Spatio-Temporal Clustering of Dengue Incidence

Nik Nur Fatin Fatihah Sapri¹, Wan Fairos Wan Yaacob^{2,*}, Yap Bee Wah¹,
Syed Sharizman Syed Abdul Rahim³

¹Faculty of Mathematical and Computer Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia

²Faculty of Mathematical and Computer Sciences, Universiti Teknologi MARA Cawangan Kelantan Kampus Kota Bharu, 15050 Kota Bharu, Kelantan, Malaysia

³Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Jalan UMS, 88400 Sabah, Malaysia

Received January 1, 2021; Revised February 10, 2021; Accepted March 20, 2021

Cite This Paper in the following Citation Styles

(a): [1] Nik Nur Fatin Fatihah Sapri, Wan Fairos Wan Yaacob, Yap Bee Wah, Syed Sharizman Syed Abdul Rahim, "Spatio-Temporal Clustering of Dengue Incidence," *Universal Journal of Public Health*, Vol. 9, No. 3, pp. 120 - 130, 2021. DOI: 10.13189/ujph.2021.090303.

(b): Nik Nur Fatin Fatihah Sapri, Wan Fairos Wan Yaacob, Yap Bee Wah, Syed Sharizman Syed Abdul Rahim (2021). *Spatio-Temporal Clustering of Dengue Incidence*. *Universal Journal of Public Health*, 9(3), 120 - 130. DOI: 10.13189/ujph.2021.090303.

Copyright©2021 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Dengue fever is a well-known vector-borne disease caused by *Aedes aegypti* mosquito. It has become a major burden to economy and society of affected country. In Malaysia, dengue incidence in Selangor has been worsening and alarming. The aim of this study is two-fold; to examine the trend and behaviour of dengue incidence across time and districts in Selangor and to cluster the endemic areas in Selangor using Wards hierarchical clustering method. The spatial and temporal analysis found that the dengue incidence is worsening in the early and middle of the year. The Wards minimum variance method was able to cluster Selangor's endemic area into high endemic areas (Gombak, Hulu Langat, Klang and Petaling), medium endemic area (Sepang) and low endemic areas (Hulu Selangor, Kuala Langat, Kuala Selangor, Sabak Bernam). The findings of the study are significant to respective local authorities in providing information for monitoring and planning the early dengue warning systems. This is important to reduce the dengue incidence in hot spot areas and to safeguard the community from dengue outbreak.

Keywords Dengue Incidence, Spatial and Temporal Analysis, Wards Hierarchical Clustering

transmitted to human body as susceptible host through biting of infective female *Aedes Aegypti* mosquitos. Commonly, people will contract dengue fever after they are infected by one of the serotypes namely DEN-V 1, DEN-V 2, DEN-V 3 and DEN-V 4 [1]. The severe infection of this virus causes dengue hemorrhagic fever or dengue shock syndrome which can cause fatality. This vector-borne disease has emerged as a global health problem since 1950s [2]. World Health Organization (WHO) [3] revealed that the global dengue cases have sharply increased from 2.2 million in 2010 to 3.2 million in 2015. Moreover, the organization added that about 2.5% of the global dengue cases are fatal. There are about 200 million people in Southeast Asia have been infected with dengue virus [4]. Dengue outbreak is endemic in tropical and sub-tropical regions which are mainly driven by climatic factors and supported by non-climatic factors. This is because the life cycle of *Aedes* mosquitos is closely related to climatic factors as it can affect the speed of virus replication, the development, survival and dispersal of *Aedes* mosquitos [5-10].

Spatial and temporal analysis has been widely used in analyzing and detecting the patterns and trends of infectious diseases such as dengue incidence across the underlying time. It is an important analysis prior to other statistical analysis to gain a better understanding on the dengue distribution especially in a specific area so that early precautions can be well-planned. Besides that, it is useful to be applied in disease surveillance, epidemiology

1. Introduction

Dengue fever is an acute viral infection which is mainly

studies and other studies as this analysis plays an important role in quantifying the geographic variation patterns.

Broadly speaking, cluster analysis involves categorization in which a large group of observations is divided into smaller groups so that the observations within each group are relatively similar (possess similar characteristics) and the observations in different groups are relatively dissimilar. Cluster analysis is another form of multivariate technique of analysis. This method is appropriate to be used if the main intention of the research is to investigate whether or not there are groups of cases in dataset and the characteristics of these groups [11]. The word 'cluster' can also be defined as grouping. It is a class of techniques used to classify cases into groups that are relatively homogeneous within themselves and heterogeneous between each other, on the basis of a defined set of variables. These groups are called as clusters. Cluster is a group of relatively homogeneous cases or observations [12]. In other words, cluster analysis is a natural grouping of something in common based on similarities and dissimilarities among objects.

According to Johnson and Wichern [13], the concept of grouping or clustering is different from classification method. In classification method, the interest is trying to identify and observe the classification pattern (form into grouping) due to pre-defined group (existing group). Meanwhile, cluster analysis deals with no assumptions being made concerning the number of groups or the group structure. Theoretically, cluster analysis is an exploratory technique in which it helps the researcher to explore the natural grouping of observations. The concept of grouping or clustering is made on the basis of similarities and distances (dissimilarities).

Besides that, cluster analysis has been widely used by researchers in many fields of study. For example, it is a popular method in market research field. Since the concept of this analysis is an exploratory technique, it is a preferable method to be used especially in market research. Cluster analysis works best in market segmentation. For instance, clustering consumers according to their attribute preferences as to observe the impact on buying customer. Cluster is also used in identifying new product opportunities. The clustering of similar brands/products can help in identifying competitors/market opportunities. Another reason is that cluster analysis is able to understand buying behaviours of the customer. These explain how important the cluster analysis is needed and why it is preferable in market research area.

Cluster analysis has also been used in epidemic studies such as in dengue study [14-20]. A study done by Vandhana and Anuradha [17] employed hierarchical clustering analysis in order to classify the dengue in various states of India. The data used in the study consisted of dengue cases and deaths for 8 years (2011-2018). A comparison on the various hierarchical clustering methods were used such as single linkage, complete linkage,

centroid linkage, median linkage, average linkage and wards linkage. The study found that out of 6 clustering methods used, wards method showed better clustering in clustering the dengue cases of various states in India. Wards method managed to classify 28 states and 7 union territories of India into 4 main clusters which were high dengue risk, low dengue risk, medium dengue risk and no dengue incidence regions.

Another study was carried out by Ponciano et al. [18] in Guatemala City of 3 main cities (Escuintla, Tiquisate and Masagua) using Wards method to cluster the weekly dengue data (2001-2013). A spatial and temporal analysis was conducted on climatic variables such as temperature, rainfall and relative humidity. The trend and behavioural of the climatic variables were observed across time. The findings obtained based on the spatial and temporal analysis revealed that the increased amount of rainfall is associated to increased dengue incidences in all three localities of Guatemala City. However, temperature and humidity were found to be flat for the whole year. Thus, rainfall was highlighted as the main variable in causing the dengue outbreak in Guatemala City. The significant finding from Wards method revealed that the method had successfully grouped the dengue data of each localities of Guatemala City into 3 major clusters, which roughly can be associated to the cases occurred during the warm, rainy and cold seasons. Furthermore, the study also clustered the health information of Escuintla City and found that there were 2 clusters which can be classified as red alert group (high prevalence of dengue) and yellow alert group (low prevalence of dengue).

Another related work of cluster analysis was done by Sun et al. [19]. The aim of the study was to examine the spatial and temporal trend of dengue and to identify the epidemic area of Sri Lanka. Sri Lanka consists of 25 districts. Five-yearly (2012-2016) data on the number of dengue cases were collected from Epidemiology Units of Sri Lanka. The epidemic area of hot spot and cold spot was identified using Getis-Ord G_i^* statistics. Hot spot area is defined as the area with high dengue transmission while cold spot is defined as the area with low dengue transmission. Based on the spatial-temporal clustering analysis, the hot spot area of Sri Lanka consists of only Colombo district, while cold spot area includes Ampara, Anuradhapura, Badulla, Batticaloa, Kurunegala, Kandy, Kegalle, Matale, Monegarala, Nuwara Eliya, Polonnarwa, Puttalam, Trincomalee and Vavuniya districts.

Mutheneni et al. [20] studied spatial distribution and cluster analysis of dengue in Andhra Pradesh, India. Andhra Pradesh is the third largest and populous state located in India. This state consists of 23 districts. Similar to a study done by Sun et al. [19], the study was conducted to classify the hot spot and cold spot areas of dengue disease using Getis-Ord G_i^* statistics. The endemicity level was illustrated using geographical information system (GIS) tools. The findings of the study revealed that

the districts of Warangal, Karimnagar, Khammam and Vizianagaram were classified as hot spot dengue areas, while Adilabad and Nizamad districts were classified under cold spot dengue areas. In order to capture the endemicity patterns in the study areas, self-organizing map (SOM) was applied in the analysis. SOM tool was used to identify the level of endemicity of dengue cases for each district of Andhra Pradesh, India. The classification based on SOM indicated that the district of Andhra Pradesh can be classified into 3 major clusters which were high, medium and low endemic areas.

Therefore, the purpose of this study is to identify the natural groups (clusters) of dengue endemic regions in Selangor using Ward's hierarchical clustering analysis. This proposed method is able to give a better understanding in assessing the cluster analysis of dengue incidence by districts in Selangor. The findings of the study are able to provide vital information to the local authorities and public health institution to take appropriate actions in planning for the effective programs to safeguard the community from dengue outbreak. In addition to this, the findings serve as prior information to the tourists and residence in a specific region of Selangor as to be aware and take initiative to avoid from being infected with dengue virus. Moreover, this study can be used as a guide in conducting further study in the future by developing the dengue early warning systems in the districts of Selangor.

Hence, the paper is organized as follows: Section 2 provides a brief discussion on the ward hierarchical cluster analysis. The methodology carried out this study is deeply explained in this section. Results and discussions from the data analysis are reported in Section 3. The paper ends with some concluding remarks and some recommendations for future study in Section 4.

2. Methodology

2.1. Descriptive Analysis

Prior to the main analysis of wards hierarchical clustering analysis, this study first presents a descriptive analysis to summarize the data used in the study. For example, summary statistics for the total number of dengue (2011 – 2017), the annual average of dengue incidence per 100,000 population and the maximum of the annual dengue incidence per 100,000 population for each district of Selangor. The analysis was carried out using R software.

2.2. Spatial and Temporal Analysis

In this study, a spatial and temporal analysis was used to analyse the whole of Selangor and the districts of Selangor. For this analysis, a temporal trend was plotted to observe the annual trend of dengue incidence per 100,000 populations. Temporal analysis is worth to investigate as it provides the timeline analysis on the dengue incidence

which could serve and relate towards dengue incidence trend, behaviour and gaps that could potentially lead to other sources of evidence across time [21]. The temporal analysis used in this study was the annual temporal analysis of dengue incidence plotted for the whole of Selangor to observe the trend of dengue incidence in Selangor across time. The annual temporal analysis was also plotted for the districts of Selangor to observe the trend of dengue incidence of each district across time. In addition to this, the monthly temporal analysis on the monthly average of dengue incidence and monthly average of temperature, rainfall and relative humidity were also analysed and presented in this study to capture the monthly temporal trend behavioural. This study also produced the spatial map of monthly average of dengue incidence per 100,000 population to indicate the dengue hot spot areas in Selangor. The spatial map produced in the study was analysed using *spatial*, *map* and *tmap* packages in R software version 3.5.2. Twelve spatial maps were produced for each of the month (January - December) with a variation of red colour in which the lightest red represents the lowest dengue incidence per 100,000 populations while the darkest red represents the highest dengue incidence per 100,000 populations in a specific region. The lightest red indicates a range of dengue incidence between 10 to 20 incidences per 100,000 populations in a particular month while the darkest red indicates a range of dengue incidence between 70 to 80 incidences per 100,000 populations in a particular month.

2.3. Euclidean Distance

Distance can be measured in a variety of ways. According to Johnson and Wichern [13], statistical distance can be measured by squared Euclidean distance, Minkowski metric, Canberra metric and Czekanowski coefficient. Among all these, Euclidean distance is the most used. By definition, Euclidean distance measures distance of straight line between two points. Euclidean distance is the most used as it is the most straightforward and generally accepted way of computing distances between objects [22]. The computational for Euclidean distance is as follows:

$$D(\mathbf{a}, \mathbf{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \quad (1)$$

Where;

$\mathbf{a}$ and $\mathbf{b}$ represent the vectors

a_1 , a_2 , b_1 and b_2 represent the element of observation of vector $\mathbf{a}$ and $\mathbf{b}$ respectively.

Eq. 1 is equivalent to Pythagoras theorem. If the vectors contain n dimensions, the Euclidean distance is calculated as follows:

$$\begin{aligned} D(\mathbf{a}, \mathbf{b}) &= \sqrt{(a_1 - b_1)^2 + \dots + (a_i - b_i)^2 + (a_n - b_n)^2} \\ &= \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \end{aligned} \quad (2)$$

Where i denotes points of coordinates.

Hence, the squared Euclidean distance is obtained by using the following formula:

$$D(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n (a_i - b_i)^2 \quad (3)$$

These sample quantities cannot be computed without prior knowledge of the distinct groups and it is the main reason why squared Euclidean distance is often preferred for clustering compared to other methods [13].

2.4. Ward Hierarchical Linkage Clustering

Ward's method is one of the agglomerative hierarchical clustering methods in which the similarity used to join the clusters is calculated as the sum of squares between the two clusters summed over all variables [23-24]. To simplify, Ward's method joins the two clusters whose merger leads to the smallest error sum of squares whereby at each step the pair of clusters with minimum between-cluster distance are merged. This method is also known as Wards minimum variance method. Such measure is derived from the Euclidean distance as in Eq. 1. It considers two clusters of Cluster A and Cluster B. Then, the sum of squared errors (SSE) is defined as follows:

$$SSE_A = \sum_{i=1}^{n_A} (\mathbf{a}_i - \bar{\mathbf{a}})' (\mathbf{a}_i - \bar{\mathbf{a}}) \quad (4)$$

$$SSE_B = \sum_{i=1}^{n_B} (\mathbf{b}_i - \bar{\mathbf{b}})' (\mathbf{b}_i - \bar{\mathbf{b}}) \quad (5)$$

$$SSE_{AB} = \sum_{i=1}^{n_A} (\mathbf{y}_i - \bar{\mathbf{y}}_{AB})' (\mathbf{y}_i - \bar{\mathbf{y}}_{AB}) \quad (6)$$

Where;

$\mathbf{a}_i$ is the i^{th} observation vector and $\bar{\mathbf{a}}$ is the centroid of Cluster A

$\mathbf{b}_i$ is the i^{th} observation vector and $\bar{\mathbf{b}}$ is the centroid of Cluster B

$\mathbf{y}_i$ is the i^{th} observation vector and $\bar{\mathbf{y}}_{AB}$ is the centroid of Cluster AB

Eq. 4- Eq. 6 imply that Wards method is actually calculating the distance between cluster members and the centroid by minimizing the sum of squared Euclidean distances between centroid point and each of the point of cluster members. The centroid can be defined as the sum of all points inside a cluster divided by the number of points in that particular cluster. Once the values are computed, the smallest sum of squared from the pair of sample units will form the first cluster. In each step of the procedure, the cluster members are grouped (clustered) such that the sum of squared error is minimized. The procedure stops when all objects are combined into a single large cluster. According to Strauss and von Maltiz [25], Wards method

can be simply written as:

$$I_{AB} = \frac{n_A n_B}{n_A + n_B} (\bar{\mathbf{a}} - \bar{\mathbf{b}})' (\bar{\mathbf{a}} - \bar{\mathbf{b}}) \quad (7)$$

Where;

$\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ present the centroids Cluster A and Cluster B respectively.

n_A and n_B present the size of Cluster A and Cluster B respectively.

The ward hierarchical linkage clustering was performed using *stats* package in R software. The clustering method employed in this study was analyzed using *hclust()* function with method= "ward.D2". The results obtained from the clustering procedure are visualized by means of a dendrogram.

Dendrogram is a two-dimensional graphical representation diagram which is used to display the results of clustering procedure. It illustrates the information in the amalgamation table in the form of a tree diagram. The branches in the tree represent clusters. The branches come together (merge) at nodes whose positions along a distance (or similarity) axis indicate the level in which the fusions occur. Dendrogram is more useful as there are smaller numbers of cases.

2.5. Study Area Coverage

According to iDengue [26], dengue incidence is alarming in Selangor state. This state is known as a dengue endemic region as it recorded the highest dengue incidence for the past 10 years as compared to other states in Malaysia. Historically, Selangor experienced two times epidemic year which were in 2015 and 2019. In the year 2015, Selangor recorded about 63, 198 dengue incidences over a total dengue incidence of 120, 834 with dengue incidence rate of 1024 cases per 100,000 populations [26]. Meanwhile, in the year 2019, the dengue incidences recorded in Selangor were about 66, 959 over total dengue incidences of 120, 871 with dengue incidence rate of 1026 cases per 100,000 populations [26].

Therefore, the study area covers only Selangor state for the analysis of spatial and temporal analysis of dengue incidence. Selangor state is one of the states in Malaysia which is located between 3.0738° North latitude and 101.5183° East latitude and it is situated at the west coast of peninsular Malaysia encircling the capital of Kuala Lumpur. The total land area of Selangor is estimated to be approximately 8,000 km². Selangor is divided into 9 districts namely Gombak, Hulu Langat, Hulu Selangor, Klang, Kuala Langat, Kuala Selangor, Petaling, SabakBernam and Sepang. The state capital of Selangor is Shah Alam and its royal capital is Klang. In 2019, the population in Selangor was estimated to be about 6.53 million [27]. Figure 1 shows the map of Peninsular Malaysia and districts of Selangor.

Figure 1. Spatial Map of Peninsular Malaysia and Districts of Selangor

2.6. The Data

Table 1 summarizes the secondary data used in the study. Noted that the study used daily data which comprised of 1st January 2011 to 31st December 2017. Daily dengue incidence refers to the number of dengue cases recorded by the hospital in which the patients were diagnosed and confirmed being infected by one of the serotypes (DENV-1, ENV-2, DENV-3 and DENV-4) based on the blood test. The data were collected at Vector Borne and Infectious Diseases Sector from the Ministry of Health Malaysia through approved medical ethics from Malaysia Medical Research and Ethics Committee.

Table 1. List of Variables

Variables	Measurement Unit
Dengue	Count
Average Temperature	°C
Rainfall	mm
Relative Humidity	%
Population	Count

Population data were collected from the Department of Statistics Malaysia (DOSM) in order to obtain dengue incidence rate (DIR). The dengue incidence rate (DIR) was computed in the study which can be explained as y_{it} which y is the number of dengue cases at districts, i ($i=1,2, \dots, I$) at time, t ($t=1,2,\dots,T$) divided by p_{it} which p refers to total population of the districts, i at time, t and then multiplied with 100,000 populations. This can be written as follows;

$$DIR = \frac{y_{it}}{p_{it}} \times 100,000 \quad (8)$$

Hence, *DIR* actually explains the number of dengue incidence reported per 100,000 populations in a specific

district at specific time.

Climatic factors such as temperature, rainfall and relative humidity were also used in analyzing the spatial and temporal trend of Selangor. Temperature is defined as the daily average temperature measured in degree Celsius. Meanwhile, rainfall is the daily amount of rainfall measured in millimeter. Relative humidity refers to the daily atmosphere's water vapour which exists in gas particles at a certain level of temperature and it is measured in percentage. These climatic factors were obtained from Malaysia Meteorological Department and NASA climate data online (<https://power.larc.nasa.gov/data-access-viewer/>). All the data used in the study were collected for every district of Selangor.

The next section displays and reports the results of the study based on analyses that have been done.

3. Results

3.1. Descriptive Analysis

Figure 2 displays the trend of dengue incidence rate per 100,000 populations in Selangor for 7 years (2011- 2017). Across 7 years, Selangor experienced the lowest DIR per 100,000 populations in 2011 which was about 158 cases per 100,000 population as compared to 1024 cases per 100,000 populations recorded in 2015. In addition to this, it was found that the reported dengue in Selangor contributed about 52% of the total dengue cases in Malaysia. This would imply that Selangor is the dengue endemic state in Malaysia.

Instead of looking at DIR of Selangor as a whole, the study is more meaningful and interesting when it explores the DIR of each district. Table 2 reports the summary of variables in the dataset for dengue incidence rate per 100,000 populations and climatic information for 9

districts in Selangor. The annual average DIR is defined as the summation of the annual DIR over 7 years while the maximum annual DIR refers to the district with the highest annual DIR recorded across 7 years. Among 9 districts, Petaling district contributed the highest maximum annual DIR with about 1243 cases per 100,000 populations followed by Hulu Langat district which contributed about 1147 cases per 100,000 populations, DIR Gombak district reported about 1026 cases per 100,000 populations and the least maximum annual DIR was SabakBernam district which recorded about 271 cases per 100,000 population.

Figure 2. Dengue Incidence per 100,000 Populations in Selangor

Table 2. Descriptive Statistics for DIR

Districts	TotalDengue (2011-2017)	Annual Average DIR	Maximum Annual DIR
Gombak	32496	605.16	1025.79 (2015)
Hulu Langat	62349	686.63	1147.00 (2015)
Hulu Selangor	5422	347.26	591.36 (2014)
Klang	37693	554.68	862.57 (2015)
Kuala Langat	5549	309.18	545.87 (2016)
Kuala Selangor	3685	224.38	381.37 (2015)
Petaling	98515	689.60	1243.27 (2015)
SabakBernam	1401	170.04	270.97 (2015)
Selangor	8043	470.93	1135.50 (2015)

The spatial and temporal trend is shown in the next section.

3.2. Spatial and Temporal Analysis of Selangor

This section reports the spatial and temporal analysis of monthly average trend for dengue incidence per 100,000 populations and the trend for the climatic factors used in the study (See Figure3 (a), (b) and (c)). Figure3 (a) exhibits the monthly average of dengue incidence per 100,000 populations in Selangor for 7 years (2011 - 2017). Based on the plot, it clearly showed that the trend of dengue outbreak in Selangor endemic was in the beginning of the year (January), middle year (July) and end of the year (December) with January having the highest incidence with about 60 cases recorded for every 100,000 populations.

Figure3 (b) and Figure3 (c) display the temporal trend of monthly average of rainfall (mm) (bar chart) and monthly average of average temperature ($^{\circ}\text{C}$) (line plot) and monthly average of relative humidity (%) respectively. For the monthly average of rainfall, Selangor received a high amount of rainfall in two cycles which were during March-May and August-November. The lowest amount of rainfall was spotted in the month of February ($<4\text{mm}$) while the highest amount of rainfall was in the month of November ($>10\text{mm}$). In addition to this, the monthly average of temperature in Selangor was averaged between 26°C to 28°C . The average temperature was high in the months of March-June with May having the highest average temperature throughout the other months. However, the monthly average temperature showed a decrease pattern from June till December. Meanwhile, the monthly average relative humidity was averaged in a range between 78% to 85% with an increasing trend. Previous studies had shown that the current dengue incidence correlate with the previous climatic factors. This implies that the dengue incidence which was high in January and July correlated to the previous rainy season that occurred in November and May as it naturally created an abundance of *Aedes* mosquito breeding sites. Since *Aedes* mosquito is a climate-dependent vector, the warm and humid conditions throughout the year have sped up the mosquito development and caused the outbreak to disperse quickly in a specific region.

Figure 3. Monthly Average of ; (a) Dengue Incidence Rate per 100,000 populations (b) Rainfall (mm) and Average Temperature (°C) and (c) Relative Humidity in Selangor from 2013 to 2017

3.3. Spatial and Temporal Analysis by Districts of Selangor

To obtain additional information on the annual trend and behavioural of dengue outbreak in a specific region, this study analysed the annual trend of dengue incidence as shown by Figure 4. This can help local authorities to identify the severity level of dengue incidence per 100,000 populations among districts. The annual trend of dengue incidence was plotted for each district across 7 years (2011 – 2017). Overall, Gombak, Hulu Langat, Klang, Kuala Selangor, Petaling, SabakBernam and Sepang districts recorded the highest dengue incidence in the year 2015. This is the reason why Selangor recorded more than 50% of the overall total dengue incidence in Malaysia in the year 2015. Meanwhile, Hulu Selangor and Kuala Langat districts recorded the highest dengue incidence in the year 2014 and 2016 respectively.

Moreover, the figure clearly showed that in the year 2011 and 2012, the annual dengue incidence for all the districts were around 200 incidences per 100,000 populations. However, as the years increased, the annual trend for some districts changed with some districts having an extreme increase in annual dengue incidence. For instance, Petaling, Gombak, Hulu Langat, Klang and Sepang districts displayed a dramatic increase in the annual dengue incidence (See Figure 4). The dramatic increase in dengue incidence of these districts was due to the potential factors such as uncontrolled and rapid development which resulted into abundance of mosquitos breeding sites to breed and migration of infected people from rural area to urban area which caused the outbreak. Thus, people who want to visit these dengue infected areas should be aware of this alarming dengue outbreak.

In addition to this, Hulu Selangor, Kuala Langat, Kuala Selangor and SabakBernam districts recorded a steady growth in the annual dengue incidence within the range 200 to 600 incidences recorded in every 100,000 populations. Throughout the years, SabakBernam district

was less infected with dengue outbreak as the annual trend of dengue incidence in this district was minor as compared to other districts. Geographically, SabakBernam district is situated at the northwestern corner of Selangor and the furthest district from the state capital of Selangor which had the smallest population among other districts [27]. The main activity of this district is agriculture activity and the number of people who migrate to this district is minimal. This could encourage the minor dengue incidence in SabakBernam district.

Figure 4. Annual Dengue Incidence per 100,000 populations by Districts of Selangor

There are many other potential factors that could contribute to the increase in dengue incidence in a specific region. The study also analysed the monthly average of dengue incidence per 100,000 population for the districts of Selangor. This will supply vital information to the local authorities as to plan early warning systems to reduce the incidences. Figure 5 portrays the spatial map of monthly average of dengue incidence per 100,000 populations for the 9 districts of Selangor. The spatial map was produced as to identify the hot spot area of dengue in Selangor so that

effective and appropriate vector program can later be planned by local authorities in specific month for a specific region. The monthly dengue incidences were averaged over 7 years (2011-2017). The endemicity level of the dengue incidences was categorized by the red colour in which the lightest red represents the low dengue incidences while the darkest red represents the high dengue incidences in a specific region.

Firstly, Figure 5 shows a variation in dengue incidence among the districts. Throughout the twelve months, the dengue incidence was endemic in the months of January and February in all districts of Selangor as Gombak, Hulu Langat, Klang and Petaling presented high dengue incidence (dark red). Nevertheless, the dengue incidence for these districts decreased in the months of April and May before the outbreak started in the months of June, July and August. Then, in the months of September, October and November, the dengue incidence dropped before it continued to worsen in the month of December. Since the international airport is located in Sepang district, thousands of people including tourists and local people enter Selangor and they are likely to visit Kuala Lumpur, Putrajaya,

Gombak, Hulu Langat, Klang and Petaling districts. Thus, the migration of people will accelerate the transmission of dengue virus which later causes an outbreak in a specific region. Furthermore, these districts experience a very rapid and uncontrolled development process which encourages and creates the abundance of natural *Aedes* mosquito's breeding sites that cause the spread of dengue outbreak. In comparison, Sabak Bernam and Kuala Selangor districts exhibited a low dengue incidence as they were covered with light red which indicated about 10 to 30 incidences per 100,000 populations throughout the twelve months. Meanwhile, Kuala Langat and Hulu Selangor districts were also the districts with less infected dengue virus throughout the twelve months. However, both of these districts had a moderate dengue incidence as they experienced a small increase in dengue incidence in the months of January and December respectively. For Sepang district, the endemicity level for this area was medium infection with dengue virus as shown by the spatial map of monthly average of dengue incidence since the district was mostly covered with light red throughout the months except for the months of February, March, May, June and July.

Figure 5. Spatial Map of Monthly Average of Dengue Incidence per 100,000 Populations in Selangor

Overall, it clearly revealed that the spatial map of monthly average and temporal trend of annual dengue incidence were able to supply preliminary information in determining the dengue hot spot areas in Selangor. The changes in dengue incidence in a specific region from time to time is driven by many factors including climatic factors and non-climatic factors which are the significant factors that cause the transmission of dengue virus in a specific region. The hot spot dengue areas that have been identified and any areas that have the tendency to be endemic area need to be frequently and extensively monitored by the local authorities so that the dengue outbreak in these areas is under control.

3.4. Clustering Analysis of Dengue Incidence by Districts in Selangor

The clustering of dengue incidence by districts in Selangor was applied using Wards Hierarchical Clustering method. From what have been explained in the previous section, this approach clusters the districts based on the minimum variance within clusters or a small increase in sum of squares. Figure 6 displays the dendrogram of 9 districts of Selangor. Based on the dendrogram obtained, there were 8 stages in the procedure. In the early stage (Stage 1), it can be seen that Kuala Selangor and Sabak Bernam districts were clustered together due to the smallest increase in the sum of squares. Then, the clustering process clustered Gombak and Hulu Langat districts (Stage 2) and Hulu Selangor and Kuala Langat districts (Stage 3). In the middle stage (Stage 4), the districts which were clustered in Stage 1 and Stage 3 were then merged to make a cluster. After that, in stage 5, Petaling district was clustered with districts in Stage 2. Noted that in stage 6, the clustering process joined Klang district with districts in stage 5 which formed another cluster. Due to principle held by Wards minimum variance method, Sepang district was the only district which made up the next cluster. During the final stage (Stage 8), stage 6 and stage 7 were joined together to form a cluster which was the process of linking all the observations.

Based on the dendrogram presented, it is now clear to justify on the number of clusters that can make up the dengue incidence in Selangor. The dendrogram was able to cluster the 9 districts of Selangor into 3 different clusters which were Cluster 1 (high endemic areas), Cluster 2

(medium endemic area) and Cluster 3 (low endemic areas) in Wards hierarchical linkage clustering method. The first cluster of Klang, Petaling, Gombak and Hulu Langat districts were clustered as high dengue endemic areas in Selangor. Sepang was classified in second cluster which acted as medium endemic area while Kuala Selangor, Sabak Bernam, Hulu Selangor and Kuala Langat districts were clustered in the third cluster which implied the regions with low endemic areas. The list of districts according to clusters obtained is shown in Table 3. Figure 7 provides a cluster plot which shows how the 9 districts of Selangor were being clustered together to form the three clusters. The cluster plot clearly shows the districts were grouped into 3 clusters which was similar to what have been portrayed by dendrogram in Figure 6.

Figure 6. Dendrogram

Table 3. Cluster of Districts

Cluster 1 High Endemic Areas	Cluster 2 Medium Endemic Area	Cluster 3 Low Endemic Areas
Klang	Sepang	Kuala Selangor
Petaling		Sabak Bernam
Gombak		Hulu Selangor
Hulu Langat		Kuala Langat

Figure 7. Cluster Plot

4. Conclusions

Dengue incidence in Selangor has worsened for the past few years. It has become the main reason of hospital admissions and deaths. The main objective of the study is to determine the dengue endemic regions in Selangor by means of hierarchical wards clustering analysis. A secondary data was collected from several sources. Firstly, this study has come out with exploratory analysis which reported the annual trend of dengue incidence for the whole of Selangor and some descriptive statistics (the total number of dengue, the annual average of dengue incidence per 100,000 populations and the annual maximum incidence of dengue) summarized for each district of Selangor. Based on the summary statistics, Gombak, Hulu Langat and Petaling districts had the highest annual dengue incidence compared to other districts in Selangor. In addition to this, this study also produced a temporal trend of the annual dengue incidence per 100,000 populations and spatial map of monthly average of dengue incidence per 100,000 populations for the 9 districts of Selangor. The graphical visualization displayed by the spatial map was able to identify the endemicity level of dengue throughout the twelve months. For the twelve months, this study found that the trend of dengue incidence was alarming in the months of January, February, June, July, August and December. The main findings obtained from Wards hierarchical linkage clustering analysis managed to group the regions of Selangor into three clusters in which Klang, Petaling, Gombak and Hulu Langat districts were grouped as cluster 1 and classified as high dengue endemic areas; Sepang district was clustered as the second cluster and

classified as medium dengue endemic area; while Kuala Selangor, Sabak Bernam, Hulu Selangor and Kuala Langat districts formed the third cluster and classified as low dengue endemic areas. These significant findings of the study can provide vital information to the readers as to be aware on this alarming issue while local authorities and public health institution can plan for effective dengue early warning systems so that the dengue outbreak in the identified hot spot dengue area can be controlled to safeguard the community from dengue outbreak. As a recommendation to future study, this study can be extended by investigating the association between climatic factors and dengue incidence in Selangor so that any potential climatic factors which contribute to the dengue incidence in Selangor can be further explained.

Acknowledgements

We would like to express our appreciation and record our gratitude to the Ministry of Higher Education (MOHE) and Universiti Teknologi MARA for financial support under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2016/STG06/UITM/02/2). The authors are also grateful to the Editor, the Associate Editor and anonymous referees for their insightful comments and suggestions.

REFERENCES

- [1] Mudin, R. N. "Dengue Incidence and the Prevention and

- Control Program in Malaysia". *The International Medical Journal Malaysia*, vol 14(1) (2015): 5-9.
- [2] Central for Disease Control and Prevention (CDC). "Dengue". <https://www.cdc.gov/dengue/>, 2016.
- [3] World Health Organization (WHO). Dengue and Severe Dengue. <http://www.who.int/mediacentre/factsheets/fs117/en/>, 2016.
- [4] Urbanova, K. "Analysis of Risk Factors Associated with Dengue in Southeast Asia". Diploma Thesis, Mendel University, 2015.
- [5] Githeko, A. K., Lindsay, S. W., Confalonieri, U. E. and Patz, J. A. "Climate Change and Vector-Borne Diseases: A Regional Analysis". *Bulletin of World Health Organization*, vol. 78(9) (2000): 1136-1147.
- [6] Morin, C. W., Comrie, A. C. and Ernst, K. "Climate and Dengue Transmission: Evidence and Implications". *Environmental Health and Perspectives*, vol. 121(11-12) (2013): 1264-1272.
- [7] Yaacob, W. F. W., Azaki, W. W., Lazim, M. A. and Wah, Y. B. "Modelling Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) Outbreak Using Poisson and Negative Binomial Model". *World Academy of Science, Engineering and Technology*, vol. 38 (2010): 903-908. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.309.7294>
- [8] Tuladhar, R., Singh, A., Varma and Choudhary, D. K. "Climate Factors Influencing Dengue Incidence in an Epidemic Area of Nepal". *BMC Res Notes*, Vol 12(131) (2019). <https://doi.org/10.1186/s13104-019-4185-4>
- [9] Arcari, P., Tappe, N. and Pfueller, S. "Regional Variability in Relationship Between Climate and Dengue/DHF In Indonesia". *Singapore Journal of Tropical Geography*, vol. 28(3) (2007): 251-272. <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9493.2007.00300.x>
- [10] Azman, N. N. and Karim, S. A. A. "Assessing Climate Factors on Dengue Spreading in State of Perak". *Journal of Physics: Conference Series*, vol. 1123(1) (2018). <https://iopscience.iop.org/article/10.1088/17426596/1123/1/012026/pdf>
- [11] Spencer, N. H. Essentials of Multivariate Data Analysis. CRC Press. Taylor & Francis Group, (2014).
- [12] Sage Publication, "Discriminant Analysis", (2015). <http://www.uk.sagepub.com/burns/website%20material/Chapter%2025%20-%20Discriminant%20Analysis.pdf>
- [13] Johnson, R. and Wichern, D. Applied multivariate Statistical Analysis, 6th edition, Pearson New International Edition, (2014).
- [14] Huang, C. C., Tam, T., Chern, Y. R., Lung, S. C., Chen, N. T. and Wu, C. D. "Spatial Clustering of Dengue Fever Incidence and Its Association with Surrounding Greenness". *International Journal of Environmental Research and Public Health*, vol. 15(9) (2018): 1869. doi:10.3390/ijerph15091869
- [15] Silva, J. C. B. D. and Machado, C. J. S. "Associations Between Dengue And Socio-Environmental Variables In Capitals Of The Brazilian Northeast By Cluster Analysis". *Ambiente & Sociedade*, 21(2019): e01332. <https://dx.doi.org/10.1590/1809-4422asoc0133r2vu18l4td>
- [16] Sangkaew, S., Tan, L. K., Ng, L. C., Ferguson, N. M. and Dorigatti, I. "Using Cluster Analysis To Reconstruct Dengue Exposure Patterns From Cross-Sectional Serological Studies in Singapore". *Parasites Vectors* 13, 32 (2020). <https://doi.org/10.1186/s13071-020-3898-5>
- [17] Vandhana, S. and Anuradha, J. "Dengue Prediction Using Hierarchical Clustering Methods". *International Conference on Design Science Research in Information Systems and Technology*, (2018): 157-168.
- [18] Panciano, J. A., Polanco, W. and Barrios, M. "Dengue Outbreaks Pattern in Southern Guatemala". 2018.
- [19] Sun, W., Xue, L. and Xie, X. "Spatial-Temporal Distribution of Dengue and Climate Characteristics for Two Clusters in Sri Lanka From 2012 To 2016". *Scientific Reports* 7, (2017):12884. doi:10.1038/s41598-017-13163-z
- [20] Mutheneni, S. R., Mopuri, R., Naish, S., Gunti, D. and Upadhyayula, S. M. "Spatial Distribution and Cluster Analysis of Dengue Using Self Organizing Maps in Andhra Pradesh, India, 2011-2013". *Parasite Epidemiology and Control*, vol. 3(1) (2016): 52-61. doi:10.1016/j.parepi.2016.11.001
- [21] Casey, E. Handbook of Digital Forensics and Investigation, Academic Press, (2009).
- [22] Murtagh, F. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?". *Journal of Classification*, vol. 31(3) (2014): 274-295. <https://doi.org/10.1007/s00357-014-9161-z>
- [23] Lattin, J., Carroll, J. D. and Green, P. E. Analyzing Multivariate Data, Thompson Learning, (2003).
- [24] Hair, J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. Multivariate Data Analysis, 4th ed, 1995, Prentice Hall.
- [25] Strauss, T. and von Maltitz, M. J. "Generalising Ward's Method for Use with Manhattan Distances". *PloS one*, vol. 12(1) (2017): e0168288. doi:10.1371/journal.pone.0168288
- [26] iDengue. "Dengue Statistics". (2020). https://idengue.mysa.gov.my/ide_v3/pdf/statistik.pdf#page=3
- [27] Department of Statistics Malaysia (DOSM). "Malaysia at a Glance", (2020). <http://www.dosm.gov.my>