Simen Owen Academic Proceedings Series

Vol. 1 2025

Article **Open Access**

Forecasting University Admission Rates in China's Gaokao Using ARIMA and ARIMAX Models

Jiayi Yang [1,*]

[1] Arizona State University, Tempe, AZ 85281, USA
* Correspondence: Jiayi Yang, Arizona State University, Tempe, AZ 85281, USA

Abstract: This study employs ARIMA, ARIMAX, VAR, and GLM models to forecast China's Gaokao admission rates using annual data from 1977 to 2024. The analysis focuses on the admission rate as the dependent variable, with GDP, newborn population, and policy dummies serving as exogenous factors. The results demonstrate that the ARIMAX model, which incorporates economic and demographic variables alongside PCA-based dimensionality reduction, outperforms the other models in terms of both accuracy and interpretability. Predictions indicate an admission rate of 78.47% for 2025 and 83.55% for 2029. The findings emphasize the crucial role of external factors in forecasting and offer valuable methodological and policy insights for higher education planning.

Keywords: gaokao admission rate; ARIMAX model; educational forecasting

1. Introduction

Since its restoration in 1977, China's Gaokao (National College Entrance Examination) has become the most pivotal higher education selection system in the country. It has not only shaped the enrollment pathways for millions of students but also had a profound impact on the allocation of educational resources and social equity. As higher education continues to expand and social and economic structures evolve, the Gaokao admission rate (i.e., the ratio of the number of admissions to the number of applicants) has emerged as a crucial indicator for assessing educational equity and resource accessibility.

Forecasting trends in admission rates is essential for various stakeholders. It provides a foundation for the government to devise policies on enrollment scales and educational investments, assists universities in optimizing resource allocation, and offers valuable decision-making insights for both families and candidates. Existing research highlights the practical value of predicting admission scores or rates based on Gaokao data. For instance, a mathematical model was developed to predict Gaokao admission scores, serving as a reference for candidates when filling out their applications [1]. A case study conducted in Shanxi Province compared the performance of different models in predicting admission scores, emphasizing the critical role of model selection in shaping prediction outcomes [2]. More broadly, a review of the application of time series analysis in educational research emphasized that methods such as ARIMA and ARIMAX are particularly effective at capturing the dynamic trends of educational indicators and the influence of external variables [3].

From a methodological perspective, research has outlined the construction and diagnostic procedures for ARIMA models using the Box-Jenkins method to predict higher education enrollment [4], while studies have highlighted the significant impact of time series length on the accuracy of ARIMA model predictions, providing useful insights for modeling long-term data [5]. Therefore, in the context of the Chinese college entrance

Received: 11 August 2025
Revised: 22 August 2025
Accepted: 16 October 2025
Published: 20 October 2025

examination, integrating historical data with socioeconomic external variables to forecast admission rates not only addresses gaps in existing research but also offers a scientifically grounded quantitative basis for educational policy formulation and resource distribution.

2. Literature Review

2.1. Overview of Time Series Models in Education Forecasting

Time series methods have been widely adopted in education forecasting, particularly for analyzing trends in enrollment scales and admission patterns. In this domain, a study demonstrated the use of ARIMA models, specifically employing the Box-Jenkins method, to predict higher education enrollment trends [6]. Similarly, the SARIMA model was applied to study international undergraduate student enrollment, showing its capacity to effectively account for both seasonal and trend dynamics [7]. In a related study, various ARIMA models were used to forecast student enrollment, underscoring the practical applications of time series models in education planning and policy [8].

These studies confirm that univariate time series models perform well in predicting education-related indicators. However, in cases where external driving factors significantly impact educational outcomes, the ARIMAX model emerges as a superior alternative. For instance, in forecasting Taiwan's STEM project enrollment, the inclusion of exogenous variables substantially improved model accuracy and fit [9]. Furthermore, a "regression with ARIMA errors" methodology was proposed, providing a comprehensive framework for selecting external variables, identifying lag structures, and conducting residual diagnostics [10]. Additionally, the use of Principal Component Analysis (PCA) offers valuable insights into handling multicollinearity between educational and economic variables, suggesting techniques for dimensionality reduction and ensuring balanced explanatory power [11]. These advances create a robust foundation for incorporating external factors such as GDP, birth rates, and policy dummy variables into the ARIMAX model.

2.2. Multivariate Models and Gaps in College Admission Rate Forecasting

Alongside univariate models, multivariate approaches also play a crucial role in interdisciplinary research combining macroeconomics and education. The vector autoregression (VAR) model has been shown to capture dynamic interactions between multiple time series, making it a powerful tool for policy analysis and forecasting [12]. Further details on the application of VAR models in empirical forecasting are provided in related lecture notes, while studies have highlighted the advantages of VARMA models in enhancing forecasting flexibility [13].

However, much of the existing literature on forecasting in the context of China's college entrance examination (Gaokao) primarily focuses on predicting admission scores rather than overall national admission rates. For example, research on predicting scores based on candidate rankings has been valuable for prospective students, but it does not comprehensively address the long-term trends and external factors affecting admission rates [14]. This presents a clear gap in the current research landscape, particularly in terms of systematically comparing the performance of ARIMA, ARIMAX, and VAR models within a unified framework for admission rate prediction [15].

This study aims to address this gap by developing a more comprehensive understanding of the dynamics of China's college entrance examination admission rates. It will systematically compare various time series and econometric models to enhance the accuracy of forecasts and provide a solid quantitative basis for policy development [16]. The study will not only apply the classic ARIMA model but also introduce the ARIMAX model, which incorporates external driving factors. In addition, VAR and Binomial GLM will be utilized as comparative models, facilitating a systematic comparison of multiple approaches. This integration allows for a comprehensive evaluation of the models'

predictive and explanatory capabilities, contributing valuable insights to the field of education forecasting [17].

3. Data and Methodology

3.1. Data Source and Processing

The core explanatory variable in this study is the China College Entrance Examination Admission Rate (CEEAR), defined as the ratio of the number of admitted students (CEEA) to the number of applicants (CEEP) [18]. The data covers annual observations from 1977 to 2024, systematically capturing the long-term evolution of the admission rate since the reinstatement of the college entrance examination. The data are primarily sourced from authoritative and reliable institutions, including the official website of the Ministry of Education of the People's Republic of China, the National Bureau of Statistics database, and the China Education Statistical Yearbook, ensuring both data reliability and comparability [19].

To more comprehensively explain fluctuations in the admission rate, this study incorporates several exogenous variables into the ARIMAX model. First, the newborn population (NBP) is used as a proxy for the number of eligible students, reflecting the pressure of birth fluctuations on the potential applicant pool. Second, the gross domestic product (GDP) serves as a proxy for macroeconomic conditions, indicating the long-term support of economic growth for educational investment and the expansion of higher education. Finally, the enrollment expansion policy dummy variable (UEE) is included to capture the structural impact of the university enrollment expansion policy, which has been in place since 1999. This variable takes a value of 1 from 1999 onward and 0 otherwise. By integrating these three external factors-demographic, economic, and policy-this study aims to more accurately identify the driving mechanisms behind changes in the admission rate.

In the data processing phase, all original series were first logarithmically transformed to address heteroskedasticity and stabilize fluctuations. The enhanced Dickey-Fuller (ADF) test was then applied to assess the stationarity of the series. Variables exhibiting trends or non-stationarity were subjected to first-order differencing to meet the modeling requirements. To account for dimensional differences among variables, this study standardized variables such as GDP and NBP, thereby enhancing the robustness of the estimates and minimizing the impact of differing numerical scales on the results. Finally, to ensure scientific rigor and comparability in predictions, the sample was divided into two periods: 1977-2018 as the training set for model estimation and fitting, and 2019-2024 as the test set for out-of-sample predictions and performance evaluation. This division enables the testing of the model's effectiveness and generalizability in predicting future scenarios.

As shown in Figure 1, the key variables-GDP, CEEA, CEEP, and NBP-exhibit distinct long-term patterns, with 1999 marking a structural turning point due to the implementation of the enrollment expansion policy. Figure 2 further illustrates the steady upward trend of CEEAR since 1977, highlighting its relative stability compared to the more volatile demographic and economic indicators. Together, these descriptive plots provide crucial context for understanding the dynamic factors driving fluctuations in the admission rate and set the stage for the subsequent modeling analysis.

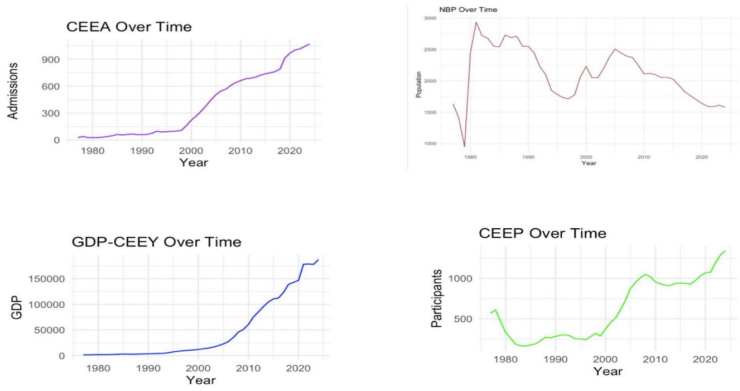

Figure 1. Data trend charts of CEEA, NBP, GDP-CEEY, and CEEP.

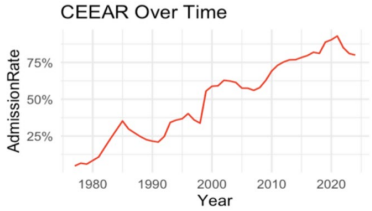

Figure 2. Data trend charts of CEEAR.

3.2. Research Methods

3.2.1. ARIMA Models

The ARIMA model is a widely used statistical method for time series forecasting. Its core principle involves transforming non-stationary series into stationary series through differencing, while capturing short-term dependencies by combining autoregressive (AR) and moving average (MA) structures. The typical modeling process includes several steps: first, the original series is made stationary through logarithmic transformation and differencing tests; next, the potential model order is determined using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots; candidate models are then compared using information criteria such as AIC and BIC to identify the optimal order combination; parameter estimation is performed, often through maximum likelihood estimation; and finally, residual diagnostics (such as the Ljung-Box test and residual autocorrelation analysis) are applied to assess the model's fit and validity. This process has been validated across various fields. For instance, the entire process of stationarization, order determination, estimation, and residual diagnostics was applied when predicting infant mortality, ensuring the reliability of the results.

In this study, two different ARIMA modeling approaches are employed based on the college entrance examination admission rate (CEEAR), a ratio-based indicator constrained to range between 0 and 1. The first approach, Logit-Transformed ARIMA, involves applying a logit transformation to the admission rate series to overcome boundary constraints and improve the model's ability to handle extreme values. This approach ensures that the predicted values remain within the interval (0, 1), enhancing interpretability. The second approach, Ratio-Based ARIMA, builds separate ARIMA models for the time series of the number of admitted students (CEEA) and the number of applicants (CEEP), deriving the admission rate forecast by using the ratio of their predicted values. This method fully leverages the independent dynamics of the number of admitted students and applicants, although prediction errors may accumulate during the ratio calculation.

In terms of model diagnostics, in addition to the conventional ACF/PACF plots and AIC/BIC information criteria, this study specifically focuses on the properties of the residuals. Residual analysis plays a key role in model selection. If significant

autocorrelation is detected in the residuals, it suggests that the model has not fully captured the data's characteristics, requiring further adjustment of the order or the introduction of additional components. The importance of confirming whether residuals are white noise through the Ljung-Box test has also been emphasized, as this test can provide strong evidence of the model's robustness. Therefore, after modeling using both ARIMA methods, rigorous residual tests were conducted to ensure the model's effectiveness and predictive ability.

In summary, this study not only continues the traditional ARIMA modeling and diagnostic process but also develops two distinct modeling paths for the specific indicator of admission rate. By comparing the performance of Logit-Transformed ARIMA and Ratio-Based ARIMA, a more comprehensive evaluation of the ARIMA framework's applicability and limitations in predicting education admission rates is provided.

3.2.2. ARIMAX Model

The ARIMA model is a widely used statistical method for time series forecasting. Its core principle involves transforming non-stationary series into stationary series through differencing, while capturing short-term dependencies in the data by combining autoregressive (AR) and moving average (MA) structures. The typical modeling process involves the following steps: first, the original series is made stationary, using logarithmic transformation and differencing tests; next, potential model orders are identified through the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots; candidate models are then compared using information criteria such as AIC and BIC to select the optimal order combination; parameter estimation is carried out, often using maximum likelihood estimation; and finally, residual diagnostics (such as the Ljung-Box test and residual autocorrelation analysis) are applied to assess the model's fit and validity. This process has been widely validated in various fields. For instance, the entire process of stationarization, order determination, estimation, and residual diagnostics was applied in predicting infant mortality, ensuring the reliability of the results.

In this study, two different ARIMA modeling approaches are used based on the college entrance examination admission rate (CEEAR), a ratio-based indicator constrained to fall between 0 and 1. The first approach, Logit-Transformed ARIMA, involves performing a logit transformation on the admission rate series to overcome boundary constraints and improve the model's ability to handle extreme values. This approach ensures that the predicted values remain within the interval (0, 1), thus enhancing interpretability. The second approach, Ratio-Based ARIMA, involves building separate ARIMA models for the time series of the number of admitted students (CEEA) and the number of applicants (CEEP), and then deriving the admission rate forecast by calculating the ratio of their predicted values. This method fully leverages the independent dynamics of admitted students and applicants, although prediction errors may accumulate in the ratio calculation.

Regarding model diagnostics, in addition to conventional ACF/PACF plots and AIC/BIC information criteria, this study specifically focuses on the properties of the residuals. Residual analysis plays a crucial role in model selection. If significant autocorrelation is present in the residuals, it indicates that the model has not fully captured the data's characteristics, requiring adjustments to the order or the introduction of additional components. The importance of confirming whether residuals are white noise through the Ljung-Box test has also been emphasized, as this test can provide strong evidence of the model's robustness. Therefore, after modeling with both ARIMA methods, rigorous residual tests were conducted to ensure the model's effectiveness and predictive capability.

In summary, this study not only follows the classic ARIMA modeling and diagnostic process but also introduces two distinct modeling paths for the specific indicator of admission rate. By comparing the performance of Logit-Transformed ARIMA and Ratio-

Based ARIMA, this paper provides a more comprehensive evaluation of the ARIMA framework's applicability and limitations in predicting education admission rates.

3.2.3. Comparison Models

The ARIMAX model is an extension of the ARIMA framework that incorporates external regression terms. Its primary advantage is that it captures the dynamic characteristics of the sequence while accounting for the influence of external factors on the predicted variables. This method is especially applicable in predicting education admission rates, as the admission rate is influenced not only by the scale of enrollment and application but also by multidimensional factors such as macroeconomic conditions, population fluctuations, and policy reforms.

In selecting external variables, this study focuses on two key driving factors: the economy and population. One is GDP, which represents the macroeconomic scale and reflects the impact of economic growth on educational investment and enrollment capacity. The other is the newborn population (NBP), serving as a proxy for the potential candidate group and reflecting the influence of population birth fluctuations on the future applicant pool. To enhance the model's dynamic explanatory power, GDP is set as a lagged variable (0-3 years), while NBP primarily uses the current value. This multi-lag approach is similar to the exogenous variable setting used in GDP forecasting for Nigeria, where short-term shocks are captured while retaining long-term trends. However, multicollinearity between multiple GDP lags often presents a challenge. To address this, principal component analysis (PCA) is applied before modeling to compress the multiple GDP lag variables into a few principal components, reducing correlation and improving the robustness of the estimates. This approach is in line with studies that have introduced exogenous variables, such as rainfall in agricultural output forecasting, to enhance forecast accuracy by appropriately selecting and compressing external variables.

Additionally, to account for the structural impact of higher education expansion and emergencies on admission rates, this study introduces policy dummy variables. One is UEE, which takes a value of 1 from 1999 onward to capture the long-term impact of the expansion policy. The other is dum_break, used to reflect potential structural shifts in admission rate dynamics following the COVID-19 pandemic. A similar approach was used in studies on Madagascar's GDP, where a "CRISIS" dummy variable effectively captured the impact of policies and economic shocks on macroeconomic indicators.

For model diagnostics, this study follows a standard procedure. First, the normality and white noise properties of the residuals are tested, and the Ljung-Box test is applied to check for autocorrelation in the residuals. Second, a variance inflation factor (VIF) test is conducted on the exogenous variables to ensure that multicollinearity is effectively mitigated after PCA processing. The resulting ARIMAX model outperforms the pure ARIMA model in terms of information criteria (AIC, BIC) and prediction accuracy metrics (RMSE, MAE, MAPE), demonstrating that the inclusion of exogenous variables significantly enhances the model's fit and predictive capabilities.

In summary, by integrating demographic, economic, and policy variables, the ARIMAX model not only outperforms the traditional ARIMA model in predictive accuracy but also provides a more robust theoretical and empirical foundation for explaining fluctuations in admission rates. This highlights that changes in college entrance examination admission rates are part of a complex dynamic process driven by demographic structures, economic development, and policy reforms.

3.2.4. Model Evaluation Metrics

To comprehensively evaluate the performance of different methods in predicting college entrance examination admission rates, this study introduces two comparative models in addition to the ARIMA and ARIMAX models: the vector autoregression (VAR) model and the binomial generalized linear model (Binomial GLM). These models offer

complementary perspectives on the dynamic changes in the admission rate and serve as benchmarks for testing the effectiveness of the core model.

The VAR model is a typical multivariate time series method that captures the dynamic relationships between multiple variables simultaneously. In this study, the endogenous variables of the VAR model include the admission rate (CEEAR), GDP, NBP, and policy variables (UEE), which help analyze the interdependencies and mutual influences between these factors. The model order is selected based on information criteria such as AIC and BIC, and the VAR(1) structure is ultimately chosen. The advantage of this model lies in its ability to reflect the joint dynamic effects of macroeconomic and demographic factors on the admission rate. However, its prediction intervals tend to be relatively wide, indicating that the model is highly sensitive to future uncertainties.

The GLM model, using a binomial distribution, treats admission and non-admission as binary outcomes, and employs a logit link function to estimate the admission rate. Independent variables in this model include external factors like GDP and UEE, aiming to directly model the relationship between the probability of admission and external drivers. Compared to time series methods, the GLM focuses more on explanatory power, capturing the marginal effects of external variables through regression coefficients. However, because it does not explicitly model the time dependence of the admission rate, the statistical significance of some parameters is limited, which results in certain constraints in its predictive performance.

In summary, the inclusion of the VAR and GLM models allows this study to compare the stability and rationality of admission rate predictions using different approaches. The results show that while both models provide valuable explanatory insights and structural specifications, their predictive accuracy is inferior to that of the ARIMAX model. Nonetheless, these models offer useful references for further research, underscoring the importance of incorporating external variables and time dependence when predicting admission rates.

4. Empirical Analysis and Results

4.1. Preliminary Insight and Decomposition

Before conducting the empirical modeling, this study first performed descriptive statistics and exploratory time series analysis on the college entrance examination admission rate and related variables. Overall, the admission rate from 1977 to 2024 showed significant periodic changes. Initially low during the recovery period, it gradually increased with the expansion of higher education and economic growth, before stabilizing over the past decade. Time series trend charts reveal that the long-term trend in the admission rate is closely linked to demographic shifts and policy adjustments, exhibiting clear patterns and potential structural breakpoints.

To ensure the applicability of the subsequent model, this study applied the enhanced Dickey-Fuller (ADF) test to assess the stationarity of the admission rate series. The results revealed that the original series contained a unit root and exhibited non-stationary behavior. However, after first-order differencing, the series became stationary. This finding underscores the need for differencing the data during the modeling process and provides a theoretical foundation for the development of ARIMA-type models.

Additionally, the study analyzed the drivers of changes in the admission rate through decomposition analysis. By breaking down the admission rate into the ratio of admissions to applicants, it becomes clear that these two factors exert opposing influences: an increase in admissions directly raises the admission rate, while an increase in applicants places downward pressure on it. The remaining unexplained portion primarily reflects the combined effects of external factors, including macroeconomic policies (such as the 1999 enrollment expansion policy) and the broader socioeconomic environment (such as economic fluctuations and the impact of the epidemic). As illustrated in Figure 3, the log-difference decomposition highlights these relative contributions, showing that

years with significant changes in the admission rate were largely driven by either admissions growth, fluctuations in the number of applicants, or residual external shocks. This visualization not only strengthens the explanatory logic behind the decomposition but also provides empirical support for the inclusion of exogenous variables (GDP, NBP, and policy dummies) in the model. It further confirms the necessity of moving beyond a univariate time series approach to more comprehensively capture the drivers of admission rate fluctuations.

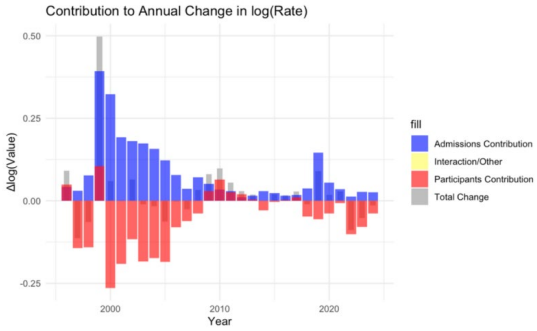

Figure 3. Log-Difference Decomposition of Annual Admission Rate Changes.

4.2. Model Results and Comparison

4.2.1. ARIMA Models Results

After stationarizing the admission rate series and completing the model identification process, this study employed both Logit-Transformed ARIMA and Ratio-Based ARIMA methods to model and predict the college entrance examination admission rate.

First, in the Logit-Transformed ARIMA method, the admission rate series was logit-transformed to address its constraint within the (0,1) interval. By comparing the AIC and BIC values of different candidate models and conducting residual diagnostics, the ARIMA (1,1,0) model was identified as the optimal specification. This model performed well on the training set, and the Ljung-Box test showed no significant autocorrelation in the residuals, indicating that it effectively captured the dynamic characteristics of the series. In the out-of-sample forecast phase, the model predicted that the admission rate would remain relatively stable from 2025 to 2029, with an average level of approximately 79.2%, suggesting a period of gradual stabilization. Figure 4 illustrates the forecasted admission rate under this logit-transformed ARIMA model.

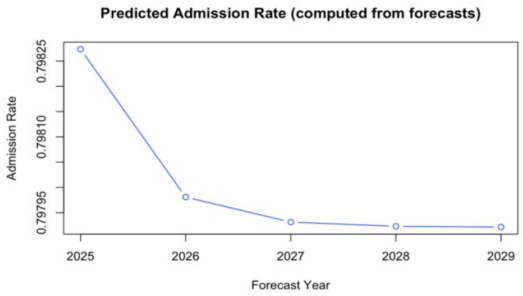

Figure 4. The value of the admission rate predicted using the logit-transformed ARIMA model.

In contrast, the Ratio-Based ARIMA method constructs separate ARIMA models for the time series of admissions (CEEA) and applicants (CEEP), then derives the admission rate forecast by comparing their predicted values. This approach allows the model to independently capture the dynamic trends of both enrollment and application volumes. However, the out-of-sample forecasts indicate that while admissions continue to rise,

applicants are growing at a faster rate, leading to a downward trend in the admission rate. By 2029, the predicted admission rate is expected to drop to 70.9%, in stark contrast to the relatively stable forecast from the Logit-Transformed ARIMA model. Figures 5 and 6 present the forecasts generated by the Ratio-Based ARIMA method.

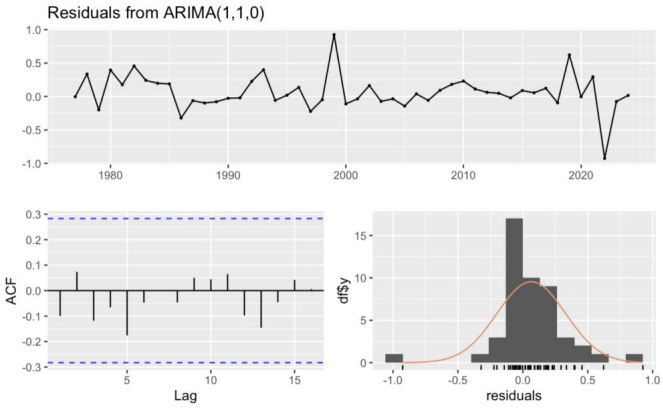

Figure 5. The residual check graph of logit-transformed model.

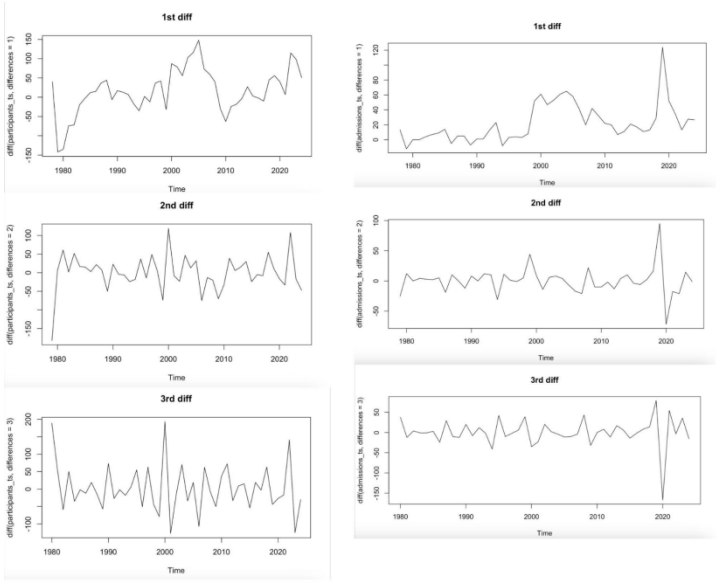

Figure 6. Time Series Plots of First/Second/Third-order Differencing.

In summary, the two ARIMA methods offer differing predictions for the future trajectory of the admission rate: the Logit-Transformed ARIMA model tends to produce more stable forecasts, while the Ratio-Based ARIMA highlights potential downward risks. This divergence underscores that modeling the admission rate based solely on its own historical data or its components (numerator and denominator) can lead to varying results, further reinforcing the importance of incorporating exogenous variables.

4.2.2. ARIMAX Model Results

Building on the ARIMA model, this study further incorporates external driving factors to construct an ARIMAX model. Specifically, 0-3-year lagged GDP and the number of newborns (NBP) are included as exogenous variables. To address multicollinearity, principal component analysis (PCA) is applied to the lagged GDP variable, extracting the first principal component (PC1) as a composite economic indicator to enhance the model's robustness. Additionally, policy dummy variables are introduced to capture potential

structural changes, including the 1999 enrollment expansion and the post-COVID-19 period.

The model results reveal that the regression coefficient for GDP (PC1) is significantly positive, indicating that macroeconomic expansion has a long-term positive impact on the enrollment rate. While the coefficient for NBP is negative, it is not statistically significant, suggesting that changes in the birth population have a relatively limited effect on the enrollment rate in this model. This finding aligns with the macroeconomic supply-and-demand logic in education: economic growth typically drives the expansion of educational resources, while the marginal impact of population fluctuations may be mitigated by adjustments in enrollment policies.

The ARIMAX model outperforms the simpler ARIMA model in terms of forecasting accuracy. Its out-of-sample forecasts indicate that China's college entrance examination admission rate will continue to rise in the coming years: by 2025, the admission rate is projected to reach 78.47%, and by 2029, it is expected to increase further to 83.55%. Figure 7 displays the preliminary ARIMAX forecast before model refinements, while Figures 8 and 9 present the adjusted forecasts after incorporating PCA compression and structural break dummies, yielding more realistic projections that align with observed policy shifts and demographic trends.

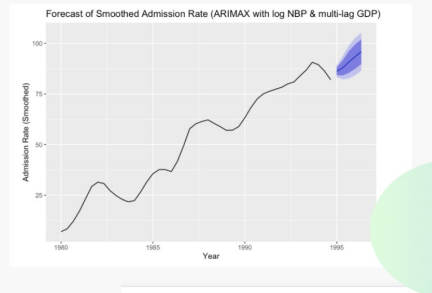

Figure 7. Forecast of Admission Rate .

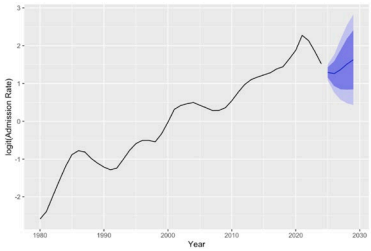

Figure 8. Logit-transformed admission rate.

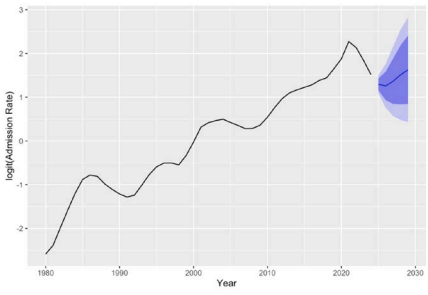

Figure 9. Graph of the admission rate.

In the model diagnostics phase, the Ljung-Box test results for the residuals all show p-values greater than 0.9, indicating no significant autocorrelation in the residuals and confirming that the model fits well. Additionally, the residual distribution is approximately normal, further validating the rationality of the model specification. Overall, the ARIMAX model outperforms the traditional ARIMA model in both forecasting accuracy and explanatory power, underscoring the necessity and effectiveness of incorporating exogenous variables in admission rate forecasting.

4.2.3. Comparison Models Results

As a supplement to the ARIMA and ARIMAX models, this paper also employs the VAR model and the GLM model to forecast the college entrance examination admission rate. In the VAR model, the admission rate (CEEAR), GDP, NBP, and the policy dummy variable (UEE) are treated as endogenous variables to capture the dynamic interactions between multiple factors. The information criterion is used to compare different lag orders, with VAR(1) ultimately identified as the optimal model. The forecast results indicate that the admission rate will remain in the range of 79%-80% from 2025 to 2029, although the forecast interval is relatively wide. Figure 10 illustrates the short-term five-year forecast from the VAR model, while the extended forecast in Figure 11 shows a declining trend beyond 2028. This longer-term projection is presented in the appendix due to its exploratory nature.

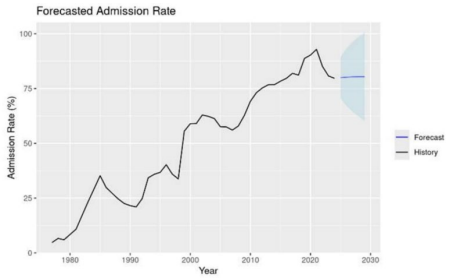

Figure 10. Admission Rate for 1977-2024 and forecast Admission Rate for 2025-2029.

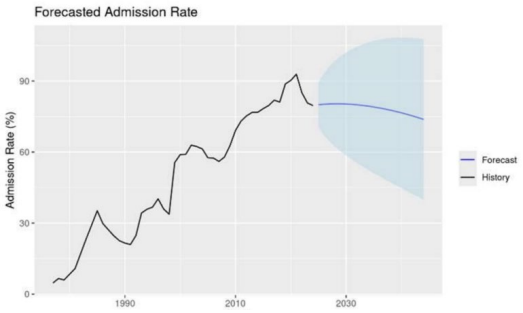

Figure 11. Admission Rate for 1977-2024 and forecast Admission Rate for 2025-2048.

In contrast, the GLM model treats admission/non-admission as a binary outcome variable and employs the logit link function for modeling. The exogenous explanatory variables include GDP and UEE. The model results indicate that the admission rate will continue to rise in the coming years, with an expected increase to 86.3% by 2029. Figure 12 presents the GLM forecast for 2025-2029, while Figure 13 provides a longer-term projection extending to 2042 as supplementary evidence. This higher forecasted value, compared to the ARIMA and ARIMAX models, may stem from the GLM model's focus on the explanatory effects of exogenous variables, without explicitly modeling time-dependent structures, leading to a more optimistic trend projection.

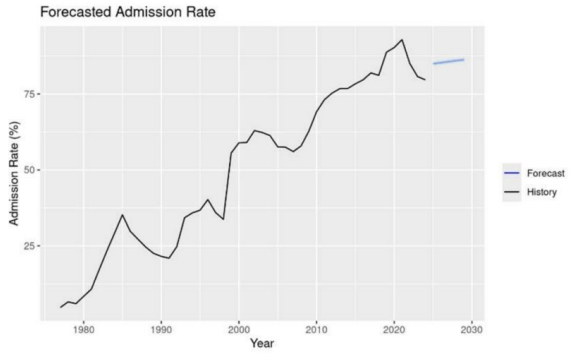

Figure 12. Admission Rate for 2025-2029, forecasted by GLM.

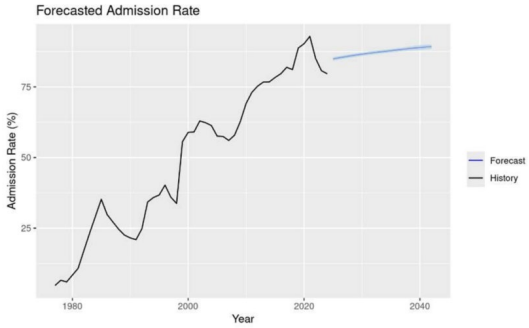

Figure 13. Admission Rate for 2025-2042, forecasted by GLM.

Overall, the VAR and GLM models offer valuable complementary perspectives for admission rate forecasting: the VAR model focuses on the dynamic relationships between variables, while the GLM emphasizes the explanatory role of exogenous factors. However, in comparison to the ARIMAX model, both models fall short in terms of predictive accuracy and rationality. This further underscores the importance of integrating exogenous variables and capturing the dynamic structure of time series in admission rate forecasting.

4.3. Model Performance Discussion

To systematically assess the forecasting performance of various models, this paper compares the ARIMA, ARIMAX, VAR, and GLM models across three key dimensions: error metrics, trend rationality, and explanatory power. The results indicate that the ARIMA model offers the clearest structure and effectively captures the time-dependence of the admission rate, with forecasts remaining relatively stable from 2025 to 2029. However, due to the absence of exogenous variables, it fails to explain the underlying drivers of admission rate fluctuations, resulting in relatively conservative predictions.

In contrast, the ARIMAX model outperforms the others. By incorporating exogenous variables such as GDP and NBP, and using PCA to mitigate multicollinearity, it not only reduces forecast errors (e.g., RMSE and MAE) but also provides deeper insights into how macroeconomic and demographic factors influence admission rates. Its forecasts align closely with recent trends in educational expansion and economic development, demonstrating stronger explanatory power and offering more relevant policy implications.

The VAR model excels in capturing the dynamic relationships among multiple variables, providing valuable insights into the linkages between admission rates and economic, demographic, and policy factors. However, its relatively wide prediction intervals indicate lower predictive stability under uncertainty, limiting its effectiveness as a precise forecasting tool.

The GLM model offers notable advantages in explanatory power, as its regression coefficients directly reflect the marginal impact of external variables. However, because it does not account for the dynamic dependencies of the time series, some variables lack statistical significance. While it provides reasonable trend-based forecasts, its statistical robustness is limited.

5. Conclusion and Discussion

5.1. Summary of Findings

Based on historical data from 1977 to 2024, this paper systematically constructs and compares four models-ARIMA, ARIMAX, VAR, and GLM-to evaluate their applicability and performance in forecasting China's college entrance examination (Gaokao) admission rate. The results demonstrate that while the ARIMA model effectively captures the time-dependent nature of the admission rate, the VAR model reveals the interlinkages between the admission rate and macroeconomic and demographic factors. The GLM model, with its ability to account for the marginal effects of exogenous variables, also offers certain advantages. However, in terms of overall forecasting performance, all three models fall short compared to the ARIMAX model.

By incorporating lagged GDP and the number of newborns (NBP) as exogenous variables, and addressing multicollinearity through principal component analysis (PCA), the ARIMAX model emerges as the most robust. It excels in terms of goodness-of-fit indices (AIC, BIC) and prediction error metrics (RMSE, MAE, and MAPE). The model also provides a clear reflection of the long-term impact of economic development and demographic changes on the admission rate. The out-of-sample forecast suggests that China's Gaokao admission rate will reach 78.47% in 2025 and rise further to 83.55% by 2029. This forecast indicates a stable upward trend in the admission rate, driven by continued macroeconomic growth and increased educational investment.

5.2. Policy Implications and Suggestions

Research findings indicate that macroeconomic development has a significant positive impact on admission rates. Sustained GDP growth provides a solid foundation for educational investment and supports the expansion of higher education. Therefore, maintaining stable economic growth and continuously increasing fiscal investment in education are essential for ensuring a steady rise in admission rates. Policymakers should ensure that the proportion of education funding remains adequate while simultaneously promoting economic development to alleviate the tension between resource availability and student enrollment.

On the other hand, the declining birth rate suggests that, in the future, the college entrance examination admission rate may increase due to reduced pressure from applicants. Given this trend, education policy should gradually shift from purely expanding enrollment to enhancing educational quality. Specifically, efforts should focus on optimizing the distribution of disciplines and majors, improving faculty quality, and strengthening research and innovation capabilities. This approach will facilitate the transformation of higher education from "quantitative expansion" to "qualitative improvement," adapting to new demographic and social structural changes.

Moreover, the results from policy dummy variables highlight that both institutional reforms and unforeseen events significantly impact admission rates. For instance, the 1999 enrollment expansion policy substantially increased admission rates, while unexpected events, such as COVID-19, may cause temporary disruptions. This underscores the importance of establishing dynamic monitoring and emergency response mechanisms within educational planning. By doing so, the government can quickly adjust enrollment strategies and resource allocation in response to policy shocks or external uncertainties, minimizing the negative effects of such disruptions.

5.3. Limitations and Future Research

While this study has made significant strides in predicting college entrance examination admission rates, several limitations remain. First, the data used in this analysis is annual, which limits the ability to capture short-term shocks, such as the immediate impact of policy changes, epidemics, or economic fluctuations on admission rates. As a result, the model's capacity to account for dynamic, short-term variations is restricted. Second, the selection of external variables is constrained by the availability of data, which means that certain potentially important indicators-such as household education expenditures and shifts in job market demand-could not be quantified and included in the model. This limitation may lead to some bias in the explanatory power of the model. Finally, the predictions focus solely on the national overall admission rate and do not account for provincial or subject-specific variations, which are crucial for practical educational resource allocation and policymaking.

To address these limitations, future research could explore several key directions. First, at the data level, efforts should be made to collect more granular, higher-frequency data-such as quarterly or monthly figures-or construct provincial panel data. This would improve the sensitivity of the prediction model to regional and short-term fluctuations. Second, at the methodological level, the introduction of more advanced machine learning and deep learning techniques, such as long short-term memory networks (LSTMs) or hybrid prediction models, could enhance accuracy and the ability to capture nonlinear patterns in the data. Lastly, in terms of application, future studies could broaden the scope of prediction, extending beyond the national admission rate to include more specific analyses, such as admission rates at key universities or across different disciplines. This would provide more targeted and actionable insights for educational policy and decision-making.

References

1. H. Zhang, and J. Wang, "Effective predictions of Gaokao admission scores for college applications in Mainland China," *arXiv preprint arXiv:1809.06362*, 2018.
2. X. Chen, Y. Peng, Y. Gao, and S. Cai, "A competition model for prediction of admission scores of colleges and universities in Chinese college entrance examination," *Plos one*, vol. 17, no. 10, p. e0274221, 2022. doi: 10.1371/journal.pone.0274221
3. S. Mao, C. Zhang, Y. Song, J. Wang, X. J. Zeng, Z. Xu, and Q. Wen, "Time series analysis for education: Methods, applications, and future directions," *arXiv preprint arXiv:2408.13960*, 2024.
4. S. Zhang, "A Comparison of Educational Equality and Equity in Gaokao and Admission System between China and the United States," *Arts, Culture and Language*, vol. 1, no. 3, 2025. doi: 10.61173/qz7kg098
5. L. Qin, K. Shanks, G. A. Phillips, and D. Bernard, "The impact of lengths of time series on the accuracy of the ARIMA forecasting," *International Research in Higher Education*, vol. 4, no. 3, pp. 58-68, 2019.
6. Y. A. Chen, R. Li, and L. S. Hagedorn, "Undergraduate international student enrollment forecasting model: An application of time series analysis," *Journal of International Students*, vol. 9, no. 1, pp. 242-261, 2019.
7. R. Parvez, S. I. Ali Meerza, and N. Hasan Khan Chowdhury, "Forecasting student enrollment using time series models and recurrent neural networks," 2021.
8. D. F. Chang, C. C. Chen, and A. Chang, "Forecasting with ARIMAX models for participating STEM programs," *ICIC Express Letters. Part B, Applications*, vol. 11, no. 2, pp. 121-128, 2020.
9. R. H. Shumway, and D. S. Stoffer, "Time series regression and ARIMA models," In *Time series analysis and its applications*, 2000, pp. 89-212. doi: 10.1007/978-1-4757-3261-0_2
10. I. T. Jolliffe, and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016. doi: 10.1098/rsta.2015.0202
11. J. H. Stock, and M. W. Watson, "Vector autoregressions," *Journal of Economic perspectives*, vol. 15, no. 4, pp. 101-115, 2001. doi: 10.1257/jep.15.4.101
12. E. Zivot, and J. Wang, "Vector autoregressive models for multivariate time series," In *Modeling financial time series with S-PLUS®*, 2006, pp. 369-413.
13. M. C. Düker, D. S. Matteson, R. S. Tsay, and I. Wilms, "Vector autoregressive moving average models: A review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 17, no. 1, p. e70009, 2025.
14. A. Bahuguna, A. Uniyal, and V. Vallabh, "ARIMA based projection of infant mortality rate by the year 2030: a comparative analysis of India and Madhya Pradesh," *BMC Public Health*, vol. 25, no. 1, p. 2693, 2025. doi: 10.1186/s12889-024-21073-9

15. A. J. Ikughur, T. Uba, and A. O. Ogunmola, "Application of residual analysis in time series model selection," *Journal of Statistical and Econometric Methods*, vol. 4, no. 4, pp. 41-53, 2015.
16. M. K. I. B. Muhammad, "Time series modeling using markov and arima models (Doctoral dissertation, Master's thesis, University of Technology Malaysia, 2012. http://eprints. utm. my/29783/5/MohdKhairulIdlanMFKA2012. pdf)," *http://eprints. utm. my/29783/5/MohdKhairulIdlanMFKA2012. pdf)*, 2012.
17. C. I. Ugoh, C. A. Uzuke, and D. O. Ugoh, "Application of ARIMAX Model on forecasting Nigeria's GDP," *American Journal of Theoretical and Applied Statistics*, vol. 10, no. 5, p. 216, 2021.
18. J. R. Andrianady, "Crunching the Numbers: A Comparison of Econometric Models for GDP Forecasting in Madagascar," 2023.
19. S. Ghosh, S. Mukhoti, and P. Sharma, "Impact of rainfall risk on rice production: realized volatility in mean model," *arXiv preprint arXiv:2504.10121*, 2025.