Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8028754, 12 pages
https://doi.org/10.1155/2022/8028754

Research Article

Methodology and Application of Fiscal and Tax Forecasting Analysis Based on Multi-Source Big Data Fusion

Lin Zhu

School of Business, Sias University, Zhengzhou 451150, China

Correspondence should be addressed to Lin Zhu; zhulin@sias.edu.cn

Received 24 April 2022; Revised 24 May 2022; Accepted 30 May 2022; Published 24 June 2022

Academic Editor: Wen-Tsao Pan

With the advent of the big data era, the use of computers has spread to all walks of life, and the finance and taxation industry is also in the middle of it. The current taxation system is huge and complex, and different tax types are inevitably linked to different economic indicators at a deep level, so tax forecasting requires personalised forecasting analysis for different tax types. This paper selects several tax types that account for a large proportion of tax revenue for prediction analysis, respectively, and conducts fusion research on multi-source big data, including business tax, corporate income tax, and personal income tax. Based on the multi-source big data fusion method, the prediction research on fiscal taxation tax types is conducted, and experiments are conducted with the taxation data of Beijing from 1995 to 2020 to predict the three tax types from 2017 to 2020. The results show that the deviation of the forecast data from the real tax data is small, controlling the forecast deviation to within 14%, indicating the effectiveness of the proposed method.

1. Introduction

Tax forecasting is the first step in budgeting not only for the Chinese government but also for other governments around the world and is an important tool for government financial management. A large number of relevant studies have been conducted in developed countries around the world, and a relatively mature tax forecasting system has been developed. The tax risk management system established by the Organisation for Economic Cooperation and Development [1] is to analyse the many factors affecting taxation and the impact of each factor on taxation, so as to best manage taxation in advance. The core idea of risk management is also the forecasting of taxes, predicting future taxes through current factors.

The US tax system has undergone hundreds of years of modification and refinement and is now very well established [2]. The US is also the first country to practice tax revenue forecasting. Tax forecasting has become a prerequisite for the US Congress to discuss changes in tax legislation and fiscal spending arrangements and plays a key role in ensuring the stability of government tax and spending policies [3]. Some areas of the US use time series to forecast tax revenues and provide a basis for changes to the fiscal system. For example, New Jersey's tax budget is based on the results of seasonally adjusted time series analysis, using seasonally indexed smoothing and seasonally adjusted ARIMA models [4, 5]. The application of such models eliminates the non-stationarity of the time series and provides a relatively high degree of predictive accuracy. They also use other models for prediction, such as neural networks and support vector machine models, but there are relatively few practical applications of these algorithms.

The Government of Canada has developed a large set of macroeconomic analysis models using estimated parameters and predictive analysis methods. The models contain hundreds of equations and the parameters are estimated using OLS estimation [6] and LS estimation [7] error correction models, which consist mainly of economic models and fiscal models. The tax forecasting model allows for forecasting and analysis of labour tax [8], corporate income tax, and personal income tax. The Spanish National Tax Administration has developed a forecasting function between taxes and GDP, based on macroeconomics [9]. By

inputting a variety of economic indicators into the forecasting model, the finance department can obtain tax forecast data, while using tax return data, actual tax revenue data, and net income from taxes for tax forecasting and in-depth analysis of the economic factors affecting tax revenue [10].

The Italian tax authorities [11] have developed more than 500 macroeconomic econometric forecasting models. The economic indicators and forecasting models can be selected according to the type of tax, such as the relationship between operating surplus and corporate income tax and the relationship between employee income and personal income tax. The models take into account not only conventional economic factors but also unexpected factors such as policy effects and natural disasters, which are adjusted to eliminate such effects when making forecasts [12]. The model uses an adaptive dynamic model, which takes into account new influences during the forecasting process and keeps the model in a constant state of transformation to obtain the most accurate forecasts.

Sun uses a linear model to fit multi-source fusion data [13] and tests the model on a small sample medical data set show that the method helps improve the accuracy of linear regression model estimates. The accuracy of the coefficients of the linear regression model is improved and a test for fungibility is proposed. Due to the problems of distribution bias and non-uniform measurement standards in datasets from different sources, it is not possible to integrate all the data directly [14]. A number of algorithms have been proposed in existing research to address distribution bias, but in the adaptive literature, the focus has been more on using specific algorithms to address discrepancies between datasets from different sources.

Some scholars have also used graphical model methods to propose hypothesis testing methods for the existence of distribution bias in different datasets that such data can be fused, and they have also considered the classification problem of fused datasets using support vector machine methods in conjunction with Alzheimer's disease datasets [15]. In recent years, with the continuous improvement of data collection techniques, research related to the problem of large-scale multi-source data fusion has become a major hotspot. When each data source stores data with a large sample size, processing data from all sensors faces high computational costs as well as storage costs. In this paper, we propose a subsampling approach to the problem of fusing large samples of multi-source data, where only a portion of the total data with high representativeness can be extracted instead of a large-scale full sample of data. Zheng [16] presented the statistical theory of leverage score importance sampling and illustrated the feasibility and effectiveness of the method through theoretical analysis and extensive random simulation studies.

China's research on tax revenue forecasting began in the 1980s and can be divided into the following three stages. The first stage is to count the data and present them using some simple charts and graphs [17, 18], focusing on the operational analysis of the correlation between tax revenue and various economic indicators. In the second stage, theoretical methods of econometrics were introduced, and with the help of scientific theories, some simple modelling algorithms were used one after another. Regression analysis began to be applied to the study of tax revenue forecasting. In the third stage, from the late 1990s onwards, more and more analytical models were applied to fiscal and tax revenue forecasting [19], and more attention was paid to the research of model algorithms. The forecasting accuracy of the models became higher and higher, and the forecasting results were correspondingly more scientific and accurate.

Taxes are the main source of revenue in China, and each year the government formulates a tax plan based on the economic development of the previous year [20]. Tax forecasts are the basis for the government to carry out the next fiscal budget. In order to develop a scientific tax plan, local governments have started to actively work on tax forecasting. The economic indicators chosen in this paper include GDP, added value of tertiary industry, fixed asset investment, industrial input, investment in real estate development, total import and export, industrial added value, industrial electricity consumption, total retail sales of social consumer goods, sales area of real estate development, and total consumer price index. The economic indicators related to different tax types also vary, and they need to be analysed operationally and verified using technical tools. The text selects several taxes that account for a large share of tax revenue for predictive analysis, namely, business tax, corporate income tax, and personal income tax, and uses a multi-source big data fusion approach as the rationale for the predictive study.

2. Requirement Analysis

2.1. Business Requirements. At the national level, fiscal revenue has an important impact on the operation of the national economy and the stable development of society. The first step in how to monitor fiscal revenue is to be clear about the sources of fiscal revenue so that it can be monitored at source. Previously, forecasts of fiscal revenue were mainly judged by experience, but due to China's fluctuating economic development, economic development is guided by new policies every year [21]. Taxes are generated in economic development and are inevitably linked to economic indicators, so after clarifying the relationship between taxes and economic indicators, tax forecasts can be made through economic indicators.

Taxes that account for a large proportion of tax revenue in the tax system include business tax, corporate income tax, and personal income tax [22]. As can be seen from the chart analysing the proportion of taxes in local taxation in Beijing in 2020, business tax accounts for 31%, while corporate income tax and personal income tax account for 10% and 28% of the analysis. These three major taxes account for close to 70% of the revenue, so it is important to make forecasts for these three taxes. The specific distribution of each tax category in 2020 for the Beijing Local Taxation Bureau is given in Table 1.

The share of these different taxes in the total value of all taxes is shown in Figure 1.

Different taxes require separate forecasting [23]. There are two starting points for forecasting taxes; firstly, the laws

Table 1: Beijing Local Taxation Bureau 2020 tax data information.

Category	Tax credits
Business tax	10679195
Corporate income tax	3339135
Personal income tax	9587911
Urban construction tax	1893719
Property tax	192524
Stamp duty	613264
Land value added tax	2456040
Vehicle and vessel use tax	277198
Deed tax	2056854

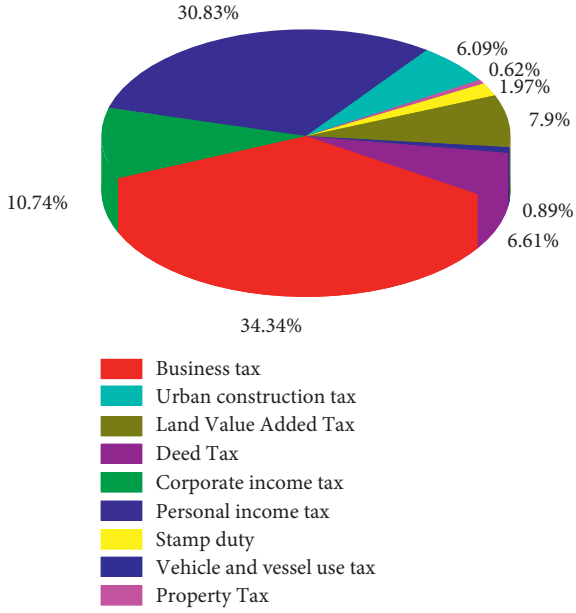

Figure 1: Map of different taxes as a proportion of the total value of all taxes.

of change in the operation of the tax itself should be explored, but only if the current environment is relatively stable, such that the tax will inevitably operate according to the laws. In fact, China's economic policies are adjusted every year, and although the overall economic operating environment is relatively stable, adjustments will be made locally. This paper forecasts business tax, corporate income tax, and personal income tax separately based on the fusion of multiple sources of big data. Because any one forecasting method has its own advantages and disadvantages, a comprehensive forecasting analysis using multiple forecasting models can filter out the models with relatively high forecasting accuracy. In order to obtain prepared forecasting results, the data collection environment is particularly important, and the datasets themselves are obtained through government-type websites such as the Beijing Municipal Bureau of Statistics and the National Bureau of Statistics.

2.2. Functional Requirements. On the basis of meeting business requirements, the application platform functions in this paper include two major functional modules, namely, regression model management and forecasting data management. Regression model management is the use of trained models for tax forecasting, i.e., the output models can be used to make separate forecasting applications, such as linear regression. Prediction data management can manage the tax data predicted by the regression model, calculate the error between the real data and the predicted data, and remove the tax data with large prediction errors, so that the regression model can reforecast the data.

2.3. Understanding of Requirements

2.3.1. Sales Tax. By analyzing the relationship between business tax and related economic indicators, it is conducive to accurate prediction and analysis of business tax. With the relevant economic indicators identified, annual forecasting analysis is applied separately to multi-source data fusion methods and the forecasting results are evaluated to determine the most scientific model and to be able to use the training model for the next stage of tax forecasting.

The tax subjects of business tax [24] are units and individuals engaged in taxable tax and transfer of intangible assets and sale of immovable property, and their taxation is based on turnover, transfer amount, and sales, and the scope of levy involves the construction industry and the tertiary industry. Based on business experience, the following economic indicators are tentatively determined, social fixed asset investment, GDP, total retail sales of social consumer goods, and sales area of commercial properties. Firstly, the correlation between these economic indicators and the business tax is carried out from the data, and the collinearity between these economic indicators is analysed to filter out the economic indicators that have a more obvious impact on the business tax.

2.3.2. Corporate Income Tax. The study analyses the relationship between the corporate income tax and the economic indicators associated with the corporate income tax that affect it and provides a forecast analysis of the corporate income tax. Corporate income tax [25] is a tax levied on the income from the production and operation of enterprises and other production units. It is levied on the income of taxpayers and ranges from sales income, labour income, property income, interest income, etc.

The following economic indicators are determined based on business experience: GDP, total industrial output value above scale, added value of tertiary industry, area of commercial properties sold, total import and export, and RMB loan balance of financial institutions. The correlation between these economic indicators and corporate income tax was first conducted from the data, and the collinearity between these economic indicators was analysed to filter out the economic indicators that have a more obvious impact on corporate income tax. Forecast analysis of the corporate income tax is carried out and the forecast results are evaluated to determine the most scientific model. At the same time, a monthly time series forecasting analysis of corporate

income tax is conducted to explore the quarterly pattern of corporate income tax.

2.3.3. Personal Income Tax. Personal income tax is a tax levied by the taxing authority on the legal income of natural persons in the country. The forecasting analysis of personal income tax requires firstly the study and analysis of the relationship between personal income tax and the economic indicators that affect personal income tax. The following economic indicators, GDP and per capita disposable income of urban residents, were determined based on business experience. The correlation between these economic indicators and personal income tax is first analysed in terms of data, as well as the covariance between the economic indicators, so as to filter out the economic indicators that have a greater impact on personal income tax and finally establish the relationship between personal income tax and the economic indicators that affect personal income tax in relation to them.

3. Model Algorithms

3.1. Multi-Source Data Fungibility Test. The question of whether the fusion of multiple sources of data helps to improve the testing of coefficient estimates of linear regression models is considered here. The single-source data linear regression model can be expressed as

$$Y = X\beta^* + \varepsilon, \tag{1}$$

where Y is the n dimensional response vector, X is the $n \times p$ dimensional design matrix, β^* is the p dimensional vector of true coefficients, and ε is the n dimensional noise variable and is assumed to be $\varepsilon \sim N(0, \delta^2 I)$. The least squares estimate of β^* is

$$\widehat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2. \tag{2}$$

Assume that there is k data source, $X_i (i = 1, 2, \ldots, k)$ denotes the design matrix for the ith data source, and Y_i denotes the corresponding response vector. If it is assumed that the potential relationships between the predictor and response variables are the same across sources, it is feasible to estimate the coefficient estimates for each data source using the same $\widehat{\beta}$. Removing this assumption, the difference between the coefficient vector for each source and the shared coefficient vector for all sources can be denoted as $\Delta\beta_i = \beta_i - \beta^*$ where $i = 1, 2, \ldots, k$. Then, for each source i, we have

$$Y_i = X_i\beta^* + X_i\Delta\beta_i + \varepsilon_i. \tag{3}$$

Assumption 1. $\varepsilon_i \sim N(0, \sigma_i^2 I)$; in the multi-source data case, the weighted least squares estimate is

$$\widehat{\beta} = \arg\min_{\beta} \sum_{k}^{i=1} \tau_i^2 Y_i - X_i\beta_2^2, \tag{4}$$

where τ_i is the weighting factor for each source.

To test whether the estimation of model coefficients for a particular data node can be improved with information from the data of other nodes, the common mean square error is used as a measure of merit. Without loss of generality, source node 1 is taken as the reference and $\tau_1 = 1$ is set; let $\beta^* = \beta_1$ in equation (4), at which point equation (4) becomes

$$\widehat{\beta} = \arg\min_{\beta} \|Y_1 - X_1\beta\|_2^2 + \sum_{k}^{i=2} \tau_i^2 \|Y_i - X_i\beta\|_2^2. \tag{5}$$

When $\tau_i = \sigma_1/\sigma_i$, the error of $\widehat{\beta}$ is minimised.

3.2. Multi-Source Data Regression Methods for Large Sample Scenarios. From equation (2), if X column is full rank, the ordinary least squares estimate $\widehat{\beta}_{\text{ols}}$ for β is

$$\widehat{\beta}_{\text{ols}} = \left(X^T X\right)^{-1} X^T Y. \tag{6}$$

The predicted value of the response vector $\widehat{Y} = X(X^T X)^{-1} X^T Y$ can be obtained from equation (6). Order $H = X(X^T X)^{-1} X^T$ represents the hat matrix, the diagonal element $h_{ii} (i = 1, 2, \ldots, n)$ of which is the leverage score of the ith sample point, which reflects the degree of influence of the ith sample on the predicted value. By performing a singular value decomposition on X, H can also be expressed as $H = UU^T$, where U is a $n \times p$ dimensional column orthogonal matrix whose columns are X left singular vectors. Thus, the leverage value score for the ith sample point can also be expressed as

$$h_{ii} = u_i^2, \tag{7}$$

where u_i is the ith row of matrix U. When the sample size n is large, the computational cost of the full sample OLS estimate will become very high. Therefore, a viable alternative is to resort to subsampling methods and then use subsamples to calculate estimates of the coefficients. The steps of the subsampling algorithm are as follows.

(1) Construct sampling probability $\{\pi_i\}_{i=1}^n$ for all original sample points and draw $r (r \ll n)$ samples from the full sample (X, Y) according to this probability distribution, denoted as (X^*, Y^*), and construct the corresponding sampling probability matrix $\Phi^* = \text{diag}\{\pi_k^*\}_{k=1}^r$. Uniform sampling: make $\pi_i = 1/n$; leveraged sampling: $\pi_i = h_{ii}/(\sum_n^{j=1} h_{ii})$.

(2) Solving the ordinary least squares model on a subsample yields the unweighted subsample estimate $\widetilde{\beta}$.

$$\widetilde{\beta} = \arg\min_{\beta} Y^* - X^*\beta^2. \tag{8}$$

3.3. Subsampling Multi-Source Big Data Regression Algorithms. In the analysis above, before integrating the k source data together, the value of the weight parameter needs to be set such that the variance of $\widehat{\beta}$ is minimised when $\tau_i = \sigma_1/\sigma_i$. Similarly, the weight parameter was set to $\tau_i = \sigma_1^*/\sigma_i^*$, where σ_i^* is the sample standard deviation of the subsample after sampling from source i.

3.4. Shared Partial Coefficient Algorithm

Step 1. Let $\{\pi_i\}_{i=1}^{n}$ be the sampling probability distribution and draw $r\,(r \ll n)$ subsamples by rows from all data from each of the k sources, with subsamples noted as $(X_i^*, y_i^*)_{i=1}^{k}$. Fuse $\tau_1(X_1^*, y_1^*), \ldots, \tau_k(X_k^*, y_k^*)$ into a new dataset (X^*, Z^*, y^*).

Step 2. Use the fused data to calculate the least squares estimate $\widetilde{\beta} = \text{argmin}_\beta \|y^* - X^*\beta - Z^*\gamma\|^2$.

4. Forecasting Methodology Research and Analysis

This section focuses on the full process of researching the forecasting and analysis methodology for different taxes, with forecasting and analysis of business tax, corporate income tax, and personal income tax, respectively.

4.1. Data Collection. Collecting data is the first step in conducting predictive analysis of the data, and in order to collect data of high quality, the source of all identified data in this paper is mainly relevant official websites. The data were divided into a training set and a validation set, with the actual range of the training set being 1995 to 2017 and the data from 2018 to 2020 serving as the validation set. Only the influence of the weight of the independent variables on the dependent variable is discussed in the data mining process, and the influence of units is not considered here.

For analysis purposes, the data need to be split into a training set and a validation set, with the actual range of the training set being 1995 to 2017 and the data from 2018 to 2020 serving as the validation set. The tax and economic indicators are not in the same units, some are in RMB billion, some are in USD, and some are in million squared. However, in the data mining process, only the influence of the weight of the independent variable on the dependent variable is discussed, so the influence of the units is not considered here.

In this paper, three data training sets are used: the annual dataset of business tax, the annual dataset of corporate income tax, and the annual dataset of personal income tax. At the same time, the validation set for the relevant tax types should be prepared. According to the conventional allocation ratio, the ratio between the training set and the validation set is 7 : 3, and due to the limited amount of data collected for the whole data, only three records for each tax type are selected as the validation set. The training sets for sales tax and related indicators are shown in Table 2.

For the years 1995–2017, a detailed comparison of the data for each category in the sales tax is shown in Figure 2. As can be seen visually in Figure 2, the size of GDP of the various taxes during the period 1995–2017 far exceeds the size of the other taxes, increasing at an increasing rate as the years progress. Sales tax accounts for the lowest share and has been at the bottom end of all taxes during the period.

The validation set for sales tax and related indicators is shown in Table 3.

The annual data for the corporate income tax and related indicator training set are shown in Table 4.

A comparison of the data for each subcategory of the corporate income tax and related indicator training set is shown in Figure 3. The sizes of the five corporate income tax amounts are compared in Figure 3 using radar images, where the area containing the larger area corresponds to the larger tax amount, giving a visual indication of the share of each subcategory in the total amount.

The annual data for the corporate income tax and related indicator data validation set are shown in Table 5.

A comparison of the data for each subcategory of sales tax and corporate income tax is given in Figure 4. The graph shows the share of each subcategory of sales tax and corporate income tax, respectively. The larger the area of each coloured area in the graph, the greater the share of its corresponding tax in the total tax revenue.

The annual data for the individual income tax and related indicator training set are shown in Table 6.

The data pairs for the 3 subclasses in the individual income tax prediction training set table are shown in Figure 5. Figure 5 shows that both GDP and disposable income have increased in their subcategories compared to personal income tax, and the relationship shows an equal incremental relationship as the year increases.

The annual data for the individual income tax and related indicator validation set are shown in Table 7.

A comparison chart of the three subcategories of the annual data for the personal income tax and related indicator validation set is shown in Figure 6.

As can be seen from Figure 6, the tax on disposable income far exceeds the other two subcategories in 2018–2020, with GDP being the second most taxed and the personal income tax subcategory the least taxed, showing that disposable income has the greatest impact on the overall tax.

Fiscal tax forecasting analysis was carried out based on multi-source big data fusion under the three tax scenarios, using a subsampling multi-source big data regression algorithm to integrate the three sources of data together for forecasting. The forecast results for the three tax categories for 2018–2020 are shown in Tables 8–10.

A comparison of the data information in these three tables is shown in Figure 7. The relationship between the true and predicted values for the three tax categories is visually represented in Figure 7, where the closer the lines are for the two categories, the more accurate the prediction is and the worse the prediction is for the two categories.

As can be seen from the data information above, the forecast results for business tax are better compared to those for corporate income tax and personal income tax, with a

Table 2: Sales tax training set.

Year	Business tax	GDP	Fixed assets	Social consumption	Tertiary sector
1995	127.95	500.80	179.20	345.10	194.38
1996	130.78	598.90	193.20	363.23	261.53
1997	135.86	709.10	256.30	386.54	314.52
1998	152.36	886.20	351.20	400.12	395.17
1999	165.24	956.32	412.02	436.23	428.63
2000	178.52	1002.56	500.32	462.35	486.32
2001	181.61	1256.32	520.32	480.23	542.36
2002	197.54	1324.21	602.35	500.35	598.65
2003	213.02	1396.58	650.78	520.65	634.52
2004	228.56	1456.23	668.23	543.63	685.96
2005	249.87	1568.25	703.65	568.35	714.52
2006	281.84	1598.63	751.23	596.3	763.23
2007	327.45	1603.21	798.23	624.56	800.88
2008	363.52	1645.58	856.23	653.23	843.21
2009	333.25	1749.52	894.52	675.85	869.85
2010	383.21	1845.98	921.20	695.32	900.23
2011	460.99	1935.23	976.23	714.52	924.32
2012	458.32	1856.32	1002.32	745.36	953.68
2013	478.62	1801.57	1050.96	778.65	978.52
2014	500.62	1905.32	1099.63	802.59	1000.25
2015	494.52	1986.52	1152.30	842.35	1053.23
2016	520.88	2000.32	1198.63	863.56	1082.52
2017	536.68	2019.65	1235.36	888.56	1100.28

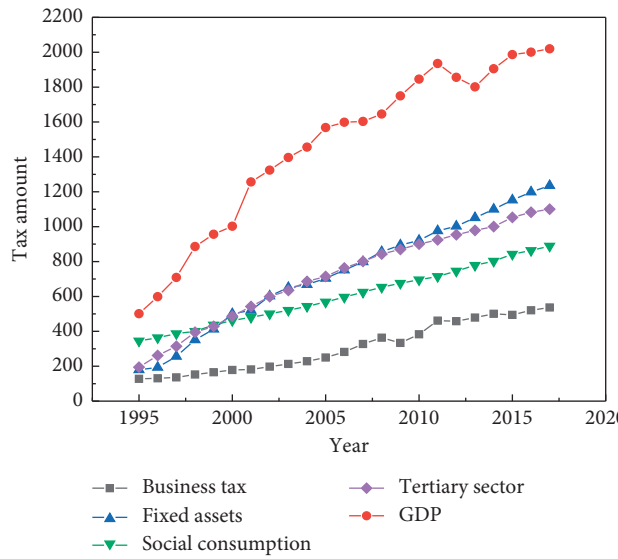

Figure 2: Visual comparison of the various categories of data in business taxation (1995–2017).

Table 3: Validation set for sales tax and related indicators.

Year	Business tax	GDP	Fixed assets	Social consumption	Tertiary sector
2018	580.95	2036.80	1279.20	925.10	1194.38
2019	615.78	2068.90	1320.20	965.23	1261.53
2020	645.86	2098.10	1386.30	1002.54	1314.52

Table 4: Corporate income tax and related indicator training set.

Year	Corporate income tax	GDP	Industrial output	Tertiary sector	Imports and exports
1995	122.55	325.32	456.23	194.55	236.23
1996	120.59	356.25	485.32	232.21	242.25
1997	119.50	368.52	502.32	256.68	249.56
1998	114.25	395.62	542.32	287.54	250.23
1999	121.23	421.36	563.23	300.21	256.32
2000	129.62	455.23	586.35	312.25	259.85
2001	137.85	486.98	599.87	332.25	262.30
2002	141.23	514.32	612.54	356.25	264.32
2003	141.68	523.25	635.65	378.65	271.02
2004	145.66	544.56	645.52	400.12	275.36
2005	158.89	563.23	665.23	412.23	279.62
2006	186.52	583.23	684.52	456.32	281.26
2007	200	600.23	700.23	486.56	283.56
2008	193.74	623.21	721.23	501.25	289.56
2009	205.62	624.25	745.62	523.25	293.54
2010	212.32	640.25	782.32	546.32	300.23
2011	215.23	653.21	800.21	568.59	305.23
2012	217.56	665.25	845.23	598.65	308.52
2013	218.65	672.52	870.65	602.32	311.56
2014	225.36	694.23	921.23	623.56	320.45
2015	227.56	714.23	946.32	653.23	330.25
2016	230.56	736.52	984.78	685.62	335.62
2017	241.20	750.26	1032.66	702.36	340.25

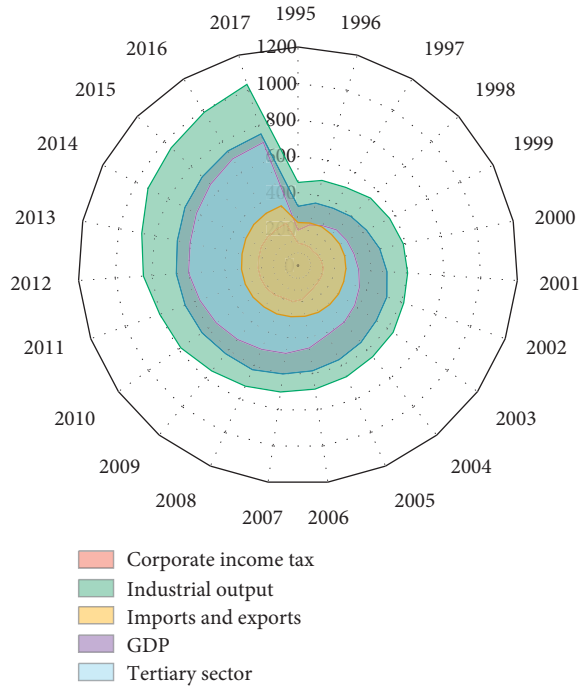

Figure 3: Comparison of data from the corporate income tax and related indicator training set.

Table 5: Corporate income tax validation set.

Year	Corporate income tax	GDP	Industrial output	Tertiary sector	Imports and exports
2018	802.12	1980.80	1579.10	1024.10	562.38
2019	915.78	2133.90	1620.80	1126.23	601.53
2020	1024.86	2198.10	1753.30	1202.54	685.52

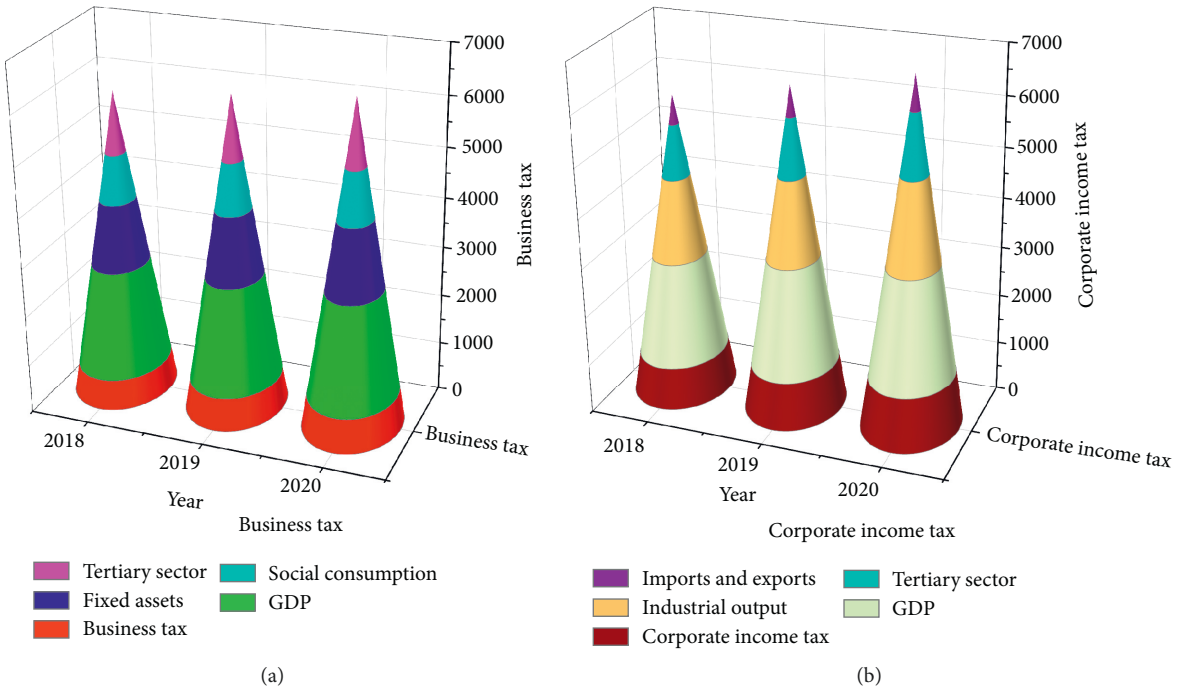

Figure 4: Comparison of data for each subcategory of business tax and corporate income tax. (a) Business tax. (b) Corporate income tax.

Table 6: Individual income tax forecast training set.

Year	Personal income tax	GDP	Disposable income
1995	102.36	500.80	302.14
1996	103.56	598.65	403.26
1997	104.25	623.25	514.58
1998	105.23	658.23	598.65
1999	116.23	700.56	654.28
2000	122.35	758.65	752.32
2001	128.36	799.23	900.23
2002	136.54	845.21	1000.25
2003	145.26	868.47	1120.23
2004	156.36	897.24	1256.00
2005	179.85	936.45	1385.25
2006	161.52	975.58	1425.33
2007	157.21	986.35	1569.78
2008	173.23	1035.26	1658.20
2009	184.56	1086.54	1799.54
2010	202.32	1126.35	1921.02
2011	265.32	1175.65	2003.33
2012	271.45	1241.25	2365.25
2013	277.65	1286.56	2653.22
2014	315.56	1352.32	3000.21
2015	372.56	1386.35	3500.21
2016	381.45	1432.52	3987.56
2017	398.78	1475.65	4533.20

lower error rate in their forecast data compared to the formal data. The forecasting results for all three tax types achieved good error rates, producing errors within acceptable limits given the large magnitude of the data, demonstrating the effectiveness of a tax forecasting methodology based on multiple sources of big data.

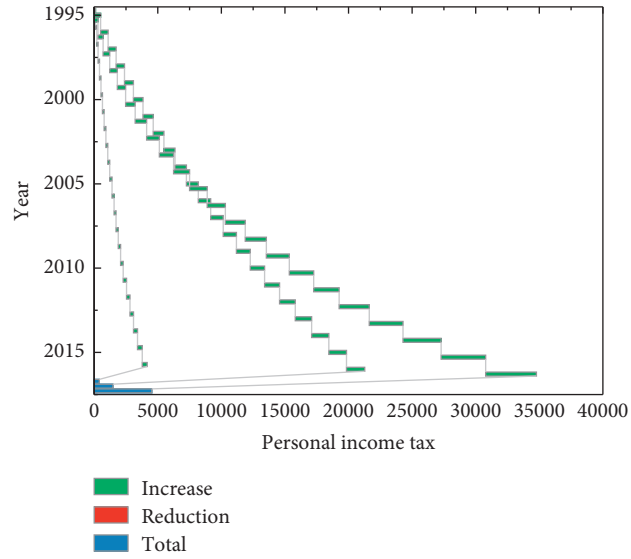

Figure 5: Comparison of the 3 subclass levels in the individual income tax prediction training set table.

Table 7: Annual data for the individual income tax and related indicator validation set.

Year	Personal income tax	GDP	Disposable income
2018	412.46	1500.70	5302.64
2019	463.56	1548.35	5603.16
2020	504.32	1653.28	5914.62

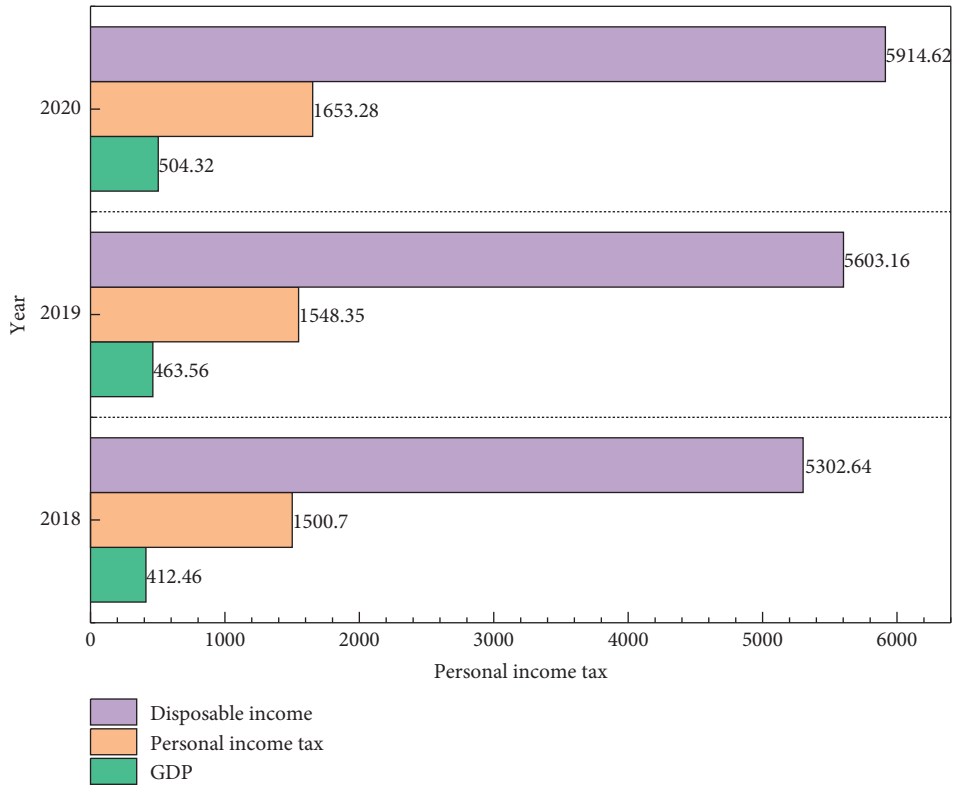

Figure 6: Comparison of the three subcategories of annual data for the individual income tax and related indicator validation set.

Table 8: 2018–2020 sales tax forecast results.

Year	Business tax	Business tax forecast	Error rate (%)
2018	580.95	598.63	3.04
2019	615.78	645.32	4.80
2020	645.86	721.32	11.68

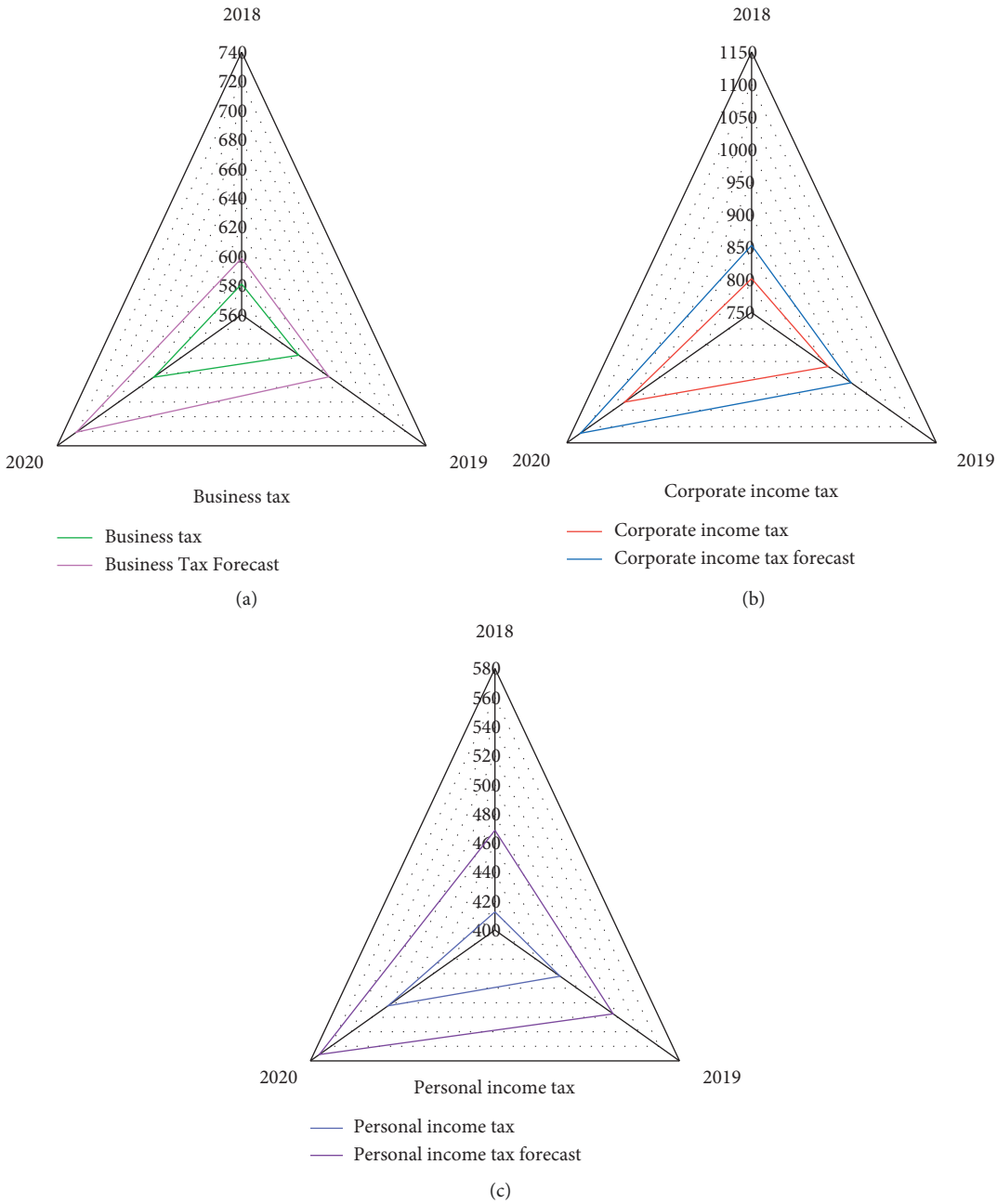

Figure 7: Comparison of three tax projection figures for 2018–2020. (a) Business tax. (b) Corporate income tax. (c) Personal income tax.

Table 9: 2018–2020 corporate income tax forecast results.

Year	Corporate income tax	Corporate income tax forecast	Error rate (%)
2018	802.12	853.23	6.40
2019	915.78	965.54	5.43
2020	1024.86	1120.36	9.32

Table 10: Individual income tax projection results for 2018–2020.

Year	Personal income tax	Personal income tax forecast	Error rate (%)
2018	412.46	468.63	13.61
2019	463.56	515.32	11.17
2020	504.32	571.32	13.29

5. Conclusion

This paper has combined statistical data and multi-source big data fusion analysis tools for tax forecasting research and evaluated the proposed subsampling method for multi-source data fusion in a big data scenario to reduce computational effort and storage costs. Compared to the method without data fusion, the estimation of the model with sampling followed by fusion is superior. In this paper, three aspects of business tax, corporate income tax, and personal income tax are predicted, respectively, and the forecast results of corporate income tax and personal income tax are compared. The whole idea of forecasting is still the principle of multi-source data fusion and the process includes six steps: business understanding, data understanding, data preparation, data modelling, model evaluation, and model release. The experimental results show that corporate income tax has the highest proportion of the three tax types, while the prediction error rate of business tax is the lowest. The impact indicators considered in this paper for tax forecasting are limited, in practice, containing more influential factors of interference, and the next work will explore more influential factors to make tax forecasting more accurate.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] J. Ebieri and D. C. Ekwueme, "Assessment of the impact of tax reforms on economic growth in Nigeria," *Journal of Accounting and Financial Management*, vol. 2, no. 2, pp. 15–28, 2016.

[2] J. U. Ihendinihu, E. Jones, and E. AmapsIbanichuka, "Assessment of the long-run equilibrium relationship between tax revenue and economic growth in Nigeria: 1986 to 2012," *The SIJ Transactions on Advances in Space Research & Earth Exploration*, vol. 2, no. 5, pp. 1–9, 2014.

[3] U. J. Charles, M. C. Ekwe, and J. U. B. Azubike, "Federally collected tax revenue and economic growth of Nigeria: a time series analysis," *Accounting & Taxation Review*, vol. 2, no. 1, pp. 24–38, 2018.

[4] W. G. Gale and A. A. Samwick, "Effects of income tax changes on economic growth," *Economic Studies at Brookings*, vol. 3, no. 2, pp. 1–16, 2014.

[5] D. E. Oriakhi and R. R. Ahuru, "The impact of tax reform on federal revenue generation in Nigeria," *Journal of Policy and Development Studies*, vol. 9, no. 1, pp. 92–108, 2014.

[6] A. C. Obasikene, "Government expenditure in Nigeria and its impact on the Nigerian economy," *Journal on Banking Financial Services & Insurance Research*, vol. 7, no. 11, pp. 1–12, 2017.

[7] D. C. Nwosu and H. O. Okafor, "Government revenue and expenditure in Nigeria: a disaggregated analysis," *Asian Economic and Financial Review*, vol. 4, no. 7, pp. 877–892, 2014.

[8] C. O. Emelogu and M. O. Uche, "An examination of the relationship between government revenue and government expenditure in Nigeria: Co-integration and causality approach," *Central Bank of Nigeria Economic and Financial Review*, vol. 48, no. 2, pp. 35–57, 2016.

[9] K. P. Saeed and S. Somaye, "Relationship between government spending and revenue: evidence from oil exporting countries," *International Journal of Economics and Management Engineering*, vol. 2, no. 2, pp. 33–35, 2012.

[10] K. Ogujiuba and T. W. Abraham, "Testing the relationship between government revenue and expenditure: evidence from Nigeria," *International Journal of Economics and Finance*, vol. 4, no. 11, pp. 172–182, 2012.

[11] Q. M. A. Hye and M. A. Jalil, "Revenue and expenditure nexus: a case study of Romania," *Romanian Journal of Fiscal Policy*, vol. 1, no. 1, pp. 22–28, 2020.

[12] R. Ali and M. Shah, "The causal relationship between government expenditure and revenue in Pakistan," *Interdisciplinary Journal of Contemporary Research in Business*, vol. 3, no. 12, pp. 323–329, 2012.

[13] W. Sun, "Research on factors affecting national fiscal revenue based on SPSS regression analysis," *China Business Theory*, vol. 2, no. 1, pp. 231-232, 2019.

[14] L. Xie, "Empirical analysis of factors affecting fiscal revenue in anhui province," *Modern Commerce and Trade Industry*, vol. 3, no. 2, pp. 115-116, 2019.

[15] S. Cai, "Analysis of influencing factors of fiscal revenue based on stepwise regression method," *Accountant*, vol. 6, no. 3, pp. 17-18, 2019.

[16] H. X. Zheng, "Analysis of influencing factors of general budget revenue in jilin province," *Guangxi Quality Supervision Guide*, vol. 5, no. 3, pp. 35–40, 2018.

[17] X. He. L. I. Xing-xu, "Analysis of the Factors Affecting the Fiscal Revenue of Yunnan," *Province. China Market*, vol. 5, no. 2, pp. 47–51, 2017.

[18] J. You, "Empirical analysis of influencing factors of China's fiscal revenue," *China Collective Economy*, pp. 100–103, 2019.

[19] E. Zhang, "Empirical analysis of factors affecting fiscal revenue in xinjiang," *Economic Forum*, vol. 3, no. 1, pp. 26–30, 2015.

[20] B. Dong, "Discussion on the Development of Township Finance and Taxation," *Finance and Economics*, Academic Edition, 2020.

[21] M. Li, *The Influencing Factors and Fiscal Revenue Forecast Analysis of Gansu Province*, Shandong University, Jinan, China, 2019.
[22] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[23] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.
[24] K. Zhang and C. T. Ng, "Adaptive LASSO regression against heteroscedastic idiosyncratic factors in the covariates," *Statistics and Its Interface*, vol. 13, no. 1, pp. 65–75, 2020.
[25] A. Rudat, J. Buder, and F. W. Hesse, "Audience design in Twitter: retweeting behavior between informational value and followers' interests," *Computers in Human Behavior*, vol. 35, pp. 132–139, 2014.