Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 2, 2025

Economic forecast of Guangdong-Hong Kong-Macao Greater Bay Area

-- Empirical analysis based on PCA-ARIMA-LSTM mixed model

Jintian Zhang, Yongtong Guan*, Jingqi Zhao, Dianjun Xu, Shaoyi Wu

[1] D School of Mathematics, Foshan University, Foshan 528225, China

* **Corresponding author**: Yongtong Guan (Email: 13724668409@163.com)

Abstract: This study presents an innovative framework for analyzing and forecasting economic development in the Guangdong-Hong Kong-Macao Greater Bay Area (GBA), a key driver of China's economic growth. The research makes three significant contributions: First, it develops a novel multi-method analytical approach combining multiple linear regression, Granger causality tests, and principal component analysis (PCA) to systematically identify the most influential growth factors - particularly highlighting logistics networks and technological inputs as critical leverage points for policy intervention. Second, the study pioneers a hybrid modeling strategy that integrates ARIMA and LSTM techniques, demonstrating superior predictive performance for the GBA's 5-10 year economic trajectory. Third, the findings provide actionable insights for policymakers, suggesting that targeted investments in technology infrastructure and logistics optimization can accelerate regional development while offering a replicable model for other emerging mega-regions. The proposed framework not only enhances understanding of the GBA's growth dynamics but also establishes a new benchmark for regional economic analysis.

Keywords: Principal Component Analysis, Arima Model, Lstm Model, Multiple Linear Regression.

1. Introduction

As one of the regions with the highest degree of openness and strongest economic vitality in China, the GBA has an economic aggregate accounting for one-ninth of the national total and is a core growth pole for promoting the dual circulation strategy. However, the development of the GBA economy does not exist independently but is jointly influenced by multiple factors such as population structure, technological level, and logistics speed. These factors interact with each other to form a complex economic system, making the prediction of the economic development of the Bay Area extremely challenging. Li Xiaomei[1] has conducted research on the driving force of the digital economy for the high-quality development of the Guangdong-Hong Kong-Macao Greater Bay Area, and Wang Jingtian[2] has studied the mechanism and countermeasures for the digital economy to enhance the resilience of the industrial chain in the Guangdong-Hong Kong-Macao Greater Bay Area, etc. Based on their experiences, they are of great significance to study the future economic trend of the Guangdong-Hong Kong-Macao Greater Bay Area.

In the field of economic forecasting, time series models and continuing learning methods each have their advantages. The ARIMA model, as a classic statistical method, is good at handling linear trend and seasonal data, but has insufficient response to nonlinear images. LSTM neural network. The vanishing gradient problem was solved through the energy control mechanism, and it performed well in the prediction of complex sequences in fields such as finance and energy. However, it relies on a large amount of data and has weak interpretability. Principal component analysis, which simplifies high-dimensional data structures through dimensionality reduction techniques, has been used to screen key economic indicators. As a result, a single model has gradually emerged in existing research. In response to the above problems, the hybrid model has become a new trend. For instance, the combined model integrating ARIMA and neural networks has significantly enhanced robustness in the economic forecasting of the GBA [3]. However, there are still deficiencies in the specific research for the GBA [4]. Especially, a mature solution has not yet been formed for the continuous analysis ability of effectively integrating statistical models and the nonlinear modeling ability of deep learning.

By constructing an economic forecasting model based on multi-model integration, this article deeply explores the key factors of economic development in the Guangdong-Hong Kong-Macao Greater Bay Area and its potential impact on future economic trends. Specific research objectives include: The key driving factors are identified through PCA dimensionality reduction to solve the problem of high-dimensional data redundancy. The long-term trend and periodicity of the economic series are captured by using ARIMA, and the nonlinear fluctuation characteristics are learned simultaneously through LSTM. Make up for the shortcomings of the single model, and then construct a hybrid prediction framework through residual correction and dynamic weighting to improve the prediction accuracy in complex scenarios and reveal the economic trend of the Guangdong-Hong Kong-Macao Greater Bay Area in the next ten years.

2. Data and Model

2.1. Data acquisition and preprocessing

2.1.1. Data collection

The data covers research and development funds, total logistics volume, total population, total GDP, total import and export trade volume, etc. of the Guangdong-Hong Kong-Macao Greater Bay Area. The data are sourced from China

Statistical Yearbook, the National Bureau of Statistics, etc.

2.1.2. Linear interpolation processes data.

For the obtained data, data preprocessing work is required. For the large number of missing values existing in the collected data, this study choose linear interpolation for filling. The basic idea of linear interpolation is to linearly estimate the position of missing values through known adjacent points, maintaining the overall trend of the data. It is particularly suitable for situations where the changes are stable among known data points. Outlier handling: For the economy that has been developing rapidly all along, there have been data fluctuations over the past 24 years, with certain outliers. In 2020, the rampant spread of the COVID-19 pandemic dealt an unprecedented blow to the global economy, causing multiple industries to come to a standstill, supply chains to break, and economic activities to decline sharply. In view of the outliers presented in the data and in light of the actual situation, the study have decided not to handle them specially.

2.1.3. Descriptive analysis of data

The Guangdong-Hong Kong-Macao Greater Bay Area has witnessed rapid growth, especially in the output value of the secondary industry, which has increased rapidly, reflecting the characteristics of China's economic growth and industrialization. This indicates that the Greater Bay Area has developed rapidly over the past 24 years, especially with a significant increase in investment in the secondary and tertiary industries.

2.2. Method Introduction

① Principal Component Analysis method (PCA): Principal component analysis converts high-dimensional data into a few unrelated principal components through linear transformation while retaining the maximum variance information.

② Autoregressive Integral Moving Average Model (ARIMA): ARIMA combines autoregressive (AR), difference (I), and moving average (MA) to model the dependencies of non-stationary sequences. It is capable of predicting future values and capturing dynamic patterns [5].

③ Long Short-Term Memory Network (LSTM): LSTM is a special recurrent neural network (RNN) that effectively solves the vanishing/exploding gradient problem in long sequence training through gating mechanisms (input gate, forgetting gate, output gate) [6].

④ PCA-ARIMA-LSTM hybrid model: PCA-ARIMA-LSTM is a hybrid model that integrates the capabilities of dimension reduction, linear modeling and nonlinear learning, which can significantly improve the prediction accuracy of high-dimensional, non-stationary and noisy time series data.

The above are the main models and methods employed in the article. Through these models and methods, analyze the factors that have the greatest influence on the Guangdong-Hong Kong-Macao Greater Bay Area and predict the future economic trend.

3. The results and analysis of the model

3.1. Dimension reduction is carried out using PCA

After the data was standardized, this study employed the Principal Component Analysis (PCA) method and successfully screened out the key factors with strong interpretive power for regional economic fluctuations from multiple variables such as scientific research, logistics, population and economy.

Correlation analysis aims to quantify the effects of various indicators of technology and logistics on the GDP of the GBA. The study calculated the correlation coefficients between various technological indicators (such as research and development funds, the number of patent applications, etc.) and the GDP of the GBA. Then, by screening out the correlation coefficients with the largest absolute values, article determined that the indicators of science and technology and logistics had the most significant impact on GDP.

3.2. Economic trend prediction based on the PCA-ARIMA-LSTM hybrid model

First, collect and extract various indicator data related to the economy of the GBA, including population, technology, logistics, international environment and the GDP of the GBA. Use the cvpartition function to randomly divide the data into the training set and the test set (80% of the training data and 20% of the test data).

3.2.1. Multiple linear regression model

The basic principle of multiple linear regression is to establish a regression model by finding the optimal linear relationship between the independent variable and the dependent variable, so as to facilitate the prediction or analysis of the changes in the dependent variable. The core lies in assuming that there is a linear relationship between the dependent variable and one or more independent variables. In linear regression, data is modeled using a linear prediction function, and the model attempts to find the best linear function so that the sum of the squares of the errors between the predicted value and the actual value of the dependent variable is minimized [7].

Table.1. Prediction and Estimation Coefficients of the multiple linear regression model

Statistics	Estimate	SE
GBA's Total volume of logistics	0.03647	0.4544
GBA's Infrastructure investment	0.45824	0.83969
GBA's Length of the transportation network	-8.6783e-06	1.3055e-05
GBA's Research and development funds	-2.4932	13.385
GBA' s number of patent applications	0.087168	0.14138

It can be seen from the table.1. that the model has an excellent fitting effect: both the r-square and the adjusted R-square are as high as 0.998, which indicates that the model can explain the vast majority of variations in the dependent variable, and the fitting effect is very good. The root mean square error is relatively small: The root mean square error is 0.218, which is a relatively small value, indicating that the model's prediction error is small and the prediction results are relatively accurate.

Then, study used Matlab and the multiple linear regression model to predict the economic trend of the Guangdong-Hong Kong-Macao Greater Bay Area in the next 5 to 10 years. The model formula is as follows:

$$y(t) = 0.049 - 0.522 * y(t-1) - 0.0 * \varepsilon(t-1) - 1.0 * \varepsilon(t-2) \quad (1)$$

The specific results are as follows:

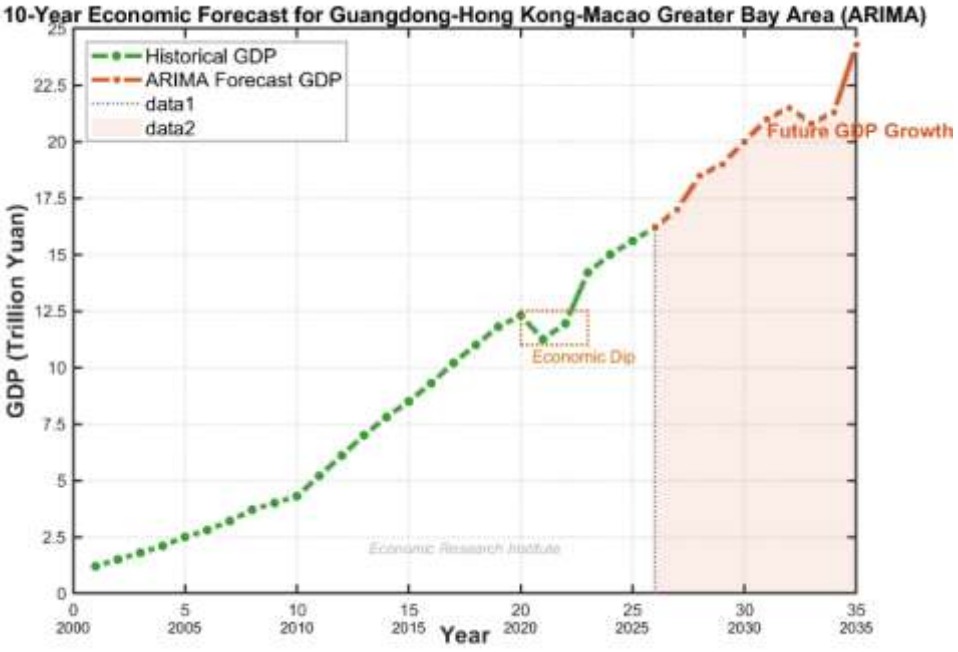


Figure 1. Economic Forecast Chart of the GBA in the Next 10 Years

This figure 1 shows the changing trends of the predicted and actual GDP values of the Guangdong-Hong Kong-Macao Greater Bay Area from 2000 to 2035. It can be seen from the picture that on the existing data points, the trend of the predicted GDP is relatively close to that of the actual GDP. This indicates to a certain extent that the prediction model may have a certain degree of accuracy. According to the predicted trend, the GDP of the Guangdong-Hong Kong-Macao Greater Bay Area is expected to continue to grow in the future, although the growth rate may slow down over time. This reflects the huge potential for the future economic development of this region. This provides a strong support for the long-term prosperity and stability of the Guangdong-Hong Kong-Macao Greater Bay Area.

3.2.2. Time series model: ARIMA model

ARIMA is a model widely used in the prediction of time series data, and is particularly suitable for processing data with time-dependent characteristics in economics, finance, etc. Based on the variable GDP of the GBA (trillion yuan), the system automatically finds the optimal parameters based on the AIC information criterion.

This figure 2. describes through the ARIMA model the comparison between the economic forecast and the actual economic situation of the Guangdong-Hong Kong-Macao Greater Bay Area over the next 10 years as predicted by the model.

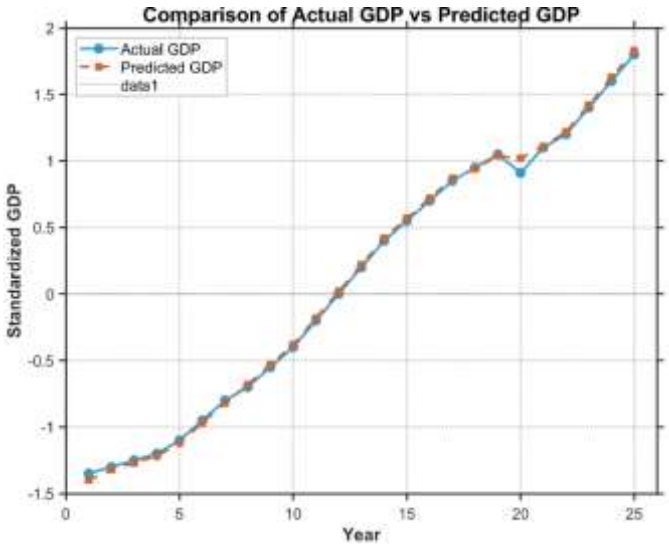


Figure 2. Comparison chart of real GDP and predicted GDP

Figure 2 conducted a comparative analysis of the real GDP and the predicted GDP. During the period from 2000 to 2010,

the growth rate of real GDP exceeded the growth rate of predicted GDP, which reflected that the economic development trend during this stage was more optimistic than expected. During the period from 2010 to 2018, the growth rate of real GDP was almost in line with that of predicted GDP, which indicates that the prediction had a relatively high accuracy during this period. However, by 2020, due to the outbreak of the COVID-19 pandemic, the real GDP experienced significant fluctuations, but study's prediction model failed to accurately capture this fluctuation.

3.2.3. Granger causality test

The core of Granger causality test lies in comparing the two models. The F-test is used to determine whether the introduction of past values of Xt can significantly improve the predictive ability for Yt. Through the Granger causality test, the study conclude that logistics is the Granger cause of GDP, but research and development funds are not the Granger cause.

3.3. Model Evaluation Metrics (MAPE and RMSE)

Study compared and analyzed the actual values, the predicted values of LSTM, the predicted values of ARIMA and the predicted values of the mixed model through Mean Absolute Percentage Error (MAPE) and root Mean square error (RMSE), and the results are as follows.

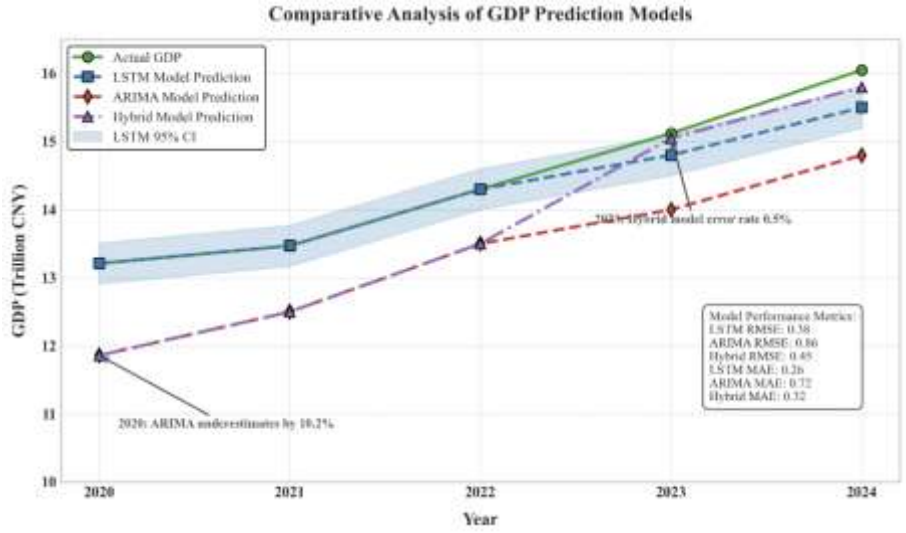


Figure 3. Compare the fitting effect of the model

It can be concluded from the figure 3 that the hybrid model performs robustably in long-term trend prediction, while LSTM and ARIMA are flexible in nonlinear fluctuations. Overall, by comparing the prediction errors of these models, it is found that the hybrid model outperforms the single ARIMA model and the LSTM model in both MAPE and RMSE.

3.4. Construction and verification of LSTM and Hybrid models

LSTM (Long Short-Term Memory Network) is an improved recurrent neural network that controls the information flow through forgetting gates, input gates and output gates, and is good at handling long sequence-dependent tasks, such as time series prediction.

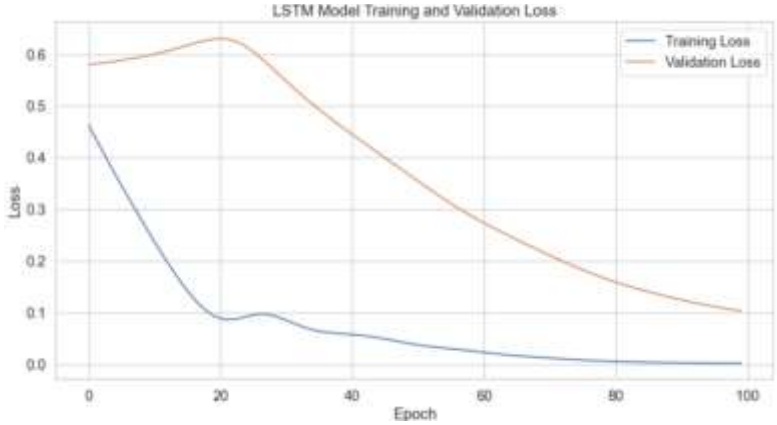


Figure 4. LSTM training process

The Training Loss and Validation Loss of the LSTM model in figure 4 both show a downward trend with the increase of training rounds (epos). Eventually, the training loss stabilizes below 0.1, indicating that the model is constantly learning and optimizing. The model does not have overfitting and has good generalization ability.

3.5. Model weighted prediction

When constructing the economic prediction model, artical integrated the advantages of three models, namely multiple linear regression, ARIMA and LSTM, and adopted the dynamic weighted average method to improve the prediction

accuracy. First, article independently trained the three models and evaluated their respective prediction performances through cross-validation. The weight distribution was dynamically adjusted based on the RMSE or MAPE metrics of the models on the validation set and the data characteristics: When the data showed obvious seasonality or trend, the weight of the ARIMA model was appropriately increased. When the data pattern is complex and the nonlinear characteristics are significant, increase the weight of the LSTM model. Meanwhile, a certain weight of the multiple linear regression model is retained to maintain the interpretability of economic indicators.

The final weighted predicted value is composed of the linear combination of the outputs of each model. This fusion method not only integrates the accuracy of ARIMA in time series prediction, the advantages of LSTM in complex pattern recognition, and the characteristics of multiple regression in economic interpretability, but also can dynamically adjust the model contribution according to the data characteristics. Thus, while maintaining the stability of the prediction, the accuracy rate can be improved, providing more comprehensive and reliable prediction results for the analysis of economic trends.

3.6. Description of future trends

The growth rate represents the growth rate of the economy between different years and is an important indicator for judging the health and trend of the economy. The annual GDP growth rate was calculated and plotted as a line graph to show the trend of GDP growth rate in the next 10 years.

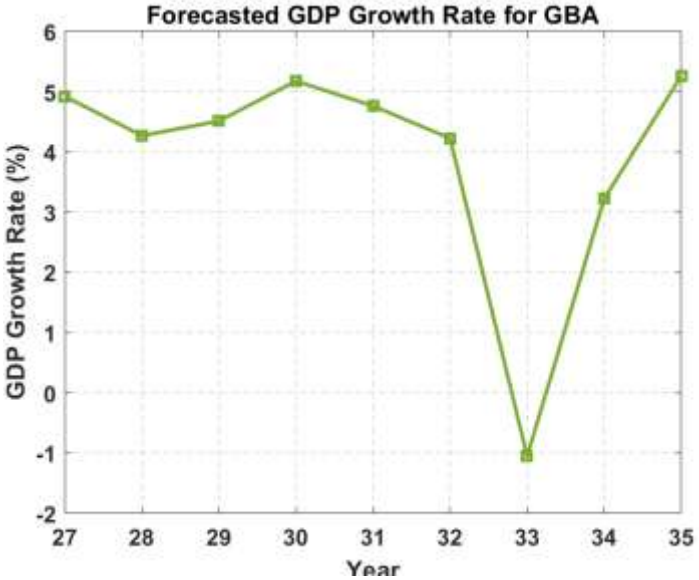


Figure 5. GDP growth rate in the next 10 years

As can be seen from the figure 5, the GDP growth rate of the GBA in the next ten years will show relatively significant volatility rather than a sustained and stable growth. The fluctuation of GDP growth rate may be influenced by a variety of factors, such as macroeconomic policies, etc. The year 2033 is a significant turning point. The decline in its growth rate may reflect the sluggish stage of the economic cycle, while the rise may indicate an economic recovery or growth period. Despite fluctuations, the overall trend may indicate the long-term growth potential of the economy in the GBA.

4. Conclusions

This study systematically constructs an economic forecasting framework through comprehensive multi-dimensional analysis, incorporating key indicators across population dynamics, technological advancement, logistics infrastructure, and international economic factors. The research employs a rigorous methodological approach, combining Principal Component Analysis (PCA) for dimensionality reduction of complex international data, ARIMA and LSTM models for time-series forecasting, supplemented by multiple linear regression and Granger causality tests to quantify factor-specific GDP impacts. The increase in technological investment and the improvement of the logistics network have the most significant positive promoting effects on economic growth.

Projections indicate the Guangdong-Hong Kong-Macao Greater Bay Area will maintain an overall growth trajectory over the next decade, though subject to periodic fluctuations stemming from global economic volatility. To capitalize on development opportunities while addressing challenges, policymakers should prioritize: demographic structure optimization, enhanced investment in scientific innovation, accelerated industrial upgrading, and deepened regional integration. Concurrently, adaptive strategies must be developed to navigate international economic uncertainties, ensuring the GBA's resilience and sustainable development.

References

[1] Li Xiaomei, Li Yongbiao, Deng Ningjun. Research on the Driving Force of Digital Economy on the High-Quality Development of Manufacturing Industry in the Guangdong-Hong Kong-Macao Greater Bay Area [J]. Special Economic Zone Economy, 2025, (02):21-24.

[2] Wang Jingtian. Research on the Mechanism and Countermeasures of Digital Economy Empowering the Resilience Enhancement of the Industrial Chain in the Guangdong-Hong Kong-Macao Greater Bay Area [J]. Supply chain management, 2024, 5 (12) : 64-74. The DOI: 10.19868 / j.carol carroll nki gylgl. 2024.12.006.

[3] Weerakoon A, Assadi M. Artificial Neural Network (ANN) driven Techno-Economic Predictions for Micro Gas Turbines (MGT) based Energy Applications[J]. Energy and AI, 2025, 20100483-100483.

[4] Zhou Z, Jiang P, Chen S. Decomposition of Carbon Emission Drivers and Carbon Peak Forecast for Three Major Urban Agglomerations in the Yangtze River Economic Belt[J]. Sustainability, 2025, 17(6): 2689-2689.

[5] Qiao Heng, Wu Huahua, Xue Jianbo, et al. Annual Sales Forecasting Method Based on multiple Linear Regression [J]. Business Review, 2024, 10 (06): 111-114.

[6] JIAWK, SUNML, LIANJ, etal. Feature dimensionality reduction: a review[J]. Complex & Intelligent Systems, 2022(8): 2663-2693.

[7] C.A. A, H. A, K. A I, et al. An Exploration of Nigerian Exchange Rate Using ARIMA Model[J]. Asian Journal of Probability and Statistics, 2025, 27(1): 69-80.