Statistical Modelling and Analysis: The study of COVID-19 in Malaysia

By

<Cheam Heng Sheong>

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE

KUALA LUMPUR

ACADEMIC YEAR

2021/2022

Statistical Modelling and Analysis: The study of COVID-19 in Malaysia

By

Cheam Heng Sheong
Supervisor: Mr. Chee Keh Niang

A project report submitted to the Faculty of Computing and Information Technology in partial fulfillment of the requirement for the **Bachelor of Science (Hons.) Management Mathematics with Computing**
Tunku Abdul Rahman University College

Department of Mathematical and Data Science
Faculty of Computing and Information Technology
Tunku Abdul Rahman University College

Kuala Lumpur
2021/2022

DECLARATION

The project report submitted herewith is a result of my own efforts in totality and in every aspects of the works. All information that has been obtained from other sources had been fully acknowledged. I understand that any plagiarism, cheating or collusion of any sorts constitutes a breach of Tunku Abdul Rahman University College rules and regulations and would be subjected to disciplinary actions.

Signature: 

Name: CHEAM HENG SHEONG

ID No: 19WMR10705

Date: 19/11/2021

APPROVAL FOR SUBMISSION

I certify that this project report entitled "**STATISTICAL MODELLING AND ANALYSIS: THE STUDY OF COVID-19 IN MALAYSIA"** was prepared by CHEAM HENG SHEONG and has met the required standard for submission in partial fulfillment of the requirements for the award of Bachelor of Science (Hons.) management Mathematics with Computing at Tunku Abdul Rahman University College.

Approved by

Signature:  ______________________

Signature: Lsg ______________________

Supervisor: Chee Keh Niang

Moderator: Lee Shu Gyan

Date: 15-12-2021

Date: 15-12-2021

ACKNOWLEDGEMENTS

The author would like to thank everyone who had contributed to the successful completion of this project. The author would like to express his gratitude to his research supervisor, Mr. Chee Keh Niang for his invaluable advice, guidance and his enormous patience throughout the development of the research. The author would also like to convey his heartfelt appreciation to Dr. Tan Yan Bin for contributing her ideas and in-depth knowledge in the field of Mathematical and Statistic Software. A warm thank is extended to Dr. Kok Wai Keong for sharing hisresources, opinions, knowledge, experience and skills in Mathematics so generously. The author would also like to personally thank his family, friends in Tunku Abdul Rahman University College and his fellow course-mates who have one way or another extended their assistance in completing this project. Last, but not least, the author wishes to acknowledge the unwavering support shown by other Mathematics lecturers.

ABSTRACT

This study is intending to work on an analysis of a COVID-19 dataset from Jan 2021 to Oct 2021 using different method of modelling to do comparison. Also, to check the relationship between each variable in order to determine the problems and the causes of cases increased. Figure 4.1 shows that out of all cluster, the most cases increased in cluster was the cluster of community followed by high risk area. This allow the government to put more effort on those cluster which have a high correlation between new cases increased. To prepare a decent decision in order to prevent this pandemic. There is different regression model had implemented in this study while the most suitable regression modelling is polynomial regression with 3 degree (cubic) with a coefficient of determination of approximate 78% after a comparison was made. Regression splines are often more accurate and reliable in forecasting. Hence, a regression splines is implemented to the previous polynomial regression for improvement of model. There are total 2 different regression splines with different random knots (3 and 4) was designed and the result was satisfying with a coefficient of determination of approximate 89% for regression splines with 3 knots and coefficient of determination of approximate 90% for regression splines with 4 knots. Selecting knots are important and difficult. Therefore, by transforming a time-series data to a complex network and do the division of the visibility graph of the dataset into connective communities to define the knots in applying into spline algorithm again by the most distant nodes within each community. The result does provide a better coefficient of determination which approximated to 91% with the lowest root mean square among other method. The overall attempt in this research is to contributes to the scientific research proposing a reliable framework of forecasting and prediction which can helps in provides insights of good policy or good decision making in order to control over the pandemic.

Keywords: Regression, Splines, COVID-19, pandemic, complex network, community detection

TABLE OF CONTENTS

DECLARATION...iii

APPROVAL FOR SUBMISSION.. iv

ACKNOWLEDGEMENTS..v

ABSTRACT.. vi

TABLE OF CONTENTS..vii

LIST OF FIGURES viii

LIST OF TABLES... ix

LIST OF APPENDIX..x

CHAPTER 1 Introduction

1.1 Understanding of COVID-19 1

1.2 Problem Caused by COVID-19 in MALAYSIA.........................2

1.3 Understanding of Statistical Modelling and Analysis 2

1.4 Problem Statement…..2

1.5 Objective...3

1.6 Project Scope…...3

CHAPTER 2 Literature Review 4-7

CHAPTER 3 Methodology

3.1 Datasets... 8

3.2 Procedures 8

3.3 Mathematics algorithms

3.3.1 Regression Analysis..9-10

3.3.2 Regression Splines…..10

3.3.3 Complex Network Analysis 11-12

CHAPTER 4 Results...13-18

CHAPTER 5 Findings and Future Work 19-21

CHAPTER 6 Conclusion…..22

REFERENCE...23-24

APPENDICES.. 25-27

LIST OF FIGURES

No	Figures	Page
1	Figure 1 : COVID-19 cases by week from 2020 to 2021	6
2	Figure 2 : Pearson's bivariate correlation	13
3	Figure 3 : p-value between variables (alpha = 0.05)	14
4	Figure 4 : Regression Analysis	15
5	Figure 5 Visualization of Cubic Regression and Regression Spline with 3 and 4 knots	16
6	Figure 6 Community detection of the number of new cases in Malaysia visibility graph	17
7	Figure 7 Visualization of Cubic Regression, Regression Spline (3 and 4 knots) and Complex network defined knots	18
8	Figure 8: Model for Regression Spline with 3 knots and 4 knots	20

LIST OF TABLES

NO	Table	Page
1	Table 1 Comparison between models in R^2 and RMSE	13
2	Table 2 Comparison between models in R^2 and RMSE [Added Regression Spline (Complex Network)]	15
3	Table 3 Comparison between Prediction and Actual Average new cases	20

LIST OF APPENDIX

NO	Table	Page
1	Dataset of Malaysia’ Covid-19 new cases from Jan 2021 to Oct 2021	25 - 27

CHAPTER 1

INTRODUCTION AND DEFINITION

1.1. Understanding of COVID-19

COVID-19, also called Coronavirus disease is a spreading disease caused by a newly discovered coronavirus which “CO” refer to corona, “VI” for virus, “D” for disease and 19 refer to the year discovered. Formally, COVID-19 is named as “2019 novel coronavirus” or “2019-nCoV” and this virus is linked to same family of viruses as Severe Acute Respiratory Syndrome (SARS) and some type of common cold [1]. The COVID-19 was reported at the beginning of December 2019 near Wuhan City, Hubei Province, China. From the phylogenetic analysis carried out with attainable full genome sequences, bats occur to be the COVID-19 virus holder, but the intermediate host(s) has not been detected till now [2]. The COVID-19 has not yet been identified in humans, where respiratory symptoms, fever, cough, shortness of breath, and loss of taste or smell are common signs of the infection of COVID-19. However, in more serious cases patientmay fall under pneumonia, severe acute respiratory syndrome, kidney failure and even death [3].

1.2. Problem Caused by COVID-19 in MALAYSIA

The COVID-19 pandemic has becoming into a health, socioeconomic and humanitarian crises of bizarre scale and impact. The Malaysia Government imposed a Movement Control Order (MCO) starting 18 March to 12 May 2020 and decided to soften the regulations under a Conditional Movement Control Order (CMCO) by 4 May to 9 June 2020. During MCO, it imposed stay-at-home orders, banned outdoor activities and shut down all of the businesses expect a few important sectors such as natural resource sectors. Although CMCO allowed most of the economic activities and public movements under strict condition but still the lockdown is turning into an economic knockout with aggravate negative impacts such as unemployed, incomes decreased and livelihoods [4]. Due to the MCO and CMCO caused by COVID-19, the country’s fiscal deficit is estimated to increase to 4% of GDP in 2020 and Malaysia's economic growth is slowed down to 4.3% in 2019. Other than that, knock-on effect of this pandemic impact on its major trading partners such as China and Singapore and falling crude prices to affect state revenue is expected [5].

1.3. Understanding of Statistical Modelling and Analysis

Statistical Modelling and Analysis is an art of collecting, exploring and presenting a large amounts of data to find the hidden patterns or trends of something which can be called as the process of applying statistical analysis to a dataset. It allows us to identify relationships between variables, making a prediction about future sets of data, visualize the data so that the collaborator and community are able to use and leverage it. One of the traditional methods for this Statistical Modelling and Analysis is from sampling data to interpreting results which is used by scientists for millennium. However nowadays data volumes make statistics ever more valuable and powerful. Most common model canbe found is Regression Models. It normally is to examine relationships between variables and to determine which independent variables have the most impact over dependent variables. The common Regression models are logistic, linear, andpolynomial regression [20].

1.4. Problem Statement

The appearance of Covid-19 was unexpected and becoming a world-wide pandemic rapidly while most of the countries unaware of the damaged it brings which caused the economics and social disruption. Policies and decision must make by government before it out of control. Hence, the factor of cases increases must be determining to prevent the spread of Covid-19. Example of factor such as sector travelling which may bring virus carrier, no proper quarantine for patient and etc. Also, further action and policy for future must make based on the situation of pandemic. Hence,forecasting the new cases to make the best decision is critical to fight against thispandemic.

1.5. Objective

To fight against this pandemic, the development of more accurate and reliable models in terms of description and prediction can help the government in better conceptualize the situation and apply proper and more effective policies. Therefore, the objective of this research is to predict the expected number of infection cases based on the historical daily number of infection cases recorded from the past. Moreover, the prediction may help as an indicator for the government to allocate the budget to each important sectors such as health care facilities and imports vaccine from other country tocontrol the pandemic over time. Also, the objective of this research is to determine the factor of rises cases by analysis the trend of a time series data. In order to control over something, factors and causes are important to be determined before further action were taken. This research is intending to work on extension of a time series based data while the study of other area on complex network.

1.6. Project Scope

In this project, the study of regression modelling especially on regression spline is included to forecast and determine which model best fit. In additional, the complex network of transforming a time series dataset into a visibility graph for community detection purposesis studied to help in determine the knots of the regression spline model to increase the accuracy. Since there is different stage of pandemic happened in Malaysia starting from Jan 2020, dataset timeline chosen was from Jan 2021 to Oct 2021 because this is the period where new cases getting fluctuated and series. The main statistical software for modelling is Python with the assist of Microsoft Excel.

CHAPTER 2

LITERATURE REVIEW

As the COVID-19 epidemic continuous, everyone hopes to contribute in solving this pandemic. Hence, the mathematician plays an important role in helping during this pandemic. They share their point of views on that models reveal about how the disease has spread and the current state of play also what work still needs to be done. The information shared by those mathematicians constructed a detailed presentation of COVID-19 instead of the analysis of data. This contributed on the decision-making for public health with sharp primary information which help COVID-19 research as an archive of statistical data [6].

In 2020, Samuel Asumadu Sarkodie and Phebe Asantewaa Owusu done an investigation on the cases of COVID-19 by using 7 models comprising of five panel data setting and two time-series [7]. They discovered there is a perfectly linear relationship between deaths and confirmed cases while the impact of confirmed cases on recovery cases is a nonlinear relationship. They detect the relationship by using an insightful time series analysis based on four health indicators of COVID-19 in China. This article contributes to the literature with an interesting case study. However, this research is restricted to the case of China and other factors such as limited number of indicators used that does not help in a decent forecasting.

Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris and Constantinos Siettos collaborated a research [8] on COVID-19 and attempted to estimate the mean values of the main epidemiological parameters with 90% confidence intervals for a small size of data and adjusted the parameters of the *Susceptible-Infectious-Recovered-Deceased* (SIRD) model to fit the reported data. They did a reverse forecasting process that builds on spreading scenarios but this is not a reliable basis for forecasting.

Simon Hames Fong, Gloria Li, Nilanjan Dey, Ruben Gonzalez Crespo and Enrique Herrera-Viedma published a journal in 2020 about Finding an Accurate Early Forecasting Model from Small Dataset [9]. They suggested five methods for machine learning model to compare the outcome with polynomial neural network with corrective feedback (PNN + cf) for forecasting. The PNN method have a satisfactorily accuracy even in a cases of small size of data availability. However, they only considered to compare the accuracy criterion and stopped the research once their objective is obtained. As this interesting approach should compare with other established algorithms of similar good accuracy.

Many researchers mostly were focusing on their own country's case-study of COVID-19 rather than the global case as the diversity that the phenomenon has in different country. For example, Asita Elengoe [10] did a research on the changes in statistical data of Malaysia after the application of MCO and CMCO. As the hypothesis testing is important in statistics, analytics and data science, Diego Giuliani, Maria Michela Dickson, Giuseppe Espa and Flavio Santi's derived an analysis builds on statistical modelling but without testing the statistical significance of the research hypothesis [13] which very fatal. It is because in epidemiologic studies, the goal is to develop an instance function, quantifying a cause and effect relation between a determinant and its result. Therefore, to test if the cause and effect relation is statistically significant is our major concern [11, 12].

Machine learning models is a powerful tool in forecasting. In 2020, R. Sujath, Jyotir Moy Chatterjee, and Aboul Ella Hassanien cooperated on different machine learning model in forecasting COVID-19 pandemic in India [18]. They performed linear regression, Multilayer perceptron and Vector auto-regression method where resulting the best prediction among these methods was Multilayer perceptron based on the comparison between the prediction value and the actual value of infected cases. However, it stated that the Vector auto- regression model is the more suitable analysis model in this multivariate time series data which helps in inferencing and analysis of policy.

An article related to complex network analysis is posted on 2021 by Sha Zhu, Meng Kou, Fujun Lai, Qingxiang Feng and Guorong Du which studies on the connectedness of the COVID-19 pandemic in the world [19]. By applied its characteristics, they obtained acomplex network without considering the direction and it is an undirected network with 122 nodes and 958 connected edges. Using this result, they generated a visual graph which showsthat the connectedness between country. One of its result is the China has connectedness with Singapore, Japan, and Thailand which has good control over the pandemic. They also further analyse some characteristics of the pandemic network and resulting the countries number of degrees and edges. This may help in decision making in policy for those countries who has high degree and edges since the higher the degree and edges, the higher the chance of the pandemic getting worse in those connectedness between country.

The dataset of COVID-19 is a Time Series based data where it is a series of data point indexed in a time order. Many traditional models require the time series to be stationary where it refers to some statistical properties such as mean, variance and serial correlation are constant over time. It makes the analysis more straightforward but modern approaches makeit possible to work with data without pre-processing for stationarity. There are components which can be decomposed by time series which is the trend (long tern direction), the seasonal(systematic), and the residual (unsystematic) [17].

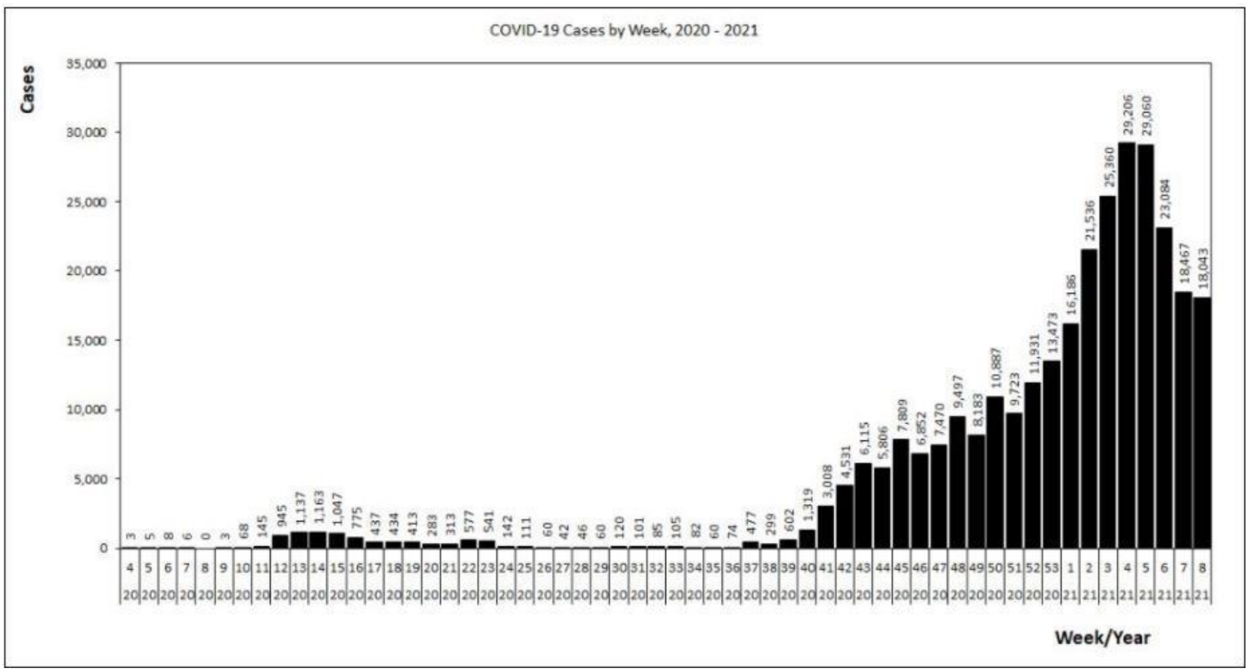

Figure 1

COVID-19 cases by week from 2020 to 2021

[obtained from Kementerian Kesihatan Malaysia [15]]

Figure 1 above shows that the trend of this COVID-19 pandemic has an obvious upward trend over time since week 39 of 2020 and there is no systematic pattern of fluctuation and changing over time irregularly.

A book titled “Forecasting principles and practice” by Rob J Hyndman and George Athanasopoulos stated that time series based Auto-regressive integrated moving average (ARIMA) model and Exponential smoothing model are the two most common used approaches for time series forecasting. While in this case of dataset, ARIMA model is preferred due to it able to describe the autocorrelations in the data but exponential smoothing models are based on a description of the trend and seasonality in the data. The biggest advantage of ARIMA model is that it can be applied in cases where the data shows evidence of non-stationarity. [21]

CHAPTER 3

METHODOLOGY

3.1 Datasets

The datasets included the first case happened in Malaysia until now were extracted from multiple resources but mainly from the official website of Kementerian Kesihatan Malaysia. The other resources included:

- Official website of Malaysiakini [16], Official website of kpkesihatan [15],
- Official website of OurWorldinData [14]

3.2 Procedures

Firstly, the COVID-19 variables (x_i) to participating in the analysis must be decided. Each variable is a time-series x_n = {x_i| i = 1, 2, . . . , n} = {x_1, x_2, . . . , x_n} , where each node i = 1, 2, . . . , n refers to a day. For example, $x_1 = date, x_2 = new\ cases$, . . . Then, to tests the structural dynamics of one variable in comparison with otheravailable variables, Pearson's bivariate correlation analysis is applied to set of the available variables and investigate the outcome. After that, we examines the time-seriespattern configured for a variable x_i by regression analysis. To improve the accuracy and determination ability of the fitting, we apply next a regression analysis based on the regression splines algorithm. Finally, transforming a time-series to a complex network followed by the five step below:

Step 1: Transform time series to visibility graph.
Step 2: Calculate the node similarity
Step 3: Find out the most similar node
Step 4: Determine node distance
Step 5: Make predictions

3.3 Mathematics algorithms

3.3.1 Regression Analysis

The available types of fitting curves tested in the regression analysis are linear, quadratic, cubic, power and logarithmic. All the available types of fitting curves can be generally described by the general multivariate linear regression model expressed by the formula:

$$\hat{y} = b_1x_1 + b_2x_2 + \cdots + b_nx_n + c = \sum b_ix_i + c \quad (1)$$

by considering that each independent variable x_i can represent a function of x, let $x_i = f(x)$, as it is:

$$\hat{y} = \sum b_if(x_i) + c \quad (2)$$

The function $f(x)$ can be either logarithmic $f(x) = (\log(x))^m$, polynomial $f(x) = x^m$, exponential $f(x) = (\exp\{x\})^m$ or any other. Within this context, the purpose of the regression analysis is to test the parameters b_1 of Model (2) that best fit the observed data y, so that able to minimize the square differences of $y_i - \hat{y}$:

$$\min\{e = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} [y_i - (\sum b_if(x_i) + c)]^2\} \quad (3)$$

the algorithm test on the beta coefficients (b_i) by using least-squared linear regression method based on the assumption that the differences e in (3) follow by a normal distribution $N(0,\sigma_e^2)$. If x_1 is selected as the independent variable, each and other available variable can be selected as the dependent variables to the models. In all cases, the simplest form of regression models that best fits the data is chosen. The simplicity criterion regards both the number of the usedterms $b_if(x_i)$ and the polynomial degree m. That is, the model with the least possible terms and the lowest possible degree $m < n - 2$ (where n is the number of observations in the dataset) is chosen if it best fits the data. The determination ability of each model is expressed by the coefficient of determination R^2, which is given by the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (4)$$

Where y_i are the observed values of the dependent variable, $\hat{y}$ are the estimated values of the dependent variable, $\bar{y}$ is the average of the observed values of the dependent variables, and n is the number of observations (the length of the variables). Next measure of fitting ability is the root mean square deviation or error (RMSE), which calculates the square root of the expected differences between the predicted ($\hat{y}$) and the observed (y) values of the model, A final measure of fitting ability used in the analysis is the relative absolute error (RAE), which calculates the relative value of the RMSE in accordance with the expected observed values, as is shown by the formula:

$$RMSE = \frac{\sqrt{(y_i - \hat{y})^2}}{\sqrt{E((y)^2)}} \tag{5}$$

where E() is the function of the expected value. The RAE is often used in machine learning, data mining, and operations management applications, and it represents the analogy of the RMSE relative to the expected value of the observed values.

3.3.2 Regression Splines

The regression spline is a special piecewise polynomial function defined in parts, which is widely used in interpolation problems requiring smoothing. For a given partition $a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = b$ of the interval $[a, b]$. A spline is a multi-polynomial function S(t) denote by the union of functions:

$$S(t) = S_1([a, t_1]) \cup S_2([t_1, t_2]) \cup \ldots \cup S_{n-1}([t_{n-2}, t_{n-1}]) \cup S_n([t_{n-1}, b]) = \cup_{i=1}^{n} S_i([t_{i-1}, t_i]) \tag{6}$$

where n is the number of knots t = (t_0, t_1, t_2, ... , t_{n-1}, t_n) dividing the interval $[a, b]$ into k-1 convex subintervals. Each function $S_i(t)$, $i = 1, 2, \ldots, n$ is a polynomial of low (usually square) degree (sometimes can also be linear) that fits to the corresponding interval $[t_{i-1}, t_i]$, $i = 1, 2, \ldots, n$, so that the aggregate spline function is continuous and smooth.

3.3.3 Complex Network Analysis

i. Visibility Graph

It is an algorithm that change a time series into a network. It may defined as two time series data (t_1, y_1) and (t_2, y_2) are considered as two visible and connected nodes in a network. If any data value (t_3, y_3) between these two data, fulfils the algorithm below:

$$y_3 < \frac{(t_2-t_3)}{(t_2-t_1)}(y_1 - y_2) + y_2 \qquad (7)$$

ii. Link Prediction

In a network, the probability that a random walker departing from node x is denoted by $\vec{\pi}_x$. In t steps (t is a large enough value to ensure the walker can fully walk in the network), $\vec{\pi}_x(t)$ meets the following requirement:

$$\vec{\pi}_x(t) = P^T \vec{\pi}_x(t-1) \qquad (8)$$

It should point out that P^T is the transpose of transition probability matrix P. The probability of moving from node x to node y in one step is denoted by $\vec{\pi}_{xy}$. $P_{xy} = \frac{a_{xy}}{k_x}$, where $a_{xy} = 1$ if point x is linked to point y, otherwise $a_{xy} = 0$. Besides, k_x means the degree of point x. $\vec{\pi}_x(0)$ is an N × 1 vector indicating the initial state of node x, where the $x\ th$ element is equal to 1 and the others are 0. The definition of the similarity between node x and node y is followed by:

$$S_{xy}^{LRW}(t) = \frac{k_x}{2|E|}\pi_{xy}(t) + \frac{k_y}{2|E|}\pi_{yx}(t) \qquad (9)$$

Where LRW refer to local random walk, $|E|$ refer to the total number of edges in the network. After that, the higher similarity between node x and node y can be obtained by adding up the results of S_{xy}^{LRW} in each step. The result is calculated by the algorithm below after a superposed random walk (SRW):

$$S_{xy}^{SRW}(t) = \sum_{l=1}^{t} S_{xy}^{LRW}(l) \qquad (10)$$

iii. Node Distance

The node distance in a visibility graph is defined by following algorithm:

$$d_{i \to j} = |t_i - t_j| \tag{11}$$

Where t_i and t_j are the corresponding time values in point (t_i, y_i) and node (t_j, y_j) respectively.

CHAPTER 4

RESULT

To tests the structural dynamics of one variable in comparison with other available variables, Pearson’s bivariate correlation analysis is applied to set of the available variables and investigate the outcome as shown in Figure 2.

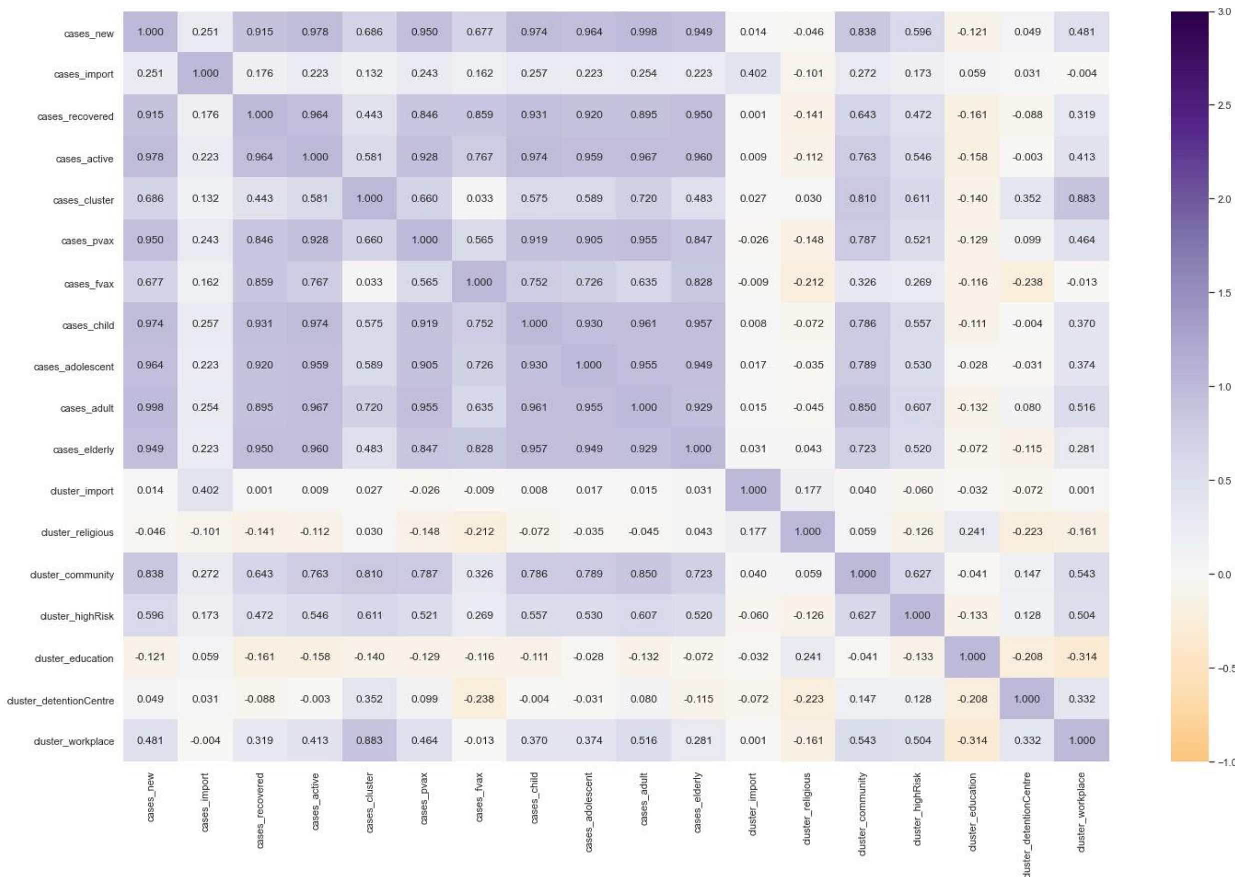

Figure 2 Pearson’s bivariate correlation

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	cases_new	c_import	c_recovered	c_active	c_cluster	c_pvax	c_fvax	c_child	c_adolescent	c_adult	c_elderly	clus_import	clus_religious	clus_community	clus_highRisk	clus_education	clus_detentionCentre	clus_workplace
1																		
2	3.3265E-66																	
3	0.3223576	3.3E-63																
4	2.6143E-56	2.3E-63	1.6187E-56															
5	9.9699E-52	8.8E-89	9.6143E-49	3.39E-62														
6	1.7259E-51	2.7E-23	2.0071E-48	3.08E-62	0.279111													
7	6.7411E-46	1.9E-30	6.5664E-43	6.83E-62	0.004423	0.006												
8	1.3272E-54	9.4E-56	1.3532E-51	1.98E-62	0.000135	0.0693	6E-06											
9	1.7905E-60	3E-51	1.8266E-57	6.67E-63	1.06E-39	3E-09	2E-16	1E-16										
10	4.7458E-09	8E-68	1.0174E-07	1.32E-58	6.84E-46	8E-44	3E-35	4E-50	2.99632E-59									
11	8.661E-59	9.1E-70	8.8407E-56	9.22E-63	3.67E-27	2E-06	7E-13	4E-09	8.92609E-05	1E-56								
12	2.5196E-66	2.1E-59	2.5352E-63	2.21E-63	1.22E-89	1E-23	9E-31	1E-56	8.9051E-53	5E-68	3.34E-71							
13	4.8268E-66	1.2E-05	4.8569E-63	2.33E-63	9.57E-88	7E-23	5E-30	1E-54	3.58902E-49	1E-67	6.49E-68	1.3357E-13						
14	6.9413E-64	1.9E-59	7.0535E-61	3.53E-63	7.2E-72	5E-17	2E-24	1E-39	2.63158E-21	3E-64	1.27E-39	1.8784E-63	4.71671E-53					
15	5.2626E-66	5.8E-20	5.2969E-63	2.35E-63	2.02E-87	1E-22	6E-30	2E-54	9.12975E-49	2E-67	2.09E-67	1.8769E-43	0.166669236	1.63787E-52				
16	6.6141E-66	9.3E-12	6.6558E-63	2.39E-63	8.35E-87	2E-22	1E-29	9E-54	1.99256E-47	2E-67	2.62E-66	1.1605E-20	0.007401545	7.68042E-48	0.019877374			
17	1.4164E-65	3E-31	1.4256E-62	2.55E-63	1.56E-84	1E-21	8E-29	2E-51	3.35217E-43	7E-67	2.88E-62	5.4803E-41	2.96479E-14	7.21288E-36	1.95812E-15	1.66611E-07		
18	1.4528E-58	4.7E-79	1.4704E-55	9.58E-63	1.36E-26	5E-06	2E-12	1E-08	6.70093E-06	2E-56	0.320045	1.4612E-80	2.74467E-77	6.9274E-47	1.14857E-76	1.34934E-75	2.23035E-71	

Figure 3 p-value between variables (alpha = 0.05)

As Figure 2 shown, the number of new cases is significantly correlated with all the other variables except for the cluster of education and religious. The cluster of community had the highest correlation with new cases among the other cluster where we can assume that new cases mostly come from the cluster of community. This correlation illustrates the analogy in cluster of community and new cases increased which allows government to handling this particular cluster seriously. Also, by observing on Figure 2 we can know that the ages that have the highest chance infected is at the age of adult as the number of new cases increased. Figure 2 also shows that the number of cases recovered is significantly correlated with the number of active cases which expressing a tendency of the Malaysia health system to get more patients recovered when more number of cases being active. In short, this Figure 2 provide a different vision of this pandemic in Malaysia are dominated by non-randomness.

Other than that, Figure 3 was designed to test the hypothesis of whether there is relationship between variables. If the p-value calculated was greater than the alpha value, we determine that the two selected variables had no relationship. Figure above explained that since most of the elderly had already retired, therefore they have no relationship with the cluster of working place. Not only that, but it also shows that although cases of patients recovered are high (since it highly correlated with active cases) but it still would not stop the new cases increasing since they have no relationship.

Hence, it shows indication that the growth of the Malaysia's pandemic system is driven by short-term linear trends. Therefore, to improve the overall system's determination and a provide a better conceptual on the pandemic, a deeper analysis is applied using regression analysis.

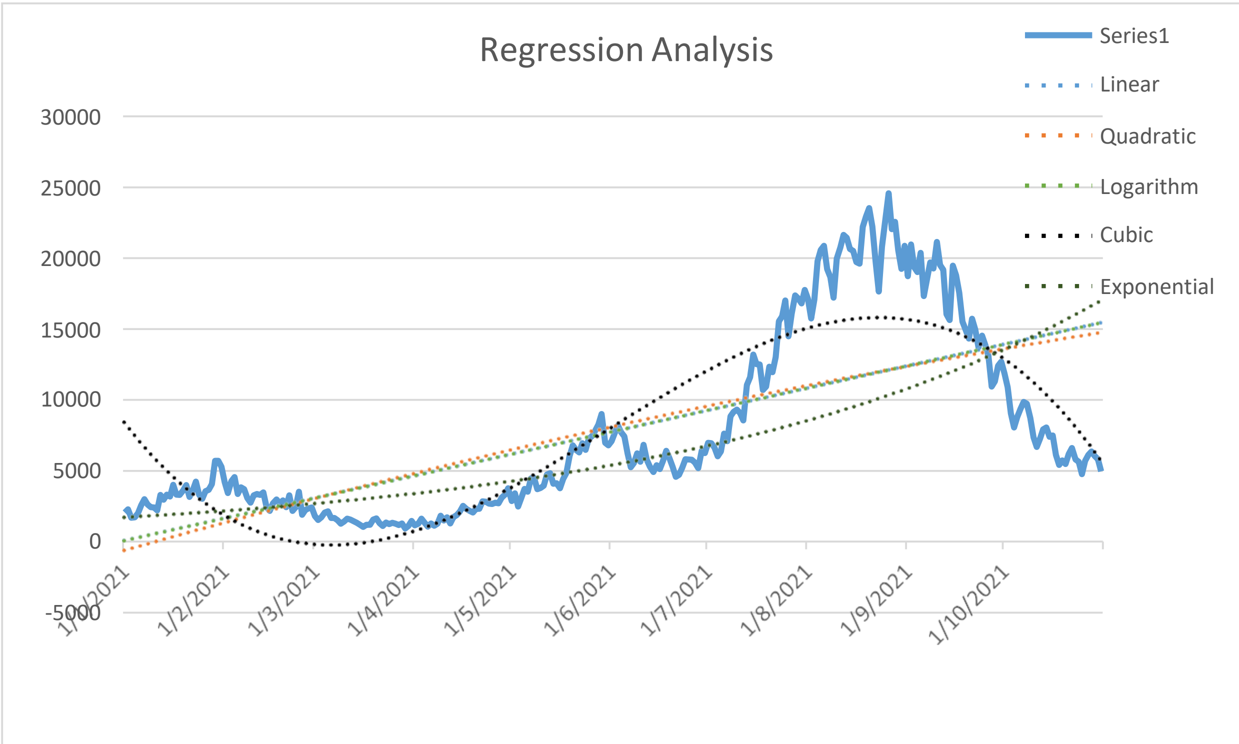

Figure 4 Regression Analysis

In the Figure 4, the date is selected to be the independent variable (x-axis) and the number of new cases as the dependent variable (y-axis). By using the build-in function in excel, multiple regression model had been calculated and compared. As it can be observed from the Figure 4 which showing the results of the regression analysis applied to specified range of dataset, the cubic fitting curve had the best describes of the data of Malaysia's COVID-19 cases. Hence, the best fit for the dataset is Cubic Regression. Although a higher order polynomial degree providesa higher R^2 but it tends to be overfitting, therefore we normally would not go up more than 3rd order polynomial degree. Therefore, to increase the accuracy of the fitting, a regression splines algorith is applied.

The regression spline model can generate a fiitings of low variance and also low bias which can minimizes the expected loss expressed by the sum of square bias, variance and noise. The most important step in applying regression spline is the determination of the knot vector. By using Python, the independent variable was designed as Month (Jan – Oct) in year 2021 only and dependent variable was the avearge number of new cases for each months. Since the knots selection is a major problem with will affect the result, therefore the knots was selected based on the quantiles of independent variable and also by trial and error in order to obtain a better result. The knots selected for Regression spline with 3 knots is (3, 6, 8) and for Regression spline with 4 knots is (3, 6, 7.25, 8). Other than that, the selection of the order of polynomial degree could smoothing the model and allows the model to be continous at the borders of the subintervals. Based on the Figure 3, the order of polynomial degree for the regression spline will be the 3rd order polynomial degree (Cubic).

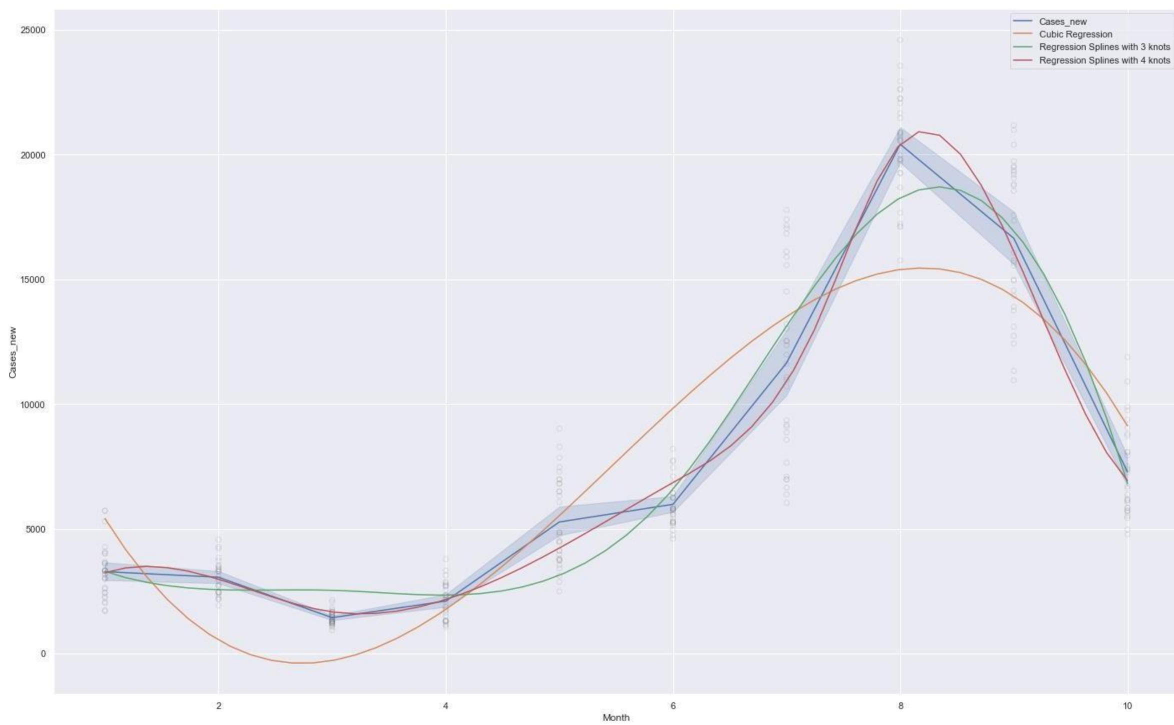

Figure 5 Visualization of Cubic Regression and Regression Spline with 3 and 4 knots

Model	R^2	RMSE
Cubic Regression	0.7780758233400923	3145.25654427303
Regression Spline with 3 knots	0.8893187160564521	2221.2163430840537
Regression Spline with 4 knots	0.9019501697492794	2090.6300526575665

Table 1 Comparison between models in R^2 and RMSE

Based on the Figure 5, we can observed that the the accuracy had increased quite a number when applied spline algorithm into cubic regression. The Regression Spline with 4 knots resulting a better determination ability and lower error terms than Cubic Regression and Regression Spline with 3 knots. From the Table 1, the improvements made for the model determination, R^2 from Cubic Regression to Regression Spline with 3 knots is approximate 11% which indicate that adding knots could increase the accuracy of the model. After the number of knots increased from

3 to 4, the model determination, R^2 had slightly increased by approximately 1%. The root mean square error (RMSE) was decreased as the knots added into the cubic regression which also proved that the model is being improved.

To further improve the model, a complex network analysis theory had suggested in this paper by transforming the dataset to a visibility graph to allow us to investigate the time series dataset as a complex network. It helps in the division of the visibility graph of the dataset into a connective communities. Hence, it allow the most distant nodes within each communities can define as the knots for us to apply in the regression spline algorithm.

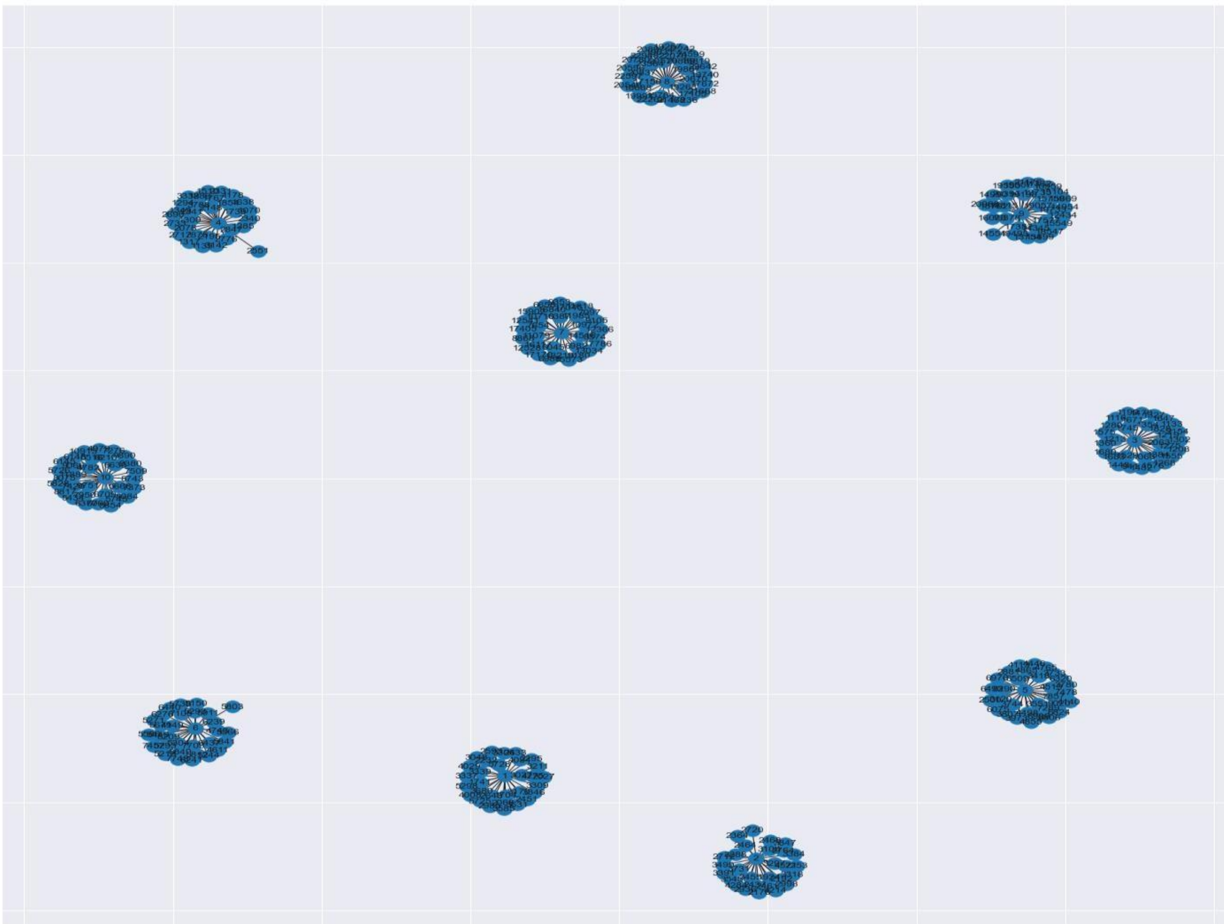

Figure 6 Community detection of the number of new cases in Malaysia visibility graph

With this approach, the visibility graph of the dataset is divided into ten modularity-based communities while the center of each network represented the Month in 2021 which noded to new cases. As the Figure 6 shows, we can let the spline knot vector to be defined as knots = (1,2, 3, 4, 5, 6, 7, 8, 9, 10). The knots obtained from the approach above will be facilitates in the useof spline regression algorithm to compare to those models previously discussed.

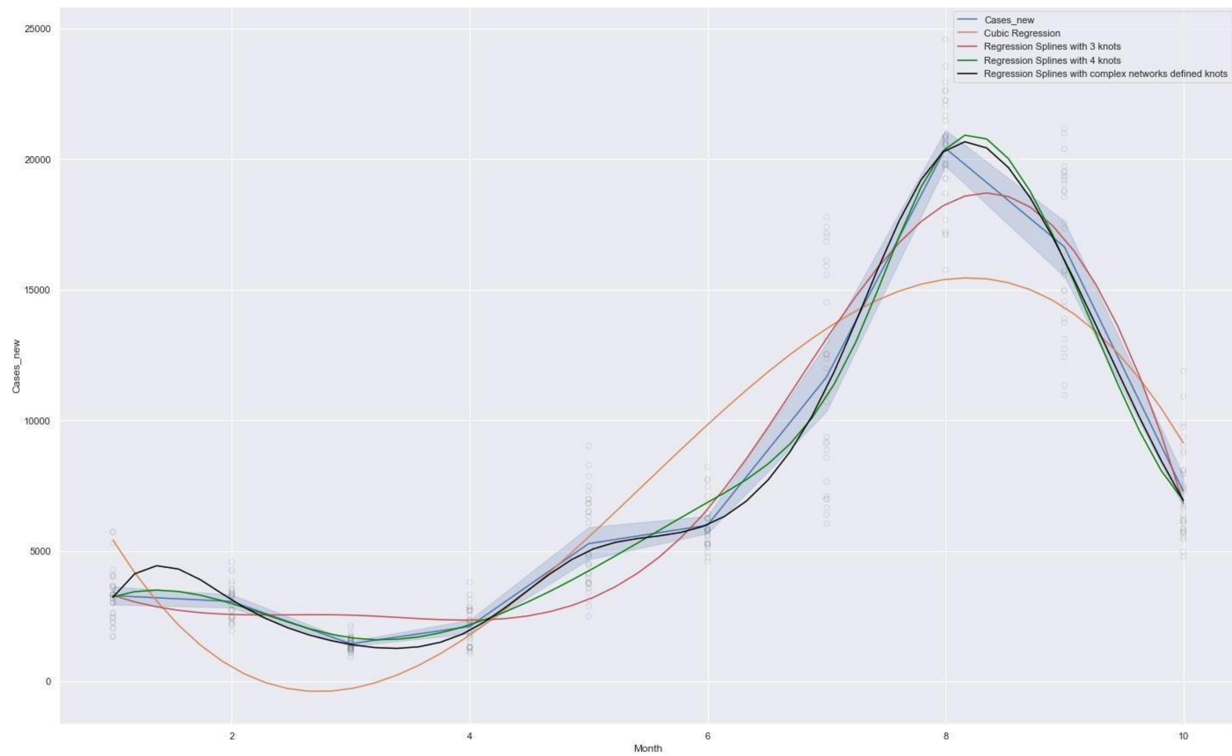

Figure 7 Visualization of Cubic Regression, Regression Spline (3 and 4 knots) and Complex network defined knots

Model	R^2	RMSE
Cubic Regression	0.7780758233400923	3145.25654427303
Regression Spline with 3 knots	0.8893187160564521	2221.2163430840537
Regression Spline with 4 knots	0.9019501697492794	2090.6300526575665
Regression Spline (Complex Network)	0.9101380765381512	2001.4356032889211

Table 2 Comparison between models in R^2 and RMSE [Added Regression Spline (Complex Network)]

From the observation on Figure 7, the Regression Spline (Complex Network) having a good result as the fitting determination and RMSE are also outperforms the other model which are alsoa models of high accuracy. The result proved that the added value of the complex network algorithm and spline algorithm.

CHAPTER 5

FINDINGS AND FUTURE WORK

5.1 Findings

Based on the Figure 2 Pearson's bivariate correlation, we can know that the decision of government for conducting online classes for education and the SOP rules made are able to avoid an incensement of new cases with a negative correlation of 12.1% between new cases and cluster of education. Other than that, the government should pay more attention on the cluster of community which contributed the most cases among other cluster with a high positive correlation of 83.8%. Also, we can know that the Malaysia Health Center are able to handle this pandemic. This finding is based onalthough there are many cases being active, but also numbers of patients recovered every day with a correlation of 96.4%. However, there are no relation between recovered cases and new cases since the p-value is 0.322 which greater than the alpha value (0.05).

In this research 3^{rd} polynomial regression was the best degree compared to other degree and model such as Linear Regression and Exponential Regression with a R^2 of 78% and RMSE of 3145.2565. The model of this Cubic Regression is:

$$\text{Model}: -194.4(x^3) + 3187(x^2) - 1.307(10^4)(x) + 1.549(10^4)$$

With the help of equation above, we are able to produce a table of prediction and comparison to real data for examination of accuracy.

Month	Predict	Actual	Different
1	5412.6	3288.677419	- 2123.922581
2	542.8	3064.035714	2521.235714
3	- 285.8	1443.483871	1729.283871
4	1760.4	2107.1	346.7
5	5515	5278.83871	- 236.1612903

6	9811.6	5987.4	- 3824.2
7	13483.8	11654.6129	- 1829.187097
8	15365.2	20418.77419	5053.574194
9	14289.4	16648.03333	2358.633333
10	9090	7288.612903	- 1801.387097
11	- 5447.455	1399.4	- 6846.86

Table 3 Comparison between Prediction and Actual Average new cases

A prediction for November 2021 was made using Cubic Regression Model. Based on the Table 3, we can have observed that the average new cases for actual data have a dramatically decreased from October to November and the predictive model does predict the trend of decreasing. However, it may not be very accurate since the number between real cases and predicted cases have a gap of 6846 cases. Hence, using a regression spline model of 3 and 4 knots, we can get a model of :

```
In [17]: fit1.params #(Model for Regression spline with 3 knots)

Out[17]: Intercept                                                        3298.135403
         bs(train, knots=(3, 6, 8), degree=3, include_intercept=False)[0]   -1031.345600
         bs(train, knots=(3, 6, 8), degree=3, include_intercept=False)[1]    -123.936576
         bs(train, knots=(3, 6, 8), degree=3, include_intercept=False)[2]   -2802.434012
         bs(train, knots=(3, 6, 8), degree=3, include_intercept=False)[3]   19718.638707
         bs(train, knots=(3, 6, 8), degree=3, include_intercept=False)[4]   13605.367243
         bs(train, knots=(3, 6, 8), degree=3, include_intercept=False)[5]    3539.194374
         dtype: float64

In [18]: fit2.params #(Model for Regression spline with 4 knots)

Out[18]: Intercept                                                              3243.346064
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[0]    984.366441
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[1]  -4355.543874
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[2]   2397.694754
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[3]   6209.774482
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[4]  25642.346823
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[5]   6903.407599
         bs(train, knots=(3, 6, 7.25, 8), degree=3, include_intercept=False)[6]   3692.019552
         dtype: float64
```

Figure 8: Model for Regression Spline with 3 knots and 4 knots

By doing the same thing as Table 3 using the model of Figure 8, we can know that the Regression Spline with 4 knots provide a better result with R^2 of 90% and RMSE of 2090.6301.

5.2 Future Work

Based on the final result, complex network defined knots for regression splines does increased the accuracy and decrease the error. However, compared to the changes from cubic regression to regression splines, it does not have a significant improvement. Hence, a further study of complex network may be proceeding in future to strengthen the model of this research. Also, there are not much similar article or journal using Malaysia' dataset for us to test and compare the accuracy of our model. Therefore, we will be using this model but different countries dataset for comparison since this topic had been done and tested by other countries' researchers. Lastly, a comparison between this model and machine learning model such as Naïve' Bayes and K-nearest neighbor (KNN) will be tested in the future to determine which model best fit the dataset (time series based dataset).

CHAPTER 6

CONCLUSION

The purpose of this research is to find the dynamics in the structure of the new cases data in Malaysia that are not visible by using the regular analysis methods on time series data. This research also meant to proposes a method to modeling and forecasting in epidemiology based on complex network analysis and regression spline algorithm. Among all the regression method tried in this research, cubic regression was the most suitable degree for this paper. However, adding knots to transform it into cubic spline could increase the accuracy of forecasting by given a smoother and smaller error interpolating polynomial.

Other than that, converting a time series data to a visibility graph followed by dividing it into connected communities that defined the spline knot vector also assigned a conceptualization to the knots based on the network connectivity. This approach advanced the regression spline algorithm which the selection of knots was still facing difficulties.

Since the accuracy in forecasting is an important task in epidemiology especially during this global pandemic. Hence models which able to contribute a more accurate forecasting is very important for the ongoing fight against the pandemic.

References

[1] Lisa Bender (Education UNICEF NYHQ). [2020] *Key Messages and Actions for COVID-19 Prevention and Control in School.*

[2] Fong, S., Dey, N., & Chaki, J. [2020]. An Introduction to COVID-19. *Artificial Intelligence for Coronavirus Outbreak*.

[3] World Health Organization, "Coronavirus,"

[4] Lin Lean, L. The socioeconomic impacts of COVID-19 in Malaysia: Policy review and guidance for protecting the most vulnerable and supporting enterprises. 2020

[5] Duddu, P. Coronavirus: Malaysia Covid-19 updates, cases, measures and impact. 2020

[6] Vespignani, A.; Tian, H.; Dye, C.; Lloyd-Smith, J.O.; Eggo, R.M.; Shrestha, M.; Scarpino, S.V.; Gutierrez, B.; Kraemer, M.U.G.; Wu, J.; et al. Modelling COVID-19. Nat. Rev. Phys. 2020

[7] Sarkodie, S.A.; Owusu, P.A. Investigating the cases of novel coronavirus disease (COVID-19) in China using dynamic statistical techniques. Heliyon 2020

[8] Anastassopoulou, C.; Russo, L.; Tsakris, A.; Siettos, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS ONE 2020

[9] Fong, S.J.; Li, G.; Dey, N.; Gonzalez-Crespo, R.; Herrera-Viedma, E. Finding an accurate early forecasting model from small dataset: A case of 2019-nCoV novel coronavirus outbreak. Int. J. Interact. Multimed. Artif. Intell. 2020

[10] Elengoe, A. *COVID-19 Outbreak in Malaysia*. 2020

[11] MEENA, S. *Hypothesis Testing*. Analytics Vidhya. 2020

[12] B.Davis, R. and J.Mukamal, K.. *Hypothesis Testing*. Circulation. 2006

[13] Giuliani, D.; Dickson, M.M.; Espa, G.; Santi, F. Modelling and predicting the spatio-temporal spread of coronavirus disease 2019 (COVID-19) in Italy. SSRN Electron. J. 2020.

[14] Our World in Data. n.d. *Our World in Data*. Available at: <https://ourworldindata.org/>

[15] COVID-19 Malaysia. *COVID-19 Malaysia Updates*. Available at: <http://covid-19.moh.gov.my/>

[16] Kini News Lab. *Covid-19 in Malaysia | Kini News Lab*. Available at: <https://newslab.malaysiakini.com/covid-19/en/patients>

[17] Abs.gov.au. 2021. *Time Series Analysis: The Basics*. [online] Available at: <https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:+the+basics>

[18] R. Sujath, Jyotir Moy Chatterjee, and Aboul Ella Hassanien. A machine learning forecasting model on COVID-19 pandemic in India. 2020

[19] Sha Zhu, Meng Kou, Fujun Lai, Qingxiang Feng and Guorong Du. The Connectedness of the Coronavirus Disease Pandemic in the World: A Study Based on Complex Network Analysis. 2021

[20] Stobierski, T. What is Statistical Modeling for Data Analysis? 2019

[21] HYNDMAN, ROB. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice*. OTEXTS.

Appendices

date	cases_new
1/1/2021	2068
2/1/2021	2295
3/1/2021	1704
4/1/2021	1741
5/1/2021	2027
6/1/2021	2593
7/1/2021	3027
8/1/2021	2643
9/1/2021	2451
10/1/2021	2433
11/1/2021	2232
12/1/2021	3309
13/1/2021	2985
14/1/2021	3337
15/1/2021	3211
16/1/2021	4029
17/1/2021	3339
18/1/2021	3306
19/1/2021	3631
20/1/2021	4008
21/1/2021	3170
22/1/2021	3631
23/1/2021	4275
24/1/2021	3346
25/1/2021	3048
26/1/2021	3585
27/1/2021	3680
28/1/2021	4094
29/1/2021	5725
30/1/2021	5728
31/1/2021	5298
1/2/2021	4214
2/2/2021	3455
3/2/2021	4284
4/2/2021	4571
5/2/2021	3391
6/2/2021	3847
7/2/2021	3731
8/2/2021	3100
9/2/2021	2764
10/2/2021	3288
11/2/2021	3384
12/2/2021	3318
13/2/2021	3499
14/2/2021	2464
15/2/2021	2176
16/2/2021	2720
17/2/2021	2998
18/2/2021	2712
19/2/2021	2936
20/2/2021	2461
21/2/2021	3297
22/2/2021	2192
23/2/2021	2468
24/2/2021	3545
25/2/2021	1924
26/2/2021	2253
27/2/2021	2364
28/2/2021	2437
1/3/2021	1828
2/3/2021	1555
3/3/2021	1745

Dataset of Malaysia' Covid-19 new cases from Jan 2021 to Oct 2021 (1)

4/3/2021	2063	25/3/2021	1360	15/4/2021	2148
5/3/2021	2154	26/3/2021	1275	16/4/2021	2551
6/3/2021	1680	27/3/2021	1199	17/4/2021	2331
7/3/2021	1683	28/3/2021	1302	18/4/2021	2195
8/3/2021	1529	29/3/2021	941	19/4/2021	2078
9/3/2021	1280	30/3/2021	1133	20/4/2021	2341
10/3/2021	1448	31/3/2021	1482	21/4/2021	2340
11/3/2021	1647	1/4/2021	1178	22/4/2021	2875
12/3/2021	1575	2/4/2021	1294	23/4/2021	2847
13/3/2021	1470	3/4/2021	1638	24/4/2021	2717
14/3/2021	1354	4/4/2021	1349	25/4/2021	2690
15/3/2021	1208	5/4/2021	1070	26/4/2021	2776
16/3/2021	1063	6/4/2021	1300	27/4/2021	2733
17/3/2021	1219	7/4/2021	1139	28/4/2021	3142
18/3/2021	1213	8/4/2021	1285	29/4/2021	3332
19/3/2021	1576	9/4/2021	1854	30/4/2021	3788
20/3/2021	1671	10/4/2021	1510	1/5/2021	2881
21/3/2021	1327	11/4/2021	1739	2/5/2021	3418
22/3/2021	1116	12/4/2021	1317	3/5/2021	2500
23/3/2021	1384	13/4/2021	1767	4/5/2021	3120
24/3/2021	1268	14/4/2021	1889	5/5/2021	3744

Dataset of Malaysia' Covid-19 new cases from Jan 2021 to Oct 2021 (2)

6/5/2021	3551	27/5/2021	7857	17/6/2021	5738
7/5/2021	4498	28/5/2021	8290	18/6/2021	6440
8/5/2021	4519	29/5/2021	9020	19/6/2021	5911
9/5/2021	3733	30/5/2021	6999	20/6/2021	5293
10/5/2021	3807	31/5/2021	6824	21/6/2021	4611
11/5/2021	3973	1/6/2021	7105	22/6/2021	4743
12/5/2021	4765	2/6/2021	7703	23/6/2021	5244
13/5/2021	4855	3/6/2021	8209	24/6/2021	5841
14/5/2021	4113	4/6/2021	7748	25/6/2021	5812
15/5/2021	4140	5/6/2021	7452	26/6/2021	5803
16/5/2021	3780	6/6/2021	6241	27/6/2021	5586
17/5/2021	4446	7/6/2021	5271	28/6/2021	5218
18/5/2021	4865	8/6/2021	5566	29/6/2021	6437
19/5/2021	6075	9/6/2021	6239	30/6/2021	6276
20/5/2021	6806	10/6/2021	5671	1/7/2021	6988
21/5/2021	6493	11/6/2021	6849	2/7/2021	6982
22/5/2021	6320	12/6/2021	5793	3/7/2021	6658
23/5/2021	6976	13/6/2021	5304	4/7/2021	6045
24/5/2021	6509	14/6/2021	4949	5/7/2021	6387
25/5/2021	7289	15/6/2021	5419	6/7/2021	7654
26/5/2021	7478	16/6/2021	5150	7/7/2021	7097

Dataset of Malaysia' Covid-19 new cases from Jan 2021 to Oct 2021 (3)

8/7/2021	8868
9/7/2021	9180
10/7/2021	9353
11/7/2021	9105
12/7/2021	8574
13/7/2021	11079
14/7/2021	11618
15/7/2021	13215
16/7/2021	12541
17/7/2021	12528
18/7/2021	10710
19/7/2021	10972
20/7/2021	12366
21/7/2021	11985
22/7/2021	13034
23/7/2021	15573
24/7/2021	15902
25/7/2021	17045
26/7/2021	14516
27/7/2021	16117
28/7/2021	17405

29/7/2021	17170
30/7/2021	16840
31/7/2021	17786
1/8/2021	17150
2/8/2021	15764
3/8/2021	17105
4/8/2021	19819
5/8/2021	20596
6/8/2021	20889
7/8/2021	19257
8/8/2021	18688
9/8/2021	17236
10/8/2021	19991
11/8/2021	20780
12/8/2021	21668
13/8/2021	21468
14/8/2021	20670
15/8/2021	20546
16/8/2021	19740
17/8/2021	19631
18/8/2021	22242

19/8/2021	22948
20/8/2021	23564
21/8/2021	22262
22/8/2021	19807
23/8/2021	17672
24/8/2021	20837
25/8/2021	22642
26/8/2021	24599
27/8/2021	22070
28/8/2021	22597
29/8/2021	20579
30/8/2021	19268
31/8/2021	20897
1/9/2021	18762
2/9/2021	20988
3/9/2021	19378
4/9/2021	19057
5/9/2021	20396
6/9/2021	17352
7/9/2021	18547
8/9/2021	19733

9/9/2021	19307
10/9/2021	21176
11/9/2021	19550
12/9/2021	19198
13/9/2021	16073
14/9/2021	15669
15/9/2021	19495
16/9/2021	18815
17/9/2021	17577
18/9/2021	15549
19/9/2021	14954
20/9/2021	14345
21/9/2021	15759
22/9/2021	14990
23/9/2021	13754
24/9/2021	14554
25/9/2021	13899
26/9/2021	13104
27/9/2021	10959
28/9/2021	11332
29/9/2021	12434

30/9/2021	12735
1/10/2021	11889
2/10/2021	10915
3/10/2021	9066
4/10/2021	8075
5/10/2021	8817
6/10/2021	9380
7/10/2021	9890
8/10/2021	9751
9/10/2021	8743
10/10/2021	7373
11/10/2021	6709
12/10/2021	7276
13/10/2021	7950
14/10/2021	8084
15/10/2021	7420
16/10/2021	7509
17/10/2021	6145
18/10/2021	5434
19/10/2021	5745
20/10/2021	5516

21/10/2021	6210
22/10/2021	6630
23/10/2021	5828
24/10/2021	5666
25/10/2021	4782
26/10/2021	5726
27/10/2021	6148
28/10/2021	6377
29/10/2021	6060
30/10/2021	5854
31/10/2021	4979

Dataset of Malaysia' Covid-19 new cases from Jan 2021 to Oct 2021 (4)