International Journal of Engineering & Technology, 7 (4.30) (2018) 1-4

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Application FCM in Modelling DIR for Selangor Using Negative Binomial GAM

Nazeera Mohamad[1], Norziha Che Him[1]*, Mohd Saifullah Rusiman[1], Suliadi Sufahani[1], Siti Afiqah Muhammad Jamil[1]

[1]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia 84600 Pagoh, Muar, Johor, Malaysia
**Corresponding author E-mail: norziha@uthm.edu.my*

Abstract

This study attempts to obtain the best fitted model among two clusters which describe the relationship between dengue incidence rate (DIR) and relevant covariates such as climatic and non-climatic variables. The significant variables include amount of rainfall and number of rainy days with lag 0 until 3 months, number of locality and population density. Fuzzy C-Means clustering (FCM) was applied in clustering DIR data based on the value of membership function. The boundary of membership function has been set as 0.5. There are two clusters identified in this study with Cluster 1 consist of 569 data and Cluster 2 consist of 43 data. Then, this study developed models to predict future dengue incidences in Selangor by using negative binomial Generalised Additive Model (GAM). The result shows that the model able to be one of tools for future development in controlling and reducing the number of dengue cases particularly in Selangor, Malaysia as well as other states.

Keywords: *AIC; DIR; Fuzzy C-Means; membership function; negative binomial GAM.*

1. Introduction

Dengue is a disease which are known to be carried by mosquitoes. There are two mosquito species, *Aedes aegypti* and *Aedes albopictus* which are significant to the dissemination of dengue virus. Dengue virus is transmitted by the bite of female mosquitoes and *Aedes aegypti* is treated as super spreader of this risk disease. Meanwhile, *Aedes albopictus* also involve in the spreading of dengue virus and might transmit other infectious diseases such as chikunguya, yellow fever and Zika fever. Both species endemic in tropical and subtropical countries and the geographic range of these species has extremely expanded from year to year [1]. Today, the frequency of dengue disease showed increment every year and at the same time, the geographical distribution also expanded. Up to now, there are about 125 countries in the world that have been affected by dengue virus including Africa, America, Asia and Pacific [2]. Among these regions, the maximum burden is borne by countries of the Asia. As there is less effective vaccine to control the dengue outbreak resulted in this disease continue to be the most widely distributed disease and a public health burden worldwide.

Dengue has become a serious threat to public health in Malaysia. According to World Health Organization (WHO) Western Pacific Region (WPRO), Malaysia ranked second among Asian countries in term of the number of dengue cases reported and the total cases was estimated to be 41,486 cases with 88 deaths in 2009. In the early 1900s, the first dengue outbreak was reported in Penang [3] and dengue as the most rapidly spreading vector borne disease occurred frequently over the last 30 years and now reached 100,028 cases in 2016, with 231 deaths in the nation. Potential risk factors related to the growth of dengue cases influenced by globalisation, population growth, socio-economic and environmental changes and unplanned urbanisation [4]. In Malaysia, the most effective strategy to control the transmission of dengue virus using traditional measure by eradicating the breeding spots of *Aedes* mosquitoes [5]. On the contrary, another dengue control innovation has been developed by evaluating new mechanisms using the concept of insecticide, biological, gene, mechanical and environmental-based [6]. They believed that these new innovations could help in reducing the occurrence of dengue cases if a multi-pronged approach is integrating these new measurements.

From the past 17 years, dengue is considered as complex and unreasonable to recognise the contributors that associated with the dengue epidemic [7]. Earlier epidemiological studies in Venezuela, Malaysia and Bangladesh have proved significant associations between dengue cases and climate factors [8, 9, 10, 19]. For example, in South East Brazil, a recent study investigated the most significant climatic factors are temperature and precipitation which have greater influenced towards the dengue incidence [11]. This is due to a humid and warm condition which encourage the development of mosquito and leads to the increase of mosquito breeding site and results in the rise of dengue cases in Brazil. Furthermore, the same authors highlighted the importance of recognition the non-climatic factors in the development of dengue disease. Non-climatic factors such as altitude and population density were found to be have an association towards DIR in Brazil. As well as in Malaysia, population density was found to have a strong significant relationship towards DIR [9, 19]. Malaysia known as one of the fastest developing countries including the growing economies and the development of technology. There are many good dengue control efforts established in Malaysia, but only several strategies effective to control the dengue outbreaks. In this paper, we develop the best potential negative binomial GAM model for monthly dengue cases with the application of FCM in 9 districts of Selangor for the period January 2010 to August 2015. The potential

independent variables investigated including climatic and non-climatic factors which has been proofing to be significant to the dengue incidence rate.

2. Description of data and model development

2.1. Description of data

Selangor is the most developed and populated in Malaysia with an estimated population of 5,874,000 in 2015. Located on the west coast of Peninsular Malaysia, Selangor consists of high population density areas and cities. Meanwhile, Selangor was chosen as the area of this study because this district recorded the highest annual number of dengue cases in Malaysia since 2008 [12]. Selangor is divided into nine administrative districts which are Gombak, Hulu Langat, Hulu Selangor, Klang, Kuala Langat, Kuala Selangor, Petaling, Sabak Bernam and Sepang. In modelling dengue fever incidence, the monthly dataset for each of the nine districts during the 68 months period between January 2010 and August 2015 were obtained Ministry of Health Malaysia (MOH). An annual population and population density in each district were obtained from Department of Statistics Malaysia (DOSM). Meanwhile, the monthly rainfall amount and monthly number of rainy days were supplied by the Department of Irrigation and Drainage Malaysia (DIDM). Dengue incidence rate (DIR) is a dependent variable for this study. According to [13], DIR can be explained as the number of new dengue cases spotted in a certain time-period divided by the population of the district at risk per 100,000 populations (see (1)).

$$DIR=\frac{y_{dm}}{\rho_{dt}}\times 100{,}000 \tag{1}$$

where is y_{dm} is the number of new dengue cases in the district, in a month, $d(d=1,2,3,...)$ in a month, $m(m=1,2,3,...)$. The value then divided by the estimation of the total population of the district, ρ_{dt} for the year, $t(t=1,2,3,...)$. Therefore, in this study, the calculation of the DIR based on the monthly DIR per 100,000 populations.

2.2. Model development

In modelling dengue cases, there are many researchers proved climatic variables play the major role in the distribution of dengue incidence worldwide [5, 9, 14]. Meanwhile, [15] agreed that due to limited non-climatic variables information, this lead to several limitation in producing better results of reducing dengue cases. Therefore, the unique features in this study, we decided to highlight both climatic and non-climatic variables that have potential in effecting dengue cases in Selangor.

After exploratory data analysis, we found that the potential non-climatic variables that show relationship either positive or negative towards DIR are number of localities, population density and DIR with lagged 1 until 3 months. Meanwhile, for climatic variables, the average monthly amount of rainfall and number of rainy days with lagged current to 3 months respectively show significant relationship towards DIR. The most important finding in this study is the interaction between two climatic variables which are the average monthly amount of rainfall and number of rainy days show strong significant relationship towards DIR in Selangor for the period of January 2010 to August 2015.

A Fuzzy C-Means Model (FCM) was first developed by Dunn [16] then was improved by [17]. The advantage of this method is the availability of one piece of data to equip in two or more clusters. Meanwhile, the main objective of FCM is to minimise the objective function as in (2).

$$\Gamma(U,V)=\sum_{i=1}^{N}\sum_{j=1}^{C}\left(\mu_{ij}\right)^{m}\left\|x_i-v_j\right\|^2 \tag{2}$$

Where N is the number of observation and C represent the number of clusters? Meanwhile, μ_{ij} represent the membership of i^{th} data to j^{th} cluster center with m is the fuzziness index and $\left\|x_i-v_j\right\|$ is the Euclidean distance between i^{th} data and j^{th} cluster center.

There are new datasets formed based on the application of FCM. Then, to develop the best potential model for dengue cases in Selangor, another framework was developed by adopting a generalised additive model (GAM). GAM was originally introduced by [18] with the aim to give valuable impact towards independent variables through smooth function. In this study, there is a problem called overdispersion exist in the datasets due to the high variability in the dengue counts throughout the study period. Therefore, negative binomial distribution was adopted to allow overdispersion [11, 19]. Hence, the final model was arranged as in (3) and (4).

$$y_{dm}\sim NegBin\left(e_{dm}=\rho_{dm}\tau_{dm},\theta\right) \tag{3}$$

$$\begin{aligned}\log e_{dm}&=\log\left(\rho_{dm}\right)+\log\left(\tau_{dm}\right)\\&=\log\left(\rho_{dm}\right)+\eta+\sum_{k=1}^{8}\phi_k a_{kdm}+\phi_{15}a_{1dm}a_{5dm}+\\&\quad\sum_{k=2}^{9}\gamma_k x_{kdm}+\gamma_{14}x_{14dm}+\gamma_{16}x_{16dm}+f_d\left(x_{1dm}\right)\end{aligned} \tag{4}$$

From (3) and (4), the observed dengue cases, y_{dm} for the district, $d(d=1,2,3,...)$ and month, $m(m=1,2,3,...)$. Then, y_{dm} considered to be negative binomial distributed where e_{dm}, represent the expected number of dengue cases is given by the multiplication of population, ρ_{dm}, and the unknown relative dengue factor, τ_{dm}, for a given district, d, and month, m. From the general τ_{dm} term in the (4), there were two different groups has been introduced. Firstly, the selected climatic variables, $\phi_k a_{kdm}$ were found to be the average monthly number of rainy days and the average monthly amount of rainfall with lag 1 to 3 months respectively and the term $\phi_{15}a_{1dm}a_{dcm}$ represent the interaction between the average monthly number of rainy days and amount of rainfall. Meanwhile, the second part is the selected non-climatic variables, $\gamma_k x_{kdm}$ which represent by the district, month, number of locality population, log DIR with lagged 1 to 3 months and year. For $f_d\left(x_{1dm}\right)$ represent smooth function of the calendar month, x_{1dm}.

3. Results and discussion

Table 1: The value of c and F value for DIR

Number of clusters, *c*	*F* value
2	0.0186
3	0.0399
4	0.0320

Firstly, the FCM was applied to the DIR data to cluster the dataset. However, several clusters need to be decided before applying FCM into the dataset. The best number of clusters can be determined based on the minimum *F* value. Therefore, we found that

the best number of clusters for this study is 2 with the minimum *F* value is 0.0186. Table 1 summarises the *F* value for cluster 2 to 4. Next, application of FCM started when the value of membership function was set as 0.5. the membership function is a benchmark for the individual data to belong either in Cluster 1 or Cluster 2.

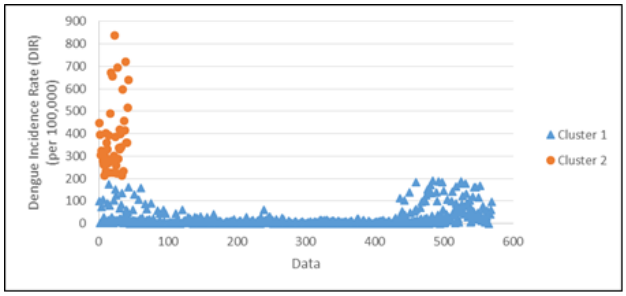

Fig. 1: Plot for DIR based on cluster

Table 2: Details of the DIR variable

	Cluster 1	Cluster 2
Number of data, *n*	569	43
Minimum	0	212
Maximum	211	837
Mean, μ	22.11	388.75
Standard deviation, σ	38.54	155.56

In this study, the individual data which has the membership value less than 0.5 is considered in Cluster 1, meanwhile the individual data which has the membership value more than 0.5 is considered in Cluster 2. Therefore, we found 569 data in Cluster 1 with ranges from 0 to 211 cases per 100,000 population whereas 43 data for Cluster 2 ranges from 212 to 837 cases per 100,000 population (see Figure 1). The membership function graph for variable DIR is shown in Figure 2. The mean and standard deviation value for DIR in Cluster 1 and Cluster 2 are 22.11 and 38.54 respectively. Meanwhile, for Cluster 2, the value of mean and standard deviation is 388.75 and 155.56 respectively (see Table 2).

Table 3: Comparison of Deviance (D), *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) using negative binomial

Model	Deviance	*AIC*	*BIC*	*UBRE*
Cluster 1	676.134	4777.038	4974.865	0.494
Cluster 2	6.088	748.266	785.734	0.131

Finally, after applying FCM into the dataset, by adopting GAM, the best potential model between the 2 clusters is Cluster 2 based on smallest value of Deviance (D), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Table 3 shows the differences of D, AIC and BIC values between both clusters. Therefore, (5) shows the best potential model which proved in this study.

$$DIR = 4.7504e^{-38}\exp\left(\begin{array}{l}0.0463\gamma_2+0.0009\gamma_3-0.0019\gamma_4+1.5560\gamma_5+\\6.3630\gamma_{11}+0.0611\gamma_{12}+0.4232\gamma_{13}+0.0004\gamma_{14}+\\0.0002\gamma_{15}+0.0004\gamma_{16}+0.0154a_1+0.0166a_2-\\0.0138a_3-0.0010a_4+0.0015a_5-0.0112a_6+\\0.0113a_7+0.0100a_8-0.0090a_{15}\end{array}\right) \quad (5)$$

It could be summarised as the influence of climatic and non-climatic variables either positive or negative towards the DIR. In this study, we found seven non-climatic variables which are year (γ_2), number of locality (γ_3), district of Petaling (γ_{11}), district of Hulu Langat (γ_{13}), DIR lag 1 month (γ_{14}), DIR lag 2 months (γ_{15}) and DIR lag 3 months (γ_{16}) show positive influence towards the increasing of the DIR in Selangor. For example, after exponent the coefficient value of γ_2, we can conclude that the value of DIR is expected to be increase by 4.71% for every increasing one year. Meanwhile, the other non-climatic variables such as population density (γ_4), district of Hulu Selangor (γ_5) and district of Sepang (γ_{12}) show negative influence towards the value of DIR. Next, focus on climatic variables, the average monthly rainfall amount in current month (a_1), the average monthly rainfall amount at lag 1 month (a_2), the average monthly number of rainy days in current month (a_5), the average monthly number of rainy days at lag 2 months (a_7) and the average monthly number of rainy days at lag 3 months (a_8) show positive influence towards the DIR. Meanwhile, other climatic variables show negative influence towards the DIR including the average monthly amount of rainfall at lag 2 months (a_3), the average monthly amount of rainfall at lag 3 months (a_4), the average monthly number of rainy days at lag 1 month (a_6) and finally, the interaction between the average monthly rainfall amount and number of rainy days which both in current month (a_{15}) show negative influence towards the DIR in Selangor for the period of January 2010 to August 2015.

4. Conclusion

The unique features in this study is the application of FCM in modelling infectious disease and this is a new development in dengue modelling especially in Malaysia. Besides, the potential of combination between climatic and non-climatic factors in modelling dengue incidence also need to be highlighted in this study. The main finding in this study is the division of cluster which based on the value of membership function that has been set 0.5 as the boundary. There are two clusters found in this study which Cluster 1 consist of 569 data and Cluster 2 consist of 43 data. Other than that, the potential non-climatic factors that have been proved show significant relationship towards the DIR are the year, month, population density, number locality and log DIR with lagged 1 to 3 months. Meanwhile, the potential climatic variables which show significant relationship towards the DIR in Selangor are the average monthly number of rainy days and amount of rainfall with lagged 1 to 3 months respectively. Therefore, the combination of clustering process and modelling development is one step ahead in providing more potential work for modelling dengue cases and other diseases in the future. Hopefully, this new framework can be a stepping stone to develop the best decision with the main purpose to reduce the number of dengue cases in Selangor and Malaysia.

Acknowledgement

The authors would like to express gratefully heartfelt thanks to the Universiti Tun Hussein Onn Malaysia and Office for Research, Innovation, Commercialization and Consultancy Management (ORICC) for the financial support under the TIER 1 research grant (U909).

References

[1] Alto BW & Bettinardi D (2013), Temperature and dengue virus infection in mosquitoes: independent effects on the immature and adult stages, American Journal of Tropical Medicine and Hygiene 88(3), pp.497-505.

[2] Murray NEA, Quam MB & Wilder-Smith A (2013), Epidemiology of dengue: past, present and future prospects, Clinical Epidemiology, 5, pp.299-309.

[3] Skae FMT (1902), Dengue fever in Penang, The British Medical Journal, 2(2185), pp.1581-1582.
[4] Qi X, Weng Y, Li Y, Meng Y, Chen Q, Ma J & Gao GF (2015), The effects of socioeconomic and environmental factors on the incidence of dengue fever in the Pearl River Delta, China, 2013, Public Library of Science Neglected Tropical Diseases, 9(10), e0004159.
[5] Wan Fairos WY, Wan Azaki WH, Mohamad Alias Y & Bee Wah Y (2010), Modelling dengue fever (DF) and dengue haemorrhagic fever (DHF) outbreak using Poisson and Negative Binomial model, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, 4(2), pp.1-6.
[6] Lee HL, Rohani A, Khadri MS, Rozilawati H, Nurulhusna AH, Nor Afizah AH, Roziah A, Rosilawati R (2015), Dengue vector control in Malaysia-challenges and recent advances, The International Medical Journal Malaysia, 14(1), pp.11-16.
[7] Gubler DJ (1998), Dengue and dengue hemorrhagic fever, Clinical Microbiology Reviews. 11(3), pp.480-496.
[8] Aura DHM & Alfonso JRM (2010), Potential influence of climate variability on dengue incidence registered in a western pediatric Hospital of Venezuela, Tropical Biomedicine, 27(2), pp.280-286.
[9] Che Him N, Bailey TC & Stephenson DB (2012), Climate variability and dengue incidence in Malaysia, Proceedings of the 27th International Workshop on Statistical Modelling, Prague 2012, Volume II, pp.435-440.
[10] Morales I, Salje H, Saha S & Gurley ES (2016), Seasonal distribution and climatic correlates of dengue disease in Dhaka, Bangladesh, The American Journal of Tropical Medicine and Hygiene, 9496, pp.1359-1361.
[11] Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CA, Carvalho MS & Barcellos C (2011), Spatio-temporal modelling of climate-sensitive disease risk: towards an early warning system for dengue in Brazil, Computers & Geosciences, 37(3), pp.371-381.
[12] Shah SA & Sani JA (2011), SP6-37 Predicting dengue fever incidence in Selangor using time-series analysis technique, Journal of Epidemiology & Community Health, 65, A464.
[13] Hassan H, Shohaimi S & Hashim NR (2012), Risk mapping of dengue in Selangor and Kuala Lumpur, Malaysia, Geospatial Health, 7(1), pp.21-25.
[14] Colon-Gonzalez FJ, Fezzi C, Lake IR & Hunter PR (2013), The effects of weather and climate change on dengue, Public Library of Science Neglected Tropical Diseases, 7(11), e2503.
[15] Moreno-Banda GL, Riojas-Rodriguez H, Hurtado-Diaz M, Danis-Lozano R & Rothenberg SJ (2017), Effects of climatic and social factors on dengue incidence in Mexican municipalities in the state of Veracruz, Salud Pública de México, 59(1), pp.41-52.
[16] Dunn JC (1973), A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, Journal of Cybernatics, 3, pp.32-57.
[17] Bezdek JC (1981), Pattern recognition with fuzzy objective function algorithm, Plenum Press, New York.
[18] Hastie T & Tibshirani R (1986), Generalized Additive Models, Statistical Science 3 (1), pp.297-310.
[19] Che Him N (2015), Potential for using climate forecasts in spatio-temporal prediction of dengue fever incidence in Malaysia (Doctoral dissertation). Retrieved from https://ore.exeter.ac.uk/repository/handle/10871/23205