Understanding Factors Affecting the Outbreak of Malaria Using Locally-Compensated Ridge Geographically Weighted Regression: Case Study in DakNong, Vietnam

Tuan-Anh Hoang(✉), Le Hoang Son, Quang-Thanh Bui, and Quoc-Huy Nguyen

VNU University of Science, Vietnam National University, Hanoi, Vietnam
{hoangtuananh88, sonlh, thanhbq}@vnu.edu.vn,
huyquoc23ll@gmail.com

Abstract. In this paper, we propose a new scheme to analyze factors that affect outbreak of malaria using the Locally-Compensated Ridge Geographically Weighted Regression (LCR-GWR). Since malaria prevalence is location dependence, the relationships between natural and social-economic factors to the development and concentration of malaria hotspots have been investigated. The proposed method is applied to DakNong province, one of the most vulnerable areas to malaria risk in Vietnam due to the lack of social infrastructure and the limited accessibility to health services. Even though mitigation campaigns were launched in the last several years, the number of new cases was found increasingly and several hotspots are still remained. The result is compared to those of several local analyses of spatial collinearity. It has been shown that LCR-GWR considerably improves the model fit and is useful to determine several factors including NDVI, DEM, distance to residential areas, distance to road that are highly associated with malaria risks. The results of this study help measuring the incidence of malaria in the context of climate change and under the impact of change in people's livelihoods.

Keywords: Malaria · Locally-compensated ridge · Geographically weighted regression · Hotspots

1 Introduction

Malaria is one of the most widespread parasitic diseases worldwide and is considered as one of the most dangerous endemic in 91 countries with 212 million cases and 429, 000 deaths [33]. Vietnam is one of those malaria - endemic countries which has 74% of the population having malaria risk mainly in the Central Coastal and Central Highland region [33]. There were several activities aim to reduce the morbidity and mortality of malaria in Vietnam, typical of which is National Malaria Control and Elimination Program launched officially by Vietnamese Government in 2011 [32]. The program achieved some successes with confirming cases in 2016 was 9331 and 3 deaths compared to 18387 cases and 8 deaths in 2012 [32, 33]. However, there are still challenges

D. Tien Bui et al. (eds.), *Advances and Applications in Geospatial Technology and Earth Resources*, https://doi.org/10.1007/978-3-319-68240-2_11

especially in the context of climate change where climate factors affect the distribution of malaria [3].

Several methods have been proposed to this problem including epidemiological expert methods [15], Remote Sensing and GIS [1, 24] and the hybrid GIS with soft computing [6, 25, 31]. Lubetzky-Vilnai et al. [19] and Mosha *et al.* [21] used statistics and spatial analysis based on time series [4]. However, the statistical approach is not sufficient to handle complex structures and nonlinearity of malaria risk datasets. Incorporating artificial intelligence, remote sensing and GIS is an alternative way to overcome this drawback [18, 24]. Specifically, Ch *et al*. [7] integrated a support vector model and Firefly algorithms to evaluate the risk of malaria. Buczak *et al.* [6] applied fuzzy logic to study malaria in Korea. Zacarias *et al.* [34] compared the support deployment models and random forest in Mozambique. Recently, Geographically Weighted Regression (GWR) and its variants have been used in the study of the relationship between malaria and geographical factors [5, 12, 25–29].

Unlike conventional regression methods that assume the relationship between malaria and geographic factors is the same across regions in a study area, GWR creates separate regressions for each set of observed data (local regression) using adjacent objects in a defined "bandwidth" distance [2]. Although GWR is important to explore spatial non-stationary data relationship, a problem found in many regression models is collinearity which affects the precision of the model. Locally-compensated ridge GWR aims to reduce the influence of the collinearity to the regression model; thus improving its accuracy [8].

Taking advantage of regression in epidemiological studies, in this paper, we propose a new scheme to analyze factors that affect outbreak of malaria using the Locally-Compensated Ridge Geographically Weighted Regression (LCR-GWR). The new method is used to ascertain the relationship between factors such as land use, distance to residence, distance to road, elevation, NDVI and the development of malaria. For variables related to geography, collinearity is used to verify two variables if they have linear relationship or highly correlated [8]. For example, the higher the elevation, the lower the temperature and contrary the higher the humidity (at the troposphere), or the age group will have a relationship with the employment. It is indeed obvious that the LCR-GWR can reduce the effects of local collinearity so as to improve the models' efficiency.

The proposed method is applied to Dak Nong province, located in the Central Highland region of Vietnam- an inhabited area for the minority community who has lacks condition and infrastructure with limited access to health services [23]. Dak Nong is determined within the geographical coordinate: $11^{\circ}45'$ to $12^{\circ}50'$ north latitude, $107^{\circ}13'$ to $108^{\circ}10'$ east longitude. Dak Nong shares border with Dak Lak province to the North and Northeast, border with Lam Dong province to the East and Southeast, border with Binh Phuoc province, "the cradle" of malaria in Vietnam [9], to the Southern and Southwestern and border with Cambodia to the West [10]. With its geographical conditions, Dak Nong has been one of the malaria hotspots. In the first 4 months of 2015, 176 cases of malaria and 175 patients were detected, with an increase of 68 cases and 72 patients compared to the same period in 2014. Although in 2016, the province set up impact mitigation campaigns for malaria and got some achievements.

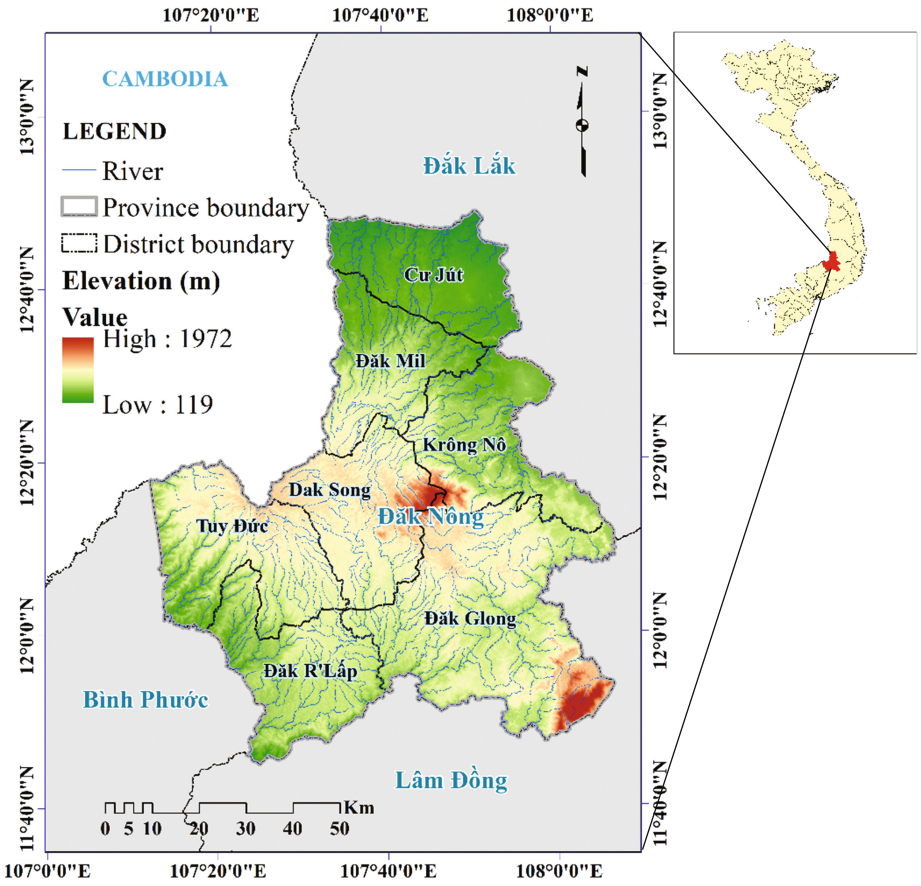

Fig. 1. Location of Dak Nong province, Vietnam

However, this region is still considered as hotspots of malaria from the past to present (Fig. 1).

The remainder of this paper is organized as follows: Sect. 2 introduces the datasets and LCR-GWR method. Section 3 presents the experimental results. Section 4 highlights some conclusions and further works of the paper.

2 Materials and Methods

2.1 Datasets

The vector for transmission of malaria is from infected mosquitoes' bite. In hot and humid climate territories, such as Vietnam, mosquitoes thrive as a favorable condition for malaria outbreaks [15]. According to WHO, the malaria parasites commonly found in Vietnam are P. vivax and P. falciparum through Anopheles mosquitoes [33]. In this

study, we do not focus on the epidemiology of malaria but would like to point out the conditions that affect the development of malaria through the collected survey data. The malaria data were collected from the Provincial Center for Preventive Medicine through a survey and inspection of more than 50,000 people from DakNong province in which 198 people were infected by malaria parasites in 2016. To assure and enhance the data quality, we have also visited the field to check and investigate more in prevalence malaria cases with a total of 209 observations which is expressed through Fig. 2a. From the data, we conduct a malaria hotspot map using Kriging algorithm. The result turns out that the malaria area is close to the border of Cambodia and Binh Phuoc province (Fig. 2b).

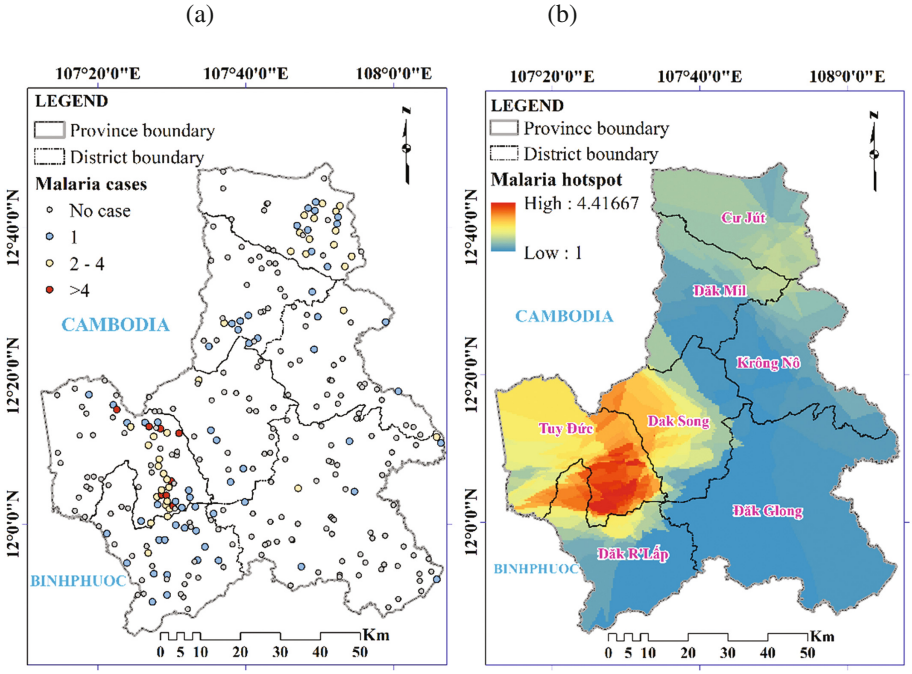

Fig. 2. (a) Distribution map of malaria; (b) Hotspot map of malaria in DakNong

2.2 Methods

In each regression model, the results indicate the level of predictor variable influence on the dependent variable. In this study, the dependent variable is the 209 malaria cases in 2016 for the whole province. The probable predictor variables were chosen based on the relation to malaria and through previous research and source of data. Several researches can provide these variables, such as vegetation, which plays a very important role in transmitting malaria. There are few indicators to express the properties vegetation but NDVI is the most widely used index. Temperature, rainfall,

humidity are the meteorological variables that are often used in predicting malaria transmission [17]. In addition, age group, gender are the social variables that are highly associated with malaria [32]. To take advantage of data collected, we selected 40 predictor variables as follows: the variables were divided into 2 groups which are natural conditions and social - economic conditions.

Natural condition variables

A DEM was provided globally from Aster Global DEM data (available at https://earthexplorer.usgs.gov/). The DEM is then used to generate Aspect and Slope using ArcGIS 10.4.1. Daily climatic data during 2016 were collected from 18 national meteorological stations and then averaged. Then we used Kriging in ArcGIS 10.4.1 to interpolate for the whole study area. NDVI, NDBI, and NDMI are calculated using bands from Landsat 8 OLI captured in Fed 12 2017 in which NDVI is Normalized Difference Vegetation Index, NDBI is Normalized Difference Built-up Index and NDMI is Normalized Difference Moisture Index. NDVI, NDMI and NDBI are defined as follow:

$$\begin{aligned} \mathrm{NDVI} &= (\mathrm{NIR}-\mathrm{R})/(\mathrm{NIR}+\mathrm{R}) \\ \mathrm{NDMI} &= (\mathrm{NIR}-\mathrm{IR})/(\mathrm{NIR}+\mathrm{IR}) \\ \mathrm{NDBI} &= (\mathrm{SWIR2}-\mathrm{NIR})/(\mathrm{SWIR2}+\mathrm{IR}) \end{aligned} \tag{1}$$

where NIR is the Near Infrared band, R is the Red band and SWIR is the Short-Wave Infrared band. The last variables are determined using Dak Nong land use map 2015 which was collected at the Dak Nong Department of Natural Resources and Environment at the scale of 1:50000. The land use map is categorized into 16 types name TTN (Religious land), SMN (Water surface), RSX (Production forest land), RPH (Protective forest land), RMP (Protective planted forest land), RDD (special forest land), OTC (Residential land), NTS (Aquaculture land), NHK (Upland land cultivate another annual crop), NKH (Other agricultural land), CSK (Productive land, non-agricultural business), CSD (Unused land), CLN (Perennial crops), CHN (Annual crop land), CDG (Specialized land), CCC (Public land) and few other point and line layers (river, road, hospital, school). Variable as forest is extracted in a combination of RSX, RPH, RPM and RDD. Agriculture land is extracted using CHN, CLN, NHK, NKH. Aquaculture land and residential land are NTS and OTC. The point and line objects as hospitals, rivers and roads were also extracted from land use map. For distance to hospital, we also selected few others which are inside 10 km distance from Dak Nong. These objects are respectively used to calculate distance using Euclidean Distance tool from ArcGIS software [13] (Table 1).

Social – economic condition variables

Data about population including population, density, gender, age group and building structure are collected from Dak Nong Statistic Yearbook 2015 at commune level. The data are then divided into sub-groups as shown in Table 2:

Due to the reason that the statistic is at commune level while the spatial malaria data are points, the statistic was calculated as the average of the whole commune. In addition, natural condition variables are also at Raster format and were added to the malaria table base on the location of the malaria point.

Table 1. Natural condition variables

Variable name	Code	Resolution/Unit/Method	Source
Aspect	Aspect	30 m	Extracted from Aster global DEM data
Elevation	DEM	30 m	Extracted from Aster global DEM data
Slope	Slope	30 m	Extracted from Aster global DEM data
Normalized Difference Built-up Index	NDBI	30 m	Calculated using SWIR 2 and NIR bands of Landsat 8 OLI Feb 12, 2017
Normalized Difference Vegetation Index	NDVI	30 m	Calculated using NIR and Red Bands of Landsat 8 OLI Feb 12, 2017
Normalized Difference Moisture Index	NDMI	30 m	Calculated using NIR and IR Bands of Landsat 8 OLI Feb 12, 2017
Distance to residence	DRe	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Resident area extracted from Landuse map 2015
Distance to road	DRo	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Roads extracted from Landuse map 2015
Distance to river	DRi	Calculated using Euclidean Distance in ArcGIS resolution 30 m	Rivers extracted from Landuse map 2015
Distance to hospital	DHo	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Hospital extracted from Landuse map 2015
Distance to aquaculture land	DAqL	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Aquaculture land extracted from Landuse map 2015
Distance to agriculture land	DAgL	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Agriculture land extracted from Landuse map 2015
Distance to wetland	DWe	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Agriculture land extracted from Landuse map 2015
Distance to forest	DFo	Calculated using Euclidean Distance in ArcGIS, resolution 30 m	Resident area extracted from Topography map 2015
Rainfall	Rain	Kriging (ArcGIS) average of daily rainfall during the year 2016 (mm)	Rainfall collected from 18 national stations

(*continued*)

Table 1. (*continued*)

Variable name	Code	Resolution/Unit/Method	Source
Temperature	Temp	Kriging (ArcGIS) average of daily temperature during the year 2016 (^{o}C)	Temperature collected from 18 national stations
Humidity	Humid	Kriging (ArcGIS) average of daily humidity during the year 2016 (%)	Humidity collected from 18 national stations

Table 2. Social – economic variables

Variable name	Code	Resolution/Unit	Source
Population	Pop	Population group includes 4 variables as Sum_pop, Urban_pop, Rural_pop, Pop_dens	DakNong Statistic yearbook 2015
Age	Age	Age is divided into 10 subs – group as 0–5 years old, 5–9 years old 10–19 years old, 20–29 years old, 30–39 years old, 40–49 years old, 50–59 years old, 60–69 years old, 70–79 years old, >80 years old (10 variables)	DakNong Statistic Yearbook 2015
Sex	Sex	Sex is divided into 2 subs – group as Pop_male, Pop_femal. (2 variables)	DakNong Statistic Yearbook 2015
Building structure	Build_struct	Building structure group includes 7 variables as Sum_build, Solid_build, Semi_sol_bui,un_so_bui,Simp_build, undef_build and Area_per_cap.	DakNong Statistic Yearbook 2015

We presume that the non-stationary relationship exists between malaria occurrences and influential factors. Therefore, the following scheme is proposed to analyze factors that affect outbreak of malaria using LCR-GWR (Fig. 3).

Step 1 is to determine the correlation coefficient between variables to eliminate those which are highly correlated. For instance, one of these variables as NDVI, NDBI, and NDMI can be eliminated if the correlation coefficient is higher than 0.7 (the correlation ranges from -1 to 1, with 0 value means there is no correlation between variables, we will give a table pairing all variables in the next section), we decide to keep those having high impact in the model. Then, in step 2, we run an OLS model to test the explanatory variables and determine how many percents of the data that potentially explain the malaria disease. However, spatial data, as variables in this research has unique characteristics. Firstly, the geographical factors are spatial auto-correlation which means the two locations that are close together should have similar characteristics compared to locations that are far away from each other. Secondly, geographic related factors for instance elevation, NDVI, population density are spatial

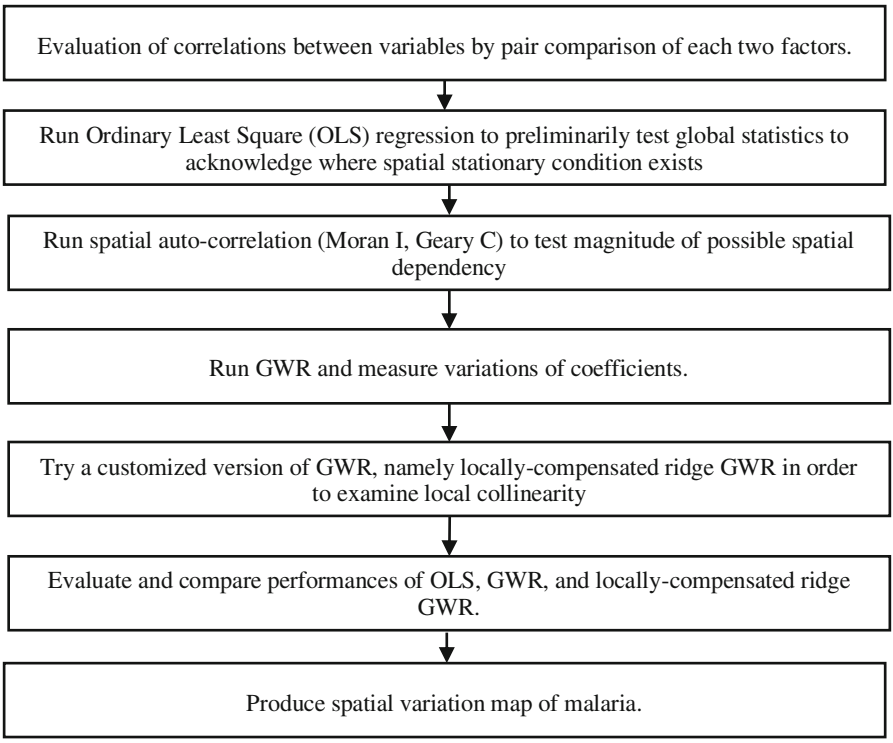

Fig. 3. Schematic overview of using LCR-GWR in prediction of malaria incidences

non-stationary. These factors value vary differently across the research area. Therefore, in step 3, spatial autocorrelation statistics as Moran I and Geary C are calculated to estimate the degree of spatial autocorrelation in a dataset [12, 22]. The Moran I runs with the residual of OLS regression. If Moran I approximately equals to 0 or less than 0, that means the data are randomly distributed or dispersed and an OLS model is fit. But if Moran Index is greater than 0, then the data are clustered and an OLS model is not appropriate. Step 4 in hence provides a GWR model will be run with the significant variables extracted from OLS model. The basic GWR model is:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \epsilon_i \tag{2}$$

where y_i is the dependent variable at location i, x_{ik} is the value of the k_{th} independent variable at location i, m is the number of independent variables, β_{i0} is the intercept term at location i, β_{ik} is the local regression coefficient for the k^{th} independent variable at location i and ε_i is the random error at location i. In this step, we need to consider the variable collinearity, which can be measured as condition number (CN) and variance inflation factors (VIFs). The condition number is used to assess the whole model while VIFs consider each variable in turn [8]. Therefore, the local condition number

(CN) was tested in this model to check if it is greater than 30, as proved in literature, means that there are local collinearities between variables [8, 16]. Step 5 is to provide a better GWR model in which AICc, R^2 and R^2 adjusted will be observed. Step 6 is to compare between OLS, GWR and LCR.GWR. In the final step, the coefficients of variables from LCR.GWR are used to calculate and provide a spatial variation map of malaria.

3 Results and Discussion

3.1 Evaluation of Correlation

Correlation coefficient (r) is a statistical indicator that measures the correlation between two variables. In this research, Pearson correlation coefficient is used to indicate the relationship of variables globally using *cor(x,y)* function in R [14]. The definition of the correlation coefficient is defined as follow:

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

where $\bar{x}$ and $\bar{y}$ is the mean value of variables x and y. Some variables were eliminated due to the correlation coefficient absolute value is higher than 0.7 (age structure group, gender group, NDBI, NDMI, Temperature). Strong correlations amongst the predictor variables indicate prominent level of collinearity. Therefore, this step is to filter the significant variables for OLS model (Table 3).

Table 3. Correlation between variables

	DEM	DAgL	DAqL	DFo	DHo	DRe	DRi	DRo	NDVI	Rain	Slope	Aspect	DWe
DEM	1	0.043	0.12	−0.32	−0.34	0.13	−0.25	0.03	0.06	0.46	0.17	0.006	0.16
DAgL	–	1	0.36	−0.19	0.02	0.68	−0.07	0.5	0.11	−0.03	0.048	0.04	0.30
DAqL	–	–	1	−0.28	0.23	0.58	0.008	0.43	0.11	−0.006	−0.000	0.001	0.25
DFo	–	–	–	1	0.06	−0.39	0.17	−0.3	−0.05	−0.28	−0.10	0.02	−0.11
DHo	–	–	–	–	1	0.04	0.14	0.09	0.08	−0.34	−0.06	−0.05	−0.13
DRe	–	–	–	–	–	1	−0.09	0.66	0.25	0.020	0.09	0.05	0.39
DRi	–	–	–	–	–	–	1	−0.03	−0.09	−0.24	−0.16	0.19	0.24
DRo	–	–	–	–	–	–	–	1	0.22	0.06	0.041	0.06	0.18
NDVI	–	–	–	–	–	–	–	–	1	0.08	0.23	0.04	−0.06
Rain	–	–	–	–	–	–	–	–	–	1	0.23	−0.006	−0.01
Slope	–	–	–	–	–	–	–	–	–	–	1	−0.08	0.05
Aspect	–	–	–	–	–	–	–	–	–	–	–	1	0.06
DWe	–	–	–	–	–	–	–	–	–	–	–	–	1

After calculating correlation between 40 preliminary input variables, only 13 variables are selected to use for regression models. They are elevation, distance to agriculture land, distance to aquaculture land, distance to forest, distance to hospital,

distance to residence, distance to river, distance to road, distance to wetland, NDVI, rainfall, slope, and aspect. These variables seem to be appropriated to explain the malaria incidence in this study area. However, we need to indicate which are the most significant variables.

3.2 Ordinary Least Square Regression (OLS) and Spatial Autocorrelation

Predictor variables listed in Table 3 were used as inputs to construct a model to predict malaria incident. ArcGIS 10.4.1 is used for generating and visualizing the OLS model. The global coefficient estimates together with their significance and VIFs for each predictor variable are shown in Table 4. Table 4 indicates significant variables to predict malaria which has p-value < 0.05 and none of VIFs are greater than 10 which mean that globally there is evidence of variable collinearity [8].

Table 4. Estimated regression coefficient for OLS model

Variables	Coefficient	Standard error	t-value	p-value	VIF [c]
Intercept	1.604376	1.239738	1.294125	0.197158	—
DEM	0.001541	0.000597	2.578767	**0.010645***	1.543405
DRe	0.000178	0.000124	2.032316	**0.049665***	3.781569
DRo	−0.000294	0.000146	−2.01361	**0.045420***	1.918792
NDVI	−1.736418	0.658739	−2.635972	**0.009060***	1.194432

Table 5. OLS summary

Number of observations	209
Akaike's Information Criterion (AICc) [d]:	714.086802
Multiple R-Squared [d]:	0.127098
Adjusted R-Squared [d]:	0.068905

Overall, summary of the OLS model in Table 5 shows that malaria is dependent on elevation, distance to residence, distance to road, and NDVI. This is reasonable because elevation influences many other variables as precipitation, temperature, slope and aspect, therefore it directly or indirectly influences the widespread of malaria. Meanwhile, NDVI represents factors such as forests and crops. For epidemiology, forestland is a good place for the development of malaria [32]. Distance to residential area and distance to roads are also explained by the activities of Central Highland people of the central highlands who have the custom to go into forests for cultivation or cutting wood, they move back and forth between regions, provinces, even across Cambodia borders for months and sometimes sleep in the forest. These people are at higher risk of malaria infection than the others [15]. Therefore, we can see that the area outside the residence, away from the main roads and close to the border have a higher incidence of malaria than normal.

There are several different measures: The R^2 is 0.127098 and the adjusted R^2 is 0.068905. The R^2 measures the proportion of the variation in the dependent variable which is accounted for by the variation in the model which has possible values range from 0 to 1. The adjusted R^2 is a preferable measure since it contains some adjustment for the number of variables in the model. In our OLS model, the value of R^2 adjusted is 0.068905 indicates that it accounts for about 6% of the variation in the dependent variable. This means this model has a substandard performance. There is a need to consider another model.

Another measure for evaluating model fit is provided by the Akaike Information Criterion (AIC) [16, 20]. Unlike the R^2 the AIC measure the 'relative distance' between the model that has been fitted and the unknown 'true' model [20]. Models with smaller values of the AIC is better fit than other with higher AIC value. The AIC in this case is 714.086802.

3.3 Moran Index of Residual

Global Moran I is used to determine autocorrelation of input variables for OLS model through interpretation of the residual.

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(x_i - \bar{x})} \tag{4}$$

where N is the number of observation, $\bar{x}$ is the mean of variable x, x_i is the variable value at location i, x_j is the variable value at location j, w_{ij} is the spatial weight. Moran I ranges from −1 (negative correlation) to +1 (positive correlation) (Table 6).

Table 6. Global Moran Index summary

Moran's Index:	**0.376422**
Expected Index:	−0.004808
Variance:	0.001222
z-score:	10.906042
p-value:	0.000000

In this research, the Moran I value is 0.376422 indicating that the variables are positively correlated. After the Moran I is computed, the Expected Index value will also be generated using the following formula:

$$E(I) = \frac{-1}{(n-1)} \tag{5}$$

where n is the number of observation. The Expected Index will also help to measure the variance. The Expected Index is then used in comparison with the Observed Index. The

z-score (standard deviation) and p-value (probability) are calculated using this comparison, which will indicate whether this data is statistically significant or not. Z-score is defined as follow:

$$z(I) = \frac{I - E(I)}{\sqrt{V(I)}} \tag{6}$$

where

$$V(I) = E(I^2) - E(I)^2 \tag{7}$$

V(I) is the variance. The value of z-score and p-value will also show whether or not to reject the null hypothesis. For this tool, the null hypothesis states that the values associated with features are randomly distributed [11]. The z-score and p-value in this situation mean that the null hypothesis will be rejected, the data is highly clustered with approximately 0% percent of being randomly distributed.

3.4 GWR and Variations of Coefficients

The results of Moran I OLS' residual indicate the consideration of using GWR while OLS can only explain 6% malaria collected data (R^2 adjusted = 0.068905). There are two main parameters required for GWR including bandwidth and kernel type. In this research, we used a corrected version of AICc to automatically specify the bandwidth. For the kernel, there are two possible choices for the Kernel type, FIXED and ADAPTIVE. FIXED kernel type is used for observations which are regularly distributed across the research area while ADAPTIVE kernel is for clustered observation [16, 20]. The Moran I have already indicated that data are clustered. Therefore, for better fit and accuracy, and also an ADAPTIVE kernel will cover most applications [20], an ADAPTIVE kernel was used. GWR function in ArcGIS (spatial statistic toolbox) was used to generate GW model.

The results from GWR model are express in Table 7 and Fig. 4.

Table 7. GWR summary

	Standard GWR	Compare to OLS model
Kernel type	Adaptive	
AICc	**679.41602964841491**	714.086802
R2	**0.36841843019175602**	0.127098
R2Adjusted	**0.23326644249726569**	0.068905

The results show a significant improvement of GWR compare to OLS: AICc decreased from 714.09 to 689.42. The lower the AICc, the better performance the model. R^2 and R^2 adjusted are highly improved indicate that GWR explained more in malaria collected data than OLS (23.32%). However, the condition number is higher than 30 in some locations which is considered to have collinearity between variables

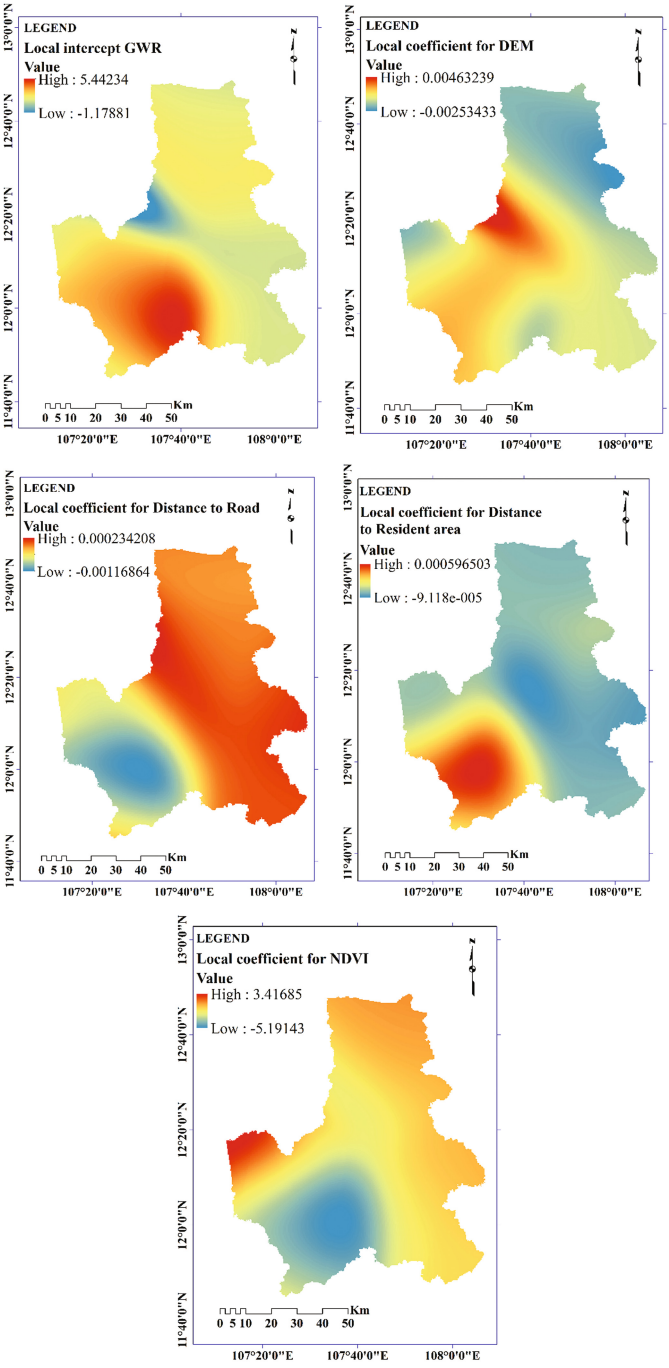

Fig. 4. Local parameters for GWR

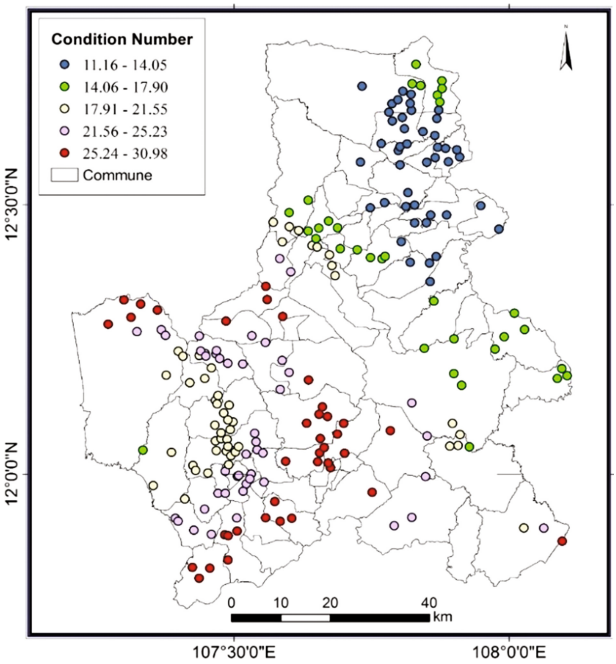

Fig. 5. Spatial variations in condition number (CN) of coefficients

[16]. The CN values shown in Fig. 5 demonstrate that the collinearity appeared near the malaria hotspots.

3.5 Locally-Compensated Ridge GWR

One of the methods to reduce collinearity in the explanatory variables of a linear model is ridge regression. The estimator of a ridge regression is altered to include a small change to values of diagonal of the cross-product matrix known as ridge shown as λ in the following equation:

$$\beta = \left(x^T x + \lambda I\right)^{-1} x^T y. \tag{8}$$

The relationship between condition number and ridge parameter is shown as follow:

$$\lambda = \{(\epsilon_1 - \epsilon_P) \,/\, (\kappa - 1)\} - \epsilon_P, \tag{9}$$

where ϵ_1, $\epsilon_2 \ldots \epsilon_P$ are the eigenvalues of the matrix $(x^T x)$, κ is the condition number. In locally-compensate ridge, the estimator for GW regression model is:

$$\beta(u_i, v_i) = \left(x^T W(u_i, v_i)x + \lambda I(u_i, v_i)\right)^{-1} x^T W(u_i, v_i)y. \quad (10)$$

$\lambda I(u_i,v_i)$ is locally-compensated value of λ at location (u_i,v_i). The weight $W(u_i,v_j)$ (weight at location i with coordinate value (u,v)) is calculated based on its distance to the center of the kernel as follows:

$$W(u_i, v_i) = 1 - \left(d(u_i, v_i))^2/b^2\right), \quad (11)$$

where $d(u_i,v_i)$is the distance in meters from the center of the kernel to the data point and b is the bandwidth [16, 30]. The same bandwidth of GWR can be applied to LCR-GWR model. LCR-GWR bandwidth was run in R environment *(bw.gwr.lcr* in GWmodel) then put into ArcGIS for generating intercept and coefficients. From the results, the bandwidth 51 which has the smallest CV score was chosen for GWR model with the CN threshold less than 30 (Table 8 and Fig. 6).

Table 8. Locally compensated ridge GWR summary

	Locally-compensated ridge GWR (LCR-GWR)	Compare to standard GWR model
Kernel type	Adaptive	Adaptive
AICc	**220.50814901236132**	679.41602964841491
R2	**0.63565935117477057**	0.36841843019175602
R2Adjusted	**0.50627226519768787**	0.23326644249726569

Variations maps of malaria occurrences

The results of LCR-GWR is improved in comparison with basic GWR with R^2 adjusted = 0.506. The intercept and coefficient will then be used in ArcGIS along with the DEM, distance to road, distance to residence, NDVI value run in the previous steps to generate the spatial variation of malaria occurrence (Fig. 7). The malaria prevalence map seems to be appropriate with the malaria occurrence near the border of Cambodia and Binh Duong at the low terrain area.

The results of this study indicated that it is necessary to focus on the collinearity among variables, especially variables that varied spatially. The regression models also show that the influence of variables on the incidence of malaria is reasonable, especially with high mountainous areas with most people being ethnic minorities who have different customs. OLS or standard GWR may explain some cases of malaria. However, when collinearity occurs between variables, the model becomes less precise and predictive. LCR-GWR was used to reduce the local collinearity by using the locally-compensated value of ridge with condition number less than 30. The results showed a significant improvement within which the established model capable of

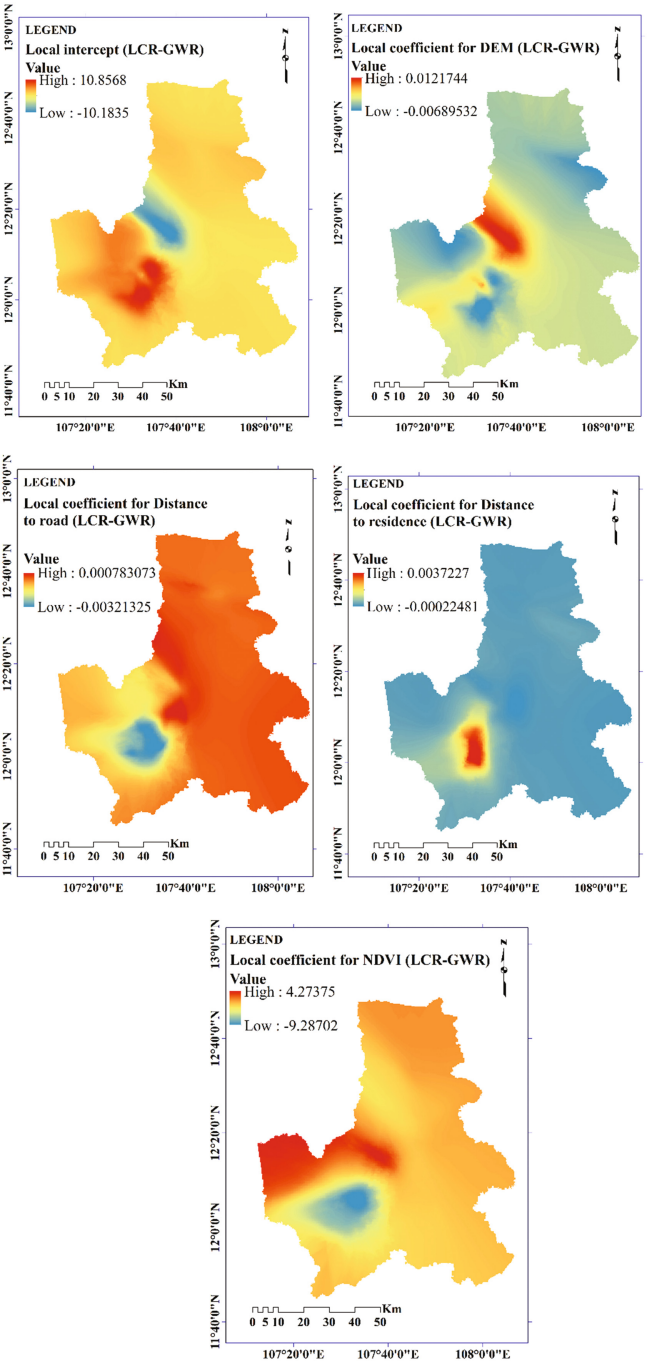

Fig. 6. Spatial variations of coefficients by locally-compensated ridge GWR

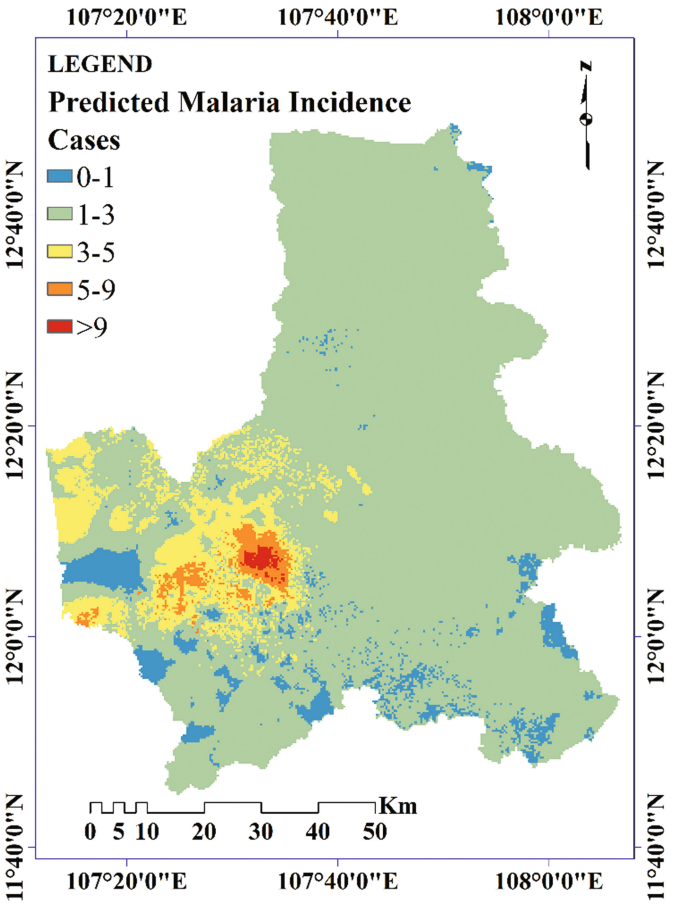

Fig. 7. Spatial variation of malaria occurrence

explaining more than 50% of cases of malaria in the study area by considering factors such as elevation, distance to roads, distance to residential areas and NDVI.

4 Conclusions

Variation of malaria hotspot is subject to change across the study area depending on local physical and environmental conditions. In this paper, we proposed a new scheme to analyze factors that affect outbreak of malaria using the LCR-GWR. Forty variables that were likely to influence the distribution of malaria occurrences were selected and filtered out by correlation to keep the most predictive power ones for modeling. The remaining variables were analyzed by the ordinary least square, GWR and GWR analysis with LCR term to result in the final variation map of malaria incidences in Dak Nong province, Viet Nam. The results showed a significant improvement from OLS to

GWR, and from GWR to LCR-GWR, where local collinearity was taken into consideration. The local collinearity between variables significantly reduced the results of GWR analysis with about twenty three percent of the malaria cases were explained compare to nearly fifty percent of LCR-GWR. However, LCR-GWR can only be necessary when local condition number was found to be greater than 30 that indicate the collinearity between variables. In addition, the results also support the fact that NDVI, DEM, Distance to residence and Distance to road are the most controlling factors to detect malaria hotspot in the study area. The local combination of the four significant variables determined magnitudes in local community that exposures to the diseases.

From the view of statistics, it shows that malaria occurrences in Dak Nong mainly distribute in the low terrain and near the border of "malaria cradle" Binh Duong and Cambodia. The application of statistics and regression model feature the advantage of eliminating the human's subjective thought but requires accurate statistical and large enough sample size data. For malaria, the factors that affect the incidence of malaria are very complex and vary in different areas. However, the regression model in this research, especially LCR-GWR accounts for approximately half of the cases in the study area. In this way to show, this tool is useful for the study of malaria in the future in Vietnam and across the world in general.

As its name, GWR is much depended on the variation of local physical and social condition. This method should be used in different geographical regions to validate its predictive capability. On the other hand, since this study employed point locations of malaria occurrences (because of data limitation), all aggregated social factors were removed from the analysis that might reduce accuracies of prediction map. There is a possibility to divide the study area into small sub parts to take explanatory capability of social statistics factors such as occupation or livelihood behaviors. The spatial variation in scale selection between point and polygons might produce new insight into malaria hotspot study.

Acknowledgment. This research is funded by the Vietnam National University, Hanoi (VNU) under project number QG.17.20

References

1. Adimi, F., Soebiyanto, R.P., Safi, N., Kiang, R.: Towards malaria risk prediction in Afghanistan using remote sensing. Malar. J. **9**, 125 (2010)
2. Comber, A., Harris, P., Quan, N., Chi, K., Hung, T., Phe, H.H.: Local variation in hedonic house price, Hanoi: a spatial analysis of SQTO theory. In: GIScience 2016
3. Beguin, A., Louis, V.R., Hales, S., Rocklov, J., Astrom, C., Sauerborn, R.: The opposing effects of climate change and socio-economic development on the global distribution of malaria. Glob. Environ. Change **21**, 1209–1214 (2011)
4. Aregawi, M., Lynch, M., Bekele, W., Kebede, H., Jima, D., Taffese, H.S., Yenehun, M.A., Lilay, A., Williams, R., Thomson, M., Nafo-Traore, F., Admasu, K., Gebreyesus, T.A., Coosemans, M.: Time series analysis of trends in malaria cases and deaths at hospitals and the effect of antimalarial interventions, 2001–2011 Ethiopia. PLoS ONE **9**, e106359 (2014)

5. Wijayanto, A.W., Purwarianti, A., Son, L.H.: Fuzzy geographically weighted clustering using artificial bee colony: an efficient geo-demographic analysis algorithm and applications to the analysis of crime behavior in population. Appl. Intell. **44**, 377–398 (2016)
6. Buczak, A.L., Baugher, B., Guven, E., Ramac-Thomas, L.C., Elbert, Y., Babin, S.M., Lewis, S.H.: Fuzzy association rule mining and classification for the prediction of malaria in South Korea. BMC Med. Inform. Decis. Mak. **15**, 1–17 (2015)
7. Ch, S., Sohani, S.K., Kumar, D., Malik, A., Chahar, B.R., Nema, A.K., Panigrahi, B.K., Dhiman, R.C.: A support vector machine-firefly algorithm based forecasting model to determine malaria transmission. Neurocomputing **129**, 279–288 (2014)
8. Brunsdon, C., Charlton, M., Harris, P.: Living with collinearity in local regression model. In: Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
9. http://www.impehcm.org.vn/noi-dung/sot-ret/binh-phuoc-cai-noi-sot-ret-va-sot-ret-khang-thuoc-cua-viet-nam.html
10. http://daknong.gov.vn/web/dak-nong-english/daknong-introduction
11. http://resources.esri.com/help/9.3/arcgisengine/java/gp_toolref/spatial_statistics_tools/how_spatial_autocorrelation_colon_moran_s_i_spatial_statistics_works.htm
12. Ge, Y., Song, Y., Wang, J., Liu, W., Ren, Z., Peng, J., Lu, B.: Geographically weighted regression-based determinants of malaria incidences in northern China. Trans. GIS (2016)
13. Gonçalves, D.N.S., de Morais Gonçalves, C., de Assis, T.F., da Silva, M.A.: Analysis of the difference between the euclidean distance and the actual road distance in Brazil. Transp. Res. Procedia **3**, 876–885 (2014)
14. Zhou, H., Deng, Z., Xia, Y., Fu, M.: A new sampling method in particle filter based on Pearson correlation coefficient. Neurocomput. J. **69**, 2138–2141 (2016)
15. Hoang, H.: Nghiên cứu thực trạng sốt rét và đánh giá kết quả can thiệp phòng chống sốt rét tại một số xã biên giới của huyện Hưzng Hóa, tỉnh Quảng Trị (Malaria situation and evaluation of malaria control interventions in several border communes of Huong Hoa district, Quang Tri province). Ph.D. in Community Medicine, vol. PhD. Hue College of Medicine and Pharmacy, Hue, Viet Nam (2014)
16. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P.: GW model: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. **63**, 1–49 (2015)
17. Kiang, R., Adimi, F., Soika, V., Nigro, J., Singhasivanon, P., Sirichaisinthop, J., Leemingsawat, S., Apiwathnasorn, C., Looareesuwan, S.: Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. Geospat. Health **1**(1), 71–84 (2006)
18. Krefis, A.C., Schwarz, N.G., Nkrumah, B., Acquah, S., Loag, W., Oldeland, J., Sarpong, N., Adu-Sarkodie, Y., Ranft, U., May, J.: Spatial analysis of land cover determinants of malaria incidence in the ashanti region Ghana. PLoS ONE **6**, e17905 (2011)
19. Lubetzky-Vilnai, A., Ciol, M., McCoy, S.W.: Statistical analysis of clinical prediction rules for rehabilitation interventions: current state of the literature. Arch. Phys. Med. Rehabil. **95**, 188–196 (2013)
20. Charlton, M., Fotheringham, A.S.: Geographically Weighted Regression - A tutorial on using GWR in ArcGIS 9.3 (2007)
21. Mosha, J.F., Sturrock, H.J.W., Greenwood, B., Sutherland, C.J., Gadalla, N.B., Atwal, S., Hemelaar, S., Brown, J.M., Drakeley, C., Kibiki, G., Bousema, T., Chandramohan, D., Gosling, R.D.: Hot spot or not: a comparison of spatial statistical methods to predict prospective malaria infections. Malar. J. **13**, 1–12 (2014)

22. Ndiath, M.M., Cisse, B., Ndiaye, J.L., Gomis, J.F., Bathiery, O., Dia, A.T., Gaye, O., Faye, B.: Application of geographically-weighted regression analysis to assess risk factors for malaria hotspots in Keur Soce health and demographic surveillance site. Malar. J. **14**, 463 (2015)
23. General Statistical Office: Statistical Yearbook of Dak lak. Dak Lak statistical office, Dak Lak (2016)
24. Masimalai, P.: Remote sensing and Geographic Information Systems (GIS) as the applied public health & environmental epidemiology. Int. J. Med. Sci. Pub. Health **3**, 1430–1438 (2014)
25. Rusk, A., Highfield, L., Wilkerson, J.M., Harrell, M., Obala, A., Amick, B.: Geographically-weighted regression of knowledge and behaviour determinants to anti-malarial recommending and dispensing practice among medicine retailers in western Kenya: capacitating targeted interventions. Malar. J. **15**, 562 (2016)
26. Son, L.H.: Enhancing clustering quality of geo-demographic analysis using context fuzzy clustering type-2 and particle swarm optimization. Appl. Soft Comput. **22**, 566–584 (2014)
27. Son, L.H.: A novel kernel fuzzy clustering algorithm for Geo-Demographic Analysis. Inf. Sci. **317**, 202–223 (2015)
28. Son, L.H., Cuong, B.C., Lanzi, P.L., Thong, N.T.: A novel intuitionistic fuzzy clustering method for geo-demographic analysis. Expert Syst. Appl. **39**, 9848–9859 (2012)
29. Son, L.H., Cuong, B.C., Long, H.V.: Spatial interaction – modification model and applications to geo-demographic analysis. Knowl. Based Syst. **49**, 152–170 (2013)
30. Son, L.H., Lanzi, P.L., Cuong, B.C., Hung, H.A.: Data mining in GIS: a novel context-based fuzzy geographically weighted clustering algorithm. Int. J. Mach. Learn. Comput. (IJMLC) **3**, 235–238 (2012)
31. Stensgaard, A.-S., Vounatsou, P., Onapa, A.W., Simonsen, P.E., Pedersen, E.M., Rahbek, C., Kristensen, T.K.: Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in Uganda: predictors of risk and geographical patterns of co-endemicity. Malar. J. **10**, 298 (2011)
32. Thanh, P.V., Van Hong, N., Van Van, N., Van Malderen, C., Obsomer, V., Rosanas-Urgell, A., Grietens, K.P., Xa, N.X., Bancone, G., Chowwiwat, N., Duong, T.T., D'Alessandro, U., Speybroeck, N., Erhart, A.: Epidemiology of forest malaria in Central Vietnam: the hidden parasite reservoir. Malar. J. **14**, 86 (2015)
33. WHO: World Malaria Report 2016 (2016)
34. Zacarias, O.P., Boström, H.: Comparing support vector regression and random forests for predicting malaria incidence in Mozambique. In: 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 217–221 (2013)