Thai Journal of Mathematics
Special Issue: The 17^{th} IMT-GT ICMSA 2021
Pages 127–134 (2022)

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

A Simple Statistical Model for Forecasting COVID-19 Infections with Application to South and Southeast Asian Countries

Ameen Mhamad[1,2], **Nurin Dureh**[1,*] **and Areena Hazanee**[1]

[1] *Department of Mathematics Computer Science, Faculty of Science Technology, Prince of Songkla University, Pattani Campus, Pattani, 94000, Thailand.*

[2] *The Halal Science Center (HSC-CU), Chulalongkorn University, Bangkok, Thailand.*
e-mail : 6220320004@email.psu.ac.th (A. Mhamad); nurin.d@psu.ac.th (N. Dureh)

Abstract Nowadays, countries worldwide have been facing crises due to the epidemic of coronavirus disease 2019 (COVID-19). This study aimed to construct a model for forecasting COVID-19 infected cases and deaths using the natural cubic spline function. The data used in this study were obtained from publicly available databases updated daily and located on GitHub run by Microsoft. The model fits the data remarkably well, with r-squared values of 0.997, 0.981, 0.992, 0.975, 0.995, 0.957, 0.973, 0.939, 0.989, 0.881, 0.943, and 0.610 in India, Pakistan, Bangladesh, Indonesia, Singapore, Nepal, Australia, Malaysia, Thailand, Sri Lanka, and Vietnam, respectively. The model generates daily change forecasts that indicate when intervention is required. Furthermore, this model is routinely applied to all such regions globally and can be extended to accommodate additional predictors such as environmental and demographic variables.

MSC: 49K35; 47H10; 20M12

Keywords: COVID-19; Natural Cubic Spline; Forecasting; South and South-East Asian

Submission date: 31.01.2022 / Acceptance date: 31.03.2022

1. Introduction

According to its global spread, the disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been a public health emergency of international concern, which was later termed COVID-19. In November 2019, the first case of COVID-19 was discovered in Wuhan (Hubei, China) as a respiratory ailment with an unknown cause [1]. Over 1.2 million deaths have been reported from COVID-19 monthly updates since the World Health Organization (WHO) designated COVID-19 a pandemic [2]. The challenge for public health is to confront emerging and reemerging diseases to protect the world population.

*Corresponding author.

Published by The Mathematical Association of Thailand.

The South and Southeast Asian region (SSEA) is one of the regions that suffer from the pandemic of COVID-19. There are seven countries in SSEA land bordering China, including Bhutan, India, Laos, Myanmar (Burma), Nepal, Pakistan, and Vietnam. The first COVID-19 cases outside China were also reported in SSEA (Thailand on January 13^{th}, 2020) [3]. Furthermore, the SSEA countries suffer from widespread economic and social despair, as seen by a high poverty rate and inadequate health facilities. Therefore, it has posed significant hurdles to public health organizations in terms of quickly implementing pandemic-prevention measures.

COVID-19 has advocated for a variety of measures from the government and public entities to combat the pandemic, including school closures, public gathering bans, travel limits, public transportation shutdowns, and international movement restrictions [4]. COVID-19 has had a significant impact on human existence, as well as the global community and economic progress. To stop the virus from spreading, a long-term and effective measuring method is required. The decision on the best moment for the disease to vanish, on the other hand, is fraught with uncertainties.

Even if effective vaccines or treatments are available, mathematical modeling studies and unambiguous data suggest that public health actions will be essential. The COVID-19 pandemic is relevant even in the slightest hint for short-term forecasting to better manage social, economic, cultural, and public health challenges in the coming month. To deal with the COVID-19 pandemic, some research have used mathematical and statistical methods, such as the time series model [5–7], deep learning methods [8–10], and machine learning methods [11]. Therefore, this study aimed to determine the pattern and trend of the COVID19 pandemic and forecast COVID-19 confirmed cases in South and Southeast Asian countries using a simple statistical model.

2. MATERIALS AND METHODS

2.1. Data Source

The data used in this study were obtained from publicly available databases updated daily reported by Johns Hopkins University coronavirus resource center and located on GitHub run by Microsoft. The dataset includes cumulative confirmed cases, recoveries cases, and the number of deaths.

2.2. Data Management

Data management is a crucial aspect in this research. This section seeks to clean up raw data by removing variables that aren't relevant to the study. Data management was divided into four steps. It all began with downloading the dataset from one of the many internet sources and picking the area to be combined and transposed. Finally, the data was organized into a 7-day bin or weekly data set, yielding 624 observations and nine variables for analysis. The R program was used to manage all of the previous steps.

The process began with the download of cumulative confirmed cases, recoveries cases, and death cases. Each dataset contained 274 rows and 439 columns of confirmed and death cases, as well as 259 rows and 439 columns of recoveries cases, with five variables comprising nation names, state names, latitude, longitude, and observation day. The second phase of data management is data collecting in specific areas. We had to subset the study area because the dataset was available for the entire world. Countries in South and Southeast Asia were chosen, including India, Nepal, Bangladesh, Philippines, Indonesia, Singapore, Nepal, Malaysia, Thailand, Sri Lanka, Vietnam, and Australia. The day of the year was computed in the third step. Daily cases were derived from the cumulative data of confirmed cases, death cases, and recovery cases. Finally, the dataset yielded 624 rows containing six variables: country name, pandemic date, weekly total confirmed cases, weekly death cases, and weekly total recovery cases.

2.3. Statistical analysis

For categorical variables, frequency and percentage were used; for continuous variables, standard deviation, minimum, maximum, mean, and graphical techniques were used to visualize the pandemic trend. To assess the severity of the COVID-19 pandemic, case fatality rates were calculated.

The cube root transformation was employed to ensure that the data met the distributional and variance homogeneity requirements. The COVID-19 epidemic was then fitted and predicted using the natural cubic spline. In addition, the model was used to forecast the cumulative confirmed cases in the short term.

The spline function was used for smoothing the interpolation of the regression model. A cubic spline with k knots $\mathrm{x}_1 < x_2 < \ldots < x_k$ is any function s(x) with continuous second derivatives comprising piecewise cubic polynomials between and beyond the knots [12]. This function is written as

$$S(x) = d_0 + d_1 x + d_2 x^3 + \sum_{i=1}^{n} c_i (x - x_i)_+^3 \tag{2.1}$$

The additional requirement for a natural cubic spline is that the function is linear for values of x outside the knots since s(x) is linear for $x < x_1$ if d_2 and d_3 terms in s(x) must also disappear for $x < x_k$, so to be natural spline, the n+4 coefficients in the cubic spline must satisfy the following two sets of equations.

$$d_2 = 0, \sum_{i=1}^{n} c_i = 0, \;\; d_3 = 0, \sum_{i=1}^{n} x_i c_i = 0 \tag{2.2}$$

The natural cubic spline produces the smoothest results when fitting a function to data because it minimizes the integral of the fitted function's squared second derivative. For knot X_k where k ranges from 1 to k, this function has the formula

$$S(x) = a + b_x + \sum_{i=1}^{n} C_k \{(x - x_k)_+^3 - d_1 (x - x_{p-1})_+^3 + d_2 (x - x_{p-2})_+^3\}, \tag{2.3}$$

Where $\mathrm{d}_1 = (x_p - x_k)/(x_p - x_{p-1}), d_2 = (x_{p-1} - x_k)/(x_p - x_{p-1})$, and $\mathrm{x}_+ = max(x, 0)$, The (positive part of x),which are 12 equi-spaced knots to fit the data.

3. RESULTS

Table 1 shows the COVID-19 outbreak in South Asian countries. Among these, the explosion in India is higher than in the other four countries. The first case in India was reported together with Sri Lanka on January 27^{th}, 2020. On the other hand, the COVID-19 pandemic started earlier in Nepal (January 23^{rd}, 2020) and later in the other two countries (February 26^{th}, 2020, in Pakistan, and March 7^{th}, 2020, in Bangladesh).

Table 1. The COVID-19 pandemic, confirmed, death, and recovered cases in selected south and Southeast Asian countries

Continents	Countries	Confirmed Cases	Death Cases	Recoveries Cases	Case Fatality Rate	First started Date
South Asia	India	10,610,883	152,869	10,265,706	1.44	27/1/2020
	Bangladesh	529,687	7,950	474,472	1.50	07/3/2020
	Pakistan	527,146	11,157	480,696	2.12	26/2/2020
	Nepal	268,310	1,975	262,642	0.74	23/1/2020
	Sri Lanka	55,189	274	47,215	0.50	27/1/2020
Southeast Asian	Indonesia	939,948	26,857	763,703	2.86	01/3/2020
	Philippines	505,939	10,042	466,993	1.98	30/1/2020
	Malaysia	169,379	630	127,662	0.37	25/1/2020
	Singapore	59,197	29	58,926	0.05	23/1/2020
	Thailand	12,795	71	9,842	0.55	13/1/2020
	Vietnam	1,544	35	1,406	2.27	23/1/2020
Australia	Australia	28,749	909	22,716	3.16	25/1/2020

In India, the total number of confirmed, recovered, and death cases of COVID-19 were 10,610,883, 10,265,706, and 152,869, respectively.

In Southeast Asian, the highest number of confirmed cases was 939,948 in Indonesia, followed by the Philippines with 505,939 cases, Malaysia with 169,379 cases, Singapore with 59,197 cases, Thailand with 12,795 cases, and 1,544 cases from Vietnam. On January 13^{th}, 2020, Thailand was the first country to announce confirmed cases. Despite the fact that Indonesia was the most recent Southeast Asian country to report COVID-19 infected cases, the number of confirmed cases has been steadily increasing, with Indonesia having the largest number of confirmed cases and the highest case fatality rate among Southeast Asian countries.

We plot time series of confirmed, recovery, and death cases data from 12 nations in the accompanying figures to establish the trend of the COVID-19 pandemic in the studied area.

Figure 1 shows the patterns of the COVID-19 new cases. The black dots represent the number of confirmed cases. The blue dots represent the number of recovered cases, and the red dots represent the number of death cases. Overall, the number of deaths and recovered cases follow the same pattern. However, the number of fatalities is shown to be lower than the confirmed and recoveries case. The graph showed that the COVID-19 pandemic could be categorized into five patterns of the pandemic.

The first pattern in Indonesia. As shown in figure 1, the number of COVID-19 pandemics significantly increased but did not reach its peak. The total number of cases was reached the highest in the last week of observation. There were 939,948 confirmed cases and 763,703 recoveries, and 2,685 deaths. The number of COVID-19 confirmed cases rose sharply over the last 11 weeks and reached the highest point at 81,905 confirmed cases in the 52^{nd} week. The number of recovered and death cases of COVID-19 is rising, similar to the confirmed cases. However, there are differences between the increasing size and the number of recoveries.

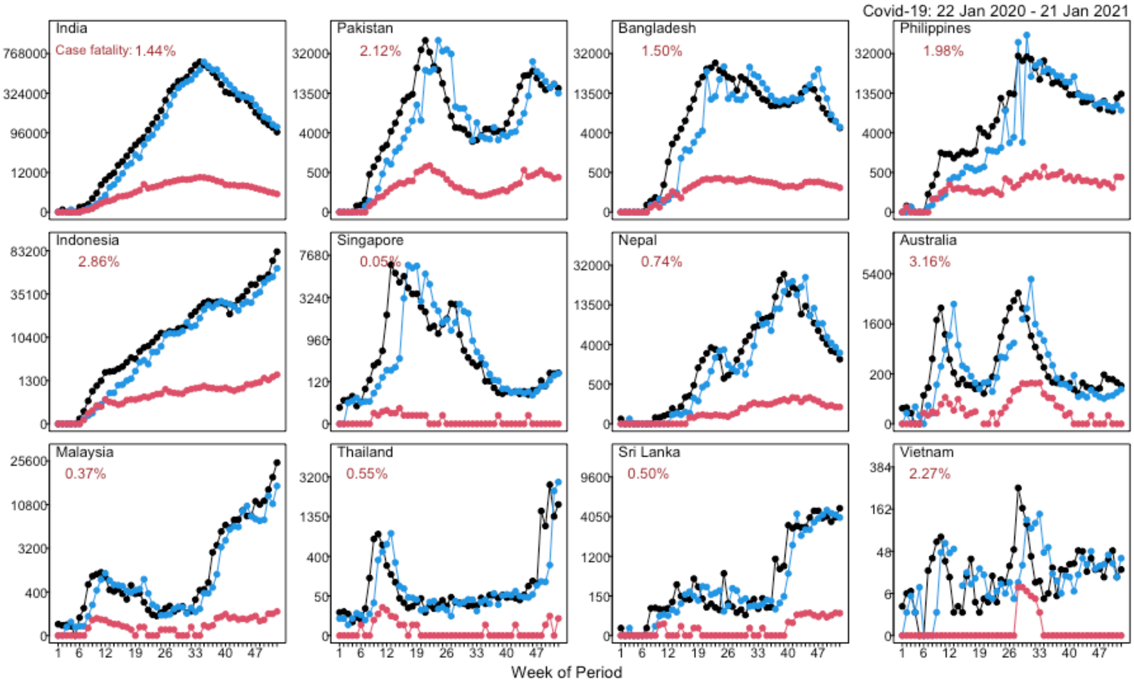

FIGURE 1. The distribution of COVID-19 pandemic cases

In the second pattern, there are five countries, including India, Bangladesh, Nepal, Singapore, and the Philippines. Figure 1 shows the confirmed cases and recoveries increased to the peak one time. After that, the confirmed cases decreased to the based line of the pandemic. The number of confirmed cases in India is higher than the other three countries, reaching the 34^{th} week with 652,390 confirmed cases. While Singapore, the confirmed cases sharply reached the peak in the 13^{th} week at 6,442 confirmed cases. After that, it gradually declines to get the baseline in the 39^{th} week of the COVID-19 pandemic. Nepal, the confirmed cases gradually peaked at 26,876, the number of confirmed cases between the 8^{th} and 39^{th} week. The cases fatality rate in the Philippines is higher than the other courtiers, 1.98 percent, while Singapore has the lowest case fatality ratio at 0.05 percent.

The third pattern shows a rise in confirmed and recovered cases followed by a drop to the baseline level. The number of confirmed and recovered cases then surged (the second wave of COVID-19 was detected). In Malaysia, confirmed cases rose dramatically to a new high on the last day of observation, with 23,861 confirmed cases, after hitting their first peak in the 11^{th} week. Despite having the highest number of confirmed cases, Malaysia's case fatality rate of 0.27 percent is lower than the other two countries. In Thailand and Sri Lanka, pandemics follow the same pattern. In Thailand and Sri Lanka, case fatalities are 0.55 percent and 0.50 percent, respectively.

The fourth pattern was found in Australia and Pakistan, as shown in Figure 1. In Australia, the number of confirmed cases peaked in the 10^{th} and 28^{th} week with 3,592 cases. In Pakistan, the number of confirmed cases of the first peak was higher than the second peak, with 40,582 confirmed cases in the 21^{st} week. The second peaked at the 46^{th} week after showing a downward trend in the last observation period. The fatality rate in Australia was higher than Pakistan, with 3.16 and 2.21 percent, respectively. The fatality rate in Australia was highest compared with all the selected countries. Therefore, we grouped Australia and Pakistan in the same group.

The last pattern of the graphical pandemic of COVID-19 is in Vietnam. There was an oscillation of the pandemic in Vietnam between the 1^{st} to 23^{rd} week of being obviated. The highest number of confirmed was 258, which was lower than the pandemic's peak in observed countries in the 28^{th} week. The number of recoveries follows the pattern of the confirmed cases. The number of deaths in Vietnam was found in the 28^{th} week, which was the same as the week that confirmed cases reached the pandemic's peak with a cases fatality rate of 2.27, as shown in Figure 1.

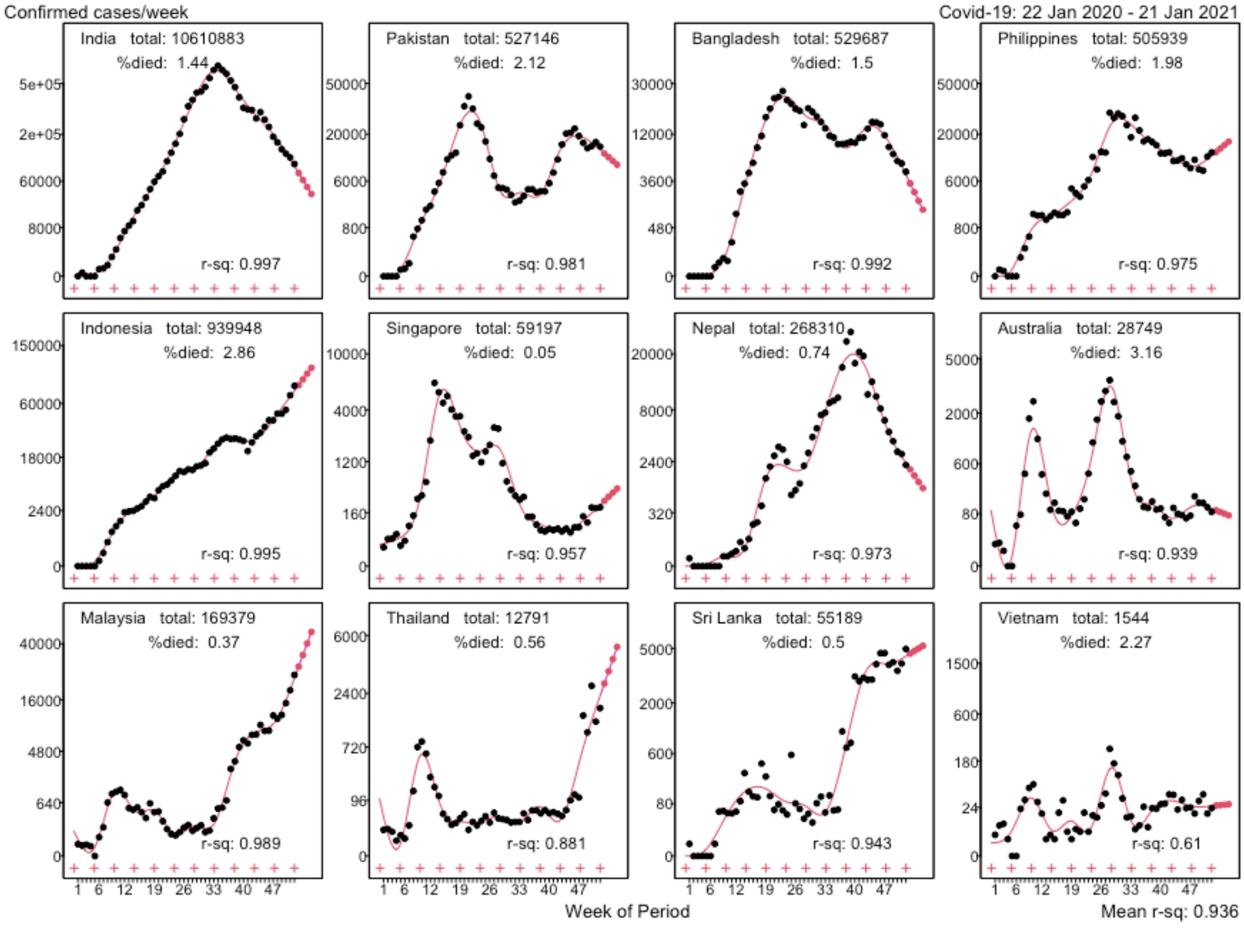

Figure 2. Spline regression model for COVID-19 confirmed cases

A natural cubic spline is linear outside the range of the knots, which can provide the linear forecasting value in the tail of the natural cubic function. So, we employed the natural cubic spline function with equi-spaced knots, which are 12 knots, to fit the data and forecast the COVID-19 confirmed cases. The results are shown in Figure 2. The black dot represents the recording of confirmed cases, the pink line is the spline model, and the pink dot is the forecasted data.

The results showed that the data natural cubic spline fit the data well with a high r-square value. However, the number of r-square values differs between countries, shown as 0.997 in India, 0.979 in Pakistan, 0.992 in Bangladesh, 0.978 in the Philippines, 0.995

in Indonesia, 0.961 in Singapore, 0.975 in Nepal, 0.942 in Australia, 0.989 in Malaysia, 0.945 in Sri Lanka, and 0.886 in Thailand. The lowest r-square was in Vietnam, and it is equal to 0.598. Moreover, the model also provides short-term forecasting of the COVID-19 confirmed cases. The forecasting trend can be categorized into three groups. The first, the COVID-19 confirmed cases, tends to increase in the next four weeks, comprising Indonesia, Malaysia, Thailand, Sri Lanka, Philippines, and Singapore. Second, the trend of the COVID-19 confirmed cases decreased in the coming months, including in India, Pakistan, Bangladesh, and Nepal. Lastly, the forecasting trend with a steady trend in the short-term was found in Australia and Vietnam.

4. DISCUSSIONS AND CONCLUSIONS

COVID-19 pandemic confirmed cases, recovered cases, and death cases were utilized to compare the pandemic pattern of COVID-19, which was categorized into five categories indicating the pandemic pattern in South and Southeast Asia in this study. Furthermore, with a high r-squared value, the natural cubic spline model with 12 equally spaced knots fit COVID-19 confirmed cases well.

In South Asia, cases fatality rate ranges from 0.5 to 2.12 percent. The lowest fatality rate was shown in Sri-Lanka. The highest fatality rate was shown in Nepal. In Southeast Asia, cases fatality rate ranges from 0.05 to 2.86 percent. The lowest case was shown in Singapore. The highest case fatality rate was shown in Indonesia. The COVID-19 case fatality rate varies in South and Southeast Asian countries, which is consistent with the findings of a previous study [13]. Case fatality rate varies for a variety of reasons. The number of confirmed cases was determined by the testing policy. As a result, the differing COVID-19 testing procedures had an impact on the number of confirmed cases in each country. Finally, in a public health crisis, collecting method of COVID-19 data had an impact on the number of deaths.

Fitting the data by natural cubic spline model with equi-spaced knots provides the model fit well with the data as shown in high r-squared values. However, some countries in this study have a low r-squared than the others. It was interpreted that the model does not fit much with the data. All mentioned was the limitation of the natural cubic spline model with equi-spaced knot [14]. To improve the accuracy of prediction knot adjusted in each country which low COVID-19 reported pandemic cases and oscillation trend on the COVID-19 data [15].

In conclusion, COVID-19 has spread worldwide, an ongoing pandemic warning a vital public-health threat, there is still a crucial need for forecasting models. That could help predict more probable pandemic situation waves. The obtained forecasting results indicate that the natural cubic spline model with equi-spaced knots available to use at any time, for any country, the region was one of the alternative measures to forecasting the short term of COVID-19 confirmed cases.

Acknowledgements

This work was supported by the faculty of science and technology and the graduate school, Prince of Songkhla University, Thailand.

References

[1] N. Zhu, D. Zhang, W. Wang, A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine, New England Journal of Medicine, 2020.

[2] S. Platto, T. Xue, E. Carafoli, COVID19: an announced pandemic, Cell Death Disease 11 (9) (2020) 1-13.

[3] S. Triukose, S. Nitinawarat, P. Satian, A. Somboonsavatdee, P. Chotikarn, T. Tammasanya, Y. Poovorawan, Effects of public health interventions on the epidemiological spread during the first wave of the COVID-19 outbreak in Thailand, Plos One 16 (2) (2021).

[4] T. Hale, N. Angrist, Variation in government responses to COVID-19 BSG-WP-2020/032, BSG Working Paper Series, 2020.

[5] O.D. Ilie, R.O. Cojocariu, A. Ciobica, S.I. Timofte, Forecasting the spreading of COVID-19 across nine countries from Europe, Asia, and the American continents using the ARIMA models, Microorganisms 8 (8) (2020) 1158.

[6] H. Iftikhar, M. Iftikhar, Forecasting daily COVID-19 confirmed, deaths and recovered cases using univariate time series models: A case of Pakistan study, MedRxiv (2020) 1-13.

[7] M. Yousaf, S. Zahir, M. Riaz, S. Hussain, K. Shah, Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan. Chaos, Solitons Fractals, 2020.

[8] A. Zeroual, F. Harrou, A. Dairi, Y. Sun, Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos, Solitons Fractals 140 (2020).

[9] P. Arora, H. Kumar, B. K. Panigrahi, Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos, Solitons Fractals 139 (2020).

[10] M. Wieczorek, J. Sika, M. Wozniak, Neural network powered COVID-19 spread forecasting model. Chaos, Solitons Fractals 140 (2020).

[11] D.P. Kavadi, R. Patan, M. Ramachandran, A.H. Gandomi, Partial derivative nonlinear global pandemic machine learning prediction of covid 19. Chaos, Solitons Fractals 139 (2020).

[12] N. McNeil, P. Odton, A. Ueranantasun, Spline interpolation of demographic data revisited. Sonklanakarin Journal of Science and Technology 33 (1) (2011) 117-120.

[13] E. B. Karnadi, T.A. Kusumahadi, Why Does Indonesia Have a High Covid-19 Case-Fatality Rate, JEJAK: Jurnal Ekonomi dan Kebijakan 14 (2) (2021) 272-287.

[14] A.R.D.S. Sousa, M.T. Severino, F. G. Leonardi, Model selection criteria for regression models with splines and the automatic localization of knots, Preprint arXiv, 2020.

[15] A. K. Iddrisu, D. Otoo, A predictive model for daily cumulative COVID-19 cases in Ghana. F1000Research 10 (2021) 343.