# Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries


1st Rahmad Kurniawan
*Department of Informatics Engineering*
*Universitas Islam Negeri Sultan Syarif Kasim Riau*
Pekanbaru, Indonesia
https://orcid.org/0000-0002-0957-9480

2nd Siti Norul Huda Sheikh Abdullah
*Faculty of Information Science and Technology*
*Universiti Kebangsaan Malaysia*
Bangi, Selangor, Malaysia
sitinorulhuda@ukm.edu.my

3rd Fitra Lestari
*Department of Industrial Engineering*
*Universitas Islam Negeri Sultan Syarif Kasim Riau*
Pekanbaru, Indonesia
fitralestari@uin-suska.ac.id

4th Mohd Zakree Ahmad Nazri
*Faculty of Information Science and Technology*
*Universiti Kebangsaan Malaysia*
Bangi, Selangor, Malaysia
mzn@ukm.edu.my

5th Akhmad Mujahidin
*Faculty of Sharia and Law*
*Universitas Islam Negeri Sultan Syarif Kasim Riau*
Pekanbaru, Indonesia
ahmadmujahidin@uin-suska.ac.id

6th Noridayu Adnan
*Faculty of Information Science and Technology*
*Universiti Kebangsaan Malaysia*
Bangi, Selangor, Malaysia
noridayu@ukm.edu.my



***Abstract*—An extraordinary outbreak of pneumonia in Wuhan City, China, was subsequently termed as COVID-19 emerged in December 2019. The virus is also known as an infectious disease inherited from a novel coronavirus. This study exposed the beginning of the unprecedented COVID-19 confirmed cases spike exponentially in the United States and 200 countries globally. Epidemiologists usually utilize conventional spread prediction via the classic clustering method. A suspected patient is likely to blow out the disease to a potential agglomerative of cases grouped in place and time. In the era of cutting edge, outbreak prediction can also generate accurate techniques to utilize unsupervised machine learning methods. We apply two prominent unsupervised learning methods, namely K-means clustering and correlation on a set Coronavirus Outbreak COVID-19 data collection dated March 27 and August 16, 2020. The K-means automatically search for unknown clusters of many countries infected with the COVID-19 rapidly. It shows that a group of *m*=5 produces an accuracy of about 97% with [The United States and Italy], [Iran, France], [Spain, German], [Indonesia, Malaysia, Philippine] as clusters. At the same time, it predicts a pertinent relationship between the total deaths and critical patients' attributes of 0.85 while correlating COVID-19 characteristics.**



***Keywords—COVID-19 epidemic, Clustering, Correlations, Machine learning, Infection control***


## I. INTRODUCTION

An unprecedented outbreak of pneumonia of unknown etiology in Wuhan City, Hubei province in China, emerged in December 2019 [1]. A novel coronavirus was identified as the causative agent and was subsequently termed COVID-19 by the World Health Organization (WHO). Regardless of rigorous and stringent global containment and quarantine efforts, the incidence of COVID-19 continues to escalate, with 90,870 laboratory-confirmed cases and over 3,000 deaths worldwide.

Besides, COVID-2019 is multiplying throughout the world and even has dispersed unintendedly to 200 countries in the world [2]. Based on the preliminary study, the probability of mortality in Indonesia continues to rise abruptly, reaching 8.7% in contrast to the preceding five-day report was 8.1% and made the probability of COVID-19 death in Indonesia among the highest spike within Southeast Asia. Alongside Indonesia, the virus has overwhelmed the health care system unintentionally in the Philippine, with a mortality rate of 6.3%. Currently, the researchers solemnly put much effort into predicting the spread and risk of transmitting this deadly virus.

For that reason, the epidemiologists usually practice their theory of cluster dispersion study with regards to predicting the spread of any infectious disease, either an epidemic or pandemic in which it differs in terms of intra-geographically or inter-geographically subsequently. They trace the possibility of infecting from a human to another human via agglomerating the host with their circle of friends underlying the first infected place and gradually adhere to the movement of subsequent levels. The epidemiologists usually predict the virulence of the host agent, mode of transmission, changes made by the susceptibility of the host to the agent as well as environmental factors. Up to date, such information of the data is still not publicized for further investigation by another third party. Additionally, Gilad Edelman in Wired News quoted the shown symptoms among COVID-19 patients reveals that the window to act the coronavirus has already passed [3]. For that reason, such an automatic prediction via artificial intelligence, precisely the machine learning method, is vital to accelerate the current study.

The data sets were employed to analyze the spread of COVID-19 [4]. Auto-Regressive Integrated Moving Average (ARIMA) model prediction was conducted on the Johns Hopkins epidemiological data to predict the epidemiological trend of the prevalence and incidence of COVID-2019. Prediction and Risk analysis of COVID-2019 is urgently needed, particularly in affected countries [5].

Mathematica 10 and SPSS 23 were used for clustering 30 COVID-19 pandemic countries [6]. The active cases per population and areas were stated for data attributes. At the same time, the last data for analysis is April 04, 2020 [6].

However, the COVID-19 confirmed cases spike exponentially in the United States and 200 countries in the world [7]. The progress of the spread and impact of the COVID-2019 occurs dynamically and exponentially in some countries due to late stringent countermeasures such as the United States, Italy, Spain, and Iran. Hence, prediction using the accurate algorithms and monitoring of the longest time is necessary.

Machine learning that utilizes informatics and statistical analysis can improve forecasting, thus widely used to solve specific trained problems. In the beginning, we deployed a popularly unsupervised machine learning method, namely clustering. This well-known clustering method able to agglomerate unlabeled data regardless of the number of attributes [8]. It is also suitable to be used on this underlying open data, namely COVID-2019 data. Due to its effectiveness and simplicity, we utilize the K-means clustering algorithm extensively to determine the infectious attribute correlation across countries worldwide. A recent study used the K-means algorithm to clustering 25 mammals. Although it used small data, it has produced excellent clustering results. The study also compared various methods to determine the best number of groupings. It indicated that the lowest amount of clusters was right [9].

Therefore, this study employs K-means clustering for clustering 200 pandemic countries in the world. Furthermore, risk analysis also needs another method for search unknown knowledge in data. A survey states that the knowledge generated from the correlation matrix is essential [10]. The correlation matrix successfully overcomes the uncertainty problems among related factors. Therefore, both clustering and correlation matrix techniques are applied to see the hidden relationship between components in data. Knowledge of the relationships between attributes is essential to know to determine appropriate prevention and mitigation methods.

## II. Materials and Methods

The data we collected from Worldometers [2], [7], dated March 27 and August 16, 2020. Even though the raw data looks missing, it indicates that no reported cases. Therefore, we have done the data pre-processing task by eliminating the row of missing value. We also have employed to calculate the probability of death in each country.

Finally, these data reveal the spread of COVID-19 conditions in 200 countries in the world comprises of 10 attributes which are *country, total cases, new cases, total deaths, new deaths, total recovered, active cases, serious critical, death probabilities (%) and recovered probabilities (%).* Fig. 1 briefly explains the methodology that has been implemented.

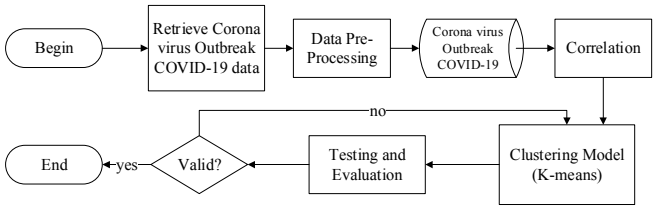


Fig. 1. The methodology of machine learning technique for cluster COVID-19

Eventually, the K-means algorithm to cluster the countries infected by COVID-19. Let the Euclidean distance between $\beta_s$ and $\gamma_t$, $\delta_s$ is the number of COVID-19 data points at the cluster, and $\delta$ is the number of cluster centers, K-means clustering algorithm satisfies $\alpha(\beta, \gamma, \delta)$, an objective function minimization of squared error function $\|\beta_s - \gamma_t\|$ as the following:

$$\alpha(\beta, \gamma, \delta) = \sum_{s=1}^{\delta} \quad \sum_{t=1}^{\delta_s} \quad (\|\beta_s - \gamma_t\|)^2 \qquad (1)$$

Given that $\beta = \{\beta_1, \beta_2, \beta_3, \dots\dots, \beta_n\}; n = 10$ be the set of COVID-19 data points comprising of *country, total cases, new cases, total deaths, new deaths, total recovered, active cases, serious critical, death probabilities, and recovered probabilities* and $\delta = \{\delta_1, \delta_2, \delta_3, \dots., \delta_m\}$ be the set of center points, and the number of the cluster that tested are $m = \{5,10,15,20,30\}$ subsequently. After defining $n$, step 1 randomly selects $\delta$ cluster centers. Then, step 2 estimates the distance between each COVID-19 data point, $\beta_s$and cluster centers, $\gamma_t$. Upon completion, step 3 substitutes the COVID-19 data point, $\beta_s$ to the cluster center, $\gamma_t$ that constitutes the lowest distance, $\alpha$ of the cluster center compared to all the cluster centers, $\delta$. Next, it re-estimates the new cluster center with the average value of $\gamma_s = \frac{1}{\delta_s}\sum_{t=1}^{\delta_s} \beta_s$, and distance between each COVID-19 data point, $\beta_s$ and the new cluster centers, $\gamma_{new}$ at Step 4 and 5 subsequently. At final step 6, it will stop the whole process if there is no COVID-19 data point that was substituted else proceed to Step 3that reason.

## III. Results and Discussion

In this section, we run five experiments, as shown in Table I. We have used the simple K-means algorithm for clustering 200 pandemic countries.

TABLE I. Experiment settings of the K-means clustering algorithm

| **Experiment settings** | **Run 1** | **Run 2** | **Run 3** | **Run 4** | **Run 5** |
|---|---|---|---|---|---|
| **Distance function** | Euclidean distance | Euclidean distance | Euclidean distance | Euclidean distance | Euclidean distance |
| **Initialization method** | Random | Random | Random | Random | Random |
| **Max iteration** | 500 | 500 | 500 | 500 | 500 |
| **Total number of items** | 200 | 200 | 200 | 200 | 200 |
| **Number of clusters, *m*** | 5 | 10 | 15 | 20 | 30 |
| **Cluster number index (Performance)** | 97% | 95% | 92% | 90% | 85% |
| **Within cluster sum of squared errors *(α)*** | 148 | 143 | 137 | 132 | 121 |
| **Time taken** | 0.02 seconds | 0.02 seconds | 0.02 seconds | 0.02 seconds | 0.02 seconds |

Based on Table I, it achieves the best cluster number index (performance) when the number of clusters m = 5. Therefore, the experiment of clustering pandemic countries has been conducted when the number of clusters *m*=5. Table II shows the generated silhouette score to calculate how close an object is to a cluster of its own compared with other clusters

TABLE II. THE EXAMPLE RESULTS OF GENERAL CLUSTERING TEN COUNTRIES WHEN THE NUMBER OF CLUSTERS M=5, A=148 AT RUN 1 (LAST UPDATED: MARCH 27, 2020)

| Country | Total Cases | Total Deaths | Total Recovered | Active Cases | Death Probabilities (%) | Recovered Probabilities (%) | Cluster | Silhouette Score |
|---|---|---|---|---|---|---|---|---|
| **USA** | 85435 | 1295 | 1868 | 82272 | 1.51 | 2.18 | C4 | 0.66 |
| **China** | 81340 | 3292 | 74588 | 3460 | 4.04 | 91.69 | C3 | 0.5 |
| **Italy** | 80589 | 8215 | 10361 | 62013 | 10.1 | 12.85 | C4 | 0.6 |
| **Spain** | 57786 | 4365 | 7015 | 46406 | 7.55 | 12.13 | C2 | 0.65 |
| **Germany** | 43938 | 267 | 5673 | 37998 | 0.60 | 12.91 | C2 | 0.59 |
| **Iran** | 29406 | 2234 | 10457 | 16715 | 7.59 | 35.56 | C5 | 0.71 |
| **France** | 29155 | 1696 | 4948 | 22511 | 5.81 | 16.97 | C5 | 0.7 |
| **Malaysia** | 2031 | 23 | 215 | 1793 | 1.13 | 10.58 | C1 | 0.73 |
| **Indonesia** | 893 | 78 | 35 | 780 | 8.73 | 3.91 | C1 | 0.74 |
| **Philippines** | 707 | 45 | 28 | 634 | 6.36 | 3.96 | C1 | 0.75 |

Based on Table II, it is inevitable that Italy and the United States possess discriminative group characteristics compared to other countries. Similarly, China also isolates independent attributes in contrast to other neighboring and across countries. The rest of the table contents can be seen in our Coronavirus Outbreak COVID-19 Dataset [9].

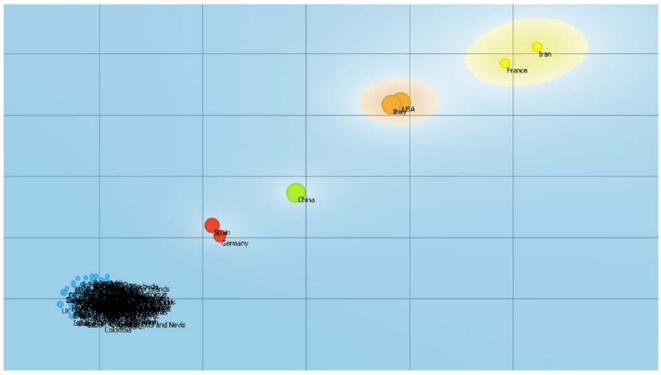


Fig. 2. Clustering of COVID-19 in the world based on the general aspect and m=5, $\alpha$=148, at running 1 (last updated: March 27, 2020)

Fig. 2 shows the general clustering of COVID-19 cases in the world. Those similar colored bubbles indicate countries with similar characteristics. Concerning general aspects such as *total cases, new cases, total deaths, new deaths, total recovered, active cases, acute or critical, death probabilities,* and *recovered probabilities;* thus, France and Iran (yellow bubbles) belong to the same group. Spain and Germany (red bubbles) also surprisingly encapsulate in a similar cluster.

The COVID-19 case in Southeast Asia and other countries with blue bubbles belong to the equivalent group, i.e., they have almost the same characteristics in general. However, Fig. 3 shows Indonesia has the same attributes as Italy, Spain, Iran, the Philippines, and other countries when grouped according to the probability of death. Italy and Spain are the four most positive COVID-19 countries in the world.

Fig. 3 also presents the four quadrants that dictate the countries that share similar characteristics regarding the probability of death. For example, in the top left quadrant, it illustrates minimal distance values between Indonesia and Italy, excluding Paraguay.


Fig. 3. The third cluster countries (C3) through a number of cluster m=5 and $\alpha$=148 at Run 1 based on the probability of death (last updated: March 27, 2020)

We have also conducted experiments to find a correlation between attributes, as shown in Table III. The correlation coefficient values can range between -1 and + 1. The closer it is to + 1 or -1, the closer the two variables are compared to each other—the correlation a negative means a constructive or an inverse relationship. Let the number of COVID-19 cases as *A* and *B* is total deaths, denoting *A'* and *B'* respectively, and standard deviations *S(A)* and *S(B)* correspondingly. The correlation computed as summation from 1 to n of the result (A(z)-A')*.(B(z)-B')* and then dividing this summation by the conclusion *(n-1).S(A).S(B)* where *n* is the total number of instances, and *z* is the increment variable of computation.

TABLE III. A CORRELATION MATRIX EXPLAINS THE RELATIONSHIP OF EACH ATTRIBUTE ON MARCH 27, 2020

| Attributes | Total Cases | New Cases | Total Deaths | New Deaths | Total Recovered | Active Cases | Serious Critical | Death Probability | Recovered Probability |
|---|---|---|---|---|---|---|---|---|---|
| **Total Cases** | 1 | 0.064 | 0.78 | 0.064 | 0.227 | 0.989 | 0.835 | 0.104 | 0.056 |
| **New Cases** | 0.064 | 1 | 0.163 | 0.999 | 0.547 | 0.013 | 0.066 | 0.223 | 0.544 |
| **Total Deaths** | 0.78 | 0.163 | 1 | 0.167 | 0.49 | 0.703 | 0.854 | 0.267 | 0.321 |
| **New Deaths** | 0.064 | 0.999 | 0.167 | 1 | 0.559 | 0.013 | 0.068 | 0.229 | 0.556 |
| **Total Recovered** | 0.227 | 0.547 | 0.490 | 0.559 | 1 | 0.156 | 0.335 | 0.465 | 0.964 |
| **Active Cases** | 0.989 | 0.013 | 0.703 | 0.013 | 0.156 | 1 | 0.776 | 0.074 | -0.001 |
| **Serious Critical** | 0.835 | 0.066 | 0.854 | 0.068 | 0.335 | 0.776 | 1 | 0.195 | 0.157 |
| **Death Probability** | 0.104 | 0.223 | 0.267 | 0.229 | 0.465 | 0.074 | 0.195 | 1 | 0.442 |
| **Recovered Probability** | 0.056 | 0.544 | 0.321 | 0.556 | 0.964 | -0.001 | 0.157 | 0.442 | 1 |

The results indicated a strong positive linear correlation 0.78 at the number of COVID-19 cases in the world with total deaths and 0.85 among the total deaths with patients categorized as critical. The positive value proves the direction of the correlation, i.e., if one of the variables increases, then the other increases too. Furthermore, Fig. 4 verify a strong correlation among the total deaths with patients categorized as critical.

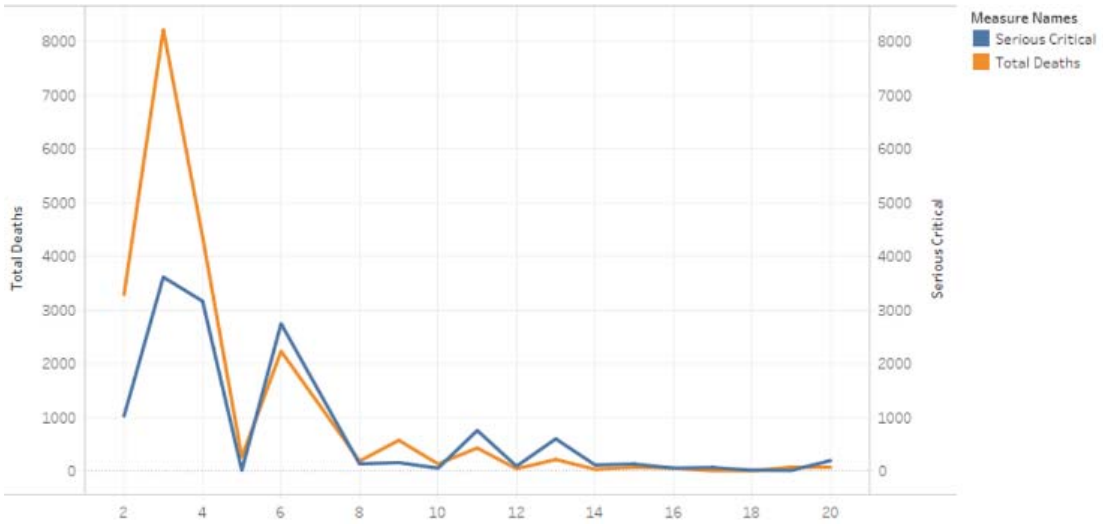


Fig. 4. The correlation graph of total deaths and severe cases in most positive COVID-19 countries (last updated: March 27, 2020)

In addition, we are interested in comparing the results of the analysis on March 27, 2020, with the current analysis using data on August 16, 2020 [2], [7]. Data on August 16, 2020, shows the unexpected results were obtained that the risk of COVID-19 has increased in general. The results indicated a stronger positive linear correlation 0.94 at the number of COVID-19 cases in the world with total deaths and 0.91 among the total deaths with patients categorized as critical.

However, based on the data [7], the probability of mortality in Indonesia decreases, reaching 4.4%, in contrast to the preceding March 27, 2020 report was 8.7%. Furthermore, the recovered probability in Indonesia continues to rise abruptly, attainment of 66.7%.

## IV. CONCLUSIONS

We have successfully used K-means clustering to search hidden or unknown clusters of many countries infected with the COVID-19 rapidly and correlation matrix to determine the relationships among a set of attributes. K-means and correlation matrix was used to analyze COVID-19 risk in Pandemic Countries. This study exposed the beginning of the unprecedented COVID-19 confirmed cases spike exponentially in the United States and 200 countries globally. Based on five months of observation in pandemic countries in the world, correlation matrix results still indicated a strong positive correlation among the total deaths with patients categorized as critical. This relation means that many patients who are classified as critical have not recovered.

Conversely, the probability of mortality in a country, i.e., Indonesia, continues to decrease, and the recovered probability in Indonesia continues to rise abruptly.

The results are based on existing data from a source. Future work is suggested to use data from various reliable and integrated sources to overcome missing values.

We are hoping these results as the knowledge that can assist the Center for Control Disease (CDC) in critical decision making such as precise mitigation response, countermeasures, and infection control from time to time.

## ACKNOWLEDGMENT

Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau and Universiti Kebangsaan Malaysia support this work.

## REFERENCES

[1] C. Sohrabi *et al.*, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *Int. J. Surg.*, vol. 76, no. February, pp. 71–76, 2020, DOI: 10.1016/j.ijsu.2020.02.034.

[2] Worldometers.info, "COVID-19 Coronavirus Outbreak," *Dadax*, 2020. https://www.worldometers.info/coronavirus (accessed March 27, 2020).

[3] G. Edelman, "Congress Needs to Get Its Act Together on the Coronavirus," *wired*, 2020. https://www.wired.com/story/congress-needs-to-get-act-together-coronavirus (accessed March 12, 2020).

[4] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset," *Data Br.*, vol. 29, p. 105340, 2020, doi: 10.1016/j.dib.2020.105340.

[5] Z. Zheng, Y. Huyan, H. Li, S. Sun, and Y. Xu, "Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan," *Sensors Actuators B. Chem.*, p. 127065, 2019, DOI: 10.1016/j.snb.2019.127065.

[6] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas, "Clustering analysis of countries using the COVID-19 cases dataset," *Data Br.*, vol. 31, p. 105787, Aug. 2020, DOI: 10.1016/j.dib.2020.105787.

[7] Worldometers.info, "Coronavirus Update (Live): 21,842,782 Cases and 773,279 Deaths from COVID-19 Virus Pandemic - Worldometer," *Dadax*, 2020. https://www.worldometers.info/coronavirus/ (accessed Aug. 16, 2020).

[8] R. Baruri, A. Ghosh, R. Banerjee, A. Das, A. Mandal, and T. Halder, "An Empirical Evaluation of k-Means Clustering Technique and Comparison," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 470–475, 2019, DOI: 10.1109/COMITCon.2019.8862215.

[9] Matt, "10 Tips for Choosing the Optimal Number of Clusters," *Towards Data Science*, 2019. https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92 (accessed April 16, 2020).

[10] A. dos Santos and R. Diniz, "The correlation matrix for the effective delayed neutron parameters of the IPEN/MB-01 reactor," *Ann. Nucl. Energy*, vol. 136, p. 107008, Feb. 2020, DOI: 10.1016/j.anucene.2019.107008.