Public Budget Revenue Model Based on SVR and Neural Network-A Case Study of Shanxi Province

Li Fang, Yuan Liu, Qing Zhao, Jing Yang, and Luoyifan Zhong(✉)

School of Computer and Information Engineering, Qilu Institute of Technology, Jinan, China
xiaobubing@qlit.edu.cn

Abstract. This paper analyzes a variety of factors that affect public budget revenue. The data set is the Shanxi Provincial Statistical Yearbook collected from 1999 to 2021 years. This study uses the lasso regression method to select multiple factors affecting public budget revenue. The key influencing factors were retained. Then the GM (1,1) model was used to predict the data of each influencing factor in 2022 and 2023. Finally, SVR and neural network were used to forecast the public budget revenue in 2022 and 2023 and the advantages and disadvantages of the two models were compared. The results show that the general economic pattern of Shanxi province will be positive in the future. The public budget revenue of Shanxi Province will further increase.

Keywords: public budget revenue · lasso · SVR · BP · GM (1,1)

1 Introduction

Local public budget revenue is an important source of state revenue also reflects the development of local economy. Rational arrangement of fiscal expenditure is of great significance to local economic development and layout. Therefore, effective forecasting of public budget revenue is of great research value.

The choice of influencing factors has a significant impact on the prediction accuracy of the model. Most researchers in past studies [1–7] selected a number of factors as influencing factors of public budget revenue, including the number of employees in employment, total wages, total retail sales of consumer goods, fixed asset investment, resident consumption index, gross regional product, primary industry, secondary industry, and tertiary industry. In Li Min's study on the impact of fiscal revenue in Gansu Province [1], the largest number of factors was as many as 39 and the smallest number of factors was 5.

There is a problem of multicollinearity among the influencing factors. This problem needs to be solved before building the model. Min Li [1] used the commonly used lasso, adaptive lasso and SCAD methods for variable selection and compared the screening results of the three methods, analyzing the advantages of the lasso method in terms of screening results. Sheng, Yifu Sheng and Jianjun Zhang [2] used lasso regression to

© The Author(s) 2023
N. Akhtar et al. (Eds.): PMIS 2023, AHIS 8, pp. 872–883, 2023.
https://doi.org/10.2991/978-94-6463-200-2_91

screen multiple variables. Peng Qin [3] used minimum angle regression to solve the adaptive Lasso estimation, which eliminates some covariates and variables with small effects. Peiyu Liu [4] used lasso and random forest for variable screening, respectively. Yang He [5] demonstrated that lasso regression has great advantages in the variable selection process.

On the construction of the research forecasting model, Peng Qin [3] used a BP neural network model to forecast fiscal revenue in Anhui province. Peiyu Liu [4] used a combined BP and RBF neural network model to predict the fiscal revenue of Hunan province. Jingjing Ren and Shangbin Gao [6] used GM(1,1) to predict each influencing factor and built a fiscal revenue forecasting model using SVR. Tong Chen and Xiaohui Zhou [7] used a four-layer neural network model for forecasting and achieved good results. Hui Yu [8] chose the discriminative model DMLP in DNN.

Based on the above study. In this paper, we use GM(1,1) model combined with SVR and BP neural network for forecasting and compare the forecasting results, respectively. The results show that the prediction effect of the model combining GM (1,1) model and BP neural network is better than that of GM (1,1) combined with SVR. The prediction results provide data support for the public budget revenue of Shanxi Province.

2 Theoretical Basis

2.1 Theory of Lasso Regression

Lasso regression [9] is a compressed estimation method. A more refined model can be obtained by constructing a penalty function. Some regression coefficients are compressed to zero when the sum of the absolute values of the variables to be evaluated is less than some fixed value. This part of the variables is eliminated and the purpose of screening variables is never achieved.

The lasso parameter estimates are defined as shown in Eq. 1.

$$\hat{\beta}(lasso) = \arg\min_{\beta} \left\| y - \sum_{i=1}^{p} x_i \beta_i \right\| 2 + \lambda \sum_{i=1}^{p} |\beta_i| \tag{1}$$

The parameter λ controls the complexity of the regression. The larger the λ, the larger the penalty to obtain a model with fewer variables.

2.2 Gm (1,1)

Gray system theory [10] was proposed by Prof. Ju-Long Deng in 1982. The model is suitable for solving uncertainty problems. The gray model is applicable when the problem data is small and the sample data carries incomplete information. Gray system theory explores the inner laws by digging deeper into the valid information carried by the small sample information data.

The general form of the GM (1,1) model.

Let the variable $x^{(0)} = \left\{x^{(0)}(i), i = 1, 2 \cdots, n\right\}$ be a non-negative monotone original data series, and the steps to build a gray prediction model are as follows.

Firstly, perform one accumulation on $x^{(0)}$ to get one accumulation sequence $x^{(1)} = \left\{x^{(1)}(k), k = 1, 2 \cdots, n\right\}$

The following first-order linear differential equation can be established for $x^{(1)}$, as shown in Eq. 2, which is the GM (1,1) model. Where t denotes the data series, α denotes the development coefficient, and μ denotes the amount of gray action.

$$\frac{dx^{(1)}}{dt} + \alpha x^{(1)} = \mu \tag{2}$$

The differential equation is solved to obtain the prediction model as in Eq. 3.

$$\hat{x}^{(1)}(k+1) = \left[\hat{x}^{(0)}(1) - \frac{\widehat{\mu}}{\widehat{\alpha}}\right] e^{-\widehat{\alpha k}} + \frac{\widehat{\mu}}{\widehat{\alpha}} \tag{3}$$

Since the GM (1,1) model obtains a single cumulative quantity, the data $\hat{x}^{(1)}(k+1)$ obtained from the GM(1,1) model is reduced to $\hat{x}^{(0)}(k+1)$ by cumulative reduction, i.e., the gray prediction model for $x^{(0)}$ is shown in Eq. 4.

$$\hat{x}^{(0)}(k+1) = (1 - e^{\widehat{\alpha}})\left[\hat{x}^{(0)}(1) - \frac{\widehat{\mu}}{\widehat{\alpha}}\right] e^{-\widehat{\alpha k}} \tag{4}$$

2.3 SVR

SVR is an application of SVM to regression problems [11]. An 'interval band' is created on both sides of the linear function with an interval of ε. No loss is calculated for all samples in the interval band. Only the support vector affects its function mode. Finally, the optimization model is derived by minimizing the total loss and maximizing the interval. The equation for SVR is shown in Eq. 5.

$$\min_{w,b} \frac{1}{2}\|W\|_2 + C\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{L}_{\varepsilon}(\mathrm{f}(\mathrm{x_i}) - \mathrm{Y_i}) \tag{5}$$

where W and b denote regressors, $\mathrm{C} \geq 0$, C is the penalty coefficient, and L_{ε} is the loss function.

2.4 Neural Networks

BP neural networks [12] are multi-layer feed-forward networks trained by error back propagation. The model is divided into two propagation phases: forward and backward. The basic idea is gradient descent. The connection weights are continuously adjusted according to the principle of making the loss function fall the fastest. The BP neural network model is a multi-layer perceptron model structure that contains not only input and output points, but also one or more hidden layers. Figure 1 shows a three-layer BP network. It consists of an input layer, an implicit layer and an output layer.

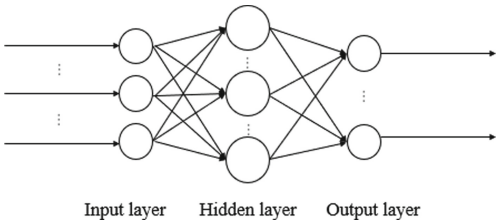

Fig. 1. BP neural network diagram.

3 Data Processing and Analysis of Findings

3.1 Data Sources and Significance

In this paper, various fiscal data of Shanxi Province from 1999–2021 were selected. 15 influencing factors affecting public budget revenue were identified based on previous studies. The data used in this paper were obtained from the Shanxi Provincial Statistical Yearbook [13]. Some data were obtained from the National Bureau of Statistics. The meanings of the variables are shown in Table 1.

Table 1. Meaning of the influencing factors.

Variable	Meaning
x1	Number of people employed in society
x2	Average wage of employed persons in urban non-private sector
x3	Total retail sales of social consumer goods
x4	Per capita disposable income of urban residents
x5	Per capita consumption expenditure of urban residents
x6	Total population at the end of the year
x7	Total social fixed asset investment
x8	Gross regional product
x9	Primary Industry Output
x10	Tax revenue
x11	Consumer Price Index
x12	Output value of secondary industry
x13	Output value of tertiary industry
x14	Per capita disposable income of the people in the province
x15	Per capita consumption expenditure of the people in the province
y	General public budget revenue

Table 2. Descriptive statistics for each variable.

Variable	MIN	MAX	MEAN	STD
x1	1392.4	1855	1655.13	171.54
x2	6065	84938	37015.3	24881.57
x3	587.1	7747.3	3698.33	2610.47
x4	4337	37433	17894.74	10700
x5	3516	21965	11582.13	6087.33
x6	3203.63	3574.11	3429.94	111.05
x7	575.4	14285	5697.01	4447.37
x8	16671000	2.26E + 08	91518452	59347387
x9	1599600	12868711	5292948	3173455
x10	925219	20947237	8032620	6132185
x11	98.4	107.2	102.07	1.94
x12	7854700	1.12E + 08	46477502	27804121
x13	7216700	1.01E + 08	39748002	29875724
x14	2622	27426	12211.48	7954.83
x15	1857	17191	8217.74	5033.79
y	1091785	28344743	11353336	8572658

3.2 Research Process

3.2.1 Data Description

The data are statistically described using minimum, maximum, mean and standard deviation. Table 2 shows that the overall public budget revenue in Shanxi Province has increased steadily, but the standard deviation is 857.2658, which is a large difference. Therefore, the relationship between variables and public budget cannot be found by data description only.

3.2.2 Correlation Analysis

Correlation analysis is an analysis of two or more related elements to measure the proximity between two characteristic elements. In this paper, Pearson's correlation coefficient, which is commonly used in statistics, was chosen for the analysis. Pearson's correlation coefficient measures the interrelationship between two characteristics X and Y. It is usually expressed using r or p and takes values in the range of [-1,1].

Using the code prepared by python, the Pearson correlation coefficients for variables x1-x15 and variable y were viewed and the results are shown in the Fig. 2.

A darker color means a stronger correlation between the variables. The lighter the color, the weaker the correlation between the variables. The graph shows that the correlation between x11 and y is very small. The variable x11 will be removed later for

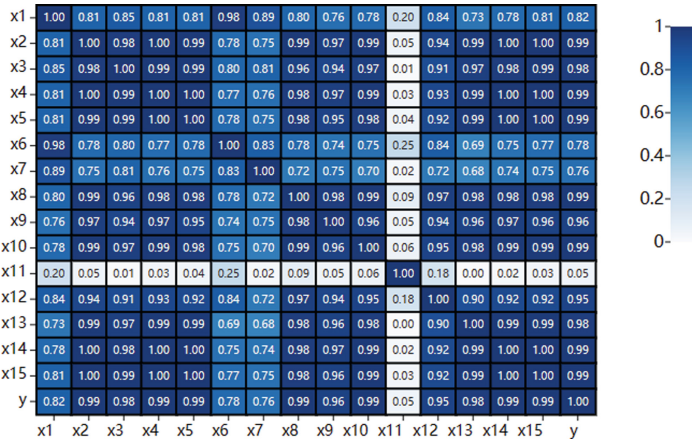

Fig. 2. Pearson correlation coefficient heat map.

Table 3. Coefficients of the variables after Lasso regression.

x1	x2	x3	x4	x5	x6	x7	x8
115.9492	0.854	0.0247	0.0169	-0.2012	-0.195	-1.5603	0.0193
x9	x10	x11	x12	x13	x14	x15	
0	-0.0001	0.0002	-0.2347	0	0	0.3202	

predictive accuracy. The variables x2, x3, x4, x5, x8, x9, x10, x13, x14, x15 and y all show a strong correlation.

3.2.3 Variable Screening

The correlation test can reveal the problem of multi-col-linearity among multiple variables. To ensure the accuracy of the prediction requires screening out the variables that meet the requirements among the original variables. Through research, we found that lasso regression is suitable for small sample data. It can compress the coefficients of certain variables to zero and will not eliminate certain variables.

The minimum angle regression algorithm is selected for the regression. The compression coefficient is encoded by python. The coefficient results of each variable after compression are shown in Table 3.

The coefficients of variables x9, x13 and x14 are compressed to zero, so these three variables can be removed, and x11 is artificially removed as it has a very low correlation with y. The remaining 11 variables were used as the basis for prediction.

3.2.4 Building the Model

Based on the excellent performance of grey prediction for small sample data, the individual influences were first predicted using a GM(1,1) model. The code prepared using

Table 4. Predicted values of each variable for GM(1,1).

Variable	x1	x2	x3	x4
2022	1916.47	106558.1	11159.79	45460.9
2023	1940.36	117330.3	12266.51	49564.41
Variable	x5	x6	x7	x8
2022	26239.5	3587.12	13237.58	249126959.5
2023	28304.81	3600.27	14139.26	273156953.5
Variable	x10	x12	x15	
2022	27132419	109549700.6	21365.49	
2023	30230164	118166517.3	23332.97	

python can obtain the predicted values of each influence factor in 2022 and 2023, based on which the SVR and BP neural network models are used to forecast the fiscal budget values respectively.

3.2.5 Comparison of Results

1. Prediction using SVR.

The LinearSVR class in the sklearn library is used for SVR prediction. The prediction results are shown in Table 5, where y is the true value and y_pred is the predicted value. Figure 3 shows that the SVR prediction is highly accurate and the curve between the true and predicted values is well fitted.

2. Using BP neural network model.

Due to the small sample size. The BP neural network in this paper consists of 2 hidden layers and uses "relu" as the activation function. The 11 filtered variable values from 1999–2021 are used as the training data for the neural network. After training, the model is used to predict the data for 2022 and 2023. The prediction results are shown in Table 6. Figure 3 and Fig. 4 show the prediction results of the SVR and BP neural network models, respectively. From the figures, it can be concluded that the BP neural network model outperforms the SVR. The error value curves of the SVR and BP neural network are given in Fig. 5. The error of the BP neural network model is significantly smaller.

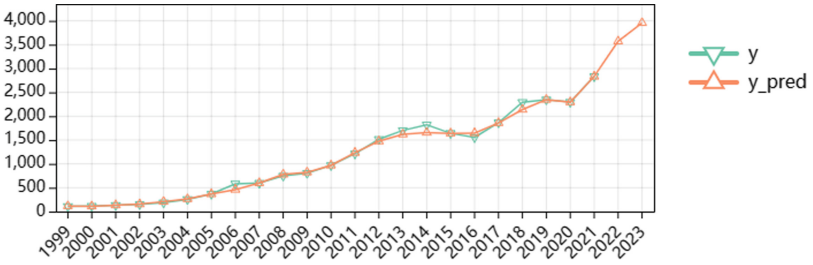

Fig. 3. Comparison of true and predicted values of SVR forecasts.

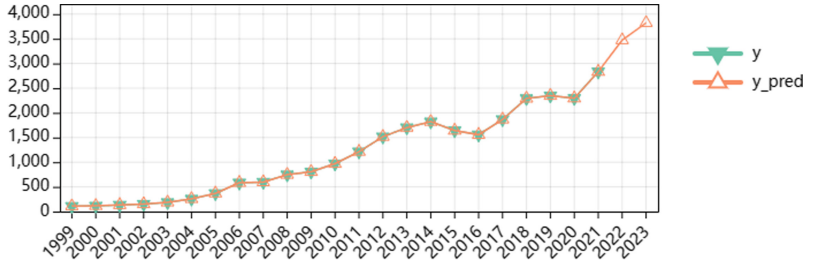

Fig. 4. Comparison of true and predicted values of BP neural networks.

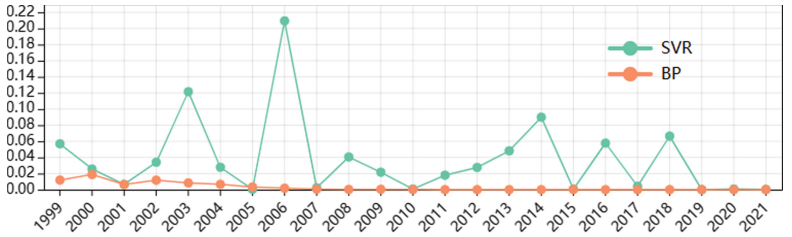

Fig. 5. Comparison of error values for the two models.

Table 5. Predicted SVR values.

Year	y	y_pred
1999	109.1785	115.3896
2000	114.4762	111.5406
2001	132.7618	133.637
2002	150.8245	155.9544
2003	186.0547	208.6675

(*continued*)

Table 5. (*continued*)

Year	y	y_pred
2004	256.3634	263.4924
2005	368.3437	368.2072
2006	583.3752	461.2044
2007	597.887	599.2461
2008	748.0047	778.3623
2009	805.8279	823.2801
2010	969.6652	968.9297
2011	1213.434	1235.231
2012	1516.378	1474.534
2013	1701.623	1619.627
2014	1820.635	1656.899
2015	1642.355	1641.106
2016	1556.997	1646.932
2017	1867.002	1859.084
2018	2292.698	2140.304
2019	2347.748	2347.748
2020	2296.567	2294.656
2021	2834.474	2833.737
2022		3571.139
2023		3957.45

Table 6. BP neural network predicted values.

Year	y	y_pred
1999	109.1785	110.48
2000	114.4762	116.6283
2001	132.7618	133.5994
2002	150.8245	152.6
2003	186.0547	187.6166
2004	256.3634	258.0676
2005	368.3437	369.5729
2006	583.3752	584.4279

(*continued*)

Table 6. (*continued*)

Year	y	y_pred
2007	597.887	598.2795
2008	748.0047	748.1863
2009	805.8279	806.2085
2010	969.6652	969.8847
2011	1213.434	1213.476
2012	1516.378	1516.363
2013	1701.623	1701.545
2014	1820.635	1820.582
2015	1642.355	1642.188
2016	1556.997	1556.906
2017	1867.002	1866.95
2018	2292.698	2292.624
2019	2347.748	2347.487
2020	2296.567	2296.449
2021	2834.474	2834.314
2022		3473.552
2023		3822.904

4 Conclusion

This paper investigates the public budget revenue model for Shanxi Province. Lasso regression method was used to screen the variables. GM (1,1) model used to forecast each influencing variable. Based on this, the public budget revenues for 2022 and 2023 were forecasted using SVR and BP neural network, respectively. The results show that the BP neural network model outperforms the SVR, and its forecasting effect is closer to the real value. The predicted data show that the public budget revenue of Shanxi Province can reach RMB 347,355.2 billion and RMB 382,290.4 billion in 2022 and 2023, respectively. Overall the revenue still maintains an upward trend. It has a greater increase than 2021.

References

1. Li M. (2019). Analysis of factors influencing fiscal revenue and fiscal revenue forecast in Gansu Province. Shandong University, MA thesis.https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201902&filename=1019055381.nh
2. Sheng Yf, Zhang Jj, Tan Ww, Wu J, Lin Hj, Sun G and Guo P. (2021). Application of grey model and neural network in financial revenue forecast, *Computers, Materials & Continua,* vol. 69, no. 3, pp. 4043-4059, https://doi.org/https://doi.org/10.1155/2022/7817264

3. Qin P. (2018) Forecasting and analysis of financial revenue in Wuhan based on data mining technology. Huazhong University of Science and Technology, MA thesis. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201901&filename=1018786130.nh
4. Liu Py. (2021) Local revenue forecasting based on grey Markov and RBF neural network models.Xiangtan University, MA thesis. DOI:https://doi.org/10.27426/d.cnki.gxtdu.2021.000388.
5. He Y. (2018) Impact analysis and forecasting of fiscal revenue in Yunnan Province based on variable selection and neural network. Yunnan University, MA thesis. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202001&filename=1018247087.nh
6. Ren Jj, Gao Sb. (2022) " SVR-based local fiscal revenue forecasting model for Lvliang city." *Information Technology and Informatization* .01:46–49. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iJTKGjg9uTdeTsOI_ra5_XZTsGUa7GDnFpYBKZ6JOc5-UEfcYR5HZO69CXCUnve25&uniplatform=NZKPT
7. Chen T, Zhou Xh. (2019) " A deep perceptron prediction model based on BP neural network." *Computers and Digital Engineering* 47.12:2978–2981+3009. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7i8oRR1PAr7RxjuAJk4dHXot2Zh5Mrmm6_ik1FJlPCQqGUY8DvMpKmabABxnz96M_m&uniplatform=NZKPT
8. Yu H. (2018) " Fiscal revenue forecasting model based on grey deep perceptron." Computers and Digital Engineering 46.01(2018):25–29. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7i0-kJR0HYBJ80QN9L51zrPwIx_MpO-tO2plEyYl99kgdk564ONS8VYmLmWZsC02K3&uniplatform=NZKPT
9. Qin Q. (2021) Analysis and Forecasting of Fiscal Revenue Influencing Factors in Hunan Province - Based on Python Software Implementation. China Market .29(2021):40-41. doi:https://doi.org/10.13939/j.cnki.zgsc.2021.29.040.
10. Deng Hl, Zhang Lj. (2016) An integrated grey and regression model for local fiscal revenue forecasting method. China Management Informatization 19.05 (2016): 145–148. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7ijP0rjQD-AVm8oHBO0FTadqcD9FxTNCrr2JMEie0pjJLz0NEu73U9vCtLtNpdaF5r&uniplatform=NZKPT
11. Luo Ll, He Pf. (2009) Application of regression support vector machine in fiscal forecasting. Science Association Forum (Second Half) .05(2009):154–155. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKgchrJ08w1e75TZJapvoLK17Z3E4AOb2gRMcrpENfsjLbAXD6Zqp_ym5XnZjWdxAKODJPaNn-Mda&uniplatform=NZKPT
12. Wang Q. (2021). Analysis of general public budget revenue forecasting based on grey forecasting and BP neural network - taking Wuxi city as an example. *China Township Enterprise Accounting*.12(2021):12–14. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iy_Rpms2pqwbFRRUtoUImHSFbczw5Fy-zOi3ohThJkyA0Uo47PanYgEtI_tDfXZAB&uniplatform=NZKPT
13. Shanxi Provincial Statistical Yearbook. Beijing: China Statistics Press, 2021, http://tjj.shanxi.gov.cn/tjsj/tjnj/

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.