Forecasting the Tourism Development of Yunnan Province with ARIMA and Double ES

Keyi Zhao

cocoa11yi@163.com

Kunming University of Science and Technology Oxbridge College, No. 1369 Yunqiao Street (Yunrui Road), Guandu District, Kunming City 650106, Yunnan Province,P.R.China

Abstract. Yunnan Province is located on the southwestern border of China, with diverse landforms, unique climate and environment, rich ethnic customs, and abundant tourism resources. It is one of the favorable tourist destinations for domestic and foreign tourists in these years. This article takes the total tourism revenue of Yunnan from 1991 to 2022 as a sample. By adopting ARIMA and Double Exponential Smoothing techniques as research modes, to forecasting the trending of total tourism revenue in Yunnan. Compared with the other model mentioned in this article, the Double Exponential Smoothing model is more fitful to forecast. The prediction based on the Double exponential smoothing model indicates that Yunnan tourism will enter a period of rapid growth from 2023 to 2027. The 2023-2025 Yunnan Tourism Development Action Plan formulated by the Yunnan Provincial Government has great potential for success.

Keywords: Foracasting, Tourism, ARIMA, Double Exponencial Smoothing, Yunnan

1 Introduction

Tourism plays an important role in the local economy development, especially for the place in Southwest China, where mostly the mountainous areas, such as Yunnan province. However, the deficient traffic and communication infrastructure are the main obstacles to obtain enough information and resources to develop tourism. For years, large amount of money having been invested in all sorts of the infrastructure projects related to tourism. Nowadays in Yunnan, the high-speed internet having been built up, highways with longer miles and cities with expanded roads. Various national roads leading to mountainous areas having been repaired and subways and high-speed railways have been built in Kunming, the provincial capital. With a faster internet speed and convenient transportation, plus the more sophisticated smart phone with many useful software, such as Tencent Map, TikTok, RED and varieties of online booking system, the more tourists can easily get where he or she wants to go. Still there are many problems in the process of Yunnan tourism development. The uneven resource allocation, the unstable tourism service quality and overexploiting the tourism resources and so forth, will all end to an unsustainable tourism development. Therefore, the tourism sustainable development of Yunnan is a critical target for the local government. Three years plan for sustainable developing tourism in Yunnan province, start from 2023, has been made. The Content of this plan is about strengthening the culture development and combine the local ethic culture and tourism, serving for building up a unique Yunnan tourist brand.

ICEMME 2023, November 17-19, Beijing, People's Republic of China

DOI 10.4108/eai.17-11-2023.2342670

The tourism market has the characteristics of complexity, seasonality, and volatility. Traditional forecasting methods such as regression forecasting and gravity model forecasting do not match the volatility data of the modern travel market, making the prediction inaccurate. [1]. To achieve our research goal, we applied the Autoregressive moving average method as it is a well famous modeling technique for predicting future values of any time series variable based on "past values. [2]. Exponential smoothing is widely used in practice and has shown its efficacy and reliability in many business applications. [3]. is a method that will take into account average (smoothing) the data of the past exponentially by repeating calculations continuously using the latest data.[4]Based on the tourism development over past 30 years in Yunnan, this article aims to forecast the tourism development trend by ARIMA model and Double exponential smoothing model in the following years, from the perspective of total tourism revenue(abbreviated as TTR).

2 Yunnan tourism from 1991 to 2022

The total tourism revenue of Yunnan province from 1991 to 2022 is considered to be a useful sample for analysis. By establishing a proper model for the total tourism revenue of Yunnan Province from 1991 to 2022, to forecast the future five years' revenue. All the data sourced from the Yunnan Provincial Statistical Yearbook and Social and Economic Development Bulletin.

From 1991 to 2019, the tourism revenue gradually grew, and the proportion of tourism revenue to GDP is increasing as well. Due to the constant effort from local government through all these years, the infrastructure was much better than before. Lots of tourist infrastructure had been reformed or upgraded. In 2019, the total tourism revenue was close to half of the GDP, about 47.5%, as shown in Fig.1. At the same year, the Western Yunnan Tourism Ring Road construction plan was brought up by the local government. There are 17 cities involving in the Tourism Ring Road, which on the map the shape looks like a number "8". Unfortunately, the unforeseeable covid-19 epidemic outbroke in December 2019. For the sake of preventing people from unexpected infection, a quarantine policy was made out by the official in 2020. During the quarantine time, the total tourism revenue exposed to a significant reduction in Yunnan province(Fig.1).

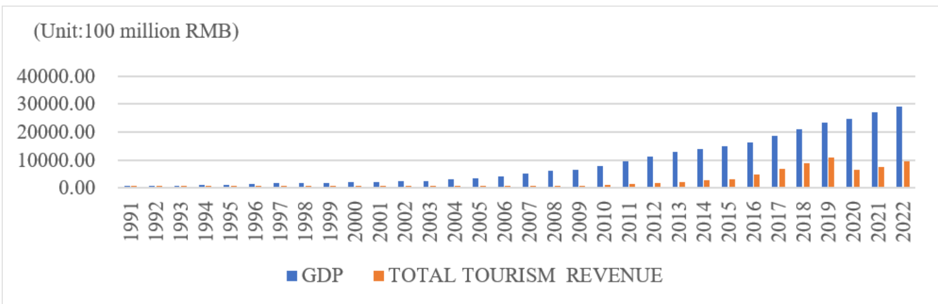

Fig. 1. GDP and Total Tourism Revenue 1991-2022 of Yunnan

3 ARIMA process

3.1 ARIMA mode

ARIMA stands for Autoregressive Integrated Moving Average. It is a popular statistical model commonly used for time series forecasting and analysis. [4] ARIMA models are suitable for data that exhibit non-stationarity, meaning the statistical properties of the data such as the mean and variance change over time. It is often represented as ARIMA (p, d, q), where p is the order of the autoregressive component, d is the degree of differencing, and q is the order of the moving average component.[2]

$$Y'_{(t)} = c + \Phi_1 Y'_{(t-1)} + \Phi_2 Y'_{(t-2)} + ... + \Phi_p Y'_{(t-p)} + \mu_{(t)} + \theta_1 \varepsilon_{(t-1)} + \theta_2 \varepsilon_{(t-2)} + ... + \theta_q \varepsilon_{(t-q)} \quad (1)$$

In the equation (1), the value $Y'_{(t)}$ represents the differenced series, c is a constant term, $\Phi_1, \Phi_2 ... \Phi_p$ are the coefficients of the autoregressive terms, $\mu_{(t)}$ is the error term at time t, and θ_1, $\theta_2 \varepsilon_{(t-2)} \ldots \theta_q$, are the coefficients of the moving average terms. The differencing component I(d) is implicitly included in $Y'_{(t)}$. As it shows in Fig.2 the Total Tourism Revenue is not smooth, specially from 2019 to 2020. In avoiding the negative influence from the incidence impact the prediction accuracy of an ARIMA model. Because in an ARIMA model with a log transformation (short for Ln-ARIMA), can help stabilize the variance of the data and linearize relationships that show exponential growth or decay.

$$LnY'_{(t)} = \mu + \phi_1 * \Delta LnY'_{(t-1)} + ... + \phi_p * \Delta LnY'_{(t-p)} + (-\theta 1 * \varepsilon^*_{(t-1)} - ... - \theta_q * \varepsilon^*_{(t-q)}) \quad (2)$$

In the equation (2), the $LnY'_{(t)}$ represents the transformed observation in logarithmic form, obtained by taking the logarithm of the original data. The $\Delta LnY'_{(t-1)}$ represents the first difference operator, indicating the difference between consecutive logarithmic transformed observations. μ is a constant term (if present). $\phi_1, ..., \phi_p$ are the autoregressive (AR) coefficients, representing the correlation between the current time point and the past p time points after logarithmic transformation and differencing. $\varepsilon^*_{(t-1)}, ..., \varepsilon^*_{(t-q)}$ are lagged error terms obtained during the fitting of the ARIMA model. $\theta_1 ... - \theta_q$ are the moving average (MA) coefficients, representing the correlation between lagged error terms and the current time point and the past q time points after logarithmic transformation and differencing. This combined formula integrates the logarithmic transformed data with autoregressive, moving average terms, and a constant term (if present) to capture the trend and relationships in the data.

3.2 ADF test

From investigating the Fig.2, Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a time series is stationary or non-stationary. The ADF test is based on the assumption that the time series can be represented by an autoregressive model. It determines whether the autoregressive coefficient in the model is significantly different from 1, which indicates non-stationarity. Except for the years affected by the epidemic, the overall tourism revenue of previous years showed a certain exponential trend, manifested as a non-stationary series. Therefore, the ADF test is applied to confirm the stationarity of the total tourism revenue sequence, represented by a variable “ttr” and the log transformed variable “lnttr”. The test result is showing in table 1, where the “D.” means performing first-order difference on variables and the “D2.” means performing second-order difference on variables.

Table 1. Dickey–Fuller test for unit root (ADF)

Variable	Statistic Z(t)	Dickey–Fuller critical value			MacKinnon approximate
		1%	5%	10%	p-value for Z(t)
ttr	-1.091	-4.325	-3.576	-3.226	0.9738
D.ttr	-5.161	-3.716	-2.986	-2.624	0
D2.ttr	-7.915	-3.723	-2.989	-2.625	0
lnttr	-3.15	-3.709	-2.983	-2.623	0.0231.
D.lnttr	-5.3	-3.716	-2.986	-2.624	0
D2.lnttr	-10.445	-3.723	-2.989	-2.625	0

The Dickey-Fuller test is used to determine if a time series data has a unit root, which implies that it follows a random walk (without drift) and is non-stationary. This test calculates the statistic and compares it with critical values to determine the significance of the test. If the statistic is less than the critical value, the null hypothesis of non-stationarity has been rejected which means the series is stationary. And reject the null hypothesis (H0) means a unit root do not exist in the data. On the other hand, if the test statistic is greater than the critical value, the series is non-stationary. A unit root implies that it follows a random walk (without drift) and is non-stationary. Each statistic compared with the critical values at different significance levels (1%, 5%, and 10%) to determine the rejection region. If the test statistic is more negative than the critical values, the null hypothesis is rejected, indicating the presence of stationarity in the data. Also, if The MacKinnon approximate p-value associated with the statistic is 0.0000, the result further supports the rejection of the null hypothesis. Therefore, in table 1, except for the variable “ttr” and “lnttr”, there are no unit roots for all the other variables, which means the sequences are stationary.

3.3 Autocorrlation and Partial Autocorrelation analysis

Autocorrelation function, short for ACF, measures the relationship between an current value and its lagged values. The autocorrelation values range from -1 to 1, where a value of 1 indicates a perfect positive autocorrelation, 0 indicates no autocorrelation, and -1 indicates a perfect negative autocorrelation. As for the partial autocorrelation, short for PACF, it helps identify the correlation between the current value and its lags after removing the effect of intermediate lags. After finding the stationary sequences, it is necessary to make autocorrectio(abriviated as AC) and partial autocorrelation(abriviated as PAC) graphs to analyze and roughly found out proper lag terms, serve for the value “p“and “q“in an ARIMA(p,d,q) model. For p value, in an ACF graph, if the lagged term with a significant peak autocorrelation coefficient, the lag term can be considered as a choice; and in a PACF graph, if the partial autocorrelation coefficient of the lagged term rapidly decays to zero, the lag order is an option. For q value, in an ACF graph, the lag term with a significant peak autocorrelation coefficient can be used; and in a PACF graph, the lag term with a partial autocorrelation coefficient rapidly decays to zero can be considered. After series attempts, for the selection of the p value, the PACF graph is more reliable. It is useful to choose the lag term that the partial autocorrelation coefficient rapidly decays to zero. The same method can be applied to the q value, only different is that it is about an autocorrelation coefficient.

Although the D.ttr sequence is stationary accouding to the ADF result, but by comparing the D.ttr graph with the D2.ttr graph of AC, the D2.ttr of AC is relatively significant. The same situation occurs between the D.lnttr and D2.linttr. so it is reasonable to choose the AC and PAC graph of “D2.ttr” and “D2.lnttr” to analyze the possible parameter p and q.

Autocorrelation values that exceed the 95% confidence intervals, are considered statistically significant. In Fig.2, there is a negative autocorrelation at the lag 1, that slightly exceed the 95% confidence intervals, and the autocorrelation number is -0.4. In the PAC graph of Fig.2, at the lag 4, the partial autocorrelation is smaller than -0.5. It suggests that the current observation is negatively related to the past observation, but in a mild way. As a result, the p value of “D2.ttr” may be selected the lag term from 0, 1and 2; the q value of “D2.ttr” may be choosed from the lag term 0,1,2 and 3. And in Fig.3, there is a negative autocorrelation at the lag 1, and the autocorrelation number is a bit over -0.5; the partial autocorrelation number is also smaller than -0.5. It also suggests that the current observation is negatively related to the past observation with a weak autocorrelation. And the p value of “D2.lnttr” may be selected the lag term 1, 2 and 4; the q value of “D2.lnttr” may come from the lag term 0, 1 and 2.

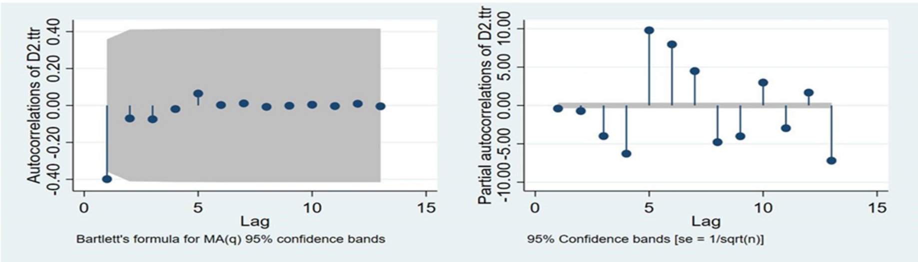

Fig.2. The AC and PAC graph of “D2.ttr”

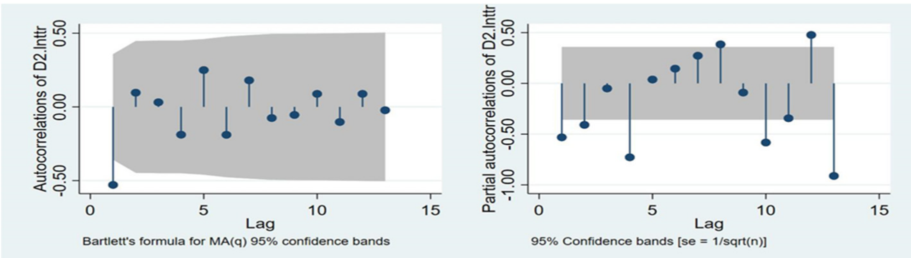

Fig.3. The AC and PAC graph of “D2. lnttr”

The possible ARIMA models are listed in table 2 and table 3, under the standard of “Prob > chi2” equal to zero. Besides, the lower the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) number, the better the model. In the words, the models were built based upon autocorrelation function (ACF) and partial autocorrelation function (PACF). And the model with the least Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values were selected.[5] According to the AIC and BIC criteria, the Arima(0,2,1) is suitble for “D2.ttr”; and the Arima(1,2,2) is acceptable for “D2.lnttr”. Refering to the equition (1) and the equition (2), the model of “D2.ttr“ can be illstrated by the equitions (3); the model of “D2.lnttr“ is the equition (4). Because the equation (4) is log transformed way of ARIMA model, it has to return to the ARIMA model, so the predictive values can be generate. So the equition of the restored predictive value is the equition (5) . $^\wedge Y'_{(t)}$, $^\wedge LnY'_{(t)}$ and $^\wedge Y''_{(t)}$ repressent the predictive value, other variables are same as the equation (1) and (2). And the standard error is the number in the column" ()".

$$\hat{}Y'_{(t)} = 31.86255 + \mu_{(t)} - \varepsilon_{(t-1)} \quad (3)$$

(27.14987) (0.166616)

$$\hat{}LnY'_{(t)} = -0.032925 + 0.7302808\Delta LnY'_{(t-1)} + (-1.795347\varepsilon^*_{(t-1)} - \varepsilon^*_{(t-2)}) \quad (4)$$

(0.0398357) (0.7302808) (0.1696533)

$$\hat{}Y''_{(t)} = 22.06735 - 0.189116\, Y'_{(t-1)} + \mu_{(t)} - 0.6958072\varepsilon_{(t-1)} - 0.3041947\varepsilon_{(t-2)} \quad (5)$$

(135.896) (9.549354) (10.38758) (10.29156)

From 1991 to 2019, compare to the actual total tourism revenue, the Arima (0,2,1) and the Ln-Arima (1,2,2) as in Fig.4, the Ln-Arima prediction is very close to the TTR (total tourism revenue) . But after 2020, both the Arima and Ln-Arima prediction is not very accurate. The Ln-Arima line shows a particular optimistic trending. Due to the low fitting of the ARIMA model from 2020, there has to be another option to predict the total tourism revenue. And the Ariam mode can only forecast one year's data in STATA17, that is not enough. Inspired by the research about stock market where severe fluctuations happen all the time in the trading hours. So the double exponencial smoothing, also known as the Holt's linear exponential smoothing, may be more fitful to analyze the fluctuation in data, caused by the pandemic.

Table 2. The possible ARIMA models of "D2.ttr"

Arima(p,d,q)	AIC	BIC	Prob > chi2
(0,2,1)	513.1996	516.002	0
(0,2,2)	515.1512	519.3548	0
(1,2,1)	515.1755	519.3791	0
(1,2,3)	515.1981	522.2041	0
(2,2,1)	515.451	521.0557	0
(2,2,2)	515.1486	522.1546	0

Table 3. The possible ARIMA models of "D2. lnttr"

Arima(p,d,q)	AIC	BIC	Prob > chi2
(1,2,0)	1.904572	6.108164	0
(1,2,2)	-2.119202	3.485587	0
(2,2,2)	-0.1952894	6.810697	0
(4,2,1)	-0.2475108	9.560871	0

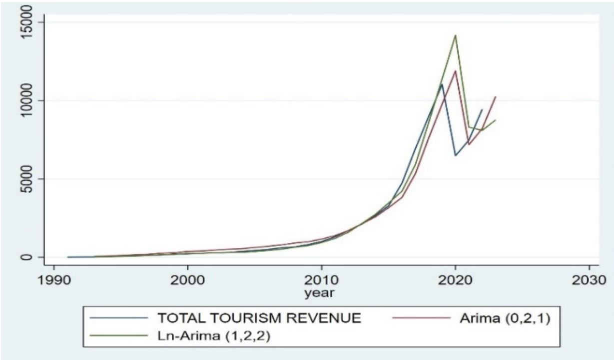

Fig. 4. Comparison of the predication line with reality

4 Double Exponential Smoothing

Double ES (Double exponential smoothing) is a forecasting method used to analyze time series data. It extends simple exponential smoothing by incorporating a trend component into the forecast. Same as the simple exponential smoothing, double exponential smoothing utilizes smoothing factors, α and β, to adjust the level and trend components of the forecast. The equation (6) is the forecasting formulars. The possible forecasting equation of total tourism revenue is the equation (7). Suppose $L_{(0)} = Y_{(0)}$, $T_{(0)} = Y_{(1)} - Y_{(0)}$.

$$F_{(t+k)} = \alpha * Y_{(t)} + (1 - \alpha) * (L_{(t-1)} + T_{(t-1)}) + k * [\beta * (L_{(t)} - L_{(t-1)}) + (1 - \beta) * T_{(t-1)}] \quad (6)$$

$$F_{(t+k)} = 0.9 * Y_{(t)} + 0.001 * (L_{(t-1)} + T_{(t-1)}) + k * [5.25 * (L_{(t)} - L_{(t-1)}) - 4.25 * T_{(t-1)}] \quad (7)$$

In the above equations, $L_{(t)}$ represents the level at time t. "(t)" represents the observed value of the time series at time t. $T_{(t)}$ represents the trend at time t. "α" is a parameter between 0 and 1 that determines the weight given to the current observation $Y_{(t)}$ versus the weighted average of previous level $L_{(t-1)}$ and trend $T_{(t-1)}$. A higher value of α, the more emphasis on the current observation, while a lower value places more emphasis on the previous level and trend. "β" is a parameter between 0 and 1 that determines the weight given to the difference between the current level $L_{(t)}$ and the previous level $L_{(t-1)}$ versus the previous trend $T_{(t-1)}$. A higher value of β places more emphasis on the recent change in level, while a lower value places more emphasis on the previous trend. $F_{(t+k)}$ represents the forecast value at time $_{t+k}$, and k denotes the future time periods.Double exponential smoothing is suitable for forecasting time series data with a linear trend. It captures the trend component and applies smoothing to reduce the impact of random fluctuations.[7] And if the data fluctuation is not significant, the parameter α is generally taken as a smaller value, such as between 0.1 and 0.5. If the data fluctuation is large and the α value is relatively large, such as 0.9.

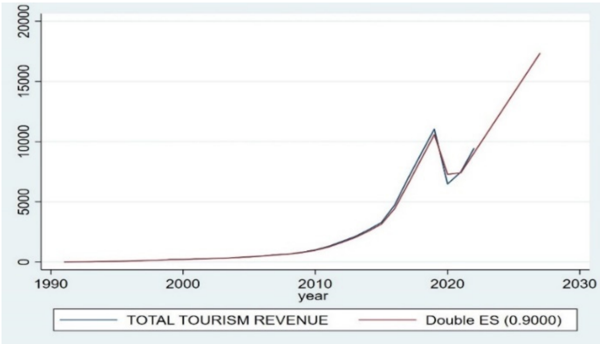

Fig. 5. Comparison of the predication line with reality

In order to find a proper perimeter α to build a Double ES (exponential smoothing) model to forecast the next five years trending of total tourism revenue by using STATA17. A relatively large α has been set as 0.9, very near 1, which means the current observation is more important than the previous observation. After the pandemic, the tourism revenue will restore soon. In Fig.5, the trending lines of Double ES and Total tourism revenue are very close to each other from 1991 to 2019, except 2020. Based on 2022, the total tourism revenue will grow in the next three years. The performance of these models for predicting is measured using the Root Mean Square Error (RMSE).[6]There are more fitful choice of α compare to the chosen one. The larger α ,such as 0.999 and 0.9999, the closer to the total tourism line. And both the SSR (sum-of-squared residuals) and the RMSE (root mean squared error) are less than the results when α equal to 0.9, which is namely 1568344 and 221.38. Still The overfitting problem of the model needs to be taken into consideration. Beacuse exponential smoothing potentially suffers from overfitting and might produce inaccurate forecasts[7] Mainly because there are only three months left for stakeholders to earn more tourism revenue, and winter is not a good season for traveling. In reality, after the three years pandemic, the curbed tourist demand for traveling was released. A large number of tourists works related to Yunnan produced by freelance tourism bloggers, as well as the television shows related to Yunnan, attract more and more tourists visit Yunnan. In the first half of 2023, Yunnan received 539 million tourists 40.0% more than the same period of last year. And achieved 639.432 billion RMB on total tourism revenue, 48.8% more than the same period of last year. Consider the influence of the summer vacation and National Day, as well as the tourism revenue earned from January to June, it is very hopeful to reach the estimated prediction revenue in 2023.

5 Model comparison

The residual formulas can be used to evaluate the performance of the model. After calculating the residual of each model, it is necessary to make a ADF test on each residual. Validation is usually done to assess the precision of the fit in estimating the observed values. [8] By observing the p-value of each residual and comparing the statistic to the critical value to decide the most suitable model for prediction. In table 4, the res1 represent the residual of ARIMA (0,1,2), the res2 represents the residual of Ln-ARIMA (1,2,2), the res3 represent the residual of Double ES (0.9). The test result reveals residuals that the statistics of the res1, res2 and res3 are significant.

It means all three models can be adopted to predict the trending. But from the perspective of p-value, the res2 is a bit more than zero. At the same time, the p-value of res1 and res3 is zero, which implies the residual sequences of ARIMA (0,2,1) and Double ES (0.9) are slightly more stationary. As it shows in Fig.6, the residual line of Double ES(0.9) is less flucuate than the others. The prediction result of Double ES (0.9) is larger than the others, it may more close to the reality, as it shows in table 5.

Table 4. ADF (Dickey–Fuller test for unit root) of Residuals

Variable	Label or the variable	Statistic Z(t)	Dickey–Fuller critical value			MacKinnon approximate
			1%	5%	10%	p-value for Z(t)
res1	ARIMA(0,2,1)	-5.411	-3.723	-2.989	-2.625	0
res2	Ln-ARIMA(1,2,2)	-4.495	-3.723	-2.989	-2.625	0.0002
res3	Double ES(0.9)	-4.666	-3.709	-2.983	-2.623	0

Table 5. Predictions of models

YEAR	ARIMA(0,2,1)	Ln-ARIMA(1,2,2)	Double ES(0.9)
2022	8217.417	8098.194	9061.731
2023	10263.39	8763.495	10719.4
2024	/	/	12377.07
2025	/	/	14034.74
2026	/	/	15692.41
2027	/	/	17350.08

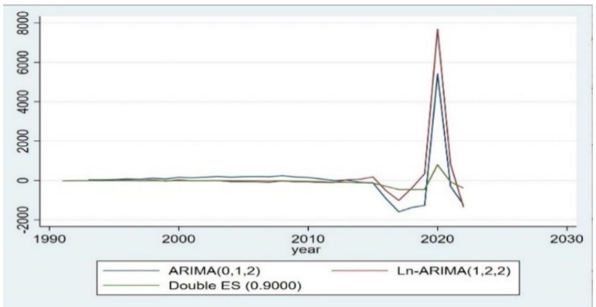

Fig. 6. Comparison the residuals of different models

6 Conclusion

All the results indicate the tourism revenue will increase in the future. In the case of the ARIMA model and the log transformed ARIMA model, the prediction lines apparently deviate from the line of Total Tourism Revenue. The double exponential smoothing techniques outperform the ARIMA models, because it adapts to the sudden changes in reality. However, select a proper parameter α is very important to the Double ES model. Therefore, the forecasting models based on double exponential smoothing techniques are an effective tool to predict the trending of tourism development after the pandemic. Although in the past three years, the COVID-19 had a great impact on the development of tourism in Yunnan Province. But overall, the long-term growth trend of tourism in Yunnan Province will not change fundamentally. With the assistant of government, it can be expected that in the absence of events like the pandemic that have a significant impact on tourism, the tourism of Yunnan will inevitably experience obvious development at least in the next three years. promote the sustainable development of tourism in Yunnan. Hope the finding will provide a useful reference for government and enterprise managers to make decision.

References

[1] Qin Q .FORECASTING TOURISM MARKET DEMAND IN HUNAN PROVINCE USING ARIMA MODEL[J].Delta Jurnal Ilmiah Pendidikan Matematika, 2021, 9(2):211.DOI:10.31941/delta.v9i2.1410.

[2] Janjua L R , Muhammad F , Sukjai P ,et al.Impact of COVID-19 pandemic on logistics performance, economic growth and tourism industry of Thailand: an empirical forecasting using ARIMA[J].Brazilian Journal of Operations & Production Management, 2021, 18(2):e2021999.DOI:10.14488/BJOPM.2021.001.

[3] Pritularga K F , Svetunkov I , Kourentzes N ,et al.Shrinkage estimator for exponential smoothing models[J]. 2023.

[4] Ahmar A S , Fitmayanti F , Ruliana R .Modeling of inflation cases in South Sulawesi Province using single exponential smoothing and double exponential smoothing methods[J].Quality & Quantity: International Journal of Methodology, 2022, 56.

[5] Patil R , Nagaraj D M , Polisgowdar B S ,et al.Forecasting potential evapotranspiration for Raichur district using seasonal ARIMA model[J].Mausam: Journal of the Meteorological Department of India,2022.

[6] Alya AtouiKamal SlimSamir Abbad AndaloussiRégis MoilleronZaher Khraibani.Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air Pollutants in Zahleh, Lebanon[J]. Atmospheric and Climate Science (English), 2022, 12(4):728-749.

[7]Barrow D K , Kourentzes N , Sandberg R ,et al.Automatic robust estimation for exponential smoothing: perspectives from statistics and machine learning[J].Expert Systems with Applications, 2020.DOI:10.1016/j.eswa.2020.113637.

[8] Saxena K K , Kamnge J S .Comparative study of exponential smoothing models and Box-Jenkins ARIMA model of partitioned data of daily stock prices of the CRDB Bank in Tanzania[J].Bulletin of Pure & Applied Sciences- Mathematics and Statistics, 2020, 39e(1):1.DOI:10.5958/2320-3226.2020.00001.6.