Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 2, 2025
DOI: 10.30829/zero.v9i2.26381
Page: 534-545

Hybrid GSTAR-Machine Learning Model for Forecasting Tourists Numbers in Yogyakarta

[1] **Gama Putra Danu Sohibien**
Politeknik Statistika STIS, Jakarta, 13330, Indonesia

[2] **Annisa Nurul Azmi**
Politeknik Statistika STIS, Jakarta, 13330, Indonesia

[3] **Wahyuni Andriana Sofa**
Politeknik Statistika STIS, Jakarta, 13330, Indonesia

[4] **Cucu Sumarni**
Politeknik Statistika STIS, Jakarta, 13330, Indonesia

[5] **Rindang Bangun Prasetyo**
Politeknik Statistika STIS, Jakarta, 13330, Indonesia

[6] **Christiana Anggraeni Putri**
Politeknik Statistika STIS, Jakarta, 13330, Indonesia

Article Info

Article history:

Accepted, 30 October 2025

Keywords:

GSTAR;
Hybrid;
KNN;
Machine Leraning;
SVR;
XGBoost.

ABSTRACT

Tourism management in DI Yogyakarta is vital to ensure tourism benefits local communities. A key challenge lies in the uncertainty and spatial interdependence of tourist visits among neighboring regions. While the GSTAR model captures spatial relationships, its accuracy decreases with outliers, non-linearity, and assumption violations. To overcome these issues, this study integrates GSTAR with machine learning. Using 168 observations of tourist visits across DI Yogyakarta's regencies/cities (January 2010–December 2023), GSTAR-GLS-XGBoost model achieved 22–34% lower RMSE than other models. Tourist numbers fluctuate greatly, with peaks in May, June, July, and December. Practically, these findings can help local governments and stakeholders optimize resource allocation, plan promotions, and prepare facilities during peak seasons for sustainable tourism management in DI Yogyakarta.

Corresponding Author:

Gama Putra Danu Sohibien
Department of Statistics
Politeknik Statistika STIS
Email: gamaputra@stis.ac.id

1. INTRODUCTION

The tourism sector contributes to increasing foreign exchange earnings, creating jobs, and boosting regional economic growth [1]. The Special Region of Yogyakarta (DI Yogyakarta) is one of the main destinations for

domestic and foreign tourists [2] [3] [4]. One of the challenges in managing the tourism sector is the uncertainty of future tourist arrivals. Therefore, forecasting tourist arrivals is a crucial aspect of strategic planning and policy-making [5]. Accurate forecasts can help governments and tourism industry players anticipate spikes in tourist arrivals. A popular univariate time series model is the ARIMA model. Research [6] predicted the number of foreign patients coming to Turkey using the ARIMA approach. Research [7] predicted the number of foreign tourists coming to India using the ARIMA model. The ARIMA model was then developed into the Vector Autoregressive (VAR) model that is not only influenced by the value of the endogenous variable itself in the past but also by the past values of other endogenous variables [8]. In addition to past data patterns, forecasting tourist numbers are often influenced by interrelated regional factors (spatial relationships), such as accessibility and cross-regional promotion [9]. The ARIMA and VAR models cannot accommodate interregional linkages. In the GSTAR model, the forecasting of the value of an endogenous variable in a location is not only influenced by past data from that location but also by past data from other locations [10] [11].

The GSTAR model cannot accommodate outliers and non-linear relationships between variables that can cause a decrease in forecasting accuracy [12] [13]. One approach that can be used for forecasting data containing outliers and nonlinear relationships between variables are machine learning models, like Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). These models have high flexibility in handling various types and patterns of data [14] [15] [16] [17]. Several studies that use machine learning models to predict the number of tourists include [18], [19], and [20]. Study [18] found that the SVR model predicted domestic tourist visits to Bali accurately, with a MAPE of 9.19%. Study [19] showed that the most accurate forecasting of the number of visitors to the Dieng area was when using one nearest neighbor in the KNN model with smallest RMSE of 0 in testing data 36. Study [20] showed that tourism forecasting can be made accurately using the XGBoost model with a MAPE of 3.92 %.

Machine learning models do not automatically recognize patterns that frequently appear in time series data. In addition, machine learning models are prone to overfitting [21]. Furthermore, machine learning models are difficult to interpret. One possible solution is to use a hybrid approach. This approach combines conventional models, such as GSTAR and machine learning models. Combining these two approaches can complement each other's strengths and weaknesses [22] [23] [24]. Based on the search conducted, no research has been found that uses a GSTAR- machine learning (KNN, SVR, and XGBoost) for forecasting the number of tourists in Indonesia, especially in the Province of DI Yogyakarta. Research conducted by [25] and [26] found that spatial hybrid and machine learning models can provide better forecasting accuracy than conventional spatial models. This study aims to model the number of tourists visiting tourist attractions in cities/regencies of the Province of DI Yogyakarta using a hybrid GSTAR-machine learning model. The best model will then be used to forecast tourist numbers for several periods ahead.

2. RESEARCH METHOD

2.1 Data

The data in this study is the number of tourists visiting to tourist attractions in cities/regencies of the Province of DI Yogyakarta. The cities and regencies that are the focus of this study are Yogyakarta, Sleman, Bantul, Kulon Progo, and Gunung Kidul. There were 168 observations used with a data period from January 2010 to December 2023. The research data were sourced from tourism statistics published by the DI Yogyakarta Provincial Tourism Office. The data was divided into two parts, namely training data consisting of 151 observations and testing data consisting of 17 observations. The training data was used to build the model, while the testing data was used to evaluate the forecasting performance of the built model. The variables used were the number of tourists in Kulon Progo (Y_{1t}), Bantul (Y_{2t}), Gunung Kidul (Y_{3t}), Sleman (Y_{4t}), and Yogyakarta (Y_{5t}).

2.2 Generalized Space-Time Autoregressive (GSTAR)

The GSTAR model is a forecasting model that can accommodate the influence of endogenous variables in one location and other locations in previous periods [27] [28] [29]. Let

$$\boldsymbol{y_t} = [Y_{1t} \quad Y_{2t} \quad Y_{3t} \quad Y_{4t} \quad Y_{5t}]',$$

$$\boldsymbol{\phi_0^k} = \begin{bmatrix} \phi_{01}^k & 0 & 0 & 0 & 0 \\ 0 & \phi_{02}^k & 0 & 0 & 0 \\ 0 & 0 & \phi_{03}^k & 0 & 0 \\ 0 & 0 & 0 & \phi_{04}^k & 0 \\ 0 & 0 & 0 & 0 & \phi_{05}^k \end{bmatrix}, \boldsymbol{\phi_1^k} = \begin{bmatrix} \phi_{11}^k & 0 & 0 & 0 & 0 \\ 0 & \phi_{12}^k & 0 & 0 & 0 \\ 0 & 0 & \phi_{13}^k & 0 & 0 \\ 0 & 0 & 0 & \phi_{14}^k & 0 \\ 0 & 0 & 0 & 0 & \phi_{15}^k \end{bmatrix}$$

$$\boldsymbol{W^k} = \begin{bmatrix} 0 & w_{12} & w_{13} & w_{14} & w_{15} \\ w_{21} & 0 & w_{23} & w_{24} & w_{25} \\ w_{31} & w_{32} & 0 & w_{34} & w_{35} \\ w_{41} & w_{42} & w_{43} & 0 & w_{45} \\ w_{51} & w_{52} & w_{53} & w_{54} & 0 \end{bmatrix}, \boldsymbol{\varepsilon_t} = [\varepsilon_{1t} \quad \varepsilon_{2t} \quad \varepsilon_{3t} \quad \varepsilon_{4t} \quad \varepsilon_{5t}]'$$

ϕ_{0i}^k is the autoregressive coefficient for the model in city/regency i, where i=1, 2, 3, 4, 5,
ϕ_{1i}^k is the space-time coefficient for the model in city/regency i, where i=1, 2, 3, 4, 5,
ε_{it} is the residual model in period t in city/regency i, where i = 1, 2, 3, 4, 5,
$w_{it}$is the spatial weighting value between regency i and regency j, where i = 1, 2, 3, 4, 5 and j=1, 2, 3, 4, 5 and i≠j, we can write the general form of GSTAR with spatial order 1, autoregressive order p, and five locations is as follows:

$$\boldsymbol{y_t} = \sum_{k=1}^{p} \left(\boldsymbol{\phi_0^k} + \boldsymbol{\phi_1^k W^k}\right) \boldsymbol{y_{t-k}} + \boldsymbol{\varepsilon_t}\,, \tag{1}$$

with a total of T observations, equation (1) can be written as follows:

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}\,, \tag{2}$$

where:

$$\boldsymbol{y} = Vec(\boldsymbol{Y})\,, \boldsymbol{Y} = \begin{bmatrix} Y_{1,p+1} & Y_{2,p+1} & Y_{3,p+1} & Y_{4,p+1} & Y_{5,p+1} \\ Y_{1,p+2} & Y_{2,p+2} & Y_{3,p+2} & Y_{4,p+2} & Y_{5,p+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{1,T} & Y_{2,T} & Y_{3,T} & Y_{4,T} & Y_{5,T} \end{bmatrix}, \boldsymbol{\varepsilon} = Vec\,(\boldsymbol{E})$$

$$\boldsymbol{E} = \begin{bmatrix} \varepsilon_{1,p+1} & \varepsilon_{2,p+1} & \varepsilon_{3,p+1} & \varepsilon_{4,p+1} & \varepsilon_{5,p+1} \\ \varepsilon_{1,p+2} & \varepsilon_{2,p+2} & \varepsilon_{3,p+2} & \varepsilon_{4,p+2} & \varepsilon_{5,p+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{1,T} & \varepsilon_{2,T} & \varepsilon_{3,T} & \varepsilon_{4,T} & \varepsilon_{5,T} \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z_1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z_2} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z_3} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z_4} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z_5} \end{bmatrix}$$

$$\boldsymbol{Z_i} = \begin{bmatrix} Y_{i,p} & Y_{i,p}^* & \cdots & Y_{i,1} & Y_{i,1}^* \\ Y_{i,p+1} & Y_{i,p+1}^* & \cdots & Y_{i,2} & Y_{i,2}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Y_{i,T-1} & Y_{i,p+T-1}^* & \cdots & Y_{i,T-p} & Y_{i,T-p}^* \end{bmatrix}, Y_{1,t}^* = w_{i1}Y_{1,t-p} + w_{i2}Y_{2,t-p} + \cdots + w_{i5}Y_{5,t-p}\,,$$

$$\boldsymbol{\beta} = \left[\phi_{01}^1 \quad \phi_{11}^1 \quad \ldots \quad \phi_{01}^p \quad \phi_{11}^p \quad \ldots \quad \phi_{05}^1 \quad \phi_{15}^1 \quad \ldots \quad \phi_{01}^p \quad \phi_{15}^p\right]'$$

The existence of error correlation between equations can cause model coefficient estimators to become inefficient and reduce the accuracy of estimates. Therefore, the GSTAR model estimation method used is Generalized Least Squares (GLS) [27]. The formula used is as follows [29] [30]:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = \left(\boldsymbol{Z'}\widehat{\boldsymbol{\Omega}}^{\boldsymbol{-1}}\boldsymbol{X}\right)^{-1}\boldsymbol{Z'}\widehat{\boldsymbol{\Omega}}^{\boldsymbol{-1}}\boldsymbol{y}\,, \tag{3}$$

where:

$$\widehat{\boldsymbol{\Omega}} = \begin{bmatrix} \hat{\sigma}_1^{\,2} & \hat{\sigma}_{12} & \hat{\sigma}_{13} & \hat{\sigma}_{14} & \hat{\sigma}_{15} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^{\,2} & \hat{\sigma}_{23} & \hat{\sigma}_{24} & \hat{\sigma}_{25} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_3^{\,2} & \hat{\sigma}_{34} & \hat{\sigma}_{35} \\ \hat{\sigma}_{41} & \hat{\sigma}_{42} & \hat{\sigma}_{43} & \hat{\sigma}_4^{\,2} & \hat{\sigma}_{45} \\ \hat{\sigma}_{51} & \hat{\sigma}_{52} & \hat{\sigma}_{53} & \hat{\sigma}_{54} & \hat{\sigma}_5^{\,2} \end{bmatrix} \otimes \boldsymbol{I_{(T-p)\times(T-p)}}\,,$$

$\hat{\sigma}_i^{\,2}$ is the estimated variance error for the equation at location i, where i = 1, 2, 3, 4, 5,
$\hat{\sigma}_{ij}$ is is the covariance estimate between errors in equations i and j, where i = 1, 2, 3, 4, 5; j=1.

2.3 Support Vector Regression (SVR)

SVR works by finding hyperparameters that minimize the margin of error, thereby producing a model capable of capturing non-linear relationships between input and output variables using an approach known as a kernel function. The general function of SVR with the RBF kernel is as follows [16]:

$$s_i(x_t) = \sum_{i=1}^{m}(\alpha_i + \alpha_i^*)\exp(-\gamma\|x_i + x_t\|^2 + b), \quad (4)$$

where:
$0 < \alpha_i, \alpha_i^* \leq C$,
α_i and α_i^*: The dual Lagrange coefficient is obtained by maximizing the optimization function,
$\|x_i + x_t\|^2$: input distance and support vector,
γ: gamma, b: bias, C: cost.

The hyperparameters used are cost, gamma, and epsilon. These three hyperparameters are selected using the Optuna library in Python. The most commonly used kernel function is the Radial Basis Function (RBF) [31]. In this study, the RBF kernel will be used. The hyperparameter tuning procedure aims to obtain the optimal combination of parameters C, $\boldsymbol{\varepsilon}$, and γ that minimize forecasting error. Optuna automatically searches for the best parameter values within predefined ranges by iteratively evaluating model performance and minimizing the objective function.

2.4 K-Nearest Neighbors (KNN)

In forecasting, KNN calculates the average target value of the k nearest neighbors based on the distance in feature space. The formula for obtaining the target value forecasting is as follows:

$$\hat{y}_t = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i},$$

where:
$\hat{y}_t$: predicted value, m: number of features, y_i: target value from neighbor i, k: number of neighbors considered, and w_i:: weight of neighbor i based on its distance to point t.

The hyperparameters used to generate the best forecasting performance are k (the number of nearest neighbors used for forecasting), weights (the contribution weight of each neighbor), and metric (the distance measurement method, namely Euclidean, Manhattan, or Minkowski).

2.5 Extreme Gradient Boosting (XGBoost)

XGBoost builds multiple decision trees incrementally, where each new tree attempts to improve upon the errors of the previous tree. The parameters of XGBoost are determined to be optimal so that they minimize the value of the objective function, which is the sum of the loss and regularization. The loss value relates to how different the forecasting results are from the actual data. The regularization value relates to how complex the model used for forecasting is. The following is the formula for the XGBoost objective function:

$$L^{(t)} = \sum_{i=1}^{n}\left(y_i \hat{y}_i^{(t)}\right) + \sum_{t=1}^{T} W(f_t),$$

where:

$L^{(t)}$ is the objective function in the t-th iteration, $\sum_{i=1}^{n} 1\left(y_i, \hat{y}_i^{(t)}\right)$ is a *loss function*, $\sum_{t=1}^{T}\Omega(f_t)$ is a regulation, n is the number of observations used to build the model, and T is the number of leaves in the decision tree.

The Loss function in this study is $1\left(y_i, \hat{y}_i^{(t)}\right) = \frac{1}{2}1\left(y_i - \hat{y}_i^{(t)}\right)$. Meanwhile, the regulatory function in this study is $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j{}^2$,
where:
γ is the gamma parameter, which is a penalty for each leaf (encouraging simple trees),
λ is the lambda parameter, which is a penalty on the weight (preventing leaf values from becoming too large).
The final forecasting results with XGBoost can be obtained in the following manner:

$$\hat{y}_i = \hat{y}_i^{(0)} + \sum_{j=1}^{T}\eta w_j\,,$$

where:
$\hat{y}_i$: Final forecasting on the i-th observation, $\hat{y}_i^{(0)}$: Initial forecasting, w_j: predicted value on the j-th leaf, and η: Learning Rate that indicates the learning rate for reducing the contribution of each tree.

2.6 Hybrid Model of GSTAR and Machine Learning

In this study, data scaling or normalization was not applied because all variables used were measured on a comparable scale. Moreover, machine learning algorithms can internally handle variations in data magnitude. The development of the hybrid GSTAR and machine learning model begins with testing the stationarity of all variables using the Augmented Dickey-Fuller (ADF) test [12]. After confirming stationarity, spatial relationships among endogenous variables are examined through cross-correlation analysis. A weighting matrix is then

constructed using both uniform weighting and inverse distance weighting approaches to represent spatial influence structures [28] [29]. The optimal autoregressive order (p) is determined based on the Akaike Information Criterion (AIC), followed by the estimation of model coefficients using the GLS method with formula (3). Next, several GSTAR-GLS models are compared, and the one with the smallest Root Mean Square Error (RMSE) is selected as the best-performing model. The residuals from this optimal GSTAR-GLS model are then used to develop a residual forecasting model using a machine learning algorithm. Finally, the hybrid model is obtained by combining the predicted values from the GSTAR-GLS model with the residual forecasts produced by the machine learning model. The GSTAR and machine learning hybrid model can be written as follows:

$$\hat{\hat{Y}}_{t,GSTAR-GLS-SVR} = \hat{Y}_{t,GSTAR-GLS} + \hat{e}_{t,SVR} \quad (5)$$

$$\hat{\hat{Y}}_{t,GSTAR-GLS-KNN} = \hat{Y}_{t,GSTAR-GLS} + \hat{e}_{t,KNN} \quad (6)$$

$$\hat{\hat{Y}}_{t,GSTAR-GLS-XGBoost} = \hat{Y}_{t,GSTAR-GLS} + \hat{e}_{t,XGBoost} \quad (7)$$

2.7 Forecasting Performance Evaluation

The forecasting performance evaluation aims to assess the accuracy of the developed models. The training data are used to build the GSTAR-GLS model, from which the residuals are modeled using machine learning techniques. The best GSTAR-GLS model is then applied to forecast the dependent variable, while the residuals are predicted using SVR, KNN, and XGBoost during the testing period. The final forecasts are obtained by combining the GSTAR-GLS predictions with the corresponding residual forecasts as expressed in equations (5) to (7). The residuals of these forecasts are then calculated for the testing period, and the Root Mean Square Error (RMSE) is computed to evaluate performance. The model with the smallest RMSE value is considered to have the best forecasting accuracy.

3. RESULT AND ANALYSIS

3.1 Overview of the Number of Tourist Visits to Regencies/Cities in the Province of DI Yogyakarta

Based on Figure 1, it can be seen that during the period of 2010-2023, there was an increase in the number of tourists visiting tourist attractions in regencies/cities in the Province of DI Yogyakarta. Despite this upward trend, in 2020 and 2021, the number of tourists visiting tourist attractions in cities/regencies in the Province of DI Yogyakarta experienced a drastic decline. During the first wave of COVID-19 in March 2020, the Indonesian government imposed Large-Scale Social Restrictions (PSBB). Meanwhile, starting in January 2021, the government-imposed Community Activity Restrictions (PPKM) in Java and Bali. These two regulations had an impact on the decline in the number of tourists in 2020 and 2021.

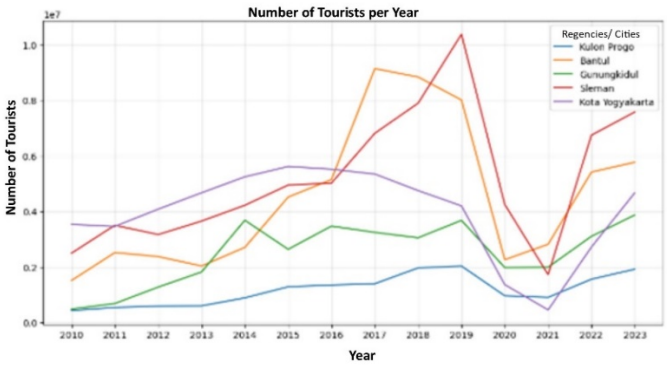
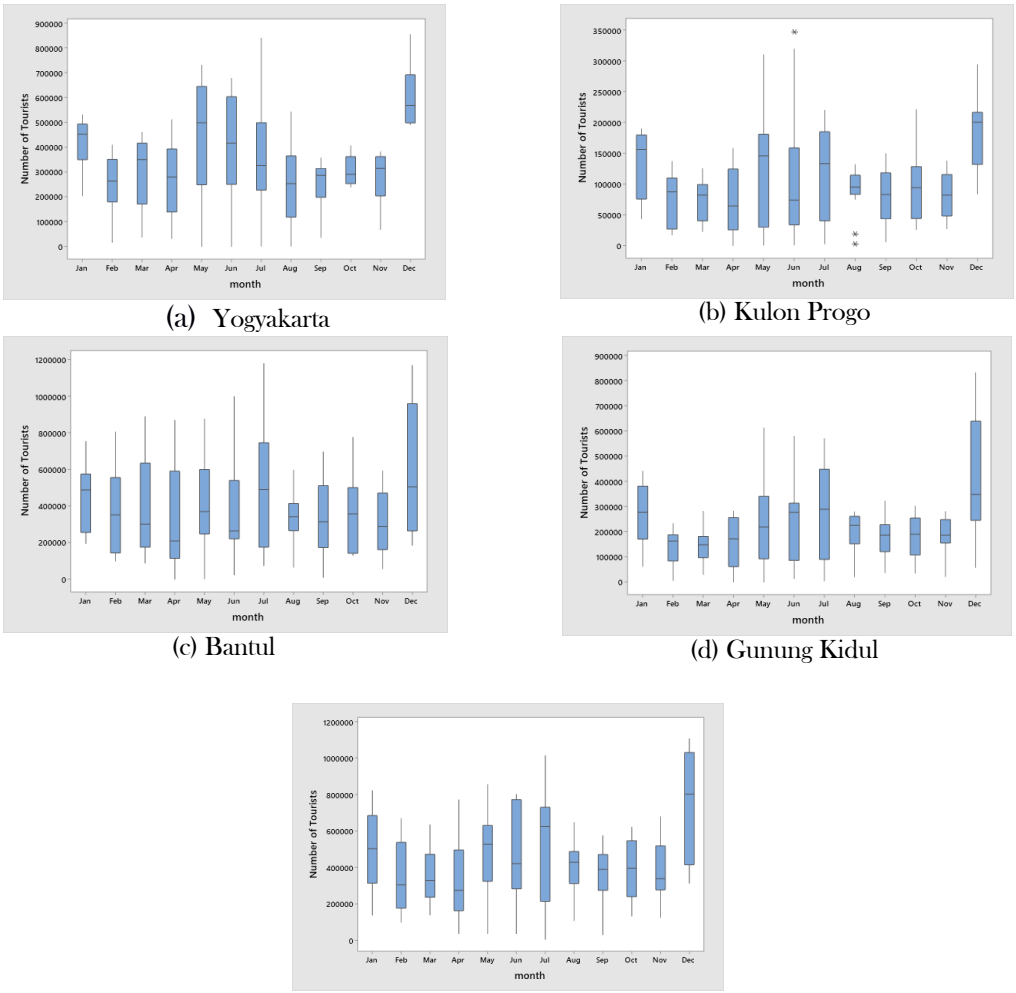


Figure 1. Number of tourist visits per year in each regency/city

There is also a seasonal pattern in the data on tourist visits to tourist attractions in cities/regencies of the Province of DI Yogyakarta. Based on Figure 2, there is a surge in tourist visits to tourist attractions in cities/regencies of the Province of DI Yogyakarta every December. The increase in tourist numbers in December coincides with the odd semester school holidays, Christmas holidays, and New Year holidays. This regularly drives an increase in tourist visits. In addition, an increase also occurs in July in Bantul, Gunung Kidul, and Sleman regencies. The increase in tourist numbers in July may be due to the even semester school holidays.



(a) Yogyakarta (b) Kulon Progo

(c) Bantul (d) Gunung Kidul

(e) Sleman

Figure 2. Boxplot of Tourist Visits by Month (a). Yogyakarta City, (b). Kulon Progo, (c). Bantul, (d) Gunung Kidul, (e). Sleman

Tourists' decisions to visit a destination are often influenced by the characteristics and popularity of the surrounding areas. Tourism promotion, infrastructure development, and perceptions of safety in one region can influence perceptions of surrounding regions. Figure 3 shows that the number of tourist visits between cities/regencies in the Province of DI Yogyakarta is directly proportional. It means that the higher the number of tourist visits in one regency/city, the higher the number of tourist visits in other regencies/cities. The highest correlation occurs between the number of tourist visits to Kulon Progo and Sleman, which is 0.794. Meanwhile, the lowest correlation occurs between the number of tourist visits to Kulon Progo and Yogyakarta.

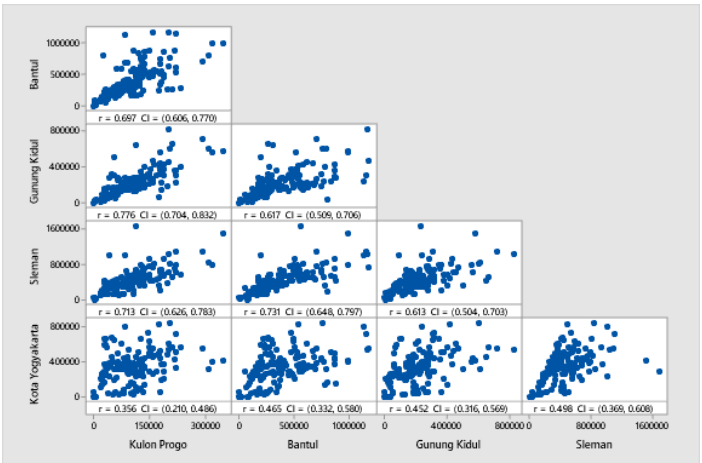


Figure 3. Scatter Plot of the Number of Tourist Visits between Regencies/Cities in the Province of DI Yogyakarta

3.2 Development of a Tourist Visit Forecasting Model

The first step in time series data modeling is to test the stationarity of the data. Table 1 shows the results of the stationarity test for all research variables using the ADF test. The results show that the tourist number data in all regencies/cities are not stationary at the level. Therefore, the research data needs to be differenced. The p-value of the stationarity test on the differenced data is less than 0.05. It indicates that the differenced data are stationary.

Table 1. Results of the Data Stationarity Test with ADF

City/ Regency	p-value before differencing	p-value after differencing
Kulon Progo	0.127	0.01
Bantul	0.657	0.01
Gunung Kidul	0.131	0.01
Sleman	0.342	0.01
Yogyakarta	0.478	0.01

The next step is to identify the relationship between the number of tourist visits to cities/regencies in the past and the number of tourist visits to cities/regencies in the present. The identification is done using cross-correlation at lag 1 to lag 12. Table 2 presents the significant lags of the dependent variable at a 5 percent significance level. Table 2 shows that tourist visits to cities/regencies of the Province of DI Yogyakarta are entirely influenced by tourist visits to other regencies/cities in previous periods. For instance, the number of tourist visits to tourist attractions in Bantul Regency in month t shows a significant correlation with the number of tourist visits to tourist attractions in several other regions in previous months. Specifically, it is correlated with tourist visits to Kulon Progo Regency in the previous 1, 4, 5, 11, and 12 months; to Bantul Regency itself in the previous 1, 4, 5, and 12 months; to Sleman Regency in the previous 1, 4, 7, 9, 11, and 12 months; and to Yogyakarta City in the previous 1, 7, 9, and 12 months.

Table 2. Lags of Dependent Variables that are Significantly Correlated with Dependent Variables in Period t by City/Regency

City/ Regency	Lag Kulon Progo	Lag Bantul	Lag Gunung Kidul	Lag Sleman	Lag Yogyakarta
Kulon Progo	-	1,12	1,12	1,12	1,4,7,8,9,12
Bantul	1,4,5,11,12	-	1,4,5,12	1,4,7,9,11,12	1,7,9,12
Gunung Kidul	1,11,12	1,7,9,12	-	1,7,9,11,12	7,9,12
Sleman	1,4,5,12	1,4,5,12	1,4,5,12	-	1,4,7,8,9,12
Yogyakarta	1,4,5,11,12	1,4,5,7,12	1,5,12	1,7,11,12	-

The determination of the autoregressive (AR) order in this study was carried out using the optimum order obtained from the Vector Autoregressive (VAR) model. Based on the calculation results, the smallest AIC value was obtained from VAR with an autoregressive order of 5. In addition, in order to accommodate seasonal effects in the model, we also used the 12th lag of the dependent variable. The spatial order used in this study is 1. Thus, the GSTAR-GLS model in this study is a model that accommodates the effects of lags 1, 2, 3, 4, 5, and 12 of the dependent variable using data differencing 1 (I(1)). The model can be written as GSTAR-GLS([1,2,3,4,5,12],1)-I(1). In this study, two weightings are used, namely uniform weighting and inverse distance weighting. The following are the uniform and inverse distance weighting matrices.

$$W_{uniform} = \begin{bmatrix} 0 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 & 0 \end{bmatrix} \qquad W_{\substack{invers \\ distance}} = \begin{bmatrix} 0 & 0.358 & 0.134 & 0.252 & 0.255 \\ 0.240 & 0 & 0.143 & 0.236 & 0.378 \\ 0.178 & 0.284 & 0 & 0.232 & 0.304 \\ 0.174 & 0.243 & 0.120 & 0 & 0.462 \\ 0.148 & 0.328 & 0.133 & 0.389 & 0 \end{bmatrix}$$

After determining the weighting matrix, the next step is to estimate the GSTAR model coefficients [1,2,3,4,5,12]1-I(1) using the GLS method on the training data using uniform and distance inverse weights. The model that produces the smallest RMSE value is selected as the best GSTAR-GLS model. Based on Table 3, the model with uniform weights is selected as the best model because it provided the best performance in three cities/regencies, while the other two cities/regencies show the best performance results with inverse distance weights. This difference shows that the effectiveness of weight types can vary between regions, but overall, uniform weights tend to provide the best performance on the analyzed data. The GSTAR-GLS model with uniform weights can be seen in equations (8) to (12).

The existence of spatial correlation can be proven by the significant space-time coefficient at a 5 percent significance level. For example, based on equation (8), the number of tourists visiting tourist attractions in Kulon Progo Regency in month t is significantly influenced by the number of tourists visiting tourist attractions in other cities/regencies in the previous 1, 4, and 5 months. Based on equation (9), the number of tourists visiting tourist attractions in Bantul Regency in month t is significantly influenced by the number of tourists visiting tourist attractions in other cities/regencies in the previous 12 months. It indicates that there is a potential for people who have visited tourist attractions in cities/regencies in DI Yogyakarta to visit other tourist attractions in cities/regencies in DI Yogyakarta again 12 months later.

Table 3. RMSE Values of the GSTAR-GLS Model according to Location Weight

City/ Regency	Uniform weight	Inverse distance weighting
Kulon Progo	49506.45	50601.56
Bantul	175511.14	184982.84
Gunung Kidul	140556.74	141290.79
Sleman	201453.86	195926.77
Yogyakarta	139095.63	135251.66

Kulon Progo

$$\begin{aligned}\Delta\hat{Y}_{1t} = {} & -0.578^{*}\Delta Y_{1,t-1} - 0.461^{*}\Delta Y_{1,t-2} - 0.354^{*}\Delta Y_{1,t-3} - 0.005\Delta Y_{1,t-4} + 0.155\Delta Y_{1,t-5} - \\ & 0.285^{*}\Delta Y_{1,t-12} + 0.00825^{*}\left(\Delta Y_{2,t-1} + \Delta Y_{3,t-1} + \Delta Y_{4,t-1} + \Delta Y_{5,t-1}\right) + 0.01875\left(\Delta Y_{2,t-2} + \Delta Y_{3,t-2} + \right. \\ & \left.\Delta Y_{4,t-2} + \Delta Y_{5,t-2}\right) + 0.001\left(\Delta Y_{2,t-3} + \Delta Y_{3,t-3} + \Delta Y_{4,t-3} + \Delta Y_{5,t-3}\right) - 0.02975^{*}\left(\Delta Y_{2,t-4} + \Delta Y_{3,t-4} + \right. \\ & \left.\Delta Y_{4,t-4} + \Delta Y_{5,t-4}\right) - 0.0365^{*}\left(\Delta Y_{2,t-5} + \Delta Y_{3,t-5} + \Delta Y_{4,t-5} + \Delta Y_{5,t-5}\right) - 0.007\left(\Delta Y_{2,t-12} + \Delta Y_{3,t-12} + \right. \\ & \left.\Delta Y_{4,t-12} + \Delta Y_{5,t-12}\right)\end{aligned} \quad (8)$$

Bantul

$$\begin{aligned}\Delta\hat{Y}_{2t} = {} & -0.735^{*}\Delta Y_{2,t-1} - 0.476^{*}\Delta Y_{2,t-2} - 0.235^{*}\Delta Y_{2,t-3} - 0.268\Delta Y_{2,t-4} - 0.099\Delta Y_{2,t-5} + \\ & 0.182^{*}\Delta Y_{2,t-12} + 0.06025^{*}\left(\Delta Y_{1,t-1} + \Delta Y_{3,t-1} + \Delta Y_{4,t-1} + \Delta Y_{5,t-1}\right) + 0.071\left(\Delta Y_{1,t-2} + \Delta Y_{3,t-2} + \right. \\ & \left.\Delta Y_{4,t-2} + \Delta Y_{5,t-2}\right) + 0.00175\left(\Delta Y_{1,t-3} + \Delta Y_{3,t-3} + \Delta Y_{4,t-3} + \Delta Y_{5,t-3}\right) - 0.01675\left(\Delta Y_{1,t-4} + \right. \\ & \left.\Delta Y_{3,t-4} + \Delta Y_{4,t-4} + \Delta Y_{5,t-4}\right) - 0.02675\left(\Delta Y_{1,t-5} + \Delta Y_{3,t-5} + \Delta Y_{4,t-5} + \Delta Y_{5,t-5}\right) + \\ & 0.09075^{*}\left(\Delta Y_{1,t-12} + \Delta Y_{3,t-12} + \Delta Y_{4,t-12} + \Delta Y_{5,t-12}\right)\end{aligned} \quad (9)$$

Gunung Kidul

$$\begin{aligned}\Delta\hat{Y}_{3t} = {} & -0.486^{*}\Delta Y_{3,t-1} - 0.416^{*}\Delta Y_{3,t-2} - 0.368^{*}\Delta Y_{3,t-3} - 0.201\Delta Y_{3,t-4} - 0.002\Delta Y_{3,t-5} + \\ & 0.158^{*}\Delta Y_{3,t-12} + 0.01625\left(\Delta Y_{1,t-1} + \Delta Y_{2,t-1} + \Delta Y_{4,t-1} + \Delta Y_{5,t-1}\right) + 0.022\left(\Delta Y_{1,t-2} + \Delta Y_{2,t-2} + \right. \\ & \left.\Delta Y_{4,t-2} + \Delta Y_{5,t-2}\right) + 0.00775\left(\Delta Y_{1,t-3} + \Delta Y_{2,t-3} + \Delta Y_{4,t-3} + \Delta Y_{5,t-3}\right) - 0.0405\left(\Delta Y_{1,t-4} + \right. \\ & \left.\Delta Y_{2,t-4} + \Delta Y_{4,t-4} + \Delta Y_{5,t-4}\right) - 0.08075^{*}\left(\Delta Y_{1,t-5} + \Delta Y_{2,t-5} + \Delta Y_{4,t-5} + \Delta Y_{5,t-5}\right) + \\ & 0.0515^{*}\left(\Delta Y_{1,t-12} + \Delta Y_{2,t-12} + \Delta Y_{4,t-12} + \Delta Y_{5,t-12}\right)\end{aligned} \quad (10)$$

Sleman

$$\begin{aligned}\Delta\hat{Y}_{4t} = {} & -0.627^{*}\Delta Y_{4,t-1} - 0.357^{*}\Delta Y_{4,t-2} - 0.137^{*}\Delta Y_{4,t-3} - 0.048\Delta Y_{4,t-4} - 0.111\Delta Y_{4,t-5} + \\ & 0.007\Delta Y_{4,t-12} + 0.0505\left(\Delta Y_{1,t-1} + \Delta Y_{2,t-1} + \Delta Y_{3,t-1} + \Delta Y_{5,t-1}\right) + 0.066\left(\Delta Y_{1,t-2} + \Delta Y_{2,t-2} + \right. \\ & \left.\Delta Y_{3,t-2} + \Delta Y_{5,t-2}\right) - 0.004\left(\Delta Y_{1,t-3} + \Delta Y_{2,t-3} + \Delta Y_{3,t-3} + \Delta Y_{5,t-3}\right) - 0.131^{\&}\left(\Delta Y_{1,t-4} + \Delta Y_{2,t-4} + \right. \\ & \left.\Delta Y_{3,t-4} + \Delta Y_{5,t-4}\right) - 0.015\left(\Delta Y_{1,t-5} + \Delta Y_{2,t-5} + \Delta Y_{3,t-5} + \Delta Y_{5,t-5}\right) + 0.1855^{*}\left(\Delta Y_{1,t-12} + \Delta Y_{2,t-12} + \right. \\ & \left.\Delta Y_{3,t-12} + \Delta Y_{5,t-12}\right)\end{aligned} \quad (11)$$

Yogyakarta

$$\begin{aligned}\Delta\hat{Y}_{5t} = {} & -0.365^{*}\Delta Y_{5,t-1} - 0.193^{*}\Delta Y_{5,t-2} - 0.207^{*}\Delta Y_{5,t-3} - 0.232^{*}\Delta Y_{5,t-4} - 0.104\Delta Y_{5,t-5} - \\ & 0.564^{*}\Delta Y_{5,t-12} - 0.000825\left(\Delta Y_{1,t-1} + \Delta Y_{2,t-1} + \Delta Y_{3,t-1} + \Delta Y_{4,t-1}\right) - 0.0065\left(\Delta Y_{1,t-2} + \Delta Y_{2,t-2} + \right. \\ & \left.\Delta Y_{3,t-2} + \Delta Y_{4,t-2}\right) - 0.011\left(\Delta Y_{1,t-3} + \Delta Y_{2,t-3} + \Delta Y_{3,t-3} + \Delta Y_{4,t-3}\right) + 0.0025\left(\Delta Y_{1,t-4} + \Delta Y_{2,t-4} + \right. \\ & \left.\Delta Y_{3,t-4} + \Delta Y_{4,t-4}\right) + 0.008\left(\Delta Y_{1,t-5} + \Delta Y_{2,t-5} + \Delta Y_{3,t-5} + \Delta Y_{4,t-5}\right) + 0.0005^{*}\left(\Delta Y_{1,t-12} + \Delta Y_{2,t-12} + \right. \\ & \left.\Delta Y_{3,t-12} + \Delta Y_{4,t-12}\right)\end{aligned} \quad (12)$$

Note: (*) significant at the 5% significance level

The residual data in the training data period is then used to form machine learning models using SVR, KNN, and XGBoost. The best SVR, KNN, and XGBoost models for each city/regency are then used to predict residuals in the testing data period. The results of tourist number forecasting from the GSTAR-GLS model are then added to the residual forecasting obtained from machine learning (SVR, KNN, and XGBoost) to produce hybrid tourist number forecasting. Based on Table 4, the smallest RMSE value for forecasting the number of

tourists to five city/regency tourist attractions was obtained from the GSTAR-GLS-XGBoost model. It indicates that the GSTAR-GLS-XGBoost model produces the best forecasting compared to the other three models. A visual comparison of the forecasting results between GSTAR-GLS-XGBoost and the actual data in the training and testing data periods can be seen in Figure 4.

Table 4. RMSE of GSTAR-GLS, GSTAR-GLS-SVR, GSTAR-GLS-KNN, and GSTAR-GLS-XGBoost models by city/regency in DI Yogyakarta Province

Location	GSTAR-GLS	GSTAR-GLS-SVR	GSTAR-GLS-KNN	GSTAR-GLS-XGBoost
Kulon Progo	49506.450	47688.931	149158.189	41114.391
Bantul	175511.140	167138.120	440573.776	153264.045
Gunung Kidul	140556.740	145132.667	339060.334	137047.849
Sleman	201453.860	178445.684	628865.223	165778.526
Kota Yogyakarta	139095.630	149360.313	379667.411	116793.225
Rata-rata	141224.764	137553.143	387464.986	122799.607

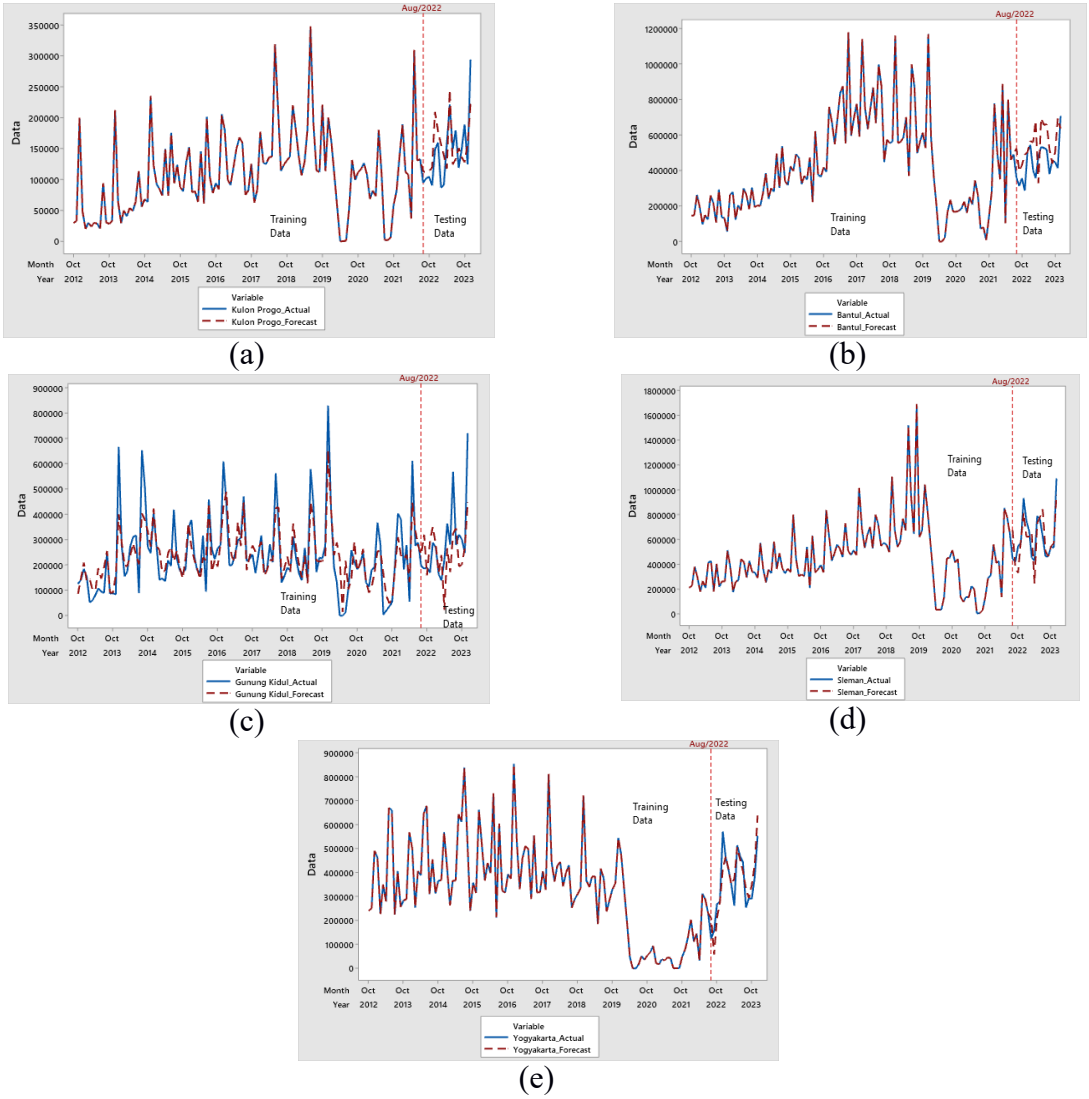


Figure 4. Time Series Plot of Forecast Results and Actual Data on the Number of Tourists Visiting Tourist Attractions in (a) Kulon Progo, (b) Bantul, (c) Guung Kidul, (d) Sleman, and (e) Yogyakarta

The next step is to forecast the number of tourists visiting tourist attractions in cities/regencies of the Province of DI Yogyakarta for the next few months using the best model, namely GSTAR-GLS-XGBoost. Forecasting is carried out from January 2025 to December 2025. The results of the tourists forecasting number visiting tourist attractions in five cities/regencies of the Province of DI Yogyakarta can be seen in Table 5.

Table 5. The Forecasting Number of Tourist Visits to Tourist Attractions by City/Regency in DI Yogyakarta Province from January 2025 to December 2025

Year	Month	Kulon Progo	Bantul	Gunung Kidul	Sleman	Yogyakarta
2025	January	189485	422638	444258	835894	274456
2025	February	216370	727654	467954	903991	314656
2025	March	191386	368731	446764	592216	211439
2025	April	176947	486443	470488	809928	341972
2025	May	207186	472695	471849	712083	474193
2025	June	194066	488388	514428	954268	375180
2025	July	213020	429318	508943	1097817	341384
2025	August	181745	532298	453350	616347	249589
2025	September	199791	501864	454011	779903	292809
2025	October	152774	410228	453054	567343	260633
2025	November	184229	444852	459247	1013183	453192
2025	December	210470	465506	421717	1076836	420823

Tourist numbers in the province of DI Yogyakarta are forecasted to be high in May 2025 due to holidays such as Labor Day, Vesak Day, and Ascension Day. Surges are also expected in June and July, coinciding with school holidays—particularly in Gunung Kidul and Sleman. December is another peak period due to Christmas, New Year's Eve, and school breaks.

4. CONCLUSION

The number of tourist visits to attractions in DI Yogyakarta has generally increased, although a sharp decline occurred in 2020–2021 due to the COVID-19 pandemic. The data reveal clear seasonal patterns, with consistent peaks in May, June, July, and December. Tourist activity in each regency is interrelated, as visits are strongly influenced by those in neighboring regions, particularly with a 12-month lag indicating annual recurrence. Mathematically, the GSTAR-GLS ([1,2,3,4,5,12],1)-XGBoost model provides the best forecasting performance, capturing both short-term (lags 1–5) and yearly (lag 12) effects while accounting for spatial interactions among directly adjacent regions. The GLS estimation improves efficiency by correcting cross-regional error correlations, yielding interpretable parameters that quantify how temporal persistence and spatial influence jointly shape tourism flows. Quantitatively, this hybrid model achieves about 22–34% lower RMSE than other models, confirming significant gains in predictive accuracy. From a policy perspective, these forecasts enable local governments to anticipate tourist surges and manage resources more effectively. Peak periods—particularly May–July and December—require enhanced crowd management, infrastructure readiness, and coordinated promotion across regencies. Although the model assumes stable spatial relationships and relies on historical data, future research could integrate additional explanatory variables, dynamic spatial weighting, or deep learning methods to further improve forecasting precision.

5. REFERENCES

[1] Kemenparekraf, "Siaran Pers: Menpar Optimistis Capaian Kinerja Pariwisata 2024 Lampaui Realisasi Tahun Sebelumnya," 2024.

[2] W. Hadi, "Menggali Potensi Kampung Wisata Di Kota Yogyakarta Sebagai Daya Tarik Wisatawan," *Journal of Tourism Economics*, vol. 2, pp. 129–139, 2024, doi: 10.36594/jtec/08yq9670.

[3] BPS, "Statistik menurut Subjek," 2024.

[4] A. Z. Yonatan, "Simak Kota Pilihan Orang Indonesia untuk Wisata," 2024.

[5] M. Hu, H. Li, H. Song, X. Li, and R. Law, "Tourism demand forecasting using tourist-generated online review data," *Tour Manag*, vol. 90, p. 104490, 2022, doi: 10.1016/j.tourman.2022.104490.

[6] N. Yilmaz, "Turkey's Health Tourism Demand Forecast: the ARIMA Model Approach," *International Journal of Health Management and Tourism*, vol. 7, no. 1, pp. 47–63, 2022, doi: 10.31201/ijhmt.1065460.

[7] M. Arora, "Forecasting Number of Inbound Tourists in India Adopting ARIMA Model," *MATRIX Academic International Online Journal of Engineering and Technology*, vol. 6, no. 1, pp. 9–17, 2023, doi: 10.21276/matrix.2023.6.1.2.

[8] Tsay, *Multivariate Time Series Analysis*. Chicago: Wiley, 2014.

[9] B. Z. et al., "A graph-attention based spatial-temporal learning framework for tourism demand forecasting," *Knowl Based Syst*, vol. 263, p. 110275, 2023, doi: 10.1016/j.knosys.2023.110275.

[10] M. Prastuti, L. Aridinanti, and W. P. Dwiningtyas, "Spatio-Temporal models with intervention effect for modelling the impact of Covid-19 on the tourism sector in Indonesia," in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/1821/1/012044.

[11] I. Adella, D. Ispriyanti, and H. Yasin, "Pemodelan Jumlah Wisatawan Di Jawa Tengah Menggunakan Metode Generalized Space Time Autoregressive - Seemingly Unrelated Regression (GSTAR-SUR)," *Jurnal Gaussian*, vol. 11, no. 2, pp. 258–265, 2022, doi: 10.14710/j.gauss.v11i2.35473.

[12] W. W. S. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed. New York: Pearson Addison Wesley, 2006.

[13] A. B. Kock and T. Teräsvirta, "Forecasting With Nonlinear Time Series Models," in *The Oxford Handbook of Economic Forecasting*, Oxford University Press, 2012, pp. 61–88.

[14] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLoS One*, vol. 13, no. 3, pp. 1–26, 2018, doi: 10.1371/journal.pone.0194889.

[15] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[16] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat Comput*, vol. 14, no. 3, pp. 199–222, 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[17] Y. Nader, L. Sixt, and T. Landgraf, "DNNR: Differential Nearest Neighbors Regression," in *Proceedings of Machine Learning Research*, 2022, pp. 16296–16317. [Online]. Available: http://arxiv.org/abs/2205.08434

[18] F. Mardliyah, E. Zukhronah, and Y. Susanti, "Forecasting the number of domestic tourist visits to Bali using support vector regression," in *AIP Conference Proceedings*, 2025, p. 20031. doi: 10.1063/5.0262760.

[19] E. H. Rachmawanto and C. A. Sari, "Visitor Forecasting Decision Support System at Dieng Tourism Objects Using the K-Nearest Neighbor Method," *Journal of Applied Intelligent Systems*, vol. 7, no. 2, pp. 183–192, 2022.

[20] X. Liu, Y. Chen, Z. Qiu, and M. Chen, "Forecast of the Tourist Volume of Sanya City by XGBoost Model and GM Model," in *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2019, pp. 166–173. doi: 10.1109/CyberC.2019.00038.

[21] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 2nd ed. Melbourne: OTexts, 2018.

[22] M. E. Nor, A. I. M. Nurul, and M. S. Rusiman, "A Hybrid Approach on Tourism Demand Forecasting," in *Journal of Physics: Conference Series*, 2018. doi: 10.1088/1742-6596/995/1/012034.

[23] D. P. M. Abellana, D. M. C. Rivero, M. E. Aparente, and A. Rivero, "Hybrid SVR-SARIMA model for tourism forecasting using PROMETHEE II as a selection methodology: a Philippine scenario," *Journal of Tourism Futures*, vol. 7, no. 1, pp. 78–97, 2020, doi: 10.1108/JTF-07-2019-0070.

[24] M. He and X. Qian, "Forecasting tourist arrivals using STL-XGBoost method," *Tourism Economics*, 2025, doi: 10.1177/13548166241313411.

[25] Suhartono, N Nahdliyah, MS Akbar, NA Salehah, and A Choiruddin, "A MGSTAR: An Extension of The Generalized Space-Time Autoregressive Model," *Journal of Physics: Conference Series. Vol. 1752*, No. 1, 2021. doi: 10.1088/1742-6596/1752/1/012015.

[26] Devi Munandar, budi Nurani Ruchjana, Atje Setiawan Abdullah, and Hilman Ferdinandus Pardede." Literature Review on Integrating Generalized Space-Time Autoregressive Integrated Moving Average

(GSTARIMA) and Deep Neural Networks in Machine Learning for Climate Forecasting Devi," *Mathematics*, 11, 2975, 2023, doi: 10.3390/math11132975

[27] Setiawan, Suhartono, and M. Prastuti, "S-GSTAR-SUR model for seasonal spatio temporal data forecasting," *Malaysian Journal of Mathematical Sciences*, vol. 10, pp. 53–65, 2016.

[28] G. Sohibien, "Perbandingan Model STAR dan GSTAR untuk Peramalan Inflasi Dumai, Pekanbaru, dan Batam," *Statistika*, vol. 5, no. 1, 2017.

[29] Suhartono, S. R. Wahyuningrum, Setiawan, and M. S. Akbar, "GSTARX-GLS Model for Spatio-Temporal Data Forecasting," *Malaysian Journal of Mathematical Sciences*, vol. 10, pp. 91–103, 2016.

[30] M. S. Akbar, Setiawan, Suhartono, B. N. Ruchjana, and M. A. A. Riyadi, "GSTAR-SUR Modeling with Calendar Variations and Intervention to Forecast Outflow of Currencies in Java Indonesia," in *Journal of Physics: Conference Series*, 2018. doi: 10.1088/1742-6596/974/1/012060.

[31] R. Rodríguez-Pérez and J. Bajorath, "Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery," *J Comput Aided Mol Des*, vol. 36, no. 5, pp. 355–362, 2022, doi: 10.1007/s10822-022-00442-9.