BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0053–0064
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0053-0064

COMPARISON OF LOCAL POLYNOMIAL REGRESSION AND ARIMA IN PREDICTING THE NUMBER OF FOREIGN TOURIST VISITS TO INDONESIA

Bagas Shata Pratama[1], Alda Fuadiyah Suryono[2], Nina Auliyah[3], Nur Chamidah[4*]

[1,2,3]*Student of Statistics Study Program, Faculty of Science and Technology, Universitas Airlangga*
[4]*Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga*
St. Dr. Ir. H. Soekarno, Mulyorejo, Surabaya, 60115, Indonesia

*Corresponding author's e-mail: * **alda.fuadiyah.suryono-2021@fst.unair.ac.id***

Article History:

Received: 9th June 2023
Revised: 12th November 2023
Accepted: 11th December 2023

Keywords:

ARIMA;
Local Polynomial;
Prediction;
Tourism

ABSTRACT

Indonesia is a country that has a variety of exotic tourist destinations and can attract tourists to visit. Currently, tourism is one of the sectors that play a major role in driving the Indonesian economy. Various domestic and foreign tourists are expected to continue to increase in number every year. Therefore, appropriate policies are needed from the government to develop the tourism sector so that it can be even better over time. This research aims to predict the number of foreign tourists visiting Indonesia using the Autoregressive Integrated Moving Average (ARIMA) model and local polynomial regression. The data used in this research is the number of foreign tourist visits per month from January 2017 to December 2022 obtained from the Kemenparekraf website. This data fluctuates, so the method of a local polynomial approach is appropriate for this study. The data analysis methods used are local polynomial regression and the ARIMA model. In the ARIMA model, there are assumptions that must be met. In this study, the ARIMA model obtained has met the assumption of residual normality but does not meet the assumption of homoscedasticity, so ARIMA modeling cannot be continued, and analysis is only carried out with local polynomial regression. The result of this study is a prediction of future tourist visits. The MAPE value of the local polynomial regression approach is 1.43%, which is categorized as a prediction with high accuracy because the value is less than 10%. Thus, the local polynomial regression approach is very well used to predict the number of foreign tourist visits to Indonesia.

How to cite this article:
B. S. Pratama, A. F. Suryono, N. Auliyah., and N. Chamidah, "COMPARISON OF LOCAL POLYNOMIAL REGRESSION AND ARIMA IN PREDICTING THE NUMBER OF FOREIGN TOURIST VISITS TO INDONESIA," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0053-0064, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Indonesia's tourism is one of the most instrumental sectors that drive the nation's economy. The tourism sector is even one of the second largest contributors to the country's foreign exchange after palm oil [1]. Throughout 2019, foreign exchange in the tourism sector reached IDR 280 trillion and contributed 4.8% to the national Gross Domestic Product (GDP) [2]. The government is still continuing to develop the tourism sector, with various policies being carried out to make Indonesian tourism more advanced and recognized in the eyes of the world.

Indonesia is a country that has a variety of tourist destinations that can attract domestic and foreign tourists to visit and choose Indonesia as a tourist destination. In 2017 - 2022, 57.04 million foreign tourists came to Indonesia, with the highest number in 2019 at 16.1 million [3]. On the other hand, the tourism sector is also able to open many jobs for the community. In 2021, there were 21.26 million people, or 16.22% of the total population in Indonesia, working in the tourism sector [4]. Based on this, it is not wrong if the tourism sector has the potential to become a driving force for the national economy.

As Indonesian tourism continues to grow and develop, the world was shocked by the COVID-19 pandemic at the end of 2019, which impacted all aspects of life, including the tourism sector. The number of tourists, especially those from abroad, has decreased drastically. In the period January 2020 to April 2020, there was a very drastic decline, from the initial number of 1.2 million to just 100 thousand [3]. This decline occurred due to government policies in controlling the rate of Covid-19, namely an appeal to be able to minimize travel both at home and abroad. The number began to show an upward trend in early 2022 when the COVID-19 pandemic had begun to subside and the government had also relaxed the previously implemented policies.

To find out the fluctuating pattern of the number of foreign tourist visits to Indonesia, modeling or prediction is needed. Forecasting or prediction is the art and science of estimating future events using a form of mathematical model [5]. One method that can be used for prediction is local polynomial regression. The advantage of a local polynomial is its adaptability to data that divides the data into certain regions and then estimates the predetermined regions [6]**.** In addition to using the local polynomial regression approach, this research also makes predictions using the Autoregressive Integrated Moving Average (ARIMA) method. The advantages of ARIMA are its flexible nature following the pattern of the data, a fairly high level of forecasting accuracy, simplicity, suitability for use in forecasting a number of variables quickly, and cheapness because it only requires historical data [7]**.** Relevant previous research related to predicting the number of foreign tourist visits to Indonesia has been conducted by [7] which examines the same topic but uses the ARIMA method only. The novelty of this research is that it uses the local polynomial method, which is able to overcome the fluctuating pattern of data quite well. Therefore, this study aims to predict the number of foreign tourist visits to Indonesia using the local polynomial regression and ARIMA. Then this research is also expected to be a reference for the government to make the right policies in the tourism sector so that it can continue to develop for the better [8].

2. RESEARCH METHODS

2.1 Data Source

The data used in this research is data obtained from the website satudata.kemenparekraf.go.id, which is one of the online sites of the Ministry of Tourism and Creative Economy of the Republic of Indonesia in providing tourism and creative economy data in one portal. This study uses data on the number of foreign tourist visits per month from January 2017 to December 2022 with 72 observations. The division of research data is carried out by dividing the proportion of 90% for training data and 10% for testing data. Therefore, the in-sample data in this study were 65 observations, and the out-sample data were 21 observations. The in-sample data used starts from January 2017 to May 2022, while the out-sample data starts from June 2022 to December 2022.

2.2 Research Variable

In this study, with the ARIMA approach, the variable used is the number of foreign tourists to Indonesia from January 2017 to December 2022. Meanwhile, for the local polynomial approach, two research variables used, namely the period (time) of foreign tourist visits to Indonesia as an independent variable and the number of foreign tourists to Indonesia as the dependent variable. Independent variables affect or cause changes or the emergence of dependent variables [9]. Meanwhile, the dependent variable is the variable that is affected or becomes the result [10].

2.3 Analysis Technique

This research was conducted using Rstudio and Minitab software using the concept of local polynomial estimator and ARIMA Model.

The analysis steps in this study are:

1. Describe the data into two, namely, in-sample data and out-sample data. In-sample data is from January 2017 to July 2022, while out-sample data is from August 2022 to December 2022.
2. Estimate and model the data using local polynomial estimators as follows:
   a. Make a scatter plot on the in-sample data. If the results of the scatter plot show a monotonous upward or downward trend, then use order 1; if it shows fluctuating ups and downs, then use order 2.
   b. Selecting the optimal bandwidth (h) and order value with minimum CV based on the equation in the in-sample data.
   c. Model the data with a local polynomial estimator based on the optimal bandwidth value and order in point b to get the best local polynomial model.
   d. Determining the model fit using goodness of fit measures such as R2, MSE, and MAPE.
   e. Comparing the estimation results with observations on out-sample data.
   f. Creating an estimation plot between the observed value and the predicted value based on the in-sample data with the best local polynomial estimation equation. Then, compare the prediction results parametrically using the auto ARIMA function and calculate the accuracy of the model using the MAPE value.
3. Data Analysis and Interpretation

   After obtaining the best local polynomial estimation equation, an analysis and interpretation of the prediction of the number of foreign tourist visits to Indonesia in weeks is carried out.

2.3.1 Cross-Validation

Cross-Validation is a statistical method to evaluate and compare learning algorithms by dividing the data into two segments, one used to learn or train the model and the other used to validate the model [11]. Cross-validation that can be used on time series is time series cross-validation. In time series cross-validation, training data only consists of observations that occur before the observations that make up the testing data. This results in no future observations that can be used in forecasting [12]. This cross-validation method can be used in selecting the optimal bandwidth (h). The cross-validation method (CV) is defined as follows:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{g}_{h,-i}(x_i)\right]^2 \quad (1)$$

Where y_i is the value of the response variable at the i-th observation and $\hat{g}_{h,-i}(x_i)$ is the estimated value of the regression function at point x_i by excluding the i-th observation [13].

2.3.2 Local Polynomial Estimators

One of the nonparametric regression methods that can be used is the estimator [6]. Estimation of the local polynomial regression model can use WLS (Weighted Least Square) so that weights are needed. The

weighting that can be used to obtain local polynomial estimates is the kernel function [14]. The kernel function K with bandwidth h is defined as follows:

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right);\ -\infty < x < \infty\ dan\ h > 0 \quad (2)$$

According to [15], there are several types of kernel functions, namely Kernel Uniform, Kernel Triangular, Kernel Epanechnikov, and Kernel Gaussian.

2.3.3 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model was first introduced by Box and Jenkins in 1970. This model can usually be applied well to data conditions that fluctuate stationary. If the time series analysis model does not fulfill stationary properties, then differencing can be performed on the original data to produce data that is closer or stationary. According to [16], the general model of ARIMA (p,d,q) is as follows:

$$\phi_p(B)(1-B)^d Z_t = \theta_q(B)\varepsilon_t \quad (3)$$

With:
$\phi_p(B)$as AR (p) operator
$\theta_q(B)$as MA (q) operator
$(1-B)^d Z_t$ as the operator of differencing (d)

2.3.4 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a statistical measurement of the accuracy of estimates or predictions in forecasting methods. Here is the formula for calculating MAPE [17]:

with:

$$MAPE = \frac{\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100}{n} \quad (4)$$

A_t : Original data value
F_t : Value of forecasting results
n : Sample size
with MAPE value categories shown in Table 1.

Table 1. MAPE Value Category Table

Range MAPE	Interpretation
$< 10\%$	Highly Accurate
$10 - 20\%$	Accurate
$20 - 50\%$	Reasonable
$> 50\%$	Inaccurate

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis of Data

From January 2017 to December 2022, there are 72 tourist visit data. The highest visit value was achieved in July 2018 with a value of 1574231, and the lowest visit value was achieved in February 2022 with a value of 105195. The average tourist visit from January 2017 to December 2022 is at point 792330. The following will show a plot of the data.

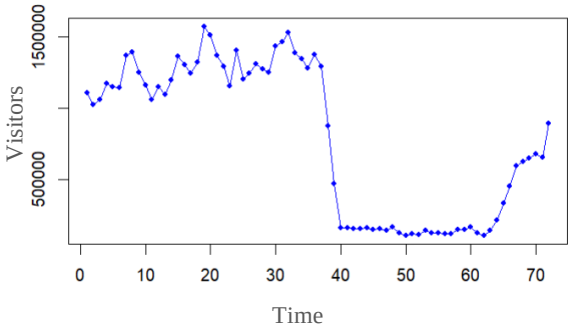

Figure 1. Plot of Tourist Visit Data

From the data, it can be seen that the data has a fluctuating upward trend, but there is a significant decline from the 38th data to the 40th data, which is an accumulation of decline from February 2020 to April 2020. The decline in visits occurred because the COVID-19 virus began to enter Indonesia that month, resulting in all activities in Indonesia being hampered. After the Covid-19 pandemic began to be resolved stably, community activities slowly returned to normal, so that tourist visits increased starting in March 2022.

3.2 Analysis Using Local Polynomial Regression

Based on **Figure 1**, it can be seen that the time series plot of tourist visit data shows a fluctuating pattern. Thus, in order for the prediction results to cover the lowest point and the highest point, a local polynomial estimator of order two can be used; the author estimates the model for the data by selecting the smoothest estimate with the CV method.

In bandwidth selection with order 0, the optimal bandwidth using the CV method is 0.154 with a minimum CV value of 5310605696.

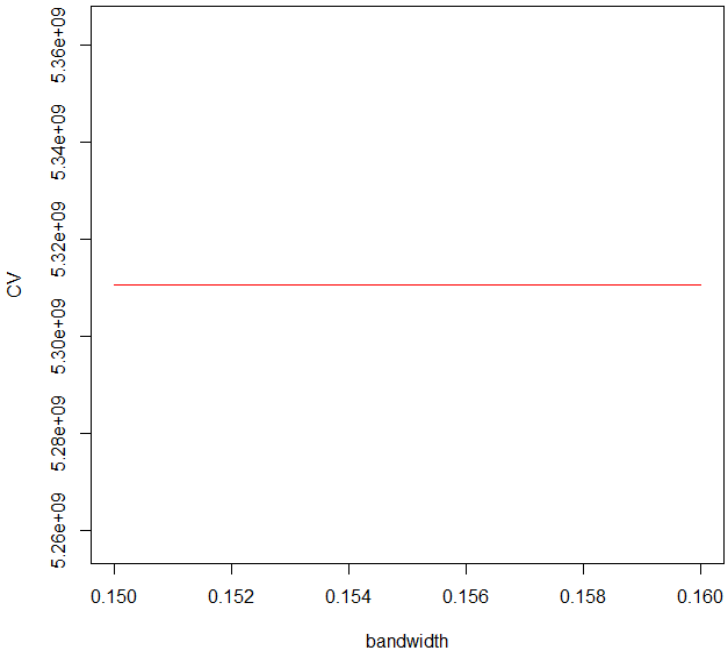

Figure 2. Plot of Optimal Bandwidth Value of CV Method with Polynomial Degree 0

In bandwidth selection with order 1, the optimal bandwidth using the CV method is 0.252 with a minimum CV value of 5241099351.

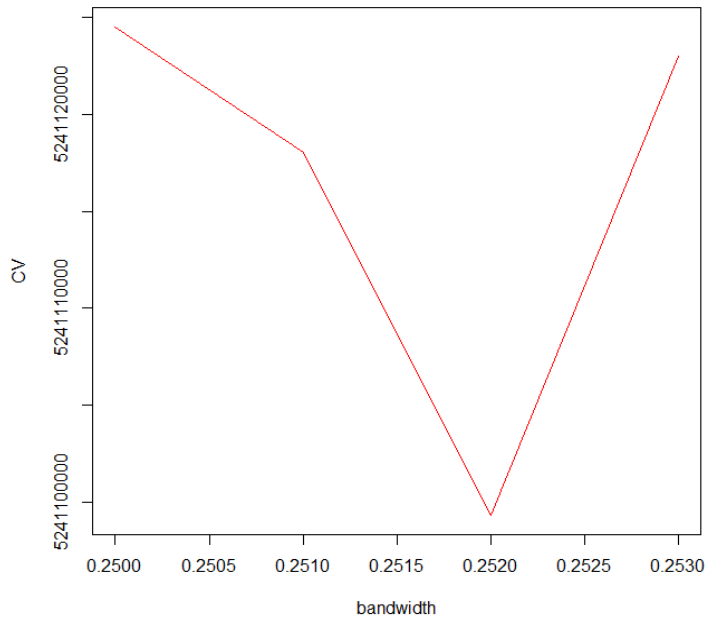

Figure 3. Plot of Optimal Bandwidth Value of CV Method with Polynomial Degree 1

In bandwidth selection with order 2, the optimal bandwidth is obtained using the CV method of 0.899 with a minimum CV value of 4556446632.

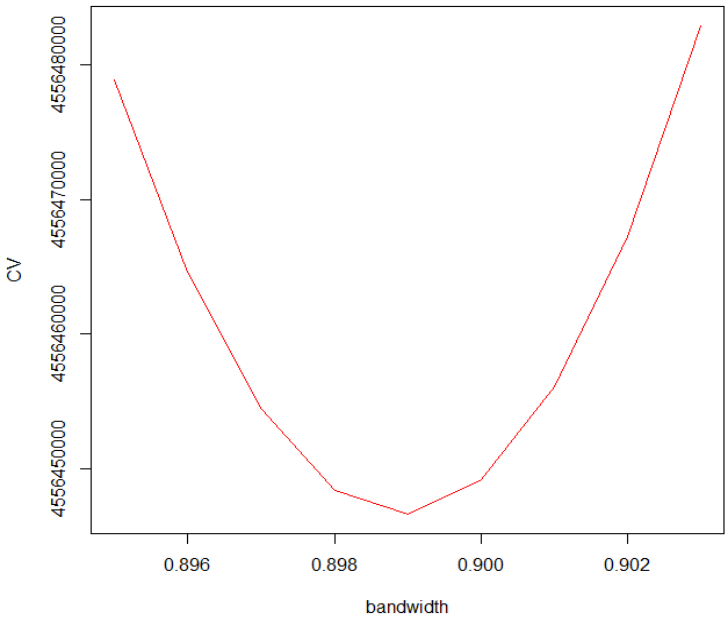

Figure 4. Plot of Optimal Bandwidth Value of CV Method with Polynomial Degree 2

The next step is to choose the most optimal bandwidth with the minimum CV value from the three bandwidth selections above so that the following results can be obtained:

Table 2. Optimal Bandwidth Selection Based on Minimum CV

	Orde		
	0	**1**	**2**
Optimum Bandwidth	0.154	0.252	0.899
CV Minimum	5310605696	5241099351	4556446632

The best model is the model that has the optimal bandwidth value with the minimum CV value. After several experiments with several orders, the most effective model was found to be the local polynomial with a polynomial degree of 2, then the optimal bandwidth value of 0.899 and the minimum CV value of 4556446632. Then, estimation is carried out on the data according to the optimal order and bandwidth that has been obtained previously. In-sample data estimation is used to form a local polynomial regression model, out-sample data estimation is used to validate the model, and all-sample estimation is used to provide an overview of the overall model performance **[18]**.

3.2.1 In-Sample Data Model Estimation and Interpretation Results

Modeling of tourist visit data is carried out after obtaining the optimal bandwidth value in the previous stage. Furthermore, estimation is carried out by determining the parameter estimator $\hat{\beta}$ using R software based on the optimal bandwidth obtained previously. After estimating $\hat{\beta}$, the monthly visit data model is obtained.

The following is a plot between the in-sample data Y and the estimated value $\hat{Y}$ – obtained after modeling.

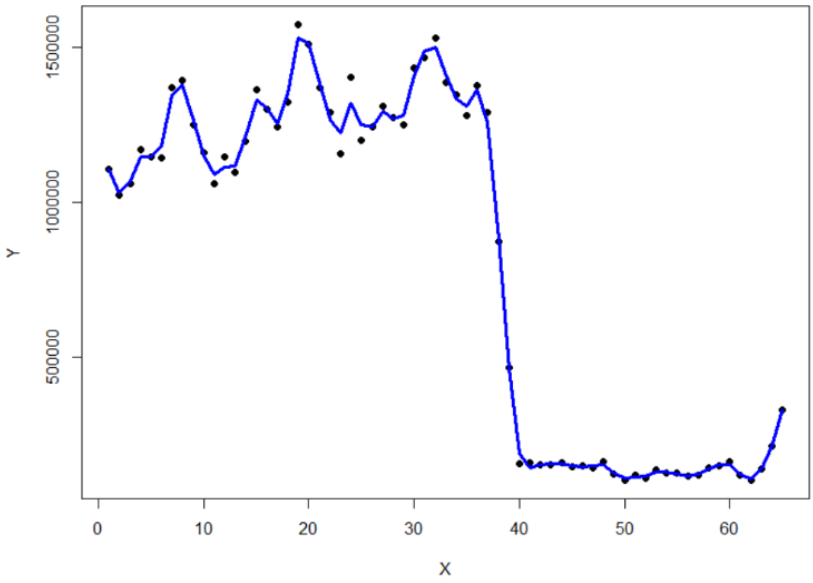

Figure 5. Plot of In-Sample Data Estimation Results

In the model estimation results with order 2, and bandwidth 0.899, the MSE of the in-sample data is 503507077, R2 is 0.9939978 or 99.34%, and the MAPE is obtained well, which is 2.535%. The forecasting results using this model will produce highly accurate forecasting because the MAPE value of 2.535% is below 10%.

3.2.2 Out-Sample Data Model Estimation and Interpretation Results

The following is a plot between the out-sample data Y called Yeval and the estimated value $\hat{Y}_{out}$ obtained after modeling.

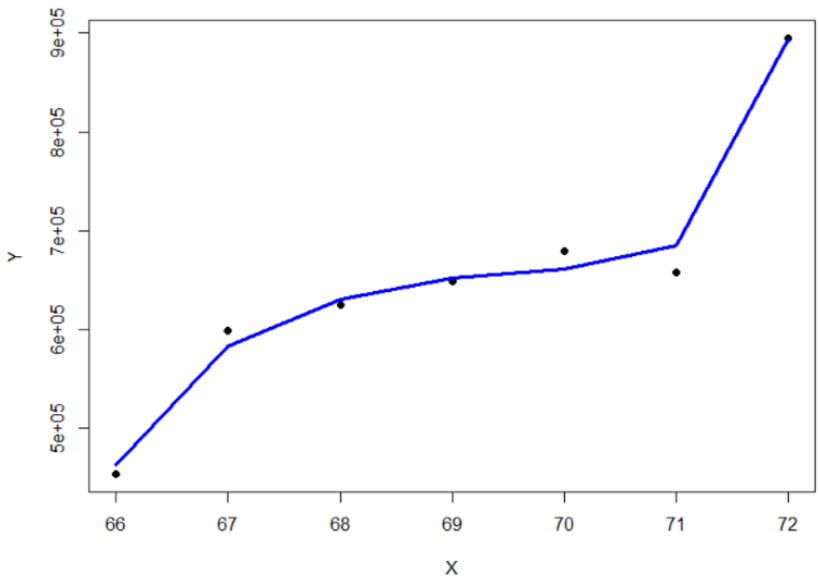

Figure 6. Plot of Out-Sample Data Estimation Results

After modeling tourist visit data on in-sample data, predictions will be made for the next 7 (months) of observations. By continuing to use the optimal bandwidth value of 0.899, the prediction results are obtained with a MAPE of 1.429%. Forecasting results using this model will produce very good forecasting (highly accurate) because the MAPE value of 1.429% is below 10%.

3.2.3 All-Sample Data Model Estimation and Interpretation Results

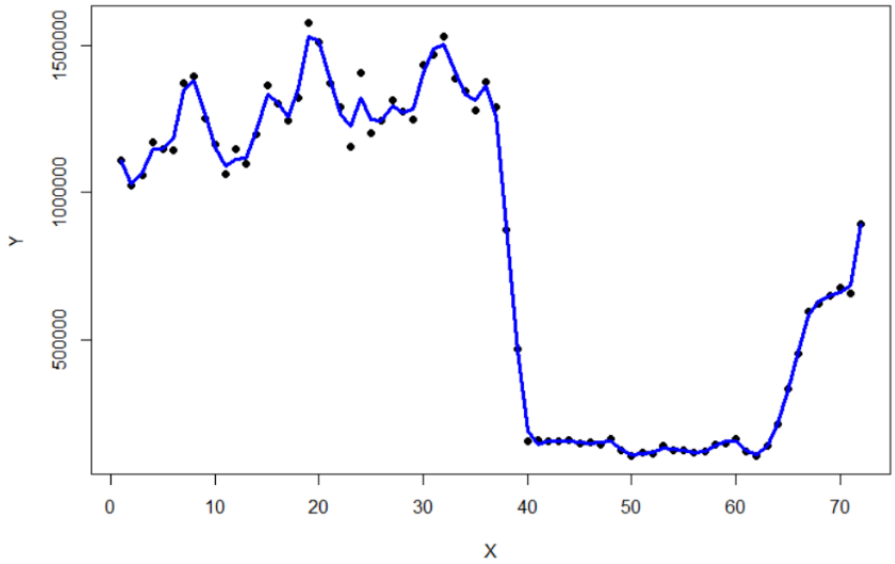

Figure 7. Plot of All-Sample Data Estimation Results

Based on the output of the R software, the results of estimating all-sample data with a second-order local polynomial estimator with an optimal bandwidth of 1.509, the MSE value of the all-sample data is 474610006, the R-Square value is 0.9938359 or 99.38%, and the MAPE value is 2.48%. Forecasting results using this model will produce highly accurate forecasts because the MAPE value is less than 10%.

3.3 Analysis of the Best ARIMA Model

In the ARIMA process, the first step is to check the stationarity of the data by looking at the time series graph or trend analysis of the data on the number of foreign tourist visits to Indonesia.

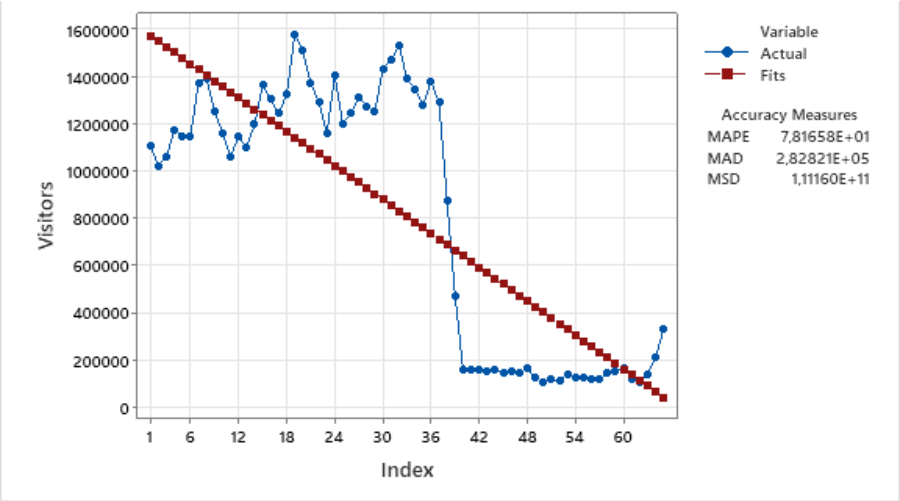

Figure 8. Plot of Trend Analysis Tourist Visit Data

It can be seen from the trend analysis plot in **Figure 8** that the data has an upward trend or is not constant, so it can be concluded that the data is not yet stationary. Furthermore, Box-Cox transformation is carried out so that the data is stationary in variance. The transformed data is then tested for stationarity in terms of the mean, and it is found that the data is not yet stationary, so one-time differencing is performed. Identification of the ARIMA model is done from the ACF and PACF plots of data that have been stationary can be seen in **Figure 9** and **Figure 10**.

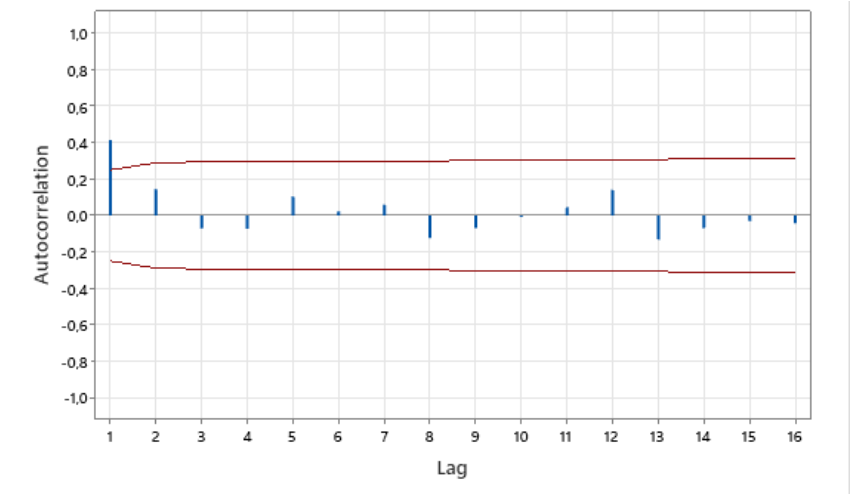

Figure 9. ACF Plot of Differencing 1 Result Data

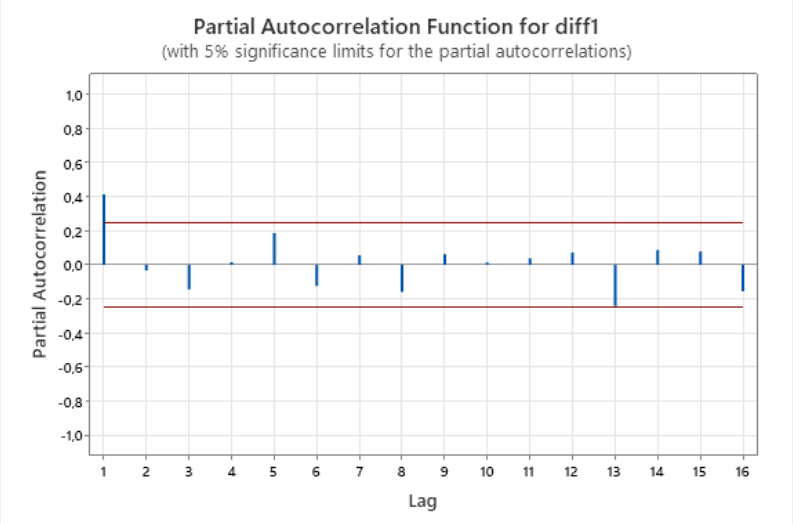

Figure 10. PACF Plot of Differencing 1 Result Data

Based on **Figure 9** and **Figure 10**, the possible time series analysis models are ARI(1,1), IMA(1,1), and ARIMA(1,1,1). The three ARIMA models are then tested to find a model that meets the requirements (the parameters are significant and characterized by p-value < α, has the smallest MSE (Mean Square Error), and has white noise residuals) **[19]**.

Table 3. ARIMA Parameter Estimation

ARIMA Model		Parameter Significance Value	MSE	White Noise (p-value Ljung-Box > 0,05)
ARI (1,1)	Deterministic	AR 1 = 0.000 Constant = 0.694	1.890	Lag 12 = 0.255 Lag 24 = 0.248 Lag 36 = 0.043 Lag 48 = 0.180 No White Noise
	Probabilistic	AR 1 = 0.000	1.865	Lag 12 = 0.323 Lag 24 = 0.288 Lag 36 = 0.052 Lag 48 = 0.204 White Noise
IMA (1,1)	Deterministic	MA 1 = 0.003 Constant = 0.556	1.961	Lag 12 = 0.242 Lag 24 = 0.328 Lag 36 = 0.050 Lag 48 = 0.205 No White Noise
	Probabilistic	MA 1 = 0.002	1.940	Lag 12 = 0.311 Lag 24 = 0.380 Lag 36 = 0.062 Lag 48 = 0.233 White Noise
ARIMA (1,1,1)	Deterministic	AR 1 = 0.138 MA 1 = 0.928 Constant = 0.689	1.921	Lag 12 = 0.183 Lag 24 = 0.199 Lag 36 = 0.032 Lag 48 = 0.151 No White Noise
	Probabilistic	AR 1 = 0.113 MA 1 = 0.949	1.895	Lag 12 = 0.245 Lag 24 = 0.236 Lag 36 = 0.040 Lag 48 = 0.174 No White Noise

Based on the results of ARIMA parameter estimation, a model has been obtained that meets the requirements (the parameters are significant and characterized by p-value < α, have the smallest MSE (Mean Square Error), and have white noise residuals, namely the probabilistic ARI (1,1) model. The model meets the assumption of residual normality but does not meet the assumption of homoscedasticity, so ARIMA modeling cannot be continued.

4. CONCLUSIONS

Based on the research results, using the ARIMA method to predict the number of foreign tourists arriving in Indonesia, the best model is ARIMA (1, 1, 0). However, the model does not meet the assumptions, namely the homoscedasticity test. So, testing using the ARIMA method was stopped. Meanwhile, in modeling using the local polynomial regression approach, the best order result is 2 with an optimal bandwidth of 0.899 and minimum CV 4556446632. The MAPE value of the local polynomial regression approach is 1.429%, which is categorized as a prediction with high accuracy because the value is less than 10%. Thus, the local polynomial regression approach is well used to predict the number of foreign tourist visits to Indonesia. Therefore, this research can be used as a reference for the government in an effort to develop Indonesia's tourism sector.

ACKNOWLEDGMENT

The authors would like to thank the Statistics Study Program, Universitas Airlangga. for providing opportunities and supporting students in carrying out research projects to implement learning materials during lectures on nonparametric regression analysis.

REFERENCES

[1] CNBC Indonesia, 2019. [Online]. Available: https://www.cnbcindonesia.com/market/20190319084140-17-61460/bi-pariwisata-bisa-jadi-penyumbang-devisa-terbesar-kedua. [Accessed 10 April 2023].

[2] C. C. Lab, W. Swesti, F. Alfian, A. N. Pajriyah, N. Bachtiar, N. Yatimah, S. Razak and J. Bramantio, Outlook Pariwisata & Ekonomi Kreatif Indonesia, Indonesia: Kemenparekraf, 2021.

[3] Satudata Kemenparekraf, 2022. [Online]. Available: https://satudata.kemenparekraf.go.id/. [Accessed 10 April 2023].

[4] S. Isnawati, Statistik Tenaga Kerja Pariwisata dan Ekonomi Kreatif 2018-2021, Jakarta: Kemenparekraf, 2022.

[5] B. G. Prianda and E. Widodo, "Perbandingan Metode Seasonal ARIMA dan Extreme Learning Machine pada Peramalan Jumlah Wisatawan Mancanegara ke Bali," *Barekeng Jurnal Ilmu Matematika dan Terapan* , vol. 15, no. 4, pp. 639-650, 2021.

[6] J. Hendrian, Suparti and A. Prahutama, "Pemodelan Harga Emas Dunia Menggunakan Metode Nonparametrik Polinomial Lokal Dilengkapi GUI R," *Jurnal Gaussian,* vol. 10, no. 4, pp. 604-616, 2021.

[7] A. Meimela, "Prediksi Jumlah Kunjungan Wisatawan Mancanegara ke Indonesia," *Media Wisata,* vol. 19, no. 1, 2021.

[8] V. Fibriyani and N. Chamidah, "Fibriyani, V., & Chamidah, N. (2021). Prediction of Inflation Rate in Indonesia Using Local Polynomial Estimator for Time Series Data," *Journal of Physis: Conference Series,* vol. 1776, no. 1, pp. 0-10, 2021.

[9] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif dan R&D, Alfabeta, 2016.

[10] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif dan R&D, Alfabeta CV, 2013.

[11] N. P. A. M. Mariati, I. N. Budiantara and V. Ratnasari, "Combination Estimation of Smoothing Spline and Fourier Series in Nonparametric Regression," *Journal of Mathematics,* vol. 2020, pp. 1-10, 2020.

[12] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice (3rd ed), Melbourne, Australia: OTexts, 2021.

[13] L. A. Yates, Z. Aandahl, S. A. Richards and B. W. Brook, "Cross Validation for Model Selection: A Review with Examples from Ecology," *Ecological Society of America,* vol. 93, no. 1, 2022.

[14] Suparti and A. Prahutama, "Pemodelan Regresi Nonparametrik Menggunakan Pendekatan Polinomial Lokal pada Beban Listrik di Kota Semarang," *Media Statistika,* vol. 9, no. 2, pp. 85-93, 2016.

[15] N. P. A. M. Mariati, I. N. Budiantara and V. Ratnasari, "Smoothing Spline Estimator in Nonparametric Regression (Application: Poverty in Papua Province)," in *7th International Conference on Research, Implementation, and Education of Mathematics and Sciences (ICRIEMS 2020)*, Atlantis Press, 2020.

[16] N. P. A. M. Mariati, I. N. Budiantara and V. Ratnasari, "The Application of Mixed Smoothing Spline and Fourier Series Model in Nonparametric Regression," *Symmetry,* vol. 13, no. 11, 2021.

[17] I. Nabillah and I. Ranggadara, "Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut," *Journal of Information System,* vol. 5, no. 2, pp. 250-255, 2020.

[18] Rory and R. Diana, "Pemodelan Data COVID-19 Menggunakan Regresi Polinomial Lokal," in *Seminar Nasional Official Statistics 2020*, 2020.

[19] Irwandi and D. P. Sari, "Analisis Metode Arima pada Peramalan Nilai Ekspor Sumatera Barat," *UNPjoMath,* vol. 6, no. 4, pp. 9-15, 2021.