Infectious Disease Modelling 9 (2024) 387–396

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

# Spatio-temporal clustering analysis of COVID-19 cases in Johor

Fong Ying Foo , Nuzlinda Abdul Rahman *, Fauhatuz Zahroh Shaik Abdullah , Nurul Syafiah Abd Naeeim

*School of Mathematical Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia*

ARTICLE INFO

*Article history:*
Received 30 December 2022
Received in revised form 17 November 2023
Accepted 28 January 2024
Available online 8 February 2024
Handling editor: Jianhong Wu

*Keywords:*
Disease mapping
COVID-19
Hot-spot areas
Sub-district level
Spatio-temporal clustering
Scan statistics

ABSTRACT

At the end of the year 2019, a virus named SARS-CoV-2 induced the coronavirus disease, which is very contagious and quickly spread around the world. This new infectious disease is called COVID-19. Numerous areas, such as the economy, social services, education, and healthcare system, have suffered grave consequences from the invasion of this deadly virus. Thus, a thorough understanding of the spread of COVID-19 is required in order to deal with this outbreak before it becomes an infectious disaster. In this research, the daily reported COVID-19 cases in 92 sub-districts in Johor state, Malaysia, as well as the population size associated to each sub-district, are used to study the propagation of COVID-19 disease across space and time in Johor. The time frame of this research is about 190 days, which started from August 5, 2021, until February 10, 2022. The clustering technique known as spatio-temporal clustering, which considers the spatio-temporal metric was adapted to determine the hot-spot areas of the COVID-19 disease in Johor at the sub-district level. The results indicated that COVID-19 disease does spike in the dynamic populated sub-districts such as the state's economic centre (Bandar Johor Bahru), and during the festive season. These findings empirically prove that the transmission rate of COVID-19 is directly proportional to human mobility and the presence of holidays. On the other hand, the result of this study will help the authority in charge in stopping and preventing COVID-19 from spreading and become worsen at the national level.



## 1. Introduction

The history of disease mapping can be traced back from 4th to 5th centuries before the Common Era, when a Greek physician named Hippocrates first proposed the interrelation between human health and the environment, more specifically the geographical location. In one of his past clinical studies, he observed that residents of north-facing cities tended to be marred by a propensity for tonsillitis and pleurisy, whereas residents of cities with southerly exposures, where fog and mist dispersed more readily, appeared to have better health (Bloss, 2008; Davies, 1971; Lawson & Williams, 2001).

* Corresponding author.
*E-mail addresses:* angelfongying@gmail.com (F.Y. Foo), nuzlinda@usm.my (N. Abdul Rahman), fauhatuz@usm.my (F.Z. Shaik Abdullah), nsyafiah@usm.my (N.S. Abd Naeeim).

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2024.01.009

Despite the early introduction of the interconnection between human health and the environment, it was not until the year 1963 that geographical analysis was fully utilised in the research field. At that time, a professor of geography named Melvyn Howe published the first National Atlas of Disease Mortality in the United Kingdom. He systematically described the geographical distribution of mortality for several diseases in the counties and towns of Scotland, England, and Wales (Howe, 1964). The most well-known application of the disease mapping technique in epidemiology is the distribution of the cholera epidemic that occurred in 1848 in London, where the death cases were plotted. The cluster of the dots on the map help to identify the root cause of the cholera epidemic as the contaminated water pump (John Snow, 1936; Barford & Dorling, 2016). As computers became more user-friendly in the late 1960s, followed by cutting-edge data processing as well as geographic capabilities, disease mapping became a popular analysis technique in analysing various diseases such as yellow fever, hepatitis B, dengue, and obesity (Shannon, 1981; Lawson & Williams, 2001; Zhong et al., 2005; Samat & Percy, 2012; Farhadian et al., 2013; Abd Naeeim et al., 2019; Abd Naeeim & Abdul Rahman, 2023).

As disease mapping is an important tool for depicting disease incidence, it can certainly be useful in analysing the COVID-19 pandemic that has spread globally. As of today, COVID-19 has become the first pandemic that originates from the coronavirus, which explodes worldwide at a speed that is much faster than any other viruses that sparked from the same subfamily of coronavirus, such as the SARS-CoV outbreak in the year 2003 and the MERS-CoV outbreak in the year 2014 (Boulos & Geraghty, 2020; World Health Organization, 2020). Since a deep understanding of the evolution of COVID-19 and how this contagious disease transmits and propagates over space is one of the key factors to shield people from the risk of getting infected with this deadly virus, disease mapping as well as geographical analysis that focuses on COVID-19 are bombarding in the research field (Boulos & Geraghty, 2020; Franch-Pardo et al., 2020).

One of the important issues in disease mapping is known as clustering. For instance, the malnutrition phenomenon among children in any community is commonly due to lack of food resources and poverty in that region, or a high record of wildfire occurrences at certain periods due to the heat wave season (Adeyeye et al., 2017; Parente et al., 2018). It is important to note that the result of the cluster analysis does not reveal any causality, thus making people unable to identify the root cause of the occurrence of that particular disease of interest (Olsen et al., 1996). Despite this limitation, clustering analysis is useful for providing hints or as an indicator on the determination of the factor that contributed to the problem of concern. The classic example would be the study of the cholera outbreak in London in the nineteenth century (John Snow, 1936). Recently, most real-world data analysis problems are inextricably linked to geographical and/or time metrics; as a result, clustering techniques that take both spatial and time dimensions into account have becoming more common in the research field (Abd Naeeim & Abdul Rahman, 2022; Amin et al., 2012; Azage et al., 2015; Kisilevich et al., 2010; Song, Wen, & Yan, 201 8, December).

As COVID-19 is the primary disease of concern in this study, numerous studies that have been carried out regarding this disease recently using a variety of statistical techniques were reviewed (Cheong et al., 2022; Guan et al., 2020; Murugesan et al., 2020; Ullah et al., 2021). For example in China, a total of 1099 laboratory-confirmed cases were taken to study the clinical properties of the COVID-19 outbreak. The results showed that the average age of the infected patient is 47 years old and that the male patient consists of 58.1% as compared to the female patient. Despite not living in Wuhan, 72.3% of the patients were found to be in close contact with city residents. 31.3% of them had visited Wuhan before contracting COVID-19. The findings also indicated that the median incubation period is about 4 days. The distribution of the confirmed cases of COVID-19 in China was visualised on a map using the ArcGIS tool (version 10.2.2) (Guan et al., 2020).

In Helsinki, the capital city of Finland, the COVID-19 infection cases at the mailing code level were used to determine the hot-spot and cold-spot clusters of COVID-19 as well as the relationship between social demographic factors and the infection of COVID-19. With the adoption of Getis-Ord Gi* hot-spot analysis, the hot-spot areas of COVID-19 are found to be centred in the eastern suburbs of Helsinki, and the cold-spot areas are located in the western parts of the city, starting from October 28, 2020, until March 24, 2021. The ordinary least squares model indicates that the COVID-19 infection in Helsinki is associated with a low median income, a high number of foreign residents, a low education level, and a high unemployment rate. This could be due to the low chance of getting a remote job among the foreign residents as well as the language barriers to understanding the safety protocols given at the work site (Siljander et al., 2022).

In other study, in Tehran, the capital city of Iran, 7043 COVID-19 death cases at neighborhood level started from December 2019 until July 2021 were used to study the COVID-19 situation and its associated risk factors. Under the retrospective space-time analysis, only one space-time cluster was detected. This cluster involved the south and southeast regions of Tehran city, and it only lasted six months, from February 2020 until October 2020. Besides, the illiteracy rate and the air pollution concentration were found to be positively associated with the COVID-19 mortality rate in Tehran. This research also showed that the female mortality rate is lower than the male mortality rate. From the age aspect, the mortality rate is 68% higher among the elderly group (those aged more than 65 years old), while the young generation (those aged 25 years old and below) has the lowest mortality rate (Mohammadi et al., 2023).

In United States, the COVID-19 case counts from 3139 United States counties, starting from January 22, 2020, until May 9, 2022, were used to study the potential drivers of the COVID-19 epidemic similarity. The dynamic time warping method that allows different starting times of epidemics was employed to compute the epidemic similarity in the form of a high-dimensional pairwise dissimilarity matrix. The t-distributed stochastic neighbour embedding (*t*-SNE) technique was then applied to forecast the low-dimensional matrix from the pairwise dissimilarity matrix. The low-dimensional matrix enables the study of the relation between the epidemic projection similarity with geographical distance and demographic traits (such

as age structure and population size). The analysis revealed that age structure was the most influential factor in epidemic similarity as compared to $R_o$, geographical distance, and demographic traits (Tad, et al., 2022).

In India, the number of positive cases of COVID-19 at state level starts from February until April 11, 2020, is used to determine the transmission pattern of COVID-19 on a small to medium geographical scale. The Inverse Distance Weighted (IDW) and Kriging interpolation methods are employed to predict the daily confirmed cases of COVID-19. As a result, there are eight states expected to be less dangerous as compared to other regions in India since the predicted daily cases of COVID-19 in these areas range from 0 to 250. The distribution of the predicted confirmed cases of COVID-19 in India was depicted using the GIS tool (Murugesan et al., 2020).

In Malaysia, clustering techniques that consider space and time metrics are deployed to study the transmission of COVID-19. For instance, the monthly cases of COVID-19 at state level starting from March 2020 until September 2020 were used to conduct a spatial analysis. As a result, the most likely cluster has shown that COVID-19 was spread from the west of Malaysia to the east of Malaysia within the investigation period (Ullah et al., 2021). Besides, the daily cumulative cases of COVID-19 at district level were adapted to study the transmission of COVID-19 under the spatio-temporal effect by using the local Moran I statistics and space-time scan statistics. The high-high risk cluster obtained by the local Moran I statistics revealed that the invasion path of COVID-19 propagated from the east of Malaysia to the west region of Malaysia. The space-time scan statistics give a most likely cluster that is centred at Jasin (Malacca district) with 36 districts involved, which starts from November 24, 2020 until February 24, 2021 (Cheong et al., 2022).

All these findings are essential in planning more specific interventions to control the COVID-19 pandemic in Malaysia at a more granular level. Also, it is certain that a more conducive plan can only be conducted when comprehensive information about the actual circumferences or information sufficiently close to the actual scenario is available. However, none of the available research regarding the analysis of COVID-19 in Malaysia was done at the sub-district level, despite the literature mentioned above. The study of how COVID-19 spread geographically and temporally at a sub-district level is definitely crucial in wrestling with the small-to-medium-scale outbreak before it turns into a national crisis that will diminish the whole nation's economy and development (Karabag, 2020; Shang et al., 2021). Hence, in this study, the spatio-temporal clustering analysis will be applied to the COVID-19 data in Johor at the sub-district level to identify the hot-spot areas of COVID-19 using an established statistical clustering technique known as spatio-temporal clustering. The application of the approach used in this research can be extended to other states in Malaysia or any other countries globally, as well as adapting to other infectious diseases.

## 2. Methodology

### 2.1. Data background

The data used in this study are the daily reported cases of COVID-19 in 92 sub-districts in Johor state, Malaysia, as well as the population size associated to each sub-district. The time frame of this research is about 190 days, which started from August 5, 2021, and ended on February 10, 2022. This time frame is subject to the accessibility of the data, as the date beyond the time frame used is cumulative cases (14 days) instead of daily cases. Due to the lack of the actual population size at the sub-district level for the years 2021 and 2022, the population size at the sub-district levels in the year 2010 is used as a baseline to obtain the population estimation for the years 2021 and 2022 (at sub-district level). All the data used are open data and can be found on the official websites of Ministry of Health Malaysia and Department of Statistics Malaysia (Department of Statistics Malaysia, 2011; Ministry of Health Malaysia, 2022).

### 2.2. Population estimation at the sub-district level

The population size at the sub-district level is essential in the identification of spatial clusters. The population size for each sub-district in a particular year is estimated based on the multiplication of the baseline (2010's census) or preceding year's population size and the annual growth rate in that particular year (Department of Statistics Malaysia, 2016). The annual population growth rate (in percentage) that will be used for the population estimation for each year starting from 2011 until 2022 is tabulated as in Table 1.

In general, the population size of a particular sub-district can be estimated based on the equation below.

Let $N_t$ be the available population size in the year $t$, and $N_{t+1}$ be estimated population size in the year $t+1$. Then,

**Table 1**
The annual population growth rate from 2011 until 2022
(Source: Department of Statistics Malaysia, 2016).

| Year | 2011 | 2012 | 2013 | 2014 | 2015 | 2016–2018 | 2019–2022 |
|---|---|---|---|---|---|---|---|
| **Annual population growth rate (100%)** | 0.019 | 0.018 | 0.017 | 0.016 | 0.015 | 0.014 | 0.013 |

$$N_{t+1} = N_t \times \text{Annual Growth Rate in year}. \quad (1)$$

### 2.3. Scan statistic using SaTScan

A scan statistic is used to determine whether a one-dimensional point process is happening in a random manner or if any clusters can be identified (Kulldorff, 1997). For purely spatial scan statistics, a circular shaped scanning window will be placed on the map where the centroid of the scanning window changes within several possible grid points located within the investigated spatial area. The size of the scanning windows changes from zero to a certain threshold that can be set based on the research purpose. Thus, the scanning window is considered to be flexible in terms of location and time parameters. The SaTScan software will generate a number of circular scanning windows, each representing a distinct potential spatial cluster that involves different geographical locations. The ellipse shape of the scanning window is sometimes preferred to detect clusters that are long and narrow (Kulldorff, 2021).

The space-time scan statistic involves both the spatial and time parameters; thus, the scanning window is in a cylindrical shape where the circular (or ellipse) base represents the geographic region and the height represents the time. Similar to the purely spatial scan statistics, the software will provide an infinite number of cylindrical scanning windows that comprise different geographical locations and time periods. Each cylindrical window represents a potential cluster. For the temporal scan statistic, the scanning window only involves one dimension, which is time. It is defined as the height of the cylindrical scanning window in the space-time scan statistic. The maximum time period can be customised based on the research situation (Kulldorff, 2021).

### 2.4. Likelihood ratio test

The likelihood ratio test is employed to determine the significant cluster. The alternative hypothesis for each location and size of the scanning window is that there is a higher danger inside the window than outside.

$H_0$ = no COVID-19 cluster within the scanning window as compared to outside
$H_a$ = COVID-19 cluster present within the scanning window as compared to outside

If the data can be modelled by a Poisson distribution, the likelihood function for a specific window is proportional to

$$\left(\frac{c}{E[c]}\right)^c \left(\frac{B-c}{B-E[c]}\right)^{B-c} I, \quad (2)$$

where $B$ is the total number of cases, $c$ is the observed number of cases within the scanning window, $E[c]$ is the covariate adjusted expected number of cases within the scanning window under the null hypothesis, $B - E[c]$ is the expected number of cases outside the scanning window, and $I$ is an indicator function.

When scanning for high-rate clusters, the indicator function is equal to one when the window has more cases than expected under the null hypothesis of equal risk within and outside the window, and equal to zero otherwise. When scanning for low-rate clusters, the indicator function is equal to one when the window has fewer cases than expected under the null hypothesis of equal risk within and outside the window, and equal to zero otherwise. When scanning for clusters with either a high or low-rate, then the indicator function is equal to one for all windows (Kulldorff, 2021).

For the Poisson model, the expected number of cases under the null hypothesis of equal risk within and outside the window is calculated using indirect standardization. The expected number of cases (spatial analysis) is computed in Equation (3).

$$E[c] = \text{population size in particular area} \times \begin{pmatrix} \text{total number of} \\ \text{cases} \end{pmatrix} / (\text{total population size}). \quad (3)$$

The likelihood function is maximised across all window locations and sizes, with the most likely cluster being the one with the highest likelihood. This gives an indication that the cluster is least likely to have occurred by chance. The greatest likelihood ratio test statistic is the likelihood ratio for this window. Its null-hypothesis distribution is calculated by repeating the same analytic exercise on a large number of random replications of the null-hypothesis data set. The $p$-value is determined by using Monte Carlo hypothesis testing, where the rank of the maximum likelihood from the real data set is compared with the maximum likelihoods from the random data sets. If this rank is R, then

$$\text{p-value} = \text{R}/(1 + \text{number of simulation}). \quad (4)$$

In this study, the number of simulations is restricted to 999. In this case, it is always clear whether to reject or not reject the null hypothesis for typical cut-off values such as 0.001. With 999 random replications, the lowest $p$-value that the Monte Carlo hypothesis testing can report is $1 \div (999 + 1) = 0.001$.

The null hypothesis of no risk difference within and outside the scanning window can be rejected when the maximum likelihood ratio computed for the most likely cluster from the read data set is higher than the maximum likelihood ratio computed for the most likely cluster from the simulation data. In this case, the inference of the presence of clusters can be drawn (Kulldorff, 2021).

Relative risk is any non-negative number. The relative risk is a measure that indicates how common the disease is in a particular region and time period compared to the benchmark. Theoretically, the relative risk is the estimated risk within the cluster divided by the estimated risk outside the cluster. A value of greater than one implies the risk is higher, while the risk is lower when a value of less than one is obtained. It is calculated as follows (Kulldorff, 2021):

$$\text{Relative Risk} = \frac{c/E[c]}{(B-c)(E[B]-E[c])} = \frac{c/E[c]}{(B-c)(B-E[c])}, \tag{5}$$

where $c$ is the number of observed cases within the cluster and $B$ is the total number of cases in the data set (Kulldorff, 2021). Note that since the analysis is conditioned on the total number of cases observed, $E[B] = B$.

The ratio of observed to expected is the observed number of cases within the cluster divided by the expected number of cases within the cluster when the null hypothesis is not rejected; that is, when there is no risk difference inside and outside the cluster. This means that it is the estimated risk within the cluster divided by the estimated risk for the study region as a whole. It is calculated as $c \div E[c]$ (Kulldorff, 2021). For ease of processing, the location ID is used in the SaTScan where the list is attached in Appendix A at: https://bit.ly/3AiS5Wu. The whole process of identifying the cluster using SaTScan is summarised in the flowchart described in Fig. 1.

## 3. Results and Discussions

The population size for the 92 sub-districts in Johor in the year 2021 and the year 2022 is estimated based on Equation (1) and parts of the output are tabulated in Table 2. The baseline population size is taken from the actual figures from the census conducted in the year 2010. The complete population size for the 92 sub-districts in Johor can be found in Appendix B at: https://bit.ly/3UYchoD.

The highest density sub-district in Johor is Plentong located in Johor Bahru district with an estimated population size of 582,647 in the year 2021 and 590,221 in the year 2022. The second dense sub-district in Johor is Pulai located in Johor Bahru district with an estimated population size of 425,227 in the year 2021 and 430,755 in the year 2022. Meanwhile, the lowest density sub-district in Johor is Lenggor located in Mersing district with an estimated population size of 779 in the year 2021 and 790 in the year 2022. The second less dense sub-district in Johor is Pulau-Pulau located in Mersing district with an estimated population size of 883 in the year 2021 and 895 in the year 2022. It is reasonable to assume that the more populous a sub-district, the higher the record of COVID-19 cases will be. Meanwhile, a lower record of COVID-19 cases will be observed in a less populous sub-district. This scenario was empirically proven in other research analyses (Irandoost et al., 2023; Wong et al., 2023). Thus, the impact of population density will also be observed in this study.

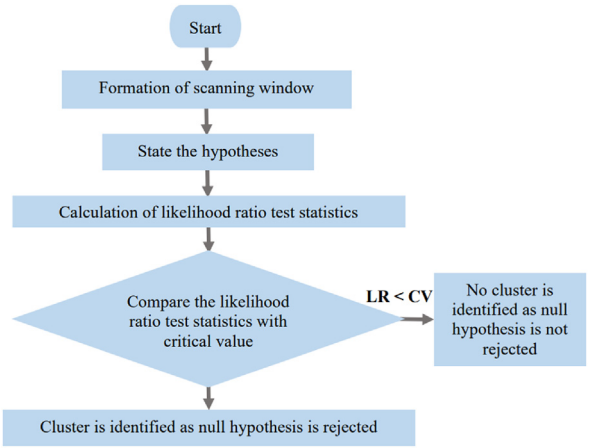


**Fig. 1.** A flowchart of the identification of clusters using SaTScan.

**Table 2**
The estimated population for selected sub-districts in Johor in the years 2021 and 2022.

| District | Sub-district | Population size in year | | | |
|---|---|---|---|---|---|
| | | 2010 | … | 2021 | 2022 |
| Johor Bahru | Bandar Johor Bahru | 124,096 | ⋯ | 146,320 | 148,222 |
| Johor Bahru | Jelutong | 14,651 | ⋯ | 17,275 | 17,499 |
| Johor Bahru | Plentong | 494,152 | ⋯ | 582,647 | 590,221 |

For illustration purposes, a word chart and a bar chart are plotted to understand how the COVID-19 has propagated in Johor across space and time in general. The aggregation of COVID-19 cumulative cases in Johor based on sub-district level is illustrated in Fig. 2, and the aggregation of the COVID-19 cumulative cases in Johor based on month is depicted in Fig. 3.

From the word cloud in Fig. 2, it can be seen that the COVID-19 cases did not spread evenly in Johor. The larger the text size, the higher the frequency. Hence, among the 92 sub-districts in Johor, Plentong has the highest record of COVID-19 cases within the investigated time frame. Thus, it is reasonable to state that there is the presence of a spatial effect in the data, which can be further explored with clustering techniques that consider the spatial metric. Besides, the bar chart in Fig. 3 also revealed that the COVID-19 cases in Johor vary across the month. Obviously, September 2021 has the highest record of COVID-19 cases, followed by August 2021 as compared to the other months within the research time frame. The spike in COVID-19 within these two months might be caused by the Delta variant that is high in transmissibility, which started to invade all the states in Malaysia since the end of July 2021, as well as the low vaccination progress in Johor (Channel News Asia, 2021; Salim, 2021; World Health Organization, 2021). The rest of the months had more or less similar numbers of COVID-19 cases (cumulative). With that, it is reasonable to presume that there is a presence of a temporal effect in the data. Thus, a clustering technique that take into account the time metric can be used to further explore this phenomenon.

In the detection of the spatio-temporal cluster, a total of 175,378 COVID-19 confirmed cases starting from August 5, 2021 until February 10, 2022 were used to build a retrospective discrete Poisson model. Only high-rate areas are included during the scanning process as the main focus in this study is to determine the presence of hot-spot areas of COVID-19 disease. The scanning window is cylindrical as the base represents the spatial dimension and the height represents the time metric. The upper limit of the spatial cluster size is set to be 10% of the population at risk, while the maximum temporal cluster size is set to be 50% of the study period. No spatial and temporal adjustments are made, and a high-rate cluster must consist of at least two cases. The $p$-value is obtained by using the standard Monte Carlo method with 999 simulations. If a spatio-temporal cluster has a $p$-value that is less than the significance level of 0.01 or its log likelihood ratio is greater than the critical value of 15.499, it can be considered as a statistically significant cluster.

As a result, thirteen spatio-temporal clusters are found. All the detected clusters are mapped as shown in Fig. 4, and their details are tabulated in Table 3. The complete results can be found in Appendix C: https://bit.ly/3hLBNPD.

The first spatio-temporal cluster is centred at (1.706363 N, 103.529297 E) with a spatial radius of 21.97 km and a total population size of 335,851. This spatio-temporal cluster is known as the most likely cluster as its log likelihood ratio of 6430.642 is ranked the highest among other spatio-temporal clusters. Also, this most likely cluster is statistically significant since its p-value is 0.001 and its log likelihood ratio is greater than the critical value of 15.499. This most likely spatio-


**Fig. 2.** Distribution of COVID-19 in johor.

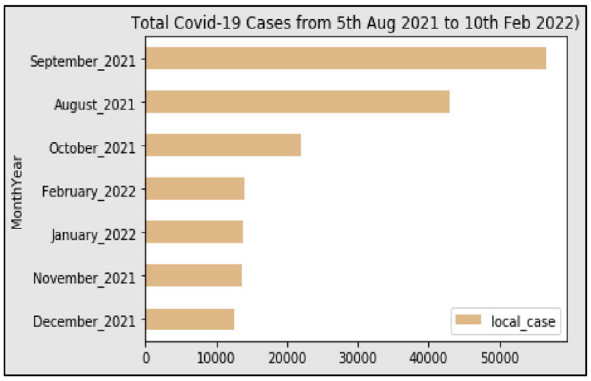


**Fig. 3.** Bar chart of COVID-19 daily cases in Johor (aggregated at sub-district level by month).

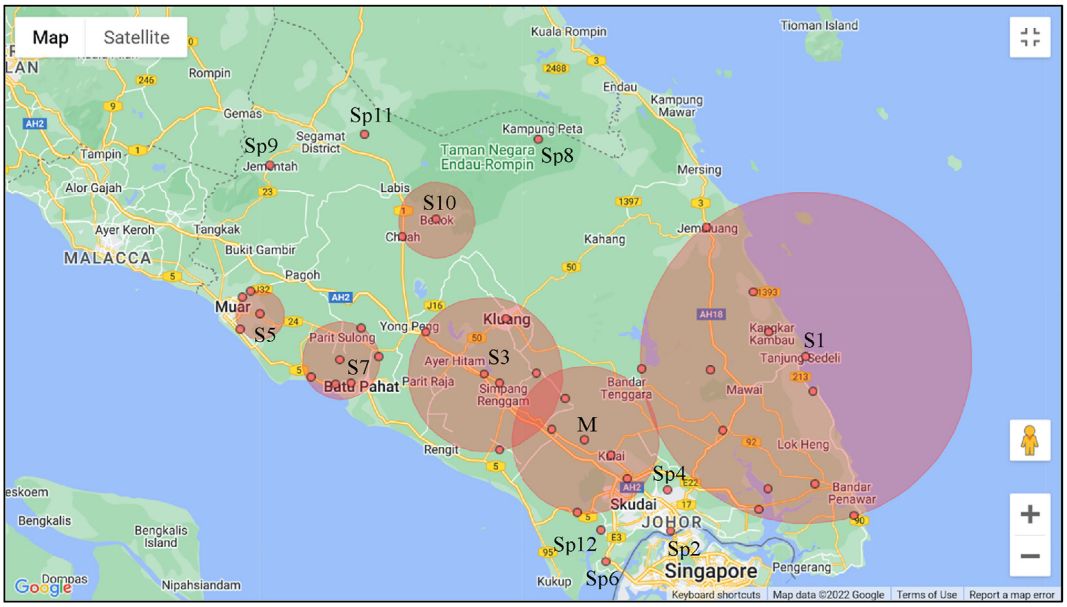


**Fig. 4.** Spatio-temporal clusters map (M: Most likely cluster; S: Secondary cluster; Sp: Secondary point cluster).

temporal cluster is constituting of all sub-districts in the Kulai districts, one sub-district from the Kluang district, and one sub-district from the Pontian district, which are Bukit Batu, Kulai, Sedenak, Senai, Layang, and Pengkalan Raja/Jeram Batu. The time frame of this most likely cluster is from August 13, 2021 until October 1, 2021, which is 50 days. Within these periods, there are three holidays, which are Malaysia's Independence Day (August 31, 2021), Almarhum Sultan Iskandar Hol Day (September 13, 2021), and Malaysia Day (September 16, 2021). The observed number of COVID-19 cases in this most likely cluster is 12,985, which is about three times higher than the expected cases of 4050 (ratio of observed to expected: 3.21). The relative risk is 3.38, which indicates the risk of getting infected by COVID-19 for people staying within this cluster is about 3.4 times higher than for those living outside this cluster.

The rest of the twelve spatio-temporal clusters are known as secondary clusters because their log likelihood ratio values are not the highest. All of them are statistically significant as their $p$-value is 0.001 and their log likelihood ratio values are larger than the critical value. There are 37 out of 92 sub-districts in Johor involved in all the twelve secondary spatio-temporal clusters. None of these clusters are overlapping. Seven out of the twelve secondary clusters are point clusters, meaning there is only one sub-district involved.

It is reasonable to state that there are some inherent factors that contribute to these sub-districts falling into the hot-spot areas of COVID-19. For instance, Bandar Johor Bahru is the main economic centre of Johor state; massive industry is located in Tebrau; an international airport in Senai; and the Malaysia-Singapore Second Link in Tanjung Kupang. All the infrastructure and socio-economic activities involved in these regions imply a high degree of human mobility, thus speeding up the transmission rate of any contagious disease such as COVID-19. However, more comprehensive research, including data

**Table 3**
The most likely of spatio-temporal clusters and its characteristics.

| Characteristics | Value |
|---|---|
| Location IDs included | Kulai_3, Kulai_2, Kulai_1, K_3, Kulai_4, P_5 |
| Coordinates/radius | (1.706363 N, 103.529297 E)/21.97 km |
| Time frame | 2021/8/13 to 2021/10/1 |
| Population | 335,851 |
| Number of cases | 12,985 |
| Expected cases | 4050.020 |
| Observed/expected | 3.210 |
| Relative risk | 3.380 |
| Log likelihood ratio | 6430.642 |
| *p*-value | 0.001 |

collection, is required to determine the stimulant of the high COVID-19 disease records in those detected sub-districts. After all, the focus of this study is to identify the hot-spot areas of COVID-19 disease, which will be useful in narrowing the scope of data collection (for a comprehensive study) as well as optimising resource allocation, which will include speeding up the vaccination rate in those high-risk areas.

## 4. Conclusion

In conclusion, the hot-spot areas of the COVID-19 disease in Johor were determined by adapting the clustering techniques that considered spatio-temporal metrics. Based on the study period, when there is a festive season or in densely crowded areas like the state's economic hub (Bandar Johor Bahru), the COVID-19 disease does seem to surge (before and after the holiday). This phenomenon was consistent with the results found in other research studies (Khan et al., 2023; Maharesi et al., 2023). These findings empirically demonstrate that human mobility and the existence of vacations are somehow directly correlated with the COVID-19 disease transmission rate. All of these discoveries are crucial for avoiding COVID-19 from spreading and become worsen specifically in the study area and in general at national level. Local governments can use this information to impose stricter regulations on these hot-spot areas, such as mask mandates, curfews, prohibiting large-scale gatherings, and even imposing small-scale lock-downs when necessary. When the hot-spot areas of COVID-19 are identified, the allocation of resources can be prioritised for these most affected regions, including the personal protection equipment for the first-line medical officers; test-kit for identification of coronavirus; clinic that deals with COVID-19 related diseases; and delivery of vaccines. Even though, currently, COVID-19 has become an endemic by World Health Organization (WHO), the application of this study can be the baseline or reference point when dealing with the disease if there is any spike or increasing number of cases especially in Malaysia (World Health Organization, 2023).

However, this study still has a limitation. Although the hot-spot areas of COVID-19 in Johor are successfully determined at a small granularity, the factors that trigger the formation of these clusters remain unclear at this stage. Despite the justifications given, other conditions that are distinct from place to place, such as weather (temperature), air humidity, main food resources, and even lifestyle habits, might also cause the residents in that particular region to be more vulnerable to the COVID-19 virus. Thus, a more comprehensive study is needed where these factors might be included in the analysis. To accomplish this, the cooperation of both the local government and the community will be required in providing more details such as demographic information of patients, disease history, individual travel tracts (including date, time, and venue), detailed records of participation in any massive gathering events, and so on. With this data, a more precise analysis can be developed to identify the reasons that contribute to those regions becoming such a high-risk area and also reveal the most close-to-reality pandemic outbreak. However, method of data collection and handling will be a tricky issue due to the confidentiality problem, utilisation of technology (devices to collect, store, and even clean the data), involvement of groups of expertise (which department or research team will be responsible for), and so forth. The possibility that any unknown party will exploit that data for any illegal or own-profit-making purpose (marketing strategies) also affects people's willingness to provide their confidential data for any research purpose. Even if people are willing to share their data for the research purposes, the reliability and integrity of the data can also be another issue that needs to be conquered, as misinformation might be given for any reason. Besides dealing with the data collection issue, approaches that are flexible in terms of considering different kinds of data, such as tolerance of distinct epidemic starting times (Tad, et al., 2022) or using the test data for the case of scarcity in reported cases in projecting the COVID-19 outbreak (Quentin & Pierre, 2021), should also be an alternative research direction which is worth to be explored.

Last but not least, all of the points raised above indicate that combating the COVID-19 pandemic is a challenging process that requires close cooperation and collaboration from all levels of society, including individuals, communities, and local governments. It can only be achieved if each individual plays their role with whole hearts and takes responsibility seriously that this pandemic can be conquered for a more promising future.

## CRediT authorship contribution statement

**Fong Ying Foo:** Formal analysis, Methodology, Software, Writing – original draft. **Nuzlinda Abdul Rahman:** Methodology, Project administration, Supervision, Validation. **Fauhatuz Zahroh Shaik Abdullah:** Data curation, Writing – review & editing. **Nurul Syafiah Abd Naeeim:** Methodology, Validation, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Special thanks to Ministry of Health Malaysia for the COVID-19 data provided in this study. We gratefully acknowledge support from Universiti Sains Malaysia Short Term Grant (304.PMATHS.6315597). The authors thank the editors of Infectious Disease Modelling for their assistance with preparation of this manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idm.2024.01.009.

## References

Abd Naeeim, N. S., & Abdul Rahman, N. (2022). Spatio-temporal clustering analysis using two different scanning windows: A case study of dengue fever in peninsular Malaysia. *Spatial and Spatio-temporal Epidemiology, 41*. https://doi.org/10.1016/j.sste.2022.100496

Abd Naeeim, N. S., & Abdul Rahman, N. (2023). Spatio-temporal clustering analysis of dengue disease in Peninsular Malaysia. *Journal of Public Health, 31*, 307–317. https://doi.org/10.1007/s10389-020-01448-z

Abd Naeeim, N. S., Abdul Rahman, N., & Muhammad Fahimi, F. A. (2019). A spatial–temporal study of dengue in Peninsular Malaysia for the year 2017 in two different space–time model. *Journal of Applied Statistics*. https://doi.org/10.1080/02664763.2019.1648391

Adeyeye, S. A. O., Adebayo-Oyetoro, A. O., & Tiamiyu, H. K. (2017). Poverty and malnutrition in africa: A conceptual analysis. *Nutrition & Food Science, 47*(6), 754–764. https://doi.org/10.1108/NFS-02-2017-0027

Amin, R., Ritter, E., & Kennedy, P. (2012). A geospatial analysis of shark attack rates for the east coast of Florida: 1994-2009. *Marine and Freshwater Behaviour and Physiology, 45*(3), 185–198. https://doi.org/10.1080/10236244.2012.715742

Azage, M., Kumie, A., Worku, A., & Bagtzoglou, A. C. (2015). Childhood diarrhea exhibits spatiotemporal variation in northwest Ethiopia: A SaTScan spatial statistical analysis. *PLoS One, 10*(12), 1–18. https://doi.org/10.1371/journal.pone.0144690

Barford, A., & Dorling, D. (2016). Mapping disease patterns. In *Wiley StatsRef: Statistics reference online* (pp. 1–15). John Wiley & Sons Ltd. https://doi.org/10.1002/9781118445112.stat06102.pub2.

Bloss, J. (2008). The contagion: Historical views of diseases and epidemics. *Reference Reviews, 32*(6), 23–24. https://doi.org/10.1108/RR-03-2018-0048

Boulos, M. N. K., & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbr. *International Journal of Health Geographics, 19*(8). https://doi.org/10.1186/s12942-020-00202-8

Channel News Asia. (2021). Sultan of Johor expresses disappointment over state's low COVID-19 vaccination rates. Retrieved from https://www.channelnewsasia.com/asia/johor-sultan-ibrahim-state-covid-19-vaccination-rates-disappointing-malaysia-2093481. (Accessed 6 August 2022).

Cheong, Y. L., Ghazali, S. M., Che Ibrahim, M. K., Kee, C. C., Md Iderus, N. H., Ruslan, Q., Gill, B. S., Lee, F. C. H., & Lim, K. H. (2022). Assessing the spatiotemporal spread pattern of the COVID-19 pandemic in Malaysia. *Frontiers in Public Health, 10*, 1–11. https://doi.org/10.3389/fpubh.2022.836358

Davies, C. (1971). Hippocrates, the founder of scientific medicine. *History Today, 21*(4). Retrieved from https://www.historytoday.com/archive/hippocrates-founder-scientific-medicine. (Accessed 16 April 2022).

Department of Statistics Malaysia. (2011). Population distribution by local authority areas and mukims. Retrieved from https://www.mycensus.gov.my/index.php/census-product/publication/census-2010/681-population-distribution-by-local-authority-and-mukims-2010. (Accessed 27 March 2022).

Department of Statistics Malaysia. (2016). Population projection (revised). Retrived from https://www.dosm.gov.my/v1/index.php?r=column/ctwo&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT0. (Accessed 27 March 2022).

Farhadian, M., Moghimbeigi, A., & Aliabadi, M. (2013). Mapping the obesity in Iran by bayesian spatial model. *Iranian Journal of Public Health, 42*(6), 581–587. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3744254/.

Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). *Spatial analysis and GIS in the study of COVID-19. A review* (Vol. 739). Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.140033

Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., Liu, L., Shan, H., Lei, C. L., Hui, D. S. C., Du, B., Li, L. J., Zeng, G., Yuen, K. Y., Chen, R. C., Tang, C. L., Wang, T., Chen, P. Y., Xiang, J., … Zhong, N. S. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine, 382*(18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Howe, G. M. (1964). A national Atlas of disease mortality in the United Kingdom: Discussion. *The Royal Geographical Society, 130*(1), 15–22. https://doi.org/10.2307/1794261

Irandoost, K., Alizadeh, H., Yousefi, Z., & Shahmoradi, B. (2023). Spatial analysis of population density and its effects during the Covid-19 pandemic in Sanandaj, Iran. *Journal of Asian Architecture and Building Engineering, 22*(2), 635–642. https://doi.org/10.1080/13467581.2022.2047983

Karabag, S. F. (2020). An unprecedented global crisis! The global, regional, national, political, economic and commercial impact of the coronavirus pandemic. *Journal of Applied Economics and Business Research, 10*(1), 1–6. https://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-164905.

Khan, M., Mottakin, S., Kabir, M., Hoque, E., Amin, M., Rashid, M., & Shariful Islam, M. (2023). Did national holidays accelerate COVID-19 diffusion during the first phase of the pandemic in Bangladesh? *International Journal of Public Health Science, 12*(2), 741–751.

Kisilevich, S., Mansmann, F., Nanni, M., & Rinzivillo, S. (2010). Spatia-temporal clustering. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed., pp. 855–874). Springer Science+Business Media. https://doi.org/10.1007/978-0-387-09823-4.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods, 26*(6), 1481–1496. https://doi.org/10.1080/03610929708831995

Kulldorff, M. (2021). SaTScan user guide for version 10.0. Retrieved from http://www.satscan.org/. (Accessed 4 April 2022).

Lawson, A. B., & Williams, F. L. R. (2001). An introductory guide to disease mapping. In *Statistical methods in medical research*. John Wiley & Sons Ltd. https://doi.org/10.1177/096228020101000607.
Maharesi, R., Kuswandari, I., & Chukwu, C. W. (2023). Effect of long vacation on daily cases of COVID-19 during partial restriction in Jakarta, Indonesia. In *Communications in mathematical biology and neuroscience*. https://doi.org/10.28919/cmbn/7882
Ministry of Health Malaysia. (2022). COVID-19 daily cases in johor at mukim level - moh. Retrieved from https://covid-19.moh.gov.my/terkini-negeri/2022/02/kemaskini-negeri-covid-19-di-malaysia-10022022. (Accessed 5 January 2022).
Mohammadi, A., Pishgar, E., Fatima, M., Lotfata, A., Fanni, Z., Bergquist, R., & Kiani, B. (2023). The COVID-19 mortality rate is associated with illiteracy, age, and air pollution in urban neighborhoods: A spatiotemporal cross-sectional analysis. *Tropcal Medicine Infectious Disease, 8*(85). https://doi.org/10.3390/tropicalmed8020085
Murugesan, B., Karuppannan, S., Mengistie, A. T., Ranganathan, M., & Gopalakrishnanya, G. (2020). Distribution and trend analysis of COVID-19 in India: Geospatial approach. *Journal of Geographical Studies, 4*(1), 1–9. https://doi.org/10.21523/gcj5.20040101
Olsen, S. F., Martuzzi, M., & Elliott, P. (1996). Cluster analysis and disease mapping: Why, when, and how? A step by step guide. *British Medical Journal, 313*(7061), 863–866. https://www.jstor.org/stable/29733060.
Parente, J., Pereira, M. G., Amraoui, M., & Fischer, E. M. (2018). Heat waves in Portugal: Current regime, changes in future climate and impacts on extreme wildfires. *Science of the Total Environment, 631–632*, 534–549. https://doi.org/10.1016/j.scitotenv.2018.03.044
Quentin, G., & Pierre, M. (2021). Clarifying predictions for COVID-19 from testing data: The example of New York State. *Infectious Disease Modelling, 6*, 273–283. https://doi.org/10.1016/j.idm.2020.12.011
Salim, S. (2021). Covid-19: 10 new Delta variant cases detected in Malaysia. Retrieved from https://www.theedgemarkets.com/article/covid19-10-new-delta-variant-cases-detected-malaysia. (Accessed 5 August 2022).
Samat, N. A., & Percy, D. F. (2012). Dengue disease mapping in Malaysia based on stochastic SIR models in human populations. In *International conference on statistics in science, business and engineering, langkawi*. https://doi.org/10.1109/ICSSBE.2012.6396640
Shang, Y., Li, H., & Zhang, R. (2021). Effects of pandemic outbreak on economies: Evidence from business history context. *Frontiers in Public Health, 9*. https://doi.org/10.3389/fpubh.2021.632043
Shannon, G. W. (1981). Disease mapping and early theories of yellow fever. *The Professional Geographer, 33*(2), 221–227. https://doi.org/10.1111/j.0033-0124.1981.00221.x
Siljander, M., Uusitalo, R., Pellikka, P., Isosomppi, S., & Vapalahti, O. (2022). Spatiotemporal clustering patterns and sociodemographic determinants of COVID-19 (SARS-CoV-2) infections in Helsinki, Finland. *Spatial and Spatio-temporal Epidemiology, 41*. https://doi.org/10.1016/j.sste.2022.100493
Snow, J. (1936). *Snow on cholera: Being a reprint of two papers*. The Commonwealth Fund. https://books.google.com.my/books/about/Snow_on_Cholera.html?id=pLQhAQAAMAAJ&redir_esc=y.
Song, J., Wen, R., & Yan, W. (2018). Identification of traffic accident clusters using kulldorff's space-time scan statistics. In *Proceedings - 2018 IEEE international conference on big data* (pp. 3162–3167). https://doi.org/10.1109/BigData.2018.8622226. Seattle.
Tad, A. D., Grant, F., Robert, L. R., & Bre, t D. E. (2022). Epidemic time series similarity is related to geographic distance and age structure. *Infectious Disease Modelling, 7*(4), 690–697. https://doi.org/10.1016/j.idm.2022.09.002
Ullah, S., Nor, N. H. M., Daud, H., Zainuddin, N., Gandapur, M. S. J., Ali, I., & Khalil, A. (2021). Spatial cluster analysis of COVID-19 in Malaysia (Mar-Sep, 2020). *Geospatial Health, 16*(1), 137–144. https://doi.org/10.4081/gh.2021.961
Wong, H. S., Hasan, M. Z., Sharif, O., & Rahman, A. (2023). Effect of total population, population density and weighted population density on the spread of Covid-19 in Malaysia. *PLoS One, 18*(4), Article e0284157. https://doi.org/10.1371/journal.pone.0284157
World Health Organization. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Retrieved from https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-covid-19-media-briefing–12-july-2022. (Accessed 17 May 2022).
World Health Organization. (2021). Coronavirus disease 2019 (COVID-19) situation report - Malaysia. Retrieved from https://www.who.int/malaysia/emergencies/covid-19-in-malaysia/situation-reports. (Accessed 27 April 2022).
World Health Organization. (2023). Coronavirus disease (COVID-19) pandemic. Retrieved from https://www.who.int/malaysia/emergencies/covid-19-in-malaysia/situation-reports. (Accessed 5 September 2023).
Zhong, S. B., Xue, Y., Cao, C. X., Cao, W. C., Li, X. W., Guo, J. P., & Fang, L. Q. (2005). Explore disease mapping of hepatitis B using geostatistical analysis techniques. *International Conference on Computational Science, 3516*, 464–471. https://doi.org/10.1007/11428862_63. Atlanta.