Original Article

Data-Driven Transmission Patterns of COVID-19 in ASEAN+6

Panyawut Sri-iesaranusorn[1], Amornpong Trakarnkulphun[2], Attawit Chaiyaroj[3], Decho Surangsrirat[4]

Abstract

Introduction: Our objective is to discover transmission patterns of COVID-19 in the group of 16 countries called ASEAN+6 which comprise the ten countries in ASEAN and China, Japan, South Korea, Australia, New Zealand, and India.

Methods: The public dataset from John Hopkins University was used in this work. The concept of the effective reproduction number ($\mathcal{R}$) based on the SIR model is used to define the wave of infection. K-means clustering, an unsupervised machine learning algorithm, is then applied to the time-series data to divide the waves into clusters.

Results: The data of the confirmed cases and fatalities were separated into four clusters. The results from the confirmed cases suggest that the countries in Cluster 1 can handle the spread of COVID-19 better than the countries in Cluster 2 for the first 20 days of their waves. The results from the fatalities data suggest that there is a pattern of each country's capacity of the public healthcare system and the effectiveness in handling the COVID-19 situation.

Conclusions: The data seems to support the idea that the clusters of the confirmed cases and deaths may be related to each country's epidemic control measures and the capacity of the public healthcare systems. Future research may consider the COVID-19 patterns in this study and compare them with the current situation for further analysis.

Keywords: COVID-19, ASEAN+6, Unsupervised learning, Transmission pattern, Effective reproduction number

Received: 5 July 2021 **Revised: 20 July 2021** **Accepted: 27 July 2021**

[1] Division of Information Science, Nara Institute of Science and Technology, Nara, Japan
[2] Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[3] School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[4] Assistive Technology and Medical Devices Research Center, National Science and Technology Development Agency, Pathum Thani 12120, Thailand
Corresponding author: Decho Surangsrirat, National Science and Technology Development Agency, Pathum Thani 12120, Thailand
Email: decho.sur@nstda.or.th

Introduction

As a highly transmissible disease, the coronavirus disease 2019 (COVID-19) has been declared a pandemic since March 20th, 2020.[1] The most common symptoms for COVID-19 include fever, dry cough, and tiredness. Most people will experience mild to moderate respiratory illness and can recover without special treatment. However, older people and those with underlying diseases are more likely to develop serious illnesses.[2-5] Assiduously dealing with the virus for over a year, the world has accumulated over 183 million confirmed cases and over 4 million fatalities at the time of writing. The epidemic devastates the public healthcare system and creates a worldwide transportation shutdown, which has a significant impact on the global economy.

To analyze the situation, previous works used public datasets to examine the information of the disease. Multiple research works intended to visualize and predict the spreading of COVID-19.[6-9] Some of the publications provided grouping strategies for separating the cluster of populations for further investigation.[10-13] Two of the most popular methods for the analysis are the SIR model and K-means clustering. For time-series analysis, the SIR model can be used to construct an infection graph.[8-10] K-means clustering is a well-known unsupervised clustering algorithm.[11-13] As part of the public effort to decode the disease, Johns Hopkins University released an up-to-date public dataset to the COVID-19 researcher.[14]

In this study, our objective is to discover transmission patterns of COVID-19 in the group of 16 countries called ASEAN+6 which comprise the ten countries in ASEAN and China, Japan, South Korea, Australia, New Zealand, and India. The public dataset from John Hopkins University was used in this work.[14] The concept of an effective reproduction number ($\mathcal{R}$) is used to define the wave of infection. K-means clustering, an unsupervised machine learning algorithm, is then applied to divide the waves into clusters. The remainder of this paper is organized as follows. Section II describes the methodology including the dataset, definition of wave in this study, and pre-processing and K-means clustering. In section III, we present the results from an unsupervised clustering algorithm for an analysis of the transmission patterns of COVID-19 based on the number of new cases and deaths. Section IV and V describe our interpretation of the results and possible extensions and improvements of the study.

Methods

A. COVID-19 Dataset

We use COVID-19 case data from Johns Hopkins University,[14] which can be accessed via a GitHub repository.[15] Due to the advantages of being up-to-date and comprehensive information, this dataset was used for the investigation on COVID-19 in multiple publications.[6, 7, 16] The data has been collected daily since January 22nd, 2020 from many governmental websites which provide an accumulated number of cases, deaths, and recovered cases of over 3,000 regions in more than 180 countries. Note that the data from some countries consist of multiple provinces or states. In this study, only the confirmed cases and deaths data of ASEAN+6 were used. ASEAN+6 consists of the 10 ASEAN countries: Thailand, Malaysia, Singapore, Brunei, Philippines, Vietnam, Cambodia, Indonesia, Laos, and Myanmar, along with six additional countries: China, Japan, South Korea, India, Australia, and New Zealand. We refer to the country Myanmar by its current name, instead of Burma as used in the dataset. ASEAN countries were chosen because we hypothesize that commonalities in cultures and climates may affect how the virus spreads and how those countries encounter it. The additional six countries were included in this investigation so that we can study countries with large populations such as China and India, as well as countries that are geographically isolated and have taken isolation measures, such as New Zealand and Australia.

B. Wave of Infection

Following epidemiology practices of infectious diseases, we used the effective reproduction number ($\mathcal{R}$) to track the dynamics of COVID-19 and define the wave of infection. To observe a growing epidemic in waves, previous works used this number to quantify the virus transmissibility, and determine the duration of a wave of infection.[17, 18] There are various ways to estimate $\mathcal{R}$ such as basic reproduction number,[19] nonparametric compartmental models,[20] and the state-space method.[21] Recently, Arroyo-Marioli et al.[22] developed a new method

to estimate the effective reproduction number $\mathcal{R}$ based on SIR model and Kalman Filtering,[23] as well as providing an interactive online dashboard.[24] Hence, considering the effectiveness and availability of data, we decided to use the estimated reproduction number $\mathcal{R}$ to determine the wave of infection in this study.

SIR model is one of the simplest compartmental models from the concept of the effective reproduction number $\mathcal{R}$. It has been used for preventing, monitoring, as well as forecasting the spread of COVID-19.[8-10] This model contains three main components. First, susceptible (S) is the people who are not infected with the disease yet. However, they are not immune to the disease and can become infected with the disease in the future. Second, infected or infectious (I) is the people who are infected with the disease and can transmit the disease to susceptible people. Last, recovered ($\mathcal{R}$) is the people who have recovered from the disease and became immune, so they can no longer be infected with the disease. The summation of these three components represents the total population size. There are two more important parameters for the SIR model: the daily transmission rate (β) and the daily transition rate from infected (γ). The effective reproduction number ($\mathcal{R}$) is derived from

$$\mathcal{R} = \beta\gamma$$

where β is the rate of daily transmission and γ is the rate of daily transition from infected. This number can indicate the growth in infection, specifically when $\mathcal{R} > 1$.

A new wave of infection is defined in our experiment based on two conditions. The first condition is that the estimated $\mathcal{R}$ is higher than 1 for 14 consecutive days. We chose 14 days for the condition to correspond with the virus's incubation period, which is the interval between exposure and symptom development. The second condition is that the wave of infection must not intersect with other waves. In the case that there is an intersection, only the first wave is considered. Following this definition, we have identified 29 waves of infection for this dataset, as shown in Table 1. According to our definition of wave and conditions, no waves for Brunei and Laos are recognized. For the clustering analysis based on the time-series, only **120 days** of data from the start of each wave were utilized.

Table 1 Wave of infection as defined by the concept of the effective reproduction number

Country	#Wave	Start date
Australia	2	2020-03-11, 2020-12-10
Myanmar	1	2020-08-12
Cambodia	1	2021-02-12
China	3	2020-01-23, 2020-05-29, 2020-10-04
India	2	2020-03-15, 2021-02-16
Indonesia	2	2020-03-16, 2020-11-06
Japan	2	2020-02-22, 2020-10-06
Malaysia	2	2020-03-10, 2020-07-16
New Zealand	3	2020-03-23, 2020-08-11, 2020-12-25
Philippines	2	2020-03-15, 2021-01-06
Singapore	3	2020-03-07, 2020-07-11, 2020-12-03
South Korea	2	2020-02-21, 2020-08-09
Thailand	2	2020-03-16, 2020-08-31
Vietnam	2	2020-07-20, 2020-12-21

C. Data Pre-processing and K-means Clustering

Due to the different population sizes of each country, normalization is essential. The original dataset expresses the number of cases or deaths in the form of accumulated sums. We first subtract the numbers of each day by the number from the first day of the wave, to eliminate the effect of cases accumulated since the beginning of the pandemic. Then from those numbers, we derived percentages of the increase in the number of cases or deaths compared to the previous day, as shown in Equation

$$p_i = \frac{s_i - s_{i-1}}{s_{i-1}}$$

where s_i is the accumulated number of cases or deaths at day i. In the case that $s_{i-1}$is equal to zero, p_i is defined as zero. This approach was applied to the analysis and comparison of waves between countries.

After obtaining the waves of infection in each country, we performed an analysis using K-means clustering. K-means clustering is one of the most well-known unsupervised learning algorithms and is generally used with non time-series datasets. Prior works used this algorithm in several clinical fields such as disease prediction,[25, 26] gene expression analysis,[27, 28] and COVID-19 analysis.[11-13] The main idea of the algorithm is to solve the problem of classifying the given data into *k* different clusters based on certain distance metrics such as Euclidean distance. To apply this algorithm to time-series data, it is required to use the distance based on the Dynamic Time Warping (DTW) algorithm as a distance metric between two time-series data points. Dynamic Time Warping (DTW) is a similarity measure between time-series.[29] DTW algorithm allows for matching of peaks in the waves of infection by reducing the effects of time and shifting distortion between the two signals in order to detect similarities between their phases and shapes, as described by Algorithm 1. Using DTW, the distance between two sequences can be defined as the *minimum* Root Mean Square Error (RMSE) between the sequences achievable by warping. In this context, a sizable difference in RMSE implies a greater difference between the two waves of infection. Roughly speaking, we use the DTW algorithm and K-means clustering to investigate the patterns of COVID-19.

In this experiment, we explored the number of clusters *k* from 2 to 10 and selected the number that minimizes the clustering objective function as the main number of clusters. By the nature of random initialization for the centroid positions in K-means clustering, it could not be guaranteed that the algorithm will find the optimal clusters. To deal with this initialization issue, we ran 1,000 separately-initialized trials for each number of clusters *k* and selected the trial with the best objective function value.

Algorithm 1 Dynamic Time Warping Algorithm

```
Data: s, t: the sequences
Result: d: the minimum distance
n = s.length() ;
m = t.length() ;
DTW = array[0...n][0...m] ;
for i ← 0 to n do
    for j ← 0 to m do
        DTW[i][j] = infinity
    end
end
DTW[0][0] = 0 for i ← 0 to n do
    for j ← 0 to m do
        cost = RMSE(s[i], t[i]) ;
        DTW[i][j] = cost + minimum(DTW[i − 1, j], DTW[i, j − 1], DTW[i − 1, j − 1])
    end
end
return DTW[n][m];
```

Results

A. Clustering Based on New Cases

According to the evaluation metrics and the interpretation from experts, the data of the confirmed cases can be divided into four clusters as shown in Figure 1. The figure represents the latent space of confirmed cases data after we applied Principle Component Analysis (PCA) for visualization. Figure 2 illustrates the characteristics of each cluster. The members of each cluster are as follows:

- Cluster 1: The first wave of Australia, Burma, China, Indonesia, Japan, New Zealand, Malaysia, Singapore, South Korea, Thailand, and Vietnam. The second wave of Australia, India, Indonesia, Japan, New Zealand, Malaysia, Philippines, Singapore, South Korea, Thailand, and Vietnam. The third wave of China, and New Zealand.
- Cluster 2: The first wave of India, the second wave of China, and the third wave of Singapore.
- Cluster 3: The first wave of Cambodia.
- Cluster 4: The first wave of Philippines.

Two of the four clusters contain only one country's wave each. This is likely due to the extreme spikes found in the waves. The first waves of Cambodia and Philippines contain days with a growth ratio of more than 900 and 2,200 percent, respectively, compared to the accumulated sum of their previous days. These outlier ratios may have placed the data points unusually far from others, mostly within a range of 0 to 100 percent, causing the clustering algorithm to produce standalone clusters for each of them. As for the other two clusters, the results suggest that Cluster 1 can handle the spread of COVID-19 better than Cluster 2 for the first 20 days of their waves. Apart from that, only small differences can be found between them despite the algorithm producing distinct clusters. Therefore, we conclude that any patterns in the confirmed cases are unclear to our current study, and more investigation is required.

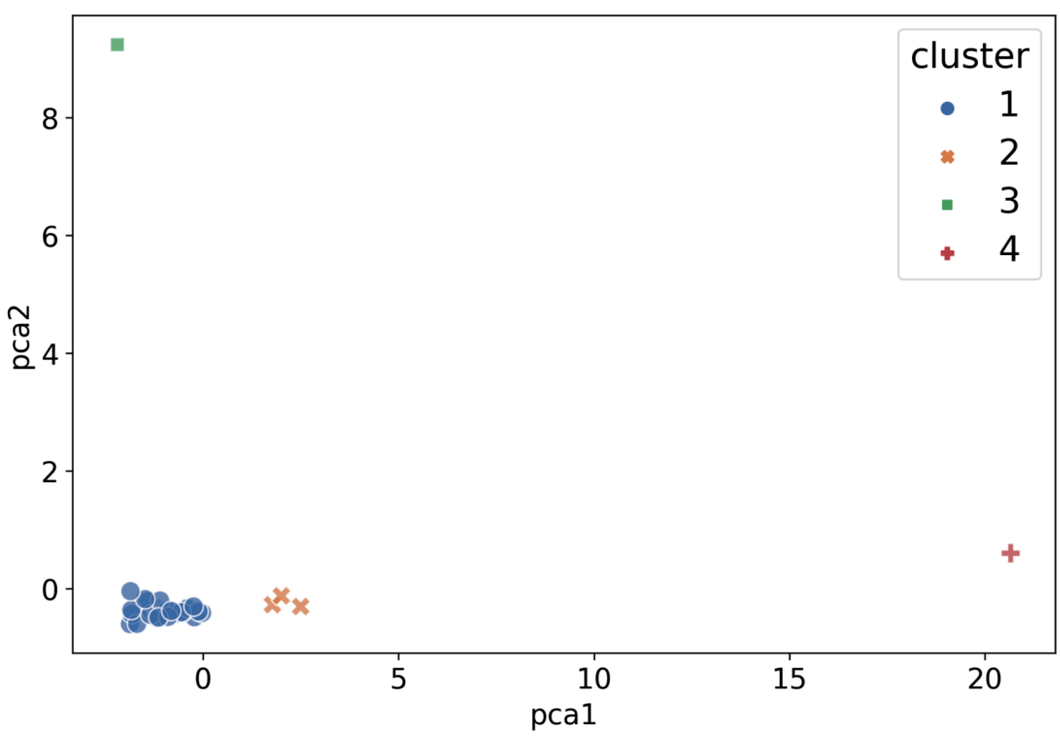

Figure 1 The latent space of the confirmed cases data after we applied Principle Component Analysis (PCA).

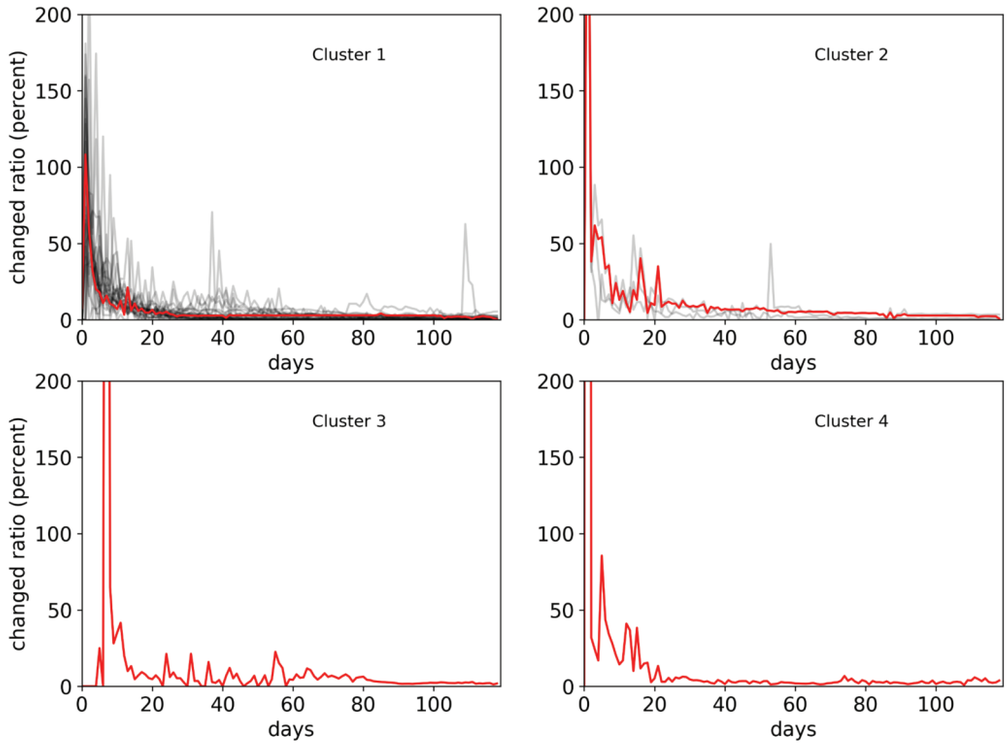

Figure 2 The four clusters derived from the K-means unsupervised clustering algorithm from the data of the confirmed cases. X-axis represents the time since the wave started, from day 1 to day 120. Y-axis represents the change in percentage of the number of new cases. The red line is the centroid of the cluster, and the grey lines are the actual values of each wave in each cluster.

B. Clustering Based on Fatalities

Similar to the results for the confirmed cases, the deaths data can be divided into four clusters as shown in Figure 3. We illustrated the characteristics of each cluster in Figure 4. The element of each cluster is as follows:

Once again, we can see one standalone cluster out of all four. Interestingly, Clusters 1, 2, and 3 seem to represent high, medium, and low levels of fatalities from COVID-19, respectively. The growth ratio of cluster 1 is higher than 100 percent for the first 30 days of their waves. We can infer that the countries in this cluster took longer to handle the COVID-19 situation. In contrast to Cluster 1, the growth ratio of Cluster 3 is lower than 50 percent for the whole wave, which shows the effectiveness of the countries' adaptations to the situation. As for Cluster 2, with the percentage change of deaths not exceeding 100 percent, the fatality rate seems to be more stable than Cluster 1. The last standalone cluster contains the first wave of Philippines. It is likely isolated for the same reasons as the standalone clusters from the previous section. One of the possible interpretations from the results of the unsupervised clustering of the fatalities data is that Clusters 1, 2, and 3 follow a pattern of each country's capacity of the public healthcare system and the effectiveness in handling the COVID-19 situation.

Discussion

The results suggest that there are four clusters of COVID-19 situations based on information about confirmed cases and deaths. These results are promising for use in investigating transmission patterns of COVID-19 in ASEAN countries. Regarding results of both confirmed cases and deaths, it is possible that the patterns are related to the ability to handle the COVID-19 spread situation of each country. More specifically, the data seems to support that the clusters of the confirmed cases could related to each country's epidemic control measures and the clusters of the fatalities could related to the capacity of the public healthcare systems. This motivates us to further investigate the differences between each cluster in the future.

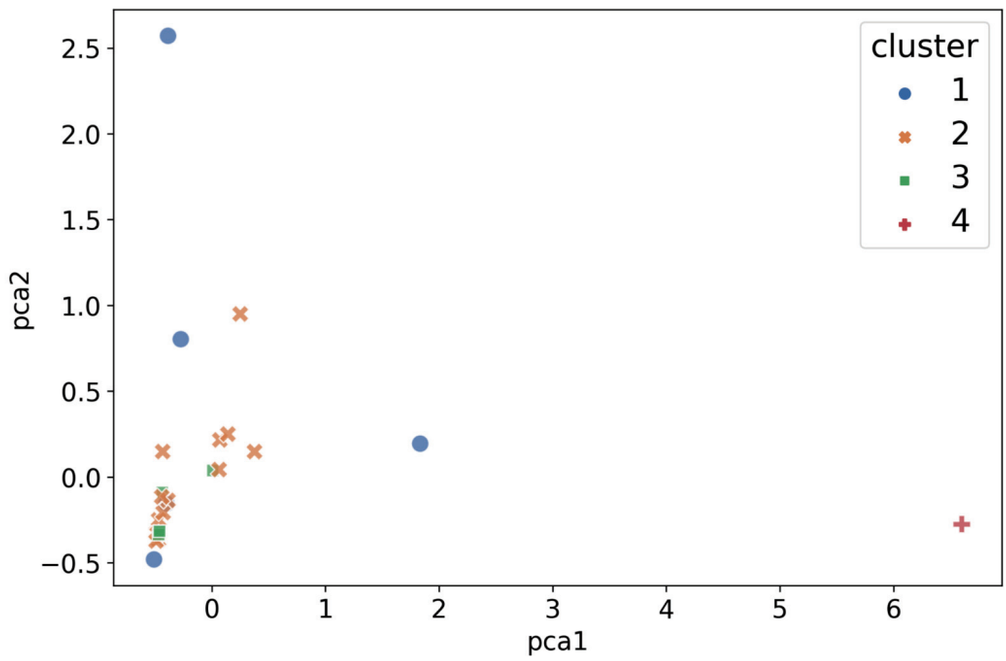

Figure 3 The latent space of the confirmed deaths data after we applied Principle Component Analysis (PCA).

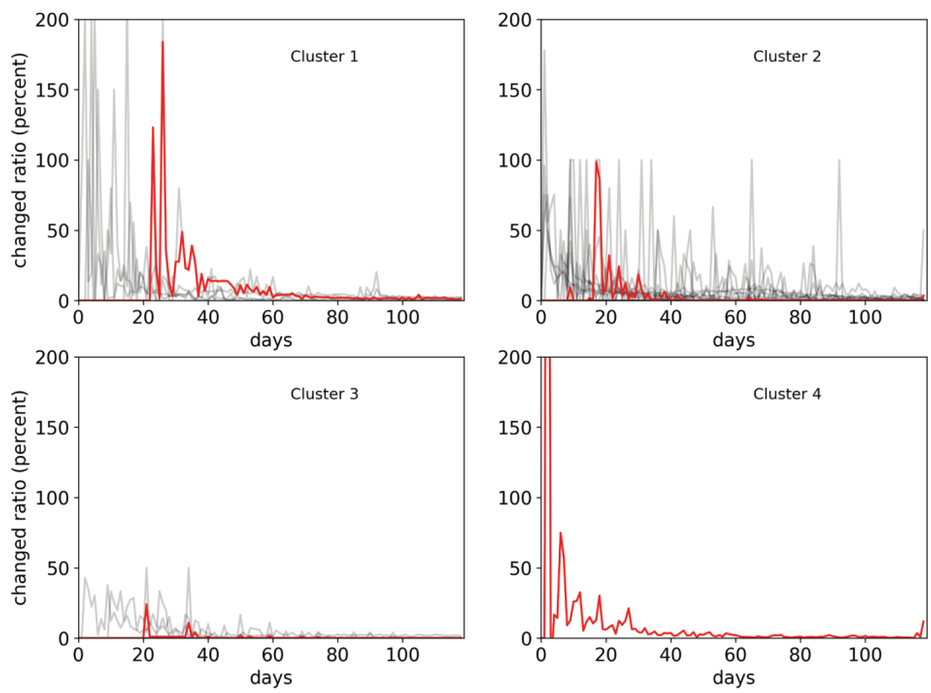

Figure 4 The four clusters derived from the K-means unsupervised clustering algorithm from the data of the death cases. X-axis represents the time since the wave started, from day 1 to day 120. Y-axis represents the change in percentage of the number of death cases. The red line is the centroid of the cluster, and the grey lines are the actual values of each wave in each cluster.

Although this analysis gives us impressive and interesting results, there are three limitations that affect the results of this study. First, there were some entries that had lower amounts than the previous days, even though they are supposed to be non-decreasing due to being cumulative sums. Furthermore, data from certain date ranges are missing from some countries, leading to incom-

plete sequences. We suspect that this might have been caused by some errors during the collection process of the dataset. In the future, we plan to use other available datasets to validate this dataset, in order to achieve a more reliable dataset. Second, the conditions in the definition for waves of infection may need to be verified. Although we provided our reasons for defining the two conditions as such, we cannot guarantee that this definition is appropriate for all situations, countries, or timings. It would be ideal to survey past definitions and expert opinions for a consensus on the definition of waves. Last, our use of K-means clustering might have produced non-optimal results due to random initialization and parameter tuning. To the best of our knowledge, running 1,000 trials of random initialization is sufficient in practice to obtain meaningful and usable results. As for parameter tuning, however, we did not perform a complete grid search for the optimal combination of hyper-parameters, so it is possible that we have missed the optimal parameters that might have been found otherwise. Still, these results should be sufficient as a preliminary study to show the effectiveness of unsupervised learning for discovering transmission patterns of COVID-19.

This paper presents an investigation for the transmission patterns of COVID-19 in ASEAN+6 countries, as discovered by an unsupervised machine learning method. We can infer from these results the effectiveness of each country in managing the spread of COVID-19. In this study, we introduce not only a definition for COVID-19 wave of infection, but also reveal some patterns among the spread of COVID-19 in each ASEAN+6 country based on the results of the data-driven analysis. Future research may consider the COVID-19 patterns in this study and compare them with the current COVID-19 situation for further analysis. To refine the results of this study, we aim to use other available datasets to validate this COVID-19 dataset. Alternative wave definitions and clustering algorithms need to be investigated for more reliable and insightful results. Further studies should include deeper investigations on the differences between each cluster including the external factors and evidence that lead to those differences, as well as combining the new cases and fatalities for a joint clustering analysis.

Author Contributions

PS and AT performed the experiments, data analysis, and manuscript preparation. AC were involved in the discussions regarding the experiments. DS was PI, designed the experiments, and was involved in all facets of the study and manuscript preparation. All authors approved the final manuscript.

References

1. Timeline: WHO's covid-19 response. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline. Published 2020. Accessed 2021.
2. Coronavirus disease (covid-19). World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-176hub/q-a-detail/coronavirus-disease-covid-19. Published 2020. Accessed 2021.
3. Alimohamadi Y, Sepandi M, Taghdir M, Hosamirudsari H. Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis. *Journal of Preventive Medicine and Hygiene.* 2020;61:304-312.
4. Singhal S, Kumar P, Singh S, Saha S, Dey A. Clinical features and outcomes of COVID-19 in older adults: a systematic review and meta-analysis. *BMC Geriatrics.* 2021. doi: 10.1186/s12877-021-02261-3.
5. Bennett S, Tafuro J, Mayer J, et al. Clinical features and outcomes of adults with COVID-19: A systematic review and pooled analysis of the literature. *International Journal of Clinical Practice.* 2020.
6. Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *Journal of Medical Virology.* 2020;92(6):632-638.
7. Shadab Far M, Mahsuli M, Sioofy Khoojine A, Hosseini V. Time-variant reliability-based prediction of COVID-19 spread using extended SEIVR model and Monte Carlo sampling. *Results in Physics.* 2021;26:104364.
8. Alanazi S, Kamruzzaman M, Alruwaili M, Alshammari N, Alqahtani S, Karime A.

Measuring and Preventing COVID-19 Using the SIR Model and Machine Learning in Smart Health Care. *Journal of Healthcare Engineering.* 2020;2020.

9. Liu M, Thomadsen R, Yao S. Forecasting the spread of COVID-19 under different reopening strategies. *Scientific Reports.* 2020;10.
10. Cooper I, Mondal A, Antonopoulos CG. A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals.* 2020;139:110057.
11. Zubair M, Asif Iqbal M, Shil A, Haque E, Moshiul Hoque M, Sarker IH. An Efficient K-Means Clustering Algorithm for Analysing COVID-19. in *Hybrid Intelligent Systems.* 2021:422-432.
12. Hutagalung J, Ginantra NLWSR, Bhawika GW, Parwita WGS, Wanto A, Panjaitan PD. COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm. *Journal of Physics: Conference Series.* 2021;1783:012027.
13. Choi YJ, Park M, Park SJ,et al.Types of COVID-19 clusters and their relationship with social distancing in the Seoul metropolitan area, South Korea. *International Journal of Infectious Diseases.* 2021;106:363-369.
14. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases.* 2020;20.
15. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com/CSSEGISand.Data/COVID-19/. Updated 2020. Accessed 2021.
16. Wijaya K, Ganegoda N, Jayathunga Y, Goetz T, Schäfer M, Heidrich P. An epidemic model integrating direct and fomite transmission as well as household structure applied to COVID-19. *Journal of Mathematics in Industry.* 2021;11.
17. Jung Sm, Endo A, Kinoshita R, Nishiura H. Projecting a second wave of COVID-19 in Japan with variable interventions in high-risk settings. *Royal Society Open Science.* 2021;8(3):202169.
18. Ranjan R, Sharma A, Verma MK. Characterization of the Second Wave of COVID-19 in India. *medRxiv.* 2021.
19. Locatelli I, Tr¨achsel B, Rousson V. Estimating the basic reproduction number for COVID-19 in Western Europe. *PLOS ONE.* 2021;16(3): 1-9.
20. Pillonetto G, Bisiacco M, Palù G, Cobelli C. Tracking the time course of reproduction number and lockdown's effect on human behaviour during SARS-CoV-2 epidemic: nonparametric estimation. *Scientific Reports.* 2021;11(1).
21. Koyama S, Horie T, Shinomoto S. Estimating the time-varying reproduction number of COVID-19 with a state-space method. *PLOS Computational Biology.* 2021;17(1):1-18.
22. Arroyo-Marioli F, Bullano F, Kucinskas S, Rondon-Moreno C. Tracking R of COVID-19: A new real-time estimation using the Kalman filter. *PLOS ONE.* 2021;16:1-16.
23. Woods J, Radewan C. Kalman filtering in two dimensions. *IEEE Transactions on Information Theory.* 1977;23(4):473-482.
24. Arroyo-Marioli F, Bullano F, Kucinskas S, Rondon-Moreno C. http://www.globalrt.live/. Updated 2021. Accessed 2021.
25. Malav A, Kadam K, Kamat P. Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology*. 2017;9:3081-3085.
26. Mahajan P, Sharma A. Role of K-Means Algorithm in Disease Prediction. *International Journal of Engineering and Computer Science.* 2016.
27. Bot́ia JA, Vandrovcova J, Forabosco P, et al. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Systems Biology.* 2017;11(1):47.
28. Souto M, Costa I, Araujo D, Ludermir T, Schliep A. Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics.* 2008;9.
29. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing.* 1978;26(1): 43-49.