Thailand Statistician
July 2024; 22(3): 594-609
http://statassoc.or.th
Contributed paper

Using k-means Clustering to Confirm Provincial COVID-19 Cases during the Omicron Epidemic in Thailand

Worrawate Leela-apiradee*, Sathinee Wareepornthep, Chutikan Premboon and Narawut Usaman
Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Pathum Thani, Thailand
*Corresponding author; e-mail: worrawate@mathstat.sci.tu.ac.th

Received: 28 April 2023
Revised: 18 September 2023
Accepted: 3 April 2024

Abstract

The Novel Coronavirus 2019 (COVID-19) pandemic has infected and killed millions of people internationally. This work uses k-means clustering and a time series k-means algorithm to present an overview of cases and deaths from COVID-19 in grouped provinces of Thailand before entering the post-pandemic period on 1 July 2022. The study is divided into two parts: the first uses k-means clustering with Euclidean distance measure to analyze confirmed cases and deaths per 100,000 population by province that cumulated from 1 January 2022 to 30 June 2022, during the Omicron (B.1.1.529) outbreak. Based on the elbow method, optimal numeric value for clusters (groups of provinces) is $k = 5$. The second cluster, consisting of two provinces: Phuket, and Samut Sakhon, is reached the highest cluster mean of the confirmed cases and deaths. We investigate the linear relationship between the confirmed cases (deaths) and 12 different feature variables associated with social, economic, health and environmental factors. Pearson correlation analysis indicates four feature variables correlated positively with confirmed cases and deaths: Gross Regional and Provincial Product (GPP) per capita; number of medical personnel per 100,000 population (pop.); average monthly household income; and number of dengue cases per 100,000 pop. In the second part, k-means clustering with dynamic time warping distance measure is applied to time series data, namely daily confirmed cases per 100,000 people by province gathered during the same time interval as the first part for 181 days, with optimal cluster number being $k = 3$. The time series of infections attained its apogee in the third cluster, consisting of three provinces: Phuket, Samut Songkhram, and Samut Sakhon. In addition, these findings provide a record of the COVID-19 pandemic in Thailand during the first half of 2022, as illustrated in choropleth maps, for potential governmental use of these provincial groupings for future public health service budget allocation decisions related to the COVID-19 pandemic.

Keywords: Dynamic time warping, k-means clustering, pearson correlation coefficient, time series k-means.

1. Introduction

COVID-19 is an infectious disease caused by the novel coronavirus. The outbreak began in December 2019 in Wuhan located in the Hubei province of central China. As of 10 March 2022 at 12:30 p.m., there were 22,984 new cases in Thailand that continues to increase day by day. With the outbreak of the Omicron (B.1.1.529) a fast-spreading coronavirus variant, most people infected with

the virus will have mild symptoms. However, the number of deaths increased by 74, while the cumulative number of confirmed cases and deaths since 2020 totaled 3,111,857 and 23,512, respectively, reported by Department of Disease Control, Ministry of Public Health of Thailand. On the same day, National Communicable Disease Committee (NCDC) approved the plan to label COVID-19 as an endemic disease in Thailand starting on 1 July 2022. Due to the resolution was announced on 18 March 2022 by the Royal Thai Government that the nationwide 77 provinces were adjusted into three color zones of COVID-19: control (orange) zones, close surveillance (yellow) zones, and tourism pilot (blue) zones. The latest announcement was eased into blue and green zones from 1 June 2022 onwards.

Throughout the COVID-19 pandemic, data mining techniques have been applied to analyze insight patterns, especially cluster analysis using k-means clustering to understand the epidemic situation in different countries. In Rizvi et al. (2021), Rizvi et al. used k-means algorithm to cluster 79 countries based on 18 different feature variables (factors) consisting of socio-economic, disease prevalence, and environmental indicators together with cumulative death cases and cumulative confirmed cases. They obtained the factors closely involved in COVID-19 spread by performing Pearson correlation analysis. The confirmed cases of COVID-19, and April 2020 deaths in Southeast Asia were grouped in Hutagalung et al. (2021) into three clusters as high, medium and low by utilizing k-means clustering method. According to COVID-19 data of confirmed, death, and recovered cases on 19 April 2020, Abdullah et al. (2022) studied the clustering of provinces in Indonesia conducted with the same method. Zubair et al. developed in Zubair et al. (2020) an efficient method of k-means clustering to analyze health care quality cluster of countries on the basis of COVID-19 datasets with high dimension, which could be applied principal component analysis to convert it into two dimensions. The initial centroids of k-means came from mean calculation after splitting the dataset into k equal parts using percentile concepts. With this method, the number of iterations and execution time were reduced when comparing with the traditional one.

Time series clustering plays a crucial role during the pandemic investigated in many researches so that it would be able to understand the patterns and design an effective policies to control and mitigate the effects of COVID-19. The paper Cerqueti and Ficcadenti (2022) combined k-means cluster analysis based on Euclidean distance and rank-size approach to investigate selected countries with a high level for the Healthcare Access and Quality Index. The high-quality goodness-of-fit parameters were attained from the approach of third-degree polynomial type that could reduce biases from the data collection phase by sorting the COVID-19 new deaths per million before ranking them. A k-means method was applied to the time series clustering of COVID-19 cases in each prefecture of Japan presented in Watanabe (2022). A clustering similarity related to COVID-19 pandemic of Italian regions was examined in publications Mattera (2022) and DUrso et al. (2022). Optimally weighted spatial and temporal dimensions of the pandemic, Mattera (2022) showed that a fuzzy clustering model could identify two groups of Italian regions with similar spread. An Exponential distance based Fuzzy c-medoids clustering algorithm on the basis of B-splines with spatial penalty term was applied in DUrso et al. (2022) to classify the spline representation over time observed from the last week of February 2020 to the first week of February 2021. The Exponential distance and B-splines are beneficial to data reduction and robustness, respectively. The c-means and fuzzy c-means algorithms were implemented in Afzal et al. (2021). The confirmed daily COVID-19 cases, recoveries, and deaths as available from January 2020 with respect to location of the globe were clustered into three main clusters whose centroids were located in the US, Brazil, and India, respectively. In accordance with data from 18 March 2020 – 18 February 2021 in weekly, the COVID-19 confirmed death trajectories in European countries including 145 regions of France, Germany, Italy, Spain, Switzerland and United Kingdom were grouped by Bucci et al. Bucci et al. (2022) into 12 clusters using a Bayesian nonparametric approach based on mixture of Gaussian processes coupled with Dirichlet process.

Most clustering algorithms utilize Euclidean distance for similarity computation. However, it may not be suitable for time series measure as a result of the one-to-one point comparison (see Keogh

and Ratanamahatana (2016) in more details). To give more flexible and robust distance measure, *Dynamic Time Warping (DTW)* determines an optimal many-to-one (and vice versa) matching between two different length trajectories, which can be found in science, medicine, industry and finance researches. The collected data of driver behavior patterns over time with four indicators as speed, acceleration, yaw rate, and sideslip angle of drivers was clustered in Yao et al. (2021) using DTW and hidden Markov model. Lee et al. studied in Lee et al. (2018) a time-series clustering approach based on DTW algorithm to measure the similarity of train doors by their temporal passenger service patterns in the busiest metro stations of Seoul, Korea. A hybrid technique for fuzzy clustering of time series using DTW distance that merges fuzzy c-means and fuzzy c-medoids clustering was developed in Izakian et al. (2015). Moreover, DTW based algorithms such as an adaptive constrained DTW, minimizing DTW, partial DTW, and weighted DTW have been proposed in the literature Cai et al. (2021); Jeong et al. (2011); Li et al. (2020, 2022); Sivaraks et al. (2015).

According to Ampornphan (2021), Ampornphan used k-means clustering to classify group of COVID-19 infection cases in Thailand from December 2020 to May 2021 as seven clusters according to the age range from 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, and 60 years or over, each of which had associated patterns determined by association rule mining. Of course, the outbreak in Thailand has a rising of infections in Bangkok and its vicinity and industrial areas. This paper examines cumulative confirmed case (the first dimension) and death (the second dimension) rates from COVID-19 pandemic per a hundred thousand people by province in Thailand during the Omicron wave from 1 January 2022 till the end of June 2022 grouping those two-dimensional data into an appropriate number of clusters by using k-means clustering. We select 12 feature variables as relevant to social, economic, health and environmental factors to test statistical relationship of Pearson correlation coefficient between the variables and the rates. Furthermore, the daily rates for confirmed COVID-19 cases per a hundred thousand people by province during the wave are investigated and clustered those trajectories data by developing time series k-means approach based on DTW measure. The provincial groups are represented as choropleth maps shown in Section 3.

The rest of the paper is organized as follows. We first describe background knowledge used in our work in Section 2 that includes k-means clustering, Pearson correlation coefficient, dynamic time warping, and time series clustering using k-means. Methodology and analysis of clustering results are presented in Section 3. The conclusion of this article is addressed in the final section.

2. Preliminaries

Throughout the paper, we let m and n be natural numbers. This section provides the details of background knowledge related to our methodology: k-means clustering, Pearson correlation coefficient, dynamic time warping, and time series clustering using k-means, which will be performed to cluster the provinces in the next section. In subsection 2.3, we add a binary integer linear program to calculate the dynamic time warping distance instead of using the traditional approach.

2.1. K-means clustering

Clustering is an unsupervised machine learning technique used to group data based on similarity of each data point by assigning similar data to the same cluster and dissimilar ones to other clusters. Assuming that the similarity between two data points $x = (x_1, x_2, \ldots, x_m)$ and $y = (y_1, y_2, \ldots, y_m)$ in $\mathbb{R}^m$ is measured by Euclidean distance, which can be computed as

$$d_E(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}.$$

In 1967, the most popular method of clustering named k-means clustering was first introduced by MacQueen in MacQueen (1967). As a result of being an exclusive method that can precisely

assign all data points to be in a particular cluster, it is commonly used in many fields nowadays. At the beginning of k-means algorithm, we need to find out the optimal number of clusters k by varying values of k from 1 to 10 and calculating the sum of the square distance between points in a cluster and the cluster centroid (Within-Cluster Sum of Square: WCSS) for each k. When a graph between the WCSS on y-axis and the values of k along x-axis is plotted, the optimal one will be found from selecting k at which distortion (or inertia) decreases look like "elbow" of the graph, i.e., it starts decreasing as a linear fashion seen for example in Figures 2 (right) and 7.

Let $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ be given n data points in m-dimensional space. The k-means algorithm would be written as an optimization problem as of the form

$$P_1: \text{ minimize } P(U,z) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}(d_E(x^{(i)}, z^{(j)}))^2$$

$$\text{subject to } \sum_{j=1}^{k} u_{ij} = 1, \ \ \forall i = 1, 2, \ldots, n,$$

where $z^{(j)} \in \mathbb{R}^m$ is a vector of variables representing a centroid of the j-th cluster. The term U is a binary matrix containing decision variables u_{ij}s. If $u_{ij} = 1$, then a point $x^{(i)}$ is assigned to cluster j; otherwise, it is not assigned to cluster j.

2.2. Pearson correlation coefficient

As known from the previous subsection, the cluster analysis is used to identify similar groups of objects in order to form clusters such as groups of customers from their shopping habits, groups of travelers related to their travel behavior, groups of patients according to symptoms or severity of a particular disease, etc. After that the concept of correlation is used to analyze relationships within a group of clustered data.

The correlation is a measure of an association between two feature variables $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. Pearson correlation coefficient is used to measure the association in terms of the linear relationship. It can be calculated by the formula (1), which assigns a score between -1 and 1.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the mean of x and y, respectively. The score r_{xy} is called **the sample correlation coefficient** interpreted as the table below.

Value of r_{xy}	Interpretation
It approaches 1 or -1.	(very) strong correlation
It approaches 0.	negligible or weak correlation
It is more than 0.	positive correlation
It equals to 0.	no linear relationship
It is less than 0.	negative correlation

To evaluate the correlation coefficient, references Asuero et al. (2006); Best and Kahn (2016); Ratner (2009); Schober et al. (2018) presented the strength of correlation split into several stratifications in different criteria. The coefficient r_{xy} is our estimate of the (unknown) population correlation coefficient denoted by ρ. To perform the hypothesis test for ρ, we state the null hypothesis $H_0 : \rho = 0$ against the two-tailed alternative hypothesis $H_A : \rho \neq 0$. For given a significance level α, we can make a decision as follows.

- If p-value (Significance: Sig.) is less than (or equal to) α, we reject the null hypothesis in favor of the alternative one. This is called **statistically significant**.
- If p-value (Significance: Sig.) is greater than α, we fail to reject the null hypothesis.

2.3. Dynamic time warping

A time series is a set of observations recorded over consecutive periods in sequential order. The time series consists of four main components: (i) trend, (ii) seasonal variations, (iii) cyclical variations, and (iv) irregular variations. We have collected data on the daily number of COVID-19 cases by province during the first six months of 2022, which contains 77 time series data in total. The measurement of similarity between two time series is a key component in cluster analysis. Readers can see more details in a survey article Liao (2005).

Given X and Y be two time series as $X = (x_1, x_2, \ldots, x_m)$ of length m and $Y = (y_1, y_2, \ldots, y_n)$ of length n. The similarity measure between them in the form of Euclidean distance for the case only when $m = n$ is defined by

$$d_E(X, Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}.$$

In this paper, we apply a matching algorithm for measuring similarity between two time series with different length called **Dynamic Time Warping (DTW)**, which is a widely used to accurately measure the distance of time series X and Y denoted here as $d_{DTW}(X, Y)$. A mapping that minimizes the distance between two time series is called a **warping path** $W = (w_1, w_2, ..., w_K)$ where $\max\{m, n\} \leq K \leq m + n - 1$.

Let us denote a symbol $\sum_{k=1}^{K} w_k$ as cumulative distance according to W. To find such a warping path, we need to create an $m \times n$ matrix as illustrated in Table 1 below.

Table 1 Accumulated cost matrix whose entries are obtained by (2).

Row Index	Time Series X					
m	x_m	$c(m,1)$	$c(m,2)$	$\cdots$	$c(m,n)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
2	x_2	$c(2,1)$	$c(2,2)$	$\cdots$	$c(2,n)$	
1	x_1	$c(1,1)$	$c(1,2)$	$\cdots$	$c(1,n)$	
		y_1	y_2	$\cdots$	y_n	Time Series Y
		1	2	$\cdots$	n	Column Index

Notice that the row indices are rearranged from the bottom to the top. The (i, j)-entry evaluated as a distance $c(i, j)$ can be calculated by a dynamic programming approach as

$$c(i, j) = (x_i - y_j)^2 + \min\{c(i-1, j-1), c(i-1, j), c(i, j-1)\} \tag{2}$$

for each $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$ together with setting the initial conditions as

$$c(0, 0) = 0 \tag{3}$$

$$c(i, 0) = \infty, \quad \forall i = 1, 2, \ldots, m \tag{4}$$

$$c(0, j) = \infty, \quad \forall j = 1, 2, \ldots, n. \tag{5}$$

When $(i, j) = (1, 2)$, the formula (2) becomes

$$\begin{aligned} c(1, 2) &= (x_1 - y_2)^2 + \min\{c(0, 1), c(0, 2), c(1, 1)\} \\ &= (x_1 - y_2)^2 + \min\{\infty, \infty, c(1, 1)\} \qquad \text{(by (4) and (5))} \\ &= (x_1 - y_2)^2 + c(1, 1). \end{aligned} \tag{6}$$

Similarly, if $(i,j) = (2.1)$, then we also get

$$c(2,1) = (x_2 - y_1)^2 + c(1,1). \tag{7}$$

It is easy to see that

$$c(1,1) = (x_1 - y_1)^2. \tag{8}$$

Let us next state the following three conditions for the warping path $W = (w_1, w_2, \ldots, w_K)$.

1. Boundary condition: The warping path starts at the bottom left corner and ends at the top right corner of the table, i.e., $w_1 = (1,1)$ and $w_K = (m,n)$.
2. Monotonicity condition: If $w_k = (m_k, n_k),\ \forall k \in \{1,2,\ldots,K\}$, then
$$m_1 \leq m_2 \leq \cdots \leq m_K \ \text{ and } \ n_1 \leq n_2 \leq \cdots \leq n_K.$$
3. Continuity condition: $w_{k+1} - w_k \in \{(1,0),(0,1),(1,1)\},\ \forall k \in \{1,2,\ldots,K-1\}$.

Here, w_k represents an entry of the table written as an ordered pair for each $k \in \{1,2,\ldots,K\}$. Moreover, $W^* = (w_1^*, w_2^*, \ldots, w_K^*) = ((m_1^*, n_1^*), (m_2^*, n_2^*), \ldots, (m_K^*, n_K^*))$ is said to be the **optimal warping path** if $\sum_{k=1}^{K} w_k^*$ is minimized. For convenience, we write x_k^* and y_k^* instead of $x_{m_k^*}$ and $y_{n_k^*}$ for each $k \in \{1,2,\ldots,K\}$, respectively. Then, we obtain $d_{DTW}(X,Y)$ from computing the Euclidean distance between two sequences $(x_1^*, x_2^*, \ldots, x_K^*)$ and $(y_1^*, y_2^*, \ldots, y_K^*)$, i.e.,

$$d_{DTW}(X,Y) = \sqrt{\sum_{k=1}^{K} w_k^*} = \sqrt{\sum_{k=1}^{K} (x_k^* - y_k^*)^2}. \tag{9}$$

In other words, the DTW is equivalent to minimizing Euclidean distance between time series under all possible matching. This paper develops a simple binary integer linear program for solving the DTW as shown in the theorem below.

Theorem 1 *Let $X = (x_1, x_2, \ldots, x_m)$ and $Y = (y_1, y_2, \ldots, y_n)$ be any two time series. Then, $d_{DTW}(X,Y) = \sqrt{P(V^*)}$ where $P(V^*)$ is the optimal value of Problem P_2 expressed as*

$$P_2: \ \text{minimize} \ \ P(V) = \sum_{i=1}^{m}\sum_{j=1}^{n} (x_i - y_j)^2 v(i,j) \tag{10}$$

$$\text{subject to} \ \ v(i,1) \leq v(i-1,1), \ \ \forall i = 2,3,\ldots,m \tag{11}$$

$$v(1,j) \leq v(1,j-1), \ \ \forall j = 2,3,\ldots,n \tag{12}$$

$$v(i,j) \leq v(i-1,j) + v(i,j-1) + v(i-1,j-1), \quad \forall i = 2,3,\ldots,m, \ \forall j = 2,3,\ldots,n \tag{13}$$

$$v(m,n) = 1 \tag{14}$$

$$v(i,j) \in \{0,1\}, \ \ \forall i = 1,2,\ldots,m, \ \forall j = 1,2,\ldots,n. \tag{15}$$

Proof: Let $v(i,j)$ be a binary decision variable such that

$$v(i,j) = \begin{cases} 1, & \text{if } (i,j) \in W; \\ 0, & \text{if } (i,j) \notin W, \end{cases}$$

for each $i \in \{1,2,\ldots,m\}$ and $j \in \{1,2,\ldots,n\}$. The continuity for $k \in \{1,2,\ldots,K-1\}$ can be replaced by (11)–(13). For $k = 1$, we have

$$w_2 - w_1 \in \{(1,0),(0,1),(1,1)\} \ \text{ and } \ w_1 = (1,1) \in W$$

Table 2 Accumulated cost matrix for Kanchanaburi and Chachoengsao series data.

Row Index	Kanchanaburi						
5	$x_5 = 29$	26	18	19	14	586	
4	$x_4 = 26$	17	25	25	10	739	
3	$x_3 = 29$	17	9	9	13	589	
2	$x_2 = 28$	8	8	12	13	638	
1	$x_1 = 28$	4	8	12	13	638	
		$y_1 = 26$	$y_2 = 30$	$y_3 = 30$	$y_4 = 27$	$y_5 = 53$	Chachoengsao
		1	2	3	4	5	Column Index

which implies $w_2 \in \{(2,1),(1,2),(2,2)\}$. From (11) – (13),

$$v(2,1) \leq v(1,1), \;\; v(1,2) \leq v(1,1) \;\text{ and }\; v(2,2) \leq v(1,2) + v(2,1) + v(1,1). \tag{16}$$

If $w_1 = (1,1)$, i.e., $v(1,1) = 1$ then (16) provides $w_2 \in \{(2,1),(1,2),(2,2)\}$. Let us next distinguish into three cases below in order to see that the boundary condition $v(1,1) = 1$ is obtained from (16) no matter what w_2 is.

Case I: When $w_2 = (2,1)$, the first inequality of (16) gives $v(1,1) = 1$.

Case II: When $w_2 = (1,2)$, the condition $v(1,1) = 1$ holds from the second inequality of (16).

Case III: For the case when $w_2 = (2,2)$, if $w_1 = (1,2)$, then $v(2,1) = 0 = v(1,1)$, which contradicts with the second inequality of (16). However, if $w_1 = (2,1)$, there is a contradiction in a similar reason. Therefore, $w_1 = (1,1)$ is guaranteed.

According to the boundary condition $w_K = (m,n)$, (14) yields. It is obvious from the objective function (10) that the square root of its optimal value becomes the DTW distance between X and Y as desired.

The following examples show how to determine the optimal warping path by using the dynamic programming approach in Example 1. We apply Theorem 1 in Example 2 to emphasize that the proposed binary integer linear program can find out the DTW distance.

Example 1 The number of confirmed cases of COVID-19 in daily of Kanchanaburi and Chachoengsao provinces from 1 - 5 January 2022 ($m = n = 5$ days) is shown in the table below.

Date	Kanchanaburi	Chachoengsao
1-Jan-22	28	26
2-Jan-22	28	30
3-Jan-22	29	30
4-Jan-22	26	27
5-Jan-22	29	53

According to Eqn. (2), we provide Table 2 in order to achieve the warping path based on its conditions.

Therefore,

$$W^* = ((1,1),(2,2),(3,3),(4,4),(5,5))$$

and

$$d_{DTW}(X,Y) = \sqrt{\sum_{k=1}^{K} w_k^*} = \sqrt{586} = 24.2074,$$

which equals to $d_E(X,Y)$ for this example.

Example 2 The following table records the number of confirmed cases of COVID-19 in daily of Trat and Nakhon Phanom provinces from 1 - 16 January 2022.

Date	Trat (X)	Nakhon Phanom (Y)
1-Jan-22	$x_1 = 4$	$y_1 = 7$
2-Jan-22	$x_2 = 7$	$y_2 = 5$
3-Jan-22	$x_3 = 8$	$y_3 = 2$
4-Jan-22	$x_4 = 6$	$y_4 = 7$
5-Jan-22	$x_5 = 13$	$y_5 = 12$
6-Jan-22	$x_6 = 50$	$y_6 = 30$
7-Jan-22	$x_7 = 61$	$y_7 = 25$
8-Jan-22	$x_8 = 75$	$y_8 = 40$
9-Jan-22	$x_9 = 51$	$y_9 = 59$
10-Jan-22	$x_{10} = 44$	$y_{10} = 49$
11-Jan-22	$x_{11} = 27$	$y_{11} = 47$
12-Jan-22	$x_{12} = 72$	$y_{12} = 46$
13-Jan-22	$x_{13} = 34$	$y_{13} = 52$
14-Jan-22	$x_{14} = 41$	$y_{14} = 49$
15-Jan-22	$x_{15} = 33$	$y_{15} = 48$
16-Jan-22	$x_{16} = 13$	$y_{16} = 58$

To accomplish the DTW distance between X and Y, we solve Problem P_2 with $m = 16 = n$ through IBM Decision Optimization CPLEX Modeling for Python, which is also known as "DOcplex" library. Hence, it provides the optimal value as $P(V^*) = 4108$. Based on Theorem 1, we concludes

$$d_{DTW}(X, Y) = \sqrt{P(V^*)} = \sqrt{4108} = 64.0937.$$

For promotional version in Python, the library limits according to the size of problem, which does not exceed 1000 variables and 1000 constraints. Note that Problem P_2 in this example has $m \times n = 16 \times 16 = 256$ variables. Moreover, we can illustrate in Figure 1 alignments of X and Y.

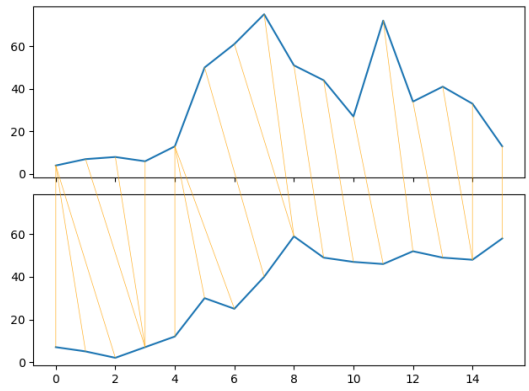


Figure 1 Alignments of daily confirmed cases in Trat and Nakhon Phanom from 1 - 16 January 2022.

The scales $0, 1, 2, \ldots, 15$ on the horizontal axis represent the date 1-Jan-22, 2-Jan-22, 3-Jan-22, and so on, while the vertical axis shows the number of cases in Trat (above) and Nakhon Phanom (below).

2.4. Time series clustering using k-means

Given $X = \{X_1, X_2, \ldots, X_n\}$ be a set of n time series such that X_i has length m written as $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ for each $i = 1, 2, \ldots, n$. Denote U as an $n \times k$ dimensional matrix, whose entries are binary variables defined as follow. When $u_{ip} = 1$ ($u_{ip} = 0$), a time series data X_i is assigned (not assigned) to cluster p. The time series k-means algorithm with DTW distance measure can be expressed as the following optimization problem:

$$P_2: \text{ minimize } P(U,Z) = \sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ip}(d_{DTW}(x_{ij}, z_{pj}))^2$$

$$\text{subject to } \sum_{p=1}^{k} u_{ip} = 1, \quad \forall i = 1, 2, \ldots, n.$$

Here, $u_{ip} \in \{0, 1\}$, while a set of variable vectors $Z = \{Z_1, Z_2, \ldots, Z_k\}$ indicates the centroids of k clusters where $Z_p \in \mathbb{R}^m$ for each $p = 1, 2, \ldots, k$. The problem P_3 could be modified as a smooth subspace clustering according to Huang et al. (2016) using weighted time stamps.

2.5. Methodology and results

We start with gathering data on the number of COVID-19 cases per day by province in Thailand from 1 January 2022 to 30 June 2022 during the first six months of the wave of Omicron variant, which is the dominant variant around the world nowadays. The data on the wave was reported daily by Department of Disease Control, Ministry of Public Health (Thailand). Before taking that information for the first part analysis, the gathered cases need to be cumulated as of 30 June 2022 by province and do it as an incidence rate per 100,000 population (pop.) by the following calculation

$$\text{Incidence rate per 100K pop. in a province} = \frac{\text{Number of cumulative cases in the province}}{\text{Pop. size of the province}} \times 100,000, \quad (17)$$

which indicates the pace at which cases occur in a population by province at the period of time. In addition, the number of deaths is performed in the same way. The efficient tool implemented for k-means clustering in this part comes from a library for machine learning named "Scikit-learn" in Python. Using the elbow method, a two-dimensional plot is depicted in Figure 2 (right) with WCSS in the y-axis and the number of clusters k in the x-axis. As seen in the figure, the value of k at which the WCSS starts decreasing in a linear function is $k = 5$. Therefore, we set 5 to be the optimal number of clusters for 77 provinces.

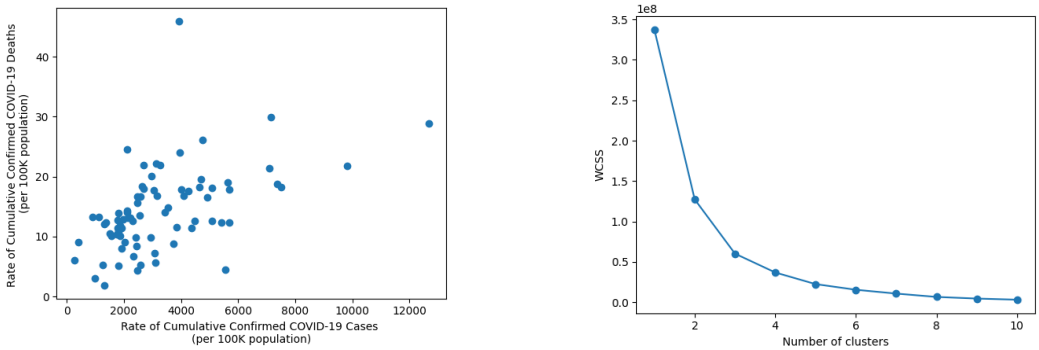


Figure 2 A graphical representation of finding the optimal number of clusters using elbow method (right) according to the dataset (left).

The k-means algorithm provides 5 clusters in total from those provinces presented as Figure 3, each of cluster contains which of the following lists below.

Cluster 1: This consists of 24 provinces such as Krabi, Kanchanaburi, Kalasin, Kamphaeng Phet, Khon Kaen, Chanthaburi, Chumphon, Nakhon Sawan, Nan, Bueng Kan, Buriram, Pathum Thani, Phitsanulok, Maha Sarakham, Yasothon, Roi Et, Loei, Songkhla, Satun, Sa Kaeo, Saraburi, Sukhothai, Surin, and Uthai Thani.

Cluster 2: This consists of 2 provinces such as Phuket, and Samut Sakhon.

Cluster 3: This consists of 19 provinces such as Chachoengsao, Trat, Nakhon Nayok, Nakhon Pathom, Nakhon Si Thammarat, Nonthaburi, Prachuap Khiri Khan, Prachinburi, Phra Nakhon Si Ayutthaya, Phang Nga, Phatthalung, Phetchaburi, Ranong, Rayong, Ratchaburi, Sing Buri, Suphan Buri, Nong Khai, and Ang Thong.

Cluster 4: This consists of 28 provinces such as Chai Nat, Chaiyaphum, Chiang Rai, Chiang Mai, Trang, Tak, Nakhon Phanom, Nakhon Ratchasima, Narathiwat, Pattani, Phayao, Phichit, Phetchabun, Phrae, Mukdahan, Mae Hong Son, Yala, Lopburi, Lampang, Lamphun, Sisaket, Sakon Nakhon, Surat Thani, Nong Bua Lamphu, Amnat Charoen, Udon Thani, Uttaradit, and Ubon Ratchathani.

Cluster 5: This consists of 4 provinces such as Bangkok, Chonburi, Samut Prakan, and Samut Songkhram.

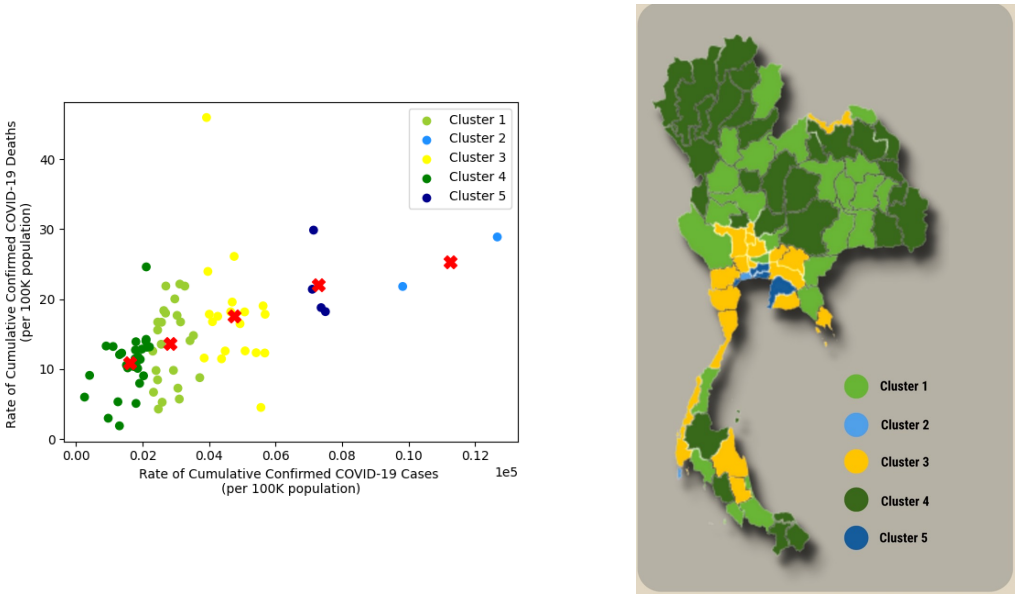


Figure 3 Clusters of provinces after using k-means clustering (left) and choropleth maps representation (right) colored as the clusters.

From the plots in Figure 4, it can be interpreted that Clusters 2 and 4 are reached the highest and the lowest cluster means of the cumulative cases and deaths as of 30 June 2022, respectively. The second cluster comprises of only two provinces: Phuket, and Samut Sakhon. The first province was open to fully vaccinated, unquarantined foreign travelers following the governmental Phuket Sandbox program. The second one is Thailand's fishing and seafood processing capital with many Burmese migrant workers. These contexts may explain why infection were on high level.

To see more insights of the clustered data, we then select twelve feature variables listed in Table 3 with the most recent data available in *2016, *2019, $^{\circ}$2020 and $^{\diamond}$2021.

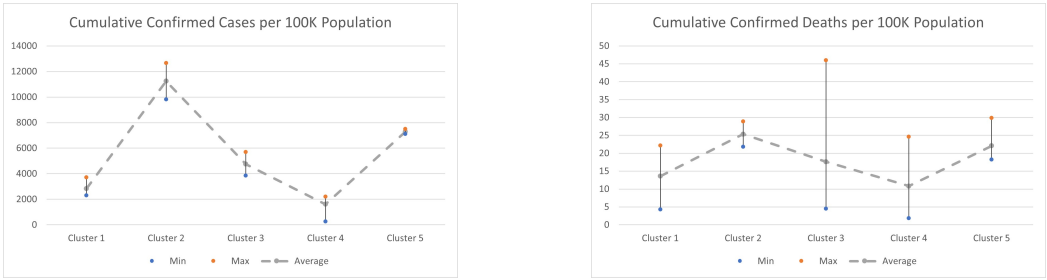


Figure 4 Average-Max-Min charts of the COVID-19 cases and deaths by cluster.

Table 3 Feature variables associated with social, economic, health and environmental factors by province, where the GPP stands for Gross regional and Provincial Product. These datasets are available in open government data of Thailand.

Social factors	Economic factors
Population size$^{\diamond}$	GPP per capita*
Unemployment rate$^{\circ}$ (%)	Number of establishments*
Birth rate$^{\circ}$ (per 100K pop.)	Average monthly household income* (Baht)
Death rate$^{\circ}$ (per 100K pop.)	Average household debt* (Baht)
Health factors	**Environmental factors**
Number of medical personnel$^{\circ}$ (per 100K pop.)	Volume of solid waste$^{\circ}$ (ton per day)
Cases of dengue* (per 100K pop.)	Percentage of forest area$^{\circ}$

Since we do not know obviously which variables are correlated to the cumulative cases and deaths, the hypothesis test for population correlation coefficient ρ is applied in this paper to learn of a linear association between the cumulative ones and those variables. At the significance level $\alpha = 0.05$, the output displayed the correlations in Table 4 tells us that there are four features: GPP per capita, number of medical personnel, average monthly household income, and cases of dengue, which are positive correlated with the cumulative cases and deaths significantly.

Therefore, we can describe in details about our five clusters from the four features plotted in Figure 5. The plots show that Clusters 2 and 5 have the first two highest means of those four feature variables. Let us take a look at the first variable, GPP, which is the standard indicator to capture provincial economy that consolidates income earned from all production of goods and services in a province during a certain period. The provinces in Cluster 2 are in a list of top ten highest ranked on the GPP per capita in 2019. The more economic growth a province has experienced, the greater one has travelers/migrants. This increases the risk of COVID-19 transmission. The average monthly household income has the highest mean in Cluster 5 including four provinces as follows.

- The capital of Thailand, Bangkok, and one of the Bangkok vicinity, Samut Prakan.
- A major tourism destination named “Pattaya” located in Chonburi.
- The smallest province of Thailand, Samut Songkhram.

As a result of being densely populated, it affects the spread of COVID-19 in these provinces from gathering events. Fortunately, the number of healthcare workers by province in Thailand is consistent with the cumulative COVID-19 cases and deaths. Last but not least, the urban environment with thickly packed housing in slums facilitates the spread of vector-borne diseases, such as dengue, or even the emerging infectious disease COVID-19.

Table 4 Pearson correlation output from SPSS for the COVID-19 cases (deaths) and other feature variables, where the cells with * serve to summarize that the null hypothesis $H_0 : \rho = 0$ is rejected since Sig. is smaller than 0.05.

		Rate of cumulative confirmed cases	Rate of cumulative confirmed deaths
Rate of cumulative confirmed cases	Pearson Correlation	1.000	0.511
	Sig. (2-talied)		0.000*
Rate of cumulative confirmed deaths	Pearson Correlation	0.511	1.000
	Sig. (2-talied)	0.000*	
Population size	Pearson Correlation	0.096	−0.141
	Sig. (2-talied)	0.409	0.223
Unemployment rate	Pearson Correlation	0.240	0.012
	Sig. (2-talied)	0.035*	0.920
GPP per capita	Pearson Correlation	0.596	0.275
	Sig. (2-talied)	0.000*	0.016*
Number of establishments	Pearson Correlation	0.195	−0.029
	Sig. (2-talied)	0.089	0.801
Number of medical personnel	Pearson Correlation	0.460	0.282
	Sig. (2-talied)	0.000*	0.013*
Volume of solid waste	Pearson Correlation	0.293	0.018
	Sig. (2-talied)	0.010*	0.878
Average monthly household income	Pearson Correlation	0.527	0.256
	Sig. (2-talied)	0.000*	0.024*
Average household debt	Pearson Correlation	−0.056	−0.103
	Sig. (2-talied)	0.631	0.372
Birth rate	Pearson Correlation	0.327	0.143
	Sig. (2-talied)	0.004*	0.214
Death rate	Pearson Correlation	−0.109	0.011
	Sig. (2-talied)	0.344	0.926
Percentage of forest area	Pearson Correlation	−0.316	−0.056
	Sig. (2-talied)	0.005*	0.627
Cases of dengue	Pearson Correlation	0.372	0.297
	Sig. (2-talied)	0.001*	0.009*

Let us now move onto the second part analysis. Based on the time series data on COVID-19 cases in daily by province collected from 1 January 2022 to 30 June 2022 that consists of overall 181 observations, each of which is transformed into the rate per 100K population according to (17) with replacing the numerator by "Number of daily COVID-19 cases in the province". For example, recalling the original time series data for Trat from Example 2, we need to convert it to become the transformed one in the table below.

Date	Number of daily COVID-19 cases in Trat	
	Original data	Transformed data
1-Jan-22	4	1.75
2-Jan-22	7	3.07
3-Jan-22	8	3.50
4-Jan-22	6	2.63
5-Jan-22	13	5.69
6-Jan-22	50	21.89
7-Jan-22	61	26.71
8-Jan-22	75	32.84

Date	Number of daily COVID-19 cases in Trat	
	Original data	Transformed data
9-Jan-22	51	22.33
10-Jan-22	44	19.27
11-Jan-22	27	11.82
12-Jan-22	72	31.53
13-Jan-22	34	14.89
14-Jan-22	41	17.95
15-Jan-22	33	14.45
16-Jan-22	13	5.69

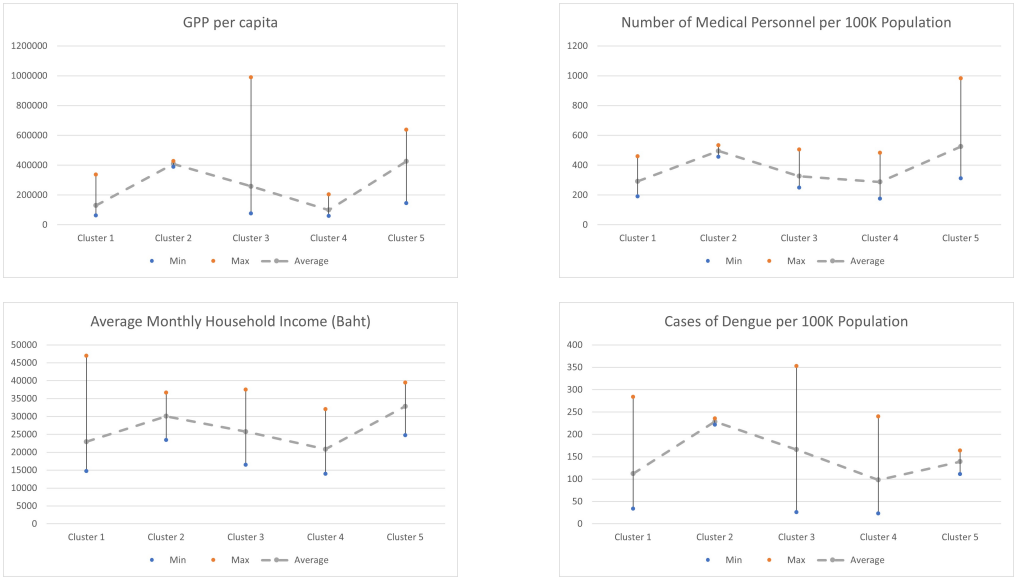


Figure 5 Average-Max-Min charts of feature variables correlated to the COVID-19 cases and deaths by cluster.

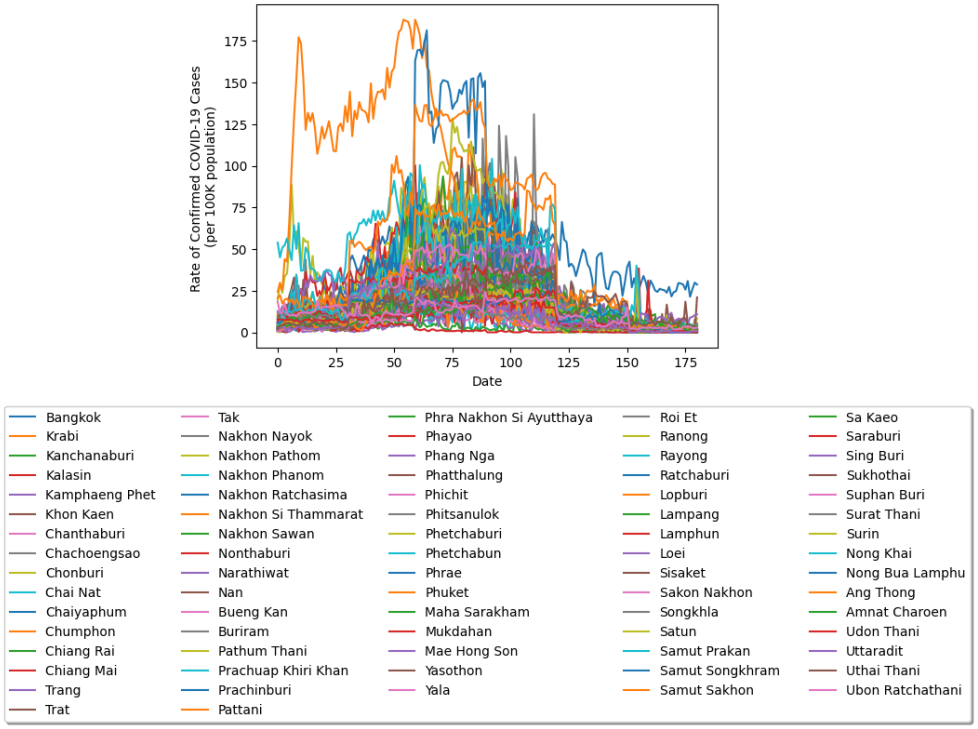


Figure 6 Daily time series of COVID-19 cases per 100,000 people for each province in 181 days.

As of 1 January 2022, the original data turns into $\frac{4}{228{,}376} \times 100{,}000 = 1.75$, where the denominator indicates population size of Trat. It is the same computation for the others. To see the trends of Omicron wave in Thailand during the first half of 2022, the transformed data for all 77 provinces is

depicted in Figure 6 as 77 time series plots.

The whole calculation in this part is carried out using "TimeSeriesKMeans" function from "tslearn.clustering" library in Python. Herein, the DTW distance plays a role as metric in time series clustering. It can be summarized from the graph in Figure 7 that there are 3 clusters in our time series data on the basis of elbow method. By applying the time series k-means clustering with DTW distance measure, we achieve Figure 8 to display three time series plots by cluster as well as choropleth maps visualization colored according to its cluster.

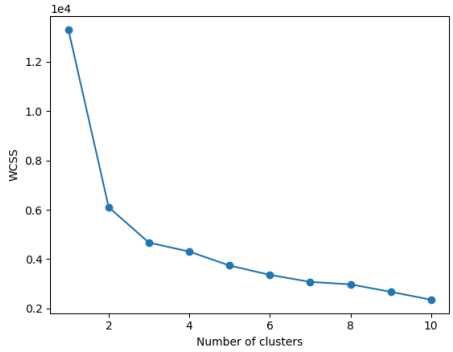


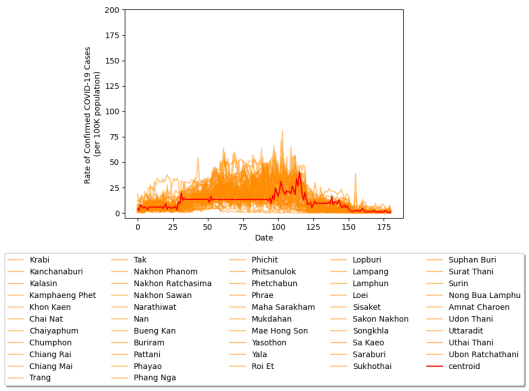
Figure 7 A graphical representation of finding the optimal 'k' in time series k-means clustering using elbow method.

3. Conclusions

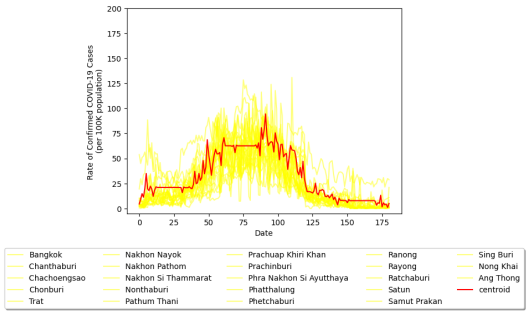
This study created a simple binary integer linear program for DTW distance calculation, which is easy to figure it out by CPLEX solver, or via "DOcplex" library in Python. We utilized k-means algorithm to cluster confirmed cases of COVID-19 infection during the Omicron wave in Thailand. In the first part, the confirmed cases were cumulated by province per 100 thousand population, and analyzed in conjunction with confirmed COVID-19 deaths, which is a provincial clustering in terms of data points in two-dimensional space with Euclidean distance measure. Using hypothesis test for the population correlation coefficient, we discovered that GPP per capita, number of medical personnel, average monthly household income, and cases of dengue are significantly correlated in a linear association with the cumulative confirmed cases and deaths. In the second part, the confirmed COVID-19 cases in daily per 100 thousand population were grouped in terms of time series data by province with DTW distance measure. Choropleth maps were used for visualization of our research findings to understand them easier. Finally, the results from both parts analysis were highlighted that Phuket, and Samut Sakhon show up in highly infectious cluster compared to their population.

Acknowledgements

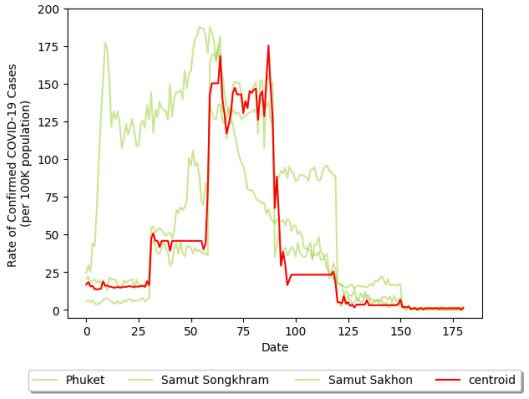
We would like to thank the referees for the careful and insightful review of our manuscript.



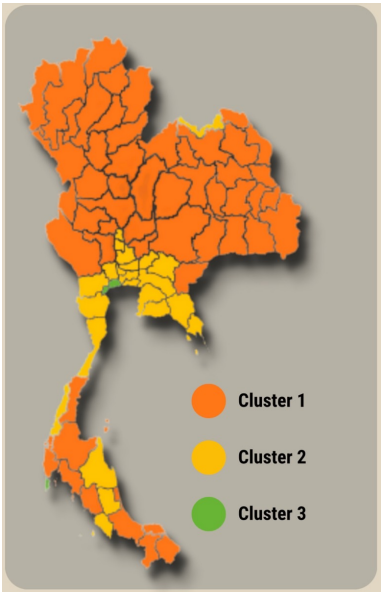
(a) Cluster 1



(b) Cluster 2



(c) Cluster 3



(c) Choropleth maps

Figure 8 Clusters (a)–(c) for rate of COVID-19 confirmed cases per 100K population based on time series k-means clustering and their geographic visualization (c).

References

Abdullah D, Susilo S, Ahmar AS, Rusli R, Hidayat R. The application of K-means clustering for province clustering in indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Qual Quant. 2022; 56: 1283-1291.

Afzal A, Ansari Z, Alshahrani S, Raj AK, Kuruniyan MS, Saleel CA, Nisar KS. Clustering of COVID-19 data for knowledge discovery using c-means and fuzzy c-means. Results Phys. 2021; 29: 104639.

Ampornphan P. Association analysis of COVID-19 outbreak in Thailand using data mining techniques. PSAKU Int J Interdisc Res. 2021; 10: 21-33.

Asuero AG, Sayago A, Gonzlez A. The correlation coefficient: An overview. Crit Rev Anal Chem. 2006; 36: 41-59.

Best JW, Kahn JV. Research in education. Pearson Education India; 2016.

Bucci A, Ippoliti L, Valentini P, Fontanella S. Clustering spatio-temporal series of confirmed COVID-19 deaths in Europe. Spat Stat - Neth. 2022; 49:100543.

Cai B, Huang G, Samadiani N, Li G, Chi CH. Efficient time series clustering by minimizing dynamic time warping utilization. IEEE Access. 2021; 9: 46589-46599.

Cerqueti R, Ficcadenti V. Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million. Chaos Solit Fractals. 2022; 158:111975.

DUrso P, De Giovanni L, Vitale V. Spatial robust fuzzy clustering of COVID-19 time series based on B-splines. Spat Stat - Neth. 2022; 49:100518.

Huang X, Ye Y, Xiong L, Lau RY, Jiang N, Wang S. Time series k-means: A new k-means type smooth subspace clustering for time series data. Inform Sciences. 2016; 367:1-13.

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. In: Annual Conference on Science and Technology Research (ACOSTER). Journal of Physics: Conference Series: IOP Publishing; 2021. p. 012027.

Izakian H, Pedrycz W, Jamal I. Fuzzy clustering of time series data using dynamic time warping distance. Eng Appl Artif Intel. 2015; 39: 235-244.

Jeong YS, Jeong MK, Omitaomu OA. Weighted dynamic time warping for time series classification. Pattern Recogn. 2011; 44: 2231-2240.

Keogh E, Ratanamahatana CA. Exact indexing of dynamic time warping. Knowl Inf Syst. 2016; 7: 358-386.

Lee J, Yoo S, Kim H, Chung Y. The spatial and temporal variation in passenger service rate and its impact on train dwell time: A time-series clustering approach using dynamic time warping. Int J Sustain Trans. 2018; 12:725-736.

Li H, Liu J, Yang Z, Liu RW, Wu K, Wan Y. Adaptively constrained dynamic time warping for time series classification and clustering. Inform Sciences. 2020; 534:97-116.

Li M, Zhu Y, Zhao T, Angelova M. Weighted dynamic time warping for traffic flow clustering. Neurocomputing. 2022; 472:266-279.

Liao TW. Clustering of time series data–a survey. Pattern Recogn. 2005; 38:1857-1874.

MacQueen J. Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability; 1967. p. 281-297.

Mattera R. A weighted approach for spatio-temporal clustering of COVID-19 spread in italy. Spat Spatiotemporal Epidemiol. 2022; 41:100500.

Ratner B. The correlation coefficient: Its values range between +1/-1, or do they? J Target Meas Anal Market. 2009; 17:139-142.

Rizvi SA, Umair M, Cheema MA. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. Chaos Solit Fractals. 2021; 151:111240.

Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. 2018; 126:1763-1768.

Sivaraks H, Sathianwiriyakhun P, Janyalikit T, Ratanamahatana C. Accurate time series classification using partial dynamic time warping. In: Second International Conference on Advances in Applied Science and Environmental Technology (ASET); 2015. p. 31-35.

Watanabe N. A k-means method for trends of time series: An application to time series of COVID-19 cases in Japan. Jpn J Stat Data Sci. 2022; 5:303-319.

Yao Y, Zhao X, Wu Y, Zhang Y, Rong J. Clustering driver behavior using dynamic time warping and hidden Markov model. J Intell Transport S. 2021; 25: 249-262.

Zubair M, Iqbal A, Shil A, Haque E, Moshiul Hoque M, Sarker IH. An efficient K-means clustering algorithm for analysing COVID-19. In: International Conference on Hybrid Intelligent Systems: Springer; 2020. p. 422-432.