talenta PUBLISHER

Journal of Research in Mathematics Trends and Technology

Journal homepage: https://talenta.usu.ac.id/jormtt

Adjusting Anomalies in International Tourist Arrivals to North Sumatra During the Peak COVID-19 Period (April 2020 to June 2022) to Enhance the Validity of Time Series Modeling

Thaswin Eddy [1], Open Darnius [*2]

[1] *Department of Mathematics, Universitas Sumatra Utara, Medan, 20155, Indonesia.*

[*]Corresponding Author: opendarnius@gmail.com

ARTICLE INFO

Article history:
Received: 03 July 2025
Revised: 04 August 2025
Accepted: 04 September 2025
Available online: 19 September 2025

E-ISSN: 2656-1514
P-ISSN:

How to cite:
Eddy, T, Darnius, O, "Adjusting Anomalies in International Tourist Arrivals to North Sumatra During the Peak COVID-19 Period (April 2020 to June 2022) to Enhance the Validity of Time Series Modeling", Journal of Research in Mathematics Trends and Technology, vol. V7 No. 2, March. 2025, doi: 10.32734/jormtt.v7i2.21718

ABSTRACT

The feasibility of time series modeling is significantly influenced by both the availability and the structural patterns of the data. Regular and continuous data collection over time is essential for constructing reliable time series models, particularly for forecasting purposes. Generally, a minimum of 50 time series data points is considered ideal to ensure the robustness and predictive power of such models. However, the presence of extreme fluctuations such as sharp spikes or drops can severely affect the integrity of the model. In the context of international tourist arrivals to North Sumatra during the peak period of the COVID-19 pandemic (April 2020 to June 2022), substantial data anomalies were observed. The results of modifying these anomalies indicate that increasing the number of adjusted data points during this period leads to a greater number of feasible time series models suitable for predictive analysis.

Keyword: modification,feasibility,time series,prediction

http://doi.org/10.32734/jormtt.v7i2.21718

1. Introduction

Time series modeling refers to a method of analysis that utilizes time-indexed data, where the periodic availability of data plays a critical role in developing a viable model. Ensuring model feasibility requires a minimum quantity of time series data[1,2,3]. Furthermore, careful attention must be given to the underlying patterns within the data. The presence of extreme upward or downward spikes in the time series can significantly influence the feasibility and reliability of the resulting model

Several commonly used time series models include the following:

- **Autoregressive (AR)**: model that uses past values of the same variable to predict future values.
- **Moving Average (MA)**: model that uses past forecast errors to predict future values.
- **Seasonal Autoregressive (SAR)**: model that utilizes past values of the same variable at the same seasonal period to predict future values.
- **Seasonal Moving Average (SMA)**: model that employs past forecast errors at the same seasonal period to predict future values.
- **Autoregressive Integrated Moving Average (ARIMA)**: model that combines AR and MA components with differencing to achieve stationarity in the data.
- **Seasonal ARIMA (SARIMA)**: An extension of the ARIMA model that incorporates seasonal patterns into the modeling process.

Time series modeling has standard operating procedures that must be fulfilled in order to produce an adequate model. In the initial stage, the conditions of stationarity in variance and then in mean must be met for the data that will be used in the modeling process. Stationarity in variance and mean plays an important role, considering that the accuracy of predictions is closely related to the patterns of time series data that exhibit both properties. Time series data that are not stationary in variance can be transformed to achieve variance stationarity, while data that are not stationary in mean can be differenced to achieve mean stationarity[5,6,7].

After the stationarity requirements are met, the next step is to obtain the best model by referring to the Akaike Information Criterion (AIC), where the best model is selected based on the lowest AIC value. The selected model is then subjected to a significance test for its parameters, and the model is considered significant if the p-values of the estimated parameters are below 0.05. Furthermore, a valid model must have diagnostic checking values greater than 0.05 at several selected lags, which are typically lag 12, lag 24, lag 36, and lag 48. If this condition is not met at those lags, the model is deemed unsuitable for forecasting because the resulting residuals are not independent[3,9].

In the final stage of model validation, a normality test is conducted on the residuals, where a model is considered to have normally distributed residuals if the p-value is greater than 0.05. Such a model can then be used to forecast future values. If multiple models meet the model adequacy requirements, the one with a lower Mean Squared Error (MSE) indicates a smaller prediction error. MSE can be used to compare the quality of different models. However, having more than one alternative time series model is expected to yield more accurate forecasts by considering the strengths of each model beyond just the MSE value.

2. Method

This study employs four time series models, namely ARIMA, SARIMA, SAR, and SMA, with data spike modifications applied to tourist arrival data to North Sumatra during the peak of the COVID-19 period, from April 2020 to June 2022, involving 27 data series that experienced spikes. The data containing spikes were categorized into five groups: 0% spike data, 60% spike data at the end, 60% spike data at the beginning, 30% spike data at the end, and 100% spike data.

The modeling equations for the four time series models are as follows:

- ✓ ARIMA modeling :
  $\phi(B)\Delta^{d}\bar{Z}_t = \theta(B)a_t$
- ✓ SARIMA modeling :
  $\Phi(B^s)\phi(B)\Delta^{D}\Delta^{d}\bar{Z}_t = \Theta(B^s)\theta(B)a_t$
- ✓ SAR modeling :
  $\Phi(B^s)\bar{Z}_t=a_t$
- ✓ SMA modeling
  $\bar{Z}_t = \Theta(B^s)a_t$

where,

$\Delta^{D}= (1 - B^s)^{D}$

$\Delta^{d}= (1 - B)^{d}$

$\Phi(B^s) = (1 - \Phi_s B^s - \Phi_{2s} B^{2s} - \ldots . - \Phi_{Ps} B^{Ps})$

$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots . - \phi_p B^p)$

$\Theta(B^s) = (1 - \Theta_s B^s - \Theta_{2s} B^{2s} - \ldots . - \Theta_{Qs} B^{Qs})$

$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots . - \theta_q B^q)$

The study was conducted on the foreign tourist arrival data (Wisman) with 60 data series, ranging from January 2019 to December 2023. Within this time frame, there were peak periods during the COVID-19 pandemic, which lasted for 27 months, from April 2020 to June 2022. During the peak period of the COVID-19 outbreak, the number of tourist arrivals experienced a significant decline, and in certain months, the number of arrivals could be considered virtually nonexistent. This is referred to as 'data spike,' where the data exhibits a sharp drop, resembling a deep sea trench. As a result, modifications were needed to adjust the 'data spike' to produce a more adequate time series pattern.

Ideally, the initial pattern of time series data should follow a normal distribution, as this constitutes a fundamental prerequisite for developing robust time series models capable of accurately forecasting future trends. In this study, a gradual modification was applied to data exhibiting sharp declines in order to evaluate the effects of reducing the presence of 'data spikes'. Specifically, the modification involved the use of the complete original dataset, which includes international tourist arrivals to North Sumatra from January 2019 to December 2023. A reduction of 30 percent of the spike data was then implemented, at the end of the spike

cluster, comprising 27 time series data points corresponding to the peak period of the COVID-19 pandemic, namely from April 2020 to June 2022.

Subsequently, 60 percent of the spike data were removed, specifically those occurring at the beginning or the end of the COVID-19 peak period. Finally, all spike data, or 100 percent, were excluded. These spike data were replaced with time series data from periods prior to January 2019. The number of time series entries replaced corresponded to the proportion of data removed: 30 percent (9 time series entries), 60 percent (18 entries), and 100 percent (27 entries), all sourced from data prior to January 2019. This procedure was undertaken to mitigate the impact of data anomalies observed during the peak period of the COVID-19 pandemic, spanning from April 2020 to June 2022.

A normality test was conducted on all five modified data conditions, namely:

- 0 percent spike data (removal of all 27 spike data points)
- 60 percent initial spike data (removal of 18 spike data points from the beginning of the spike period)
- 60 percent end spike data (removal of 18 spike data points from the end of the spike period)
- 30 percent end spike data (removal of 9 spike data points from the end of the spike period)
- 100 percent spike data retained (no removal of spike data)

Normality testing was conducted using two statistical methods: the Kolmogorov–Smirnov test and the Shapiro Wilk test. According to standard criteria, if the significance value exceeds 0.05, the data are considered to follow a normal distribution. This condition is essential for ensuring that the residuals resulting from time series modeling also follow a normal distribution. However, it is possible for the original dataset, prior to modeling, to not follow a normal distribution, while the residuals obtained after modeling do. This can occur because time series modeling involves ensuring the stationarity of the data both in terms of variance and mean so that, after modeling, the residuals exhibit a pattern consistent with a normal distribution.

The time series data required to replace the spike data consist of 27 data points prior to January 2019, specifically starting from October 2016; 18 data points starting from July 2017; and 9 data points starting from April 2018. The availability of these replacement data plays a critical role in enabling various forms of time series modeling, such as ARIMA, SARIMA, SAR, and SMA. Meanwhile, multiplicative modeling is considered a last resort alternative when none of the standard time series models are deemed suitable. The multiplicative model incorporates four data components simultaneously: trend, seasonality, cyclicality, and randomness.

3. Result and Discussion

Based on the modeling results obtained through the modification of spike data during the peak of the COVID-19 period from April 2020 to June 2022 several time series models were identified as suitable for forecasting the number of international tourist arrivals to North Sumatra Province in the months following January 2024[20]. Time series forecasting models are generally most accurate for short term periods, typically less than one year or 12 months. Several models deemed appropriate based on the various spike data modifications may serve as viable options for predicting future international tourist arrivals to North Sumatra.

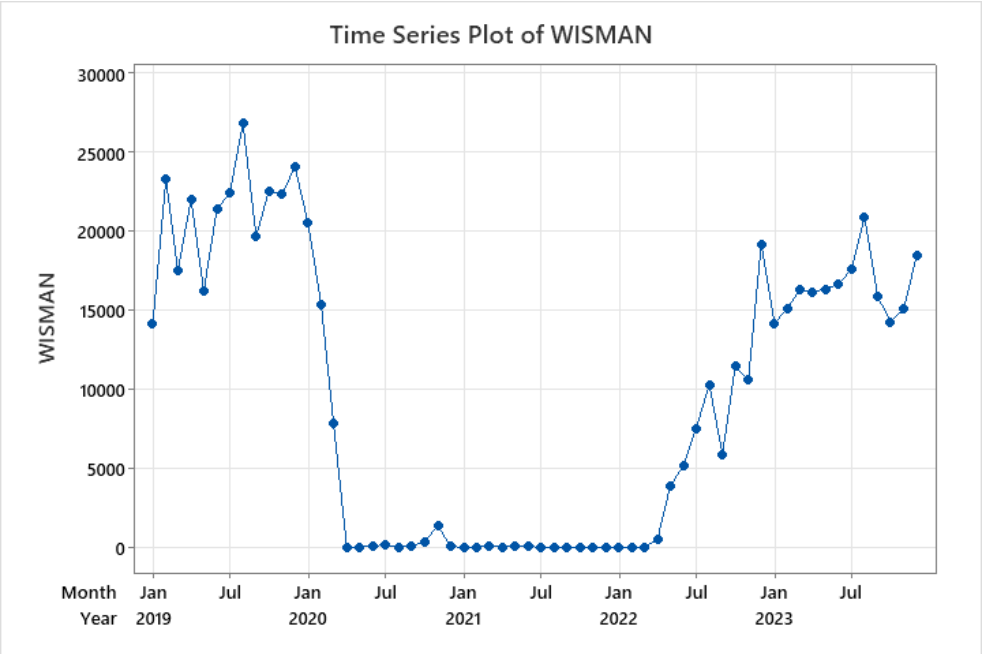

Figure 1 : Spike Data During Covid-19 Peak Period

Based on the spike data modification involving 0 percent spike data (removal of all 27 spike data points), three models were identified as suitable for forecasting: ARIMA(p=1, d=1, q=1), SAR(p=1, d=1, q=0, P=0, D=1, Q=0) with lag 4, and SMA(p=0, d=1, q=0, P=0, D=1, Q=1) also with lag 4. These models demonstrated significant parameter estimators, with p-values for each parameter estimator being less than 0.05 meeting the criterion for statistical significance. Diagnostic checking indicated that the residuals of these models were independent, as shown by p-values greater than 0.05. Furthermore, the residuals of these validated models also exhibited normality, with residual plot p-values exceeding 0.05, indicating that the residuals followed a normal distribution.

For the spike data modification involving 60 percent of spike data removed from the end (removal of 18 spike data points), two models were identified as suitable for forecasting: ARIMA(p=1, d=1, q=1) and SMA(p=0, d=1, q=0, P=0, D=1, Q=1) with lag 4. The same models were also found to be appropriate under the scenario where 60 percent of the spike data were removed from the beginning. In both cases, the models exhibited statistically significant parameter estimators, with p-values for each parameter being less than 0.05, fulfilling the requirement for model significance. Diagnostic checking further confirmed the independence of residuals, as indicated by p-values greater than 0.05. Additionally, the residuals of these models followed a normal distribution, with residual plot p-values exceeding 0.05, thus validating the assumption of normality in the residuals.

For the spike data modifications involving 30 percent of spike data removed from the end (removal of 9 data points) and 100 percent of spike data retained (no removal), no suitable models were found for forecasting purposes. The unsuitability of these models was attributed to several factors: non-significant parameter estimators, residuals that did not follow a normal distribution, and diagnostic checking results indicating p-values below 0.05, suggesting that the residuals were autocorrelated and not independent. Under such conditions, only the multiplicative model could be applied to these two types of spike data. The multiplicative model serves as a last resort alternative when standard time series models are deemed unsuitable. However, it is important to note that the residuals of the multiplicative model inherently do not follow a normal distribution, as the model integrates all four components of time series data trend, seasonality, cyclicality, and randomness.

Table 1. Suitable Model Resulting From The Modification of Spike Data on Foreign Tourist Visits to North Sumatra During The Peak Period of Covid-19 (April 2020 s/d June 2022)

Spike Data Category	Decent Model	Eligibility Category
0 Percent Data Spike	ARIMA(p=1,d=1,q=1)	Significance of Parameter Estimator, Residual Independence and Residual Normality
	SAR(p=1,d=1,q=0,P=0,D=1,Q=0) Lag 4	Significance of Parameter Estimator, Residual Independence and Residual Normality
	SMA(p=0,d=1,q=0,P=0,D=1,Q=1) Lag 4	Significance of Parameter Estimator, Residual Independence and Residual Normality
60 Percent End Data Spike	ARIMA(p=1,d=1,q=1)	Significance of Parameter Estimator, Residual Independence and Residual Normality
	SMA(p=0,d=1,q=0,P=0,D=1,Q=1) Lag 4	Significance of Parameter Estimator, Residual Independence and Residual Normality
60 Percent Initial Data Spike	ARIMA(p=1,d=1,q=1)	Significance of Parameter Estimator, Residual Independence and Residual Normality
	SMA(p=0,d=1,q=0,P=0,D=1,Q=1) Lag 4	Significance of Parameter Estimator, Residual Independence and Residual Normality

4. Conclusion

Time series modeling using various approaches such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Seasonal Autoregressive (SAR), and Seasonal Moving Average (SMA) that are deemed suitable for forecasting is closely related to the normality of the initial data prior to modeling. Data characterized by sharp fluctuations, commonly referred to as 'data spikes', can cause the data pattern to deviate from a normal distribution. The greater the reduction of spike data, the more likely it is to yield alternative time series models that are appropriate for forecasting, even if the original dataset does not conform to a normal distribution. Utilizing unmodified spike data as in the case of international tourist arrivals to North Sumatra during the peak of the COVID-19 pandemic will not result in time series models that are suitable for future forecasting.

References

[1] S. Aktivani, "Uji Stasioneritas Data Inflasi Kota Padang Periode 2014-2019," Statistika, vol. 20, no. 2, pp. 83–90, 2020.

[2] A. Pankratz, Forecasting With Univariate Box-Jenkins Model: Concepts and Cases. New York: John Wiley & Sons, 1983.

[3] S. Anwar, "Peramalan Suhu Udara Jangka Pendek di Kota Banda Aceh dengan Metode Autoregressive Integrated Moving Average (ARIMA)," Malikussaleh Journal of Mechanical Science and Technology, vol. 5, no. 1, pp. 6–12, 2017.

[4] V. P. Ariyanti and T. Yusnitasari, "Comparison of ARIMA and SARIMA for Forecasting Crude Oil Prices," JURNAL RESTI, vol. 7, no. 2, pp. 405–413, 2023.

[5] A. Ashril, S. Agrippina, and A. Yusuf, "Smoothing Data Ketinggian Air Di Saluran Irigasi Menggunakan Cubic Spline," JSN: Jurnal Sains Natural, vol. 3, no. 1, pp. 34–44, 2025.

[6] E. N. S. Dewi and A. A. Chamid, "Implementation of Single Moving Average Methods For Sales Forecasting of Bag in Convection Tas Loram Kulon," Jurnal Transformatika, vol. 16, no. 2, pp. 113–124, 2019.
[7] N. Fauziah, Y. I. Ningsih, and E. Setiarini, "Analisis Peramalan (Forecasting) Penjualan Jasa Pada Warnet Bulian City di Muara Bulian," Eksis: Jurnal Ilmiah Ekonomi Dan Bisnis, vol. 10, no. 1, p. 61, 2019.
[8] F. Husnita, S. Wahyuningsih, and D. A. Nohe, "Analisis Spektral Dan Model ARIMA Untuk Peramalan Jumlah Wisatawan Di Dunia Fantasi Taman Impian Jaya Ancol," Jurnal EKSPONENSIAL, vol. 6, no. 1, pp. 21–30, 2015.
[9] I. Soelaeman, Analisis Runtun Waktu. Jakarta: Universitas Terbuka, 2016.
[10] V. Jadhav, B. V. Chinnappa Reddy, and G. M. Gaddi, "Application of ARIMA model for forecasting agricultural prices," Journal of Agricultural Science and Technology, vol. 19, no. 5, pp. 981–992, 2017.
[11] L. N. Kasanah, "Aplikasi Autoregressive Integrated Moving Average (ARIMA) untuk Meramalkan Jumlah Demam Berdarah Dengue (DBD) di Puskesmas Mulyorejo," Jurnal Biometrika Dan Kependudukan, vol. 5, pp. 177–186, 2016.
[12] M. B. Pamungkas and A. Wibowo, "Aplikasi Metode Arima Box-," The Indonesian Journal of Public Health, vol. 13, pp. 181–194, 2018.
[13] F. Ramadhani, K. Sukiyono, and M. Suryanty, "Forecasting of Paddy Grain and Rice's Price: An ARIMA (Autoregressive Integrated Moving Average) Model Application," SOCA: Jurnal Sosial Ekonomi Pertanian, vol. 14, no. 2, p. 224, 2020.
[14] D. A. Rezaldi and Sugiman, "Peramalan Metode ARIMA Data Saham PT. Telekomunikasi Indonesia," Prisma, vol. 4, pp. 611–620, 2021.
[15] M. F. E. Saputra and M. Rizky, "Forecasting Number of Cases Acute Respiratory Infection (Ari) in 2019 Using Arima Method," Jurnal Biometrika Dan Kependudukan, vol. 8, no. 2, pp. 138–145, 2019.
[16] T. Septia and R. Wahyu, "Penggunaan Analisis Deret Berkala Dalam Meramalkan Kinerja Turbin," JURNAL TEKNOLOGI TERAPAN, vol. 2, no. 1, pp. 127–134, 2018.
[17] P. Somboonsak, "Forecasting Dengue Fever Epidemics using ARIMA Model," ACM International Conference Proceedings Series, pp. 144–150, 2019.
[18] E. Syafwan, M. Syafwan, and S. Tresnawati, "Pengembangan Metode Interpolasi Splin Kubik Terapit Dan Aplikasinya Pada Masalah Pelacakan Trajektori Objek," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 9, no. 5, pp. 943–950, 2022.
[19] S. Wardah and Iskandar, "Analisis Peramalan Penjualan Produk Keripik Pisang Kemasan Bungkus (Studi Kasus: Home Industry Arwana Food Tembilahan)," Jurnal Teknik Industri, vol. 9, no. 3, pp. 135–142, 2016.
[20] A. M. Windhy and A. S. Jamil, "Peramalan Harga Cabai Merah Indonesia: Pendekatan ARIMA," Jurnal Agriekstensia, vol. 20, no. 1, pp. 78–87, 2021.