Contribution Analysis of Regional Tax Factors Based on RBF Machine Learning

Yifan Wang [1, #], He Wang [2, #], Wenli Chen [3, #, *]

[1] China Academy of Public Finance and Public Policy, Central University of Finance and Economics, 100098, Beijing, China

[2] School of Public Finance and Taxation, Dongbei University of Finance and Economics, 116025, Dalian, China

[3] School of Economics, Guangzhou City University of Technology, 510800, Guangzhou, China

* Corresponding author: 2504758035@qq.com

Abstract. Tax revenue is one of the necessary conditions for ensuring economic stability and development in market economy countries, which affects the rational allocation of social resources, and the influence of tax relationships has become a necessary research topic. The artificial neural network has been widely used in many fields as an effective method to deal with complex nonlinear problems. Based on the data on the local economy, population, environment, and policy in North China from 2011 to 2020, this paper analyzes the four factors affecting tax revenue, establishes an artificial neural network-radial basis function(ANN-RBF)model, and uses SPSS to calculate the importance of normalization of regional tax contribution in four aspects of the local economy, population, environment, and policy. This study has particular enlightening significance for the country to formulate and use tax policies for macro-control.

Keywords: ANN-RBF model, contribution analysis, influencing factors, regional taxation.

1. Introduction

Taxation is a fundamental part of local fiscal revenue and an important means of regulating the macroeconomy. It is closely related to economic development and people's life. Tax revenue is affected by various factors. However, people hope to get quantitative conclusions through empirical analysis on how much each element affects tax growth. Taxation plays an important role in the economic operation and stable development. Through empirical analysis, the main influencing factors of local taxation are obtained, which is of great practical significance for national macroeconomic regulation and control, redistribution of national income, rational allocation of resources, and industrial development.

Many scholars have made valuable explorations on the influencing factors of tax revenue. Fang [1] studied the aspects of gross domestic product, total consumption, total import and export. Aiming at the problems of multicollinearity and autocorrelation, the model is modified to establish a regression model that affects national tax revenue. The factors that have a great influence on the national tax revenue are *GDP*, fiscal expenditure and total retail sales of social consumer goods. Wang [2] introduced the MIMIC model and used the structural equation model to examine the factors that affect the efficiency of personal income tax. It is concluded that the level of education, urbanization level, collection and management technology investment, tax service level, inspection level and so on affect the efficiency of the declaration system. Among them, collection and management technology investment, economic development level and education level are the most influential indicators. Yi [6] and Wang [7] discussed the relationship between taxation and environmental pollution. It is found that there is a positive correlation between tax scale and per capita environmental pollution emissions. The increase in tax scale is not conducive to the reduction of per capita environmental pollution emissions. There is a certain relationship between tax revenue and environmental pollution. Chen [12] used the linear data measurement model to analyze the regression model of tax revenue from five indicators such as *GDP*, fiscal expenditure and commodity retail price index. It is concluded that the main factors affecting national tax revenue are *GDP*, fiscal expenditure

and commodity retail price index. Liang [9] studied the relationship between digital inclusive financial development and local taxation by using the panel data of the digital inclusive financial index and local taxation measured by Peking University from 2011 to 2018. Through quantile regression research, it is found that digital inclusive finance can significantly promote local tax growth. And digital inclusive finance has little impact on local taxation at low quantiles and weak statistical significance. At high quantiles, inclusive finance has a great impact on local taxation and strong statistical significance.

The above research shows that the factors affecting tax revenue are multifaceted, and they all affirm the importance of economic factors, but focus on exploring the economic factors affecting tax revenue, without a comprehensive analysis of the factors affecting tax revenue. Therefore, to reveal the intrinsic relationship between taxation and the economy more profoundly, and to give full play to the functions of taxation in organizing income, regulating the economy and regulating the distribution, this paper intends to conduct a more in-depth discussion on the specific application of the index system method in statistics in the field of tax analysis based on the research results of relevant scholars. This paper constructs a model of the regional tax influencing factor index system, selects R^2 and *MSE* as two indicators for test sets, and analyzes tax contribution based on *RBF*. To provide theoretical guidance for the formulation and improvement of tax policies, and use policies to adjust the structure of economic development, to allocate social resources more reasonably.

2. Construction of an Index System of Regional Tax Influencing Factors

2.1. Local economy

The local economy (X_1) is an important factor affecting local taxation and the basis for the existence and development of taxation. Gross regional product (*GDP*) refers to the outcome of the production activities of all resident units in the region over a while. Local consumption is the source of many taxes. Total retail sales of social consumer goods (*TRS*) refers to the total amount of consumer goods sold directly in wholesale and retail industries, reflecting social purchasing power and related to value-added tax. The consumer price index (*CPI*) is the reaction of price changes and economic operations. The total fixed asset investment (*TFA*) of the whole society is expressed in monetary terms, and the workload of building and purchasing fixed asset activities. The total national economy will increase with the increase in fixed investment, which will lead to an increase in tax revenue. Foreign trade is one of the " troika " that drives national economic growth. It can not only drive economic growth, but also drive tax growth.

In terms of the local economy, this paper selects regional *GDP* (X_{11}), regional total retail sales of social consumer goods (X_{12}), consumer price index (X_{13}), local total investment in fixed assets (X_{14}), total import and export of foreign-invested enterprises (X_{15}) to measure the impact of the local economy on tax.

2.2. Population

The relationship between taxation and population is very close. The quantity and quality of population (X_2) have an important impact on the total amount and structure of taxation. The resident population refers to the population who often lives in a certain area. It is an important source of personal income tax and an important manifestation of regional development. The natural growth rate of the resident population refers to the ratio of the natural increase of the population to the average total population in a certain period. The natural population growth rate is an indicator reflecting the trend and speed of natural population growth. The urban registered unemployment rate refers to the proportion of the number of urban registered unemployed at the end of the reporting period in the total number of urban employees at the end of the period and the number of urban registered unemployed at the end of the period. The number of industrial and commercial registered private enterprises has an impact on population employment, investment scale, consumer demand and so on, which in turn affects the size of the tax base and the total personal income tax revenue.

In terms of population, this paper selects the number of permanent residents (X_{21}), the natural growth rate of permanent residents (X_{22}), the registered unemployment rate of towns (X_{23}), and the number of registered industrial and commercial households of private enterprises (X_{24}) to measure the impact of population on taxation.

2.3. Environment

Environment (X_3) is an important reference factor for population mobility and a reflection of regional development. The total investment in industrial pollutant control can reflect the degree of investment in pollution control in the region. The per capita green area reflects the livability of the regional environment, and can also reflect the local commercial model and the implementation of environmental protection tax. Population density represents the average crowding degree of a region. Reflects the local population absorptive capacity and capacity.

This paper selects the total investment in industrial pollutant control (X_{31}), population density (X_{32}), and per capita green area (X_{33}) to measure the impact of the environment on taxation.

2.4. Policy

Policy (X_4) affects the taxation environment and has an important impact on taxation. Fiscal expenditure (*PFE*), also known as public expenditure, is the total amount of control and utilization of social public resources collected by the government from the private sector to meet the public needs of society. Fiscal expenditure has a direct relationship with tax revenue, and its scale determines the growth of tax revenue, because the main source of fiscal revenue is tax revenue. At the same time, a basic criterion of taxation is to make ends meet, and the increase in fiscal expenditure needs to be met by an increase in taxation. The digital inclusive finance index is instrumental data reflecting the development status and evolution trend of digital inclusive finance. The implementation of inclusive financial policies can promote regional development and increase tax sources. Non-tax revenue is also an important aspect of the national financial system, which can affect tax revenue from a policy perspective.

This paper selects general fiscal expenditure (X_{41}), digital inclusive financial index (X_{42}) and non-tax revenue (X_{43}) to measure the impact of policies on tax revenue.

To sum up, the influencing factors index system constructed in this paper is shown in Table 1 :

Table 1. Index system of influencing factors

First level indicator	second level indicator	Indicator units	Reference
Local Economy(X_1)	Gross Regional Product(X_{11})	Billion	[1]
	Regional total retail sales of consumer goods(X_{12})	Billion	[1]
	Consumer Price Index(X_{13})	-	[1]
	Total local fixed asset investment(X_{14})	Billion	[1]
	Total import and export volume of foreign-invested enterprises(X_{15})	Million	[1]
Population(X_2)	Number of inhabitants(X_{21})	million people	[2]
	Natural population growth rate(X_{22})	-	[2]
	Urban Registered Unemployment Rate(X_{23})	-	[3]
	Number of Industrial and Commercial Registered Private Enterprises(X_{24})	household	[4]
Environment(X_3)	Total investment in industrial pollutant treatment(X_{31})	million	[5]
	Population density(X_{32})	People / hectare	[7]
	Per capita green area(X_{33})	Square meters / person	[7]
Policy(X_4)	Fiscal expenditure(X_{41})	Billion	[1]
	Digital Inclusive Finance Index(X_{42})	-	[8], [9]
	Non-tax revenue(X_{43})	Billion	[1]

3. Artificial neural network

An Artificial neural network (*ANN*) is an analog logic algorithm realized by simulating the processing of information in the human brain, which has excellent communication. Each connection is similar to a synapse between neurons for information transmission between neurons; neurons and neurons are interconnected to form a neural network to get the final feedback. Psychologist McCulloch and logician Pitts 1943 proposed a mathematical study of neural cells by simulating biology [10], called the M-P model. This model marks the birth of an artificial neural network. This paper adopts *RBF* (radial basis function) neural network structure for research.

RBF neural network is a three-layer forward network composed of an input layer, hidden layer, and output layer. The topology of the RBF neural network is shown in Figure 1.

In the hidden layer, the radial basis function is generally used to convert the input vector space into the confidential layer space, so the actual linear inseparable situation is transformed into a linear separable position. The learning process of the RBF neural network is also a supervised learning method. RBF neural network has the advantages of simple results, fast learning convergence, and approximation of any nonlinear function.

The typical *RBF* structure usually follows the rules to select the center of the basis function based on experience. As long as the distribution of the training samples can represent the given problem, the evenly distributed q centers can be selected according to the background, and the spacing is d. The width σ of the Gaussian basis function is chosen as:

$$\sigma = \frac{d}{2q} \tag{1}$$

The K-means clustering method selects the basis function and takes the center of each cluster as the basis function center. Because the output layer is a linear element, its weight can be calculated directly by the least square method.

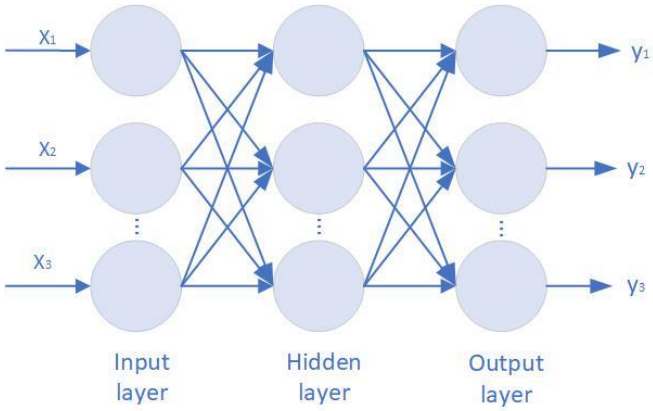

Figure 1. Topological structure of RBF neural network

The first layer of the input layer, composed of the signal source node, only plays the role of data transmission, the input information without any transformation. The second hidden layer: hidden layer neuron kernel function is a Gaussian function, the input information space mapping transformation. The third output layer responds to the input mode. The action function of the output layer neuron is linear, and the information output by the hidden layer neuron is linearly weighted and output as the output result of the entire neural network. The activation function of radial basis neural network can be expressed as:

$$R^2 = (x_p - c_i) = exp\left(-\frac{1}{2\sigma^2}\|x_P - c_i\|^2\right), j = 1,2,\ldots,n \tag{2}$$

The structure of x_p radial neural network can get the output of the network is:

$$y_j = \sum_{i=1}^{n} w_{ij}\, exp\left(-\frac{1}{2\sigma^2}\|x_P - c_i\|^2\right), j = 1,2,\ldots,n \quad (3)$$

4. Setting test indicators

This paper has two indicators of determining coefficient (R^2) and mean squared error (*MSE*) to evaluate and analyze different influencing factors. Mean squared error (*MSE*) is used to measure the degree of deviation between the predicted value and the actual value of the model [11]. The closer it is to 0, the closer the model's expected value is to the true value. The closer the coefficient of determination (R^2) value is to 1, the better the fitting effect of the model is. The calculation method of each index is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (5)$$

In the formula y_j in the measured, $\hat{y}_i$ is the model's predicted value, $\overline{y_i}$ is the average value of the model, and n is the number of test samples. From the above expression, it can be seen that the value range of MSE is $[0, +\infty)$, and the value range of R^2 is$[0,1]$. The optimal algorithm is selected by comparing the two test indexes of each influencing factor.

5. Empirical analysis

5.1. Research object selection and data collection

This paper selects North China as the research object. North China is connected to Northeast China in the north, East China and Central China in the south, Northwest China in the west and Bohai Sea in the east. The total area of the region is about 560,000 square kilometers, of which the municipal area is about 75,000 square kilometers. The administrative division of North China includes Beijing, Tianjin, Hebei, Shanxi and Inner Mongolia. It is the most important agricultural and animal husbandry production and commodity grain base in China. It is also an important industrial manufacturing center and urban agglomeration center. The economic development and industrial structure of the five provinces in North China are quite different. The diversity of regional economic development ensures the dispersion and diversity of data. This paper collects various indicators data of five provinces and cities in North China from 2011 to 2020 based on data sources such as the National Bureau of Statistics, Ruisi Database and China Economic Database. Descriptive statistics of indicators are shown in Table 2:

Table 2. Indicators Descriptive Statistics

Indicators name / unit	Min	Max	Mid	Mean
Gross regional product / billion yuan	811.250	3601.380	1604.945	1869.573
Regional Total retail sales of social consumer goods / billion yuan	287.857	1506.365	552.220	734.057
Consumer price index	93.00	106.90	102.48	102.91
Total local investment in fixed assets / billion yuan	427.660	4013.980	1185.387	1400.387
Total import and export volume of foreign-invested enterprises/million US dollar	92768.41	8014514.90	1330702.71	3229445.08
Number of inhabitants/million	134.100	746.384	243.800	337.428
Natural population growth rate /‰	0.07	6.95	3.61	3.66
Urban registered unemployment rate/%	1.21	3.80	3.51	3.11
Number of industrial and commercial registered private enterprises/million	1.327	17.336	4.161	5.912
Total investment in industrial pollutant treatment / million yuan	512.246	98753.900	24593.175	27955.490
Population density / person per square kilometer	1135.53	5016.30	3045.39	2986.18
Per capita green area / m2 per person	9.21	19.77	14.17	13.83
Fiscal expenditure / billion yuan	179.633	902.279	412.655	449.013
Digital inclusive financial index	28.89	368.54	232.65	219.15
Non-tax revenue / billion yuan	14.659	129.918	58.918	62.478

5.2. ANN calculation

Based on the SPSS software, this paper introduces the data of the above five provinces obtained by interpolation. The radial basis function algorithm is used to set the step size to 1, and the number of neurons is from 10 to 20. The ANN training is carried out by adjusting the number of hidden layer units after the iteration. Based on the predicted value and real value of tax revenue, MSE and R2 are calculated to compare the ANN results. The specific results are shown in Table 3:

Table 3. Neural network training results table

Number of hidden layers	R2	MSE
10	0.995	13826.41
11	0.996	10329.77
12	0.966	89533.88
13	0.997	6872.43
14	0.997	9862.50
15	0.979	56002.55
16	0.989	29347.95
17	0.947	138047.6
18	0.944	148714.5
19	0.961	108893.9
20	0.996	11150.11

Based on the above data results, it is found that when the number of hidden layer units is 13, the model accuracy is the highest, R2 is 0.997, MSE is 6872.43, indicating that the accuracy of the model prediction is high, and the tax contribution can be analyzed comprehensively. The model passes the convergence judgment.

5.3. Contribution analysis

According to the calculation of 5.2, when the neuron is 13, the influence and contribution of the independent variable to the dependent variable are calculated. Based on the ANN-RBF contribution analysis, the histogram is shown in Figure 2:

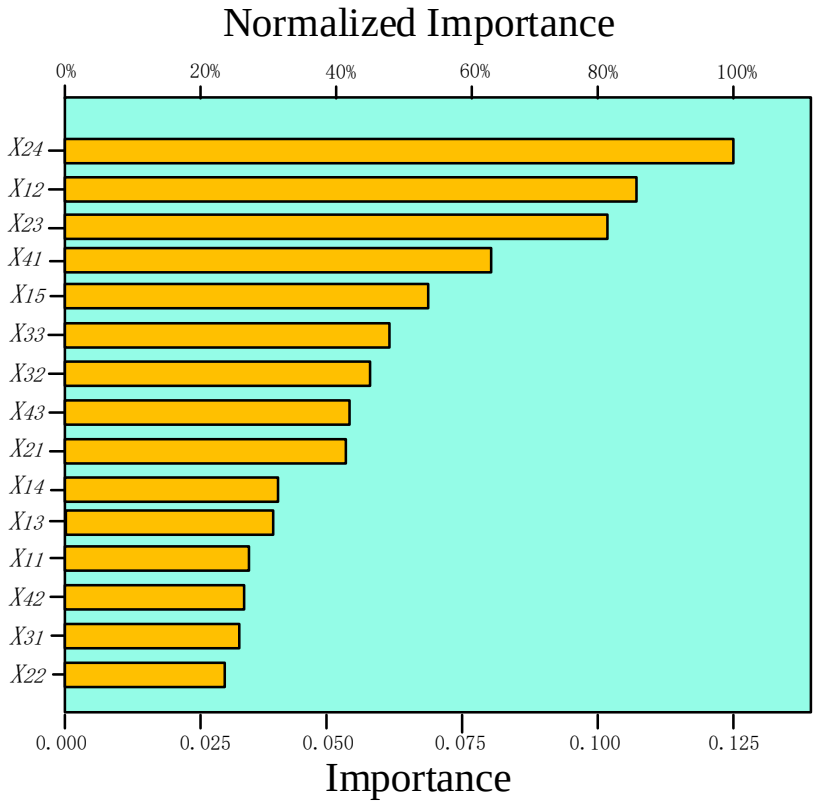

Figure 2. ANN-RBF contribution analysis results

As can be seen from the data, the main indicators that account for more than 50 % of the importance of tax contribution normalization are the number of industrial and commercial registration households of private enterprises, total retail sales of social consumer goods, urban registered unemployment rate, fiscal expenditure and total import and export volume of foreign-invested enterprises in descending order of importance. Among them, the importance of individual indicators of the number of industrial and commercial registration accounts of private enterprises exceeds 12.5 %, which is consistent with the reform of replacing business tax with value-added tax in China since 2011. This can well explain the actual situation in China, retail sales of social consumer goods, production sales and total imports, should be the formation of tax revenue ' troika'.

According to China 's 2020 tax source data, VAT, corporate income tax and consumption tax revenue are the top three of the total revenue, accounting for 37.2 %, 23.8 % and 9.7 % respectively. VAT is mainly collected directly from enterprises and consumers, mainly reflected by the number of industrial and commercial registration households of private enterprises and the total retail sales of social consumer goods, while corporate income tax and consumption tax are directly linked to the two. The urban unemployment rate represents the development level of surplus labor and urban dual structure to a certain extent, so it also has a certain impact on tax revenue. Fiscal revenue and expenditure is a dynamic balance process in theory, so the change of fiscal expenditure will lead to the change of tax revenue under the relatively complete transfer payment system. China as the world 's second largest foreign direct investment after the United States, foreign investment can create a large number of jobs and production and consumption, foreign-invested enterprises import and export volume has an important impact on tax revenue is very reasonable.

6. Conclusion

The *ANN* algorithm is constructed on this basis by making the index system of regional tax influencing factors, and the two indexes of R^2 and *MSE* are selected for the test. MSE and R^2 are obtained to test the ANN results. It is found that when the number of hidden layer units is 13, the

model has the highest accuracy, and the accuracy of model prediction is higher. Based on *ANN-RBF*, the tax contribution is analyzed. The conclusions are as follows:

(1) The number of registered households of private enterprises for business and industry, total retail sales of social consumer goods, urban registered unemployment rate, fiscal expenditure, and total import and export of foreign-invested enterprises have a regular contribution of over 50% importance, indicating that retail sales of social consumer goods, sales of production materials and total imports are essential factors in forming tax revenue.

(2) The number of registered private enterprises and total retail sales of social consumer goods are essential components of tax revenue; the urban unemployment rate has a specific influence on tax revenue as it represents, to a certain extent, the level of development of the surplus labor force and the dual structure of the towns; fiscal expenditure and revenue will also lead to changes in tax revenue due to changes in fiscal expenditure under China's relatively complete transfer payment system; the input of foreign investment can create a large number of The investment of foreign capital can create a large number of jobs and stimulate production and consumption. Its total exports have a significant impact on tax revenues. It is therefore recommended to promote production and consumption, to promote the employment of surplus labor, to ensure the sustainability of fiscal expenditure, and to continue to attract foreign investment.

References

[1] Fang Hongli. Analysis on Influencing Factors of Tax Revenue in China [J]. Economic Research Guide, 2022 (07): 120-122 + 147.

[2] Wang Xiaojia. Research on the personal income tax declaration system in China [D]. Northeast University of Finance and Economics, 2021.

[3] Han Xiulan, Zhao Nan. Study on the Impact of Population Aging on Individual Income Tax Revenue in China [J]. Tax Research, 2019 (03): 34 - 39.

[4] Hausman J A, Ruud P. Family labor supply with taxes [J]. 1984.

[5] Zhu Jialiang. Empirical Study on the Relationship between Urbanization and Finance [J]. Urban Development Research, 2014, 21 (09): 5 - 11.

[6] Yi Longsheng, Zeng Xiangang, Chen Songling. From the combination of environmental law and tax system to levy environmental tax: not only for environmental decompression [J]. Environmental economy, 2008 (08): 51 - 56.

[7] Wang Juan, Wang Weiyu. An Empirical Study on Taxation and Environmental Pollution - From the Perspective of Environmental Federalism [J]. Tax Research, 2016 (04): 50 - 54.

[8] Guo Feng, Wang Jingyi, Wang Fang, Kong Tao, Zhang Xun, Cheng Zhiyun. Measuring the development of digital inclusive finance in China: index compilation and spatial characteristics [J]. Economics (Quarterly), 2020,19 (04): 1401 - 1418.

[9] Liang Xiaoqin. Empirical Research on the Impact of Digital Inclusive Finance on Local Taxation [J]. Audit and Economic Research, 2020, 35 (05): 96 - 104.

[10] Zhao Chongwen. Summary of artificial neural networks [J]. Shanxi Electronic Technology, 2020 (03): 94 - 96.

[11] Li Jiazheng, Xu Xiaona, Zhang Huayong. Study on growth prediction model of Betula platyphylla in northern Hebei [J]. Journal of Inner Mongolia University (Natural Science Edition), 2021, 52 (03): 257 - 263.

[12] Chen Ran. Empirical Analysis of Economic Factors Affecting Tax Revenue in China [J]. Journal of Southwest Jiaotong University (Social Sciences), 2011, 12 (03): 64 - 66.