Analysis of California's Accommodation GDP Based on the ARIMA Model

Chenhao Ye
School of Mathematics and Statistics, Changchun University, Jilin, China
yechenhao027@gmail.com

Abstract. This study focuses on forecasting the future trajectory of California's annual accommodation GDP using an ARIMA model. This study utilizes annual accommodation GDP data for California from 1997 to 2023, comprising 27 observations, sourced from the Federal Reserve Economic Database (FRED) maintained by the Federal Reserve Bank of St. Louis. The ARIMA(1,1,0) with drift model is established and reasonable predictions are made about future changes in California's annual accommodation GDP. In building the model, this study employed a variety of tests, such as the stationarity test of the data, the white noise test of the data, the significance test of parameter estimates, the significance test of the model, etc., to construct the most appropriate model. The model chosen in this study is ARIMA(1,1,0) with drift, and the prediction shows that California's annual accommodation GDP will still show a significant upward trend in the future. Based on the research results, this study proposes that people need to prepare more funds to deal with the rising cost of accommodation in California or look for a partner to share the cost of accommodation.

Keywords: Time Series Forecasting; ARIMA Model; ADF test; Model Caliber; California's Annual Accommodation GDP.

1. Introduction

In recent years, influenced by education, technology, etc., California has become a popular state for international students in the United States. A growing number of international students are flocking to California, which has led to greater accommodation costs in the state and deterred many families of international students. Regarding academic domain, there is insufficient research shows that have quantitatively analyzed accommodation costs in California alone. During the literature review, it was found that scholarly articles specifically analyzing accommodation costs are notably scarce, whereas the majority of existing research focuses on the analysis of overall GDP fluctuations [1]. To enable families of international students to better understand the changes in accommodation costs in California in advance and prepare the corresponding funds in advance, this study takes the annual accommodation GDP of California as its research object, with the research question of predicting its future trends. Through an in-depth statistical analysis of its development, the author seeks to understand the internal mechanisms driving its changes and thus develop effective means to forecast its future accommodation GDP. This research offers substantive practical implications as it contributes to bridging an existing gap in literature.

2. Theoretical Knowledge

The ARIMA model is an exceptionally powerful statistical model that is significant in time series analysis, as it can make reasonable predictive analyses of future values of a time series by studying its past values. The research object in this study is California's annual accommodation GDP, and the question is about the future outlook for California's annual GDP in the accommodation industry. In the data used in this study, the annual accommodation GDP of California is arranged in chronological order, so the research object of this study belongs to a time series, and this study can use the model to summarize the currently known past values of the research object and explore the mathematical patterns therein. Ultimately, the model is used to analyze and predict the future values of the research subject to solve the research problem.

The following structure of the model is called an autoregressive moving average model, denoted as ARIMA (p, d, q), where p is the regression term, q is the moving average term, and d represents the difference order used to extract the trend in the time series [2].

$$\nabla^d x_t = \frac{\theta(B)}{\varphi(B)}\varepsilon_t + C \tag{1}$$

In the formula, ε_t is the series of white noise with a mean of 0, $\theta(B)$ is the autoregressive coefficient polynomial of the stationary reversible ARMA(p, q) model, $\varphi(B)$ is the moving average coefficient polynomial of the stationary reversible ARMA(p, q) model, ∇^d represents the d order difference of the series, x_t represents the value of the subject x in this study at time t, C represents the drift term of the model [3].

When dealing with time series containing trendiness, the difference method is usually used to eliminate the trendiness of the series, transforming the series from non-stationary to stationary for the next step of fitting, and the difference process can be expressed as:

$$\nabla^p x_t = \nabla^{p-1} x_t - \nabla^{p-1} x_{t-1} \tag{2}$$

Where ∇^p represents the p order difference of the series, and when the series shows a linear trend, the first order difference is required, $p = 1$, then x_{t-1} represents the value of the subject x in this study at the time $t - 1$.

3. California Accommodation GDP Analysis and Modeling

This section is devoted to the analytical and modeling procedures applied to the collected dataset, as detailed in the following steps.

3.1. Data Overview

The data used in this study is sourced from the Bureau of Economic Analysis of the United States, taken from FRED and the Federal Reserve Bank of St. Louis. It selects the GDP generated by accommodation in California each year from 1997 to 2023. The dataset comprises a total of 27 records, with the unit being million USD. The dataset is structured at an annual granularity, meaning that monthly data is not available.

3.2. Trend Analysis

This study first drew a time series chart of the GDP generated by accommodation in California each year from 1997 to 2023 for trend analysis. The results are shown on Fig. 1. It is evident that this data has a clear trend, showing an overall upward trend. It is initially judged that this series has obvious non-stationarity. After 2020, Due to the impact of COVID-19, the GDP declined to some extent, but the overall impact was not significant [4]. It did not cause a sustained decline in California's annual accommodation GDP. On the contrary, after that, California's annual accommodation GDP saw a considerable increase. In the analysis and modeling of this study, the seasonality of the data is not considered for the time being, since the data used is annual data.

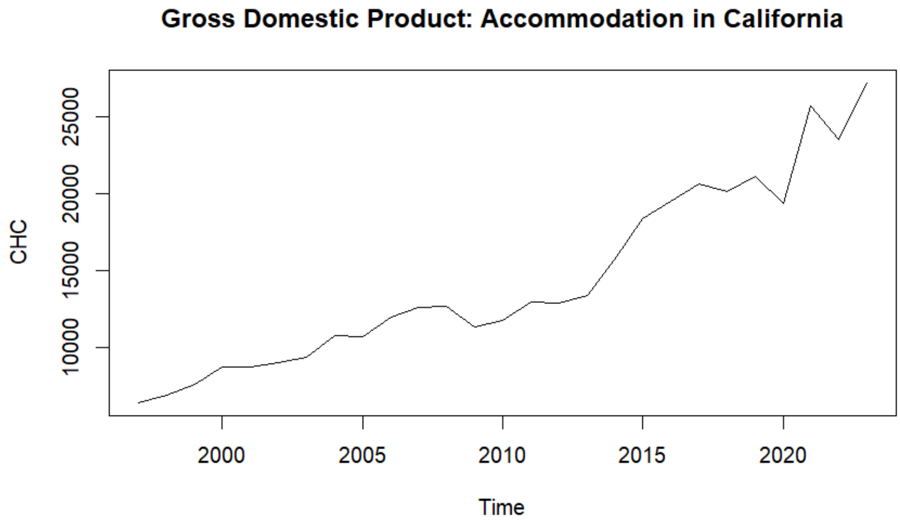

Fig. 1 Time series chart of Accommodation GDP in California [5]

So as to more rigorously explore the characteristics of the initial data, the author used software to create the autocorrelation coefficient graph and the partial autocorrelation coefficient graph of the data, as shown in Fig 2, which shows that the autocorrelation coefficient of this series did not rapidly drop to 0 but instead showed an overall slow downward trend [6]. Therefore, the series is a non-stationary series. At present, this series cannot be used for model fitting because one of the prerequisites for fitting with the ARIMA model is that the series needs to be a stationary series, and the non-stationarity in it has been eliminated before fitting. Therefore, the author believes that with the aim of extracting the linear trend of the series, the first-order difference should be considered first.

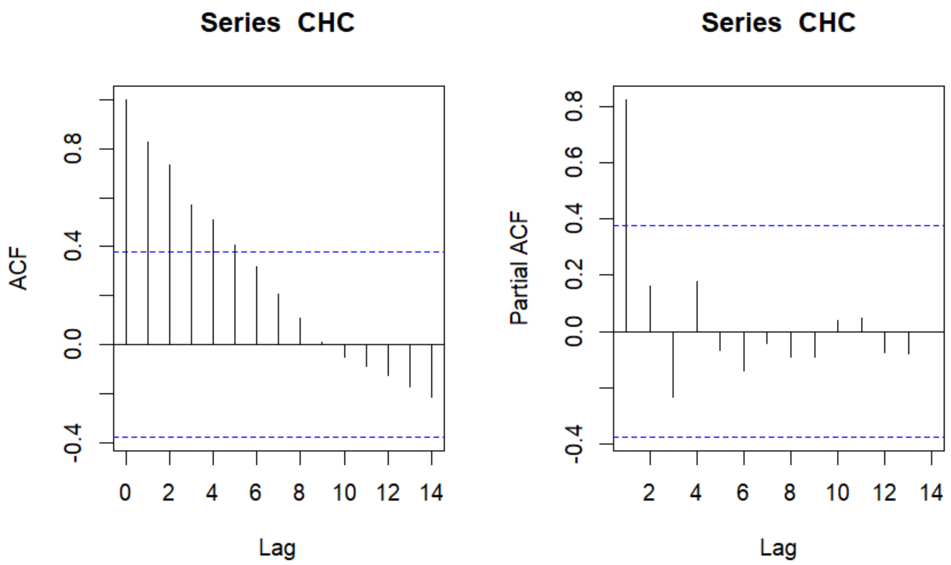

Fig. 2 ACF plots and PACF plots

Before conducting the difference, this study also carried out a white noise test on the series. If the series is a white noise series, then there is no significance for further analysis of the series because there is no correlation between the series, and it is impossible to summarize the patterns in the series through mathematical modeling and perform extrapolation prediction. In this study, the LB test was used to conduct a white noise test on the series. The null hypothesis of the LB test is that it assumes that the test quantities are all white noise series in terms of the lag order of the test, there is no autocorrelation [7]. The test results are shown in Table 1 below, which indicates that the P-values of the test results are all much less than 0.05, rejecting the null hypothesis. Therefore, it is considered that this series is a non-white noise test It has the value of further analysis.

Table 1. Results of the LB test

Lag	LB statistics	P-value
6	66.342	2.295e-12
12	69.794	3.5e-10

3.3. Preliminary Data Processing

Plot the series after the first-order difference, as shown in Fig 3.

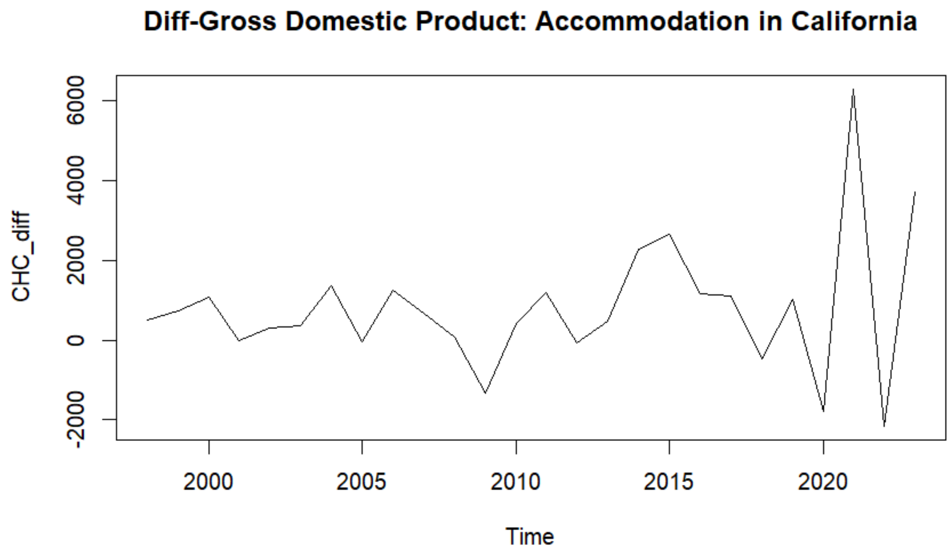

Fig. 3 Series diagram of the first-order difference California Accommodation GDP

The results show that the series after the first-order difference has no obvious trend, and the trend of the series has been eliminated through the first-order difference. To better make digital judgments on the trend of series, the author conducted a stationarity ADF test on the series after first-order difference. The null hypothesis was to assume that the series were non-stationary series [8]. The test results are shown in Table 2, which indicates that the P-values of the model in the ADF test are less than 0.05 in most cases, rejecting the null hypothesis. Therefore, the series after first-order differencing is stationary, and its original non-stationarity has been eliminated.

Table 2. ADF Test Results

Types	Lag	Tau	P-value
No drift no trend	0	-6.08	0.01
	1	-2.02	0.04
	2	-1.85	0.06
With drift no trend	0	-8.45	0.01
	1	-3.08	0.04
	2	-3.32	0.02
With drift and trend	0	-8.76	0.01
	1	-3.31	0.09
	2	-3.64	0.04

Further, the white noise LB test was conducted on the series after the first-order difference, as shown in Table 3, which indicates that when the delay order of the series is 6, the P-value of the LB test result is less than 0.05, but when the delay order is 12, the P-value of the LB test result is greater than 0.05. Therefore, the author believes that the series after the first-order difference only has short-term correlation. When the delay order is within the 6th order, the series after the first-order difference is not a white noise series and has further research value. However, when the delay order exceeds the 6th order, the series after the first-order difference may become a white noise series. There is no obvious correlation among the research objects, and the rules among the research objects cannot be summarized through mathematical models, thus losing the research value of the series. So, in the following process of predicting series, this study only selects series from recent years for extrapolation prediction.

Table 3. LB test Results

Lag	LB statistics	P-value
6	15.216	0.01864
12	16.659	0.1629

3.4. ARIMA Model Construction

In this section, we first select a candidate ARIMA model based on a subjective interpretation of the ACF and PACF plots. However, the parameter estimates for this model are statistically insignificant. Consequently, we adopt an automated approach by employing the 'auto-arima' algorithm to identify and rank potential models. The final model is selected based on the significance of its parameters and diagnostic analysis of its residuals. This optimally fitted model is then used for extrapolation and forecasting [9].

3.4.1 ARIMA Model Order Determination

First, the autocorrelation coefficient graph and the partial autocorrelation coefficient graph of the first-order difference series are made to positively determine the order of the model. As shown in Fig 4, the partial autocorrelation coefficient is not significantly 0 when the delay order is 1 but rapidly drops to within the blue standard line after the first order. However, the autocorrelation coefficient does not show a rapid downward trend. The trailing nature of the autocorrelation coefficient, which differs markedly from the sharp cutoff of the partial autocorrelation coefficient, empirically justifies the use of an ARIMA(1,1,0) model for the series.

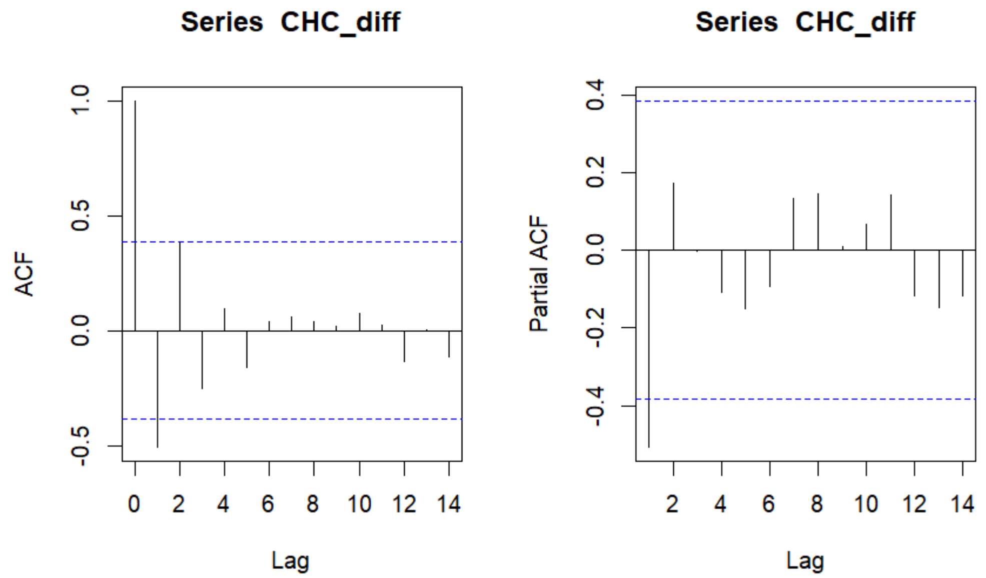

Fig. 4 ACF plot and PACF plot

3.4.2 Preliminary Parameter Estimation of the ARIMA Model

In this study, the least squares method is used to estimate the parameters of the model, and the estimation results are shown in Table 4 below. The results show that the estimated values of the model parameters are not significantly greater than twice the standard error, and the parameters are not significantly non-zero, which makes the parameter test of the model fail, and there are certain problems with the fitting results. To further optimize the model, the author used the automatic order determination method to fit the model.

Table 4. Fitting Results of ARIMA Model

Parameters	AR1
Estimates	-0.2862
Standard error	0.2005
Model variance =3096069	AIC = 466.46

3.4.3 ARIMA Model Optimization

The author used 'auto-arima' algorithm to automatically order the series [10]. The results, presented in Table 5, indicate that the series is still suitable for an ARIMA(1,1,0) model, but requires the inclusion of a drift term. The parameter values in the final fitting results are all greater than 2 times the standard error. This suggests that the parameters in the fitting results are significantly not 0.

Table 5. Model Fitting Results of the ARIMA model with drift term

Parameters	AR1	Drift
Estimates	-0.5553	762.5733
Standard error	0.1693	176.9364
Model variance =2,070,836	N/A	AIC = 456.2

To further test the significance of the parameters, this study uses PT function to digitally test the parameters. The null hypothesis assumes that the parameters in the model are significantly 0. The test results are shown in Table 6, which indicates that the P-values of the parameters are all much less than 0.05. The null hypothesis is rejected. Finally, it is determined that the parameters in the ARIMA(1,1,0) with drift model are significantly non-0. The next step can be carried out.

Table 6. PT Test Results

Parameter	AR1	Drift
P-value	0.0015235978	0.0001114041

3.4.4 ARIMA Model Checking

Finally, for the purpose of judging the basic assumptions of the model, this study first conducts a white noise test on the residuals of the fitted model, as shown in Table 7. It is found that when the lag order is 6 and 12, the P-values of the test results are much greater than 0.05. Accepting the null hypothesis, this demonstrates that under the significance level of 0.05, the residuals after model fitting are white noise series. The results of the model fitting already contain all the information in the data.

Table 7. LB Test Results

Lag	LB statistics	P-value
6	3.7392	0.7119
12	7.9733	0.7872

Further diagnosis of residuals in the fitted model is shown in Fig 5. It is evident that the autocorrelation coefficient of the residuals fluctuates within the blue standard line all the time, and it is found that the estimated curve of the residuals is basically similar to the normal distribution curve. Therefore, this demonstrates that the residuals test of the model is passed, the model fitting results are excellent, and the model can be used for prediction.

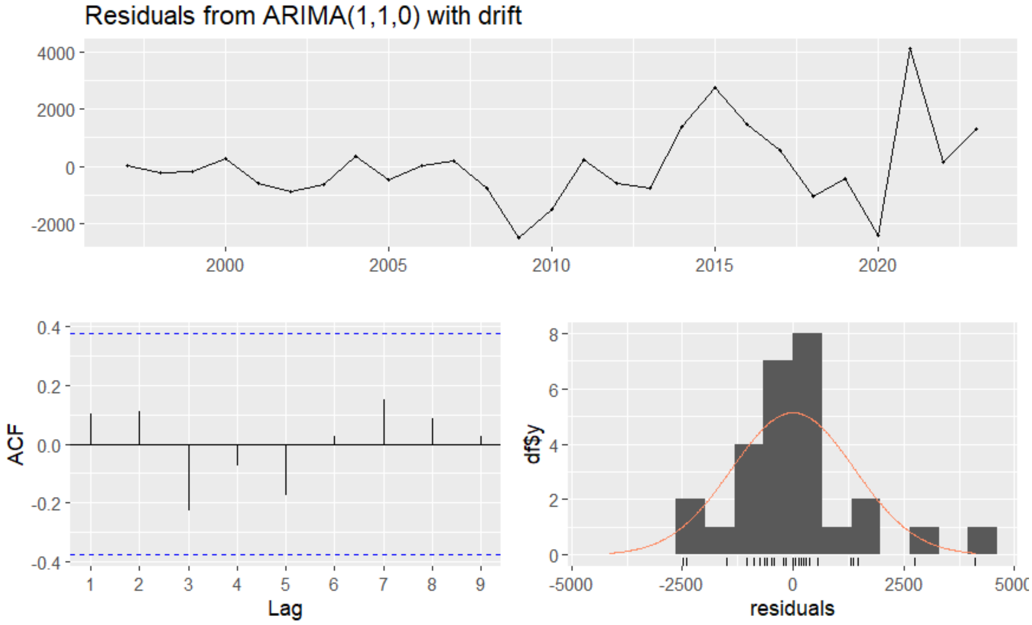

Fig. 5 Residual test results of the model

3.4.5 ARIMA Model Predictions

The equation for the final model is:

$$x_t - 0.447x_{t-1} - 0.553x_{t-2} = \varepsilon_t + 762.5733 \tag{3}$$

The extrapolation prediction of California's annual accommodation GDP after 2023 was conducted using the constructed ARIMA(1,1,0) with drift model [11]. The prediction results are shown in Fig 6. As evidenced by the point forecasts and the comparatively narrow prediction intervals (80% and 95%), the time series exhibits a distinct and consistent upward trajectory in its future values. However, the slope of the curve has not changed significantly compared to history. This indicates that the annual accommodation GDP of California will increase with the increase of years, but the overall growth rate tends to stabilize. There won't be any significant ups and downs in the coming years.

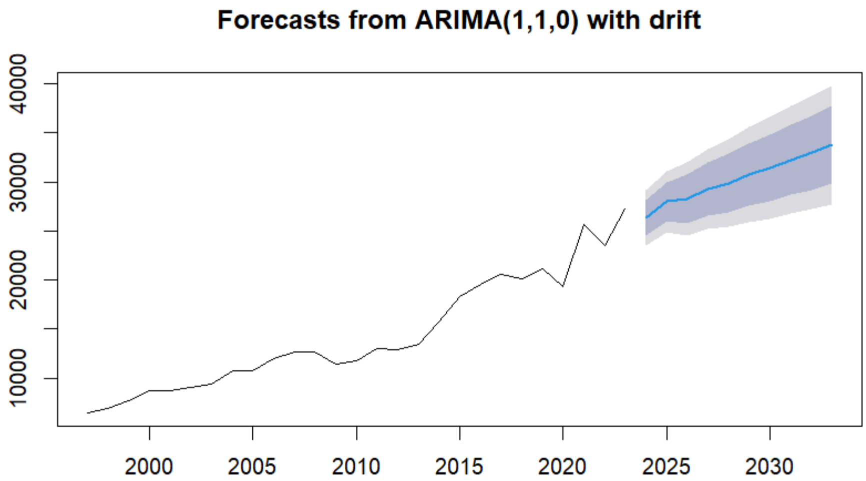

Fig. 6 Model prediction results

To present the prediction results of the model fitting more intuitively, this study presents the obtained prediction results in a tabular form. Due to the test results above in this study, the author found that the series only have short-term correlations. Therefore, this study only shows the annual accommodation GDP of California from January 2024 to January 2026. The specific numerical prediction results are shown in Table 8 below:

Table 8. Model Prediction Results

Year	Point estimates	Down 80%	Up 80%	Down 95%	Top 95%
Jan 2024	26331.51	24487.30	28175.71	23511.04	29151.97
Jan 2025	27999.89	25981.58	30018.19	24913.16	31086.61
Jan 2026	28259.42	25809.46	30709.39	24512.53	32006.32

4. Solutions

California is located in the western part of the United States, close to the Pacific Ocean, with a pleasant climate. Its population and economic aggregate rank among the top in the United States. As a result, the accommodation costs in California have always been at the leading level in the United States. Moreover, in recent years, due to the influence of a large number of foreign students, there have been signs that the accommodation costs in California will further increase. By consulting online literature, the author found that most of the factors influencing accommodation costs in California are related to population, such as technology, entertainment, education, etc. The closer to Silicon Valley, prestigious universities and tourist attractions, the greater the accommodation costs people spend. This is not hard to understand. Most people come to California in pursuit of opportunities there. This proximity facilitates access to a greater pool of opportunities and networks of like-minded individuals, which are concentrated in regions characterized by dense populations, advanced economies, and high-quality educational infrastructure. Moreover, life in these areas is convenient, and the surrounding medical and educational systems are well-developed with a high safety factor. The people's happiness index is high, which is suitable for people's long-term life. During the COVID-19 period, California's annual accommodation GDP declined to some extent, but then it rose rapidly, with the increase exceeding historical changes. This suggests that California's annual accommodation GDP has not been significantly affected by COVID-19 in the long term. COVID-19 only affected the short-term trend changes in California's annual accommodation GDP, and these trends will gradually be smoothed out over time.

5. Conclusion

This study uses the ARIMA(1,1,0) with drift model to systematically predict and model the annual accommodation GDP of California. Taking the data of the annual accommodation GDP of California from 1997 to 2023 as samples for fitting analysis, the ARIMA model with the best effect is constructed, which is ARIMA(1,1,0) with drift Use this model to predict California's annual accommodation GDP from 2024 to 2026. The results show that in the coming years, the accommodation GDP of California will still show an upward trend, but the increase will not be too high. It will still show the same changing trend as in the past and will not be affected by factors such as COVID-19 and politics. Based on the above conclusion, the author suggests that families of international students considering studying in California should prepare more funds to deal with the constantly increasing accommodation costs in California or contact local host families in advance to try to reduce the accommodation expenses for studying in California by sharing a house. The author believes that this research provides a quantitative basis for the subsequent development of California's annual accommodation GDP, and from a small perspective, it can also offer corresponding cases for the accommodation GDP of other states in the United States. A limitation of this study is that the use of annual data precludes the analysis of seasonal variations in accommodation GDP. The author infers that by looking at the monthly accommodation GDP data of California, it might be possible to discover obvious seasonal effects, extract the seasonal trends, and thus enable the model to accurately predict the changing trends of California's monthly accommodation GDP. It is also well known that accommodation costs in California are closely related to the place of residence. For instance, in popular cities like San Francisco and Los Angeles, accommodation costs have consistently been at the leading level in California or even in the United States, remaining high for a long time. This study does not conduct a detailed analysis of each city but rather provides an overall analysis of California. Therefore, the prediction results of this study are only applicable to the overall level of California and not to some popular cities.

References

[1] Wu, Y., Yang, Y., Wu, Y., & Boshuai, Y. Predicting the influence of COVID-19 on GDP with ensembled machine learning models. In: Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), Guangzhou, China, 2021. IEEE.

[2] Singh, C., & Garg, A. R. Enhancing solar power output predictions: Analyzing ARIMA and S-ARIMA models for short-term forecasting. In: Proceedings of the 2024 IEEE 11th Power India International Conference (PIICON), Jaipur, India, 2024. IEEE.

[3] Pachpatte, D. Analysis and forecasting of COVID-19 infections in India using ARIMA model. In: Computational Intelligence for Managing Pandemics. Berlin: De Gruyter, 2021.

[4] Kang, H. Analysis of the impact of COVID-19 on the U.S. economy based on big data processing and experimental analyses. In: Proceedings of the 2021 2nd International Conference on Big Data Economy and Information Management (BDEIM), Sanya, China, 2021. IEEE.

[5] U.S. Bureau of Economic Analysis. Gross Domestic Product: Accommodation (721) in California [CAACCOMDNGSP]. Retrieved from Federal Reserve Bank of St. Louis (FRED): https://fred.stlouisfed.org/series/CAACCOMDNGSP, accessed September 1, 2025.

[6] Sirisha, U. M., Belavagi, M. C., & Attigeri, G. Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: A comparison. IEEE Access, 2022, 10, 115000–115012.

[7] Dash, S., Chakraborty, C., Giri, S. K., Pani, S. K., & Frnda, J. BIFM: Big-data driven intelligent forecasting model for COVID-19. IEEE Access, 2021, 9, 105890–105905.

[8] Nie, Z., Qian, X., Gao, Y., Zhou, J., & Zhang, M. Research on EMD-SVR model with ADF test for short-term prediction of ship motion. In: Proceedings of the 2025 5th International Conference on Mechanical, Electronics and Electrical and Automation Control (METMS), Chongqing, China, 2025. IEEE.

[9] Zhang, S., & Liu, Y. Forecasting model of total import and export based on ARIMA algorithm optimized by BP neural network. In: Proceedings of the 2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2023. IEEE.

[10] Anwar, F., & Setiawan, J. Unleashing the power of time-series method: Illuminating Indonesia's open unemployment rate based on educational attainment. In: Proceedings of the 2023 7th International Conference on New Media Studies (CONMEDIA), Bali, Indonesia, 2023. IEEE.

[11] Singh, B., Kumar, P., Sharma, N., & Sharma, K. P. Sales forecast for Amazon sales with time series modeling. In: Proceedings of the 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2020. IEEE.

[12] Xie, Y., et al. Real-time prediction of Docker container resource load based on a hybrid model of ARIMA and triple exponential smoothing. IEEE Transactions on Cloud Computing, 2022, 10(2), 1386–1401.

[13] Falowo, O. I., & Abdo, J. B. 2019–2023 in review: Projecting DDoS threats with ARIMA and ETS forecasting techniques. IEEE Access, 2024, 12, 26759–26772.

[14] Shankar, P., Rohith, K. N., & Karthikeyan, M. Design and development of an ensemble model for stock market prediction using LSTM, ARIMA, and sentiment analysis. In: Deep Learning Tools for Predicting Stock Market Movements. Hoboken, NJ: Wiley, 2024.

[15] Kong, L., et al. Time-aware missing healthcare data prediction based on ARIMA model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024, 21(4), 1042–1050.

[16] Sharma, B., Prakash, C., Dhyani, B., & Joshi, C. Public investments and forecasting in renewable energy technologies: A comparative analysis using ARIMA and GARCH models. IEEE Access, 2025, 13, 124941–124954.