Economic Forecasting for Greater Bay Area: Ridge Regression & ARIMA

JingyingGuo1,* XuefenZhong[1] and SiqiZhou[1]
1* Guangdong University of Education, Huadu District, Guangzhou 510303, Guangdong, China
*Corresponding author(s). E-mail(s): 3313544899@qq.com;
Contributing authors: 545893337@qq.com; zxfnynla@163.com;3135280539@qq.com
t These authors contributed equally to this work.

Acknowledgments

This work was supported by the 2023 Guangdong Provincial Quality Engineering Project: Financial Big Data Industry-Education-Research Practical Teaching Base at Guangdong University of Education*. I would like to express my sincere gratitude to Assistant Professor Lü Hao for his dedicated guidance, insightful advice, and continuous encouragement throughout the entire research process. His professional expertise greatly contributed to the successful completion of this study.
I am also deeply thankful to my friends Zhou Siqi and Zhong Xuefen for their valuable assistance during data collection, technical discussions, and manuscript preparation. Their support and companionship have been instrumental in overcoming various challenges encountered during the project.
With heartfelt appreciation, I extend my thanks to all individuals and institutions who have provided help and support. Their contributions have been vital to this research.

Abstract: This study explores the future economic trajectory of the Guangdong-Hong Kong-Macao Greater Bay Area by applying ridge regression and time series modeling techniques. To improve the reliability of the models and minimize multicollinearity issues, principal component analysis is employed to identify the most influential variables. The results suggest a stable and sustained economic growth trend for the region. A comparative analysis with Tokyo Bay reveals that while Tokyo demonstrates strengths in trade and the service sector, the Greater Bay Area shows greater potential in infrastructure expansion and technological advancement. Based on these findings, the study proposes forward-looking policy initiatives such as the development of a “digital twin” economic framework, a “talent free trade zone,” dedicated “green innovation zones,” and a “health economy corridor.” These recommendations aim to enhance the region’s economic resilience and global competitiveness. The research contributes valuable insights to support long-term planning and sustainable development in the Greater Bay Area.
Keywords: Kmeans and Bartlett test,principal component analysis,Vif inspection,Ridge regression model,ARIMA time series

1 Introduction

The Guangdong-Hong Kong-Macao Greater Bay Area stands as one of China's most vibrant and internationally connected economic hubs.[3] Comprising nine cities—Guangzhou, Shenzhen, Zhuhai, Foshan, Huizhou, Dongguan, Zhongshan, Jiangmen, and Zhaoqing—alongside the two Special Administrative Regions of Hong Kong and Macao, this region plays a vital role in driving national innovation and economic growth.[22] Currently, its gross domestic product (GDP) accounts for nearly one-ninth of the country's total output.

To strengthen its position in the global landscape, the Greater Bay Area is implementing proactive strategies to support emerging and future industries.[24] These include planning industrial development paths, enhancing policy incentives, and cultivating highly skilled talent. The region is also investing in the research and development of cutting-edge fields such as 6G networks, quantum technologies, life sciences, and humanoid robotics.

Leveraging its comparative advantages—including international financial services, a robust talent pool, strong foundations in basic research, and world-class manufacturing—the region is accelerating innovation and market transformation. In this study, we collect and analyze economic data from both the Guangdong-Hong Kong-Macao Greater Bay Area and the Tokyo Bay Area spanning 2000 to 2023. The objective is to extract key insights and provide strategic recommendations to help the Greater Bay Area maintain a competitive edge in future global industries.

2 Research Questions

To build a robust model for forecasting regional economic trends, this study collected extensive data from various official statistical platforms, including those of the Guangdong-Hong Kong-Macao Greater Bay Area and the Tokyo Bay Area, covering the years 2000 to 2023. Through comprehensive data screening and comparison, we aim to address the following core research questions:

Task 1:

Conduct an in-depth analysis of the historical development of the Greater Bay Area, focusing on its openness to global markets, population trends, trade performance, scientific research capacity, infrastructure development, and its position within the global economy. From this, identify and rank the most critical factors influencing economic growth, and determine which will be most impactful over the next 5 to 10 years.

Task 2:

Using the key variables identified in Task 1, forecast the economic development trend of the Greater Bay Area for the next 5 to 10 years. Based on these projections, propose practical and strategic policy recommendations that can support sustained and high-quality growth.

Task 3:

Apply the same modeling techniques to economic data from other bay areas—particularly the Tokyo Bay Area—for comparative analysis. The goal is to understand different development models, identify commonalities and divergences, and assess the relative strengths of each region.

Task 4:

Recognizing the critical role of the Greater Bay Area in China's national economy, this task focuses on using mathematical and statistical models to uncover the driving forces behind its economic growth.[17] It aims to forecast long-term trends and offer policy insights that can guide regional planning and enhance economic resilience.

3 Preliminaries

3.1 Model assumptions

To ensure the applicability and reliability of the model, this study establishes the following reasonable assumptions based on historical and contextual economic data:

It is assumed that the key economic indicators within the Greater Bay Area will exhibit consistent and stable growth over the next 5 to 10 years, without being significantly disrupted by unforeseen events such as major policy shifts or global crises.

The global economic environment is presumed to remain relatively stable, with no drastic fluctuations in global GDP or international trade. This allows variables related to global openness and international markets to remain meaningful for forecasting purposes.

The relationship between dependent and independent variables, such as GDP versus openness and employment levels, is assumed to follow a linear pattern as modeled in ridge regression.

The time series model is based on the assumption that historical economic trends will persist into the future, allowing past data patterns to inform forward-looking projections

Symbolic representation

symbol	meaning
σ^2	The variance of the residual
Y	Dependent variables (such as GDP, total output and other economic indicators)
$X_1, X_2 \ldots . X_n$	Independent variables, which represent the factors affecting economic development, such as population and R&D input
β_I	Regression coefficients of the independent variables
β_0	The intercept term (constant term) of the regression model
λ	Regularization parameters for ridge regression
α	The intercept term of the regression model
ε	Random error term
t	A time point in a time series
R^2	Determination coefficient, used to measure the fit of the model
ρ	The autocorrelation coefficient of a time series measures the temporal dependence of the data
φ_p	Autocorrelation coefficient in AR(p) model
θ_q	MA (q) model moving average coefficient
p, d, q	The class parameters of ARIMA model represent the autoregressive, difference and moving average terms respectively

3.2 Data and variables

3.2.1 Data sources

The data used in this research are drawn from official statistical databases and cover the years 2000 to 2023. These datasets include economic, demographic, technological, and infrastructure-related information for both the Guangdong-Hong Kong-Macao Greater Bay Area and the Tokyo Bay Area. The data collection process prioritized accuracy and consistency to ensure the validity of the analysis.

3.2.2 Variable definition

According to the economic development characteristics of the Greater Bay Area, we classify the variables into the following categories:

· Degree of openness to the outside world: export value, import value and total import and export value.

· Population factors: total population, employed population

· Economic factors: total value of primary, secondary and tertiary industries

· Investment in scientific research: R&D expenditure (R&D)

· Logistics: infrastructure investment, length of transportation network, total logistics

· Global environment: total global GDP

3.3 Missing value processing

A line graph of the data will be obtained to find that there are breakpoints in the image, that is, there are missing values in the data. As shown in Figure1:

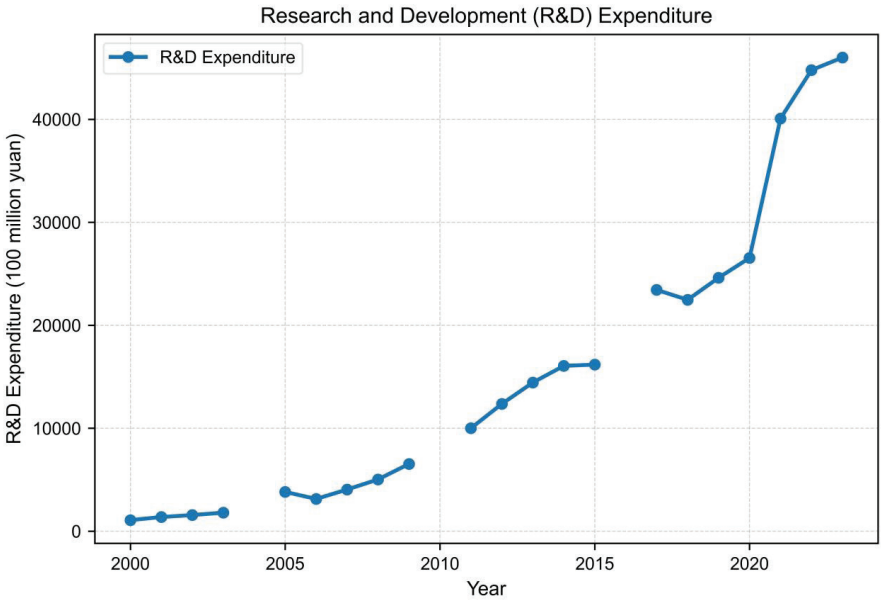

Fig.1 Research and Development (R&D) (100 million yuan)

To address this issue, this paper employs the mean, median, mode, and interpolation methods to fill in missing values. It also creates line graphs of the data obtained from different methods to observe the trends in the data images. By comparing these images with real economic development and current socio-economic conditions, the interpolation method is identified as the optimal approach.[9] The resulting image is shown in Figure2:

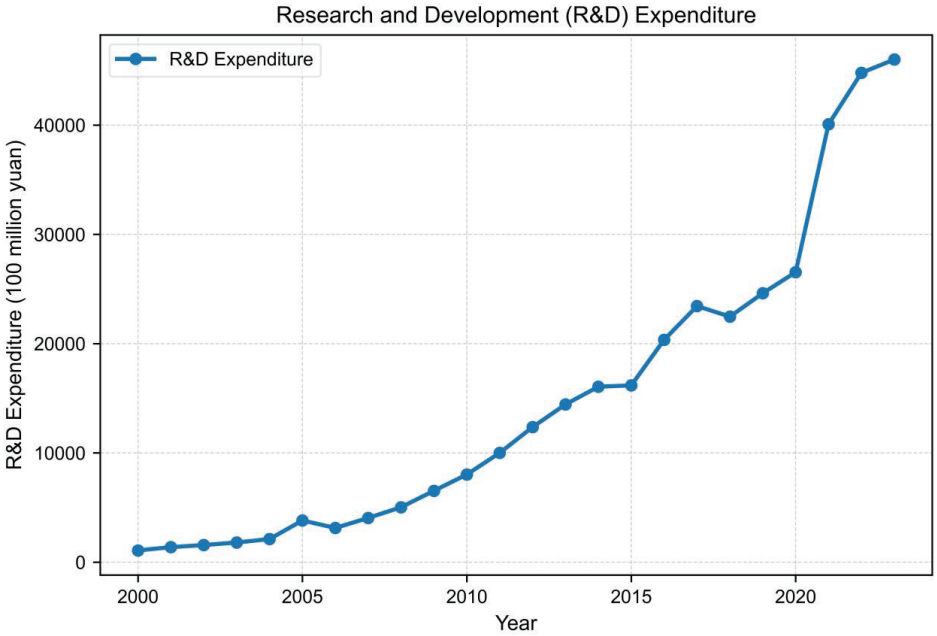

Fig.2 Research and Development (R&D) expenditure (100 million yuan)

3.4 Analysis of outliers

To assess anomalies in the dataset, visual tools such as bar and line charts were employed. A notable observation was a dip in the GDP of the Greater Bay Area in 2021, which diverged from the overall upward trend. Rather than treating this as a data error, it was retained in the analysis because it aligns with real-world conditions—specifically, the economic disruptions caused by the COVID-19 pandemic, which led to widespread lockdowns and reduced consumption.[8]

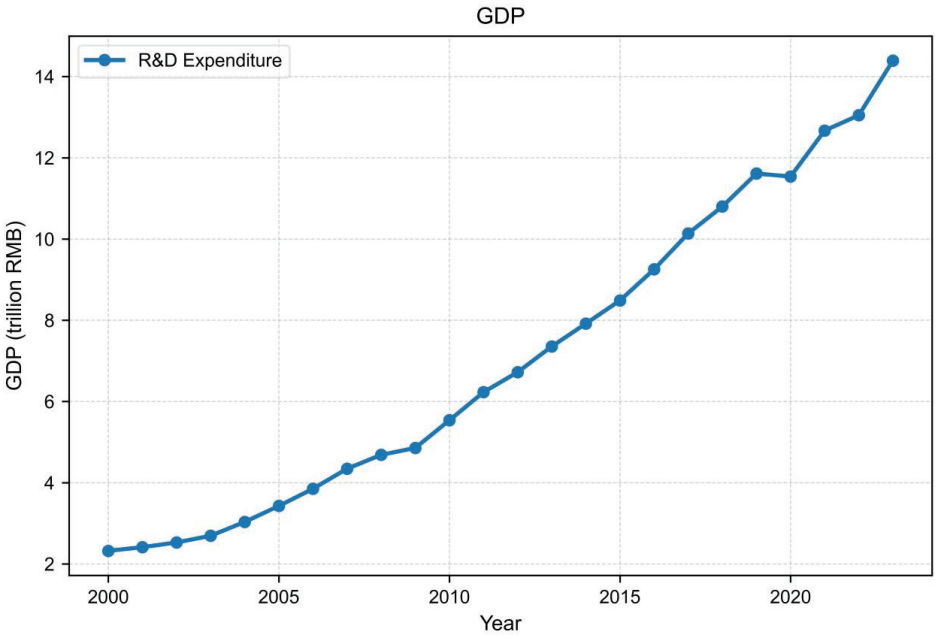

Fig.3 GDP (trillion yuan) Figure

Similarly, a sharp drop was detected in the Tokyo Bay Area's GDP in 2009. Historical context reveals this was during the global financial crisis, which also justifies keeping the data point. In both cases, retaining such outliers allows the model to capture the full scope of economic fluctuations and ensures that forecasts remain grounded in real economic events.

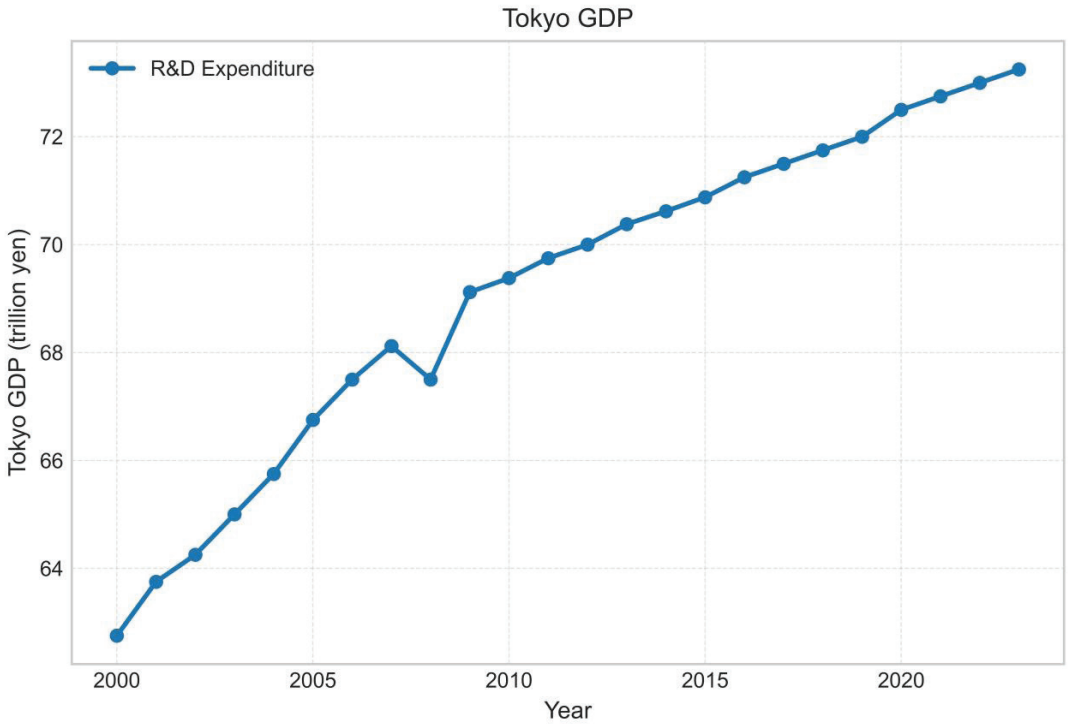

Fig.4 Tokyo GDP (trillion yen) Figure

4 Methodology

4.1 Task 1: Model establishment and solution

4.1.1 Correlation test

To determine which factors most significantly influence the Greater Bay Area's economic performance, the study utilized Principal Component Analysis (PCA) as a dimensionality reduction tool. Correlation matrix analysis showed that many variables—such as import and export values, industrial outputs, and population—exhibited high levels of positive correlation, with coefficients nearing 1.

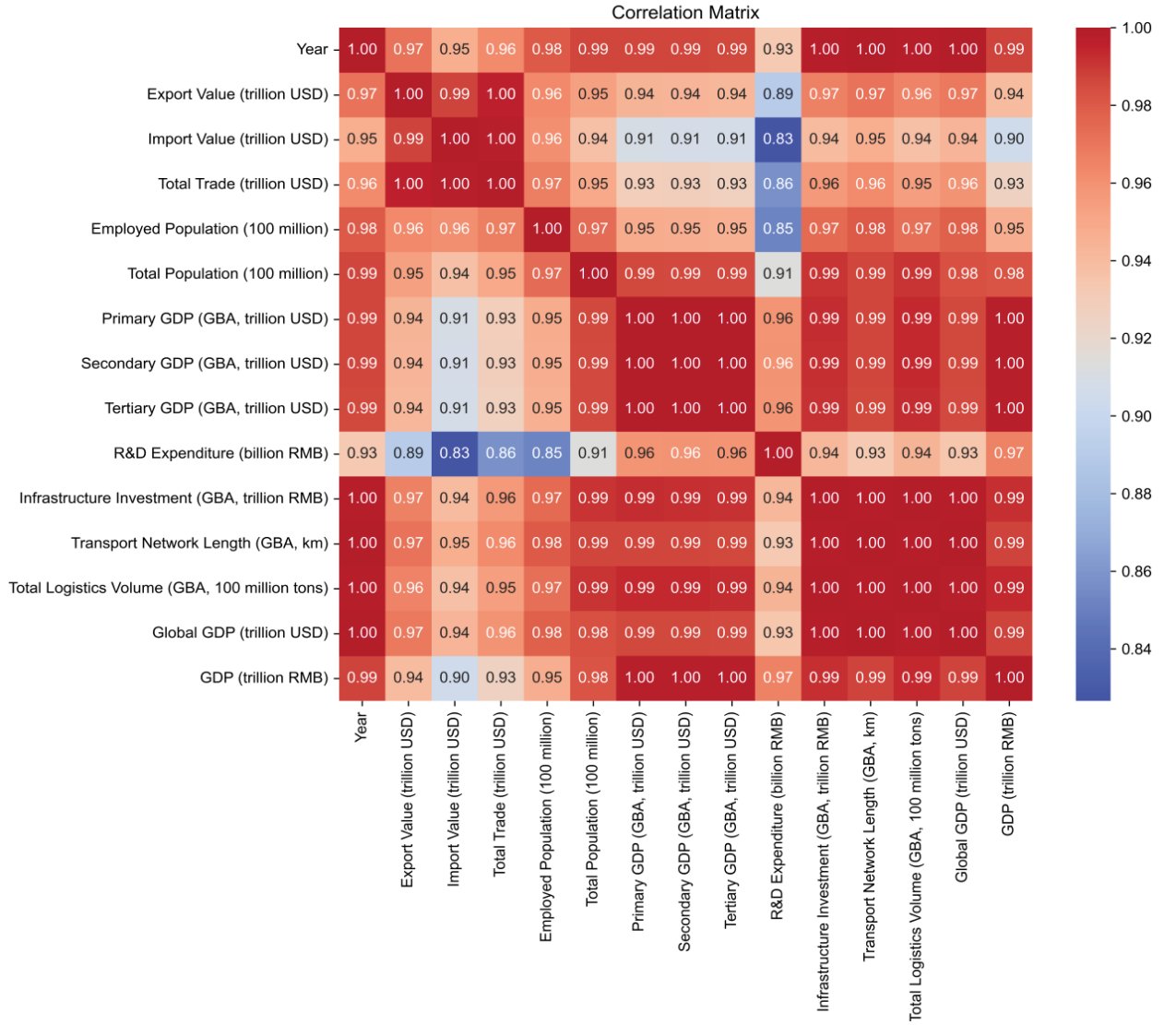

Fig.5 Coefficient matrix of variables in Guangdong-Hong Kong-Macao Greater Bay Area

As can be seen from Figure 6.1:

This high intercorrelation suggests strong multicollinearity among variables, making PCA a suitable method for extracting key components. By compressing redundant information into fewer principal components, PCA simplifies the regression process while retaining most of the original data's explanatory power.

4.1.2 Multiple collinearity problem

In this paper, the Variance Inflation Factor (VIF) is used to test for multicollinearity.

$$VIF_j = \frac{1}{1-R^2} \tag{1}$$

The coefficient R^2 of determination for the linear regression model of the j-th independent variable with all other independent variables. A higher VIF value indicates that the j-th independent variable is more difficult to estimate independently. In economic applications, a VIF value greater than 10 is typically considered a sign of severe multicollinearity, which requires the variable to be addressed.

The VIF values for each variable obtained through the VIF test are as follows:

Table 1. Variance inflation factor VIF of variables in Guangdong-Hong Kong-Macao Greater Bay Area

Variable	VIF
Primary GDP (GBA, trillion USD)	11712.059
Secondary GDP (GBA, trillion USD)	73881.335
Tertiary GDP (GBA, trillion USD)	52988.299
Infrastructure Investment (GBA, trillion RMB)	29456.614
Global GDP (trillion USD)	78659.763
Transport Network Length (GBA, km)	80601.610
Total Population (100 million)	9880.071
R&D Expenditure (billion RMB)	104.403
Employed Population (100 million)	2283.463
Export Value (trillion USD)	3541.550
Total Trade (trillion USD)	31159.761
Import Value (trillion USD)	14883.780
Total Logistics Volume (GBA, 100 million tons)	31122.526

As shown in Table 1, according to the VIF test results, all the above independent variables exceed the threshold of 10, indicating that there is multicollinearity.[20] In order to deal with the problem of multicollinearity, ridge regression method is considered to be used for model estimation in the following paper to reduce the influence of multicollinearity on the estimation of model parameters.[5] [10]

4.1.3 KMO test and Bartlett test

To mitigate the adverse effects of multicollinearity on the model, this paper conducts KMO and Bartlett tests for each variable category. The Bartlett test for each category shows that the p-value is greater than 0.05. For categories with a KMO value greater than 0.7, principal component analysis is performed to reduce dimensionality. The KMO values for each category are listed in Table 2:

Table 2.KMO test table

variable classes	KMO test value	Whether it is suitable for principal component analysis
economic class	0.753	yes
Degree of openness	0.738	yes
Logistics	0.745	yes
research and development	0.5	deny
population	0.5	deny

According to the analysis of the above table, the KMO value of economic category, logistics category and degree of openness is greater than 0.7, which is suitable for principal component analysis.[21]

4.1.4 principal component analysis PCA

(1) Create the data matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \quad (2)$$

among

· Economic variables: total x_{11} value of primary x_{12} industry x_{13}, total value of secondary industry and total value of tertiary industry

· Logistics variables: x_{21} import x_{22} and export x_{23} values

· Open type variables: infrastructure x_{31} investment x_{32}, length x_{33} of transportation network and total logistics

(2) Calculate the covariance matrix

The covariance matrix $\Sigma = [\sigma_{ij}]_{p\times p}$, σ_{ij} Is the covariance between the ith variable and the jth variable. The formula for calculating variance σ_{ij} is:

$$\sigma_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j}) \quad (3)$$

(3)eigenvalue decomposition：

$$\sum v = \lambda v \quad (4)$$

(4)Select the principal component

The expression for the principal component matrix is:

$$X = xv \quad (5)$$

The following is the linear expression of the original variable X of the principal component:

$$X_1 = a_{11}x_{11} + a_{12}x_{12} + a_{13}x_{13} \quad (6)$$

$$X_2 = a_{21}x_{21} + a_{22}x_{22} + a_{23}x_{33} \quad (7)$$

$$X_3 = a_{31}x_{31} + a_{32}x_{32} + a_{33}x_{33} \quad (8)$$

Among them, the a_{ij} coefficient representing x_{kp} the principal X_k component determines the weight of different principal components.

X_1 : Represents the overall impact of the economy

X_2 : Represents the comprehensive impact of logistics

X_3 : Represents the comprehensive impact of opening to the outside world, including infrastructure investment

The three principal components of economic variables is X_1, X_2 and X_3, logistics variables and opening to the outside world explain most of the variance in each category, which is convenient for subsequent regression analysis.

4.1.5 Establishment and solution of multiple regression prediction model

After the principal component analysis, seven variables were selected for regression analysis to quantify and evaluate their impact on the GDP of the Guangdong-Hong Kong-Macao Greater Bay Area. This paper employs ordinary least squares (OLS) to conduct linear regression analysis on

these variables, aiming to quantify their influence on the GDP. The following is an explanation of the principle of ordinary least squares.

$$min\beta \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 = min\beta \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_7 x_{i7})) \quad (9)$$

The linear regression analysis of these variables is carried out to quantify their impact on GDP. The established multiple linear regression model has the following mathematical expression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon \quad (10)$$

among：

Y: The dependent variable, namely the GDP of Guangdong-Hong Kong-Macao Greater Bay Area

X_1: The combined influence of the main components of the economy, the gross value of the primary, secondary and tertiary industries

X_2: The main components of logistics, the comprehensive influence of infrastructure investment, transportation network length and total logistics

X_3: The main component of the degree of opening to the outside world, and the comprehensive influence of export, import and import and export value

X_4: The total global GDP and the macro impact of the global economic environment on the economy of the Greater Bay Area

X_5: The total population, as the basis of economic development

X_6: Research and development funding (R&D) has an important impact on regional innovation and long-term development

X_7: The employed population represents the size of the labor force in economic activity

β_0 For the $\beta_1, \beta_2, \beta_3 \ldots \beta_7$ intercept term, for ε the regression coefficient, and for the error term

The results of the regression analysis are shown in Table 3:

Table 3. Regression analysis results table

variable	coefficient	standard error	t price	p price
economy	1.3048	0.308	4.234	0.001
Degree of openness	0.2052	0.096	2.146	0.048
logistics	-2.6704	1.779	-1.501	0.153
Global GDP	0.2665	0.165	1.619	0.125
population	9.8403	6.110	1.610	0.127
Research and development funding	0.007	0.000	4.003	0.001
working population	4.7372	4.240	1.117	0.280
constant term	-17.5932	12.515	-1.406	0.179

Through Python analysis, the R-squared value is 0.998, indicating that the model can well explain the change of GDP in Guangdong-Hong Kong-Macao Greater Bay Area. The p-value of explanatory variables such as economy, opening to the outside world and research and development funds is less than 0.05, indicating that they have a significant impact on GDP in Guangdong-Hong Kong-Macao Greater Bay Area.

4.1.6 Grading of important factors

From the results of multiple regression analysis, this paper determines which explanatory variables have a significant impact on the dependent variable GDP by looking at the p value.

· Economy (coefficient = 1.3048): has a significant positive impact on GDP.

· Openness to the outside world (coefficient = 0.2052): has a positive impact on GDP.

· Research and development expenditure (coefficient = 0.0007): has a significant positive impact on GDP.

Secondary influencing factors (0.05<p value <0.15)

· Global GDP (coefficient = 0.2665): has a certain positive impact, but does not reach the significant level.

· Population (coefficient = 9.8403): has a certain positive influence, but does not reach the significant level.

Non-significant factors (p value> 0.15)

· constant term

· logistics

· working population

Based on the above analysis, economic factors, openness to the outside world, and R&D funding are key factors significantly impacting the GDP of the Guangdong-Hong Kong-Macao Greater Bay Area. These factors are expected to play a crucial role in shaping the region's economic trajectory over the next 5-10 years, while global GDP and population are considered secondary factors.

4.2 Establishment and solution of the second task model

4.2.1 Future independent variable processing

To use the established economic forecasting model to predict the economic trends over the next 5 to 10 years, we first need to estimate the future values of each independent variable x in the model. This involves a detailed analysis of the historical trends and changes of each key factor, and calculating their respective annual growth rates. Assuming these variables maintain steady growth, we can estimate their values for the next 5 to 10 years. These estimated values will serve as inputs to the model, aiding in predicting future economic development trends.neural networks may provide nonlinear forecasting benefits not captured by traditional models.[11] So consider the following model

4.2.2 Ridge regression model

In ridge regression modell, the independent variables are infrastructure investment, global GDP, transportation network length, population, employed population, R&D expenditure, total logistics, export value, import value, total import and export value, primary industry production value, secondary industry production value and tertiary industry production value. The objective function of ridge regression is established as follows:

$$J(\beta) = \sum_{i=1}^{m} (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \quad (11)$$

In order to solve the ridge regression objective function, we first solve the estimated value of ridge regression coefficient. We first calculate the partial derivatives of the objective function with respect to β_0 and β_j (j=1,2,3,,n):

$$\frac{\partial J(\beta)}{\partial \beta_0} = -2\sum_{i=1}^{m}(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}) = 0 \quad (12)$$

$$\sum_{i=1}^{m} y_i - m\beta_0 - \sum_{i=1}^{m}\sum_{j=i}^{m}\beta_j x_{ij} = 0 \quad (13)$$

$$\beta_0 = \bar{y} - \sum_{j=1}^{n}\beta_j \bar{x}_j \quad (14)$$

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \quad (15)$$

$$\bar{x_j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \quad (16)$$

$$\frac{\partial J(\beta)}{\partial \beta_j} = -2\sum_{i=1}^{m}(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij})x_{ij} + 2\lambda\beta_j = 0 \quad (17)$$

$$\sum_{i=1}^{m}(y_i - \bar{y})x_{ij} - \sum_{j=1}^{n}\beta_j \sum_{i=1}^{m}(x_{ij} - \bar{x_j})x_{ij} + \lambda\beta_j = 0 \quad (18)$$

$$\sum_{i=1}^{m}(x_{ij} - \bar{x_j})(y_i - \bar{y}) = \sum_{j=1}^{n}\beta_j[\sum_{1=1}^{m}(x_{ij} - \bar{x_j})^2 + \lambda] \quad (19)$$

Change β_j the partial derivative of the vector to matrix form:

$$(X - I_m\bar{x}^T)^T(y - \bar{y}I_m) = \beta[(X - I_m\bar{x}^T)^T(X - I_m\bar{x}^T) + \lambda I_n] \quad (20)$$

The estimated value of the ridge regression coefficient obtained by solving the equation is:

$$\hat{\beta} = [(X - I_m\bar{x}^T)^T(X - I_m\bar{x}^T) + \lambda I_n]^{-1}(X - I_m\bar{x}^T)^T(y - \bar{y}I_m) \quad (21)$$

The ridge regression model is trained with historical data, and the optimal regularization λ parameter is selected through cross-validation to obtain:

$$\lambda = 0.001 \quad (22)$$

The obtained model is predicted to make the effect diagram of the model trained with known data, as shown in Figure 6:

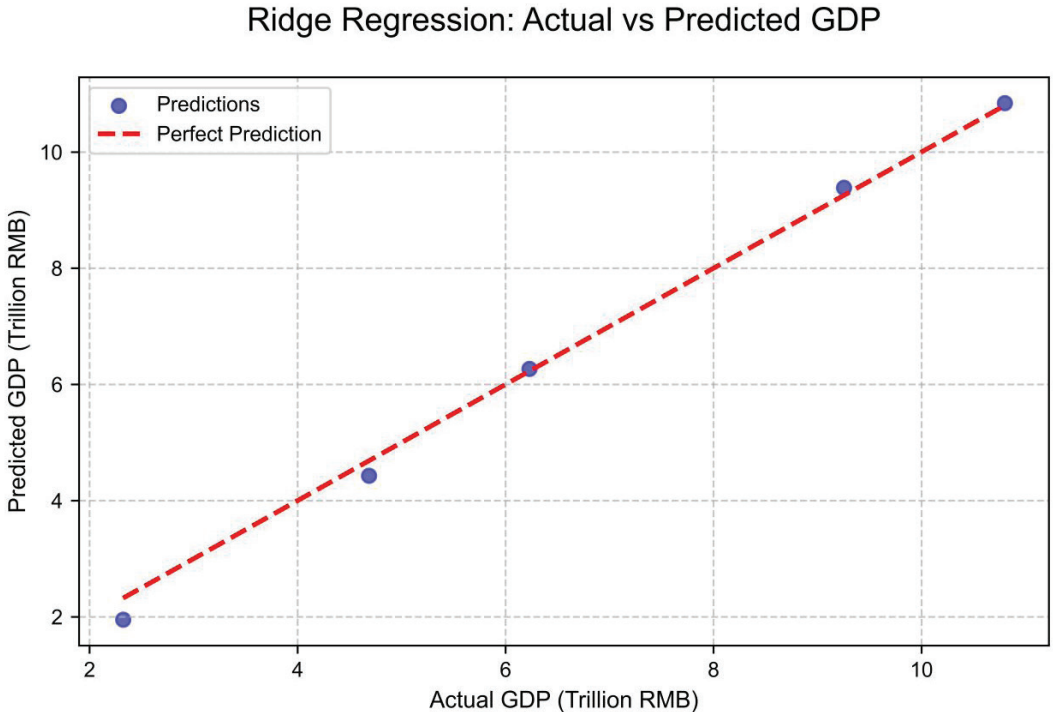

Fig.6 Actual and predicted values of the Lingguangdong-Hong Kong-Macao Greater Bay Area regression model training

Figure 6.2 shows that the predicted values of the ridge regression model and the original data R^2 (i.e., the true value) are highly overlapping on the graph, and the value $\beta_0 = 0$ is as high as 0.9988, which is very close β_j to 1, indicating that the model effect is good. Through calculation, the coefficients of each variable in the model are obtained by substituting the derivation formula as shown in the following table:

Table 4. Variable coefficient table of Linghui model in Guangdong-Hong Kong-Macao Greater Bay Area

variable name	Model coefficients
Value of primary industry	-1.1732
Value of secondary industry	+0.6519
Total value of the tertiary industry	+0.7066
infrastructural investment	-0.0237
Global GDP	+0.1745
Length of transportation network	-0.000018
population	+3.6989
working population	+3.1914
Research and development funding	+0.0007365
Total volume of logistics	-0.0129
export value	-5.7668
import value	-2.3592
total import and export value	+3.6181

From Table 4, we can get:
The coefficients derived from the ridge model revealed:
Positive contributors: Secondary and tertiary industries, global GDP, population, R&D, and total trade
Negative contributors: Primary industry, infrastructure investment, logistics volume, import/export values individually
This suggests the region's economic future will be shaped largely by services, industrial modernization, and global integration.

4.2.3 Summation Autoregressive Moving Average Model (ARIMA)

To capture GDP trends without relying on external variables, the ARIMA (AutoRegressive Integrated Moving Average) model was used.[13] This model excels at modeling temporal patterns in economic data.[6]

$$\varphi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \quad (23)$$

When the ARIMA (p, d, q) $_{d \neq 0}$ model is present, the specific expansion of the sequence $z_t = (1-B)^d y_t$ after d-order difference is:

$$z_t = \varphi_1 z_{t-1} + \ldots + \varphi_2 z_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \quad (24)$$

The steps for building an ARIMA model are as follows:

1. Stationarity Test:

The time series chart of the Guangdong-Hong Kong-Macao Greater Bay Area shows a clear upward trend in GDP. However, the p-value of the ADF test (0.186789) is less than 0.05, indicating that the time series is not stationary and requires differencing. Over the entire time range, the differenced GDP values show a clear upward or downward trend. The p-value of the ADF test (less than 0.05) indicates that after first-differencing, the time series may exhibit stationarity. The time series chart after differencing is shown in Figure 7.

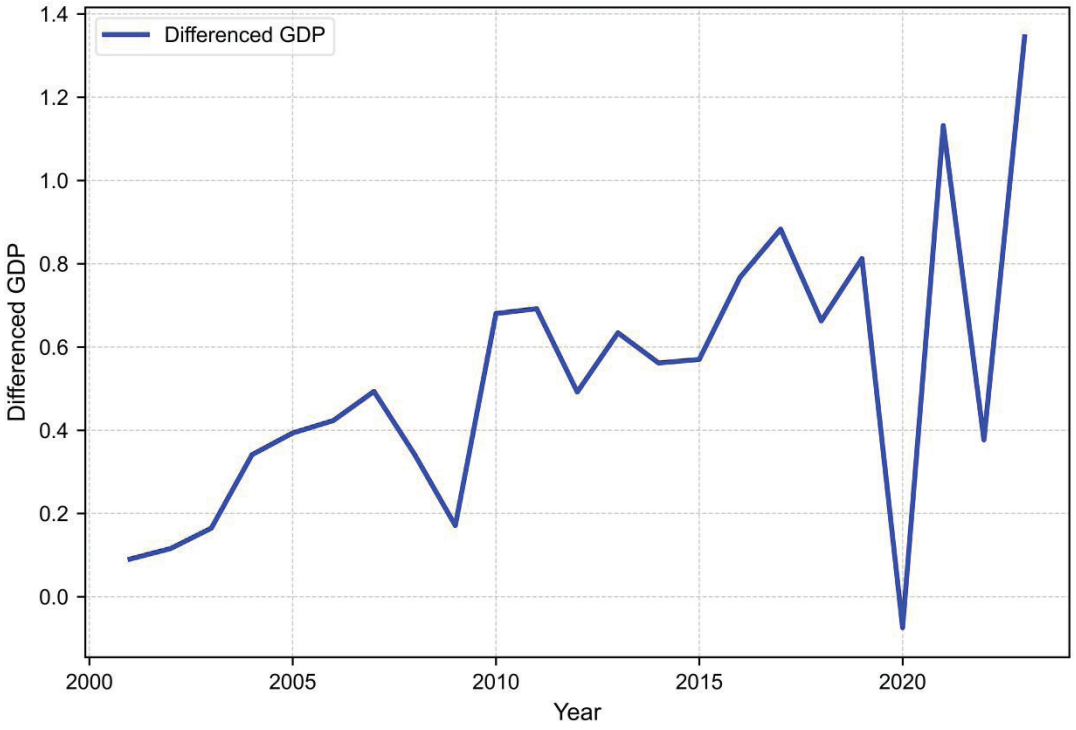

Fig.7 Time series of GDP difference in Guangdong-Hong Kong-Macao Greater Bay Area

2. Determine the model parameter:

The ACF and PACF plots are shown in Figure 8. The order d of the difference is determined by the white noise test, the order q of the autoregressive model is determined by the ACF plot, and the order p of the moving average model is determined by the PACF plot. The optimal orders p and q are verified using information criteria. The results indicate that: since the data has been differenced once, the d value is set to 1; the PACF plot clearly shows it is one-step truncated, so the p value is set to 1; the ACF plot also indicates it is one-step truncated, thus the q value is set to 1. Therefore, the ARIMA model is identified as ARIMA(1,1,1).

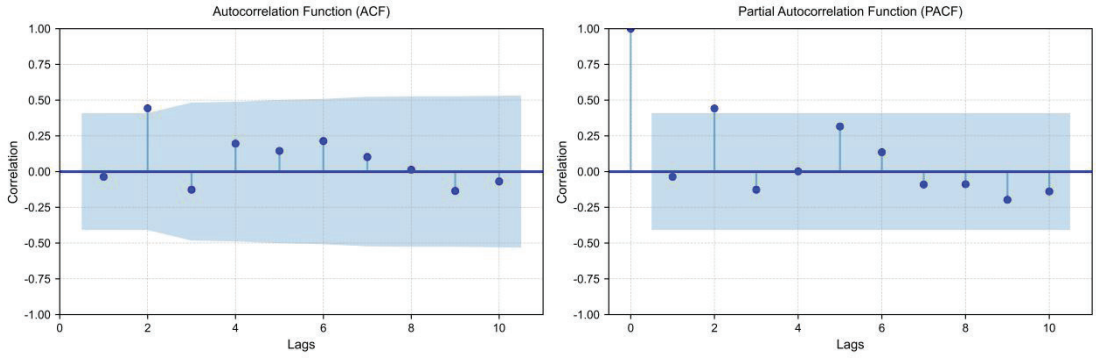

Fig.8 ACF and PACF of Guangdong-Hong Kong-Macao Greater Bay Area

According to the above tests, the ARIMA (1,1,1) model can be used to forecast the GDP of Guangdong-Hong Kong-Macao Greater Bay Area

4.2.4 Predicting future economic trends

1. Multiple regression prediction

Using the 7 important influencing variables selected in Task 1, a multiple regression model was established to predict the GDP of Guangdong-Hong Kong-Macao Greater Bay Area in the next 10 years and draw the result figure, as shown in Figure9:

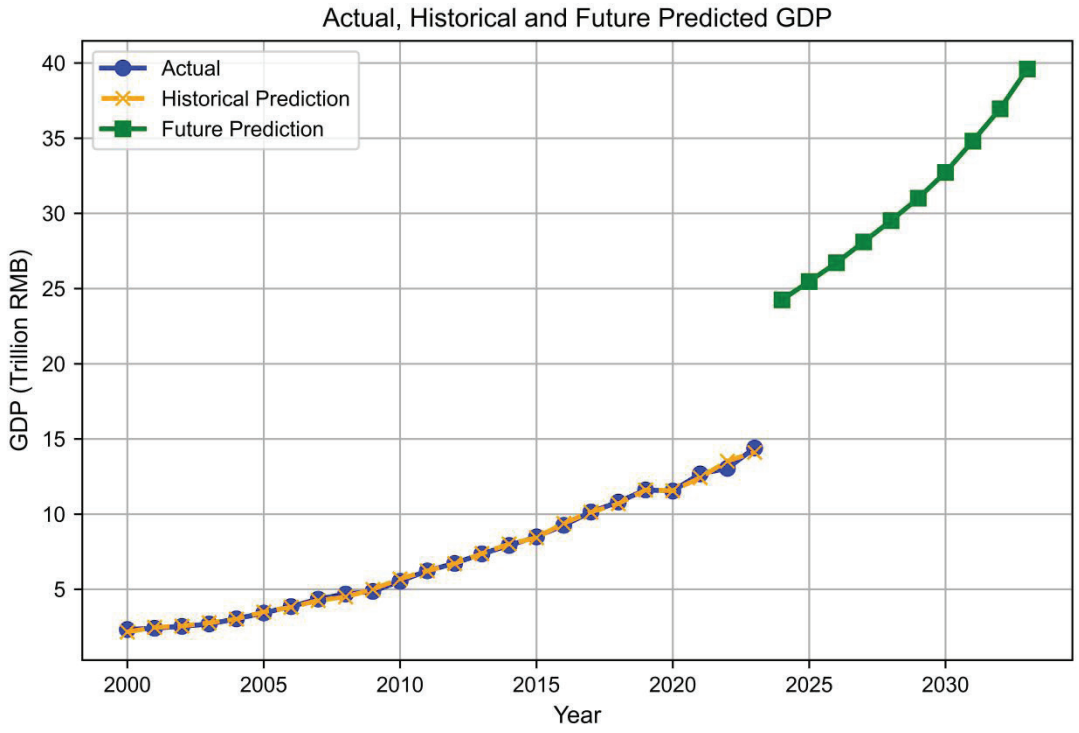

Fig.9 Comparison of actual values, historical forecast values and forecast values for the next 10 years

Figure 9 shows that the multiple regression model overestimates the GDP growth of the Guangdong-Hong Kong-Macao Greater Bay Area over the next decade. The prediction in 2023 significantly deviates from the actual value, with a significant gap at the connection point. This discrepancy

is due to multicollinearity among the variables used in this study, which affects the stability of the model and makes it highly sensitive to minor input changes, leading to increased volatility in future predictions. Economic growth is often influenced by multiple complex factors in a non-linear manner, especially in long-term forecasts, where linear regression models may fail to accurately capture these nonlinear trends. Additionally, multiple linear regression models cannot account for the dependencies in time series data, which are crucial for understanding long-term economic trends. The discontinuity in the model when transitioning to future predictions may be due to its inability to smoothly transition to changes in future data. Therefore, the multiple regression model is not suitable for predicting the GDP of the Guangdong-Hong Kong-Macao Greater Bay Area.[4]

2. Ridge regression prediction

The trained ridge regression model is used to predict the GDP of the next 10 years, and the estimated values of the above future independent variables are input into the model for prediction. Finally, the predicted GDP results of the next 10 years are plotted to intuitively show the future economic trend, as shown in Figure 10:

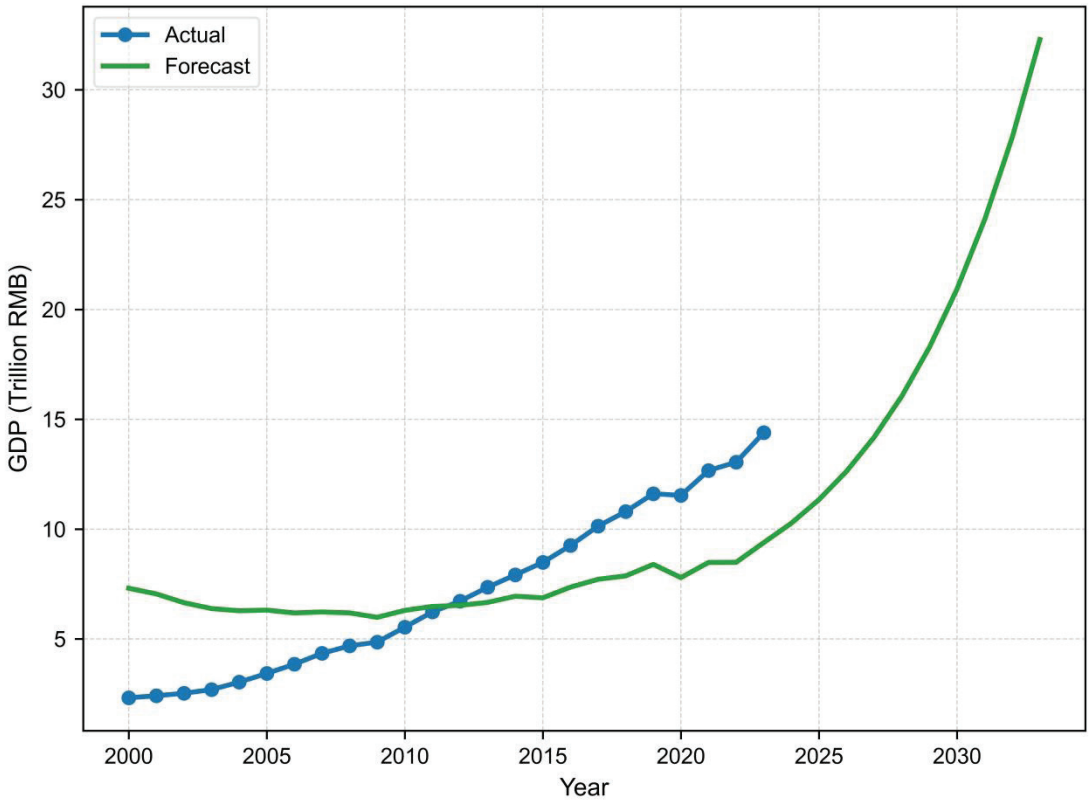

Fig.10 Actual GDP of Guangdong-Hong Kong-Macao Greater Bay Area and ridge regression forecast trend chart

As shown in Figure 10, the trend of the original data (blue solid line) and the ridge regression prediction (green solid line) diverges over the next few years. The ridge regression prediction indicates a rapid increase in future GDP, showing an accelerating upward trend. While this trend highlights the positive impact of various variables, it does not align with the actual economic growth

rate. Ridge regression may overestimate the growth rate when capturing the impact of variables on economic growth. This is because ridge regression relies on the relationships between multiple variables; if some variables exhibit nonlinear changes or future trends are unclear, it can lead to significant prediction errors. Therefore, when the data pattern is complex or highly uncertain, ridge regression results may be overly optimistic.

3. Time series prediction

In order to solve the problem of excessive growth rate predicted by ridge regression, this paper introduces time series model for trend prediction, and the results are shown in Figure 11:

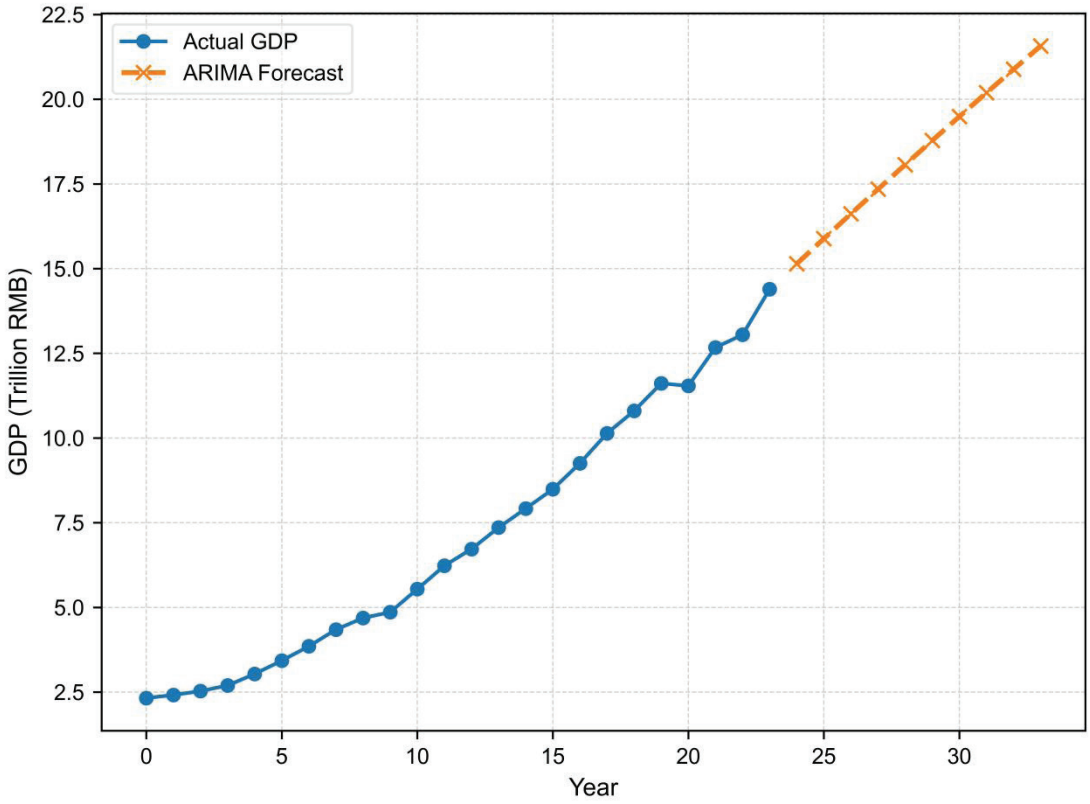

Fig.11 Comparison of actual and predicted values in Guangdong-Hong Kong-Macao Greater Bay Area

The chart compares the actual values (blue solid line) with the predictions from the time series model (orange dashed line) from 2000 to 2030. It clearly shows that the growth rates of the original data and the time series model predictions are highly consistent. The time series model, which uses historical GDP data to predict future trends, fully considers the autocorrelation of GDP without relying on other variables. This method maintains the stability of economic growth, thus being less affected by fluctuations in a single variable.

4. Comparison of three models

The prediction results of multiple regression model, time series model and ridge regression model are plotted by drawing as shown in Figure 12:

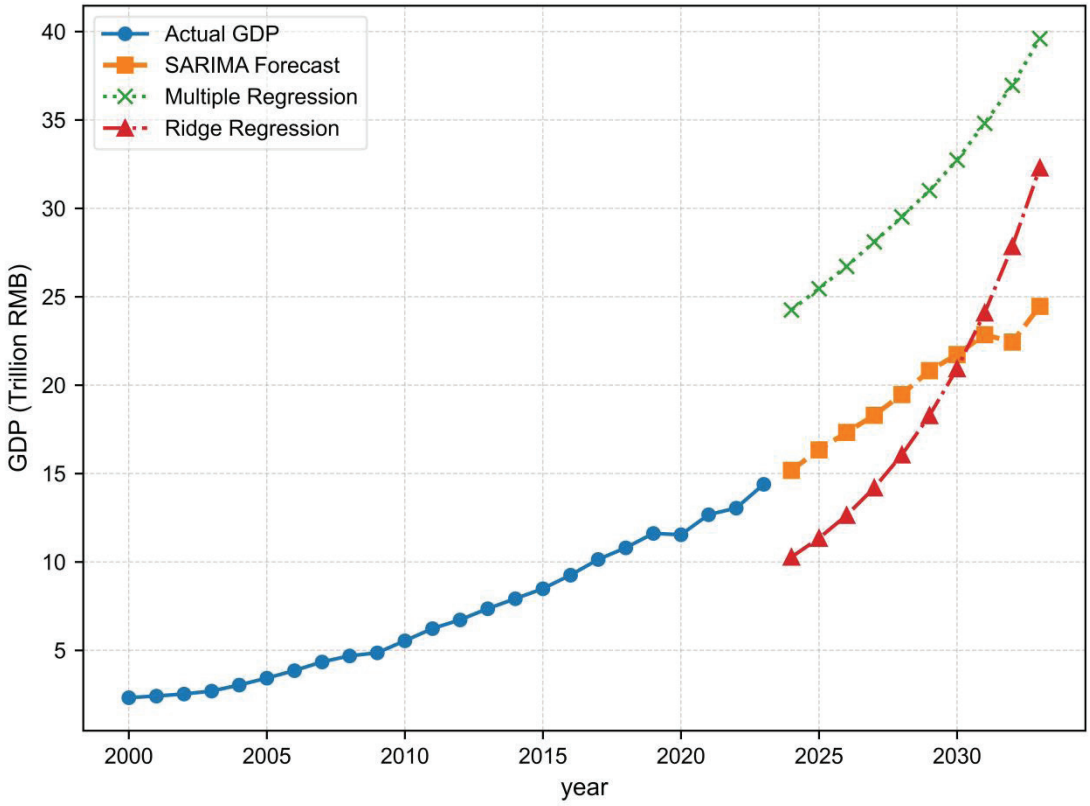

Fig.12 Real GDP of Guangdong-Hong Kong-Macao Greater Bay Area: Multivariate regression, ridge regression simulation and ARIMA model prediction trend chart

Through the analysis of the above figure, three models have been derived to predict future GDP growth trends. The multiple regression model has an overestimated trend and a significant gap at the connection point; the ridge regression model predicts a rapid growth trend; the time series model closely matches the actual growth rate, with a small discrepancy from the actual values, demonstrating strong predictive power[15][18]. In summary, ridge regression is suitable for multi-factor analysis and factor quantification, while time series models are better suited for predicting stable trends, making them effective in forecasting future GDP values.

According to the above analysis, the forecast value of time series is reasonable. The following is the GDP forecast of Guangdong-Hong Kong-Macao Greater Bay Area in the next 5-10 years, as shown in Table 5:

Table 5. GDP forecast of Guangdong-Hong Kong-Macao Greater Bay Area in the next 5-10 years

a particular year	Future GDP forecast of Guangdong-Hong Kong-Macao Greater Bay Area (RMB trillion)
2024	15.184311
2025	16.333721
2026	17.323576
2027	18.286183

a particular year	Future GDP forecast of Guangdong-Hong Kong-Macao Greater Bay Area (RMB trillion)
2028	19.463588
2029	20.813570
2030	21.720871
2031	22.848533
2032	22.436690
2033	24.453197

4.2.5 Strategy Suggestions

According to the forecast results of the hybrid model, the economy of Guangdong-Hong Kong-Macao Greater Bay Area will maintain steady growth in the next 5-10 years. In order to promote the sustainable development of regional economy, the following suggestions are put forward:

1. Strengthen investment in scientific and technological innovation: increase r&d spending and improve technological innovation capacity to support future economic growth.
2. Optimize the policy of opening to the outside world: further expand the scale of foreign trade, enhance the openness of economy and international competitiveness.
3. Improve infrastructure construction: increase investment in transportation network and logistics facilities to improve economic operation efficiency.
4. Optimize the industrial structure: accelerate the development of the tertiary industry, promote the transformation of the economy from manufacturing to service, and improve the quality of economic development.

4.3 Task 3: Model establishment and solution

4.3.1 Variable selection

Tokyo Bay Area variables:

1. Industrial output value: Output value of the primary, secondary and tertiary industries in Tokyo (unit: trillion yen)
2. Demographic factors: employed population (unit: 10,000)
3. Foreign trade: value of exports, value of imports and value of imports from Tokyo (in US dollars)
4. Others: Global GDP (unit: trillion US dollars), R&D expenditure (unit: trillion yen)

4.3.2 Predict future economic trends

Based on the findings from Task 2, the ARIMA (AutoRegressive Integrated Moving Average) model demonstrated strong predictive performance.[1] To evaluate whether this model is also applicable to the Tokyo Bay Area, the following validation procedures were conducted:

1. Stationarity test:The Augmented Dickey-Fuller (ADF) test was used to assess the stationarity of the GDP time series. Initially, the p-value exceeded 0.05, indicating non-stationarity. However, after applying a first-order differencing transformation, the p-value dropped below 0.05. This suggests that the GDP series became stationary after differencing, which satisfies one of the ARIMA modeling requirements, as shown in Figure 13:

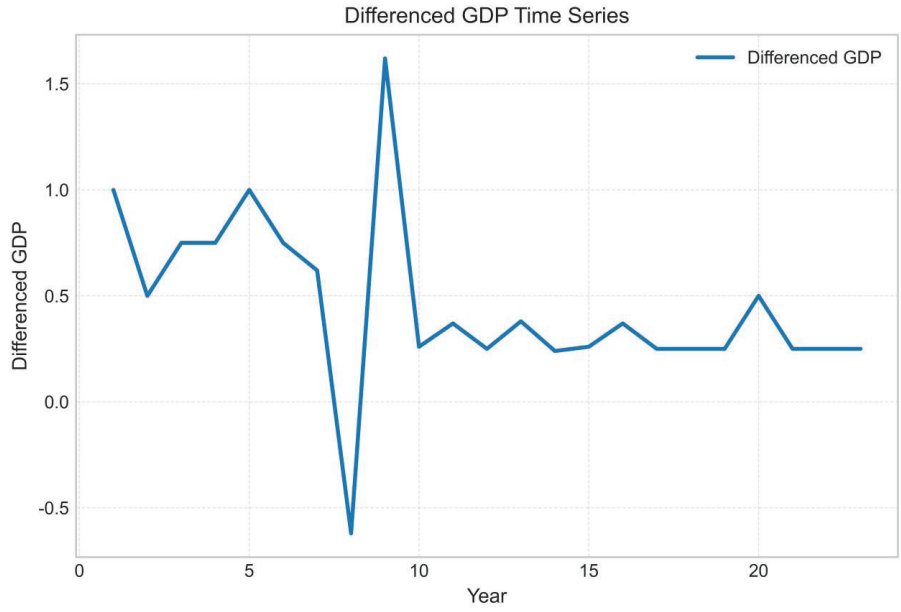

Fig.13 Time series of Tokyo Bay Area data after data differentiation

2. White noise test:it was conducted on the residuals of the differenced series. Since the p-value was greater than 0.05, we failed to reject the null hypothesis that the residuals represent white noise. This indicates that the model adequately captures the underlying trend, and no systematic patterns remain in the residuals. Therefore, the model is considered valid without needing additional modifications

3. Determine model parameters:

The autocorrelation function (ACF) plot showed a significant spike at lag 1, indicating a potential moving average component of order 1 (MA(1)). After this lag, the autocorrelation values rapidly approached zero. This behavior supports the selection of q = 1.

Similarly, the partial autocorrelation function (PACF) plot displayed significant values at lags 1 and 2, suggesting the presence of an autoregressive component of order 1 or 2 (AR(1) or AR(2)). As a result, both p = 1 and p = 2 were considered for model testing.

After testing both ARIMA(1,1,1) and ARIMA(2,1,1), the latter was selected due to its lower AIC and BIC scores, which indicate a better model fit. Although a few parameters in ARIMA(2,1,1) lacked statistical significance, the residuals passed the white noise test, confirming the model's overall validity. Therefore, ARIMA(2,1,1) was chosen for forecasting Tokyo's GDP.

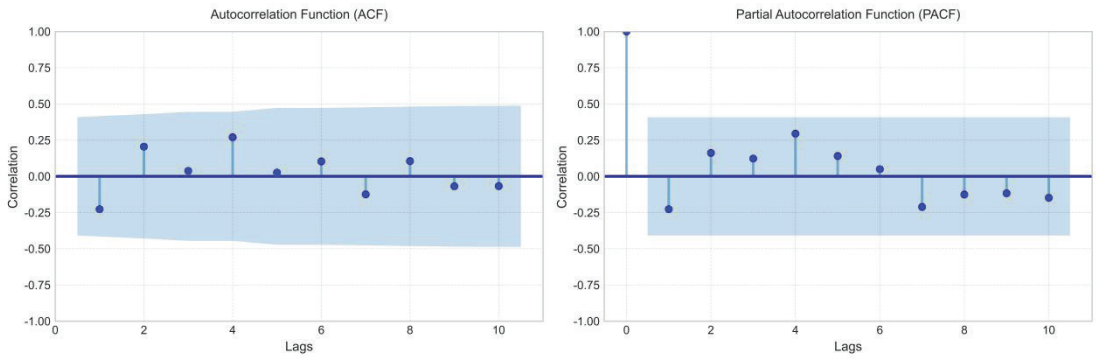

Fig.14 ACF and PACF of Tokyo Bay Area

According to the above test, the data of Tokyo Bay Area meet the requirements of ARIMA time series analysis, and the ARIMA (2,1,1) model is established as follows:

$$\varphi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \tag{25}$$

The actual GDP and GDP forecast trend of Tokyo Bay Area are drawn as follows: Figure 15

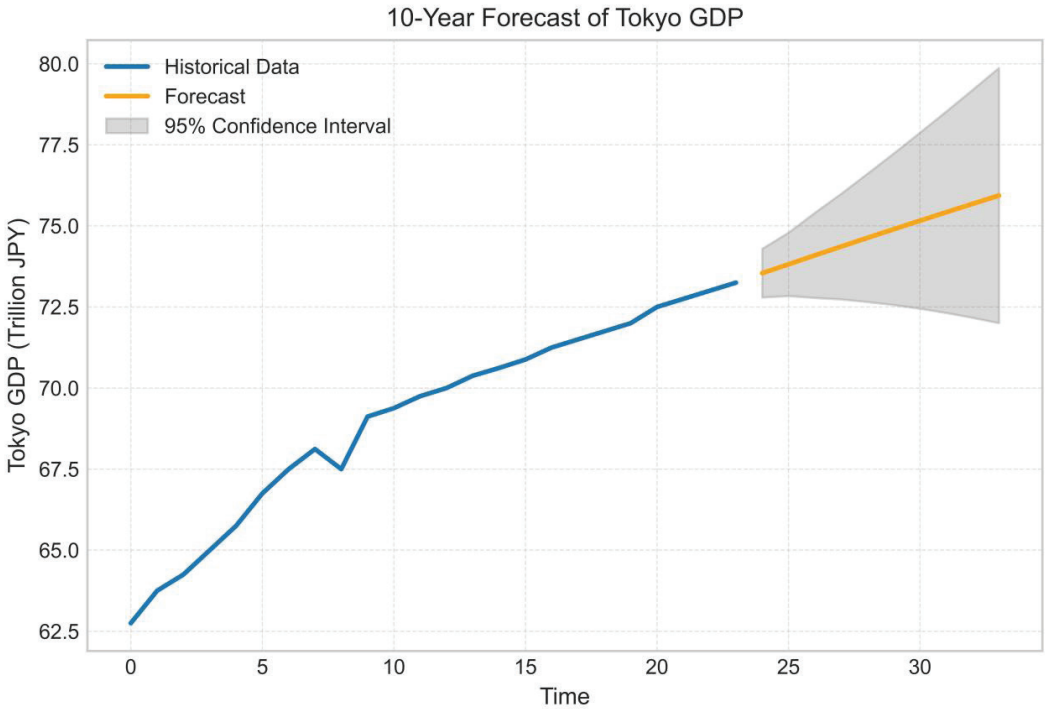

Fig.15 Actual GDP of Guangdong-Hong Kong-Macao Greater Bay Area and SARIMA model forecast trend chart

The following is the GDP forecast of Tokyo Bay Area in the next 5-10 years, as shown in Table 6:

Table 6. GDP forecast of Tokyo Bay Area in the next 5-10 years

a particular year	Future GDP forecast of Tokyo Bay Area (trillion yen)
2024	73.544075
2025	73.814881
2026	74.091975
2027	74.362439
2028	74.631974
2029	74.898119
2030	75.161989
2031	75.423133
2032	75.681784
2033	75.937877

4.3.3 Differences and similarities of different Bay Areas

In order to ensure the stability and reliability of the model's parameter estimation, and the model has good explanatory power and prediction ability, it is necessary to first conduct VIF test to determine whether there is multiple collinearity among the variables in the Tokyo Bay Area. The results are shown in Table 7:

Table 7 Variance inflation factor VIF of variables in Tokyo Bay Area

Varible	VIF
Output value of Tokyo's primary industry (trillion yen)	40332.32
Output of Tokyo's secondary industry (trillion yen)	126328.7
Output of tertiary industry in Tokyo (trillion yen)	165671
World GDP (in US trillion dollars)	121.2948
Research and development expenditure (R&D) (trillion yen)	1261.002
Employed population (10,000)	3942.753
Value of exports from Tokyo (billion US dollars)	inf
Value of imports and exports of Tokyo (billion US dollars)	inf
Value of imports from Tokyo (in $ million)	inf

As can be seen from the variance inflation factor (VIF) value table in Table 7, the VIF values of some variables are very high, even reaching infinity (inf). This indicates that there is multicollinearity among these variables, that is, some variables are highly correlated, which leads to the instability of the regression model. The specific situation analysis is as follows:

1. Variables with high VIF values

· The output value of the primary industry, the secondary industry and the tertiary industry in Tokyo all have extremely high VIF values (between 40,000 and 160,000), indicating that there is significant multicollinearity among them. This may be because the output value of these industries is related to each other in the economy, for example, the growth of one industry will drive the growth of other industries.

· The total global GDP and research and development (R&D) funds also have higher VIF values (100 and 1261), which may be related to other economic indicators.

· The VIF value of the employed population also reached 3942, indicating that the employed population may be highly correlated with other economic indicators.

2. Infinite VIF value

· The VIF value of the export value, import and export value and import value of Tokyo is infinite (inf), which indicates that there is a complete linear dependence relationship between these variables (it may be because the import and export value is directly related to the total import and export value, resulting in the collinearity problem).

When analyzing the variables in the Tokyo Bay Area, it was found that there is a strong multicollinearity among the explanatory variables, particularly the high correlation between the output values of different industries and the data on import and export trade. To address the

multicollinearity issue, this paper employs the ridge regression model. Ridge regression introduces regularization parameters to constrain the model, thereby reducing the volatility of the regression coefficients. The objective function of ridge regression is:

$$\min_{\beta} \sum_{i=1}^{n} (y_4 - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 \quad (26)$$

The independent variable matrix consists of the output value of Tokyo's primary, secondary and tertiary industries (unit: trillion yen), the employed population (unit: 10,000 people), the export value of Tokyo, the import value of goods (unit: billion US dollars), the total global GDP (unit: trillion US dollars), and the R&D expenditure (unit: trillion yen).

The ridge regression model is established $\lambda_2 = 0.01$ to obtain the regularization parameters, and the figure is drawn as shown in Figure 16:

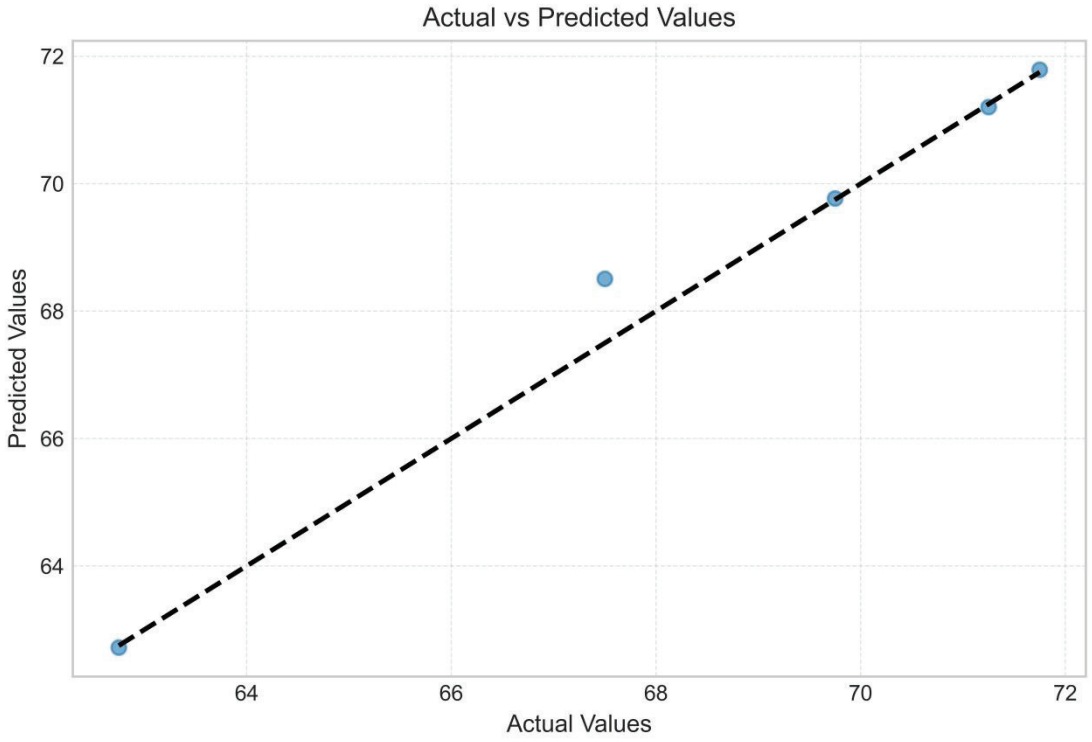

Fig.16 Actual and predicted values of ridge regression model training in Tokyo Bay Area

The overlap degree between the predicted results and $R^2 = 0.98099$ the actual results is high, and the model $\beta_0 = 0$. It can be seen that the ridge regression model has a good effect. Substitute to obtain the constant term and list the model coefficients β_J of each variable, as shown in Table 8:

Table 8. Variable coefficients of ridge regression model

variable name	Model coefficients
Output value of Tokyo's primary industry (trillion yen)	-0.001070848
Output of Tokyo's secondary industry (trillion yen)	-0.486404889
Output of tertiary industry in Tokyo (trillion yen)	1.61818532

variable name	Model coefficients
World GDP (in US trillion dollars)	-0.010433462
Research and development expenditure (R&D) (trillion yen)	0.03845146
Employed population (10,000)	0.001773864
Value of exports from Tokyo (billion US dollars)	3.40454E-05
Value of imports and exports of Tokyo (billion US dollars)	1.32271E-05
Value of imports from Tokyo (in $ million)	-2.08184E-05

4.3.4 disparate paindts：

1. Economic structure difference: The economy of Tokyo Bay Area is more diversified, dominated by the tertiary industry, and relies heavily on international trade. In contrast, Guangdong-Hong Kong-Macao Greater Bay Area has a stronger advantage in the secondary industry, and benefits from the support of China's domestic market.

2.GDP Growth Trend: The Tokyo Bay Area is expected to see a relatively stable and low GDP growth rate in the future. The region's industries and markets are already well-established. In contrast, the Guangdong-Hong Kong-Macao Greater Bay Area has significant growth potential in areas such as infrastructure investment and technological innovation, and is projected to maintain a high growth rate over the next 5-10 years.

common ground：

1. The economies of the two Bay areas are significantly affected by factors such as population R&D input and foreign trade.

2. Model effect comparison: Ridge regression and time series hybrid model both show good results in the prediction of the two Bay areas, which can effectively balance the influence of various variables. Especially in the treatment of highly correlated explanatory variables, the introduction of ridge regression significantly improves the stability of the model.

5 Results

5.1 summary

Based on the results from the principal component analysis and regression modeling, the following conclusions were drawn:

Ridge regression revealed that global GDP, total population, R&D expenditure, openness to foreign trade, and output from the secondary and tertiary sectors all have a positive impact on GDP growth.

In contrast, infrastructure investment and total trade volume showed a negative correlation with GDP, possibly reflecting diminishing marginal returns or structural inefficiencies.

The model achieved an R-squared value of 0.998, indicating an excellent fit between the variables and GDP.

Among all forecasting models tested, the ARIMA model produced the most stable and accurate GDP predictions over a 10-year horizon.[7]

A comparison with the Tokyo Bay Area shows that while the Greater Bay Area still lags in absolute economic volume, its growth potential is significantly higher.[23]

5.2 innovative policy suggestions and countermeasures

To ensure the continued growth of the Guangdong-Hong Kong-Macao Greater Bay Area, it is recommended that policy makers take action from the following innovative perspectives:

Create "digital twin"Greater Bay Area: Establish a regional economic simulation system based on digital twin technology, monitor and analyze the economic dynamics of the Greater Bay Area in real time, and combine big data and artificial intelligence for real-time prediction and resource optimization. Improve the regional early warning and forward-looking decision-making ability.

Building a global "talent free trade zone": Establishing a talent free trade zone within the region, implementing flexible work visas, remote office support, global workplace certification and other policies to attract high-end talents from all over the world to enter and leave freely. Promoting innovation and economic integration by introducing international cooperation institutions, top universities and research centers.

Develop "green innovation special zones": Set up green innovation special zones, encourage the development of carbon capture, new energy and other green industrial chains, provide tax incentives for enterprises that save energy and reduce emissions, support the issuance of green bonds, and drive the transformation of the Greater Bay Area to sustainable development.[25]

Building the "Bay Area Health Economy Corridor": Leveraging the population and medical resources of the Greater Bay Area, we aim to create an economic corridor centered on smart health, elderly care services, and life science innovation. By utilizing intelligent healthcare, telemedicine, and health big data, we will promote the development of the health economy, meet the needs of an aging society, and attract biotech companies to the region, thereby providing new growth opportunities for the local economy.

6 Model evaluation

6.1 Advantages

(1) Combination of multiple factors: The model combines the advantages of ridge regression and time series analysis. Ridge regression is used to deal with the problem of multicollinearity, and time series is used to capture historical trends.[2][19] The combination of the two provides a more comprehensive prediction for economic development.

(2) Variable dimension reduction: Through principal component analysis (PCA), the problems caused by multicollinearity are effectively reduced, the main information is retained, and the stability and explanatory power of the model are improved.

(3) Robustness of the hybrid model: By combining ridge regression and time series prediction through weighted average, a hybrid model is formed to effectively balance the influence of each variable and enhance the robustness and accuracy of the prediction.[12][14]

(4) Policy guidance value: The model prediction results clearly reveal the main driving factors of economic growth, and provide a scientific basis for policy formulation in the Guangdong-Hong Kong-Macao Greater Bay Area, especially in terms of policy recommendations on infrastructure, technological innovation and opening up to the outside world.

6.2 Disadvantages

(1) Strong dependence on data: The model depends on historical data to predict future trends,

so the accuracy of prediction may be affected when unforeseen events (such as global economic fluctuations) occur.[16]

(2) Subjectivity of weight setting of the hybrid model: The weight allocation of ridge regression and time series is based on data characteristics, but no dynamic adjustment is made, which may lead to the prediction results failing to fully adapt to the actual situation in some economic situations.

7 REFERENCES

[1] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.

[2] Chatterjee, S., & Hadi, A. S. (2012). Regression analysis by example (5th ed.). John Wiley & Sons.

[3] Fujita, M., & Hu, D. (2001). Regional disparity in China 1985–1994: The effects of globalization and economic liberalization. The Annals of Regional Science, 35(1), 3–37. https://doi.org/10.1007/s001680000020

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer. https://web.stanford.edu/~hastie/ElemStatLearn/

[5] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[6] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts. https://otexts.com/fpp2/

[7] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

[8] Wang, S., & Zhao, D. (2020). Detecting outliers in economic time series using visual and statistical methods. Open Journal of Statistics, 10(5), 433–449. https://doi.org/10.4236/ojs.2020.105029

[9] Yuan, F., & Li, Q. (2022). Data imputation in economic modeling: An application of linear interpolation and PCA. Econometrics, 10(1), 10. https://doi.org/10.3390/econometrics10010010

[10] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[11] Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

[12] Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), Handbook of Economic Forecasting (Vol. 1, pp. 135–196). Elsevier. https://doi.org/10.1016/S1574-0706(05)01004-9

[13] Perone, G. (2021). Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast COVID-19 hospitalizations in Italy. Health Economics Review, 11(1), 8. https://doi.

org/10.1007/s10198-021-01347-4
[14] Ahmadianfar, I., et al. (2025). A hybrid framework: SVD + kernel ridge regression for river water level forecasting. Scientific Reports, 15(1), 1654. https://www.nature.com/articles/s41598-025-90628-6
[15] Zhang, T., & Wang, Y. (2023). Mixed ARIMA-LSTM model for energy consumption forecasting in China. Energies, 16(4), 1987. https://doi.org/10.3390/en16041987
[16] Taleb, N. N. (2007). The black swan: The impact of the highly improbable. Random House.
[17] Chien, T., & Zhang, L. (2021). Spatiotemporal prediction of regional economic growth using machine learning. Sustainability, 13(19), 10987. https://doi.org/10.3390/su131910987
[18] Liu, Y., & Yu, H. (2022). Comparative study of ARIMA and Prophet for GDP forecasting in emerging markets. Forecasting, 4(2), 300–315. https://doi.org/10.3390/forecast4020017
[19] Chen, S., & Shi, Y. (2020). Forecasting urban economic growth using hybrid models: A comparative study. Applied Economics Letters, 27(20), 1697–1701. https://doi.org/10.1080/13504851.2019.1707772
[20] Han, X., & Wang, M. (2023). Ridge regression for regional GDP modeling under multicollinearity. Econometrics, 11(1), 6. https://doi.org/10.3390/econometrics11010006
[21] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
[22] OECD. (2020). International migration outlook 2020. OECD Publishing. https://www.oecd-ilibrary.org/social-issues-migration-health/international-migration-outlook-2020_ec98f531-en
[23] Xu, X., & Zhang, M. (2020). Comparative development of global Bay Areas: The case of Tokyo and GBA. Sustainability, 12(12), 4991. https://doi.org/10.3390/su12124991
[24] Chen, Y., & Lu, M. (2021). China's regional economic transformation and the role of Bay Area economies. Sustainability, 13(6), 3451. https://doi.org/10.3390/su13063451
[25] Rennings, K. (2000). Redefining innovation — eco-innovation research and the contribution from ecological economics. Ecological Economics, 32(2), 319–332. https://doi.org/10.1016/S0921-8009(99)00112-3

Statements and Declarations

8 Algorithms, Program codes and Listings

Name of software used: Python Excel SPSSPRO

Task 1: correlation coefficient matrix, VIF, principal component analysis and multiple linear regression

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
import numpy as np
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
# reading data
Data = pd.read_excel('Guangdong-Hong Kong-Macao Greater Bay Area.xlsx')
# Calculate the correlation coefficient matrix
correlation_matrix = data.corr()
# Draw a heat map of the correlation coefficient matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix,annot=True,fmt=".2f",cmap="coolwarm",cbar=True, square=True)
plt.title("Correlation Matrix")

from statsmodels.stats.outliers_influence
import variance_inflation_factor
Prepare the data
X = data['Guangdong-Hong Kong-Macao Greater Bay Area's First Gross Domestic Product (USD billion) ',' 
        Guangdong-Hong Kong-Macao Greater Bay Area's Second Gross Domestic Product (T yuan)', 
        'Guangdong-Hong Kong-Macao Greater Bay Area's Third Gross Domestic Product (USD billion) 
        ',' Guangdong-Hong Kong-Macao Greater Bay Area's Infrastructure Investment (T RMB)', 'Global 
        GDP Total (USD billion)', 'Guangdong-Hong Kong-Macao Greater Bay Area's Transportation 
        Network Length (km) ',' Population (100 million) ',' Research and Development (R&D) Funding 
        (Billion Yuan) ',' Employed Population (100 million) ',' Export Value (USD billion) ',' Total 
        Import and Export Value (USD billion) ',' Import Value (USD billion) ',' Guangdong-Hong Kong-
        Macao Greater Bay Area's Total Logistics Volume (billion tons)']])'
```

```
# count VIF
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print(vif_data)

# linear regression
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
Prepare the data
X1 = data['GDP of the Guangdong-Hong Kong-Macao Greater Bay Area (US $ trillion)', 'GDP of the
        Guangdong-Hong Kong-Macao Greater Bay Area (US $ trillion)', 'GDP of the Guangdong-Hong
        Kong-Macao Greater Bay Area (US $)']]
X2 = data['export value (trillion US dollars)', 'total import and export value (trillion US dollars)', 'import
value (trillion US dollars)']]
X3 = data['Length of transportation network in Guangdong-Hong Kong-Macao Greater Bay Area (km)',
        'Total logistics volume in Guangdong-Hong Kong-Macao Greater Bay Area (100 million tons)',
        'Infrastructure investment in Guangdong-Hong Kong-Macao Greater Bay Area (trillion RMB)']]
X1_scaled = StandardScaler().fit_transform(X1)
X2_scaled = StandardScaler().fit_transform(X2)
X3_scaled = StandardScaler().fit_transform(X3)
# PCA
pca = PCA (n_components=1) # Select the appropriate number of components
X1_pca = pca.fit_transform(X1_scaled )
X_pca_1 = pd.DataFrame (X1_pca, columns=['principal component 1'])
X2_pca = pca.fit_transform(X2_scaled)
X_pca_2 = pd.DataFrame (X2_pca, columns=['principal component 2'])
X3_pca = pca.fit_transform(X3_scaled)
X_pca_3 = pd.DataFrame (X3_pca, columns=['principal component 3'])
# regression
a=pd.concat([X_pca_1, X_pca_2, X_pca_3], axis=1)
B = data['Global GDP (trillion US dollars)', 'Population (100 million people)', 'Research and development
        expenditure (R&D) (100 million yuan)', 'Employed population (100 million people)']]
X=pd.concat([a,b],axis=1)
Y = data['GDP (trillion yuan)']
# Add a constant term
X = sm.add_constant(X)
# Build a linear regression model
model = sm.OLS(y, X).fit()
print(model.summary())
```

Task 2: Ridge regression and summing autoregressive moving average model (ARIMA)

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from sklearn.linear_model import RidgeCV
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
```

```
# reading data
Data = pd.read_excel('Guangdong-Hong Kong-Macao Greater Bay Area.xlsx')
Prepare the data
X1 = data['GDP of the Guangdong-Hong Kong-Macao Greater Bay Area (US $ trillion)', 'GDP of the
        Guangdong-Hong Kong-Macao Greater Bay Area (US $ trillion)', 'GDP of the Guangdong-Hong
        Kong-Macao Greater Bay Area (US $)']]
X2 = data['export value (trillion US dollars)', 'total import and export value (trillion US dollars)', 'import
value (trillion US dollars)']]
X3 = data['Length of transportation network in Guangdong-Hong Kong-Macao Greater Bay Area (km)', 'Total
volume of logistics in Guangdong-Hong Kong-Macao Greater Bay Area (100 million tons)',
        Infrastructure investment in the Guangdong-Hong Kong-Macao Greater Bay Area (trillion yuan)']]
X1_scaled = StandardScaler().fit_transform(X1)
X2_scaled = StandardScaler().fit_transform(X2)
X3_scaled = StandardScaler().fit_transform(X3)
# PCA
pca = PCA (n_components=1) # Select the appropriate number of components
X1_pca = pca.fit_transform(X1_scaled )
X_pca_1 = pd.DataFrame (X1_pca, columns=['principal component 1'])
X2_pca = pca.fit_transform(X2_scaled)
X_pca_2 = pd.DataFrame (X2_pca, columns=['principal component 2'])
X3_pca = pca.fit_transform(X3_scaled)
X_pca_3 = pd.DataFrame (X3_pca, columns=['principal component 3'])
# regression
a=pd.concat([X_pca_1, X_pca_2, X_pca_3], axis=1)
B = data['Global GDP (trillion US dollars)', 'Population (100 million people)', 'Research and development
        expenditure (R&D) (100 million yuan)', 'Employed population (100 million people)']]
X=pd.concat([a,b],axis=1)
Y = Data['GDP (trillion yuan)']
Add a constant term
X = sm.add_constant(X)
# Build a linear regression model
model = sm.OLS(y, X).fit()
print(model.summary())

# calculate
Historical data = data['Global GDP (trillion US dollars)', 'Population (100 million people)', 'Research and
                development expenditure (R&D) (100 million yuan)', 'Employed population (100
                million people)']]

growth_rates = historical_data.pct_change (mean () * 0.8) # Calculate the average growth rate
# Get the last value of the current independent variable
latest_values = historical_data.iloc[-1]
# Generate the estimated values of the independent variables for the next 10 years
future_data = []
for i in range(1, 11):
   future_year = latest_values * (1 + growth_rates)**i
   future_data.append(future_year)
# Convert future data to a DataFrame
future_data = pd.DataFrame(future_data)
Standardize and apply PCA
scaler_X1 = StandardScaler().fit(X1)
```

```
scaler_X2 = StandardScaler().fit(X2)
scaler_X3 = StandardScaler().fit(X3)
pca_X1 = PCA(n_components=1).fit(X1_scaled)
pca_X2 = PCA(n_components=1).fit(X2_scaled)
pca_X3 = PCA(n_components=1).fit(X3_scaled)
# Use previous standardization and PCA to transform future data
X1_future_scaled = scaler_X1.transform(X1)
X2_future_scaled = scaler_X2.transform(X2)
X3_future_scaled = scaler_X3.transform(X3)

X1_future_pca = pca_X1.transform (X1_future_scaled) # X1 principal component
X2_future_pca = pca_X2.transform (X2_future_scaled) # X2 principal component
X3_future_pca = pca_X3.transform(X3_future_scaled) # X3 principal component
# Merge principal components and future estimated variables
X_future = pd.concat([pd.DataFrame(X1_future_pca, columns=[' 主成分 1'])
                pd.DataFrame (X2_future_pca, columns=['principal component 2']),
                pd.DataFrame (X3_future_pca, columns=['principal component 3']),
                            future_data], axis=1)
# Add a constant term
X_future = sm.add_constant(X_future)
Use a linear regression model to predict future GDP
y_future = model.predict(X_future)
y_future_10 = y_future[:10]
# Generate future years
FutureYears = np.arange (data['Year'].iloc[-1]+1, data['Year'].iloc[-1]+11)
predictions = model.predict(X)
# Draw actual values, predicted values, and future predictions
plt.figure(figsize=(12, 6))
plt.plot (data['year'], y, label= 'actual value', marker= 'o', color= 'b') # Actual value
plt.plot (data['Year'], predictions, label= 'Historical prediction', marker= 'x', linestyle= '--', color= 'orange') #
Historical prediction
plt.plot (future_years, y_future_10, label= 'Future prediction', marker= 's', linestyle= '-', color= 'green') #
Future prediction
Add graphic titles and labels
plt.title (actual value, historical forecast value and future 10-year forecast value comparison)
plt.xlabel (Year)
plt.ylabel (GDP (trillion yuan)')
plt.legend()
plt.grid(True)
plt.show()


# ridge regression
from sklearn.linear_model import RidgeCV
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
# Select the independent and dependent variables
        'World GDP (in trillion dollars),'
        Length of transportation network in Guangdong-Hong Kong-Macao Greater Bay Area (km)
```

```
        Population (in 100 million),
        “Research and development expenditure (R&D) (in billion yuan)”,
        Total logistics volume of Guangdong-Hong Kong-Macao Greater Bay Area (100 million tons)
        ‘Employed population (100 million)’,
        ‘Export value (in trillion dollars)’,
        ‘Value of imports and exports (in trillion US dollars)’,
        [Imports (in trillion dollars)’]]
Y = Data[‘GDP (trillion yuan)’]
# Split the training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Use ridge regression to select the appropriate λ value
ridge = RidgeCV(alphas=np.logspace(-6, 6, 13), store_cv_values=True)
ridge.fit(X_train, y_train)
# Print the optimal λ value
Print (f ‘Optimal λ value: {ridge.alpha_}’)
# calculate
y_pred = ridge.predict(X_test)
# Visualize actual and predicted values
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], ‘k--’, lw=2)
plt.xlabel (Actual value)
plt.ylabel (predicted value)
plt.title (Actual vs. forecast)
plt.grid()
plt.show()
# Print the model coefficients
Print (“Model coefficients:”, ridge.coef_)

Model evaluation
from sklearn.metrics import r2_score
r2 = r2_score(y_test, y_pred)
print(r2)

# Ridge regression predicts the future
# Generate future independent variable data (assuming a simple growth rate forecast)
future_years = np.arange(2024, 2034)
future_X = pd.DataFrame(index=future_years)
# Calculate the average annual growth rate
growth_rates = X.pct_change().mean()
last_values = X.iloc[-1]
# Generate future independent variable data
```

```
for column in X.columns:
    future_values = [last_values[column] * (1 + growth_rates[column]) ** (year - 2023) for year in future_years]
    future_X[column] = future_values
# Use the ridge regression model to make predictions for the next 10 years
future_y_pred = ridge.predict(future_X)
# Combine the historical fit and future forecast into a continuous line
CombinedYears = np.concatenate([data[‘Year’].values, futureYears])
combined_gdp_pred = np.concatenate([y_pred_historical, future_y_pred])
```

```
Create a chart
plt.figure(figsize=(12, 6))
# Draw the actual historical GDP
plt.plot (data['Year'], y, label= 'Real GDP (trillion RMB)', marker= 'o', color= 'blue')
# Plot the fitted and predicted values of the ridge regression model as a continuous line
plt.plot (combined_years, combined_gdp_pred, label= 'ridge regression fit + GDP prediction', linestyle= '-',
color= 'green')
Add legends and labels
plt.title (Real GDP of Guangdong-Hong Kong-Macao Greater Bay Area and Ridge regression prediction trend)
plt.xlabel (Year)
plt.ylabel (GDP (trillion yuan)')
plt.legend()
plt.grid()
plt.xticks(np.arange (min(data['year']), max(future_years) + 1,1) # Set the x-axis scale
plt.tight_layout()
plt.show()

# time series model
X = data['year']
Y = Data['GDP (trillion yuan)']
from statsmodels.tsa.stattools import adfuller
# Perform the ADF test
adf_test = adfuller(y)
print('ADF Statistic:', adf_test[0])
print('p-value:', adf_test[1])
# If the p-value is greater than 0.05, the data is nonstationary

# first difference
y_diff = y.diff().dropna()
# Draw the time series again after the difference
plt.figure(figsize=(10, 6))
plt.plot (x[1:], y_diff, label= 'Differenced GDP')
plt.xlabel (Year)
plt.ylabel (Differenced GDP)
plt.title (Time series of GDP after differential)
plt.legend()
plt.grid(True)
plt.show()

# Perform ADF test again to confirm the stationarity
adf_test_diff = adfuller(y_diff)
Print ("Differenced ADF Statistic:",adf_test_diff[0])
print('p-value:', adf_test_diff[1])
```

```
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
# Calculate the reasonable maximum lag order
max_lags = min(10, len(y_diff) // 2 - 1)
# Draw ACF and PACF graphs
fig, axes = plt.subplots(1, 2, figsize=(16, 6))
plot_acf(y_diff, ax=axes[0], lags=max_lags)
```

```
plot_pacf(y_diff, ax=axes[1], lags=max_lags)
plt.show()
from statsmodels.tsa.arima.model import ARIMA
Model = ARIMA (y, order=(1,1,1)) # Original data is directly used with ARIMA (1,1,1)
model_fit = model.fit()
residuals = model_fit.resid
# Output model summary
print(model_fit.summary())
# Predict the next 10 years
forecast = model_fit.forecast(steps=10)
forecast_years = np.arange(y.index[-1] + 1, y.index[-1] + 11)
# Plot actual and predicted values
plt.figure(figsize=(12, 6))
plt.plot(y.index, y, label= 'Actual value', marker= 'o')
plt.plot (forecast_years, forecast, label= 'Future Forecast', marker= 'x', linestyle= '--', color= 'orange')
plt.xlabel (Year)
plt.ylabel (GDP (trillion yuan)')
plt.title (Real GDP versus projected GDP for the next 10 years)
plt.legend()
plt.grid(True)
plt.show()

# white noise
from statsmodels.stats.diagnostic import acorr_ljungbox
residuals = model_fit.resid
ljung_box_results = acorr_ljungbox(residuals, lags=[10], return_df=True)
Print ("Ljung-Box white noise test results:")
print(ljung_box_results)
# Different model predictions are shown on the same graph
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.statespace.sarimax import SARIMAX
```

```
data = pd.read_excel('Guangdong-Hong Kong-Macao Greater Bay Area.xlsx') # Ensure the path is correct
data.dropna(inplace=True)
# Fit the SARIMA model
Model = SARIMAX (data['GDP (trillion yuan)'], order=(1,1,1), seasonal_order=(1,1,1,12)) # Adjust
parameters according to the data
model_fit = model.fit(disp=False)
Get the fitted value of the model for historical data
historical_fitted_values = model_fit.fittedvalues
# Predict the next 10 years
forecast_steps = 10
forecast = model_fit.forecast(steps=forecast_steps)
# Construct future forecast years
forecast_years = np.arange (data[' 年份 '].iloc[-1]+1, data[' 年份 '].iloc[-1]+1+forecast_steps)
# Combine historical years and forecast years, as well as corresponding historical fits and future forecasts
CombinedYears = np.concatenate([data['Year'].values, forecastYears])
combined_gdp_estimates = np.concatenate([historical_fitted_values, forecast])
# plot
plt.figure(figsize=(12, 6))
```

```
# Draw the actual GDP
plt.plot (data[‘Year’], data[‘GDP (trillion RMB)’], label= ‘Real GDP (trillion RMB)’, marker= ‘o’)

# Plot the historical fit and future forecast values of the SARIMA model as a continuous line
plt.plot (forecast_years, forecast, label= ‘ARIMA prediction of GDP’, linestyle= ‘:’, marker= ‘s’)
plt.plot (future_years, y_future_10, label= ‘multiple regression’, linestyle= ‘--’, marker= ‘x’) # Future forecast value
plt.plot (future_years, future_y_pred, label= ‘Ridge regression prediction of GDP’, linestyle= ‘-.’, marker= ‘^’)
Add legends and labels
plt.title (Real GDP of Guangdong-Hong Kong-Macao Greater Bay Area, ridge regression simulation, multiple regression and ARIMA model prediction trend)
plt.xlabel (Year)
plt.ylabel (GDP (trillion yuan)’)
plt.legend()
plt.grid()
plt.xticks(np.arange (min (combined_years), max (forecast_years) + 1,1) # Set the x-axis scale
plt.tight_layout()
```

Task 3: Time series and ridge regression

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams[‘font.sans-serif’]=[‘SimHei’]
plt.rcParams[‘axes.unicode_minus’]=False
Data = pd.read_excel(‘Tokyo Bay Area.xlsx’)

X = data[‘year’]
Y = Data[‘Tokyo GDP (trillion yen)’]
from statsmodels.tsa.stattools import adfuller
# Perform the ADF test
adf_test = adfuller(y)
print(‘ADF Statistic:’, adf_test[0])
print(‘p-value:’, adf_test[1])

# first difference
y_diff = y.diff().dropna()
# Draw the time series again after the difference
plt.figure(figsize=(10, 6))
plt.plot (x[1:], y_diff, label= ‘Differenced GDP’)
plt.xlabel (Year)
plt.ylabel (Differenced GDP)
plt.title (Time series of GDP after differential)
plt.legend()
plt.grid(True)
plt.show()

# Perform ADF test again to confirm the stationarity
adf_test_diff = adfuller(y_diff)
Print (“Differenced ADF statistic:”,adf_test_diff[0])
print(‘p-value:’, adf_test_diff[1])
```

```
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
# Calculate the reasonable maximum lag order
max_lags = min(10, len(y_diff) // 2 - 1)
# Draw ACF and PACF graphs
fig, axes = plt.subplots(1, 2, figsize=(16, 6))
plot_acf(y_diff, ax=axes[0], lags=max_lags)
plot_pacf(y_diff, ax=axes[1], lags=max_lags)
plt.show()

import statsmodels.api as sm
# Use ARIMA (1,1,1) model fitting
model_1_1_1 = sm.tsa.ARIMA(y, order=(1, 1, 1))
result_1_1_1 = model_1_1_1.fit()

# Use ARIMA (2,1,1) model fitting
model_2_1_1 = sm.tsa.ARIMA(y, order=(2, 1, 1))
result_2_1_1 = model_2_1_1.fit()

# Output a summary of the model results, including AIC and BIC values
Print ("ARIMA (1,1,1) model:")
print(result_1_1_1.summary())
Print ("ARIMA (2,1,1) model:")
print(result_2_1_1.summary())

# Compare AIC and BIC values
Print ("Model comparison:")
print(f'ARIMA(1,1,1) - AIC: {result_1_1_1.aic}, BIC: {result_1_1_1.bic}")
print(f'ARIMA(2,1,1) - AIC: {result_2_1_1.aic}, BIC: {result_2_1_1.bic}")

# According to AIC and BIC, select the model with lower AIC and BIC
if result_1_1_1.aic < result_2_1_1.aic and result_1_1_1.bic < result_2_1_1.bic:
Print ("\n It is recommended to use the ARIMA (1,1,1) model")
   best_model = result_1_1_1
else:
Print ("\n It is recommended to use the ARIMA (2,1,1) model")
   best_model = result_2_1_1
```

```
# Predict the next 10 time points
forecast = best_model.get_forecast(steps=10)
forecast_values = forecast.predicted_mean
conf_int = forecast.conf_int()
# Visualize the prediction results
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
plt.plot (y, label= 'Historical data')
plt.plot (range(len(y), len(y) + 10), forecast_values, label= 'predicted value', color= 'orange')
plt.fill_between(range(len(y), len(y) + 10), conf_int.iloc[:, 0], conf_int.iloc[:, 1], color='gray', alpha=0.3)
plt.xlabel (Time)
plt.ylabel (Tokyo GDP (trillion yen)')
plt.title (Tokyo GDP Forecast for the next 10 years)
```

```
plt.legend()
plt.legend()
plt.show()

from statsmodels.stats.diagnostic import acorr_ljungbox
ljung_box_results = acorr_ljungbox(result_2_1_1.resid, lags=[10], return_df=True)
# Output the test results
Print ("Ljung-Box white noise test results:")
print(ljung_box_results)
```

Task 1: correlation coefficient matrix, VIF, principal component analysis and multiple linear regression

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
import numpy as np
plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus']=False
# reading data
Data = pd.read_excel('Guangdong-Hong Kong-Macao Greater Bay Area.xlsx')
# Calculate the correlation coefficient matrix
correlation_matrix = data.corr()
# Draw a heat map of the correlation coefficient matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix,annot=True,fmt=".2f", cmap="coolwarm",cbar=True, square=True)
plt.title("Correlation Matrix")

from statsmodels.stats.outliers_influence
import variance_inflation_factor
Prepare the data
X = data['Guangdong-Hong Kong-Macao Greater Bay Area's First Gross Domestic Product (USD billion) ',' Guangdong-Hong Kong-Macao Greater Bay Area's Second Gross Domestic Product (USD billion)', 'Guangdong-Hong Kong-Macao Greater Bay Area's Third Gross Domestic Product (USD billion) ',' Guangdong-Hong Kong-Macao Greater Bay Area's Infrastructure Investment (RMB trillion)', 'Global GDP Total (USD billion)', 'Guangdong-Hong Kong-Macao Greater Bay Area's Transportation Network Length (km) ',' Population (100 million) ',' Research and Development (R&D) Funding (Billion Yuan) ',' Employed Population (100 million) ',' Export Value (USD billion) ',' Total Import and Export Value (USD billion) ',' Import Value (USD billion) ',' Guangdong-Hong Kong-Macao Greater Bay Area's Total Logistics Volume (billion tons)']])'
# count VIF
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print(vif_data)
# linear regression
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
Prepare the data
X1 = data['GDP of the Guangdong-Hong Kong-Macao Greater Bay Area (US $ trillion)', 'GDP of the
        Guangdong-Hong Kong-Macao Greater Bay Area (US $ trillion)', 'GDP of the Guangdong-Hong
        Kong-Macao Greater Bay Area (US $)']]
```

```
X2 = data[‘export value (trillion US dollars)’, ‘total import and export value (trillion US dollars)’, ‘import value (trillion US dollars)’]]
X3 = data[‘Length of transportation network in Guangdong-Hong Kong-Macao Greater Bay Area (km)’,
          ‘Total logistics volume in Guangdong-Hong Kong-Macao Greater Bay Area (100 million tons)’,
          ‘Infrastructure investment in Guangdong-Hong Kong-Macao Greater Bay Area (trillion RMB)’]]
X1_scaled = StandardScaler().fit_transform(X1)
X2_scaled = StandardScaler().fit_transform(X2)
X3_scaled = StandardScaler().fit_transform(X3)
# PCA
pca = PCA (n_components=1) # Select the appropriate number of components
X1_pca = pca.fit_transform(X1_scaled )
X_pca_1 = pd.DataFrame (X1_pca, columns=[‘principal component 1’])
X2_pca = pca.fit_transform(X2_scaled)
X_pca_2 = pd.DataFrame (X2_pca, columns=[‘principal component 2’])
X3_pca = pca.fit_transform(X3_scaled)
X_pca_3 = pd.DataFrame (X3_pca, columns=[‘principal component 3’])
# regression
a=pd.concat([X_pca_1, X_pca_2, X_pca_3], axis=1)
B = data[‘Global GDP (trillion US dollars)’, ‘Population (100 million people)’, ‘Research and development
          expenditure (R&D) (100 million yuan)’, ‘Employed population (100 million people)’]]
X=pd.concat([a,b],axis=1)
Y = Data[‘GDP (trillion yuan)’]
# Add a constant term
X = sm.add_constant(X)
# Build a linear regression model
model = sm.OLS(y, X).fit()
print(model.summary())
```