Cluster Pattern Identification of the COVID-19 Pandemic in Thailand

[1]Maythawee Jantha, [2]Pathana Rachavong and [3]Tanyaluck Chansombat

Department of Natural Resources and Environment,
Faculty of Agriculture Natural Resources and Environment, Naresuan University,
99 Moo 9 Tanbon Thapho Amphoe Muang, Phitsanulok, Thailand
[1]Email: maythweej63@nu.ac.th, [2]Email: pathanar@gmail.com, [3]Email: tanyalaks@nu.ac.th

ABSTRACT

The novel coronavirus (COVID-19) pandemic poses a serious threat to human health worldwide. Thailand has more than 300,000 infected as of June 2021. To understand disease transmission patterns, it is necessary to know outbreak patterns. In this study, we survey COVID-19 cases in Thailand from March 1, 2020, to July 28, 2021 using Open Government Data of Thailand. Country-level cases and disease rates are mapped using a geographic information system (GIS). Overall disease trends in Thailand and in 77 provinces are analyzed using K-means clustering analysis in R.

Based on the results from K-means analysis, the results show that the optimal k value to form a cluster of 2 is accepted. Cluster 1 contains 1 province which is Bangkok, Cluster 2 consists of 62 provinces, and Cluster 3 contains 9 provinces based on COVID-19 data.

There are specific patterns of disease curves and are assigned to clusters. The results of this study provide insights into creating disease control and mitigation strategies.

1. INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (WHO, 2020). In early December 2019, an outbreak of COVID-19 originally occurred in Wuhan City, Hubei Province, China and later on January 30, 2020 the World Health Organization declared the outbreak as a pandemic incident with Public Health Emergency of International Concern (Harapan et.al., 2020). In Thailand, the first COVID-19 patient was found on 12 January 2020 in Bangkok. Since then, the disease has spread to several provinces in the country. Thailand has faced the third wave of the pandemic outbreak which started from 1 April 2021. As of 30 July 2021, more than 578,375 cases of COVID-19 have been reported in Thailand. (CDC Thailand, 2021)

The K-means algorithm is helpful in segmenting a heterogeneous population into more homogeneous subgroups and it also offers a better view of applicant characteristics and needs, which may lead to more targeted rehabilitation options (Armstrong, et.al., 2012). This technique is a cluster algorithm that provides several competence advantages (Li and Haiyan 2012). Pattern recognition plays a vital role in exploring meaningful information and detecting complex relationships from a large data set, such as identifying risk factors and illustrating the trend of a disease. Cluster analysis is one of the approaches to pattern recognition which groups objects with similar attributes into the same cluster. Therefore, objects have high similarity within cluster while low similarity between clusters. Identifying geospatial patterns of COVID-19 is also essential for disease mitigation because it helps to illustrate the extent and impact of the pandemic, develop public health policies and aid decision making and community action. The objective of this study is to identify the patterns of the COVID-19 cases in Thailand by applying K-means clustering analysis to identify these patterns (Wu, J. and Sha, S., 2021).

2. Methods and Statistical Analysis

2.1 Data Collection and Processing

The COVID-19 data was obtained from Website of Department of Disease, Thailand, which are publicly assessable from the website. The daily-country level data that started from 12 January 2020 to the present are updated daily. Considering there were only very few cases before January 2020, the data were removed before 1 April 2021. Therefore, the data in our analysis start on 1 March 2021 up until 28 July 2021. Spatially, the data from 77 provinces were included.

Two datasets were generated from the dataset for separate analyses. For spatial pattern analysis, the data were grouped by the date to obtain the total number of cases in each province on the latest day of the study period. For K-means clustering and time-series analysis, the data were aggregated by province and risk to obtain the daily new case data during the study period in each state. Data were analyzed using the K-Means Clustering method in R Software version 3.6.3.

2.2 Spatial Pattern Analysis

To demonstrate the spatial distribution of COVID-19 cases, the total number of cases (up to 28 July 2021) in each province with a geographic information system (GIS) was mapped. The rate of the disease using the total number of cases in each province divided by the population in that province is also calculated.

2.3 Temporal Trend Analysis

To achieve a trend of daily cases in Thailand during the third wave of the COVID 19 outbreak, time series data are plotted in descending order based on the total case number and genders from 1 April – 28 July 2021.

2.4 K-Means Clustering

K-means clustering is applied to explore the pattern of disease clusters in different provinces. K-means clustering is a simple and common unsupervised machine learning algorithm for exploratory data analysis to get an intuition about data structure which divides the data into subgroups or clusters. The province with a similar disease cluster curve will be grouped into the same cluster. The K-means algorithm functions as following processes. First, the algorithm specifies the number of clusters (K); second initialize K centroids and calculate the distance between each centroid and each data point; third, the data points that have the shortest distance to a centroid are grouped in the same cluster; fourth, calculate the new centroid based the data points in a cluster; and fifth, repeat the process from until the centroids are stable (Likas et.al., 2003)

3. Result and Discussion

3.1 Spatial Pattern Analysis

The total number of cases (from 1 April to 28 July 2021) in each province and the rate of the disease were mapped using GIS. The number of cases and infection rate are divided into five categories.

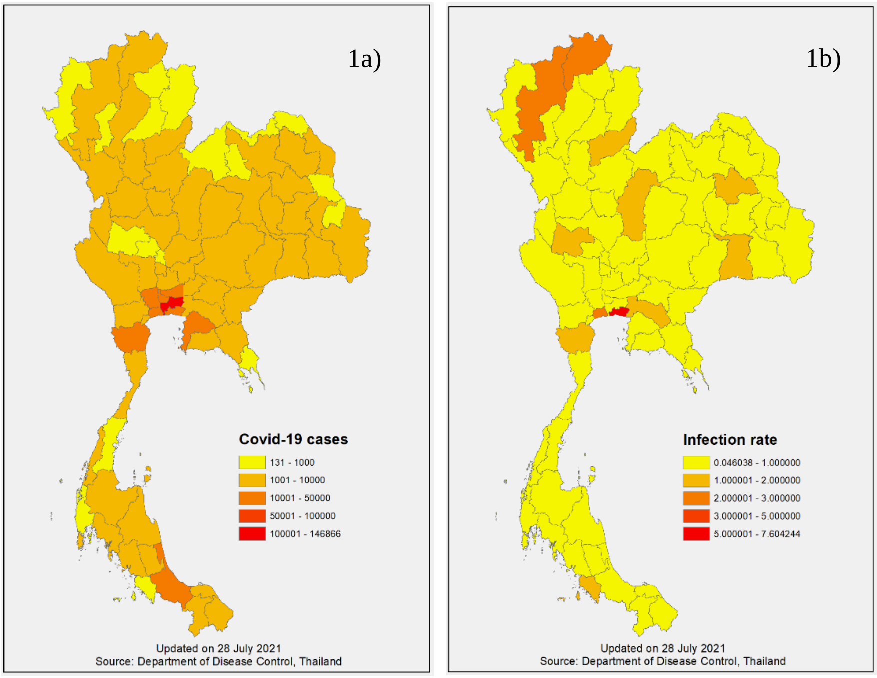

Figure 1. The country-level of COVID-19 cases Thailand on 28 July 2021. The map shows the number of COVID-19 cases (1a) and infection rate in each province (1b). The number of cases and infection rate is divided into five categories and illustrated by colors from light yellow (low number) to deep red (high number), respectively.

3.2 Temporal Trend Analysis

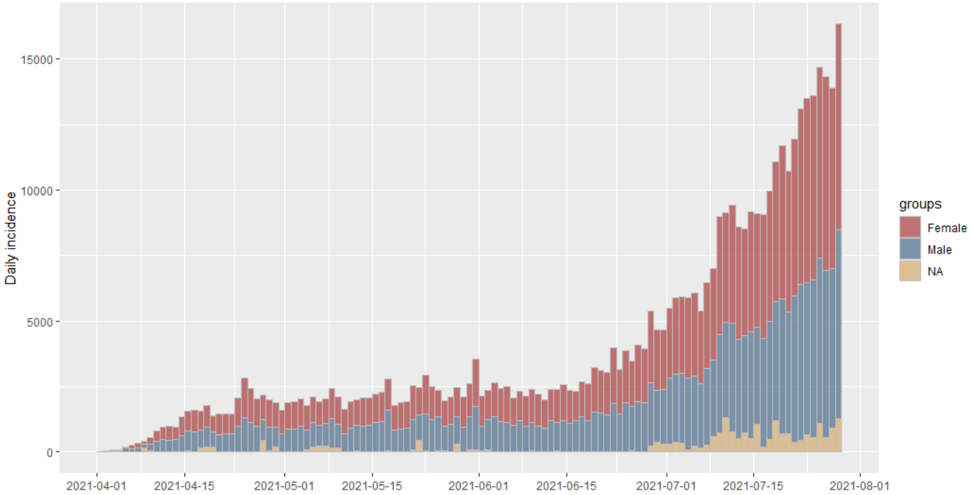

Figure 2. Trend of daily cases in Thailand during the third wave of the COVID 19 outbreak, time series data are plotted in descending order based on the total case number and genders from 1 April – 28 July 2021.

The trends of daily cases in 77 provinces are demonstrated in Fig. 2. In the early of state (1 April –31 April), the disease curve was quite low, However, the case number increased quickly and formed a large peak in mid-May and increasingly risen in June and July.

3.3 K-Means Clustering

The total within-cluster sum of squares (WSS) and Silhouette method were plotted. Based on the spatial pattern analysis of COVID-19 in Thailand, two to five main clusters in study period were observed. From the WSS plots, the K value as two for the whole period analysis was selected. K-means algorithm to initialize the centroids of clusters was applied using R package.

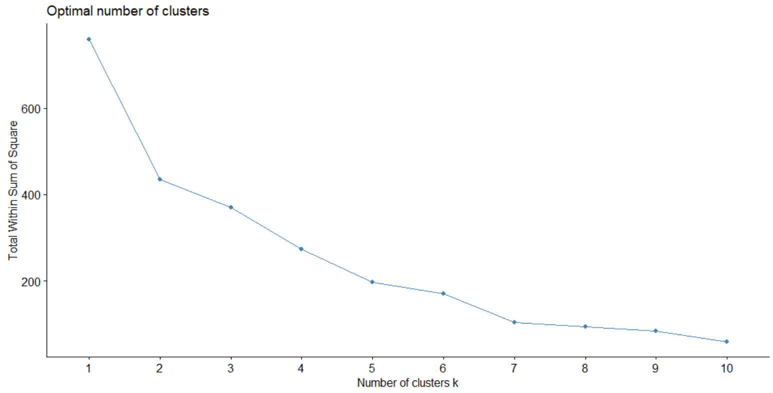

Figure 3. K value cluster analysis using the total within-cluster sum of squares (WSS) method.

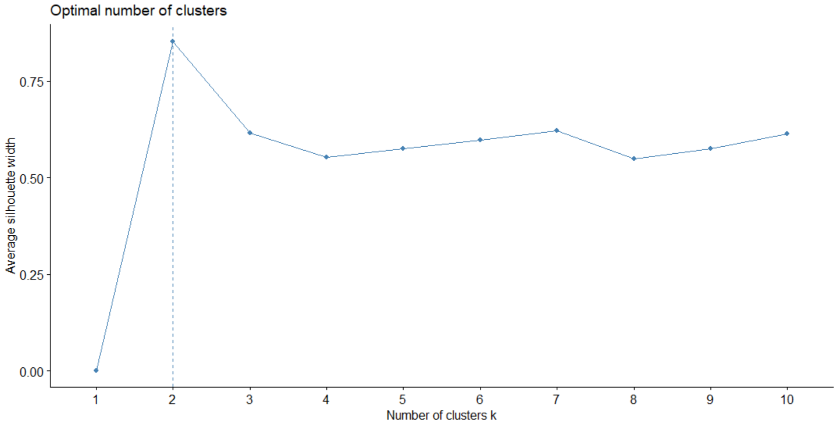

Figure 4. K value cluster analysis using the Silhouette method.

In Fig. 1, Bangkok apparently is the biggest cluster in Thailand, the province became the epicenter of the outbreak. This is due to Bangkok is the capital city and the most densely populated city of Thailand. The optimal number of k groups was determined using commonly used methodologies namely within-cluster sum of squares (WSS) and Silhouette Statistics. The results can be seen in Fig. 3 and 4.

Based on Fig. 3 and 4, the WSS method and the Silhouette method obtained optimal clusters at k = 2. Therefore, based on the results from K-means analysis, the optimal k value to form a cluster is 2 is accepted. As shown in Table 1, Cluster 1 contains 1 province which is Bangkok, Cluster 2 consists of 62 provinces, and Cluster 3 contains 9 provinces based on COVID-19 data (Fig. 5).

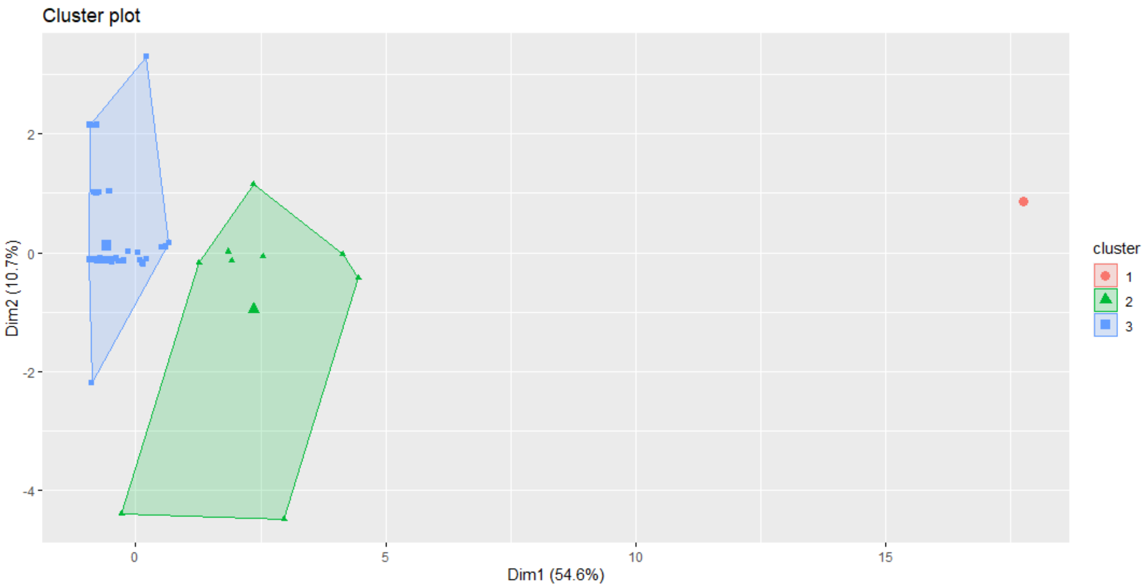

Figure 5. Cluster analysis using K value = 2

Table 1. Cluster analysis using K-means analysis

Cluster 1	Cluster 2			Cluster 3
Bangkok	Amnatchareon	Ang-thong	Buengkarn	Ayuddhaya
	Buriram	Chacherngsoa	Chainat	Chonburi
	Chaiyaphum	Chantaburi	Chiangmai	Nonthaburi
	Chiangrai	Chumpon	Loei	Pathumthani
	Lopburi	Mahasarakam	Meahongson	Petchaburi
	Mukdahan	Nakornnayok	Nakornpanom	Samutprakan
	Nakornpathom	Nakornratchasima	Nakornsawan	Samutsakorn
	Nakornsithammarat	Nan	Narathiwat	Songkhla
	Nongbualamphu	Nongkhai	Pattani	Tak
	Petchabun	Phang-nga	Phattalung	
	Phayoa	Phichit	Phitsanulok	
	Phrea	Phuket	Prachinburi	
	Prachuabkhirikhan	Ranong	Ratchaburi	
	Rayong	Roi-ed	Sakonnakhon	
	Roi-ed	Sakonnakhon	Samutsongkram	
	Saraburi	Satun	Singburi	
	Srakaew	Srisaket	Sukhothai	
	Suphanburi	Suratthani	Surin	
	Trad	Trung	Ubonratchathani	
	Udonthani	Uthaithani	Uttaradit	
	Yala	Yasotorn		

4. CONCLUSION

The result of this study is expected to provide input to the government in making policies related to restrictions on community activities and to provide insights into creating disease control and mitigation strategies of COVID-19 in Thailand. This study is according with Zarikas, et.al. (2020), Azarafza, et.al. (2021), and Dahlan, A., et.al. (2021) which stated that clustering active cases in a region is useful for drawing conclusions about the disease impact which spreads rapidly in an area and the pattern of transmitting infection between provinces can be estimated using the clustering method. Therefore, it can be determined that K-means clustering method is one of effective methods to illustrate disease spread patterns and also provide alternative solutions related to this distribution pattern of the COVID-19 outbreak in Thailand.

5. DATA AVAILABILITY

The data in this study can be accessed at: Website of Department of Disease, Thailand. (https://ddc.moph.go.th/covid19-dashboard/)

6. REFERENCES

Armstrong, J.J., Zhu, M., Hirdes, J.P., Stolee, P., 2012. *K-Means cluster analysis of rehabilitation service users in the home health care system of ontario: examining the heterogeneity of a complex geriatric population*. Arch. Phys. Med. Rehabil. **93**(12), 2198–2205

Azarafza, M., Azarafza, M., Akgun, H., 2021. *Clustering method for spread pattern analysis of coronavirus (covid-19) infection in iran*. J. Appl. Sci. Eng. Technol. Educ. **3**(1), 1–6 (2021)

Dahlan, A., S. Susilo, Ansari, S.A., R. Rusli and Rahmat, H., 2021 *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. https://doi.org/10.1007/s11135-021-01176-w

Department of disease control, Ministry of Public Health, Thailand., 2021. *COVID-19 Thailand Data*, https://ddc.moph.go.th/covid19-dashboard/

Likas, A., Vlassis, N., Verbeek, J.J., 2003. *The global k-means clustering algorithm*. Pattern Recognit 36, 451–461.

Wu, J. and Sha, S., 2021. *Pattern Recognition of the COVID-19 Pandemic in the United States: Implications for Disease Mitigation*. Int. J. Environ. Res. Public Health 2021, 18, 2493. https://doi.org/10.3390/ ijerph18052493

Zarikas, V, Poulopoulos, S.G., Gareiou, Z., Zervas, E., 2020. *Clustering analysis of countries using the covid-19 cases dataset*. Data in Brief **31**, 105787