract: Since the third scientific and technological revolution, information and communication technology and related industries have made important contributions to the macro economy, meanwhile the digital economy has also developed vigorously. From the perspective of communication, this paper explores the influence of digital economy on macroeconomic fluctuations. This paper uses principal component analysis to construct digital infrastructure early warning indexes and conventional early warning indexes, combined with Random Forest Regression (RFR), calculates the importance of digital infrastructure, capital formation, human resources and other factors to China's economic growth, and carries out forecasting research. The research finds that compared with the construction of digital infrastructure, its application plays a more obvious role in driving the economy. The importance of digital infrastructure early warning factor has regional heterogeneity, and it has the greatest influence on the economically developed eastern region. In terms of research methods, Random Forest is an effective machine learning method.

1. Introduction

Information-related infrastructure is the foundation of the information and communication industry and even the digital economy. After the "new infrastructure" was written into the government work report, the digital infrastructure integrated with 5G network and artificial intelligence became a powerful measure to ease the downward pressure on the economy. Based on existing literature, existing studies have extensively discussed the role of information infrastructure in promoting China's economic growth. Chen et al. (2011) established a simultaneous equation model to investigate the contribution of information infrastructure to economic growth and found that the promotion effect of information infrastructure on economic growth was heterogeneous[1]. Xu and Liu (2014) used the CGE model to point out that increasing information infrastructure investment had no obvious direct effect on Shanghai's economic growth, but promoted the upgrading of industrial structure, thus making positive contributions to economic transformation[2]. Jiang et al. (2020) derived the investment multiplier based on the general equilibrium analysis framework of Keynesian economy, and then empirically studied the importance of China's investment multiplier and infrastructure investment in China[3].

It can be seen that some studies have used various models to measure the contribution of information infrastructure to economic growth, but few scholars have combined machine learning research methods to investigate the importance of digital economy infrastructure. This paper uses machine learning method to explore the impact of digital infrastructure on regional GDP. Digital infrastructure mainly includes information infrastructure and digital transformation of physical infrastructure. This paper explores the effect of digital infrastructure on economic growth from two perspectives of infrastructure construction itself and infrastructure application. The construction and application of digital infrastructure have direct and indirect effects on consumption, government purchase, investment and import and export, thus driving macroeconomic development. The transmission mechanism of digital infrastructure to economic growth is shown in Figure 1.

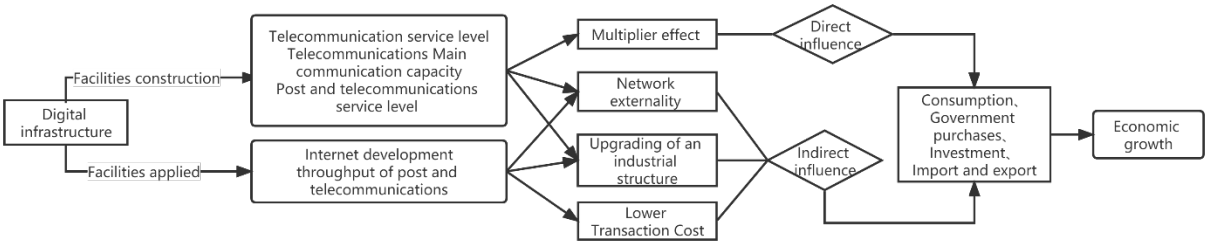

Fig. 1 Transmission mechanism of digital infrastructure to economic growth

2. Random Forest Regression (RFR)

The basic idea of Random Forest (RF) is to extract multiple sub-samples from the original samples with Bootstrap, model each sub-sample for decision tree, and then combine the prediction results of multiple decision trees with average method or voting method to determine the final prediction results[4]. Compared with ordinary least squares regression, Random Forest Regression (RFR) can obtain a higher corrected R-square value[5]. At the same time, ensemble learning algorithms such as Random Forest are well suited to large data sets because they do not use all predictive variables at the same time and therefore can operate efficiently[6].

2.1 Using RFR for prediction

Based on the research of Fang et al.(2014)[7], it is assumed that training sets 1, 2, 3... K is independently and randomly extracted from the training sample set (represented by random vectors Y and X), so multiple sub-decision trees related to random vector θ can be used to generate a Random Forest regression model. Therefore, the prediction result of RFR is the average of the prediction results of K sub-decision trees, namely:

$$H(x) = \frac{1}{k}\sum_{i=1}^{k} h_i(x) \tag{1}$$

Where, $H(x)$ is the result of combined regression model, h_i is the result of single decision tree regression model.

In addition, about 37% of the data in each sample will not be selected, also known as out-of-bag (OOB) data. The remaining data can be used for internal error estimation, that is, each classification tree can get an OOB error estimate, and take its mean value as the generalization error of the model. In the training process, when the number of trees is large enough, the Random Forest error is approximately equal to the out-of-bag error[8].

2.2 Using RFR for sorting

Random Forest can also rank the importance of features and provide some reference for the selection of feature variables[9]. The specific steps are as follows :(1) Train a Random Forest model through the training set, and record the OOB error of each data point during the training process.(2) Averaging over the entire forest. (3) Scramble the value of the JTH feature to be measured, and measure the OOB error of the scrambled data again. Then the significance score of the Jth feature is the average value of the OOB error difference before and after the scramble. In addition, when STATA is used, all importance scores are standardized so that the importance of the most important variable is always 100%[6].

3. Data and variables

In this paper, panel data of 31 provinces (municipalities and autonomous regions) in China from 2009 to 2019 were adopted, and the data came from WIND, China Statistical Yearbook and National Bureau of Statistics, etc., and the missing values were supplemented by moving average method. Table.1. is the variable explanation table, and Table.2. is the descriptive statistics.

Referring to the evaluation system of digital economy constructed by predecessors[10], data such as Internet server, post and telecommunications represent information infrastructure conditions of digital economy to a certain extent. Combined with the availability of data, this paper selects ten sub-indexes from X1 to X10 to measure the level of digital infrastructure, and divides them into five groups according to the classification of the National Bureau of Statistics. Among the sub-indexes, X1, X2 and X3 represent the main indicators of Internet development, X4, X5 and X6 are subdivided indicators of post and telecommunications business volume, X7 reflects the level of telecommunications services, X8 and X9 represent the main communication capabilities of telecommunications, and X10 represents the level of posts and telecommunications services. In this paper, after standardizing these indexes, principal component analysis is carried out according to the above grouping, and corresponding indexes are synthesized after dimensionality reduction. Factors affecting the economic growth of a region also involve talents, capital, innovation, system, industrial structure and other aspects. This paper selects different indicators to measure these five aspects respectively. After standardizing these indicators, this paper conducts principal component analysis according to the analysis framework in Table.1., and synthesizes corresponding indexes after dimensionality reduction.

Table.1. Variable interpretation table

Type	Indicator categories	Short Name	Meaning
Explained Variable	*economic growth*	Y	GDP
Core explanatory variable	*Internet development*	X1	the internet broadband port
		X2	Number of broadband Internet access users
		X3	PC view percentage
	throughput of post and telecommunications	X4	Mobile SMS service volume
		X5	Mobile phone call duration
		X6	Number of mobile phone users at year-end
	Telecommunication service level	X7	Popularization Rate of Mobile Telephones
	Telecommunications Main communication capacity	X8	Capacity of Mobile Telephone Exchanges
		X9	Length of optical cable
	Post and telecommunications service level	X10	The average population served by each branch of business
Other variables	*human resource*	X11	The number of people with higher education
	capital formation	X12	gross capital formation
	system	X13	household registration system
		X14	opening to the outside world
	Innovation	X15	R&D expenditure of industrial enterprises above designated size

		X16	Full-time equivalent of R&D personnel in industrial enterprises above designated size
	Industrial structure	X17	Proportion of secondary industry

Table.2. Descriptive statistics

Variable	Sample size	Mean	SD	Min	Max
economic growth	341	9.5433	1.0297	6.0996	11.5898
Internet development	341	0.0000	1.0000	-1.1507	4.3081
throughput of post and telecommunications	341	0.0000	1.0000	-1.3327	4.0904
Telecommunication service level	341	0.0000	1.0000	-1.3078	4.0383
Telecommunications Main communication capacity	341	0.0000	1.0000	-2.0504	3.6223
Post and telecommunications service level	341	0.0000	1.0000	-1.4425	3.1160
human resource	341	0.0000	1.0000	-1.5597	3.4122
capital formation	341	0.0000	1.0000	-1.6732	3.2909
system	341	0.0000	1.0000	-0.7237	5.1256
Innovation	341	0.0000	1.0000	-1.3044	4.2559
Industrial structure	341	0.0000	1.0000	-3.1769	2.4511

4. Results

4.1 Prediction error variation

Firstly, data points are randomly sorted in order to ensure the randomness of training data. Seed values are set at the same time to obtain repeatable results, and then the data set is evenly divided into two subsets: 50% for training and 50% for test[1].

Then, this paper adjusts the hyperparameters to obtain the optimal test accuracy, mainly adjusting the number of iterations and variables. OOB error converges as the number of iterations increases, which means that when the number of iterations is large enough, changes in the hyperparameter NUMVars can be considered to be the main cause of changes in OOB error and Validation error. In this paper, the number of iterations is set from 10 to 300. For simplicity, set the number of features to 1 initially. At the end of the loop, the actual root mean square Error (RMSE) is obtained using the data from the test set and placed on a graph with the OOB Error. When the sample range is the whole country and the eastern, central and western regions respectively, the scatter diagrams of out-of-bag error and validation error changing with the number of iterations are shown in Figure 2, Figure 4, Figure 6 and Figure 8.

Next, this paper selects an appropriate number of iterations to make OOB error and Validation error smaller and tend to be stable[2].On this basis, adjust the number of features numVars to observe changes in OOB error and Validation error. The scatter diagram of out-of-bag error and validation error

changing with the number of variables is shown in Figure 3, Figure 5, Figure 7 and Figure 9. The minimum error and the corresponding number of variables can be automatically output in combination with the frame command. For the whole country and the eastern, central and western regions, the corresponding number of characteristic variables are 2, 2, 4 and 3, respectively.

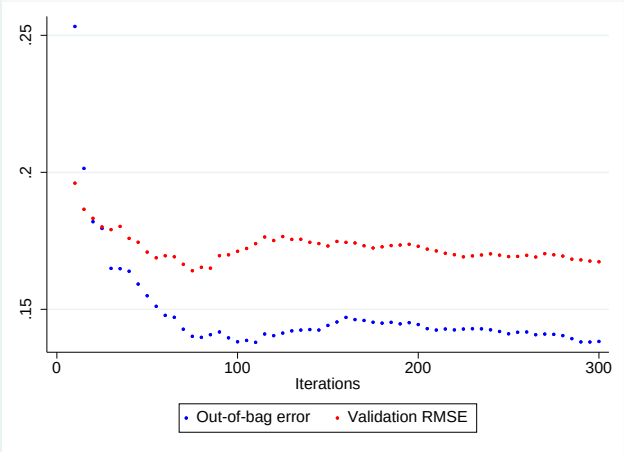

Fig. 2 PE changing with number of iterations (N)

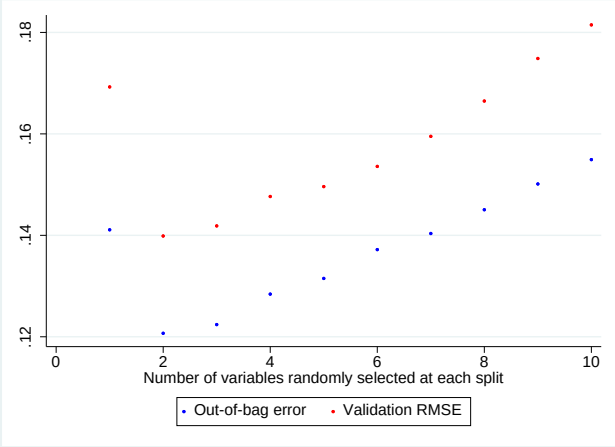

Fig. 3 PE changing with number of variables (N)

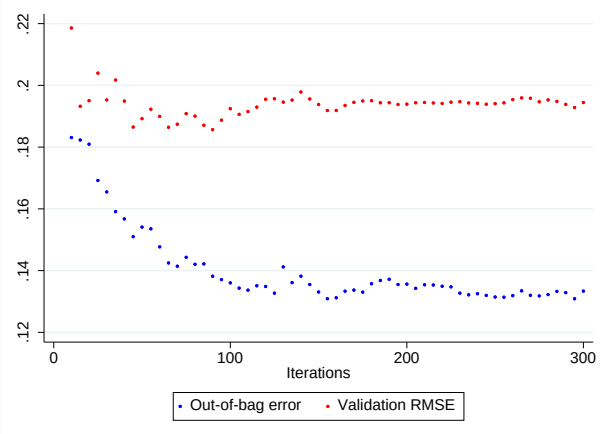

Fig. 4 PE changing with number of iterations (E)

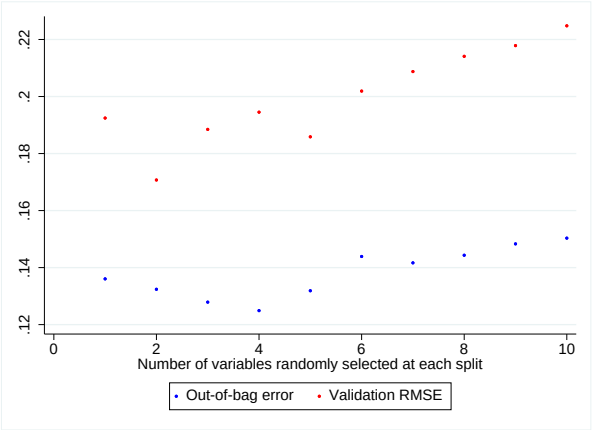

Fig. 5 PE changing with number of variables (E)

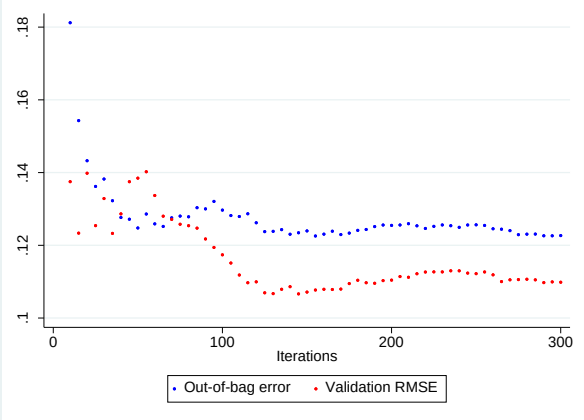

Fig. 6 PE changing with number of iterations (M)

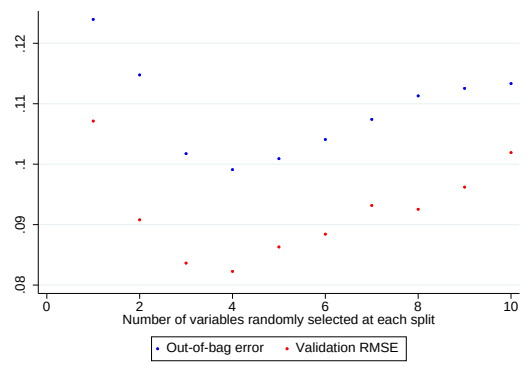

Fig. 7 PE changing with number of variables (M)

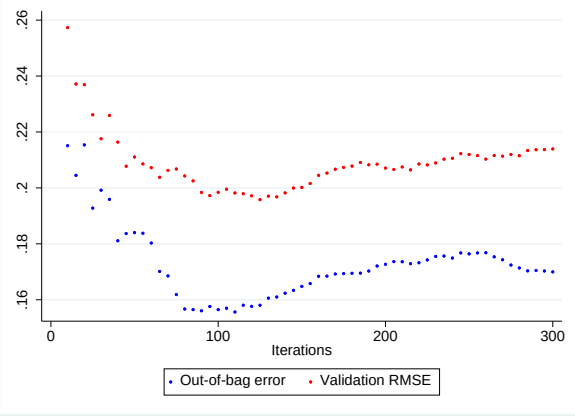

Fig. 8 PE changing with number of iterations (W)

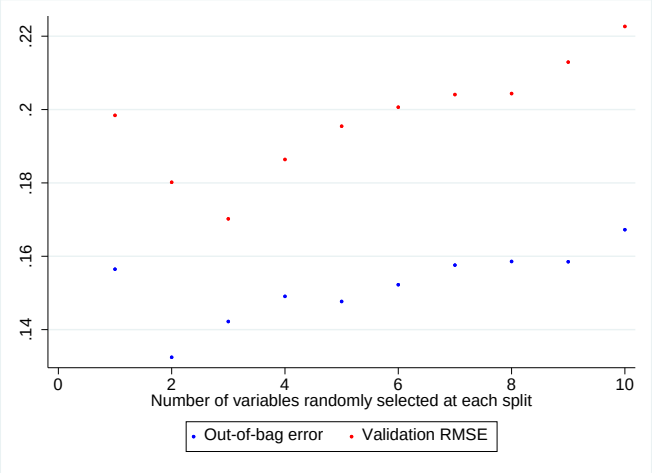

Fig. 9 PE changing with number of variables (W)

（Note：PE is short for prediction error. N is short for nationwide, E is short for Eastern Region, M is short for Middle Region, W is short for Western Region）

4.2 Importance score

Random Forest is a black box. This paper uses importance score graph to understand which variables affect the prediction effect. The data set selected in this paper contains 10 independent variables, so only the results with an importance score greater than 40% are shown. Figures 10 to 13 show the importance scores of the predictors nationwide and in the eastern, central and western regions.

First, on a national scale, Human resources, Internet development and Capital formation have a greater impact. It can be seen that both physical capital and human capital are decisive factors to promote economic development[11]. What is more noteworthy is that Internet development ranks second in importance score, reflecting the outstanding achievements of China's new infrastructure since the 18th National Congress of the Communist Party of China, which has been widely applied across the country and has gradually shown its role in promoting high-quality economic development[12].

Secondly, for the eastern region, Capital formation and Throughput of post and telecommunications play a great role in driving the information industry and the overall economy, and the promotion effect of Throughput of post and telecommunications on economic growth shows obvious regional differences, among which the impact on the eastern region is the biggest. This may be because the more population and industry concentration in prosperous regions, the more productivity increases will result from increased postal and telecommunications traffic[13].

Thirdly, for the central region, the fundamental driving force of economic growth still comes from domestic investment, which confirms the opinion of Liu (2013)[14]. The main influencing factors of economic development in the central region include Innovation and Human resources, which to some extent reflects the continuous optimization of resource allocation efficiency in the central region after the implementation of the strategy of the rise of the Central Region [15]. However, in the index of digital infrastructure, only the importance of Throughput of post and telecommunications scored more than 0.4, indicating that the central region should further improve the level of postal and telecommunications services and make them play a greater role in driving economic growth.

Finally, in the western region, the importance scores of Internet development and Post and telecommunications business volume are 0.76 and 0.79 respectively, ranking third and fourth respectively. In the East, both scored more than 0.8. By contrast, it can be found that the promotion effect of digital infrastructure-related indicators on economic development in central and western regions is weaker than that in eastern regions, which may be because the development characteristics of information infrastructure directly lead to the widening gap between the development and application of information networks in different regions[16].

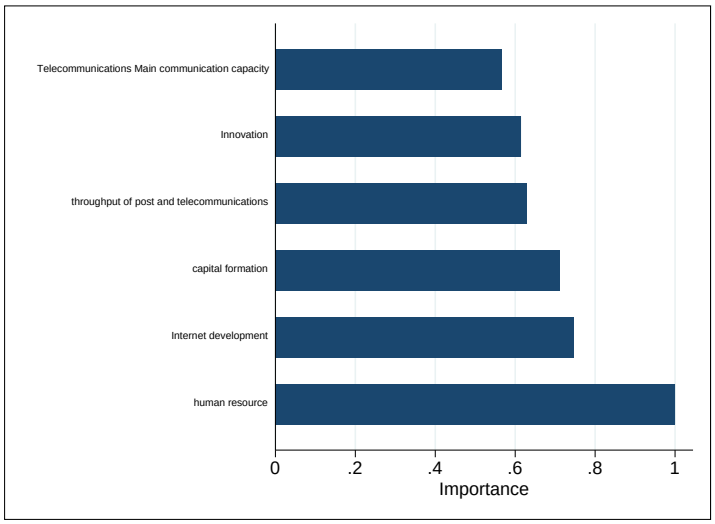

Fig. 10 Importance score of predictive factors(N)

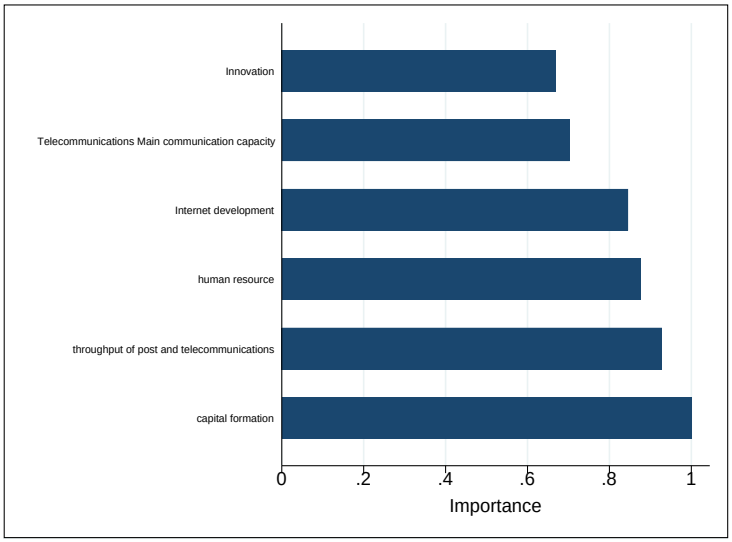

Fig. 11 Importance score of predictive factors(E)

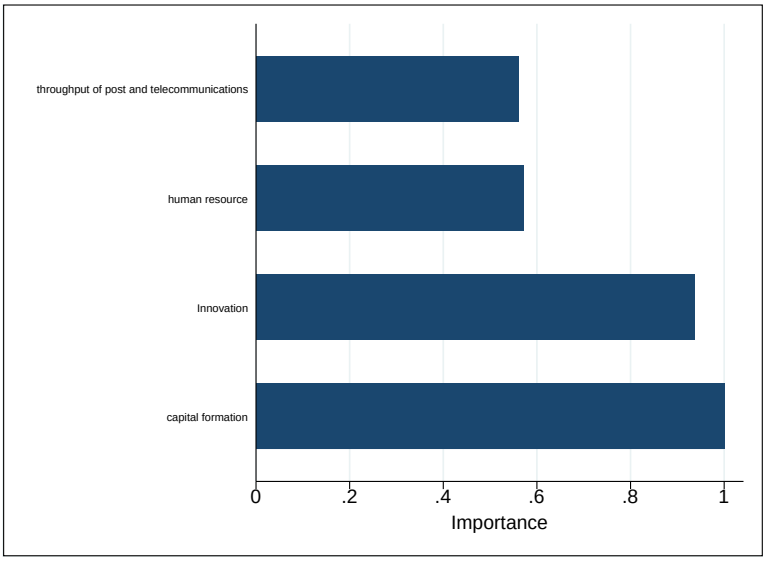

Fig. 12 Importance score of predictive factors(M)

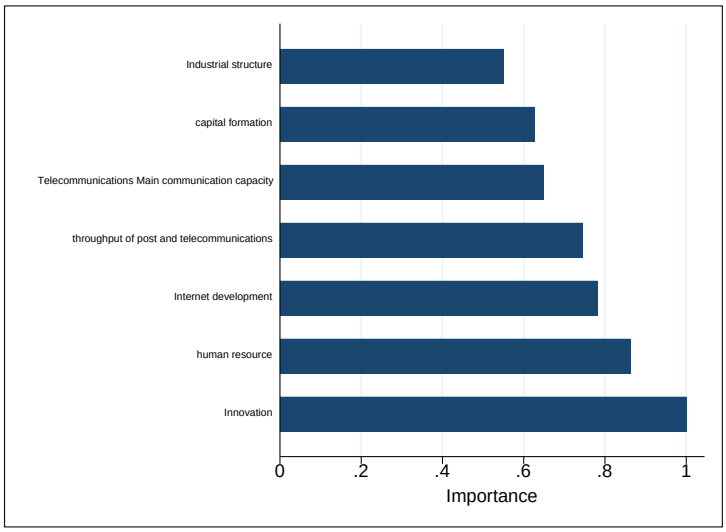

Fig. 13 Importance score of predictive factors(W)

According to Table.3., the results of OLS estimation are basically consistent with the importance score of Random Forest, and the P values of Human resources, Internet development, Telecommunications Main communication capacity, Innovation, Capital formation, Throughput of post and telecommunications, System and other variables are all less than 0.01 with large coefficients. In the eastern region, Innovation, Human resources, Industrial structure, Capital formation and Throughput of post and telecommunications have significant effects on economic development. Human resources, Innovation, System and Post and telecommunication service level in the central region are important variables affecting the macro economy. Innovation,Throughput of post and telecommunications and System in the western region are of great significance to GDP growth. However, one defect of OLS estimation compared with Random Forest is that when the significance levels and coefficients of multiple variables are similar, the importance of these variables to dependent variables cannot be compared.

Table.3. OLS estimation results of digital infrastructure on macroeconomic growth

Variable	Nationwide	East	Middle	West
	loggdp	loggdp	loggdp	loggdp
Internet development	-0.606***	-0.0948	-0.181	-0.254
	(0.153)	(0.168)	(0.156)	(0.294)
throughput of post and telecommunications	0.322***	0.160*	0.127	2.117***
	(0.0789)	(0.0835)	(0.115)	(0.300)
Telecommunications Main communication capacity	0.607***	0.114	0.156	-0.129
	(0.120)	(0.135)	(0.114)	(0.202)
Telecommunication service level	-0.0860	0.121	-0.0833	-0.0859
	(0.0549)	(0.0854)	(0.0672)	(0.0796)
Post and telecommunications service level	-0.0364	-0.0205	-0.167***	-0.0351
	(0.0428)	(0.0614)	(0.0415)	(0.0618)
system	0.362***	0.0930	-0.386***	1.212***
	(0.0438)	(0.0579)	(0.123)	(0.134)
Innovation	-0.501***	-0.264***	0.620***	-1.498***
	(0.0669)	(0.0878)	(0.175)	(0.547)
capital formation	0.415***	0.243**	0.0502	0.0776
	(0.0695)	(0.102)	(0.0675)	(0.137)

Industrial structure	0.201***	0.213***	0.0353	-0.0259
	(0.0343)	(0.0431)	(0.0315)	(0.0869)
human resource	0.574***	0.490***	0.303***	0.163
	(0.0702)	(0.0772)	(0.0899)	(0.213)
_cons	9.502***	9.614***	9.685***	9.865***
	(0.0264)	(0.0551)	(0.0651)	(0.204)
N	170	60	61	66
R^2	0.910	0.935	0.943	0.941
adj. R^2	0.904	0.922	0.932	0.930

Note: *, ** and *** represent significance levels of 10%, 5% and 1% respectively, and the numbers in brackets are the standard deviations of the coefficients of the corresponding variables.

This paper sorts out the accuracy of Random Forest model and Ordinary Least Square model in predicting GDP of the whole country and eastern, central and western regions, as shown in Table.4. It can be seen that the prediction error of Random Forest Regression is much smaller than that of OLS regression, no matter the sample scope is the whole country, eastern, central or western regions, which reflects the advantages of RFR combinational learning algorithm to some extent and confirms the conclusion of Fang et al. (2014)[7]. In order to further illustrate the accuracy of the model, this paper randomly selects the logarithmic values of real GDP and the predicted values of the model in 4 provinces (Shanghai, Jiangsu, Zhejiang and Anhui). The prediction results of the Random Forest model and OLS model are shown in Figure 14 and 15 respectively. The two figures intuitively show that the prediction result of RFR model is closer to the real value than that of OLS model, especially for zhejiang and Anhui, RFR shows obvious superiority over OLS.

Table.4. Comparison of accuracy between RFR and OLS prediction models

Sample Range	RFR		OLS	
	OOB-error	Test set error	Training set error	Test set error
Nationwide	0.1207	0.1399	0.3327	0.3808
East	0.1324	0.1717	0.2338	0.2898
Middle	0.0991	0.0823	0.1384	0.1118
West	0.1422	0.1702	0.2702	0.4782

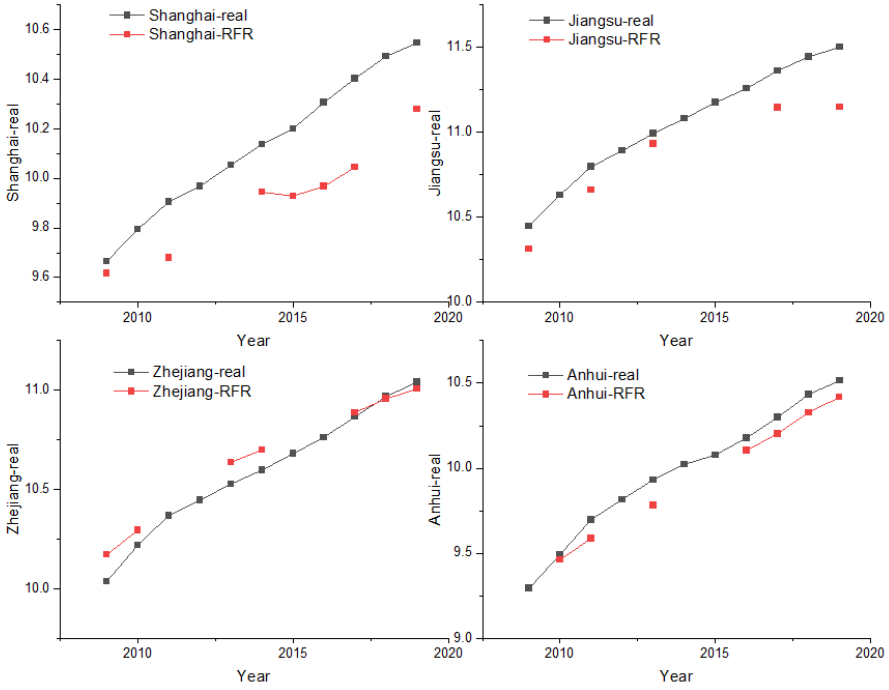

Fig.14 Comparison of the logarithm of GDP with the predicted value of RFR model

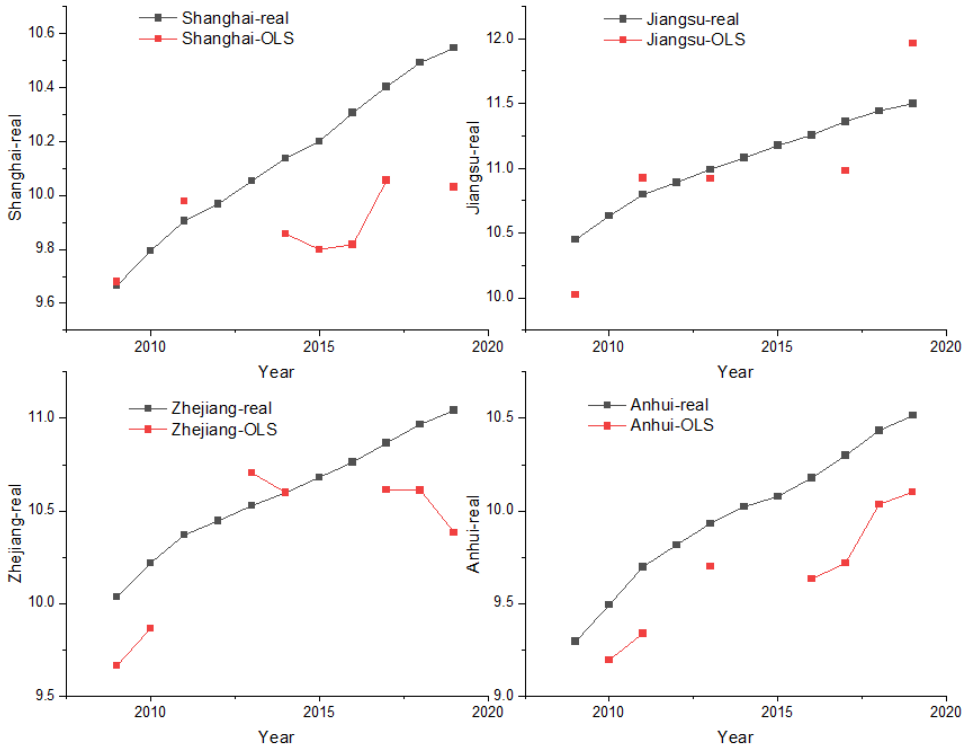

Fig.15 Comparison of the logarithm of GDP with the predicted value of OLS model

5. Conclusions

From the perspective of communication, this paper empirically examines the influence mechanism of digital economy on macroeconomic growth through digital infrastructure. Based on the panel data of provinces from 2009 to 2019, this paper constructs five digital infrastructure early-warning indexes and five conventional early-warning indexes for 31 provinces (municipalities and autonomous regions) by using principal component analysis. Combined with Random Forest regression (RFR), this paper forecasts the GDP of China, and compares its prediction results and accuracy with ordinary least square method. Therefore, the following conclusions are drawn:

First, the digital economy can have an impact on GDP through the construction and application of digital infrastructure, and the digital infrastructure early-warning index can be used as a predictor of GDP. At the national level, digital infrastructure plays a more significant role in driving the economy than digital infrastructure construction. Internet development ranks second only to Human resources in importance score.

Secondly, the importance of digital infrastructure early-warning indexes has regional heterogeneity, and it exerts the greatest influence on the economically developed eastern region. For the eastern region, Capital formation and Throughput of post and telecommunications play a great role in driving the information industry and the overall economy. The fundamental driving force of economic growth in the central region still comes from domestic investment, while other major factors include Innovation, Human resources and Post and telecommunication service level. In the western region, the importance of Internet development and Throughput of post and telecommunications ranks third and fourth respectively, while Innovation and Human resources remained the most important factors.

Third, Random Forest is an effective machine learning method. Compared with ordinary least square method, it can not only identify the importance of variables more effectively, so as to better analyze macro problems, but also have better prediction effect. In both the national and local regions, the out-of-bag error and test set error of Random Forest were less than 0.2, while the test set error of OLS regression was generally greater than 0.2. No matter how large the sample range is, the prediction accuracy of RFR is higher than that of OLS in terms of numbers or images.

According to the conclusion, this paper puts forward the following suggestions. First of all, compared with the eastern region, it is more necessary for the central and western regions to strengthen the construction of digital infrastructure, so that the digital economy can play a bigger role in driving economic growth. While developing digital infrastructure, appropriate consideration should also be given to the application of digital infrastructure to upgrade the development of the Internet and increase the volume of postal and telecommunications services. Secondly, in the knowledge age of aging population, digital industrialization is in full swing, high-quality talents are playing a more indispensable role in economic growth. While enlarging the stock of human capital, China should also pay attention to improving the efficiency of human capital use and transforming it into productive forces to realize the sustainable development of economy. Finally, new research methods and new economy-related data need to be introduced into research departments' macroeconomic forecasting models. In this paper, we prove the effectiveness of Random Forest Regression in dealing with macro - forecast problems in China.

Acknowledgements

The authors gratefully acknowledge the support from Professor Pu.

References

[1] CHEN L, LI J, XU C. Information Infrastructure and Economic Growth: An analysis based on Chinese provincial data[J]. Management Science, 2011,24(1): 98-107.

[2] XU X, LIU L. The impact of information infrastructure construction on Shanghai's economic transformation: Based on regional CGE simulation analysis[J]. East China Economic Management, 2014,28(7): 11-14.

[3] JIANG W, FAN J, XIAOLAN Z. China's "new infrastructure" : Investment multipliers and their effects[J]. Nanjing Social Sciences, 2020(04): 20-31.

[4] BREIMAN L. Random Forests[J]. Machine Learning, 2001,45(1): 5-32.

[5] LIU X, WU D, ZEWDIE G K, et al. Using machine learning to estimate atmospheric Ambrosia pollen concentrations in Tulsa, OK[J]. Environmental Health Insights, 2017,11: 1-10.

[6] SCHONLAU M, ZOU R Y. The random forest algorithm for statistical learning[J]. Stata Journal, 2020,20(1): 3-29.

[7] FANG Q, WU J, XIE B. Profit contribution of insurance customers based on random forest[J]. Application of Statistics and Management, 2014,33(06): 1122-1131.

[8] CAWLEY G C, TALBOT N. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation[J]. Journal of Machine Learning Research, 2010,11(1): 2079-2107.

[9] KUHN M. Buiding predictive models in R using the caret package[J]. Journal of Statistical Software, 2008,28(5): 1-26.

[10] ZHANG X L, JIAO Y X. China's digital economy development index and its application[J]. Zhejiang Social Sciences, 2017(04): 32-40.

[11] HUANG Y, DING X, CHEN R, et al. An empirical Analysis of the contribution of Human Capital and material Capital to economic growth —— Commemorating the 60th anniversary of the birth of human Capital Theory [J]. journal of east china normal university, 2020,38(10): 21-33.

[12] TIAN J, YAN D. New Infrastructure and industrial Internet: "Roads and vehicles" for the accelerated digital economy after COVID-19 [J]. Journal of Shandong University, 2020,240(03): 7-14.

[13] WANG W, ZHANG H. Information Infrastructure and regional economic growth: Empirical evidence from 252 prefecture-level cities in China [J]. East China Economic Management, 2018,7(32): 75-80.

[14] LIU A. Evaluation model and demonstration of economic growth in central China[J]. Statistics and Decision making, 2013(6): 60-62.

[15] MIAO Y, YANG D. Analysis of the allocation efficiency of science and technology resources in central China based on comprehensive evaluation method [J]. China Soft Science, 2020,351(3): 139-154.

[16] HAN P, YAN G. The Gap and influence of Internet Information Network application in east and West China[J]. Forum on Science and Technology in China, 2007(6): 81-86.