An Empirical Study on Influencing Factors of China's Tax Revenue Based on Principal Component Regression Model

Lijuan Zeng(✉) and Linxin Liu

School of Economics and Trade, Guangzhou Xinhua University, Guangzhou, China
selenazlj@126.com

Abstract. This paper analyzes the influencing factors of tax revenue from an empirical point of view, adopts the principal component analysis method to reduce the dimension and simplifies the regression model, and on this basis introduces the BP neural network method to predict the tax revenue. The empirical research conclusion shows that the national fiscal expenditure has the greatest impact on tax revenue, with a corresponding elasticity coefficient of 43.89%, followed by the total fixed asset investment and the added value of the tertiary industry, and the second industry's added value and social consumer goods retail sales have a relatively small impact. The results can provide a useful reference for the formulation of China's tax policy.

Keywords: Tax income · Influencing factor · Principal component analysis · BP neural network method

1 Preface

As a financial guarantee for national governance, tax revenue has played an important and active role in optimizing domestic resource allocation, promoting social welfare, and improving residents' living standards. In the context of the new normal of the economy under the influence of COVID-19, it is crucial for tax revenue to play a role in the government's performance of public functions as a source of funds and the display of government financial resources. This paper takes China's tax revenue from 1990 to 2020 as the research object, based on the principal component analysis method to empirically analyze the influencing factors of tax revenue, and use the BP neural network method to make predictions, to provide actionable advice for adjusting China's tax revenue and promoting economic growth.

2 Literature Review

Most scholars use the multiple regression analysis method to analyze the influencing factors of tax revenue. Tang Xianbin (2012) [1] analyzed relevant data in China from 1990 to 2008 and found that GDP and fixed asset investment in the whole society had

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 615–627, 2023.
https://doi.org/10.2991/978-94-6463-198-2_63

the greatest impact on tax revenue and proposed that fixed asset investment should be strengthened to increase tax increases. Liu Junhang (2020) [2] used a multiple regression analysis model to conduct a statistical analysis of the tax data from 1990 to 2016. The model study concluded that GDP, money supply, fiscal expenditure and total retail sales of social consumer goods have a significant impact on tax revenue. Cao Qianqian et al. (2020) [3] conducted multiple regression analysis on the macroeconomic data from 2000 to 2018, and concluded that the added value of the tertiary industry and the total fixed asset investment had a significant impact on the total tax revenue, and believed that traditional influencing factors of GDP, total import and export, and fiscal expenditure have no significant impact on tax revenue, and propose suggestions for promoting the development of the national economy from the perspectives of investment and the development of the tertiary industry. Zhao Shanshan (2020) established a multiple regression analysis model on the macro data from 1990 to 2019 and found that the tertiary industry has a significant impact on China's tax revenue and believes that the retail price index of commodities has no significant impact on China's tax revenue [4].

At present, there are two main types of influencing factors of tax revenue. The first type held by scholars believes that GDP, fiscal expenditure, and total retail sales of consumer goods have a significant impact on tax revenue. The second type held by scholars believes that the development of the tertiary industry has a significant impact on tax revenue. After a comprehensive analysis, considering the impact of the epidemic, this paper combines the existing research results, gathers available influencing factors to conduct multiple regression analysis, and selects variables that have a significant impact on tax revenue.

3 Data Selection and Testing

3.1 Factor Selection and Data Collection

In order to specifically analyze the impact of various factors on China's tax revenue, this case collects relevant data from 1990 to 2020 in China by consulting relevant literature and combining previous research on tax revenue, and research is conducted on China's tax-related influencing factors.

All explanatory variables are shown in Table 1, which are mainly divided into two parts: financial factors and national economic factors.

3.2 Descriptive Statistical Analysis

To reduce the influence of heteroscedasticity, a logarithm is performed on all variables, and the scatter plot is used to verify whether the impact of the selected factors on tax revenue is positively correlated with the economic theory, as shown in Fig. 1.

To further understand the strength of the linear correlation, the software is used to calculate the correlation coefficient matrix, as shown in Table 2.

According to the correlation coefficient matrix in Table 2, the correlation coefficients between tax revenue lnY and each explanatory variable are all above 0.99, indicating that there is a strong positive correlation between tax revenue and their respective variables. The judgment is consistent. The correlation coefficients between the independent

Table 1. Variable selection

Variable type		Variable name	Unit
Explained variable		Tax revenue (Y)	RMB 100 million
Explanatory variable	Fiscal factor	National fiscal expenditure (X_1)	RMB 100 million
	Economic factor	Added value of the second industry (X_2)	RMB 100 million
		Added value of the tertiary industry (X_3)	RMB 100 million
		Total investment in the fixed asset (X_4)	RMB 100 million
		Total retail sales of consumer goods (X_5)	RMB 100 million

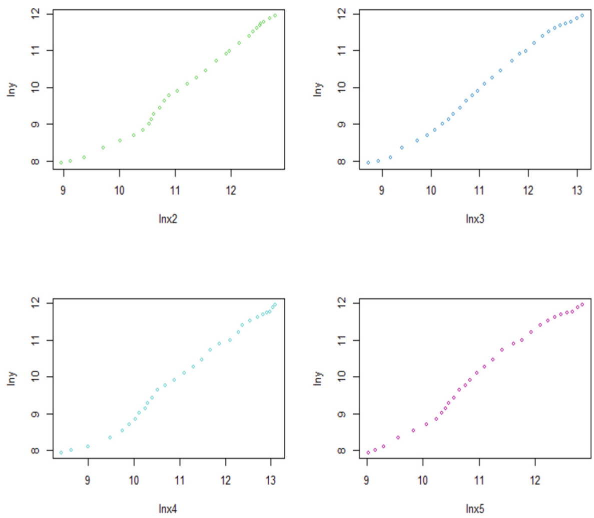

Fig. 1. Scatter plots of various factors and tax revenue (lnY)

variables lnX_1, lnX_2, lnX_3, lnX_4, $andlnX_5$ all exceed 0.99, so serious multicollinearity problems may occur when establishing a double logarithmic regression model.

3.3 Stationary Test

To prevent "pseudo-regression", the unit root test is performed on each variable one by one, and the unit root test is performed from the original sequence, the first-order difference sequence, and the second-order difference sequence in turn, until the difference sequence is stationary. According to the software output results, it is found that the original sequence and the first-order sequence of lnX_1, lnX_2, lnX_3, lnX_4, andlnX_5 all

Table 2. Correlation coefficient matrix

	lnY	lnX_1	lnX_2	lnX_3	lnX_4	lnX_5
lnY	1	0.999	0.991	0.995	0.992	0.993
lnX_1	0.999	1	0.990	0.996	0.992	0.994
lnX_2	0.991	0.990	1	0.997	0.997	0.995
lnX_3	0.995	0.996	0.997	1	0.997	0.998
lnX_4	0.992	0.992	0.997	0.997	1	0.996
lnX_5	0.993	0.995	0.995	0.998	0.996	1

Table 3. Unit root test result

Variable	Difference Order	ADF value	1% critical value	5% critical value	10% critical value	Conclusion
lnY	2	−7.48	−2.65	−1.95	−1.60	stationary
lnX_1	2	−5.63	−2.65	−1.95	−1.60	stationary
lnX_2	2	−4.93	−2.65	−1.95	−1.60	stationary
lnX_3	2	−4.31	−2.65	−1.95	−1.60	stationary
lnX_4	2	−5.40	−2.65	−1.95	−1.60	stationary
lnX_5	2	−3.15	−2.65	−1.95	−1.60	stationary

have unit roots, and the sequence after the second-order difference rejects the original hypothesis H_0, that is, the second-order sequence is stationary. Use the *eviews* software to output the unit root test results, the results are shown in Table 3.

3.4 Johansen Cointegration Test

It can be seen from the above analysis that there is an integration of order two between all variables. To prove whether there is a statistical long-term equilibrium relationship between variables and effectively avoid the problem of false regression, this paper conducts the Johansen test on the variables. The Johansen test is based on the eigenvalue test under the VAR model, so the VAR model needs to be established to determine the optimal lag order. The output results of the *eviews* software are shown in Table 4.

According to the results in Table 4, the first column of the table represents the lag order of the selection of each factor. According to the five criteria of LR, FPE, AIC, SC and HQ, select the row with the most asterisks, and after comprehensive judgment, establish an order one vector autoregression VAR(1). In the stationary test, lnY, lnX_1, lnX_2, lnX_3, lnX_4, andlnX_5 are integrated to order two, so the Johansen test can be performed, and the results are shown in Table 4.

According to Table 5, when the significance level is 5%, the null hypothesis of None, At most 1, and At most 2 are rejected, indicating that the model has at least 3 cointegration

Table 4. Optimal lag order selection

Lag	LogL	LR	FPE	AIC	SC	HQ
0	369.41	NA	1.90e−18	−23.82	−22.10	−23.29
1	450.92	93.15*	1.0e−19*	−27.07	−23.64*	−26.02*
2	492.86	91.64	1.96e−19	−27.49*	−27.35	−25.92

Table 5. Maximum eigenvalue test results

Number of Cointegrations	Eigenvalue	Max eigenvalue statistics	5% critical value	P value
None *	0.8845	62.599	40.077	0.0000
At most 1 *	0.7512	40.350	33.876	0.0074
At most 2 *	0.6197	28.039	27.584	0.0437
At most 3 *	0.4528	17.488	21.131	0.1502
At most 4 *	0.3332	11.754	14.264	0.1202
At most 5 *	0.2259	7.426	3.841	0.0064

equations, and the null hypothesis of At most 3 is not rejected. There are at most three cointegration equations between variables. In summary, there are three cointegration equations between the variables, which means that there is a long-term stable relationship between the variables. At the same time, combined with economic theory, it is known that there is also an economic long-term relationship between variables, so pseudo-regression can be effectively avoided when building a model.

4 Research on Influencing Factors of Tax Revenue Based on Principal Component Analysis

4.1 Establish Principal Component Regression Model

Principal component regression is an improvement on ordinary least squares estimation, and its parameter estimation is still a biased estimation. Principal component analysis based on multivariate statistical analysis proposes principal component regression, which inherits the advantages of principal component analysis and ensures that each principal component is mutually orthogonal, so there is no multicollinearity problem, but it also produces a new problem of loss information which causes the estimation to be biased. When the number of independent variables is large and the variables are highly correlated, this method is very suitable. Dimension can greatly simplify the model while avoiding multicollinearity. Next, the model is solved by R software, the data from 1990 to 2018 is used for modeling, and the data in 2019 and 2020 is used for prediction.

4.2 Model Results Analysis

According to the extraction results of the principal components in Table 6, the first column is the standard deviation of the five principal components, that is, the arithmetic square root of the eigenvalues corresponding to the principal components. The second column is the proportion of the variance of each principal component, reflecting the proportion of the variation of the data that can be explained by principal components, that is, the proportion of information that contains the original data. The third column is the cumulative proportion, and the variance percentage of the first principal component F_1 is 99.68%, which contains almost all the information of the original data. Based on the principle that the cumulative contribution rate is more than 80%, one is selected. Combined with the gravel diagram as shown in Fig. 2, it can be seen that the second principal component reaches the inflection point, so one principal component is extracted.

According to the principle of extracting a principal component, the principal component coefficient table obtained by R software is shown in Table 7, so

$$F_1 = 0.4466lnX_1 + 0.4469lnX_2 + 0.4477lnX_3 + 0.4474lnX_4 + 0.4473lnX_5$$

The first principal component F_1 mainly extracts the common properties of variables lnX_1, lnX_2, lnX_3, lnX_4, and lnX_5, focusing on the macroeconomic situation of economic development. Next, establish a univariate regression model of F_1 and lnY, first use R software to draw the scatter plot of F_1 and lnY as shown in Fig. 3.

From the scatter plot of F_1 and lnY, the principal component F_1 and tax income lnY show a strong positive correlation, so a univariate linear regression model is established

Table 6. Principal Component Extraction Results

Component	stdev	% stdev	Cumulative contribution rate
1st component	2.2325	0.9968	0.9968
2nd component	0.0983	0.0019	0.9987
3rd component	0.0602	0.0007	0.9994
4th component	0.0455	0.0004	0.9998
5th component	0.0253	0.0001	1.00000

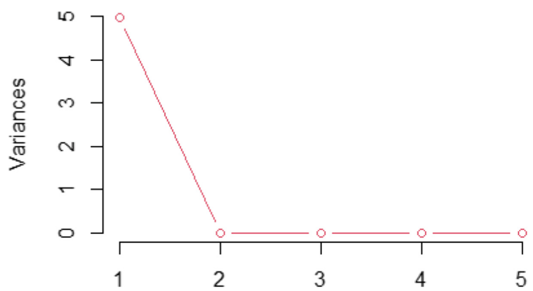

Fig. 2. Principal component gravel diagram

Table 7. Principal component coefficient table

Variable	F_1
lnX_1	0.4466
lnX_2	0.4468
lnX_3	0.4477
lnX_4	0.4474
lnX_5	0.4473

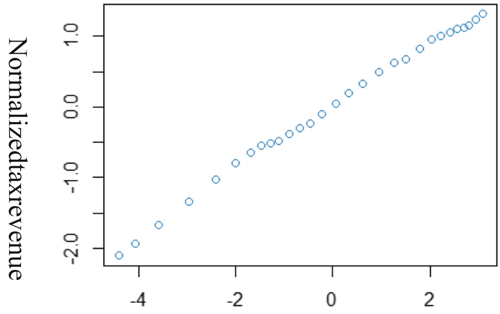

Fig. 3. Scatter plot of F_1 and lnY^*

for the principal component F_1 and tax income lnY, and the regression results are output by R software as shown in Table 8.

According to Table 8, the regression equation is $\widehat{lnY}^* = 0.4469 * F1$. Bring the principal component F_1 of the above formula into $\widehat{lnY}^* = 0.1995lnX_1^* + 0.1997lnX_2^* + 0.19994lnX_3^* + 0.19994lnX_4^* + 0.19989lnX_5^*$, the corresponding unnormalized equation is: $\widehat{lnY} = -0.7279 + 0.1908lnX_1 + 0.1438lnX_2 + 02318lnX_3 + 0.18877lnX_4 + 0.23166lnX_5$

For the univariate linear regression model, the variable significance test and the equation significance test are equivalent. The corresponding p-value of the t-test is less than 0.05, that is, the null hypothesis is rejected, so F_1 has a significant impact on lnY, and the goodness of fit $R^2 = 0.9943$, That is, about 99.43% of the change in tax revenue can be explained by the principal component F_1, and the image combined with the fitted

Table 8. Principal component regression analysis

	Coef	Stdev	T value	P value
F_1	0.4469	0.0059	75.56	0
	R^2=0.9953	F=5710	n=29	

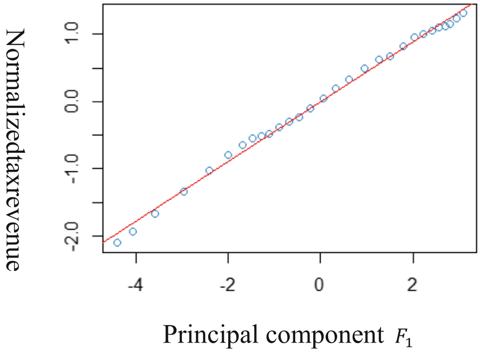

Fig. 4. True value vs regression curve

regression curve and the true value is shown in Fig. 4. It can be seen that the fitting is good.

From normalized equation coefficients in $ln\widehat{Y}^* = 0.1995lnX_1^* + 0.1997lnX_2^* + 0.19994lnX_3^* + 0.19994lnX_4^* + 0.19989lnX_5^*$, it can be seen the coefficients are very close, so variable lnX_1, lnX_2, lnX_3, lnX_4, and lnX_5 have a significant effect on lnY, and the effect is similar.

4.3 Analysis of Influencing Factors of Principal Component Regression

In this paper, a principal component regression model is established to resolve multicollinearity, and a principal component is selected according to the cumulative contribution rate to establish a univariate linear regression model, and the standardized equation of principal component regression is obtained as follows:

$$ln\widehat{Y}^* = 0.1995lnX_1^* + 0.1997lnX_2^* + 0.19994lnX_3^* + 0.19994lnX_4^* + 0.19989lnX_5^*$$

According to the standardized regression equation obtained by principal component regression, it can be seen that under the condition that other conditions remain unchanged:

1) For every 1% increase in national fiscal expenditure (X_1), tax revenue (Y) will increase by an average of 0.1995%;
2) For every 1% increase in the added value of the secondary industry (X_2), the tax revenue (Y) will increase by an average of 0.1997%;
3) For every 1% increase in the added value of the tertiary industry (X_3), the tax revenue (Y) will increase by an average of 0.19944%;
4) For every 1% increase in fixed asset investment (X_4) of the whole society, tax revenue (Y) will increase by an average of 0.19994%;
5) For every 1% increase in the total retail sales of consumer goods (X_5), the tax revenue (Y) will increase by 0.19989% on average;

The conclusion drawn by the principal component regression is that the national fiscal expenditure (X_1), the added value of the secondary industry (X_2), the added value

Table 9. Principal component regression error analysis

	Y true value	Y predicted value	Absolute error	Relative error
2019	158000.5	165263.6	7263.085	0.0459
2020	154312.3	166975.5	12663.23	0.082062

of the tertiary industry (X_3), the investment in fixed assets of the whole society (X_4) and the total retail sales of consumer goods (X_5) are significantly related to each other. The effects of tax revenue (Y) are very close, and it can be considered that each variable has the same effect on tax revenue (Y).

The multiple regression model based on principal component analysis completely overcomes the problem of multicollinearity and greatly reduces the complexity of the model by extracting the principal components. The absolute error and relative error are shown in Table 9.

It can be seen from Table 9 that the error of principal component regression is relatively ideal, ranging from 0.04 to 0.09. By consulting relevant literature, most former scholars have an error of about 6% in tax prediction, indicating that this method is more suitable for tax prediction. This assumes that there are two reasons for the forecast error of more than 0.08 in 2020, First, due to the outbreak of COVID-19 in 2020, tax revenue has experienced negative growth for the first time in nearly 30 years, with large fluctuations, resulting in a decrease in estimation accuracy; second, the selected factors cannot fully reflect the impact of tax revenue fluctuations, there are still some less influential factors that have not been taken into account in the model. There are many assumptions in classical regression analysis. In many cases, it is difficult to meet the assumptions of multicollinearity. In this paper, the assumption of multicollinearity is not satisfied, resulting in a biased estimator. The BP neural network method of data mining will be used to avoid such problems, the biggest advantage of this method is that there is no requirement for the data itself, which can avoid the assumptions and shortcomings of the classical regression analysis.

4.4 Prediction of Tax Revenue by BP Neural Network

In this paper, BP neural network is used to predict the tax revenue to verify the correctness of the index of the principal component analysis model. The input layer is five factors (lnX_1, lnX_2, lnX_3, lnX_4, and lnX_5), the output layer is tax revenue (Y), and the 1990–2018 data tax revenue is also used as a training sample, with 2019- 2020 as a test sample. Since the BP neural network is different from the classical regression analysis, it is not necessary to require data stationarity and logarithmization. It is required to normalize the data and then perform the BP neural network modeling. The R language modeling results are shown in Fig. 5 and Table 10 show:

Figure 5 shows that a total of 64 iterations have been performed. At the end of the iteration, the loss function is 0.003, and the *compute* function is used to predict the tax revenue in 2019 and 2020. The results are shown in Table 10:

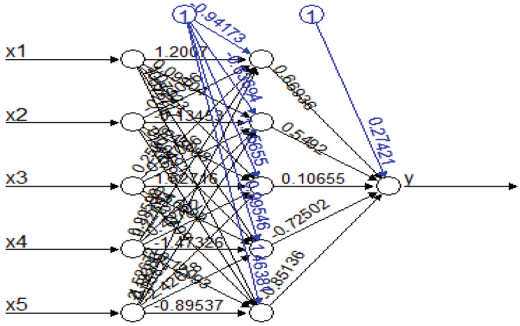

Fig. 5. Neural network topology illustrated by *neuralnet* function

Table 10. *BP* neutral network error analysis

	Y true value	*Y* predicted value	Absolute error	Relative error
2019	158000.5	156357.2	−1643.5	−0.0104
2020	154312.3	157795.6	3483.3	0.0226

The *BP* neural network has a better effect on the prediction of this model, and the prediction accuracy is higher than that of the traditional regression analysis. Some unique advantages are found in its application process. The first advantage is the strong self-learning and adaptive ability, when the BP neural network is running, it can automatically generate appropriate rules, and can memorize the content to be learned according to the weight of the network, so it has strong self-learning and adaptive ability. The second advantage is nonlinear mapping capability, the essence of the BP neural network is to realize a mapping function from input to output. The mathematical theory proves that BP neural network with more than three layers can approximate all nonlinear continuous functions with arbitrary precision. The third advantage is fault tolerance, which can continue to work normally even when the system is partially damaged. In this model, BP neural network has obvious advantages compared with regression analysis, but BP neural network also has some obvious shortcomings compared with regression analysis. First, BP neural network is only suitable for prediction, and cannot study the influencing factors of explained variables like regression analysis. Second, since the BP neural network initializes each connection weight and the bias weight of the output node at the beginning, the default value is a random number that obeys a uniform distribution between -0.5 and 0.5, so the results of each operation are different. It is the biggest disadvantage of this method. Although the results of this experiment are very close to the real value, because of its randomness, its predicted value is unstable and has a large fluctuation range, and the reliability of the final result is relatively low reliability. With the help of the law of large numbers, the model is repeated n times, and the results of the n runs are averaged. After continuous attempts, it was found that when n was 5000, the results of multiple runs were very close. The final results are shown in Table 11:

Table 11. Optimized *BP* neural network error analysis

	Y true value	Y predicted value	Absolute error	Relative error
2019	158000.5	158093.5	93	0.00058
2020	154312.3	159161.6	4849.3	0.0314

According to Table 11, the relative error of prediction for 2019 is 0.058%, and the relative error of prediction for 2020 is 3.14%. The prediction accuracy of the BP neural network for 2019 is very high, but the prediction accuracy in 2020 is relatively low. There are two possible reasons. First, as the prediction time is longer, the prediction accuracy will gradually decrease. Second, due to the impact of COVID-19 in 2020, tax revenue has also fluctuated greatly, and it is the first time in nearly 30 years that tax revenue has declined. After comparing the forecast value in 2020 and the forecast value in 2019, the difference between the forecast value is not large, indicating that the selected factors can reflect tax revenue well to a certain extent.

5 Conclusions

Based on Principal Component Regression Model, this model has obvious advantages. First, it solved the multicollinearity problem in a better way. Second, it reduced the complexity of the model and the amount of calculation by compressing the five variables into one principal component through dimension reduction. In the case where the variance contribution rate of selecting one principal component has reached 99.68% (>85%), selecting two principal components will increase the complexity and interpretation of the model. Thus, one principal component was selected in this paper, and the heterogeneity between variables were not reflected. Third, the model can be tested by Principal Component Regression, but dimension reduction resulted in loss of information and biased the estimator. Finally, the factors that affect tax revenue (Y) are found to be the national fiscal expenditure (X_1), the added value of the secondary industry (X_2), the added value of the tertiary industry (X_3), the total fixed asset investment (X_4), and the total retail sales of consumer goods (X_5).The empirical research conclusion shows that the national fiscal expenditure has the greatest impact on tax revenue, with a corresponding elastic coefficient of 43.89%, followed by the total fixed asset investment and the added value of the tertiary industry, with elastic coefficients of 15.93% and 15.53%, respectively. The added value of the secondary industry and the total retail sales of consumer goods have relatively little impact on tax revenue.

6 Related Advice

6.1 Optimizing the Industrial Structure and Guaranteeing Tax Revenue

Since the outbreak of COVID-19, the added value of the secondary industry has steadily increased, while the added value of the tertiary industry has declined year by year. The data show that the tertiary industry is more sensitive and resilient to emergencies.

Under the new normal of the epidemic, the upgrading of the industrial structure should be accelerated, the optimization of the internal structure of the industry should be promoted, the progress and innovation of science and technology should be promoted, the overall technical level of the industry should be improved, the development of the environmental protection industry should be promoted, and the sustainable development of the economy should be maintained by reducing pollution. At the same time, the government can make every effort to develop the tertiary industry, rationally formulate preferential policies for the financial and service industries, promote the rapid development of the tertiary industry, and ensure tax revenue.

6.2 Paying Attention to the Overall Investment Channels and Increasing the Investment in Infrastructure

The analysis of the influencing factors of tax revenue shows that fixed asset investment has a strong positive correlation with tax revenue. To coordinate the promotion of epidemic prevention and control and economic and social development, the government should give full play to the key role of effective investment, give full play to the role of private investment, strengthen research on new investment plans, increase investment in public health and emergency supplies, and accelerate national major projects. The progress of infrastructure construction, such as the planning and construction of 5G networks, has contributed to the growth of tax revenue.

6.3 Streamlining Fiscal Spending and Reducing the Fiscal Deficit Ratio

The main source of fiscal revenue comes from tax revenue. Under the circumstance that fiscal revenue has been greatly reduced due to the impact of the epidemic, to prevent and control the epidemic and recover the economy, and avoid further increase in the fiscal deficit rate, it is necessary to reduce those unnecessary and inefficient fiscal expenditures, improve the management level of fiscal expenditure, improve the efficiency of capital use, and at the same time, the government can issue national bonds to achieve the purpose of reducing the fiscal deficit rate.

6.4 Actively Carrying Out Tax Analysis and Forecasting to Give Full Play to the Government's Functions

Tax forecasting is related to the formulation of tax planning, the accuracy of tax forecasts, the scientific decision-making of the tax department and the scientific policy. Excessive tax revenue may lead to illegal tax collection, while too low tax revenue may lead to loose collection and management, so tax analysis and prediction can be actively carried out, the accuracy of tax prediction can be improved, and its functions can be fully utilized.

References

1. Tang Xianbin. Influencing factor analysis of China's tax revenue [J]. Economist, 2012 (05): 180-181.

2. Liu Junhang, Ren Da. Multi-variate regression analysis and prediction of factors influencing tax revenue [J]. Journal of Gansu Sciences, 2020,32(01):129-136.
3. Cao Qianqian, Zhu Jiaming, Jin Jing. Quantitative analysis of factors influencing China's tax revenue based on multivariate regression analysis [J]. Journal of Natural Sciences of Harbin Normal University, 2020,36(04):9-15.
4. Zhao Shansha, Liu Wei. Empirical analysis of factors influencing China's tax revenue [J]. Tax, 2021,15(16):23-24.
5. Ouyang Yuqi. Tax revenue prediction model and empirical analysis [D]. Jilin University of Finance and Economics, 2019.
6. Ren Shuang, Cui Haibo. Application SARIMA time series in tax revenue prediction of Hubei [J]. Journal of Hubei University (Natural Science Ed.),2021,43(01):80-85.
7. Yang Wei. Empirical analysis of factors influencing China's tax revenue [J]. Investment and Entrepreneurship, 2020,31(19):123–125+140.