'ULŪM ISLĀMIYYAH THE MALAYSIAN JOURNAL OF ISLAMIC SCIENCES
VOL.33 S4 (2021) ISSN:1675-5936 | eISSN:2289-4799

COVID-19: Pushing the Limits of Time Series Big Data

Norita Md Norwawi [1]

[1] Computer Science Program, Faculty of Science and Technology, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia

* Corresponding author: norita@usim.edu.my

Abstract

In Malaysia, the pandemic coronavirus disease 2019 (COVID-19) was first detected on 25th January and has been spreading massively and reported to have reached more than 20,000 new cases per day from July to August 2021. COVID-19 data is voluminous describing the pandemic trend around the globe. How does Big Data help decision-makers understand the pandemic behaviour which is very crucial in responding to the situation? How do data analytics on the COVID-19 spreading pattern which is time-series in nature may provide insight into the situation that may lead to a better response through forecasting future trends? This paper aims to explain the concept of Big Data and its applications that demonstrates its potential for responding to the pandemic. COVID-19 data analytic is proposed using sliding window time-series forecasting method and demonstrated using data from 25th January until 10th October 2020 obtained from the Malaysian Ministry of Health and Department of Statistics Malaysia website. The data analytics demonstrated the value gain for useful insights.

Keywords: *Big Data, data analytics, predictive analytics, time-series*

Manuscript Received Date: 08/10/21 Manuscript Acceptance Date: 5/12/21 Manuscript Published Date: 25/12/21

DOI: 10.33102/uij.vol33noS4.416

COPE COMMITTEE ON PUBLICATION ETHICS

Universiti Sains Islam Malaysia
uijournal.usim.edu.my

1.0 Introduction

Novel Coronavirus disease is an infectious disease first detected in Wuhan on the 31st December 2019 caused by a newly discovered coronavirus known as COVID-19. It has the symptoms of a severe acute respiratory syndrome such as coughing, persistent chest pain and difficulty in breathing (WHO, 2012). The disease was first reported in Malaysia on 25 January 2020 on 3 travellers from China gradually lead to the first sporadic infected person reported on 11th March. This implies that he did not contract the disease due to travelling overseas or close contact. Since then, there was a sharp increase in the trend of people infected. In Malaysia, a country of 31 million people, over 15,657 have been infected with COVID-19, 69.70 per cent have been cured with 157 fatalities implying a fatality rate of 1.0% reported until 11th October 2020. These statistics prove that Malaysia's strategies in attempting to break the chain of infection through several MCO strategies since 18 March 2020 show a remarkable result (MOH, 2020; DOSM, 2020). Analysis of the current situation, its trend and patterns are very crucial in supporting and facilitating government decision-making especially at a major critical turning point of an event.

2.0 Problem Statement

In Malaysia, the first reported cases imported from China travelers was on 25 January and a local sporadic infected person was detected on 11 March 2020. Movement Control Order (MCO) was declared on 18th March upon the trace of the first spike in the epidemic trend. The third wave took place end of September with much higher positive cases compared to the maximum from the second wave, especially in Sabah and Kedah. With the daily voluminous data across the country and the globe, how will Big Data be used to forecast the pandemic spreading pattern in Malaysia thus assist in the decision-making of the Health Ministry and National Security Division? What is the duration of the onset of the pandemic until recovery or fatality? How would time series forecasting be useful in determining the duration of an infection?

3.0 The Role of Big Data in Managing the Covid19 Pandemic

With the advent of the Internet and Web 2.0 technology, voluminous data are being created every millisecond. However, these massive data remain useless in their raw form unless transformed into a useful form to assist for problem solving, decision-making, providing useful insights that eventually can become an economic commodity.

Examples of big data applications by key players in the industry are Amazon.com strategized its business strategy using personalized marketing and recommendations system with a targeted segment of customers that has similar interests identified through books search and bought. Besides that, digital marketers are benefitting from the competitive edge by pushing ads to targeted social media customers such as Facebook, Twitter, Instagram and YouTube by studying their online behaviour through sentiment analysis. The mass volume of big data is characterized with 5 elements which are Volume, Velocity, Variety, Veracity and Value known as 5V (O'Reilly Media, 2012; Hamdan et al., 2018;) describe as follows :

Volume Huge size of data is generated every second, for instance, emails, websites, Internet of Things sensors, Whatsapp messages, status posting on social media such as Facebook, Twitter, Instagram and so on. Due to the mass volume of the data collected, alternative processing and analysis techniques are required to handle the size.

Velocity The speed of the data being created and distributed is rapid almost an instance of time such as sending messages, photos, social media status, stories and live streaming over the Internet. Cloud and mobile platform. The current big data technology can execute real-time data processing such as radar tracking for air flight, Global Positioning Sensor (GPS) tracking on transports, apps, Google map and laptop.

Variety The data in the current Algorithm Age or the Algorithmic Economic are various formats such as text, images, video, audio, structured and unstructured in contrast with the Information Age that analyses structured data.

Veracity Since 80% of the data is unstructured, extraction for hidden patterns is very challenging due to the dynamicity, fast-changing volatile data, data loss and noises, thus influence the accuracy, trustworthiness, and data quality in general.

Value The patterns extracted from the data give value for decision-making and problem-solving which is the key importance of Big Data. For example, during the Obama election campaign in 2012, many data scientists were hired to mine data and conduct sentiment analysis to gauge the inclination of the voters towards certain issues which gave a huge advantage to him (Hamdan, 2018).

Due to the 5V characteristics of Big Data, analysing those voluminous pandemic data using conventional techniques such as databases, Statistics and Data Engineering is no more efficient and effective. A special database is required with higher capacity suitable for millions of data being collected and analysed within minutes to give useful information to users.

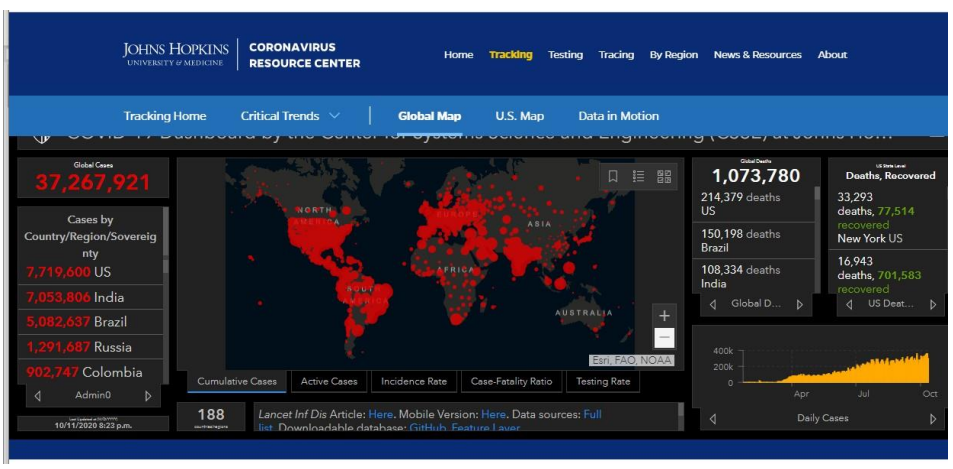
3.1 Big Data Applications for COVID 19

Some examples of Big Data applications are shortest route search using Google Map or Waze, recommendation systems such as Amazon.com for books, Trivago for hotels and weather forecast that can give an early warning on an expected storm, earthquakes and so on. Dr Kamran Kahn, an epidemiologist and practising physician trained in advanced data analytics launched BlueDot, a Toronto-based startup that developed proprietary software as a service capable of locating, tracking, and predicting the spread of infectious disease in 2015. The BlueDot engine searched data every 15 minutes every day gathering over 150 diseases and syndromes around the world. He builds the world's first Artificial Intelligence (AI)-based infectious disease surveillance equipped with Big Data functionality with a global early warning system capable to track and contextualize infectious disease risks. BlueDot was the first to detect the epidemic on a cluster of "unusual pneumonia" on 31 December 2019 from articles in Chinese that reported 27 pneumonia cases happening around a market that had seafood and live animals in Wuhan, China (Bragazzi et. al, 2020; McCall, 2020).

BlueDot was even able to anticipate the spread based on the highest volume air flights movements from Wuhan to Bangkok, Hong Kong, Tokyo, Taipei, Phuket, Seoul, and Singapore which in actuality was also the first places to record COVID-19 cases. BlueDot demonstrated the capability of AI and Big Data to co-exist with the decision-makers to give useful insights and predictive analytics that can speed up decisions, facilitate strategic problem solving and innovate solutions.

Another prominent example of using Big Data for monitoring the pandemic is John Hopkins University, the United States of America that has specially dedicated a website with important sources of information on COVID 19 with real-time global data visualization on the spreading of the virus using computational techniques through its CoronaVirus Resource Centre as shown in Figure 1.

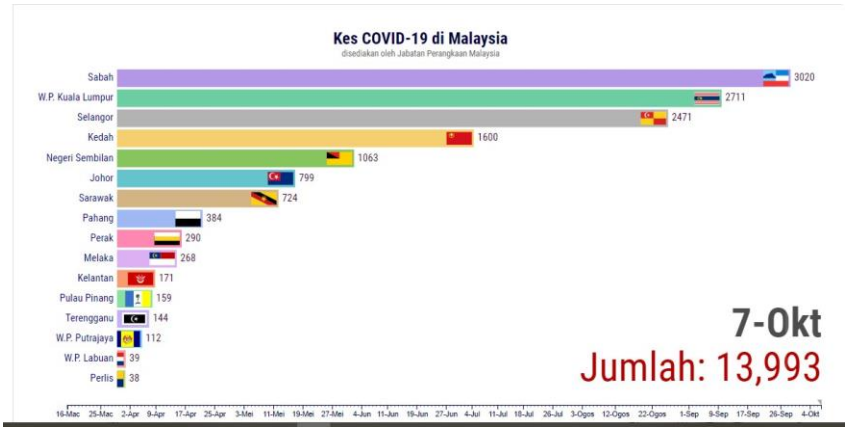
Figure 1: John Hopkins Coronavirus Resource Centre
(https://coronavirus.jhu.edu/map.html)



In terms of research, Qin et al. (2020) exploited Big Data to detect new COVID-19 suspected cases in 6–9 days and confirmed cases in 10 days whereas Yang et al. (2020) predicted the COVID-19 epidemic peaks and sizes. In Iran, Ahmadi et al. (2020) study on the correlation between climatology parameters on the COVID-19 outbreak using sensitivity analysis. He found out coronavirus highly infected patient may survive due to low values of wind speed, humidity, and solar radiation exposure. He also concluded that locations with high population density, intra- provincial movements and humidity rate are more at risk.

Besides that, the Ministry of Health and National Security Division and the Department of Statistics Malaysia also has a dedicated web page for the Corona Virus updates as shown in Figure 2.

Figure 2: COVID 19 Cases in Malaysia until 7 Oktober 2020
(https://ukkdosm.github.io/covid-19)



ESRI (https://coronavirus-nsesrimy.hub.arcgis.com/) is another website that also displays the currentCOVID 19 situation using Geographical Information System (GIS) data is as in Figure 3.

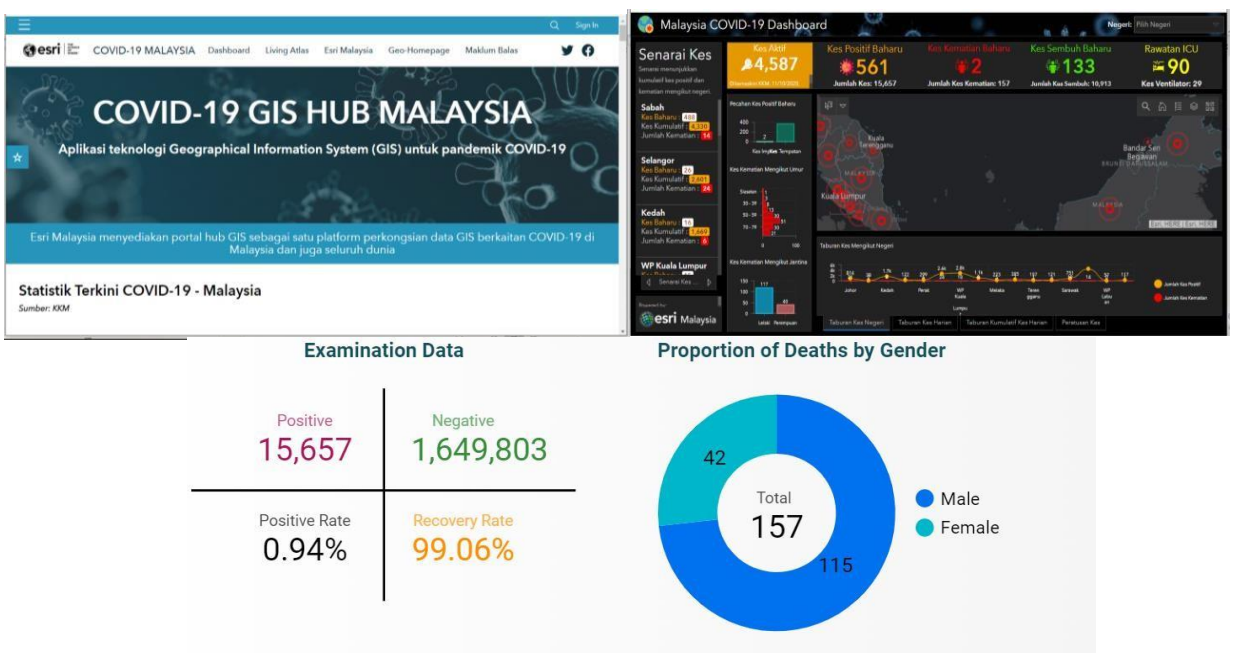


Figure 3: ESRI web site with GIS data

An interesting website (https://outbreak.my) shows the visualization of the spreading of the epidemic and patients infected with Corona Virus as depicted in Figure 4.

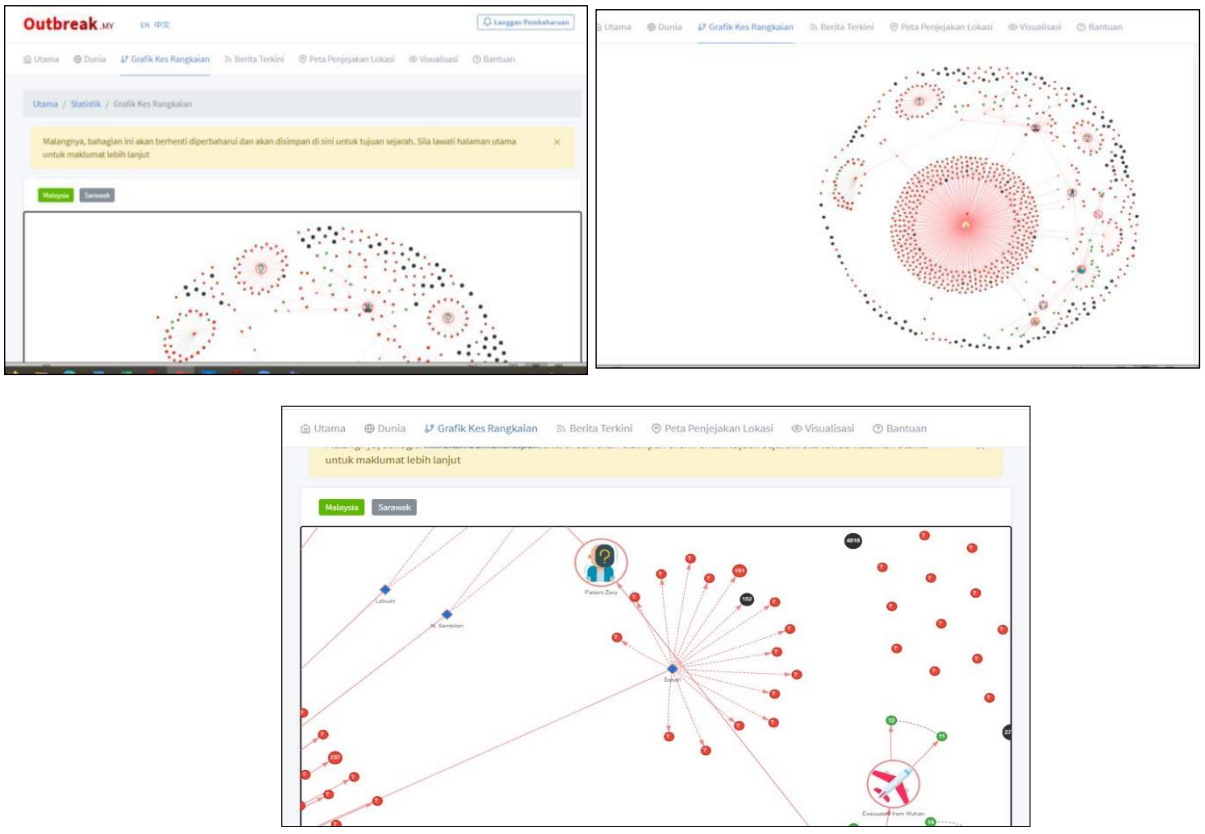


Figure 4: Visualization of COVID 19 outbreak (https://www.outbreak.my/cases)

Figure 5 shows the daily total cases in Malaysia recorded on Worldometer.

.

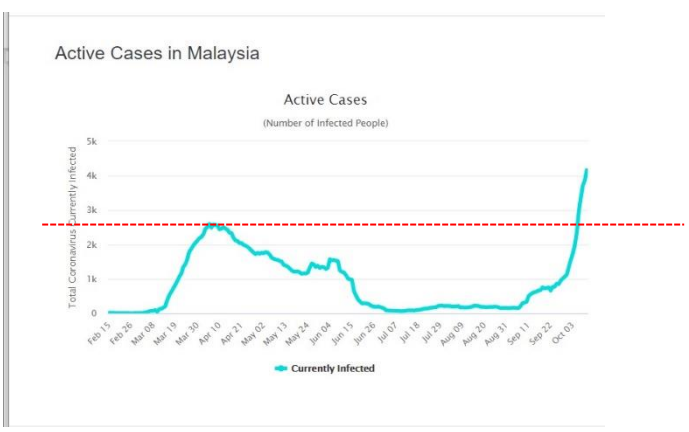


Figure 5: Active Cases in Malaysia until 11 October 2020

(https://www.worldometers.info/coronavirus/country/malaysia/)

The chart shows the second and the third wave of the coronavirus outbreak where the numbers of active cases are higher than the peak during the second wave. The MCO in the third wave focused on the locality with a red zone with a strict movement for that community observing the standard operating procedures enforcement such as wearing a face mask, keeping physical distance and home quarantine for those in close contacts with an infected person while waitingfor the swab test result as recommended by the Ministry of Health and National Security Division.

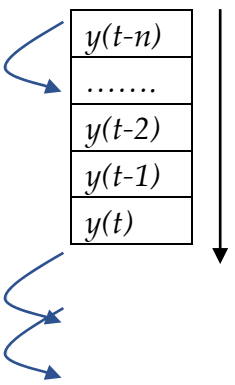
Nevertheless, to gain value from the Big Data, data analytics is very much the integral component in giving insights and forecast trend based on the outbreak pattern. Next section, the analytical part of COVID19 outbreak data in Malaysia will be presented using time series analysis and forecast the possible future trend with predictive analytics.

4.0 Methodology

Time-series data analysis is used to forecast the pandemic spreading pattern. Since the infection has a duration of a maximum of 14 days as claimed by WHO, a sliding window approach is being adopted for the data preparation to capture the range of days of infection. Then cleansed data in a sliding window format will be analysed using the Multiple Linear Regression (MLR) technique which will be compared to Artificial Neural Network (ANN) performance. This study also indicates the importance of historical data predicting future trends adopted by Norita et al. (2005) and Wan Hussain et al. (2011) through manipulationof sliding window method for time-series data to make early decision-making in a flood emergency management of a water reservoir based on daily rainfall data and reservoir water level.

4.1 Sliding Window Time Series Data Analytics and Experimentation

Time series data are usually analyzed as a single variable (univariate) varying over time where event happening at time *t* will be defined by the previous event on the time scale. It is dependent on historical data used to predict the next output on the timeline such that $y(t) = f(y(t-n))$ where n is the size of the window indicates the time frame being investigated as illustrated in Figure 6 for window size =1. The time frame may indicate the temporal delay or lagging that has useful information that must be captured.



window size = 1

Figure 6: Sliding window conceptual view for window width =1

The outcome of current time *t* is based on the output of prior time, *t-1* as in Equation 1.

$$y_t = f(y_{t-1}) \quad (1)$$

If the window width is 2 , then the equation is defined as

$$y_t = f(y_{t-1}, y_{t-2}) \quad (2)$$

Table 1a and 1b is an example of a univariate time series dataset where one variable is observe varying over time and the restructured data with window width = 1 and 2.

Table 1: Restructuring the dataset with window width=1 and width =2

1a. Data restructuring for window width=1

Time	**y_t**
1	15
2	40
3	60

Time	**y_{t-1}**	**$yt = f(y_{t-1})$ width =1**
1	?	15
2	15	40
3	40	60

1b. Data restructuring for window width=2

Time	**y_t**

Time	**y_{t-2}**	**y_{t-1}**	**$yt = f(y_{t-1}, y_{t-2})$ width =2**

1	15
2	40
3	60

1	?	15	40
2	15	40	60
3	40	60	?

Multistep forecasting can also be carried out to predict more than one future step as shown in Equation 3.

$$<y_t, y_{t-1}> = f(y_{t-1}, y_{t-2}) \quad (3)$$

Table 2 shows two steps multi forecasting with the data restructuring for the sliding window of width=1.

Table 2: Example of two steps multi forecasting

Time	**y t**
1	15
2	40
3	60
4	80

Time	**y_{t-2}**	**y_{t-1} width =1**	**y_t width =1**
1	?	15	40
2	15	40	60
3	40	60	80
4	60	80	?
5	80	?	?

A multivariate time series data can also benefit from the sliding window method.

Assume variable *a* and *b* at time *t*, where $y_t = f(a_t, b_t)$. Value of *b* at time *t* can be predicted as in Equation 4 and the example shown in Table 3.

$$y_t = b_t = f(a_{t-1}, b_{t-1}, a_t) \quad (4)$$

Table 3: Restructuring data with sliding window width = 2 for multivariate time series dataset to predict variable *b*.

Time	**a_t**	**b_t**
1	15	0.30
2	40	0.04
3	60	0.23
4	80	0.15

Time	a_{t-1}	b_{t-1}	$y_t = f(a_t, b_t)$ width = 1	
1	?	?	15	0.30
2	15	0.30	40	0.04
3	40	0.04	60	0.23
4	60	0.23	80	0.15
5	80	0.15	?	?

We may also predict more than one output variable such as in ANN, for example predicting both variables *a* and *b* at time *t* as shown in Equation 5.

$$\langle a_t, b_t \rangle = f(a_{t-1}, b_{t-1}) \quad (5)$$

Table 4 presents the dataset restructuring to predict both variables *a* and *b*.

Table 4: Restructuring data with sliding window width =1 for multivariate time series

dataset to predict variable a_t and b_t.

Time	a_t	b_t.
1	15	0.30
2	40	0.04
3	60	0.23
4	80	0.15

Time	a_t	b_t.
1	15	0.30
2	40	0.04
3	60	0.23
4	80	0.15

Time			$y_t = a_t$ width = 1	$y_t = bt.$ width = 1
1	?	a_{t-1}	$b_{t.-1}$	0.30
2	15	0.30	40	0.04
3	40	0.04	60	0.23
4	60	0.23	80	0.15
5	80	0.15	?	?

Time	a_{t-1}	b_{t-1}	a_t	$y_t = bt$ width = 1
1	?	?	15	0.30
2	15	0.30	40	0.04
3	40	0.04	60	0.23
4	60	0.23	80	0.15
5	80	0.15	?	?

Determination on the window width and forecasting steps will depend on the natureof the problem or questions asked for the analysis. The accuracy of the analysis will determine which window size gave the best performance.

In this study, two experiments on time series data analytics were conducted using (a) multiple regression technique (b) multi-layer perceptron, a feedforward artificial neural network with the restructured dataset using the sliding window approach.

5.0 Time Series Data Analytics and Forecasting On Covid 19 Outbreak In Malaysia

A data analytic begin with the primary interest of the outbreak based on the question: ***How different is the trend of the third wave outbreak compared to the second wave***?

COVID19 outbreak data in Malaysia from 25 January 2020 until 11 Oktober 2020 (261 days) is used for the data analytics and forecast the future trends with (a) polynomial estimation (b) sliding window time series forecasting with multiple regression and multi-layer perceptron (MLP) feedforward artificial neural network (ANN) architecture.

Due to the simplicity and intuitiveness of the sliding window approach, temporal sequences are recorded based on a predefined time frame and transformed into a classification problem (Norita, 2004; Al-Turaiki, 2016). Analysis and experimentation were done using MS Excel Data Analytics and Weka 3.8.3 Forecasting to generate the predictive models.

5.1 Polynomial Estimation on The General Trend on COVID-19 Cases in Malaysia

During the second wave, MCO was declared by the government with enforcement through the police and army forces at the national level. Movements were restricted to only one member of the family preferably the male to get their livelihood needs while observing the SOP. However, during the third wave, MCO is confined to a particular community within an identified red zone area. Currently, the state of Sabah and Kedah are declared with enhanced conditional MCO.

Based on the data collected. a polynomial trendline estimated the projection of cumulative cases through curve-fitting of actual data using the 5th order as shown in

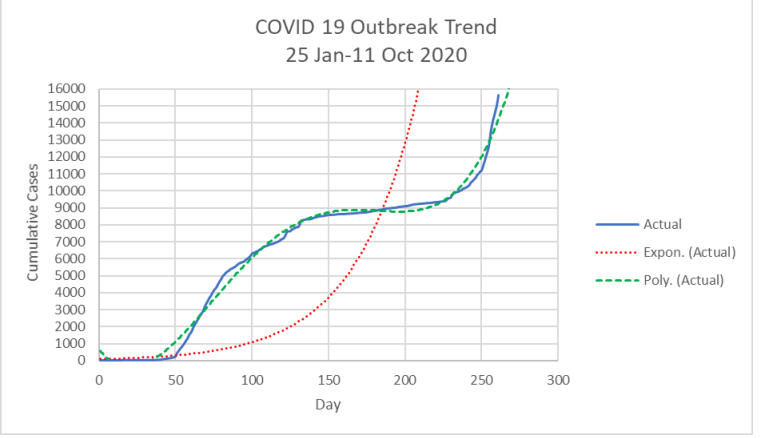


Figure 7.

Figure 7: Cumulative COVD 19 Cases Since 25 January until 11 October 2020 in Malaysia

The projection of cumulative cases may be estimated from the trendline polynomial equation generated through curve-fitting of actual data using the 5th order. The estimated trend line equation is given by Equation 6 with $R^2 = 0.991$

$$y = -8E\text{-}09x^5 + 5E\text{-}05x^4 - 0.0226x^3 + 3.488x^2 - 115.36x + 695.5 \quad (6)$$

where y and x represent total cases and days, respectively. The number of total cases by 12 October 2020, which is the in the second wave of the outbreak, i.e. day 262 according to Equation 6 is estimated to 29170 persons infected. The projection shows an escalating spreading rate for total positive cases. However, the polynomial predictive model seems to be unrealistic with an 80% increase in cumulative cases. However, it shows that the community must strictly observe the physical distancing,wearing the facemask and adhere to self-quarantine if necessary, and heightened their hygiene practices as indicated by the high-risk situation.

Figure 8 illustrates the pattern of the new cases reported and shows that spike of current new cases detected the current second wave is much higher than the maximum during the MCO1 and MCO2. The community must discipline themselves withthe SOP to break the chain as soon as possible.

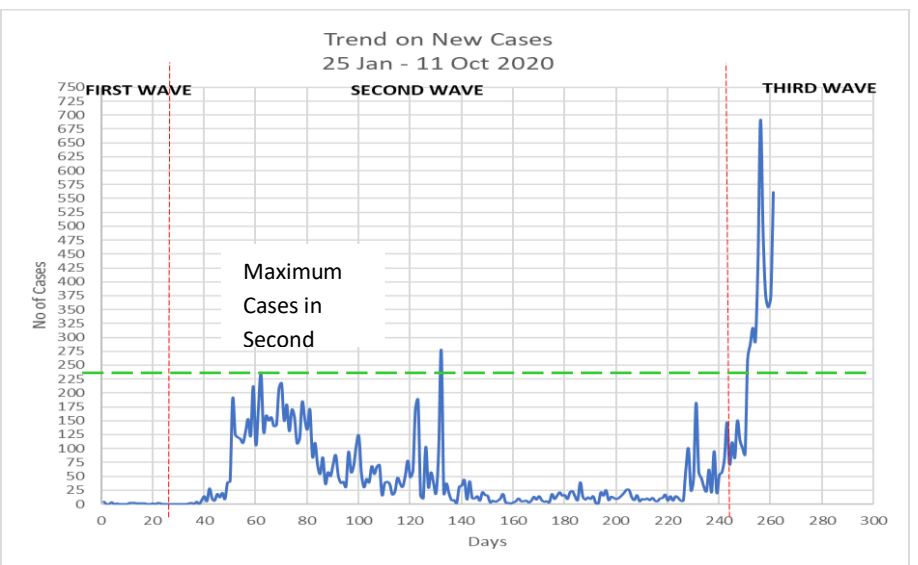


Figure 8: The Pattern of New Cases Reported.

Figure 9 shows the number of cases according to states as of 11 October 2020.

Cases by States — Oct 11, 2020

#		TOTAL		DEATHS
Sabah		4,330	↑ 488	33
Kuala Lumpur		2,754	↑ 15	18
Selangor		2,601	↑ 26	24
Kedah		1,669	↑ 16	6
Negeri Sembilan		1,069		8
Johor		814	↑ 3	21
Sarawak		751		19
Pahang		385		7
Perak		299	↑ 1	6
Melaka		273	↑ 3	5
Penang		172	↑ 1	3
Kelantan		171		3
Terengganu		157	↑ 3	1
Putrajaya		117		1
Labuan		57	↑ 5	0
Perlis		38		2
Total		15657		157

Figure 9: Number of Cumulative Positive Cases in the Various States In Malaysia (https://outbreak.my)

Next section will present the sliding window approach used for data preprocessing before training the predictive model with the restructured dataset.

5.2 Data Analysis and Experimentation

A. Data Preparation with Sliding Window Representation

In this experiment, the analysis was conducted on the first and second wave data specifically recovery cases of COVID-19 patients from 25th January until 30th April 2020 indicated by the cumulative number of infected persons being discharged from the hospital.

The purpose of the experiment is to answer

a. Is there a specific timeframe from the previous event that may influence the future outcome?

b. What is the optimum sliding window width that gives good predictive analytics performance?

c. Is it consistent with other technique?

The result of time series forecasting using multiple regression is compared to multilayer perceptron with feedforward ANN architecture. The multiple regression analysis was conducted using Microsoft Excel Data Analysis tool since the data are numerals with real values. Table 5 and Table 6 shows examples of restructured data in the data preparation phase using the sliding window technique to represents a classification problem.

Table 5: Example of Data Restructuring using a window of width 1,2 and 3.

y_{t-3}	y_{t-2}	y_{t-1}	y_t	Actual data
			0	0
		0	0	0
	0	0	1	1
0	0	1	1	1
0	1	1	1	1
1	1	1	2	2
1	1	2	3	3
1	2	3	7	7
2	3	7	8	8
3	7	8		
7	8			
8				

Table 6: Restructured Data in Sliding window format were

y_{t-i} as the predictor variables for y_t. 3 sets of data with varying window width.

Sliding Window Restructured Data								
width =1		width =2			width =3			
yt-1	yt	y t-2	yt-1	yt	yt-3	y t-2	yt-1	yt
0	0	0	0	1	0	0	1	1
0	1	0	1	1	0	1	1	1
1	1	1	1	1	1	1	1	2

1	1	1	1	2	1	1	2	3
1	2	1	2	3	1	2	3	7
2	3	2	3	7	2	3	7	8
3	7	3	7	8				
7	8							

6.0 Result and Discussion

A. Sliding Window Time Series Forecasting with Multiple Regression

The sliding window size ranging from 2 until 16 was used due to the limit of predictor variables of the MS Excel for multiple regression. Table 7 presents the performance of the multiple regression according to the window size *n* with a 95% confidence level where *n* is statistically significant with F value < 0.05.

Table 7: Performance Evaluation of the Multiple Regression Model with different window widths

Window width *n*	**Performance**	
	R Square	Observations
*** 0	0.654935	97
1	0.998887	96
2	0.999418	95
3	0.999563	94
4	0.99958	93
5	0.999643	92
6	0.999643	91
7	0.999674	90
8	0.999681	89

Window width *n*	**Performance**	
	R Square	Observations
9	0.999682	88
10	0.999685	87
11	0.999688	86
12	0.999708	85
13	0.999721	84
14	0.999757	83
15	0.999755	82
16	0.999754	81

The performance of the regression model using the sliding window perform far better than the polynomial regression, y=f(t), of window width = 0 as presented Equation 6. The experiment shows that historical time series or temporal data do influences the current outcome due to its continuity on the timeline. Thus, we may conclude that the number of discharge patients at time *t* may be explained by an event at *t-n* where *n* is the n^{th} previous time frame. The experiment result also shows that that window of width 5 and 14 gave the best performance with *p*-value at each window width is less than 0.05 and statistically significant. This result is also consistent with current practice of days if quarantine due to infection or close contact.

Thus, we may conclude that the regression model for the window of width = 5 and width =14 with significant predictor variables is given by Equation 7 and 8 where recovery r

$r_t = 1.2028 * r_{t-1} + -0.4239 * r_{t-5}$
(7)

And for window of width = 14,

$$r_t = 1.0999 * r_{t-1} - 0.5743 * r_{t-5} + 0.5353 * r_{t-7} - 0.6185 * r_{t-10} + 0.5469 * r_{t-14}$$
(8)

The predicted value for using a window with width 5 and 14 is as shown in Table 8.

Table 8: Projected value using the sliding window multiple regression model.

Date	Day	*w*=5	*w*=14
30 Apr	97	4171	4171
1 May	98	3379.953	3929.629
2 May	99	2388.145	3671.179
3 May	100	1163.322	3391.380

B. Sliding Window Time Series Forecasting with Multilayer Perceptron Using Lag

Based on the performance in the regression model, the Multilayer Perceptron (MLP) analysis was conducted with window of width = 14 using lag in Weka 3.8.3. Data from Excel was converted to CSV format and save into *.arff* extension. The results of the four experiments conducted are tabulated in Table 9.

Table 9: Result of Time Series Forecasting Using Artificial Neural Network

Multi- Layer Perceptron Architecture with Lag or Sliding Window

Data	**MSE**	**Predicted**			
		1 May	**2 May**	**3 May**	**4 May**
Original Data without lag	29.2578	4314.8850	4379.3578	4437.7216	4492.7702
Original Data with lag set to minimum 1, maximum 14	24.2193	4252.9409	4302.119	4342.7921	4377.3397
Original Data with lag set to minimum 5, maximum 14	23.4196	4279.4511	4338.5092	4393.4741	4441.0433
Restructured data with window width 14	33.4426	4245.7669	4280.4892	4312.1067	4341.2878

Based on Table 9, the best result with the lowest Mean Square Error uses the original discharge data with lag set to 5 at the minimum and 14 at maximum using the result from the regression model where the MLP Neural Network architecture is as below

- No of lag (minimum) 5
- No of lag (maximum) 14
- No of Transformed Input Data 23
- No of Hidden Nodes 12
- No of output 1
- Activation function Sigmoid
- No of instances 97

The predictive model developed is data dependent. Table 10 shows the comparison between the predicted value with actual data.

Table 10: Performance comparison of the Predicted Model of MR and MLP

Date	Day	*w*=5	*w*=14	Residual	Lag min=4 max =14	Actual Value	Residual
		MR	MR		MLP		
30 Apr	97	4171	4171		4171	4171	
1 May	98	3379.95	3929.63	-280.37	4279.45	4210	-69.45
2 May	99	2388.14	3671.18	-654.82-	4338.51	4326	12.51
3 May	100	1163.32	3391.38	-1021.62	4393.47	4413	-19.53

MR – Multiple Regression; MLP – Multilayer Perceptron

The window width =14 implies that the current outcome may be predicted using a two weeks' time frame. This agrees with the current official estimated range for the incubation of novel coronavirus COVID-19 is 2-14 days. (Worldometer, 12 March). Further, evaluation is needed to confirm whether the window width or lag size do represent the incubation period of the coronavirus. This result shows how data analytics facilitate decision-makers, health practitioners, environmentalist to understand the situation through the patterns emerging from the analysis.

7.0 Conclusion

COVID19 data outbreak has created interest in studying the trend, tracking the spread which demonstrated the capabilities of big data through visualization of the situation, data analytics with statistical, mathematical model and machine learning algorithm to investigate the behavior of the epidemic. The voluminous data will not gain its value without the analytical parts that will give insights into the situation thus facilitating decision-makers, health practitioners, environmentalist to understand the situation through the patterns emerging from the analysis.

This study also demonstrates the use of sliding window time series forecasting whichis a convenient method to present the data as a classification problem for further analysis. Thus, in the data preparation, the actual data was restructured such that it represents a temporal classification or regression problem. Multilayer Perceptron Neural Network shows superior performance compared to multiple regression in forecasting future trends of the coronavirus outbreak.

Acknowledgement

The author would like to thank the Ministry of Health, Department of Statistics Malaysia and John Hopskins University for the public access on the published data on COVID19.

References

Al-Turaiki, I., AL-Shahrani, M, Al-Mutairi, T. (2016) Building Predictive Models for MERS- CoV Infections Using Data Mining Techniques. Journal of Infection and Public Health. 9, pp 744-748.

Bragazzi, N.L.; Dai, H.; Damiani, G.; Behzadifar, M.; Martini, M.; Wu, J. (2020). How Big Data and Artificial Intelligence Can Help Better Manage the COVID-19 Pandemic. Int. J. Environ. Res. Public Health, 17, 3176.

Department of Statistics Malaysia. (DOSM) . (2020). COVID-19 by State in Malaysia. [updated 2020 oct 11; cited 2020 Apr 30] . Available from https://ukkdosm.github.io/covid-19.

Hamdan, A.R., AhmadNazree, M.Z, Abu Bakar, A. (2018). Pengenalan Sains Data in Hamdan, A.R., AhmadNazree, M.Z, Abu Bakar (eds). Sains Data Penerokaan Pengetahuan Dari Data Raya . UKM Press.

McCall, B. (2020). COVID-19 and artificial intelligence: Protecting health-care workers and curbing the spread. Lancet Digit. Health, 2, e166–e167

Ministry of Health,Malaysia(MOH). (2020). COVID-19 : Maklumat Terkini. [updated 2020 Oct 11; cited 2020 Apr 30] . Available from https://www.moh.gov.my/index.php/pages/view/2274

Norita Md Norwawi, Ku Ruhana Ku Mahamud, Safaai Deris. (2005). Recognition decision-making model using temporal data mining technique. Journal of ICT, 4 (1). pp. 37-56. ISSN 1675-414X

Norita Md Norwawi (2004). Computational Recognition-Primed Decision Model Based on Temporal Data Mining Approach in a Multiagent Environment for Reservoir Flood Control Decision. PhD. thesis, Universiti Utara Malaysia. 2004.

O'Reilly Media .(2012). Big Data Now. O'Reilly Media Inc.

Qin, L.; Sun, Q.; Wang, Y.; Wu, K.-F.; Chen, M.; Shia, B.-C.; Wu, S.-Y. (2020). Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. Int. J. Environ. Res. Public Health, 17, 2365.

Wan Hussain Wan Ishak, Ku Ruhana Ku Mahamud, Norita Md Norwawi. (2011). Mining Temporal Reservoir Data Using Sliding Window Technique. CiiT International Journal of Data Mining and Knowledge Engineering, Vol 3, No 8, 473-478

World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report. 2020.

[updated 2020 Apr 30; cited 2020 Apr 30] vailablefrom https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

Worldometer. (2020). Covid-19 Coronavirus Pandemic. [Updated 11 Oct cited on 11 Oct 2020]

Available online from https://www.Worldometers.Info/Coronavirus/#Countries.

Yang, Z.; Zeng, Z.; Wang, K.; Wong, S.S.; Liang, W.; Zanin, M.; Liu, P.; Cao, X.; Gao, Z.; Mai, Z.; et al. (2020). Modified SEIR and AI prediction of the trend of the epidemic of COVID-19 in China under public health interventions. J. Thorac., 12, 165–17