# Modelling COViD-19 Daily New Cases using GSTAR-ARIMA Forecasting Method:

## Case Study on Five Malaysian States

Siti Nabilah Syuhada Abdullah[1], Ani Shabri[1], Faisal Saeed[2], Ruhaidah Samsudin[3], and Shadi Basurra[2]

[1] Department of Mathematics, Faculty Science,
Universiti Teknologi Malaysia, 81300, Johor Bahru, Johor, Malaysia.
[2] DAAI Research Group, Department of Computing and Data Science,
School of Computing and Digital Technology, Birmingham City University,
Birmingham B4 7XG, UK
[3] School of Computing, Faculty of Engineering,
Universiti Teknologi Malaysia, 81300, Johor Bahru, Johor, Malaysia.
nabilah1991@graduate.utm.my, ani@utm.my, faisal.saeed@bcu.ac.uk

**Abstract.** On March 11, 2020, the World Health Organization declared COVID-19 to be a pandemic after the number of confirmed cases had surpassed 118,000 cases in more than 110 countries worldwide. To aid decision-makers in battling the epidemic, accurate modelling and forecasting of the spread of confirmed and recovered COVID-19 cases is essential. The non-linear patterns that are frequently seen in these situations have inspired us to create a system that can record such alterations. A hybrid method was approached in this study. Using hybrid models or combining several models has been a common practice to increase forecasting accuracy. Here, an error dataset was obtained from the GSTAR model previously and the error data for each location was modelled using ARIMA model. The final goal of this research is to develop a technique for predicting new COVID 19 cases using a hybrid GSTAR-ARIMA model. From March 16, 2020, to July 23, 2021, a case study was conducted on the number of daily confirmed COVID-19 cases in five Malaysian states. Global Change Data Lab at Oxford University furnished the dataset. GTAR-ARIMA with Uniform weights proves to be a viable forecasting option, ultimately proving to be the best model for forecasting daily new confirmed cases of COVID-19.

**Keywords:** GSTAR, ARIMA, Hybrid Model, Forecasting

## 1 Introduction

Corona-virus Disease 2019 (COVID-19) was discovered in Wuhan, China, towards the tail end of the year. The World Health Organization pro-claimed a pandemic on March 11th, 2020, following the reporting of 118,000 cases across 110 countries. There were more patient flows, which led to a lack of hospital beds nationwide and highly stressful situations. It is essential to comprehend the trend and dissemination of this virus to help governments and healthcare agencies to make urgent decision [1-3].

Many modelling, estimation, and forecasting techniques have been used to understand and control this epidemic. For instance, some studies used time-series methods including Auto-Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing to analyze and forecast patterns in the COVID-19 epidemic across many countries, including China, India, and Italy [4–9].

Numerous daily occurrences were connected not only to past occurrences but also to the places or the area in which they occurred. Time series analysis studies and spatial analysis studies were previously considered separately. When there are fixed sites spread throughout a number of locations, the time series analysis is used. Contrarily, the spatial analysis is used when a large number of locations are unknown but the time is. That said, the space-time analysis gained traction as science and technology developed.

The Generalized Space Time Autoregressive (GSTAR) model was initially brought by Ruchjana in 2002. Prior to this, [10] used the term GSTAR for a separate project in 1995; this project concerned the STAR model with a spatial correlation that happens at the same time and parameters in each location. Ruchjana described GSTAR as a STAR model in heterogeneous locations with different parameter values at each location instead [11]. To avoid ambiguous implications, this study uses Ruchjana's interpretation of GSTAR.

One of the most popular time series models is the autoregressive integrated moving average (ARIMA) model. The success of the ARIMA model can be attributed to its statistical properties and the well-known Box-Jenkins model-building process [12]. A number of exponential smoothing methods may be created using ARIMA models [13]. Whilst ARIMA models may explain a wide variety of time series, their fundamental disadvantage is the assumed linear structure of the model. Examples of these time series include pure moving average (MA), pure autoregressive (AR), and combination AR and MA (ARMA) series. The ARIMA model is unable to identify any nonlinear patterns since it is assumed that the time series values have a linear correlation structure.

Using hybrid models or combining several models has become a customary technique to boost the forecasting accuracy since the well-known M-competition [14], in which the integration of forecasts from more than one model frequently results in increased forecasting performance. The volume of literature on this topic has greatly expanded since the early work of Reid [15] and Bates and Granger [16]. Clemen [17] provided a detailed analysis and literature on this topic. Utilizing each model's unique feature to find diverse patterns in the data is the basic idea underlying model combinations in forecasting. According to theoretical and empirical studies [18–21], combining several approaches can be an effective and beneficial strategy to improve forecasts.

The goal of this research is to develop a forecasting method for COVID 19 cases. A hybrid method was approached. This is where an error dataset was obtained from the GSTAR model previously and the error data for each location was modelled using ARIMA model. In this research, a case study was conducted, about the daily reported cases of COViD-19 in the five most populous Malaysian states from 16 March 2020 to 23 July 2021. A training set makes up 80% of the data set, while a test set makes up 20%.

The paper is structured as follows. We evaluated the experimental setup in the next section (2). In Section 3, we investigated the case study by describing the data. After that, Section 4 presents the empirical findings. The final reflections are included in Section 5.

## 2 Framework of Study

The basic structure of the proposed methods for forecasting is shown in Figure 1. A hybrid method was approached. Here, an error dataset was obtained from the GSTAR model previously and the error data for each state was modelled using ARIMA model. A rough representation of the procedure carried out is as follows:

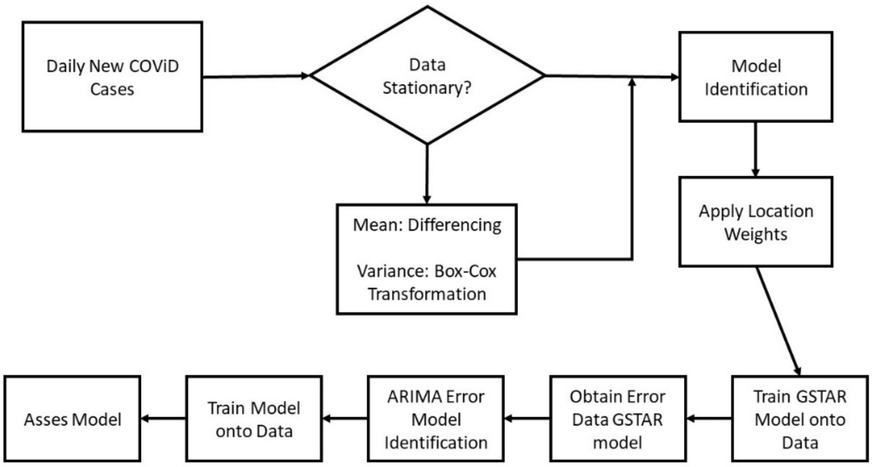

**Fig. 1.** GSTAR-ARIMA Procedure

The Augmented Dickey Fuller (ADF) test or examination of the MACF and MPACF cross correlation matrix schemes was used to initially determine whether or not the data are stationary. Establishing the temporal and the spatial order comes next when the data has reached a steady state. The determination of location weights is a challenge that frequently arises in GSTAR modelling [22]. There are three main commonly used weights. In this study, however, uniform location weights are used. Uniform location weight is defined as [22, 23]:

$$w_{ij} = \frac{1}{n_i} \tag{1}$$

where $n_i$ is the state's count of nearby sites to location $i$ in the spatial lag 1. The attributes of the weight in this model are

$$W_{ij} > 0, W_{ii} = 0, \sum_{j=1}^{N} W_{ij} = 1, \forall_i, \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} = N \tag{2}$$

The weight value provided here was equally assigned for each location. As a result, this location weight was frequently applied to data that is homogeneous or that uses the same distance for each place [10, 22]. The weight of $W_{ij}$ in lag 1 is expressed by W in the form of $n \times n$ matrix as follows:

$$W = \begin{bmatrix} 0 & W_{12} & \dots & W_{1N} \\ W_{21} & 0 & \dots & W_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & \dots & 0 \end{bmatrix} \tag{3}$$

Once the GSTAR model was fully developed, a dataset of error values obtained from comparing the GSTAR model output with the original data was obtained. This error dataset was used for error modelling using ARIMA.

An autoregressive integrated moving average, or ARIMA, was designed by Box and Jenkins [24, 25]. The $ARIMA\ (p, d, q)$ method can be determined using a time series of the actual value, $y_t$, where $t$ is the time period, and the process is given by:

$$y_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \tag{4}$$

Where $c$ is a constant, $\varphi_1, \varphi_2, \dots, \varphi_p$are parameters of autoregressive (AR), whereas $\theta_1, \theta_2, \dots, \theta_q$ are moving average (MA) parameters. The random errors, $\varepsilon_t$ are assumed to be independently and identically distributed with zero mean and constant variance, $\sigma^2$. Theoretically, ARIMA models are the most diverse category of forecasting models for the time series that may be transformed to become stationary using techniques like differencing [25].

The order of ARIMA was determined using PACF and the ACF. All potential models are mentioned for Ljung-Box test diagnostic verification. All of the large p-values for the Ljung-Box statistics indicate a good model. The fact that there are no patterns in the residues suggest that all of the information has been retrieved [25]. The accuracy of the created model were be assessed using the Root Mean Square Error (RMSE). RMSE was used to assess the precision of the model developed:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2} \tag{5}$$

This study compared the forecasting accuracy obtained from modelling COVID-19 confirmed cases using the GSTAR-ARIMA model with those obtained from GSTAR.

# 3 Case Study: Five Malaysian States

The key goal of this work is to forecast COVID 19 and predict the infection spread. This analysis is based on routine data from confirmed cases received in five most populated states in Malaysia. The states and population number reported by The Department of Statistics Malaysia in their official website as of 30th June 2021 are Selangor (6,573,862), Sabah (3,812,391), Johor (3,806,270), Sarawak (2,827,624) and Perak (2,509,587) [25]. The data of the daily confirmed cases used were in between 16th March 2020 and 23rd July 2021. The COVID-19 data in Malaysia are now made readily available online through The COVIDNOW initiative website (https://covidnow.moh.gov.my). The Ministry of Health (MoH) and the COVID-19 Immunization Task Force's open data efforts have been a novel and pleasant experience for the MoH and for everyone in Malaysia. Public and commercial stakeholders may now assess regulations, exchange new perspectives and analyses of the issue, and uphold better standards, thanks to open data. To assess the model's competitiveness, the COVID-19 data set was divided into an 80:20 training and testing set to measure the model effectiveness.

The graphic display of the data is shown in Figure 2. This disease has remained relatively under control up until September 2022. Until now, it has risen dramatically since its emergence, hitting a sizeable number of confirmed cases of 2525 on 31 December 2020. Table 1 includes a description of the data used in this analysis. The values for Minimum (Min.), Maximum (Max.), Mean, Median, 1st and 3rd Quartile (1st Qu., 2nd Qu.), are rounded up as they represent New Cases of COVID in each state. Here, we can infer that Selangor state had the highest number of new cases daily maximum value. Selangor also had the largest variance meaning that the data fluctuated greatly from day to day. In the 1st quartile, Perak had a value of 0 indicating that it was the last state out of the 5 to have COVID cases.

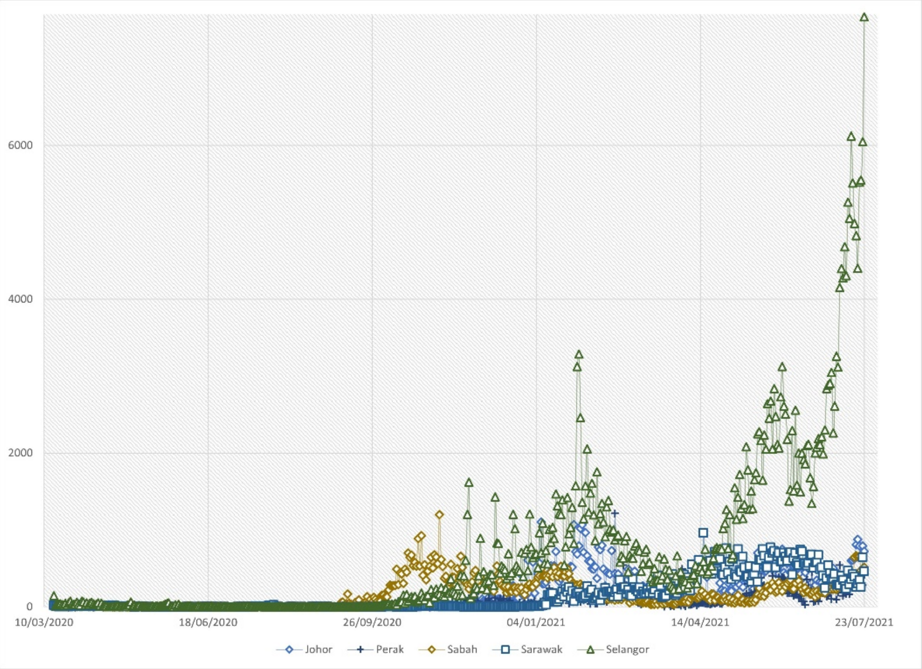

**Fig. 2.** Daily Confirmed new Cases of COVID-19 in Malaysia

**Table 1.** Summary of Dataset

| | Johor | Perak | Sabah | Sarawak | Selangor |
|---|---|---|---|---|---|
| Min. | 0 | 0 | 0 | 0 | 0 |
| 1st Qu. | 1 | 0 | 2 | 1 | 6 |
| Median | 26 | 17 | 91 | 8 | 221 |
| Mean | 167 | 66 | 158 | 148 | 710 |
| 3rd Qu. | 310 | 83 | 263 | 244 | 958 |
| Max. | 1103 | 1215 | 1199 | 960 | 7672 |
| Variance | 53 093.75 | 12 398.57 | 33 829.72 | 47 586.27 | 1 259 441.00 |
| Std. Dev. | 230.42 | 111.35 | 183.93 | 218.14 | 1 122.25 |

# 4 Results and Findings

## 4.1 Stationary Check

When there is no consistent change in the mean or variance values, the data is considered to be stationary in a time series. According to Markidakis et al. (1992), the visual representation of a time series data plot is frequently adequate to determine if the data is stationary or otherwise. Additionally, the Augmented Dickey Fuller (ADF) test or the MACF and MPACF cross correlation matrix methods were used to explicitly determine the stationarity of the data. The data requires differencing if the MACF and

MPACF plots show a steady decline, indicating that the data is not stationary to the mean [22]. In contrast, if the superior and inferior bounds of the lambda (λ) are smaller than zero, the data is not stable and resistant to variations, necessitating the use of a Box Cox transformation to make the data stationary.

**Table 2.** Stationary Check

| | Stationarity of Data | ADF p-value | ndiffs(data) | lambda |
|---|---|---|---|---|
| 1 | **Johor** | 0.5839 | 1 | 1 |
| 2 | **Perak** | 0.0513 | 0 | 1 |
| 3 | **Sabah** | 0.6436 | 1 | 1 |
| 4 | **Sarawak** | 0.5952 | 1 | 1 |
| 5 | **Selangor** | 0.9900 | 2 | 1 |

Table 2 summarizes the stationarity of data. As shown in this table, p-values below 0.05 demonstrate stationarity, whereas p-values upwards of 0.05 imply non-stationarity. In this study, only Perak's data was readily stationary. All other states' data and differencing were added to these data. Johor, Sabah and Sarawak went through 1 differencing while Selangor had to undergo differencing twice to reach stationarity.

### 4.2 Construction of GSTAR Model

The next stage was to establish the temporal and the spatial order after the data is stationary. The order spatial employed here is order spatial 1, as a higher order spatial results in a more complicated model and less interpretability. When it comes to order time, the Vector Autoregressive (VAR) technique is used, which involves examining the least Akaike's Information Criterion (AIC) value.

**Table 3.** Time Order Selection using VAR

| | **AIC(n)** | **HQ(n)** | **SC(n)** | **FPE(n)** |
|---|---|---|---|---|
| **1** | 49.797 | 49.899 | 50.057 | 4.23E+21 |
| **2** | 48.657 | 48.844 | 49.133 | 1.35E+21 |
| **3** | 48.124 | 48.396 | 48.816 | 7.94E+20 |
| **4** | 47.755 | 48.113 | 48.664 | 5.50E+20 |
| **5** | 47.423 | 47.865 | 48.548 | 3.94E+20 |
| **6** | 46.927 | 47.454 | 48.268 | 2.40E+20 |
| **7** | 46.693 | 47.305 | 48.250 | 1.90E+20 |
| **8** | 46.623 | 47.321 | 48.397 | 1.77E+20 |
| **9** | 46.495 | 47.277 | 48.485 | 1.56E+20 |
| **10** | 46.374 | 47.242 | 48.581 | 1.39E+20 |
| **P Selection** | **10** | **10** | **7** | **10** |

Based on the table above, p=10 is AR (10), hence the GSTAR model formed is GSTAR (1; 10). Next, the study tackles the task of selecting the weight to be used in the GSTAR Model. The uniform location weight used in the study is shown below:

$$w_{ij} = \frac{1}{n_i} = \frac{1}{5} = 0.2 \tag{6}$$

where $n_i$ =5 declares the 5 states (Johor, Perak, Sabah, Sarawak, Selangor) in the spatial lag 1.

The weight value provided by this location weight is equally assigned to each location. As a result, this location weight is frequently applied to data that is homogeneous or that uses the same distance for each place. The weight of $W_{ij}$ in lag 1 is expressed by W in the form of 5 × 5 matrix as follows:

$$W = \begin{bmatrix} 0 & \cdots & 0.2 \\ \vdots & \ddots & \vdots \\ 0.2 & \cdots & 0 \end{bmatrix} \tag{7}$$

## 4.3 GSTAR-ARIMA

**Table 4.** ARIMA Error Modeling

| State | ARIMA Model |
|---|---|
| **Johor** | ARIMA(4,0,0) |
| **Perak** | ARIMA(4,1,1) |
| **Sabah** | ARIMA(2,1,3) |
| **Sarawak** | ARIMA(4,1,0) |
| **Selangor** | ARIMA(2,1,2) |

To further improve the forecasting model, a hybrid method was approached. Here, an error dataset was obtained from the GSTAR model previously and the error data for each state was modelled using ARIMA model.

## 4.4 Performance Model

**Table 5.** Models' Performance Comparison

| | Training Set | | Testing Set | |
|---|---|---|---|---|
| **State** | **GSTAR** | **GSTAR-ARIMA** | **GSTAR** | **GSTAR-ARIMA** |
| **Johor** | 92.367 | 86.559* | 248.807 | 116.607* |
| **Perak** | 72.812 | 72.048* | 106.895 | 58.690* |
| **Sabah** | 80.349 | 68.340* | 119.358 | 49.188* |
| **Sarawak** | 92.714 | 80.932* | 331.158 | 135.286* |
| **Selangor** | 203.709* | 206.471 | 591.977 | 450.528* |

Table 5 illustrates the RMSE values for Johor, Perak, Sabah, Sarawak and Selangor for both GSTAR and GSTAR-ARIMA models during the training and testing stages. Models were identified using the training portion of the data. In the training stages, the GSTAR-ARIMA method outperformed the basic GSTAR model for all states except in Selangor. However, the testing stage has demonstrated astounding capabilities of the GSTAR-ARIMA model where all states performed better using the hybrid model compared to basic GSTAR. It was found that the hybrid model improves the performance significantly compared to the GSTAR models.

## 5 Conclusion and Recommendations

It is safe to conclude that GSTAR-ARIMA is suitable to be used in forecasting daily new confirmed cases of COVID. In the conducted case study, using data from Johor, Perak, Sabah, Sarawak and Selangor, this paper found that GSTAR (1,10) was the best Error Modeling conducted using ARIMA for all five states. GTAR-ARIMA with uniform weights proves to be the best model. Despite the fact that the hybrid model has performed well, it is unfortunate that the spread is increasing. Meanwhile, the incidence of infections is rising exponentially, and the number of infections is increasing. The accuracy of these predictions depends on a number of external factors. Further studies incorporating other Malaysian states should be carried out to get a more holistic view of the spreading trend. Furthermore, this model could also be adopted to studies of other epidemics such as HIV-AIDS, Polio, etc.

## References

1. Velásquez RMA , Lara JVM . Forecast and evaluation of COVID-19 spreading in USA with reduced-space gaussian process regression. Chaos Solitons Fractals 2020:109924 .
2. Yousaf M , Zahir S , Riaz M , Hussain SM , Shah K . Statistical analysis of fore- casting COVID-19 for upcoming month in Pakistan. Chaos Solitons Fractals 2020:109926 .
3. Ribeiro MHDM , da Silva RG , Mariani VC , dos Santos Coelho L . Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. Chaos Solitons Solitons Fractals 2020:109853 .
4. Dehesh T, Mardani-Fard H, Dehesh P. Forecasting of COVID-19 confirmed cases in different countries with ARIMA models. medRxiv 2020. doi: 10.1101/2020.03. 13.20035345 .
5. Gupta R , Pal SK . Trend analysis and forecasting of COVID-19 outbreak in India. medRxiv 2020 .
6. Chintalapudi N , Battineni G , Amenta F . COVID-19 disease outbreak forecast- ing of registered and recovered cases after sixty day lockdown in Italy: a data driven model approach. J Microbiol Immunol Infect 2020.
7. Kucharski AJ , Russell TW , Diamond C , Liu Y , Edmunds J , Funk S , et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020 .

8. Wu JT , Leung K , Leung GM . Nowcasting and forecasting the potential domes- tic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020;395(10225):689–97 .
9. Zhuang Z , Zhao S , Lin Q , Cao P , Lou Y , Yang L , et al. Preliminary estimation of the novel coronavirus disease (COVID-19) cases in iran: a modelling analysis based on overseas cases and air travel data. Int J Infect Dis 2020;94:29–31 .
10. Terzi S 1995 Maximum likelihood estimation of a generalized STAR(p;1p) model Journal of The Italian Statistical Society 3 pp 377-393.
11. Ruchjana B N 2002 Suatu model generalisasi space-time autoregresi dan penerapanny pada produksi minyak bumi Disertation of Doctoral Program Institut Teknologi Bandung
12. G.E.P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, CA, 1970.
13. E.D. McKenzie, General exponential smoothing and the equivalent ARMA process, J. Forecasting 3 (1984) 333–344.
14. S. Makridakis, A. Anderson, R. Carbone, R. Fildes, M. Hibdon, R. Lewandowski, J. Newton, E. Parzen, R. Winkler, The accuracy of extrapolation (time series) methods: results of a forecasting competition, J. Forecasting 1 (1982) 111–153.
15. D.J. Reid, Combining three estimates of gross domestic product, Economica 35 (1968) 431–444.
16. J.M. Bates, C.W.J. Granger, The combination of forecasts, Oper. Res. Q. 20 (1969) 451–468.
17. R. Clemen, Combining forecasts: a review and annotated bibliography with discussion, Int. J. Forecasting 5 (1989) 559–608.
18. S. Makridakis, Why combining works? Int. J. Forecasting 5 (1989) 601–603
19. P. Newbold, C.W.J. Granger, Experience withforecasting univariate time series and the combination of forecasts (withdiscussion), J. R. Statist. Soc. Ser. A 137 (1974) 131–164.
20. F.C. Palm, A. Zellner, To combine or not to combine? Issues of combining forecasts, J. Forecasting 11 (1992) 687–701.
21. R. Winkler, Combining forecasts: a philosophical basis and some current issues, Int. J. Forecasting 5 (1989) 605–609.
22. Fadlurrohman, A. (2020). Integration of GSTAR-X and Uniform location weights methods for forecasting Inflation Survey of Living Costs in Central Java. Journal of Intelligent Computing and Health Informatics (JICHI), 1(1), 20-25.
23. Suhartono and Subanar 2006 The Optimal Determination of Space Weight in GSTAR Model by Using Cross-correlation Inference, Journal of Quantitative Methods. Journal Devoted the Mathematical and Statistical Application in Various Field, 2(2) pp 45-53
24. G.E.P. Box, & G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, San Fransisco, 1976.
25. Abadan, S., & Shabri, A. (2014). Hybrid empirical mode decomposition-ARIMA for forecasting price of rice. Applied Mathematical Sciences, 8(63), 3133-3143.
26. Population Clock by State. Department of Statistics Malaysia Official Portal. (n.d.). Retrieved March 26, 2022, from https://www.dosm.gov.my/v1/index.php?r=columnnew%2Fpopulationclock