Statistical Analysis and Clustering of Dengue Incidents and Weather Data of Bangladesh Using K-Means Clustering

Md. Monirul Islam[1], Fazly Rabbi[2], Javed Meandad[3], K. M. Rafiqul Alam[2], and Jia Uddin[4](✉)

[1] Department of Computer Science and Engineering,
University of Information Technology and Sciences, Dhaka 1212, Bangladesh
monirul.islam@uits.edu.bd
[2] Department of Statistics, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh
[3] Department of Meteorology, University of Dhaka, Dhaka 1000, Bangladesh
[4] AI and Big Data Department, Endicott College, Woosong University, Daejeon, South Korea
jia.uddin@wsu.ac.kr

Abstract. In recent years, Dengue incidents have become a big issue in the public health sector in Bangladesh. This study mainly attempts to investigate the association among different weather parameters and the number of Dengue incidents in monthly frequency. The weather data has been collected from Bangladesh Meteorological Department and several Dengue incidents have been collected from the Institute of Epidemiology, Disease Control and Research (IEDCR), finally, the merged dataset has been used in this study. We found the Rainfall, Minimum Air Temperature, and Month have the highest association with Dengue incidents. Increasing the value of these variables increases the chance of Dengue incidents. After performing K-Means Clustering, we found 4 number of weather clusters with different numbers of Dengue incidents. Dengue incidents perform 431.23 on an average with 7.96 on an average value of Month, 32.22 on an average value of Maximum Temperature, 25.04 on an average value of Minimum Temperature, 84.84 on an average value of Humidity, and 49.94 on an average value of Rainfall.

Keywords: Dengue incident · K-means clustering · Weather data of Bangladesh · EDA · Anova test

1 Introduction

Dengue fever is the most widespread, lethal, and symbol mosquito-borne viral disease in Bangladesh [1]. It is a climate-sensitive disease spread by the bite of female Aedes mosquitos infected with Flaviviridae dengue virus serotypes [2–4]. According to a recent study, approximately 390 million dengue cases are reported globally each year. Dengue virus distribution and burden are primarily concentrated in tropical as well as subtropical regions [5, 6]. The triggering factors for dengue dynamics are said to be influenced by 3 central risk factors including biological, sociological, along environmental [7]. The environmental aspect, specifically meteorological variables, was examined in this

A. K. M. M. Islam et al. (Eds.): ICBBDB 2021, CCIS 1550, pp. 64–75, 2022.
https://doi.org/10.1007/978-3-031-17181-9_6

study to better understand the association with dengue epidemics. Aedesalbopictus and Aedesaegypti transmit the five-dengue virus serotypes, which are highly sensitive to environmental factors. Temperature, rainfall, as well as humidity, can all have an impact on dengue epidemiology by increasing mosquito development, population growth rate, and virus reproduction [8, 9]. Infestation usually starts during the monsoon season, when the humidity along with temperature conditions are ideal for the development of the mosquito embryo.

Bangladesh appears to be a suitable habitat for the primary dengue vector and its increased spread. The primary cause of this disease in Bangladesh has been identified as environmental effects (average rainfall, humidity, and temperature) as well as inappropriate urbanization [10]. In recent years, the country's rising atmospheric temperature has likely contributed to the development of mosquito larvae and pupa habitats, particularly in the cities of Dhaka and Chattogram, leading to an increase in the frequency of dengue contagion.

Although the association between dengue incidence as well as weather patterns has been extensively considered, little attention has been paid to determining the threshold properties of weather variability on dengue spread. To reduce dengue risk effectively, a suitable alert system for dengue incidents based on the threshold effect of weather variables is required. The goal of this study is to investigate the nonlinear relationships between weather variables as well as dengue incidence in Bangladesh. To meet the research objectives, the study's specific objectives have been defined as follows:

- To investigate the relationship between meteorological variables as well as the number of dengue cases atmonthly scale; and
- To develop an algorithm for predicting dengue incidence based on meteorological parameters.

The rest of the paper is structured as follows. Section 2 states the literation review. The proposed methodology is discussed in Sect. 3. Section 4 dictates the experimental setup and result. Finally, we conclude the paper in Sect. 5.

2 Literature Review

A lot of research has been done on the relation between dengue and meteorological data in China. In [11], The authors investigated the relationship between dengue occurrence and potential risk factors to progress a predicting model for dengue control along with prevention in Guangzhou, China. The authors [12] predicted that a time series analysis of Mosquito Density and Meteorological Factors on Dengue Fever Epidemics would be conducted. In Zhongshan, a time-series regression tree model was used to estimate the hierarchical relationship between reported autochthonous DF cases and potential risk factors such as DF surveillance system timeliness like median time interval between symptom onset date and diagnosis date, mosquito density, imported cases, and meteorological factors, China from 2001 to 2013 in [13]. The authors proposed in [14] to investigate individual risk factors for dengue virus infection in Guangdong Province and to provide a scientific foundation for future DF prevention and control. The authors [15] used boosted regression trees (BRT) to investigate the delayed effect of meteorological factors, as well as the relationships between five climatic variables and the risk for DF, from 2005 to 2011, to control the best timing as well as a strategy for familiarizing such protective actions.

In Bangladesh, the impact of meteorological factors on dengue mosquitoes is not researched. That is why; we tried to explore the relationship between the dengue and meteorological data of Bangladesh.

3 Methodology

Figure 1 demonstrates the detailed block diagram of the proposed methodology. We utilized two datasets including the dengue and weather datasets in the merge. Then we did the preprocessing, EDA, standardization, and k-means clustering part.

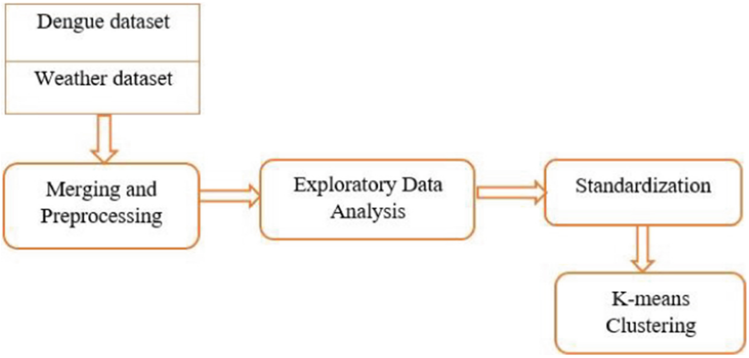

Fig. 1. Block diagram of proposed methodology.

3.1 Data Description

The dataset was collected from the Bangladesh Meteorological Department and Institute of Epidemiology, Disease Control and Research (IEDCR). The dataset is grouped into two parts containing the dengue dataset from IEDCR and the weather dataset from Bangladesh Meteorological Department. There are 7 columns including year, month, minimum average air temperature, maximum average air temperature, humidity, rainfall, and several dengue patients and 135 rows. The dataset contains the record from 2008 to 2019.

3.2 Merging and Preprocessing

We merged both datasets for analyzing our experiment. The first step is the data preprocessing. For the experimental setup, we used the python programming language. In this step, we import all necessary libraries and the proposed dataset.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a process to explore the unseen insights from the data. Through EDA, we can extract various information from the data set including null values, outliers, the relation among the dataset, etc. which helps make various decisions as well as in selecting the machine-learning model. EDA is done through various types of statistical methods, data visualization, etc.

Descriptive Statistics
Table 1 displays the descriptive statistics of the variables. We found RAINFALL and DENGUE have the highest Standard Deviations that means the most variability of these variables. On the other hand, MAX has the lowest variability that means the Tropical weather condition of Bangladesh.

Table 1. Descriptive statistics of dataset

	Year	Month	Min	Max	Humidity	Rainfall	Dengue
Count	13	134	134	134	134	134	134
Mean	2013	6.42	20.94	30.84	80.11	27.84	211.58
Std	3.23	3.49	4.83	2.85	5.25	98.39	451.21
Min	2008	1	10.59	23.52	67.54	0	0
25%	2010	3	16.40	29.27	77.18	0.547	0
50%	2013	6	22.93	31.98	80.09	6.00	36
75%	2016	9	25.27	32.67	84.77	12.77	187
Max	2019	12	26.48	35.76	88.38	689.16	3087

Range and Outliers of the Variables
Figure 2 shows the visualizing of the Box Plot. We found DENGUE has the largest range among other variables. At the same time, DENGUE and RAINFALL show some outliers. Especially DENGUE contains most of the extreme values. MIN, MAX, and HUMIDITY variables have the lowest range without any outliers.

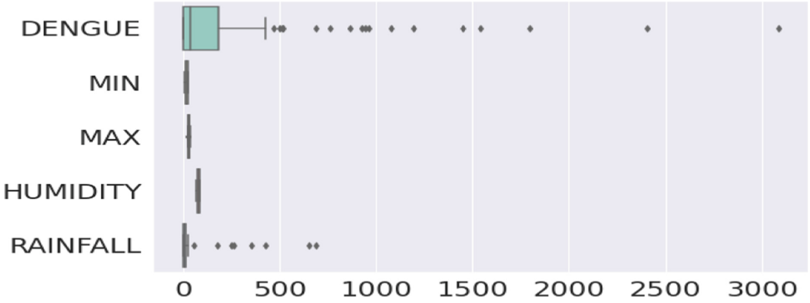

Fig. 2. Box plot of the dataset

Correlation Checking

Table 2 shows the weak correlation with DENGUE incidents and weather parameter variables, but performing the Hypothesis test for association checking using Pearson's correlation shows the statistically significant relationship with the given variables. This result shows DENGUE incidents have the highest relationship with RAINFALL.

Table 2. Pearson's correlation

Variable 1	Variable 2	Pearson's correlation	Result	Decision
MIN	Dengue	0.3013	stat = 0.301, p = 0.000403	Dependent
MAX	Dengue	0.1778	stat = 0.178, p = 0.039837	Dependent
Humidity	Dengue	0.2753	stat = 0.275, p = 0.001281	Dependent
Rainfall	Dengue	0.3686	stat = 0.369, p = 0.000012	Dependent

Monthly Dengue & Weather Parameter Pivot Table

Table 3 shows the MONTH-wise Median value of MAX, MIN, RAINFALL, HUMIDITY, and DENGUE.

Table 3. The month-wise median value of dengue and weather parameters

	Median value				
Month	Dengue	Humidity	Max	Min	Rainfall
1	3.0	78.743318	24.944470	12.248925	0.030415
2	1.5	71.853008	28.915002	14.943736	0.438776
3	2.0	70.321659	32.391073	19.581158	0.771429
4	0.0	75.647619	33.902323	22.782796	2.713333
5	8.0	78.580927	33.821106	24.189299	8.815668
6	28.0	84.183810	32.930571	25.457972	16.608571
7	171.0	86.428571	31.914706	25.465268	17.993548
8	346.0	86.494009	32.157235	25.560893	14.363134
9	334.0	85.295238	32.556765	25.294470	9.651429
10	184.0	82.778802	32.106728	23.334911	6.281106
11	36.0	78.782857	30.070476	18.194654	0.379048
12	11.0	80.924424	26.389862	14.305625	0.025806

Most Dengue Incident by Month

Figure 3 describes the monthly median DENGUE incidents. It shows August and September are the most DENGUE incidental months. We can say that dengue broke

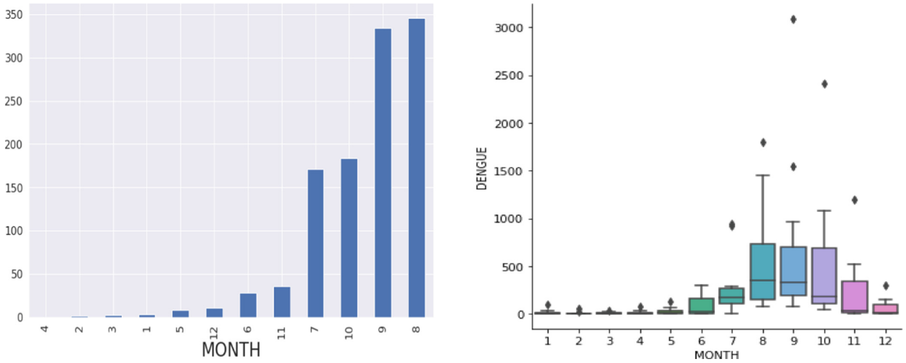

Fig. 3. Monthly DENGUGE incidents (median)

out more in August, September, October, and November of the year. First 4 months of the year, the number of dengue patients is less.

Figure 4 clearly describes that Dengue incidence increase with the increasing value of Humidity. Most of the Dengue incidents occur within the Humidity range of 80 and above.

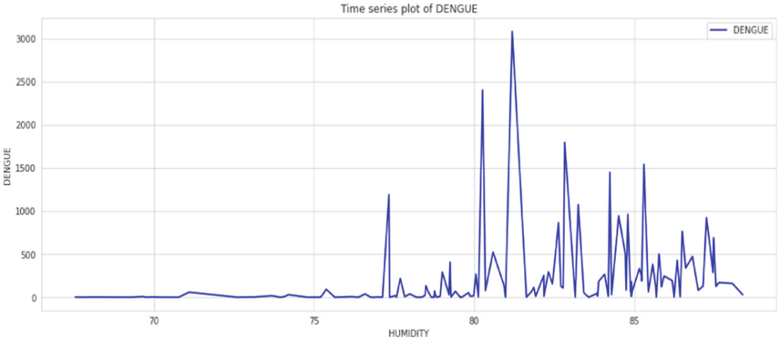

Fig. 4. Time series plot of dengue versus humidity

Figure 5 clearly describes that most of the Dengue incidents occur when the minimum temperature cross 22 °C.

Figure 6 describes that the number of Dengue incidence has increased over the year and in 2018 the highest dengue incidents occurred in Bangladesh. The second highest Dengue incidents occurred in 2016.

ANOVA Test

To test the difference of weather parameters in each month, we performed an ANOVA test. Table 4 shows the test result. As per the test result of each separate variable, we found that Null Hypothesis is Rejected for each test. DENGUE incidents, MIN temperature, MAX temperature, and HUMIDITY aren't the same in each month.

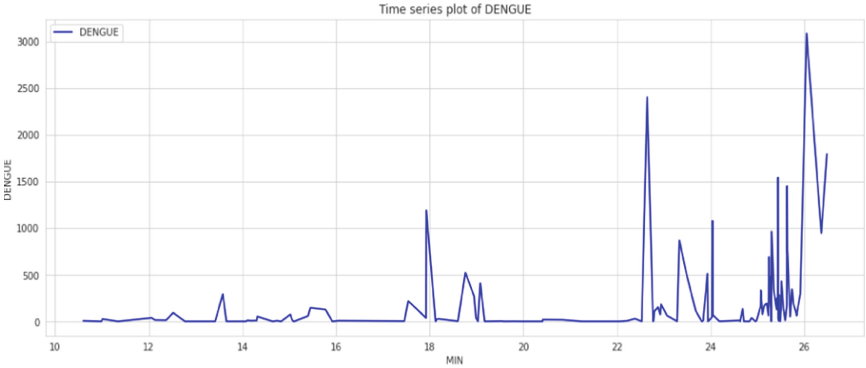

Fig. 5. Time series plot of dengue versus minimum air temperature

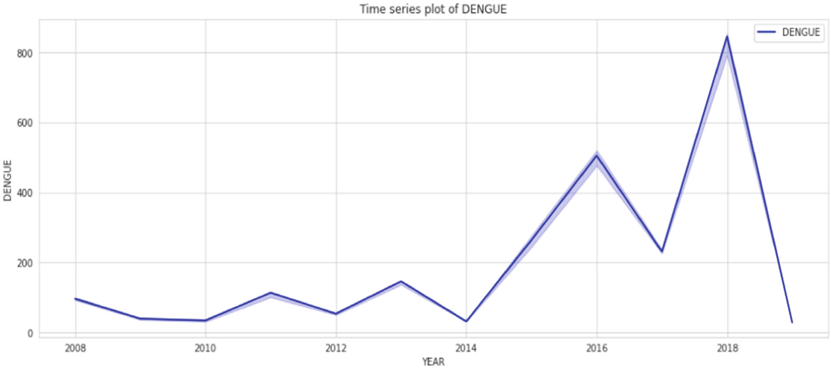

Fig. 6. Time series plot of dengue versus year

Table 4. ANOVA test result

Variable	Hypothesis	Result	Decision
Dengue	Are the Dengue incidents same in every Month?	stat = 4.601, p = 0.000	Null Hypothesis REJECTED
Min	Is the MIN temperature the same in every Month?	stat = 124.502, p = 0.000	Null Hypothesis REJECTED
Max	Is the MAX temperature the same in every Month?	stat = 567.597, p = 0.000	Null Hypothesis REJECTED
Humidity	Is the HUMIDITY same in every Month?	stat = 55.193, p = 0.000	Null Hypothesis REJECTED

3.4 Standardization

Min-Max Normalizer scales using the maximum and minimum values of this variable, so that all data moves within a certain range to the normalized state. In this method each data is subtracted from its minimum value (of that variable), the subtraction is divided by the range. This is how each data is scaled.

Range: Range is the difference between the maximum and minimum values of a variable.

$$X_{new} = \frac{X_i - min(X)}{max(X) - min(X)} \tag{1}$$

K-means Clustering Algorithm

Figure 7 shows K-Means Clustering schematic view. It divides the unmarked dataset into dissimilar clusters. In this case, K denotes the number of pre-defined clusters that must be formed during the process; for example, if K is equal to 3, there will be 3 clusters; for K is equal to 4, there will be 4 clusters, and so on [16]. It enables us to cluster the data into various groups and provides a suitable method for discovering the group categories in the unmarked dataset but without training. It is a centroid-based model, with each cluster having its centroid. This algorithm's main goal is to lessen the sum of distances between data points along with their conforming clusters.

In short, this model starts with an unlabeled dataset, splits it into k clusters, and then recaps the process until it does not discovery the best clusters. In this algorithm, the k's value should be prearranged.

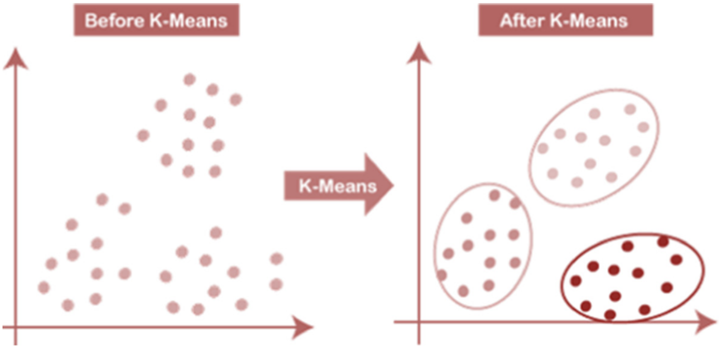

Fig. 7. K-means algorithm

The main task of the k-means clustering model is as follows.

- Iteratively determines the optimal value for K centre points or centroids.
- Appoints all data point to the nearest k-center. A cluster is formed by data points that are close to a specific k-center.

The Working Procedure of the K-Means Algorithm

- To determine the number of clusters, choose the number K.
- Choose a random set of K centroids.
- Designates all data point to the centroid that is nearest to it, which will form the predefined K clusters.
- Estimate the variance and move the cluster centroid.
- Repeat the 3 step, reassigning all data point to the cluster's current closest centroid.
- If any reassignment occurs, proceed to 4 number step; otherwise, proceed to complete.
- The model is finished.

In K-means clustering, we utilize Euclidean distance [17] to calculate the distance between any 2 data points shown in Eq. 2.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \cdots + (x_n - x'_n)^2} \tag{2}$$

4 Result and Discussion

The performance of the K-means clustering technique is predicated on the extremely efficient clusters that it produces. However, identifying the ideal number of clusters is a tough task. There are numerous methods for obtaining the best number of clusters, but we utilize the elbow method here because it is the most appropriate method for estimating the number of clusters or the value of K.

4.1 Elbow Method

The Elbow method is one of the most widely used methods for determining the optimal number of clusters. This technique makes use of the WCSS value idea. Within Cluster Sum of Squares (WCSS), we get the total variations within a cluster. The equation WCSS is shown in 3 containing 3 clusters.

$$\begin{aligned} WCSS = {} & \sum P_{i \text{ } in \text{ } Cluster1}(P_iC_1)^2 + \sum P_{i \text{ } in \text{ } Cluster2}(P_iC_2)^2 + \sum P_{i \text{ } in \text{ } Cluster3}(P_iC_3)^2 \\ & + \sum P_{i \text{ } in \text{ } Cluster4}(P_iC_4)^2 \end{aligned} \tag{3}$$

In the above equation of WCSS,

$\sum P_{i \text{ } in \text{ } Cluster1}(P_iC_1)^2$ is the sum of the squares of the distances between all data point and its center within a cluster1 and the same for other three term.

The elbow method employs the following steps to determine the optimal value of clusters:

- It does K-means clustering on a prearranged dataset for several K values.
- Computes the value of WCSS for each point of K.
- Draws a line connecting the designed WCSS values as well as the number of clusters K.

- If a point on the scheme look like an arm, that point is considered the most like value of K

The Fig. 8 shows that when the value of K drops from 1 to 2, the sum of squared distance decreases from 40 to 28. After that, after going down from 2 to 3, this value has come down from 28 to 18. Again after going down from 3 to 4, this value has come down from 28 to 10. So we will take the value of K as 4 optimal values.

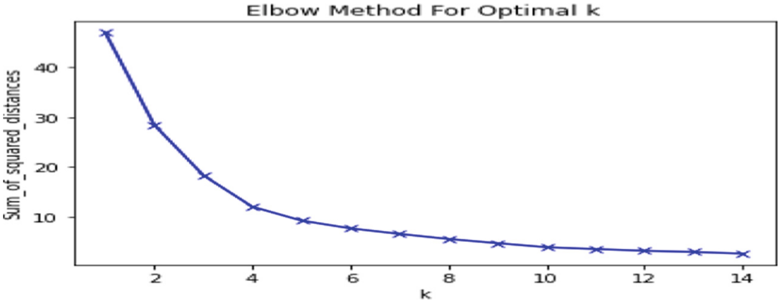

Fig. 8. Elbow method of optimal number of K

The clustering result has described in Table 5. It shows 4 clusters. They are- 0, 1, 2, and 3. In the 1st Cluster, an average 155 number of Dengue incidents occur when on average the Maximum Temperature is 28.15, Minimum Temperature is 16.50, Humidity is 79.98 and Rainfall is 1.457. In the 2nd Cluster, an average 15.71 number of Dengue incidents occur when on average the Maximum Temperature is 33.24, Minimum Temperature is 22.3, Humidity is 75.01 and Rainfall is 25.72.

In the 3rd Cluster, an average 12.08 number of Dengue incidents occur when on average the Maximum Temperature is 26.90, Minimum Temperature is 13.56, Humidity is 75.98 and Rainfall is 3.27. In the 4th Cluster, an average 431.23 number of Dengue incidents occur when on average the Maximum Temperature is 32.22, Minimum Temperature is 24.04, Humidity is 84.84 and Rainfall is 49.94.

Table 5. Clustering result

	Cluster mean					
Cluster	Month	Max	Min	Humidity	Rainfall	Dengue
0	11.500000	28.152681	16.503607	79.987957	1.457495	155.000000
1	3.968750	33.244019	22.354495	75.019204	25.724617	15.718750
2	1.500000	26.903963	13.569695	75.989150	3.274906	12.083333
3	7.946429	32.229173	25.041458	84.846397	49.942636	431.232143

5 Conclusion

Dengue incidents in Bangladesh have increased over the years. After performing Exploratory Data Analysis, it was found that dengue incidents increase every July to October and mostly it becomes extreme in August and September. Dengue Incidents have a relationship among the Minimum Air Temperature, Humidity & Rainfall. The clustering result shows four numbers of clusters with different values of the variables. The most extreme cluster shows on an average 431 dengue incidents with 7.96 on an average value of Month, 32.22 on the average value of Maximum Temperature, 25.04 on an average value of Minimum Temperature, 84.84 on an average value of Humidity, and 49.94 on an average value of Rainfall. Other clusters show 155, 15, and 12 Dengue incidents accordingly. The more environmental variable including sunshine, sosquito control program, and air quality could be helpful to identify the association of dengue incidents. We will try to include more variables and observations to expand our research findings. Time series forecasting using statistical models and artificial neural networks could be included to forecast dengue incidents based on an exogenous environmental variables.

References

1. Mutsuddy, P., Tahmina Jhora, S., Shamsuzzaman, A.K.M., Kaisar, S.M., Khan, M.N.A.: Dengue situation in Bangladesh: an epidemiological shift in terms of morbidity and mortality. Can. J. Infect. Dis. Med. Microbiol. (2019)
2. Huang, X., Clements, A.C., Williams, G., Milinovich, G., Hu, W.: A threshold analysis of dengue transmission in terms of weather variables and imported dengue cases in Australia. Emerg. Microbes Infect. **2**(1), 1–7 (2013)
3. Tran, B.L., Tseng, W.C., Chen, C.C., Liao, S.Y.: Estimating the threshold effects of climate on dengue: a case study of Taiwan. Int. J. Environ. Res. Public Health **17**(4), 1392 (2020)
4. Xu, J., et al.: Forecast of dengue cases in 20 Chinese cities based on the deep learning method. Int. J. Environ. Res. Public Health **17**(2), 453 (2020)
5. Duarte, J.L., Diaz-Quijano, F.A., Batista, A.C., Giatti, L.L.: Climatic variables associated with dengue incidence in a city of the Western Brazilian Amazon region. Revista da SociedadeBrasileira de Medicina Tropical **52** (2019)
6. Pinto, E., Coelho, M., Oliver, L., Massad, E.: The influence of climate variables on dengue in Singapore. Int. J. Environ. Health Res. **21**(6), 415–426 (2011)
7. Carvajal, T.M., Viacrusis, K.M., Hernandez, L.F.T., Ho, H.T., Amalin, D.M., Watanabe, K.: Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila. Philippines. BMC Infectious Diseases **18**(1), 1–15 (2018)
8. Iguchi, J.A., Seposo, X.T., Honda, Y.: Meteorological factors affecting dengue incidence in Davao. Philippines. BMC Public Health **18**(1), 1–10 (2018)
9. CorreiaFilho, W.L.F.: Influence of meteorological variables on dengue incidence in the municipality of Arapiraca, Alagoas, Brazil. Revista da SociedadeBrasileira de Medicina Tropical **50**(3), 309–314 (2017)
10. Noor, R.: Reemergence of dengue virus in Bangladesh: current fatality and the required knowledge. Tzu-Chi Med. J. **32**(3), 227 (2020)

11. Sang, S., et al.: Predicting local dengue transmission in Guangzhou, China, through the influence of imported cases, mosquito density and climate variability. PLoS ONE **9**(7), e102755 (2014). https://doi.org/10.1371/journal.pone.0102755.PMID:25019967;PMCID:PMC4097061
12. Shen, J.C., et al.: The impacts of mosquito density and meteorological factors on dengue fever epidemics in Guangzhou, China, 2006–2014: a time-series analysis. Biomed. Environ. Sci. **28**(5), 321–329 (2015). https://doi.org/10.3967/bes2015.046. PMID: 26055559
13. Liu, K.K., et al.: Risk assessment of dengue fever in Zhongshan, China: a time-series regression tree analysis. Epidemiol Infect. **145**(3), 451–461 (2017). https://doi.org/10.1017/S095026881600265X. Epub 2016 Nov 22 PMID: 27873572
14. Liu, J., et al.: Risk factors associated with dengue virus infection in Guangdong Province: a community-based case-control study. Int. J. Environ. Res. Public Health **16**(4), 617 (2019). https://doi.org/10.3390/ijerph16040617.PMID:30791547;PMCID:PMC6406885
15. Gu, H., et al.: Meteorological factors for dengue fever control and prevention in South China. Int. J. Environ. Res. Public Health **13**, 867 (2016). https://doi.org/10.3390/ijerph13090867
16. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern Recogn. **36**(2), 451–461 (2003)
17. Islam, M.M., Uddin, J., Kashem, M.A., Rabbi, F., Hasnat, M.W.: Design and implementation of an IoT system for predicting aqua fisheries using arduino and KNN. In: Intelligent Human Computer Interaction. IHCI 2020. Lecture Notes in Computer Science, vol. 12616. Springer, Cham. https://doi.org/10.1007/978-3-030-68452-5_11