Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Geographically Weighted Regression on dengue epidemic in Peninsular Malaysia

To cite this article: Ayuna Sulekan *et al* 2021 *J. Phys.: Conf. Ser.* **1988** 012099

Geographically Weighted Regression on dengue epidemic in Peninsular Malaysia

Ayuna Sulekan[a], Jamaludin Suhaila[b] and Nurmarni Athirah Abdul Wahid[c]

[a,b,c]Department of Mathematical Science, Faculty of Science,
[b]UTM Centre of Industrial and Applied Mathematics
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

E-mail: suhailasj@utm.my

Abstract. Dengue has been a global epidemic since World War II, with millions of individuals being infected every year. Repetitive dengue epidemic is one of the main health problems that, due to its rapid spread and geographically widespread, has become a major concern for the government authorities in dealing with this disease. In Malaysia, cases of dengue are reported annually. To keep cases under control, it is important to examine the possible factors that help the growth of the virus. Climatological factors such as rainfall, temperature, wind speed, and humidity are expected to have high potential to increase the growth of the virus in this study, and their spatial variation is associated with cases of dengue. The result revealed that Ordinary Least Square was not an effective method for modelling the relationships between dengue cases and climate variables, as climate variables in different spatial regions act differently. During the analysis, there could be some issues of non-stationarity since the geographical aspect and spatial data were involved. Hence, the Geographically Weighted Regression (GWR) is implemented due to its capability to identify the spatial non-stationarity behavior of influencing factors on dengue incidence and integrate the geographical location and altitude for the spatial analysis. GWR analysis found that the influenced factors exhibited a significant relationship with dengue incidence. GWR also shows a significant improvement in Akaike Information Criteria (AIC) values with the lowest value and the highest adjusted R square. It is expected that the developed model can help the local hygienic authorities design better strategies for preventing and controlling this epidemic in Malaysia.

Dengue epidemic; spatial variation; non-stationarity; climatological factors; geographically weighted regression.

1. Introduction

Generally, dengue fever is an illness brought by a female mosquito known as Aedes *aegypti* species. Dengue fever is categorized as Zoonotic disease, which means it is an animal-borne infectious disease that can spread to human and cause illness. The patients with dengue fever might facing symptoms such as high fever, headache, vomiting, muscle and joint paint, and characteristic skin rash. Dengue can cause a broad spectrum diseases which means that it can range from the subclinical disease with major of the infected patient are not aware that they are infected to severe flu-like symptoms [1,2].

The life-threatening dengue hemorrhagic fever (DHF) will be developed from having the antibodies of dengue virus from past infection of weaken immune system. The patients that having DHF [3,4] will experience low blood platelets, plasma leakage, gums and nose bleeding and liver enlargement. The symptoms of DHF can include dengue shock syndrome, which can cause to severe bleeding in the worst-case scenario if it is not detected and threated promptly.

According to World Health Organization (WHO), dengue fever has become a significant health hazard that has health authorities concerned due to its fast spread and geographical spread throughout the world. Since the Second World War, dengue fever has become global problem and spreading throughout countries with tropical and subtropical climates, primarily in urban and semi-urban region [5–7]. The fact that around 40% of the world's population is now at danger of contracting the dengue fever [8]. This may be due to the changing in climate factors which give significant influence towards the probability of getting infected by dengue fever. Previous researcher state that in Thailand [9], the recorded incidence of dengue was associated with the changes in temperature, relative humidity and rainfall season. Some researchers has studied the factor of urbanization rate in tropical and subtropical countries like Malaysia [10] will increased the dengue cases despite the influence of climate factors. According to Malaysian Meteorological Department, Malaysia is a developing country with an average annual rainfall of 250 centimeters and a temperature of 27 degrees, which supports the reproduction of Aedes mosquitos that aided by the hot and humid climate condition.

Since there is no specific dengue therapy and medication accessible for the dengue fever, the number of deaths continues to climb up every year. The development of dengue vaccine is also in unstable state and the effectiveness of this vaccine is unsure [11,12]. A statistical approach is required to control and prevent the spread and growth of dengue incidence. A statistical model helps to give a guideline for the researchers to observe the pattern and analysis the data as well as predict the future outcome. The regression techniques are the general technique used for examining the relationship between geographical variables. This technique has been used in virtually countless publication [13,14] to investigate the link relationship between dengue incidence and climate variables. Yue *et.al.* [13] used Ordinary Least Square (OLS) technique on the vector-borne disease to study the risks factors that give significant effect toward dengue fever. Recent study on OLS by Withanage *et.al.* [15] did not show any correlation between the dengue and the independent variable. A study by Pham *et.al.* [16] also did not give a conclusion that climate factors give significant effect toward the increasing of dengue cases. Based on these finding, it is possible to conclude that the association between dengue cases and climate factors varies by region. This implies that the non-stationarity is exist when modeling the relationship between the dependent variable (dengue cases) and independent variable (climate factors). Hence, it proves that OLS is not a good choice for studying the relationship between these two variables.

To cater this non-stationarity problem, Geographically Weighted Regression (GWR) model is the best choice where the spatial heterogeneity of data is present [17]. GWR is a spatial analysis method that take into account the non-stationary variable such as climate factors and environmental characteristics. The GWR is an extension that based on OLS in which it adds modeling complexity by allowing the relationship between independent and dependent variables to differ depending on location. GWR also detects the spatial autocorrelation of variables and is valuable as an explanatory tool for its predictive capabilities. This approach is intended to support the assumptions that contextual variables will affect the robustness and relationship between the dependent variables and predictors. The major purpose of this study is to examine the readiness of GWR in analyzing the spatial relationship between dengue incidence and influence factors when dealing with the issue of non-stationarity. This study also focuses on the impact of climatological condition on the rise of dengue fever cases. The finding of this study gives an essential help for the authorities coping with dengue epidemic.

2. Methods

2.1. Data collection
This study was conducted in Peninsular Malaysia which focused on year 2012. Data of dengue hot spots area was obtained from the open source's website (data.gov.my). In regression modeling, this dengue incidence data was chosen as dependent variable (y) while climatological factors such are temperature, rainfall, humidity, and wind speed, were chosen as the independent variable (x) as listed in table 1. These climatological factors were purchased from the Malaysian Meteorological Department. All the involved data were in form of weekly dengue cases and weekly climate data.

Table 1. List of variables.

Variables	Explanation
y	Dengue cases (no. of cases)
x_1	Mean temperature (ºC)
x_2	Mean humidity (%)
x_3	Mean rainfall (mm)
x_4	Mean wind speed (m/s)

2.2. Statistical method
Ordinary Least Square (OLS) regression and Geographically Weighted Regression (GWR) analyses were conducted on the set of data to determine the link relationship between dengue incidence and climatological factors. OLS is a regression method that examine the effect of the independent factors on the dependent variable throughout the whole study area. However, local variations of the influencing factors are not taken into consideration when applying the OLS. Since temperature, rainfall, humidity, and wind speed are spatially correlated, GWR is the most appropriate modeling process for the situation.

To handle the spatial varying, it is important to employ a technique that takes spatiality into account in the analysis. GWR is the successful technique used over many years [18]. GWR is an analytical technique that primarily aims to identify the location of non-stationarity takes place. In geographical analysis, simple linear regression is widely employed as modeling tools with the dependent variable being modelled as a linear function of a set n independent variable. A global regression model can be written as:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \tag{1}$$

where y_i is the i^{th} observation of the dependent variable, x_{ik} is the i^{th} observation of the k^{th} independent variable, ε_i are independent normally distributed error term with zero means and each β_k must be determined from a sample of n observation. GWR is a relatively simple technique that expands the standard regression model of equation (1) by enabling the local parameter to be estimated rather than global parameter, rewriting the model as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{2}$$

where (u_i, v_i) is the coordinate of i^{th} point in space and $\beta_k(u_i, v_i)$ is the implementation of continuous function $\beta_k(u, v)$ at point i, allowing for development of a continuous surface of parameter values and measurement of this surface at specific point signals the surface's spatial variability. The parameter surface of equation (1) is assumed to be constant in space.

The GWR expression in equation (2) identifies every spatial variation in the relationship that may exist and provides a way to access it. The equation (2) is calibrated based on implicitly observed data close to location i which may have greater impact on the estimation of the $\beta_k(u_i, v_i)$ than data located further away from i. Basically, equation (2) is used to calculate the relationships in the model that exists around each point i.

The crucial step in GWR is determining the latitude and longitude coordinates (u_i, v_i) for each location. This geographic coordinate was used to specify the distance between observed data in location i and observed data in location j:

$$d_{i,j} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \tag{3}$$

The distance in equation (3) is called as fundamental background in weighting the data to calculate the parameter of the GWR model. The closer the distance between estimated point to locations, the larger the weight of data during the parameter estimation. In this study, the weighting of data was conducted by means of Gaussian function:

$$\psi_{i,j} = \exp\left({-d_{i,j}^2} \middle/ {2b^2} \right) \tag{4}$$

where $b > 0$ is the constant bandwidth done by cross-validation method. The Gaussian function is an exponential function that gives the estimated parameter data a weight of one. The weight will continue reducing as the distance among the location increases. Gaussian function is applied in form of weighted matrix:

$$\mathbf{W}(u_i, v_i) = \operatorname{diag}\left(\psi_{i,(1:1)}, \psi_{i,(1:2)}, \ldots, \psi_{i,(1:q)}, \ldots, \psi_{i,(m:1)}, \psi_{i,(m:2)}, \ldots, \psi_{i,(m:q)} \right) \tag{5}$$

where $0 \le \psi_{i,(j:1)} \le 1$ is weight data for location j to estimate the parameter in location i.

The weighted least square will then provide a fundamental understanding of how the GWR performs. In weighted least square, before minimizing each square difference, a weighting factor is applied to each square difference so that the inaccuracy of some estimations is penalized more than others. An observation in GWR is weighted proximally on location i, so that the weighting of an observation differs with i. Data obtained from observations that close to i are weighted more than the data obtained from observation far away. In the estimated parameter, every observed data has one weighted matrix. By algebraic matrix approach, the estimation of, the estimation of the parameter $\hat{\beta}(u_i, v_i) = \left(\beta_0(u_i, v_i), \beta_1(u_i, v_i), \beta_2(u_i, v_i), \beta_3(u_i, v_i), \beta_4(u_i, v_i) \right)^T$ in location i by means of weighted least squares method is expressed as:

$$\beta(u_i, v_i) = \left[\hat{\mathbf{x}}^T \mathbf{W}(u_i, v_i) \mathbf{x} \right]^{-1} \mathbf{x}^T \mathbf{W}(u_i, v_i) y \tag{6}$$

where $\hat{\beta}$ is an estimate of β and $\mathbf{W}(u_i, v_i)$ is a matrix of n by n whose off-diagonal elements are zero and whose diagonal elements defined the geographically weighted of observed data for point i. The estimated parameter is the same for all observed data. This is because of the data weighted solely considers the distance between the locations.

The significance of the OLS regression and GWR model is determined via sum of squared errors (SSE) and adjusted coefficient of determination (R^2). The Akaike Information Criterion (AIC) will be used to compare the GWR model and global regression model. According to this criterion, the regression model with the least AIC value is defined as the best regression model [19].

3. Results

3.1. Results of OLS regression model

The regression modeling of data sets using OLS yields a value of *p* = 0.069, which is this value of *p* is greater than the significance level of 0.05. The result indicates that the model contains autocorrelation. The OLS regression model can be written as:

$$\hat{y} = 27864.1 + 430.6x_1 - 457.2x_2 + 827.4x_3 - 4130.8x_4 \tag{7}$$

where $\hat{y}$ is the dengue cases , x_1 is the temperature, x_2 is the humidity, x_3 is the rainfall, and x_4 is the wind speed. In the summarized table 2 revealed that all the *p*-values of regression coefficient in the model were greater than 0.05 except the rainfall (*p* = 0.051) and wind speed (*p* = 0.025). Hence, the rainfall and wind speed have an impact on the dengue cases in Peninsular Malaysia.

Table 2. Test of OLS regression coefficient.

Variables	Coefficient	Std. Error	*t*-value	*p*-value
Intercept	27864.1	48638.0	0.573	0.585
Temperature, x_1	430.6	1221.2	0.353	0.735
Humidity, x_2	-457.2	244.6	-1.869	0.104
Rainfall, x_3	821.4	349.6	2.350	0.051
Wind speed, x_4	-4130.8	1448.5	-2.852	0.025

The model has $SSE = 16657403$, $AIC = 215.776$, and adjusted $R^2 = 75.42\%$. Further analysis of this model shows that the error approach the normal distribution. Thus, this proved that OLS analysis conducted under this study appeared to be ineffective in modeling the link relationship between dengue incidence and influencing factors.

3.2. Result of GWR Model

The estimation of GWR regression model on the data set provide smaller *p*-values which is $p = 0.0053$. This result shows that the GWR regression model is free from the spatial autocorrelation. Based on the cross-validation criteria, the modeling of GWR gives an optimum bandwidth $b = 0.360079$. Thus, the Gaussian function can be expressed as:

$$\psi_{i,(j:1)} = \exp\left({-d_{i,(j:1)}^2} \Big/ {2(0.360079)^2} \right) \tag{8}$$

Table 3. Summary of coefficient for GWR model.

Variables	Min.	1st Quatiler	Median	3rd Quatile	Max.
Intercept	-23.4696	1.8363	19.2793	32.9066	33.8614
Temperature, x_1	-0.2586	-0.2275	0.0837	0.4953	1.0974
Humidity, x_2	-0.2802	-0.2671	-0.1834	-0.0931	0.0333
Rainfall, x_3	0.2153	0.2516	0.3139	0.4103	0.4507
Wind speed, x_4	-2.2289	-1.5854	-1.1156	-0.5772	-0.4146
AICc : 78.8544					
AIC : 8.634					

Adjusted R^2 : 0.92504

The GWR regression model yields sum of squared error, $SSE = 0.77857$ which is smaller values than the OLS regression model. This reveals that the estimation value of $\hat{y}$ is closer to y compared to the OLS regression model. The adjusted coefficient of determination, R^2 of GWR regression model also increases significantly high with 92.50% than the OLS regression model. In addition, the value of AIC obtained from GWR regression model analysis also provides significant value with $AIC = 8.634$ which is smaller than the value of AIC from OLS regression model. It implies that GWR regression model is better appropriate for explaining data than the OLS regression model.

4. Discussion

The finding of this study discloses that the dengue incidence in Peninsular Malaysia is connected to the climatological factors. Since Malaysia is listed in the country of subtropical climate, so the higher the risk of changing climates, the higher risk of spreading dengue incidence. This is because of the country that experience in subtropical climate is the favourable region for the reproduction of dengue. Hence, the seasonal pattern that happen in Malaysia gives good support for the growth of mosquito reproduction. The geographic dispersion, reproduction and feeding pattern of mosquitos are all affected by ambient temperature, humidity, rainfall, and wind speed. Aedes mosquito is a living microorganism that can adapt very quickly in various conditions.

Generally, rainfall is a common indicator that help to increase the breeding process of the mosquito but with heavily rainfall or storm may destroyed the breeding process and all the eggs or larvae will simply flush away from the breeding place is the worst scenario that may be happened. During rainy seasons, an Aedes mosquito can stay active and boost the reproduction with the help of ambient and support from the surrounding temperature. High temperature can interfere with mosquito mating and reproduction whilst low temperature aid in longing survival. In other hand, both high and low temperature really give contribution to the growth of mosquito and spreading the dengue cases. Furthermore, low level of wind speeds influenced the mosquito development, with a large number of dengue cases supporting the statement [20,21].

The GWR modelling indicates that the relationship between the dengue cases and climate factors in Peninsular Malaysia is spatially different due to the condition of climate factors were different from one location to another. Understanding the spatial varying relationship can helps the policymakers to halt the dengue spreading and infection. One of the statistical criteria that implies in the modelling is the adjusted coefficient of determination $\left(R^2\right)$,which indicates the percentage of variation in the dengue incidence. The GWR model provides significantly high R^2 compared to OLS model. This means that GWR model will help the local hygienic authorities to spot clearly of the region that has high risk of dengue incidence and able to predict which hot spot area have high possibility in increasing of dengue cases. Despite's climatological factors, there are other factors that also impactful towards the dengue incidence such as environment, human density, social and economic dynamics.

5. Conclusion

The climatological factors do give influence and correlated to the dengue incidence not just in Malaysia but also worldwide especially subtropical country. OLS regression technique is not an appropriate choice in modelling the relationship between dengue incidence and influencing factor because of some parameters are not fit for OLS global model as the influence factors behave differently in different spatial region. Hence, the relationship of dengue incidence and climate factors are varying in each geographical area. Every area will have high possibility of dengue infection if the citizens are not taking care of their hygiene and the government did not take serious aspect to the unplanned urban overpopulation areas.

Acknowledgement
The work was financially supported and funded by Ministry of Higher Education under Fundamental Research Grant Scheme FRGS/1/2020/STG06/UTM/02/3 and Research University Grant with vote no: 19J58 (QJ130000.2526.17H95 and RJ130000.7854.5F311) under Universiti Teknologi Malaysia.

References

[1] Laksono I, Porter KR, Erlin, Hayes CG, Tan R, Sutaryo, *et al*. 2017 A prospective seroepidemiologic study on dengue in children four to nine years of age in Yogyakarta, Indonesia I. studies in 1995-1996 *Am J Trop Med Hyg*. 61(3):412–9
[2] Raza MA, Khan MA, Ejaz K, Haider MA and Rasheed F 2020 A Case of Dengue Fever With Hemorrhagic Manifestations *Cureus.* 12(6):1–7
[3] Vasquez Velasquez C, Roman AD, Lan NTP, Huy NT, Mercado ES, Espino FE, *et al.* 2015 Alpha tryptase allele of Tryptase 1 (TPSAB1) gene associated with Dengue Hemorrhagic Fever (DHF) and Dengue Shock Syndrome (DSS) in Vietnam and Philippines *Hum Immunology*. 76(5):318–23
[4] Matthias AT, Apsara S and Epa A 2019 A case report of dengue haemorrhagic fever complicated with psoas haematoma requiring blood transfusion *BMC Infectious Diseases.* 19(1):2–5
[5] Andraud M, Hens N and Beutels P 2013 A simple periodic-forced model for dengue fitted to incidence data in Singapore *Mathematical Biosciences.* 244(1):22–8
[6] Cho H-W and Chu C 2016 A Disease Around the Corner *Osong Public Health and Research Perspective.* 7(1):1–2
[7] Fagbami AH and Onoja AB 2018 Dengue haemorrhagic fever: An emerging disease in Nigeria, West Africa *J. Infection and Public Health.* 11(6):757–62
[8] WHO WHO 2009 *Dengue: guidelines for diagnosis, treatment, prevention, and control. Spec Program Res Train Trop Dis* (France: WHO Press)
[9] Ishikawa H and Chikaki E 2009 A dengue transmission model in Thailand considering sequential infections with all four serotypes *J. Infect Dev Ctries.* 3(9):711–22
[10] Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, *et al*. 2016 The global burden of dengue: an analysis from the Global Burden of Disease Study 2013 *Lancet Infect Dis.* 16(6):712–23
[11] Wan SW, Lin CF, Wang S, Chen YH, Yeh TM, Liu HS, *et al.* 2013 Current progress in dengue vaccines *J. Biomed Sci.* 20(1):1–9
[12] Ghosh A and Dar L 2015 Dengue vaccines: Challenges, development, current status and prospects *Indian Journal of Medical Microbiology* 33(1):3-15
[13] Yue Y, Sun J, Liu X, Ren D, Liu Q, Xiao X, et al 2018 Spatial analysis of dengue fever and exploration of its environmental and socio-economic risk factors using ordinary least squares: A case study in five districts of Guangzhou City, China, 2014 *International Journal of Infectious Diseases* 75(2018):39–48
[14] Majid NA, Nazi NM and Mohamed AF 2019 Distribution and spatial pattern analysis on dengue cases in Seremban District, Negeri Sembilan, Malaysia *Sustainability* 11(13):3572
[15] Withanage GP, Viswakula SD, Nilmini Silva Gunawardena YI and Hapugoda MD 2018 A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka *Parasites and Vectors* 11(1):1–10
[16] Pham DN, Aziz T, Kohan A, Nellis S, Binti J, Jamil A, *et al.* 2015 An Efficient Method To Predict Dengue Outbreaks in Kuala Lumpur *Journal of Artificial Intelligence and Computer Science* AICS(2015)2289–5965
[17] Fotheringham AS, Charlton ME and Brunsdon C 1998 Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis *Environment and Planning A.* 30(11):1905–27

[18] Brunsdon C, Fotheringham S and Charlton M 1998 Geographically Weighted Regression *J. R. Statistical Society: Series D (The Statistican)* 47(3):431–433
[19] Fotheringham ASS, Brunsdon C and Charlton M 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (United Kingdom: Wiley-Chichester)
[20] Cheong YL, Burkart K, Leitão PJ and Lakes T 2013 Assessing weather effects on dengue disease in Malaysia *Int. J. Environmental Research Public Health* 10(12):6319–6334
[21] Ling CY, Gruebner O, Krämer A and Lakes T 2014 Spatio-temporal patterns of dengue in Malaysia: Combining address and sub-district level *Geospatial Health* 9(1):131-140