DOI: https://doi.org/10.63944/fhg01.rcs
Vol.2 No.2

Drivers of Economic Growth in the Guangdong–Hong Kong–Macao Greater Bay Area: Evidence from Ridge Regression

Lv Hao

Department of Statistics, School of Mathematics ,

Guangdong University of Education, China | lvhao@gdei.edu.cn

Abstract: This study investigates the main factors associated with economic growth in the Guangdong–Hong Kong–Macao Greater Bay Area from 2000 to 2023. Based on official statistical data, the paper constructs a multidimensional indicator system covering economic output, population and labor, science and technology input, infrastructure, openness, logistics, and the global economic environment. In the data preprocessing stage, missing values are treated through linear interpolation, and multicollinearity among explanatory variables is examined using the variance inflation factor. The results show that most explanatory variables have high VIF values, indicating serious multicollinearity in the dataset. To reduce dimensionality and improve the stability of model estimation, KMO and Bartlett tests are conducted, and principal component analysis is applied to suitable variable categories. On this basis, ridge regression is introduced to estimate the relationship between different explanatory factors and GBA GDP.

The empirical results show that industrial structure, population and labor scale, innovation input, openness, and the external economic environment are closely associated with changes in GBA GDP. In particular, the secondary and tertiary sectors show positive coefficients in the ridge regression model, indicating that manufacturing and service industries remain important supports for regional economic growth. Population and employed population also show positive associations with GDP, suggesting that labor scale and demographic agglomeration provide a basic foundation for economic expansion. R&D expenditure has a positive coefficient, reflecting the role of innovation input in promoting economic upgrading. At the same time, the results should be interpreted cautiously because the explanatory variables are highly correlated, and ridge regression coefficients mainly reflect the relative direction and stability of variable associations rather than strict causal effects. Overall, this study provides a quantitative reference for identifying key drivers of economic growth in the GBA and offers policy implications for industrial upgrading, innovation enhancement, openness expansion, and regional coordination.

Keywords: Guangdong-Hong Kong-Macao Greater Bay Area; economic forecasting; ridge regression

1. Introduction

The Guangdong–Hong Kong–Macao Greater Bay Area is one of the most economically dynamic regions in China and plays an important role in national economic development, international trade, technological innovation, and regional integration. With the continuous advancement of the Greater Bay Area strategy, the region has gradually formed a development pattern characterized by strong manufacturing capacity, expanding modern service industries, active international trade, and increasing cross-city economic connections. The Outline Development Plan for the Guangdong-Hong Kong-Macao Greater Bay Area provides the fundamental policy background and strategic framework

for understanding the regional integration, industrial coordination, innovation-driven development, and high-quality growth of the GBA. [1]Understanding the main factors associated with economic growth in the GBA is therefore important for evaluating its development mechanism and for providing quantitative support for future regional policy design.

Economic growth in the GBA is influenced by multiple dimensions, including industrial structure, population and labor supply, science and technology input, infrastructure construction, logistics capacity, openness, and the global economic environment. These factors do not operate independently. For example, the expansion of the secondary and tertiary sectors is usually accompanied by infrastructure investment, logistics development, population agglomeration, and trade growth. Similarly, R&D investment and industrial upgrading are often closely connected with the development of modern services and high-end manufacturing. Liu and Hou's study on industrial evolution and spatial reconstruction in the Tokyo Bay Area offers an important comparative reference for analyzing the transformation of bay-area economies from manufacturing concentration toward service-oriented and innovation-based spatial structures. [2] Saxenian's analysis of Silicon Valley and Route 128 highlights the importance of regional culture, industrial networks, institutional openness, and knowledge spillovers in shaping the long-term competitiveness of innovation-driven regional economies. [3] As a result, when these variables are included in the same regression model, strong correlations may exist among explanatory variables. This creates the problem of multicollinearity, which may make ordinary regression coefficients unstable and reduce the reliability of economic interpretation.

The Guangdong Economic Development Report (2024) provides recent empirical evidence and policy-oriented analysis on Guangdong's economic structure, industrial upgrading, innovation capacity, and regional development trends, offering a practical basis for interpreting the economic dynamics of the GBA.[4] To address this issue, this study uses official statistical data from 2000 to 2023 and constructs an empirical framework for analyzing the drivers of GBA economic growth. First, the study preprocesses the data by standardizing variable definitions and applying linear interpolation to deal with intermittent missing values. Second, the variance inflation factor is used to test multicollinearity among explanatory variables. The results show that many variables have VIF values far above the conventional threshold, indicating serious multicollinearity. Third, KMO and Bartlett tests are conducted for different variable categories. For categories that are suitable for dimensionality reduction, principal component analysis is used to extract representative components and reduce redundant information. Finally, ridge regression is introduced to estimate the relationship between the selected explanatory variables and GBA GDP.

Ridge regression is suitable for this study because it can reduce the instability of coefficient estimates under multicollinearity by adding a regularization term to the least squares objective function. Liu, Zhang, and Li's empirical study on the drivers of economic growth in the GBA provides direct literature support for examining how industrial structure, innovation input, openness, and factor agglomeration affect regional economic performance.[5] Mao's comparative study of world bay-area development models helps clarify the distinctive development path of the Guangdong-Hong Kong-Macao Greater Bay Area by comparing it with mature international bay areas in terms of industrial structure, governance mechanisms, and regional coordination.[6] Compared with directly removing highly correlated variables, ridge regression retains more information from the original indicator system while improving the stability of model estimation. This is particularly useful for regional economic analysis, where many macroeconomic variables are naturally correlated due to common growth trends and structural interactions. Therefore, ridge regression provides a practical method for identifying the relative direction and importance of different factors associated with GBA GDP growth.

Zhang and Li's ARIMA-based GDP forecasting study for the Yangtze River Delta provides a useful methodological reference for regional economic forecasting and illustrates the applicability of time-series models in predicting macroeconomic growth trends.[7] The contribution of this study lies in

three aspects. First, it constructs a multidimensional indicator system for analyzing GBA economic growth, covering industrial output, infrastructure, population and labor, R&D investment, trade openness, logistics, and the global economic environment. Second, it explicitly diagnoses the multicollinearity problem through VIF tests and applies KMO, Bartlett tests, and PCA to improve the structure of explanatory variables. Third, it uses ridge regression to identify the main factors associated with GBA GDP growth and provides policy implications for industrial upgrading, innovation-driven development, openness enhancement, and regional coordination.

The remainder of this paper is organized as follows. Section 2 describes the data sources, variable selection, preprocessing methods, multicollinearity diagnosis, and dimensionality reduction procedure. Section 3 introduces the construction of the ridge regression model and presents the empirical results. Section 4 summarizes the main conclusions and proposes policy recommendations for promoting high-quality economic development in the Guangdong–Hong Kong–Macao Greater Bay Area.

2. Data Preprocessing

The China Statistical Yearbook provides authoritative and continuous macroeconomic data for China, serving as a key data source for measuring economic output, population, employment, industrial structure, infrastructure, and other variables related to the GBA.[13] The Japan Statistical Yearbook provides official statistical data for Japan and offers a reliable basis for constructing comparative indicators when examining the development experience of the Tokyo Bay Area.[14] Gujarati and Porter's Basic Econometrics provides the classical econometric foundation for regression analysis, model assumptions, multicollinearity, hypothesis testing, and the interpretation of empirical relationships among economic variables.[15]

This study uses official statistical data from 2000 to 2023 for the Guangdong-Hong Kong-Macao Greater Bay Area and the Tokyo Bay Area. GBA data come mainly from Guangdong, Hong Kong, Macao, and city-level statistical sources. To ensure comparability, the study standardizes data definitions, converts currencies using annual average exchange rates, and adjusts prices to constant 2015 levels. The variables are grouped into six dimensions: economic output, population and labor, science and technology input, infrastructure, openness, and the global economic environment. For missing values, linear interpolation is applied because economic variables generally change continuously and the missing years are intermittent, allowing the time-series trend to be preserved. The formula is:

$$x_t = x_{t-1} + (x_{t+1} - x_{t-1})/2$$

Multicollinearity refers to a strong linear correlation between two or more independent variables in a dataset. This correlation makes it difficult for a model to determine the independent contribution of each explanatory variable. This paper uses the variance inflation factor (VIF) to test for multicollinearity. The mathematical expression for the VIF test is:

$$VIF_j = 1/(1 - R_j^2)$$

Here, R_j^2 is the coefficient of determination of the linear regression model in which the jth independent variable is regressed on all other independent variables. The higher the VIF value, the more difficult it is to estimate the jth independent variable independently. In economics applications, a VIF value greater than 10 is usually regarded as indicating serious multicollinearity, requiring treatment of the variable. The VIF values obtained for each variable are shown in Table 1.

Table 1 VIF values of GBA variables

Variable	VIF
GBA primary-sector output (USD trillion)	11712.059
GBA secondary-sector output (USD trillion)	73881.335
GBA tertiary-sector output (USD trillion)	52988.299
GBA infrastructure investment (RMB trillion)	29456.614
GBA transportation network length (km)	80601.610
Total global GDP (USD trillion)	78659.763
Research and development expenditure (R&D, RMB 100 million)	104.403
Population (100 million persons)	9880.0713
Employed population (100 million persons)	2283.463
Export value (USD trillion)	3541.550
Import value (USD trillion)	14883.780
Total import and export value (USD trillion)	31159.761
Total GBA logistics volume (100 million tons)	31122.526

According to the VIF results, all of the above independent variables exceed the threshold of 10, indicating the presence of multicollinearity. To address this problem, this paper later considers using ridge regression for model estimation in order to reduce the effect of multicollinearity on parameter estimates. To reduce the adverse effect of multicollinearity on the model, this paper conducts KMO and Bartlett tests for each variable category. Bartlett tests for each category show that all Bartlett test p-values are less than 0.05. Categories with KMO values greater than 0.7 are then selected for principal component analysis to reduce dimensionality. The KMO values of each category are shown in Table 2.

Table 2 KMO test results

Variable category	KMO value	Suitable for PCA
Economic variables	0.753	Yes
Openness	0.738	Yes
Logistics variables	0.745	Yes
R&D	0.500	No
Population	0.500	No

As shown above, the KMO values of the economic, logistics, and openness categories are greater than 0.7, indicating that they are suitable for principal component analysis. Principal component analysis (PCA) is a statistical method that constructs new uncorrelated variables, namely principal components, through linear combinations of original variables. The purpose is to reduce data dimensionality while retaining as much variance information from the original data as possible. For categories with KMO values greater than 0.7, namely economic variables, logistics variables, and openness variables, PCA is used to reduce each category to one dimension. After PCA dimensionality reduction, three principal component variables and four original variables jointly form seven independent variables for regression analysis, with the aim of quantifying and evaluating the influence of these variables on GBA GDP.

3. Construction of Economic Forecasting Models for the Guangdong-Hong Kong-Macao Greater Bay Area

Gao's textbook on econometric analysis and EViews modeling provides a systematic methodological foundation for regression modeling, diagnostic testing, parameter estimation, and empirical interpretation in applied economic research.[8] Li and Wang's comparative study of ARIMA, SVR, and LSTM models shows that different forecasting methods have distinct advantages in regional economic prediction, thereby supporting the need to select appropriate models according to data characteristics and research objectives.[9] Because multicollinearity exists among variables, OLS regression coefficients may become unstable, thereby affecting forecasting performance. To address this problem, this paper introduces the ridge regression model. Ridge regression stabilizes regression coefficients by adding a regularization term and improves model generalization, thereby producing more reliable forecasts under multicollinearity. Jolliffe's Principal Component Analysis establishes the theoretical and methodological basis for using PCA to reduce dimensionality, extract common information from correlated variables, and improve the structure of explanatory indicators in empirical modeling.[16] Hoerl and Kennard's seminal study on ridge regression provides the core methodological foundation for addressing multicollinearity by introducing a penalty term to stabilize coefficient estimation in regression models with highly correlated explanatory variables.[17]

Ridge regression is a linear regression method that adds a regularization term to least squares regression. The independent variables in this model are infrastructure investment, total global GDP, transportation network length, population, employed population, R&D expenditure, total logistics volume, export value, import value, total import and export value, primary-sector output, secondary-sector output, and tertiary-sector output. The objective function of ridge regression is:

$$J(\beta) = \sum_{i=1}^{m} (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

To solve the ridge regression objective function, the estimated ridge regression coefficients are first obtained by taking partial derivatives of $J(\beta)$ with respect to β_0 and β_j (j = 1, 2, 3, ..., n). The partial derivative with respect to β_0 is:

$$\frac{\partial J(\beta)}{\partial \beta_0} = -2 \sum_{i=1}^{m} (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij}) = 0$$

After simplification:

$$\sum_{i=1}^{m} y_i - m\beta_0 - \sum_{i=1}^{m} \sum_{j=1}^{n} \beta_j x_{ij} = 0$$

Thus $\beta_0 = \bar{y} - \sum_{j=1}^{n} \beta_j \bar{x}_j$, where $\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$ is the mean of the dependent variable and $\bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}$ is the mean of the jth independent variable. The partial derivative with respect to β_j is

$$\frac{\partial J(\beta)}{\partial \beta_j} = -2 \sum_{i=1}^{m} (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij}) x_{ij} + 2\lambda \beta_j = 0$$

Substituting the above expression for β_0 and simplifying the partial derivative with respect to β_j gives:

$$\sum_{i=1}^{m} (y_i - \bar{y}) x_{ij} - \sum_{j=1}^{n} \beta_j \sum_{i=1}^{m} (x_{ij} - \bar{x}_j) x_{ij} + \lambda \beta_j = 0$$

Further rearrangement gives:

$$(X-1_m\bar{x}^T)^T(Y-\bar{Y}1_m)=\beta[(X-1_m\bar{x}^T)^T(X-1_m\bar{x}^T)+\lambda I_n]$$

The ridge regression coefficient estimate is:

$$\hat{\beta}=[(X-1_m\bar{x}^T)^T(X-1_m\bar{x}^T)+\lambda I_n]^{-1}(X-1_m\bar{x}^T)^T(y-\bar{y}1_m)$$

Historical data are used to train the ridge regression model, and cross-validation is used to select the optimal regularization parameter. The trained model is then used for prediction, Substituting into the derived equation gives the coefficients β_j of each variable in the model, as shown in Table 3.

Table 3 Variable coefficients of the GBA ridge regression model

Variable	Model coefficient
Primary-sector output	-1.1732
Secondary-sector output	0.6519
Tertiary-sector output	0.7066
Infrastructure investment	-0.0237
Total global GDP	0.1745
Transportation network length	-0.000018
Population	3.6989
Employed population	3.1914
R&D expenditure	0.0007365
Total logistics volume	-0.0129
Export value	-5.7668
Import value	-2.3592
Total import and export value	3.6181

The variable coefficients in Table 3 show that the output values of the secondary and tertiary sectors both have positive coefficients, with the tertiary sector having a slightly greater effect than the secondary sector. The primary sector has a negative coefficient. Total global GDP has a positive coefficient. Population and employed population both have positive coefficients. R&D expenditure also has a positive coefficient. Total import and export value is positive, while export value and import value are negative. It should be noted that, because the variables are highly correlated, ridge regression coefficients reflect the direction of association between each factor and GDP to a certain extent, but their estimates are affected by variable definitions, standardization methods, and the regularization parameter. Therefore, they should not be directly used as strict economic interpretations. The above coefficients are more suitable as references for the relative direction and stability of each factor's influence.

Overall, the ridge regression results show that industrial structure, especially the secondary and tertiary sectors, innovation input, population and labor scale, and the external economic environment have strong associations with changes in GBA GDP. This finding provides a quantitative reference for identifying key drivers of regional economic growth, but the influence of multicollinearity among variables must be treated cautiously in economic interpretation.

Shen's comparison of the world's three major bay areas provides theoretical and practical insights for understanding how the GBA can learn from international bay-area experiences in industrial upgrading, innovation systems, openness, and regional governance.[10] The trained ridge

regression model is used to forecast GDP for the next ten years, with the future independent-variable estimates described above entered into the model. Finally, the predicted GDP values for the next ten years are plotted to visually show the future economic trend, as shown in Figure 1.

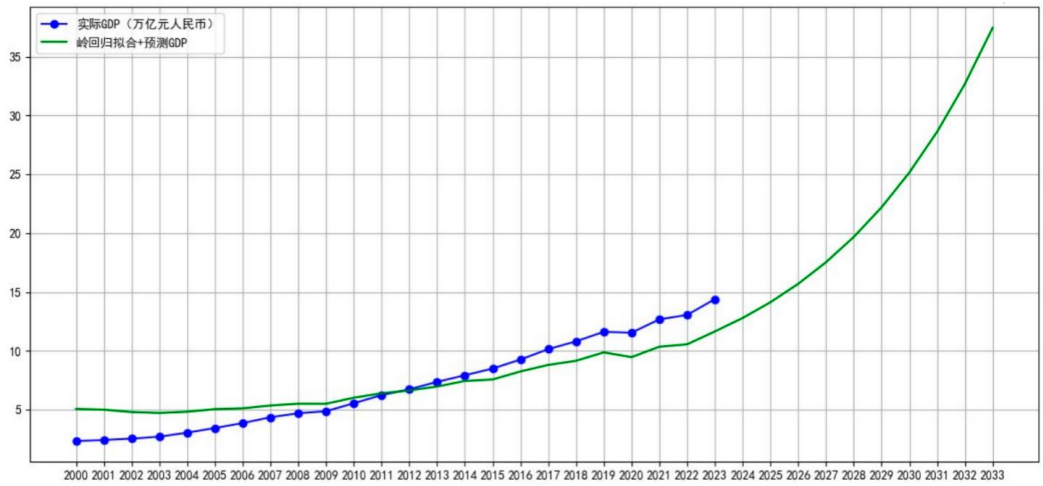

Figure.1 Actual GBA GDP and ridge regression forecast trend

As shown in Figure 1, the original data (blue solid line) and ridge regression forecast (green solid line) diverge in the next several years. The ridge regression forecast shows rapid future GDP growth and an accelerating upward trend. Fujita's study on industrial development and regional disparities in Japan provides a comparative perspective for examining how industrial concentration, spatial inequality, and regional policy interact within the Tokyo Bay Area.[11] Chen's work on advanced econometrics and Stata applications offers important technical guidance for empirical model construction, multicollinearity diagnosis, regression estimation, and robustness analysis in applied economic studies.[12] Although this trend reflects the positive effects of the variables, it does not conform to real economic growth rates. Ridge regression may overestimate the growth rate when capturing the effects of variables on economic growth. This occurs because ridge regression relies on relationships among multiple variables.

4. Conclusions

Based on ridge regression analysis, this study examines the main factors associated with economic growth in the Guangdong–Hong Kong–Macao Greater Bay Area from 2000 to 2023. By introducing ridge regression to address potential multicollinearity among explanatory variables, the paper identifies the relative importance and direction of different economic factors in explaining changes in regional GDP. The main conclusions are as follows.

First, the results show that industrial structure is one of the most important factors associated with economic growth in the GBA. The secondary and tertiary sectors both show strong explanatory power in the ridge regression model, indicating that the economic growth of the GBA is closely related to the coordinated development of manufacturing and service industries. The secondary sector continues to provide an important foundation for regional economic expansion, while the tertiary sector reflects the increasing role of modern services, finance, technology services, logistics, and the digital economy in supporting high-quality development.

Second, innovation input plays a positive role in promoting regional economic growth. The ridge regression results suggest that R&D investment is closely associated with GDP growth, showing that technological innovation has become an important driving force for economic upgrading in the GBA. This finding indicates that the region's future development should rely not only on factor input and

industrial scale expansion, but also on improvements in innovation capacity, technological transformation, and productivity enhancement.

Third, population and labor-related factors also provide important support for regional economic growth. The results indicate that labor scale and demographic factors are closely connected with the expansion of economic output. As a highly urbanized and economically active region, the GBA benefits from a large labor force, population concentration, and cross-city factor mobility. These factors help strengthen market demand, industrial agglomeration, and regional production capacity.

Fourth, the external economic environment and openness are also important factors affecting the economic performance of the GBA. The ridge regression results show that indicators related to trade and openness are associated with changes in GDP, reflecting the strong outward-oriented characteristics of the GBA economy. As an important gateway for international trade and cross-border economic cooperation, the GBA's economic growth is influenced by both domestic industrial development and external market conditions.

Overall, the empirical results show that the economic growth of the GBA is driven by multiple factors, including industrial structure, innovation input, labor scale, and openness. Ridge regression provides a suitable method for analyzing these relationships under multicollinearity among explanatory variables. The findings suggest that the future development of the GBA should focus on industrial upgrading, innovation-driven growth, factor mobility, and regional coordination, so as to promote a more sustainable and high-quality pattern of economic development. Based on the ridge regression results, the GBA should further promote high-quality economic development through industrial upgrading, innovation enhancement, openness expansion, and regional coordination. Specifically, the region should strengthen the integrated development of advanced manufacturing and modern services, increase R&D investment and improve the transformation efficiency of scientific and technological achievements, deepen institutional connectivity among Guangdong, Hong Kong, and Macao, and promote the efficient flow of labor, capital, technology, and public services across cities. Meanwhile, infrastructure development should shift from scale expansion to efficiency improvement, with greater emphasis on transportation integration, logistics optimization, digital infrastructure, and smart governance. These measures will help the GBA improve industrial competitiveness, enhance innovation capacity, release the potential of city clusters, and build a more coordinated, resilient, and sustainable regional economic system. Box, Jenkins, and Reinsel's Time Series Analysis provides the classical framework for time-series modeling and forecasting, offering methodological support for analyzing economic trends, dynamic structures, and future GDP prediction.[18] Future study of this article could combine time series with ridge regression to dig deeper how economic factors are affected by each other.

References

[1] Central Committee of the Communist Party of China and State Council. Outline Development Plan for the Guangdong-Hong Kong-Macao Greater Bay Area [Z]. 2019.

[2] Liu Yungang, Hou Lulu. Research on Industrial Evolution and Spatial Reconstruction in the Tokyo Bay Area [J]. Scientia Geographica Sinica, 2020, 40(8): 1245-1254.

[3] Saxenian A. Regional Advantage: Culture and Competition in Silicon Valley and Route 128 [M]. Cambridge: Harvard University Press, 1994.

[4] Guangdong Academy of Social Sciences. Guangdong Blue Book: Guangdong Economic Development Report (2024) [M]. Beijing: Social Sciences Academic Press, 2024.

[5] Liu Ziyan, Zhang Hua, Li Ming. An Empirical Study on the Drivers of Economic Growth in the Guangdong-Hong Kong-Macao Greater Bay Area [J]. Economic Geography, 2022, 42(5): 45-53.

[6] Mao Yanhua. Comparison of World Bay Area Development Models and the Construction of the Guangdong-Hong Kong-Macao Greater Bay Area [J]. Journal of Sun Yat-sen University (Social Science Edition), 2020, 60(3): 1-10.

[7] Zhang Hua, Li Ming. GDP Forecasting Research for the Yangtze River Delta Based on the ARIMA Model [J]. Journal of Applied Statistics and Management, 2023, 42(3): 456-465.

[8] Gao Tiemei. Econometric Analysis Methods and Modeling: EViews Applications and Examples, 4th Edition [M]. Beijing: Tsinghua University Press, 2020.

[9] Li Hong, Wang Wei. Comparative Study of Regional Economic Forecasting Models: Empirical Analysis Based on ARIMA, SVR, and LSTM [J]. Inquiry into Economic Issues, 2023, 44(2): 112-121.

[10] Shen Yong. Comparison of the Development Models of the World's Three Major Bay Areas and Implications for the Guangdong-Hong Kong-Macao Greater Bay Area [J]. Urban Development Studies, 2021, 28(5): 28-35.

[11] Fujita K. Industrial Development and Regional Disparities in Japan: A Comparative Study of Tokyo Bay Area [J]. Regional Studies, 2018, 52(4): 512-525.

[12] Chen Qiang. Advanced Econometrics and Stata Applications, 2nd Edition [M]. Beijing: Higher Education Press, 2019.

[13] National Bureau of Statistics of China. China Statistical Yearbook (2001-2024) [M]. Beijing: China Statistics Press.

[14] Statistics Bureau, Ministry of Internal Affairs and Communications of Japan. Japan Statistical Yearbook (2001-2024) [M]. Tokyo: Japan Statistical Association.

[15] Gujarati D. N., Porter D. C. Basic Econometrics, 5th Edition [M]. New York: McGraw-Hill, 2009.

[16] Jolliffe I. T. Principal Component Analysis, 2nd Edition [M]. New York: Springer, 2002.

[17] Hoerl A. E., Kennard R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems [J]. Technometrics, 1970, 12(1): 55-67.

[18] Box G. E. P., Jenkins G. M., Reinsel G. C. Time Series Analysis: Forecasting and Control, 5th Edition [M]. New York: Wiley, 2015.