Potential New Hybrid Models Of DIR By Using GAM And FCM

Norziha Che Him, Nazeera Mohamad, Mohd Saifullah Rusiman

Abstract:Dengue is one of popular infectious disease where mortality rate nowadays recorded one-third of the world's population lived in the high risk areas of dengue infection. This study proposed a new hybrid models of dengue incidence rate (DIR) by using two statistical models known as negative binomial Generalised Additive Model (GAM) and Fuzzy C-Means (FCM) Model. The data used consists of response variable known as monthly DIR and monthly climatic and non-climatic variables that covers Selangor state of Malaysia for the period of January 2010 to August 2015. This study has successfully presents the statistically significant values for climatic and non-climatic as explanatory variables that influenced DIR. Statistical results show that the climatic factors which are rainfall at current month up to 3-month and number of rainy days at current month up to lag 3-month are significant to DIR. Besides, an interaction between rainfall and number of rainy days presents strong positive relationship to DIR. In addition, non-climatic factors such as population density, number of locality and lag DIR from 1-month to 3-month also describe statistical significant relationship towards DIR. Meanwhile, for both of clustering techniques applied which are district data clustering and FCM data clustering, four models have been developed known as Model B, Model C, Model D and Model E with Model A is from the original dataset. Comparison values of Deviance (D), Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) conclude that two new models with lowest values of D, AIC and BIC known as Model C and Model E could potentially present dengue incidence in Selangor, Malaysia from January 2010 to August 2015.

Index Terms: Hybrid Model, DIR, GAM, FCM, Selangor

1. INTRODUCTION

Dengue incidence rate (DIR) has grown severely in Malaysia and now significantly become a major public health concerned [1][2]. Public health agencies in Malaysia had spent attention and put more priority to this issue to reduce the number of dengue cases that reported with dramatically increased over years. However, the symptoms of dengue become slightly identical with other infectious disease and therefore this situation put dengue cases difficult to be confirmed. The dengue cases reported keep on increasing because the geographical distribution of dengue expands year by year. This leads to the detection of new dengue risk area in rural and also in urban areas [2][3]. The number of dengue reported keep increasing and numerous programs have been introduced involving community and health authorities since the 1970s [4]. However, there has been slightly research on modelling dengue by using climate and other covariates across the state that is recorded as the hotspot area in Malaysia. There has been only a few researches on modelling dengue by using climatic and non-climatic covariates across the state that is recorded as the "hotspot" area in Malaysia. A major problem arises in pursuing this study is the availability of information regarding non-climatic factors in Malaysia. This problem also occurred in many different countries such as Vietnam [5] and Ecuador [6]. Therefore, the results from previous research could be the benchmark for further study to plan the exploration in finding the relationship between climatic and non-climatic factors and dengue incidence in Malaysia. The application of statistical analysis in modelling dengue cases has been applied worldwide. Generalised Linear Model (GLM) techniques are widely used in modelling dengue cases [7]. However, due to the presence of overdispersion, Negative Binomial regression model adopted to overcome this problem. In modelling DIR, Generalised Additive Model (GAM) could be applied in order to identify a pattern for observed dengue count in Malaysia [10,11,12]. Even clustering technique is quite new in modelling dengue fever [8] and this technique is strongly suggested in order to determine the highest potential dengue risk area. This study has diversified unique features that can contribute to minimise the dengue risk problem in Malaysia. This study considered a long amount of time as a monthly basis of dengue data in Selangor. Then, the application of clustering in modelling dengue in Selangor with the adoption of Negative Binomial GAM able to deal with the next dengue outbreak by made of several months ahead of preparation by referring to the potential model developed in this study.

- *Norziha Che Him is a Senior Lecturer in Department of Mathematics & Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, 84600 Pagoh, Johor, Malaysia. E-mail: norziha@uthm.edu.my*
- *Nazeera Mohamad is a Science Officer in Malaysian Science and Technology Information Centre (MASTIC), Ministry of Energy, Science, Technology, Environment and Climate Change (MESTECC) E-mail: nazeera@mestecc.gov.my*
- *Mohd Saifullah Rusiman is Associate Professor in Department of Mathematics & Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, 84600 Pagoh, Johor, Malaysia. E-mail: saifulah@uthm.edu.my*

2 METHODOLOGY

2.1 Research Procedure

There are five stages of the research procedures with the first stage is to determine early potential variables of climatic and non-climatic variables. The second stage is to identify the new datasets to develop appropriate DIR model. Third stage are the new approach used known as clustering process based on district and Fuzzy C-Means (FCM) categories. Firstly, the data clustered by district based on the value of annual DIR to classify the dengue risk categories. Secondly, FCM algorithm was applied to cluster the value of DIR.

Then, the DIR models that have been developed will assess by comparing the value of deviance (D), Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Finally, the validation process towards all potential models that have been developed. Therefore, the new models with the lowest value of D, AIC and BIC from an existing models selected as the best potential models to explain the climatic and non-climatic factors that contribute to DIR in Selangor, Malaysia from January 2010 to August 2015.

3 RESULT AND DISCUSSION

3.1 Determine variables based original dataset

Previous researchers have proved climatic and non-climatic variables are significant towards dengue incidence worldwide [9][10], therefore, the main objective of this stage is to investigate the strength of relationship between two variables by referring to the correlation coefficient value. Overall, the relationship between logarithm of monthly DIR and population density, number of localities, amount of rainfall with current until lag 2-month, number of rainy days with current until lag 2-month and interaction between rainfall and number of rainy days show positive relationship. Meanwhile, explanatory variables which are DIR lagged 1-month until 3-month, amount of rainfall with lag 3-month and number of rainy days with lag 3-month show negative relationship towards logarithm of monthly DIR.

3.2 District Data Clustering

Dataset of this study has been divided into two clusters where Cluster 1 consists of district's data with mean annual DIR from 0 to 200 cases per 100,000 populations that considered to represent the low dengue risk area. The other cluster, known as Cluster 2 consists of district data with mean annual DIR from 200 to 500 cases per 100,000 populations that considered to represent the high dengue risk area. Table 1 summarise the division of districts in Selangor based on cluster.

TABLE 1

THE DIVISION OF THE DISTRICTS IN SELANGOR BY CLUSTERING

Cluster 1	Cluster 2
Gombak Hulu Selangor Klang Kuala Langat Kuala Selangor Sabak Bernam	Hulu Langat Petaling Sepang

3.3 FCM Data Clustering

FCM was applied to cluster DIR data. The selection number of clusters based on the minimum value of F could be seen in Table 2. It's summarised the values of F for cluster 2, 3 and 4. In conclusion, the best cluster is 2 with the minimum value, F is 0.0186. The membership function concluded that Cluster 1 consist of data ranges from 0 to 211 cases per 100,000 populations and Cluster 2 consist of data ranges from 212 to 837 cases per 100,000 populations.

TABLE 2

THE VALUES OF C AND F FOR DIR

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

4 MODELLING FRAMEWORK

The objective of this section is to develop the modelling framework for monthly dengue incidence in Selangor. The model development divided into three parts in which the first model by using the existing dataset. Meanwhile, the other two parts of model development using the new dataset clustered by district and application of FCM.

4.1 Model Development Based On Original Dataset

A Poisson GLM model has been developed by using existing dataset. However, due to overdispersion as a common problem for count data, this study adopts a negative binomial GLM known as Model A. Here, y_{dm} represents the observed dengue cases for the district, $d\,(d=1,2,\ldots,9)$ and month, $m\left(m=1,2,...,68\right)$ then considering these observed dengue cases to be negative binomial distributed with the use of a GLM. The general form of negative binomial GLM as in (1) and (2),

$$y_{dm} \sim NegBin\,(e_{dm} = p_{dm}\upsilon_{dm}, \phi) \tag{1}$$

$$\begin{aligned}\log e_{dm} &= \log(p_{dm}) + \log(\upsilon_{dm}) \\ &= \log(p_{dm}) + \alpha + \sum_{k=1}^{p}\beta_k x_{kdm} + \sum_{k=1}^{p}\gamma_k z_{kdm}\end{aligned} \tag{2}$$

Notice that, e_{dm} represent the expected number of dengue cases where it is the multiplication of population p_{dm} and the unknown relative dengue factor, ν_{dm}, for a given district, d, and month, m. The general $\beta_k x_{kdm}$ term in (2) has been divided into two different groups. Firstly, the selected non-climatic covariates, $\beta_k x_{kdm}$, which are referring to a factor of the month, year, district, the number of locality, population density and log DIR lagged 1, 2 and 3-month. Secondly, the terms $\gamma_k z_{kdm}$, represent the selected climatic covariates. All the predictor variables and their subsets involved in the model development were explored.

Then, to overcome the non-linear problem with explanatory variable in modelling DIR, this study adopted a negative binomial GAM as in (3). It's a similar and basic equation with (2) except for the smooth function of the calendar month that represent by $f_d(x_{kdm})$.

$$\begin{aligned}\log e_{dm} &= \log(p_{dm}) + \log(\upsilon_{dm}) \\ &= \log(p_{dm}) + \alpha + \sum_{k=1}^{p}\beta_k x_{kdm} + \sum_{k=1}^{p}\gamma_k z_{kdm} \\ &\quad + f_d(x_{kdm})\end{aligned} \tag{3}$$

TABLE 3

VALUES OF D, AIC AND BIC FOR POISSON GLM AND NEGATIVE BINOMIAL GLM (MODEL A) BASED ON ORIGINAL DATASET

Statistics Test	Poisson	Negative Binomial
D	658503	1258.610
AIC	141567	5947.000
BIC	141675	6059.472

Table 3 shows the values of D, AIC and BIC where the main

result shows Model A developed based on negative binomial GLM with all statistics test are less than compared to Poisson GLM.

4.2 Model Development Based On Dataset Clustered By District

A negative binomial GAM has been applied to this clustered dataset by district. Model B represent DIR model for Cluster 1 with year (β_2) and locality number (β_3) have positive relationship with DIR instead population density (β_4) has a negative relationship with DIR. Meanwhile, for district, Hulu Selangor (β_5), Kuala Langat (β_7) and Kuala Selangor (β_8) have negative relationships with DIR except district of Klang (β_6) has a positive relationship with DIR. In addition, DIR lag 1-month (β_{14}), DIR lag 2-month (β_{15}) and lag 3-month (β_{16}) shows positive relationships with DIR. For the mean rainfall from current month (γ_1) to lag 3-month (γ_4), there have positive relationships with DIR. As for a mean number of rainy days, only lag 3-month (γ_8) has a negative relationship with DIR compared to the current month (γ_5) up to lag 2-month (γ_7) have positive relationship to DIR. The interaction between mean rainfall and mean number of rainy days (γ_{15}) has a negative relationship with DIR. Model C represent DIR model for Cluster 2 with year (β_2) and number of locality (β_3) have positive relationships with DIR. District of Petaling (β_{11}) has a positive relationship with DIR negative relationship to district of Sepang (β_{12}). DIR lag 1-month, 2-month and 3-month present positive relationships with DIR. Meanwhile for climatic variables, mean rainfall, from a current month up to lag 3-month show positive relationships with DIR. This is due to heavy rainfall in earlier month and therefore strongly influenced the density of mosquito by building a new habitat then cause the increased of mosquito population size [7]. Next, mean number of rainy days at the same month until lag 3-month have positive relationships with DIR except for lag 2-month. The interaction between mean rainfall and mean number of rainy days also has a negative relationship with DIR.

4.3 Model Development Based On Dataset Clustered By FCM

The dataset clustered by using FCM technique adopt a negative binomial GAM to develop Model D that represent DIR for Cluster 1 with year (β_2) and number of locality (β_3) present positive relationships with DIR. District of Klang (β_6), Petaling (β_{11}) and Hulu Langat (β_{13}) have positive relationships with DIR and the rest of districts have negative relationships with DIR. Meanwhile only DIR lag 3-month (β_{16}) has a negative relationship with DIR instead DIR lag 1-month (β_{14}) and 2-month (β_{15}) have positive relationships with DIR. The mean rainfall lag 3-month (γ_4) has a negative relationship with DIR but mean rainfall in the same month (γ_1) up to lag 2-month (γ_3) have positive relationship with DIR. The mean number of rainy days in the same month (γ_5), lag 1-month (γ_6), lag 2-month (γ_7) and lag 3-month (γ_8) show positive relationships with DIR. Finally, an interaction between mean rainfall and mean number of rainy days (γ_{15}) has a negative relationship with DIR. Model E represent DIR model for Cluster 2 which the year (β_2) and the locality number (β_3) have positive relationships with DIR but the population density (β_4) has a negative relationship with DIR. For district Petaling (β_{11}) and Hulu Langat (β_{14}), both districts have positive relationships with DIR. However, district Hulu Selangor (β_5) and Sepang (β_{12}) have negative relationships with DIR. Next, log DIR shows positive relationship for all variables from lag 1 (β_{14}), lag 2 (β_{15}) and lag 3-month (β_{16}) with DIR. For the mean rainfall, current month (γ_1) and lag 1-month (γ_2) have positive relationships with DIR instead mean rainfall lag 2-month (γ_3) and lag 3-month (γ_4) have negative relationships with DIR. The mean number of rainy days in the same month (γ_5), lag 2-month (γ_7) and lag 3-month (γ_8) have positive relationships with DIR but for lag 1-month (γ_6) has a negative relationship with DIR. Finally, an interaction between mean rainfall and mean number of rainy days (γ_{15}) has a negative relationship with DIR.

5 MODEL COMPARISON

Table 4 shows the comparison values of D, AIC and BIC for five models. Based on Table 4, this study concluded that the best two new hybrid models which are Model C with value of D, AIC and BIC are 232.686, 2401.617 and 2481.91 respectively and Model E with value of D, AIC and BIC are 6.088, 748.266 and 785.734 respectively that developed from group clustering data by district and group FCM are the lowest values then become two potential models could represent the dengue incidence in Selangor, Malaysia from January 2010 to August 2015.

TABLE 4

COMPARISON VALUES OF D, AIC AND BIC BY USING A NEGATIVE BINOMIAL GAM CLUSTERING METHODS

Model	D	AIC	BIC
A (original dataset)	964.413	5954.191	6133.971
B (Cluster 1 by district)	484.473	3246.222	3379.65
C (Cluster 2 by district)	232.686	2401.617	2481.91
D (Cluster 1 by FCM)	676.134	4777.038	4974.865
E (Cluster 2 by FCM)	6.088	748.266	785.734

6 SUMMARY AND CONCLUSION

Overall, this study has successfully presented a combination of unique features that contribute to minimise the identified dengue risk problem in Selangor, Malaysia. In addition, this study has considered a long period time in monthly basis of dengue data in Selangor, Malaysia. It's also produced new datasets by using clustering techniques known as district and FCM. Then, the statistical modelling of DIR successfully develops two new hybrid models known as Model C and Model C with lowest value of D, AIC and BIC by using FM and negative binomial GAM. High expected to health department in Malaysia especially in handling future dengue outbreak in preparing and consider the potential attributed for several months ahead based on these two models.

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] A.H. Mohd-Zaki, J. Brett, E. Ismail and M. L'Azou, "Epidemiology of Dengue Disease in Malaysia (2000-2012): A Systematic Literature Review", Public Library of Science Neglected Tropical Disease, vol. 8, no. 11, e3159, 2014.

[2] M.H. Juni, K.S. Hayati, C.M. Cheng, G.S. Pyang, N.H. Abd Samad, and Z.S. Zainal Abidin, "Risk Behaviour Associated with Dengue Fever among Rural Population in Malaysia", International Journal of Public Health and Clinical Sciences, vol. 2, no. 1, pp. 114-127, 2015.

[3] M. Aloka, H.L. Premaratne, and M.G.N.A.S. Fernando, "Towards an Early Warning System to Combat Dengue", International Journal of Computer Science and Electronics Engineering, vol. 1, no. 2, pp. 252-256, 2013.

[4] N. Che Him, T.C. Bailey and D.B. Stephenson, "Climate Variability and Dengue Incidence in Malaysia", 27th International Workshop on Statistical Modelling (IWSM'27), vol. 2, 2012.

[5] W.P. Schmidt, M. Suzuki, V.D. Thiem, R.G. White, A. Tsuzuki, L.M. Yoshida, H. Yanai, U. Haque, L.H. Tho, D.D. Anh and K. Ariyoshi, "Population Density, Water Supply and The Risk of Dengue Fever in Vietnam: Cohort Study and Spatial Analysis", Public Library of Science Medicine, vol. 8, no. 8, e1001082, 2011.

[6] A. Stewart-Ibarra and R. Lowe, "Climate and Non-Climate Drivers of Dengue Epidemics in Southern Coastal Ecuador," The American Society of Tropical Medicine and Hygiene, vol. 88, no. 5, pp. 971-981, 2013.

[7] W.Y. Wan Fairos, W.H. Wan Azaki, L. Mohamad Alias, and Y. Bee Wah, "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, vol. 4, no. 2, pp. 1-6, 2010.

[8] K. Shaukat, N. Masood, A. Shafaat, K. Jabbar, H. Shabbir, and S. Shabbir, "Dengue Fever in Perspective of Clustering Algorithms", Data Mining in Genomics & Proteomics, vol. 6, no. 3, pp. 1-5, 2015.

[9] P.C. Wu, H.R. Guo, S.C. Lung, C.Y. Lin, and H.J. Su, "Weather as an Effective Predictor for Occurrence of Dengue Fever in Taiwan", Acta Tropica, no. 103, pp. 50-57, 2007.

[10] N. Che Him, M.G. Kamardan, M.S. Rusiman, S. Sufahani, M. Mohamad and N.K. Kamaruddin, "Spatio-Temporal Modelling of Dengue Fever Incidences in Malaysia", Journal of Physics: Conference Series, 995, 1-7, 2018.

[11] N. Mohamad, N. Che Him, M.S. Rusiman, S. Sufahani and S.A Muhammad Jamil, "Application FCM in Modelling DIR for Selangor Using Negative Binomial GAM", International Journal of Engineering & Technology, 7(4.30), 1-42,2018.

[12] N. Che Him, N. Mohamad, M.S. Rusiman and K. Khalid and M.A. Shafi, "Dengue Incidence Rate Clustering by District in Selangor", International Journal of Engineering & Technology, 7(4.30), 416-418,2018.