MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Abd Rahman *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 20 (2024) 972-984

RESEARCH ARTICLE

Spatio-Temporal Model to Forecast COVID-19 Confirmed Cases in High-Density Areas of Malaysia

Nur Haizum Abd Rahman[a]*, Saidatul Nurfarahin Muhammad Yusof[b], Iszuanie Syafidza Che Ilias[c], Kathiresan Gopal[c], Hannuun Yaacob[d], Noraishah Mohammad Sham[e]

[a]Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, Malaysia; [b]Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia; [c]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia; [d]Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia; [e]Environmental Health Research Centre, Institute for Medical Research, Shah Alam, 40170, Selangor, Malaysia

Abstract The coronavirus 2019 disease has spread across the world. The number of coronaviruses 2019 (COVID-19) cases throughout Malaysia is high in the densely populated state of Selangor. In assisting the early preventive measures, this study utilises time series methods to model and forecast the number of daily positive cases in three Selangor districts: Petaling, Hulu Langat, and Klang. Specifically, the study compares the effectiveness of the Autoregressive Integrated Moving Average (ARIMA), a univariate model and the Generalized Space-Time autoregressive integrated (GSTARI), a multivariate model. For the GSTARI model, uniform and inverse distance weights represent the relationship between locations. The analysed data are from January to August 2021, and the lowest root mean square error (RMSE) is chosen as the best model. The results show GSTARI (1,1) with both spatial weights outperformed ARIMA (0,1,1) in Petaling and Klang but not in Hulu Langat. However, the average RMSE values show that the most accurate and effective for forecasting the number of daily confirmed positive cases in Selangor is using GSTARI. In conclusion, by utilising advanced time series methods such as spatial analysis, this study provides important insights into forecasting trends of infectious diseases like COVID-19 and can help in early preventive measures.

Keywords: Spatio-temporal model, forecasting, Generalized STAR, COVID-19.

***For correspondence:**
haizum@umpsa.edu.my

Received: 13 Jan. 2024
Accepted: 21 Aug. 2024

Introduction

The discovery of the coronavirus disease, or COVID-19, stunned the globe at the end of 2019, which was in December. This ongoing COVID-19 pandemic originated from the Hunan seafood animal market in Wuhan, China [1]. Since then, COVID-19 has spread its arms and affected nearly all nations worldwide. On March 2020, on the 11th, the World Health Organization (WHO) declared COVID-19 a global pandemic. Two hundred three (203) countries had been hit by the COVID-19 pandemic by 9 April 2020, which affected 1,476,819 people and resulted in 87,816 deaths [2]. COVID-19 is a zoonotic illness where the animal virus leads to mutation and reproduction within the human body, leading to over 4.5 million fatalities between January 2020 and September 2021. Up until today, COVID-19 variants have been circulated around the world.

The ongoing pandemic has hit Malaysia significantly, causing thousands of people to be exposed to the virus and eventually be affected. Due to the dramatic increase in positive COVID-19 cases, the disease

has become a primary worldwide public health concern and got priority attention from the Malaysian government. As time progressed, it has been noticeable that COVID-19 became a global pandemic due to the pattern of how the virus spread [3]. There was no cure for COVID-19, and it was hard to kill a novel virus. Hence, the WHO developed a guideline to prevent COVID-19 from being contagious. However, the daily case count remained a cause for concern. Forecasting the number of daily positive cases with the employment of quantitative approaches is the key to improving the guidelines and preparing action plans. The results could provide important insights into the pattern of COVID-19 spread. In addition, the results enable more information for further actions.

COVID-19 has caught the eyes of many researchers. Diverse techniques, such as epidemic modelling time series and distribution modelling, were used to analyse COVID-19. Until today, these analyses are still being carried out. Besides understanding the dynamics of COVID-19, mathematical and statistical modelling may also be employed to estimate future values [4]. Infectious disease mathematical modelling can aid in overcoming delays and uncertainty [5]. The model of Susceptible-Infectious-Recovered (SIR) has been used most frequently to simulate a contagious disease epidemic's outbreak trajectory and intensity [6]. As expected, the SIR model has been widely used in several studies [7] and its further extensions, such as the SEIR known as the Susceptible–Exposed–Infectious–Recovered mode [8], [9]. Sun and Weng [10] developed a modified model of the SIR model by including two new features: the recovery threshold behaviour and the asymptomatic population.

The COVID-19 data is one type of spatio-temporal data because the data depends on events of the previous time and locations [11], [12]. Most infectious diseases result in high space and time trends, which are of the utmost importance to theoretical study [13]. In 2020, Ceylan [14] constructed a model based on univariate time series to estimate the COVID-19 epidemiological trend in Europe's most impacted countries: Italy, Spain, and France. The model is known as the ARIMA or Autoregressive Integrated Moving Average model. Subsequently, research on assessing and forecasting the epidemiology trend of COVID-19 continues to evolve using univariate or multivariate time series analysis. Mishra *et al.* [15] found the best ARIMA and seasonal ARIMA (SARIMA) models using COVID-19 cases in India. The model was used to forecast the daily confirmed cases and the total deaths. The model was further used to assist India's plans to fight against COVID-19. In 2021, Sun [16] modified the model of ARIMA to forecast the COVID-19 pandemic in Alberta, Canada. Yamamota *et al.* [17] proposed a spatio-temporal approach to include locations for assessing COVID-19 regionally in compliance with US COVID-19 mitigation initiatives. Furtado [18] incorporates regression and ARIMA of 20 countries in predicting the COVID-19 pandemic infection curves.

A spatio-temporal model is a multivariate approach in time series modeling. There are varieties of models that consider both location and time, generally known as spatio-temporal models. For example, Space-Time Autoregressive (STAR), Generalized Space-Time Autoregressive (GSTAR), and the most basic model is Vector Autoregressive (VAR). These spatio-temporal models cover various applications in various fields, such as disease transmission, data mining, economic growth, ecology, agriculture, and population growth. In the COVID-19 area of study, Sukarna [19] estimated and forecasted the COVID-19 cases in Sulawesi Island, Indonesia, using the Generalized Space-Time Autoregressive Integrated Moving Average (GSTARIMA) model. The result showed the necessity of the differencing since the non-stationary exists in the number of cases with the model, which was suitable for up to three days ahead but not further than that. Another study was also conducted in Indonesia Bandung province [20]. The model used was Generalized Space-Time Autoregressive Integrated (GSTARI), where the result shows good performance for a maximum of two days ahead. Both studies mentioned using only one weight approach: inverse distance and uniform weight, respectively.

The study on spatio-temporal in Malaysia was limited. Many studies still focus on the univariate time series model, ARIMA, in forecasting COVID-19 cases [21], [22]. However, one study found by Abdullah *et al.* [23]. The authors model the COVID-19 daily new cases using the GSTAR-ARIMA model, a hybrid model. The study focuses on the spatio-temporal between five states in Malaysia: Selangor, Sabah, Johor, Sarawak, and Perak. The study compares the performance of GSTAR and hybrid GSTAR-ARIMA based on uniform weights. The results show that the hybrid model gave better forecasting performance than the GSTAR model. Nevertheless, the GSTAR model remains the primary model in many applications since the model permits variable autoregressive parameters and spatial variation by region, which is more practical and realistic in application [24].

Based on the literature, most existing time series forecasting studies rely on univariate approaches like ARIMA and SARIMA, which have limited capacity to capture the complex spatio-temporal dynamics of disease transmission. To overcome these limitations, this study employs the GSTARIMA family, GSTARI, to analyse COVID-19 cases that exhibit nonstationary characteristics. While previous research typically utilised either inverse distance weights or uniform weights independently, this study will use

both approaches and compare the weight types to understand the spatial impact on disease spread better. These weights are crucial for accurately modelling the spatio-temporal relationships within the data, enabling the model to capture complex interaction patterns across space and time. Besides, previous research on COVID-19 forecasting in Malaysia faced challenges in establishing connections between large distant regions, such as west and east Malaysia, as shown by low disease transmission links. This study addresses this issue by enhancing the choices of the locations based on a correlation approach, thereby improving the understanding of spatio-temporal patterns. Therefore, this study aims to develop ARIMA and GSTARI models to forecast daily positive COVID-19 cases across different locations in high-density areas in Malaysia. The performance of these models will be compared to determine which provides the most accurate forecasts.

Materials and Methods

Data Source

The statistics of daily positive COVID-19 cases data recorded by the Ministry of Health Malaysia's (MOH) governance were obtained from MOH's official website. The number of issues were retrieved in Selangor, Malaysia, from 1 January 2021 until 7 August 2021. This dataset is categorised into two sets: the in-sample data comprises information from January 1, 2021, to July 31, 2021, and is used to build a COVID-19 model. The out-sample data, covering August 1, 2021, to August 7, 2021, is employed to assess the model's accuracy.

Box-Jenkins Methodology

Box-Jenkins method is used to develop the ARIMA model and GSTAR model. There are five primary steps involved. The first step is detecting stationarity in the data, followed by model identification. The next steps are estimation, diagnostic checking, and forecasting [24]. The variance and mean of the data are tested for stationary using the Box-Cox plot and augmented Dickey-Fuller (ADF) test, respectively. Data transformation is needed if the data is not stationary in variance. Meanwhile, the differencing approach is used to achieve stationary in the mean.

Identifying an ARIMA model depends on partial autocorrelation function (PACF) and autocorrelation function (ACF) plots. It should be considered simultaneously in assessing whether the patterns are cut off or die out. As for the GSTAR model, the space-time partial autocorrelation function (STPACF) and the space-time autocorrelation function (STACF) plots have the same function as ACF and PACF, are used to identify the GSTAR model. Subsequently, the least squares method is used to estimate the parameters. The diagnostic checking involved analysing residuals using a test known as Ljung-Box to verify that the residuals are independent. The processes will be repeated if the model is insufficient until it achieves a satisfactory ARIMA and GSTAR model.

Univariate: ARIMA Model

The ARIMA model with parameters (p,d,q) integrates elements from both the autoregressive (AR) model, usually denoted as p, and the moving average (MA) model, usually denoted by q, with an additional differencing order denoted as d. Generally, this model can be written as a backward shift operator as follows:

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)\varepsilon_t \tag{1}$$

where $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$. The p represents the order, which signifies the number of time lags for the autoregressive model. The q corresponds to the order, indicating the number of time lags for the moving average model; meanwhile, the differencing order is denoted as d. The ϕ_p represents the autoregressive parameter order p, θ_q denotes the moving average parameter of order q, and ε_t, represents the white noise with $\varepsilon_t \sim N(0, \sigma^2)$.

Multivariate: GSTAR Model

Let at location $i = 1,2,\ldots,N$ and time $t = 1,2,\ldots,T$ the series can be denoted as $Y_i(t) = (Y_1(t), Y_2(t), \ldots, Y_N(t))'$ and follows the GSTAR $(p; \lambda_1, \lambda_2, \ldots, \lambda_p)$ model with time order p and spatial $\lambda_1, \lambda_2, \ldots, \lambda_p$ that can be written as follows:

$$\boldsymbol{Y_i(t)} = \sum_{s=1}^{p}\left[\boldsymbol{\phi_{s0}Y_i(t-s)} + \sum_{k=1}^{\lambda_s}\boldsymbol{\phi_{sk}W_{ij}^{(k)}Y_i(t-s)}\right] + \boldsymbol{\varepsilon_i(t)} \quad (2)$$

where

$\boldsymbol{Y_i(t-s)}$ is the observed value at time lag s,
s is time autoregressive order,
k is spatial autoregressive order,
p is the time order of $p-th$ autoregressive term,
λ_s is the spatial order of $s-th$ autoregressive term,
$\boldsymbol{W_{ij}^{(k)}}$ represents the weight of $k-th$ order spatial,
$\boldsymbol{\phi_{s0}}$ are the diagonal matrices characterised by diagonal elements corresponding to the autoregressive values at different time lags for each location,
$\boldsymbol{\phi_{sk}}$ are the diagonal matrices that have diagonal elements that serve as space-time parameters, encompassing both spatial lag and time lag,
$\boldsymbol{\varepsilon_i(t)}$ is the white noise.

Nonetheless, when the model lacks stationarity in its mean, it becomes necessary to apply a differencing process. This process leads to the creation of another model known as the Generalized Space-Time Autoregressive Integrated (GSTARI) model. For instance, the GSTARI (1,1) model, with both time and spatial orders set at one, can be expressed as:

$$\boldsymbol{Y_i(t)} = \boldsymbol{\phi_{10}^{l}Y_i(t-1)} + \boldsymbol{\phi_{11}^{1}}\sum_{j=1}^{N}\boldsymbol{W_{ij}\,Y_j(t-1)} + \boldsymbol{\varepsilon_i(t)} \quad (3)$$

where $\boldsymbol{Y_i(t)} = \boldsymbol{Y_i(t)} - \boldsymbol{Y_i(t-1)}$, and $\boldsymbol{Y_i(t-1)} = \boldsymbol{Y_i(t-1)} - \boldsymbol{Y_i(t-2)}$.

The number of surrounding observed sites in spatial order influences spatial weight. Two types of spatial weight were used: uniform weight and inverse distance weight. Uniform weight is a form of weight that gives the same amount of weight value for each site. The weight can be calculated by using the formula below:

$$w_{ij}{}^{(k)} = \begin{cases} \dfrac{1}{n_i^{(k)}} & ; j \text{ is neighbour } i \text{ in } k-th \text{ order} \\ 0; \text{others} \end{cases} \quad (4)$$

where $w_{ij}{}^{(k)}$ is the weight between i and j. $n_i^{(k)}$ is the number of neighbours sites with a site.

The weight of the inverse distance method computes the real distance between geographic locations, which in this study are represented by latitude and longitude. The distance between these locations is defined as follows:

$$w_{ij} = \frac{1/d_{ij}}{\sum_{i\neq j} 1/d_{ij}} \quad (5)$$

where d_{ij} is the distance between the location i and j.

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (6)$$

where u and v represent the latitude and longitude coordinate location, respectively.

Accuracy Measure

The forecasting accuracy will be assessed using the Root Mean Square Error (RMSE). In general, RMSE can be formulated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (7)$$

where y_t is the actual data at time t, $\hat{y}_t$ is the forecast data at time t, and n is the number of observations.

The better model indicates a smaller RMSE value.

Results and Discussion

The Federation of Malaysia comprises thirteen states, with Selangor being the most developed, wealthy, and populated state. The rising number of COVID-19 cases in Selangor has garnered national attention due to concerning data indicating a significant spread of the virus in the region. Selangor's three most correlated districts are Petaling, Hulu Langat, and Klang. Figure 1 displays a time series plot illustrating the daily count of confirmed COVID-19 cases in these three districts. Based on Figure 1, a noticeable increasing and decreasing trend can be seen in the daily positive cases in all three districts. In addition, the trends are similar in all districts. This indicates that the direction of the number of daily positive cases in one location is highly correlated with the increasing and decreasing numbers in other locations. The results in Table 1, presenting the correlation of the number of daily positive cases between three locations, concurred with the indication.

The positive correlation of positive cases between locations will also increase in the other location and vice versa. Table 1 shows Hulu Langat is highly correlated with Petaling and Klang with a correlation coefficient (r) of 0.8140 and 0.7798, respectively. Klang and Petaling show a correlation with r = 0.7386. There would be higher chances of COVID-19 cases in Petaling and Klang if Hulu Langat reported positive cases. This is due to the densely populated area in these three districts, as the locations are strategic in both industries and residential areas.

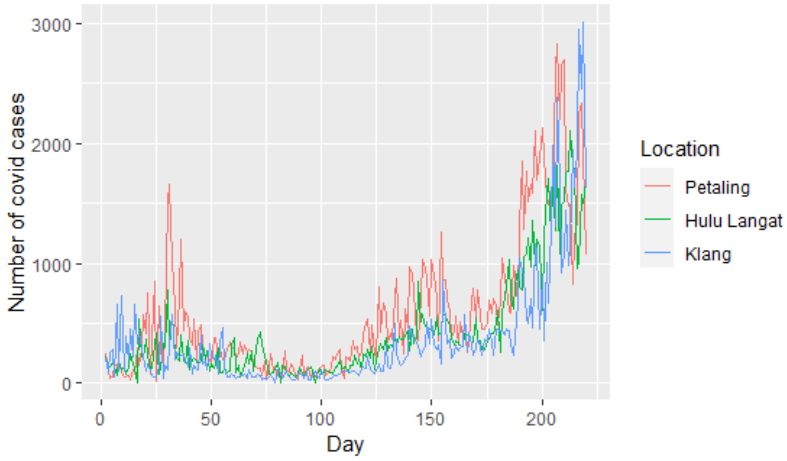


Figure 1. COVID-19 daily cases for three districts in Selangor

Table 1. Correlation on the number of positive cases in three Selangor districts

Districts	Petaling	Hulu Langat	Klang
Petaling	1	0.8140	0.7386
Hulu Langat	0.8140	1	0.7798
Klang	0.7386	0.7798	1

Summary for the daily number of positive COVID-19 cases in Petaling, Hulu Langat and Klang or the descriptive statistics are presented in Table 2. The statistics will provide insights into the distribution of daily confirmed COVID-19 cases. According to the average and standard deviation, Petaling showed the most significant number of cases, followed by Hulu Langat and Klang. This implies that the data

distribution in Petaling is widely spread. These results suggest that Petaling's highly dense population could cause a high number of positive cases.

Table 2. Descriptive statistics for the number of positive cases in three districts

Districts	Petaling	Hulu Langat	Klang
Min	15	7	5
1st Qu	193	134.5	89
Median	379	285	217
Mean	594.4	424	364.3
Std. Dev	590.5468	436.1509	488.5964
3rd Qu	778.5	462.5	408.5
Max	2832	2106	3006

The data's variance is stationary if the rounded lambda value from the Box-COX plot is 1. Table 3 shows the values of λ from the Box-Cox plot before and after transformation for all three locations. Based on Table 3, the variance of the data is not stationary since λ is equal to 0.5 for Petaling and Hulu Langat and λ = 0 for Klang. Thus, the data needs transformation. After being transformed accordingly, the data variance is said to be stationary, having λ = 1 for all three locations.

Table 3. The values of lambda from the Box-Cox plot before and after the transformation

Districts	Before	After	
	λ	Transformation	λ
Petaling	0.5	$y^{0.5}$	1.0
Hulu Langat	0.5	$y^{0.5}$	1.0
Klang	0	ln y	1.0

Subsequently, the stationarity of the data's mean is tested using the augmented Dickey-Fuller (ADF) test. Using the transformed data from all locations, the results in Table 4 show that the mean is not stationary for all locations (p-value > 0.05). Hence, a differencing approach with p-value > 0.05 was carried out to achieve stationary. Based on Table 4, the result shows that the p-values of all three locations are less than 0.05 after differencing. This suggests significance in achieving stationary for the mean of the data.

Table 4. The p-value of the ADF test before and after differencing

Districts	*p*-value before	*p*-value after
Petaling	0.7393	0.01
Hulu Langat	0.9835	0.01
Klang	0.5661	0.01

ARIMA model for each location can be identified through ACF and PACF plots in Figure 2. From Figure 2, the plot shows a die-down pattern in PACF plots for all three locations. This indicates the characteristics of the moving average (MA) model. Referring to ACF plots in Figure 2, all locations show ACF plots cut off at lag 1; hence, the model is confirmed to be MA(1). However, all the data went through different approaches to achieve stationary. Therefore, the proposed model to forecast the positive number of COVID-19 data for Petaling, Hulu Langat, and Klang can be denoted as the ARIMA (0,1,1) model. Table 5 represents the parameter estimation for each location.

Table 5. ARIMA(0,1,1) parameter estimations for three districts

Districts	Parameter	Coefficients
Petaling	θ_1	0.6348
Hulu Langat	θ_2	0.6752
Klang	θ_3	0.7362

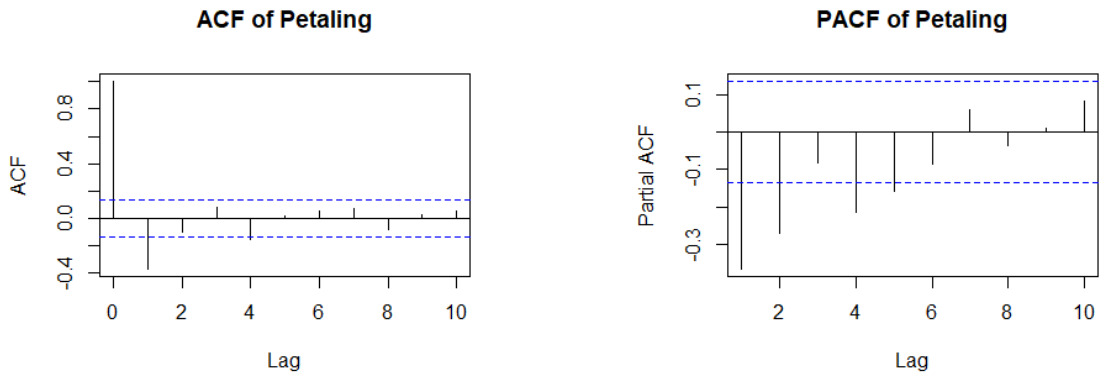
Diagnostic checks from the developed model are shown in Table 6. From Table 6, the significance level or p-values are 0.3397, 0.4590 and 0.4469, greater than 0.05. This indicates that the residuals are independent and uncorrelated. Therefore, ARIMA (0,1,1) is an appropriate univariate model to forecast the daily positive cases of COVID-19 in the three districts in Selangor.

Table 6. Results of Ljung- box Test for ARIMA (0,1,1) model

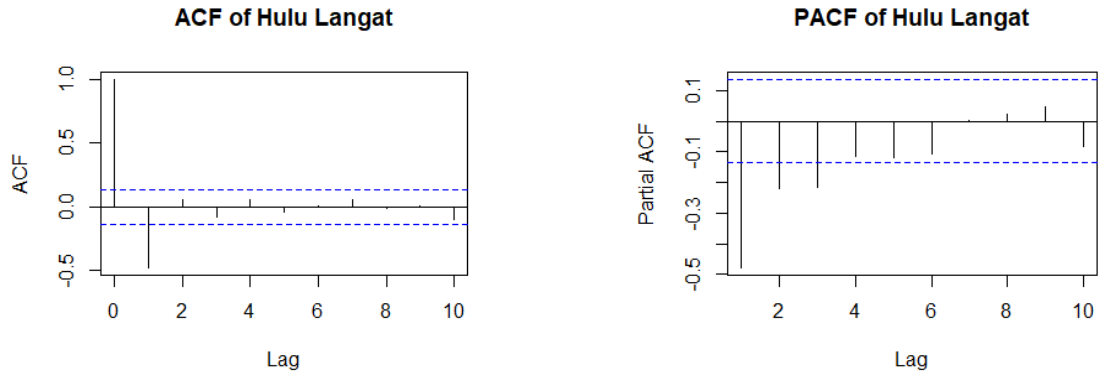
Districts	χ^2	p-value
Petaling	0.91164	0.3397
Hulu Langat	0.54844	0.4590
Klang	0.57846	0.4469

Model identification for multivariate modelling using the GSTAR model with both spatial weight uniform and inverse distance begins with observing the STACF plot and STPACF plot simultaneously. Note that the data has gone through a differencing process to achieve stationary, the model known as the Generalized Space-Time Autoregressive Integrated (GSTARI) model instead. The uniform weight matrix on COVID-19 data assumes that the COVID-19 cases in one location have the same effect on the COVID-19 cases in other areas since equal weight is given to each location. Therefore, the uniform weighting matrix, W_{ij} with their three locations, the number of locations near the location, i, are two could be written as follows:

$$W_{ij} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

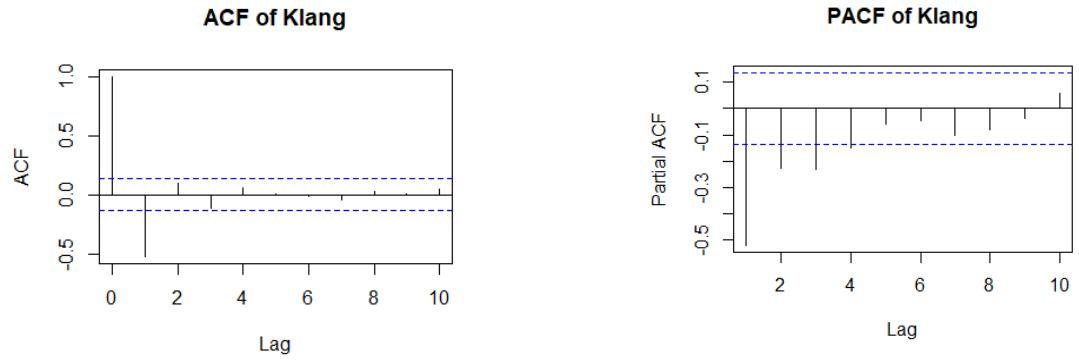


(a)



(b)

Continue to next page



(c)

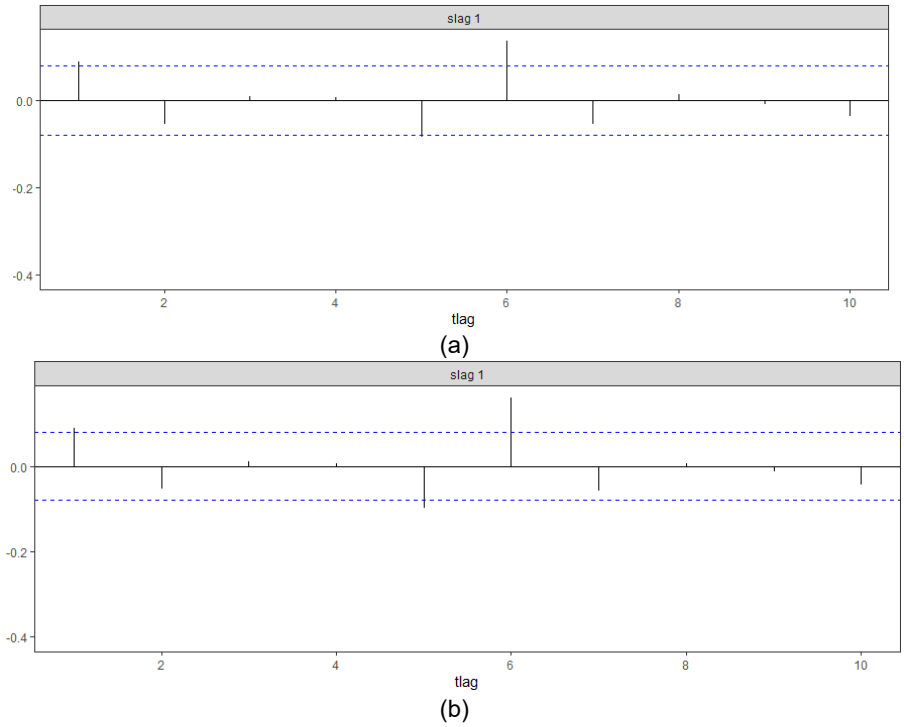
Figure 2. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plot of (a) Petaling, (b) Hulu Langat, and (c) Klang districts

Meanwhile, inverse distance weight assumes that the COVID-19 cases in one location are affected by distance or the closeness with other locations. The latitude and longitude for Petaling is (3.0833, 101.5833), Hulu Langat is (3.0833, 101.8333) and Klang is (3.0833, 101.4167). The distances between locations are then calculated using the Euclidean distance formula.

Therefore, the inverse distance weight matrix can be represented as follows:

$$W_{ij} = \begin{bmatrix} 0 & 0.3999 & 0.6001 \\ 0.6250 & 0 & 0.3750 \\ 0.7143 & 0.2857 & 0 \end{bmatrix}$$

As seen in Figure 3 and Figure 4, the STACF and STPACF plots show a die-down pattern for both spatial weights. Therefore, by the principle of parsimony, the multivariate model for both spatial weights is first identified as the GSTARI (1,1) model. The parameter estimation for GSTARI (1,1) for uniform spatial weight and GSTARI (1,1) for inverse distance spatial weight are presented in Table 7.



(a)

(b)

Figure 3. Space-time autocorrelation function (STACF) of (a) Uniform weight (b) Inverse distance weight

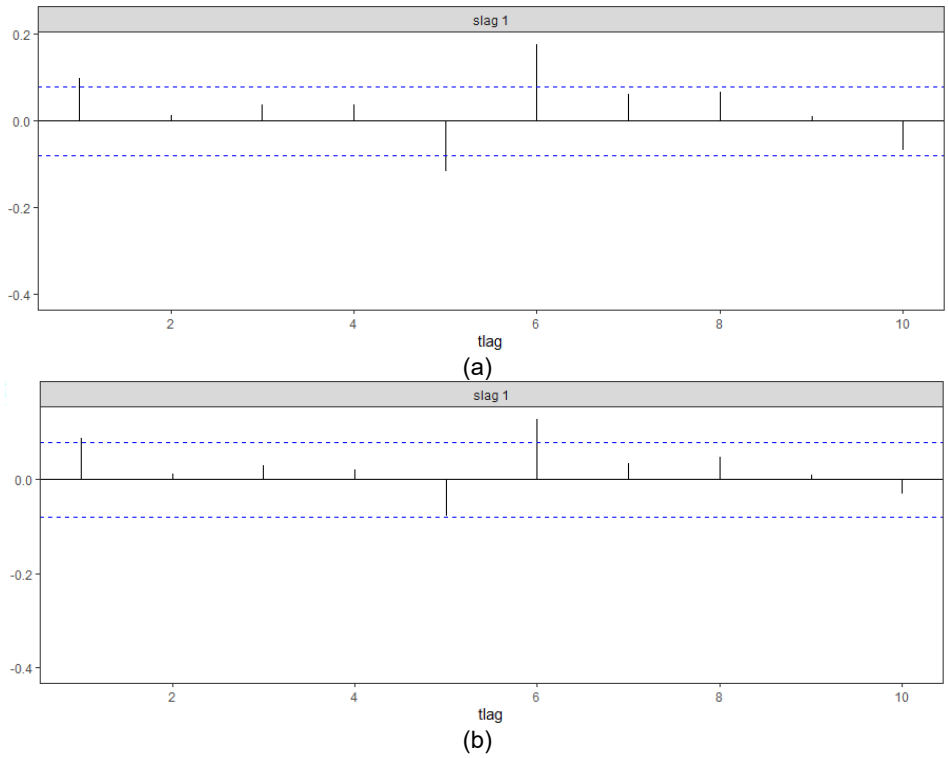


Figure 4. Space-time partial autocorrelation function (STPACF) of (a) Uniform weight (b) Inverse distance weight

Table 7. Parameter estimation of GSTARI (1,1) model

Districts	Parameter	Coefficients GSTARI(1,1)	
		Uniform	Inverse
Petaling	$\emptyset_{10}{}^{1}$	-0.3632	-0.3633
	$\emptyset_{11}{}^{1}$	0.0581	0.0454
Hulu Langat	$\emptyset_{10}{}^{2}$	-0.4633	-0.4638
	$\emptyset_{11}{}^{2}$	0.2075	0.2612
Klang	$\emptyset_{10}{}^{3}$	-0.5151	-0.5189
	$\emptyset_{11}{}^{3}$	0.0281	0.0292

Diagnostic checking of the residuals was done using the Ljung-box, and the results are presented in Table 8. For both models, the results are insignificant (p-value > 0.05) for all three districts. This verifies that the residuals are uncorrelated and independent. Consequently, GSTARI (1,1) for uniform spatial weight and GSTARI (1,1) for inverse distance spatial weight can be used for forecasting the daily positive cases of COVID-19 in the three districts in Selangor.

Table 8. Results of Ljung- box Test for GSTARI (1,1) model

Districts	GSTARI (1,1)- Uniform		GSTARI (1,1)- Inverse	
	χ^2	p-value	χ^2	p-value
Petaling	1.0554	0.3043	1.0511	0.3052
Hulu Langat	0.60659	0.4361	0.45565	0.4997
Klang	0.14065	0.7076	0.56519	0.4522

The GSTAR (1,1) was modelled as follows:

$$\boldsymbol{Y_i(t) = \left(\phi_{10}^{i} + \phi_{10}^{i} W_{ij}\right) Y_i(t-1) + \varepsilon_i(t)} \quad (8)$$

However, the series is non-stationary. Thus, a differencing approach is needed, and the series can be modelled by using the GSTARI (1,1) and can be written as:

$$\boldsymbol{Y_i(t) - Y_i(t-1) = \left(\phi_{10}^{i} + \phi_{11}^{i} W_{ij}\right)\{Y_i(t-1) - Y_i(t-2)\} + \varepsilon_i(t)} \tag{9}$$

From the estimated parameters shown in Table 7, a matrix equation of the GSTARI (1,1) model with uniform weight based on (9) can be formed as follows by denoting Petaling, Y_1, Hulu Langat, Y_2, and Klang, Y_3.

$$\begin{bmatrix}\hat{Y}_1(t)\\ \hat{Y}_2(t)\\ \hat{Y}_3(t)\end{bmatrix} = \left\{\begin{bmatrix}-0.3632 & 0 & 0\\ 0 & -0.4633 & 0\\ 0 & 0 & -0.5151\end{bmatrix} + \begin{bmatrix}0.0581 & 0 & 0\\ 0 & 0.2075 & 0\\ 0 & 0 & 0.0281\end{bmatrix}\begin{bmatrix}0 & 0.5 & 0.5\\ 0.5 & 0 & 0.5\\ 0.5 & 0.5 & 0\end{bmatrix}\right\}\begin{bmatrix}Y_1(t-1) - Y_1(t-2)\\ Y_2(t-1) - Y_2(t-2)\\ Y_3(t-1) - Y_3(t-2)\end{bmatrix} + \begin{bmatrix}Y_1(t-1)\\ Y_2(t-1)\\ Y_3(t-1)\end{bmatrix}$$

Thus, the equation of GSTARI (1,1) for uniform weight for each location is as follows:

In Petaling:

$$\hat{Y}_1(t) = -0.3632\{Y_1(t-1) - Y_1(t-2)\} + 0.0291\{Y_2(t-1) - Y_2(t-2)\} + 0.0291\{Y_3(t-1) - Y_3(t-2)\} + Y_1(t-1))$$

In Hulu Langat:

$$\hat{Y}_2(t) = 0.1038\{Y_1(t-1) - Y_1(t-2)\} - 0.4633\{Y_2(t-1) - Y_2(t-2)\} + 0.1038\{Y_3(t-1) - Y_3(t-2)\} + Y_2(t-1))$$

In Klang:

$$\hat{Y}_3(t) = 0.0140\{Y_1(t-1) - Y_1(t-2)\} + 0.0140\{Y_2(t-1) - Y_2(t-2)\} + 0.5151\{Y_3(t-1) - Y_3(t-2)\} + Y_3(t-1))$$

Using the estimated parameters from Table 7 and equation (9), the matrix equation for the GSTARI (1,1) model with inverse distance weighting can be shown as follows:

$$\begin{bmatrix}\hat{Y}_1(t)\\ \hat{Y}_2(t)\\ \hat{Y}_3(t)\end{bmatrix} = \left\{\begin{bmatrix}-0.3633 & 0 & 0\\ 0 & -0.4638 & 0\\ 0 & 0 & -0.5189\end{bmatrix} + \begin{bmatrix}0.0454 & 0 & 0\\ 0 & 0.2612 & 0\\ 0 & 0 & 0.0292\end{bmatrix}\begin{bmatrix}0 & 0.3999 & 0.6001\\ 0.6250 & 0 & 0.3750\\ 0.7143 & 0.2857 & 0\end{bmatrix}\right\} \begin{bmatrix}Y_1(t-1) - Y_1(t-2)\\ Y_2(t-1) - Y_2(t-2)\\ Y_3(t-1) - Y_3(t-2)\end{bmatrix} + \begin{bmatrix}Y_1(t-1)\\ Y_2(t-1)\\ Y_3(t-1)\end{bmatrix}$$

Thus, the equation of GSTARI (1,1) for inverse distance weight is as follows:

In Petaling:

$$\hat{Y}_1(t) = -0.3633\{Y_1(t-1) - Y_1(t-2)\} + 0.0181\{Y_2(t-1) - Y_2(t-2)\} + 0.0272\{Y_3(t-1) - Y_3(t-2)\} + Y_1(t-1))$$

In Hulu Langat:

$$\hat{Y}_2(t) = 0.1632\{Y_1(t-1) - Y_1(t-2)\} - 0.4638\{Y_2(t-1) - Y_2(t-2)\} + 0.0980\{Y_3(t-1) - Y_3(t-2)\} + Y_2(t-1))$$

In Klang:

$$\hat{Y}_3(t) = 0.0209\{Y_1(t-1) - Y_1(t-2)\} + 0.0083\{Y_2(t-1) - Y_2(t-2)\} - 0.4638\{Y_3(t-1) - Y_3(t-2)\} + Y_3(t-1))$$

From these three equations from both uniform weight and inverse distance weights, the daily number of positive cases of COVID-19 at time t correlates with the data at the previous time, $t-1$ and $t-2$ and is influenced by the COVID-19 cases at other places. Specifically, the daily positive COVID-19 cases in Petaling, Y_1 , Hulu Langat, Y_2, and Klang, Y_3 influence each other.

A negative coefficient in the GSTARI (1,1) model with a uniform weight equation implies that the previous period's positive COVID-19 cases negatively influenced the current period's positive cases. Conversely, a positive coefficient indicates a positive impact of past COVID-19 cases on the current period's positive patients. For example, the GSTARI (1,1) model with uniform weight in Petaling, Y_1 can be interpreted if the number of positive cases of COVID-19 in Petaling increased by 1 case, while at other locations and at other times, it was constant, the number of positive cases of COVID-19 in the Petaling district in the next period will decrease by 36%. This interpretation is the same as in the Hulu Langat, Y_2 and Klang, Y_3 districts equation.

Then, the RMSE measures are computed to compare the forecast performances of the ARIMA (0,1,1) model and GSTARI (1,1) model with uniform weight and distance weight shown in Table 9. The model with the lowest RMSE will be chosen as the best model. Overall, there are a few differences in the RMSE values. Based on locations, GSTARI (1,1) for inverse distance spatial weight is the best model to forecast the number of positive cases in Petaling and Klang. However, the best model for Hulu Langat is GSTARI (1,1) for uniform spatial weight.

Table 9. Results of accuracy measure based on RMSE

Districts	GSTARI (1,1)		ARIMA(0,1,1)
	Uniform	Inverse	
Petaling	640.77	637.80	643.81
Hulu Langat	426.17	430.02	388.17
Klang	729.39	712.95	823.52
Average	598.78	593.59	618.50

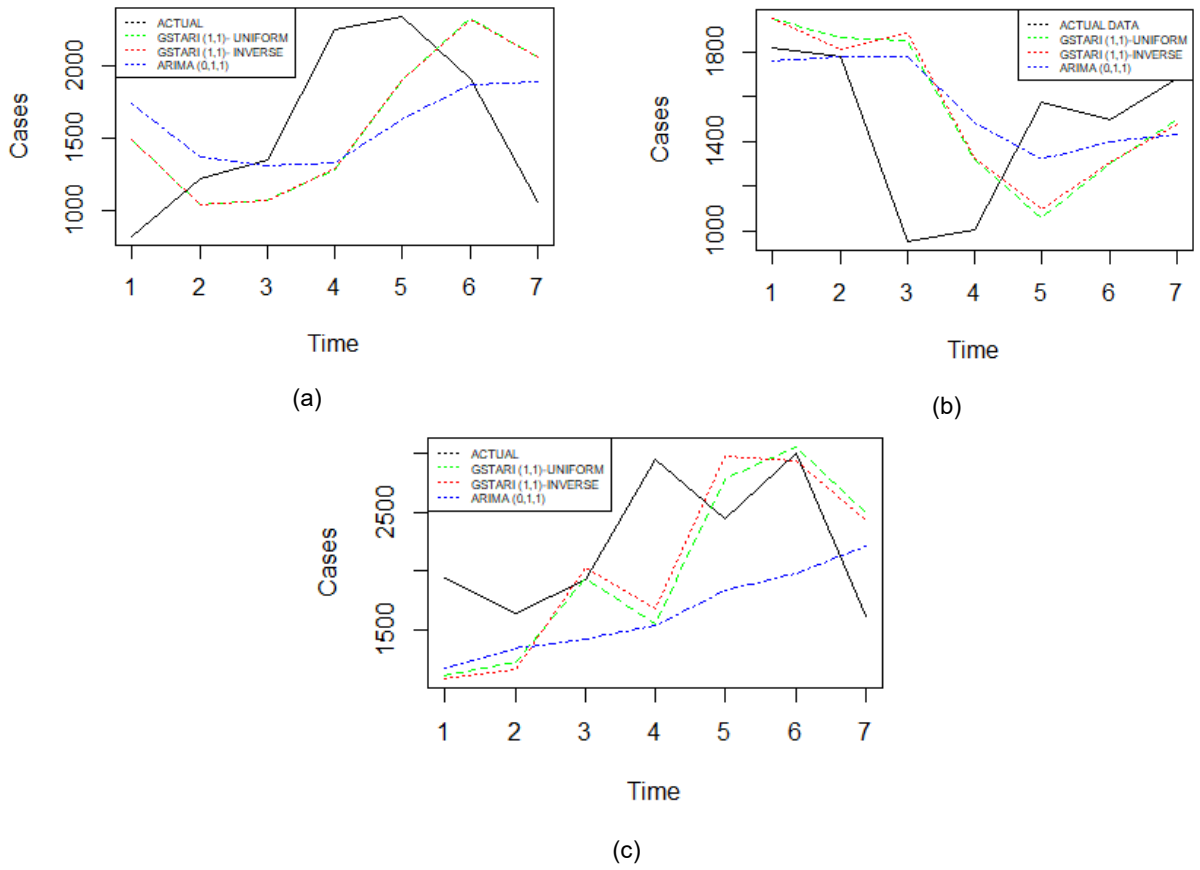


Figure 5. The plots of the ARIMA model and GSTARI models in (a) Petaling, (b) Hulu Langat, and (c) Klang districts

In selecting the best model to forecast the number of positive COVID-19 cases for all locations, the average RMSE for each model are considered. Based on the rank, the average RMSE, GSTARI (1,1) for inverse distance spatial weight is ranked first with RMSE = 593.59, followed by GSTARI (1,1) for the uniform spatial weight (RMSE = 598.78) and ARIMA (0,1,1) with RMSE = 618.50. As a result, the best model to forecast the number of positive COVID-19 cases, based on RMSE, is GSTARI (1,1) by using inverse distance spatial weight.

In addition to RMSE values, for each location, the actual data were plotted together with fitted data of each model. The fitted plots for all models show an almost similar pattern, especially in Petaling and Hulu Langat. However, based on Figure 5(c), the fitted plot for ARIMA (0,1,1) shows a different pattern and is far behind the actual data. From Figure 5, we can observe Figure 5 indicates that GSTARI (1,1) for both spatial weights show a consistent pattern to the actual data in all three locations.

Conclusions

Emerging infectious diseases, especially those caused by novel viruses, have become a major public health concern worldwide. Ever since the declaration of COVID-19 as a public health emergency worldwide, every country has taken action to control the outbreak. Forecasting the number of positive COVID-19 cases is crucial to monitor the potential distribution of COVID-19 infection, especially in the most affected location, such as Selangor. The forecasted numbers enable the COVID-19 management team to prepare appropriate plans of action to control this outbreak not only by locality but also nationally.

Through visualisation (Figure 1), the number of positive COVID-19 daily cases from 1 January 2021 until 7 August 2021 shows a possible correlation between three districts in Selangor: Petaling, Hulu Langat, and Klang. Therefore, a multivariate spatio-temporal model is convenient for forecasting the daily numbers of COVID-19 cases. This model pays attention to space and time aspects. The GSTAR model is widely used for forecasting spatio-temporal data since it can model heterogeneous locations.

Initially, the data in this analysis was not stationary in both mean and variance. Hence, the data transformation is needed to achieve stationary in variance, and a differencing approach is required to attain stationary in mean. Consequently, the GSTAR model is now known as the GSTARI model. Selecting the best model between the ARIMA model and the GSTARI model is based on its accuracy performance, which the RMSE values can determine. GSTARI models with different spatial weights provide better performance in Petaling and Klang.

On the other hand, ARIMA performed better than GSTARI models for Hulu Langat. As mentioned in Section 3, selecting the best model for all locations is ideal. Hence, according to the average RMSE values, GSTARI is the best model to be used as a forecasting tool for the positive number of COVID-19 cases in Selangor.

The outperformance of the GSTARI model over the ARIMA model could be because the GSTARI model considers the spatial weights. This indicates that the spatial effect is vital in forecasting the number of positive COVID-19 cases. The attention now turns to comparing GSTARI models with different spatial weights. The RMSE values for the GSTARI (1,1) model with uniform weight and inverse distance are relatively the same. However, to select the best model among those two, the average RMSE shows that the GSTARI (1,1) model with the inverse distance weight is the best model. Furthermore, the GSTARI (1,1) model with the inverse distance weight fits well because the parameters estimated are different across the model compared to the GSTARI (1,1) uniform weight model.

To conclude, the GSTARI model was used to forecast the number of positive COVID-19 cases in three Selangor districts for the short term. Although the results still need to be improved for long-term forecasting, the GSTARI model can forecast spatio-temporal data in real-world situations. In addition, the results can prove the interaction of observation between space and time simultaneously. Hence, the spatial weight considered in modelling GSTARI gave vital information to build the model from the data. One advantage of the GSTARI model is that it took less time to forecast the number of positive COVID-19 cases than the ARIMA model since the model building considered all locations simultaneously.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

This research was supported by Geran Putra (GP-IPM 9697800), funded by Universiti Putra Malaysia (UPM) through the Institute for Mathematical Research (INSPEM) Interim Researcher Initiative.

References

[1] Gill, B., *et al*. (2020). Modelling the effectiveness of epidemic control measures in preventing the transmission of COVID-19 in Malaysia. *International Journal of Environmental Research and Public Health, 17*, 1–13. https://doi.org/10.3390/ijerph17155509

[2] Umair, S., Waqas, U., & Faheem, M. (2020). COVID-19 pandemic: Stringent measures of Malaysia and implications for other countries. *Postgraduate Medical Journal, 97*, postgradmedj-2020. https://doi.org/10.1136/postgradmedj-2020-138079

[3] Mahalle, P., Kalamkar, M., Dey, N., Chaki, J., Hassanien, A. E., & Shinde, G. (2020). Forecasting models for coronavirus disease (COVID-19): A survey of the state-of-the-art. https://doi.org/10.36227/techrxiv.12101547

[4] Lai, D. (2022). Monitoring the SARS epidemic in China: A time series analysis. *Journal of Data Science, 3*(3), 279–293. https://doi.org/10.6339/JDS.2005.03(3).229

[5] Nishiura, H., Klinkenberg, D., Roberts, M., & Heesterbeek, J. A. P. (2009). Early epidemiological assessment of the virulence of emerging infectious diseases: A case study of an influenza pandemic. *PLoS ONE, 4*(8), e6852–. https://doi.org/10.1371/journal.pone.0006852

[6] Chen, D., Moulin, B., & Wu, J. (2014). *Analyzing and modeling spatial and temporal dynamics of infectious diseases*. Wiley. https://books.google.com.my/books?id=L-mNBQAAQBAJ

[7] Jia, W., *et al*. (2020). Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China. *Frontiers in Medicine (Lausanne), 7*, 169. https://doi.org/10.3389/fmed.2020.00169

[8] Nasab, S., Zahiri, A.-P., & Roohi, E. (2020). Prediction of peak and termination of novel coronavirus COVID-19 epidemic in Iran. *International Journal of Modern Physics C, 31*. https://doi.org/10.1142/S0129183120501521

[9] Tang, S., Wang, L., Li, D., Bragazzi, N. L., Xiao, Y., & Wu, J. (2020). Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of Clinical Medicine, 9*, 462. https://doi.org/10.3390/jcm9020462

[10] Sun, T., & Weng, D. (2020). Estimating the effects of asymptomatic and imported patients on COVID-19 epidemic using mathematical modeling. *Journal of Medical Virology, 92*. https://doi.org/10.1002/jmv.25939

[11] Borovkova, S., Lopuhaä, H., & Ruchjana, B. (2008). Consistency and asymptotic normality of least squares estimators in generalized STAR models. *Statistica Neerlandica, 62*, 482–508. https://doi.org/10.1111/j.1467-9574.2008.00391.x

[12] Wutsqa, D. U., Suhartono, & Sutijo, B. (2010). Generalized space-time autoregressive modeling. https://api.semanticscholar.org/CorpusID:12920082

[13] Siettos, C. S., & Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence, 4*. https://doi.org/10.4161/viru.24041

[14] Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of The Total Environment, 729*, 138817. https://doi.org/10.1016/j.scitotenv.2020.138817

[15] Mishra, P., *et al.* (2020). Modelling and forecasting of COVID-19 in India. *Journal of Infectious Diseases and Epidemiology, 6*, 162. https://doi.org/10.23937/2474-3658/1510162

[16] Sun, J. (2021). Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models. *Computer Methods and Programs in Biomedicine Update, 1*, 100029. https://doi.org/10.1016/j.cmpbup.2021.100029

[17] Yamamoto, N., Jiang, B., & Wang, H. (2021). Quantifying compliance with COVID-19 mitigation policies in the US: A mathematical modeling study. *Infectious Disease Modelling, 6*, 503–513. https://doi.org/10.1016/j.idm.2021.02.004

[18] Furtado, P. (2021). Epidemiology SIR with regression, ARIMA, and Prophet in forecasting COVID-19. *Engineering Proceedings, 5*, 52. https://doi.org/10.3390/engproc2021005052

[19] Sukarna, S., Syahrul, N., Sanusi, W., Aswi, A., Abdy, M., & Irwan, I. (2023). Estimating and forecasting COVID-19 cases in Sulawesi Island using generalized space-time autoregressive integrated moving average model. *Media Statistika, 15*, 186–197. https://doi.org/10.14710/medstat.15.2.186-197

[20] Alawiyah, M., Kusuma, D. A., & Ruchjana, B. N. (2021). Application of generalized space time autoregressive integrated (GSTARI) model in the phenomenon of COVID-19. *Journal of Physics: Conference Series, 1722*(1), 012035. https://doi.org/10.1088/1742-6596/1722/1/012035

[21] Singh, S., *et al.* (2020). Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. *The Journal of Infection in Developing Countries, 14*(09), 971–976. https://doi.org/10.3855/jidc.13116

[22] MA, E., ZA, M. A., & AR, J. (2020). Forecasting Malaysia COVID-19 incidence based on movement control order using ARIMA and expert modeler. *IIUM Medical Journal Malaysia, 19*(2). https://doi.org/10.31436/imjm.v19i2.1606

[23] Abdullah, S. N. S., Shabri, A., Saeed, F., Samsudin, R., & Basurra, S. (2023). Modelling COVID-19 daily new cases using GSTAR-ARIMA forecasting method: Case study on five Malaysian states. In *Advances in Data Science and Management* (pp. 439–448). https://doi.org/10.1007/978-3-031-36258-3_39

[24] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. Wiley Series in Probability and Statistics. Wiley. https://books.google.com.my/books?id=rNt5CgAAQBAJ