pagepress

Geospatial Health 2023; volume 18:1184

Understanding the spatial non-stationarity in the relationships between malaria incidence and environmental risk factors using Geographically Weighted Random Forest: a case study in Rwanda

Gilbert Nduwayezu,[1,2] Pengxiang Zhao,[1] Clarisse Kagoyire,[1,3] Lina Eklund,[1] Jean Pierre Bizimana,[4] Petter Pilesjo,[1] Ali Mansourian[1,5]

[1]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden; [2]Department of Civil, Environmental and Geomatics Engineering, University of Rwanda, Kigali; [3]Centre for Geographic Information Systems and Remote Sensing, University of Rwanda, Kigali; [4]Department of Geography and Urban Planning, University of Rwanda, Kigali, Rwanda; [5]Lund University's Profile Area: Nature-based Future Solutions, Sweden

Correspondence: Ali Mansourian, Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, 223 62 Lund, Sweden. E-mail: ali.mansourian@nateko.lu.se

Key words: geographically weighted random forest; variable importance; partial dependent plot; malaria incidence; Rwanda.

Funding: this research was funded by the Swedish International Development Agency (SIDA)-Rwanda Bilateral Programme Contribution No: 11227.

Contributions: GN: conceptualized the study, developed the methodology, formal analysis, writing (review and editing), and visualization. PZ: contributed to the methodology and revised the manuscript. CK: contributed to the methodology and revised the manuscript. LE: contributed to the methodology and revised the manuscript. JPB: interpreted the findings and revised the manuscript. PP: critically revised the manuscript. AM: conceptualized the study,developed the methodology, and critically revised the manuscript. All authors read and approved the final version of the manuscript.

Conflict of interest: the Authors declare no conflict of interest.

Availability of data and materials: all data generated or analyzed during this study are included in this published article.

Acknowledgements: we express our gratitude to the Swedish International Development Agency (SIDA) through the SIDA-University of Rwanda program, undernutrition sub-program, with the collaboration of Lund University and the University of Rwanda. The authors are also grateful to the RBC (Rwanda Biomedical Centre) for providing the malaria incidence dataset used for this paper. We would like to thank the anonymous reviewers who reviewed the manuscript.

Received for publication: 9 January 2023.
Accepted for publication: 28 April 2023.

Geospatial Health 2023; 18:1184
doi:10.4081/gh.2023.1184

Abstract

As found in the literature on health studies, the levels of ecological association between epidemiological diseases have been found to vary across regions. Due to limited research, little is known about how spatial environmental factors influence the variability of malaria incidence at smaller scales. We implemented the geographically weighted random forest (GWRF) machine-learning algorithm to analyze ecological disease patterns caused by spatially non-stationary processes using a malaria incidence dataset as well as a suite of diverse resolution environmental covariates for Rwanda. We first compared the geographically weighted regression (GWR), the global random forest (GRF), and the geographically weighted random forest (GWRF) to examine the spatial non-stationarity in the non-linear relationships between malaria incidence and risk factors. We used the Gaussian areal Kriging model to disaggregate the malaria incidence at the local administrative cell level to understand the relationships at a fine scale since the model goodness-of-fit was not satisfactory to explain malaria incidence due to the limited number of sample values at the health centre catchment level. Our results show that in terms of the coefficients of determination and prediction accuracy, the GWRF model outperforms the GWR and GRF models. The coefficients of determination of the GWR (R^2), the GRF (R^2), and the GWRF (R^2) were 0.47, 0.76, and 0.79, respectively. The local R^2 showed that the GWRF algorithm had higher performance in explaining the spatial variations of the non-linear relationships between malaria and the underlying factors, which could have implications for supporting local initiatives formalaria elimination in Rwanda.

Introduction

Malaria, a protozoan disease of the red blood cell transmitted by the bite of an infected female Anopheles mosquito, is a major public health risk in Rwanda, where children and pregnant women are the most vulnerable groups. It remains the main cause of morbidity and mortality among children in Rwanda (Habyarimana and Ramroop, 2020). Previous studies have denoted an association of environmental variables, e. g., elevation (Hasyim *et al.*, 2018), precipitation (Midekisa *et al.*, 2015), relative humidity (Chirebvu *et al.*, 2016) and the normalized difference vegetation index (NDVI) (Kibret *et al.*, 2019). In all the reviewed research on this

topic, using linear parametric models, various environmental predictors have been found to be correlated with malaria incidence in Rwanda. For instance, testing temperature and rainfall as malaria predictors using an autoregressive lag effects equation Loevinsohn (1994) revealed that temperature predicted incidence best at higher altitudes. Gasana *et al.* (1996) found that the populations located near Nyabarongo River and other water bodies have an increased incidence of malaria based on the Spearman rank correlation coefficient, while rainfall was identified as inversely correlated to malaria. In addition, Bizimana *et al.* (2015) developed a composite index of social vulnerability to malaria in a spatial multicriteria analysis environment. They discovered a strong influence of population density on malaria incidence. Using the same predictors, Bizimana *et al.* (2016) applied a spatially explicit approach to delineate homogeneous regions of social vulnerability to malaria, which revealed high levels of social vulnerability to malaria in the highland areas as well as in remote areas of Rwanda. The impact assessment of socioeconomic and environmental factors on malaria persistence in Rwanda by using a logistic regression model revealed that the increased malaria prevalence is due to lower altitudes and proximity to irrigated farmland (Habyarimana & Ramroop, 2020; Kateera *et al.*, 2015; Rudasingwa & Cho, 2020). Murindahabi *et al.*, (2021) used a multiple regression model combining the digital elevation model (DEM), the NDVI and the normalized difference water index NDWI, population density and distance to marshland, to determine the main predictors of malaria vector abundance. Their findings showed that the distance to river network and elevation played a key role in explaining malaria mosquito abundance.

All previous studies in Rwanda have denoted an association of environmental variables with malaria using linear parametric models. Examples include a linear regression model (Hakizimana *et al.*, 2018), Bayesian geostatistical models (Semakula *et al.*, 2020), a multinomial logit model (Rudasingwa & Cho, 2020; Murindahabi *et al.*, 2021), a logistic model (Kateera *et al.*, 2015) and a spatial multicriteria analysis (Bizimana *et al.*, 2015, 2016). These models have their own assumptions and require pre-defined underlying relationships between the response and explanatory variables. The multinomial logistic model, for example, assumes that the choice probabilities of each pair of alternatives are independent of the presence of all other alternatives (Rudasingwa & Cho, 2020). Violation of these assumptions leads to inconsistent parameter estimates and biased predictions. Another serious weakness in the statistical regression model is that the relative spatial effects of explanatory variables on the response variable are not assessed (Cheng *et al.*, 2019).

Geographically weighting regression (GWR) is a robust algorithm that has been successfully used in regression analysis. The GWR model was used to examine the local relationship between the response and predictor variables. Its usefulness, however, is debatable (Peng *et al.*, 2019). The unbalanced distribution of the phenomenon under investigation can be analyzed using spatial regression models capable of predicting the correlations between causative variables and response variables in defined geographic regions (Georganos *et al.*, 2020). However, in some cases, these predictor and response variables do not necessarily have linear relationships (Quiñones *et al.,* 2021). Simple linear or traditional regression models such as the GWR, on the other hand, fail to capture these nonlinear relationships accurately due to their susceptibility to local collinearity that can yield unreliable results (Maiti *et al.*, 2021). The linear model is susceptible to outliers; therefore, strong assumptions about the relationships between response variables (linearity) and the predictors (collinearity) are required. Random forest (RF) and other nonlinear non-parametric models do not need to account for multicollinearity and can examine all explanatory variables without screening (Breiman, 2001). They can therefore be used for regression research to identify nonlinear relationships between variables even in high-dimensional settings with complex interactions (Cheng *et al.*, 2019). RF has some advantages that make it the proper choice for our studies, such as being easy to compute, tolerant of missing multicollinear data and not particularly prone to overfitting. It calculates error estimates without requiring validation data, the variables are ranked, and the influence of each variable on the outcome is calculated (Breiman, 2001). The RF is also more tolerant of data noise and outliers, has a higher fitting accuracy than support vector machine (SVM) and has fewer adjustment parameters (Breiman, 1996b; Genuer *et al.*, 2010). To our knowledge, there are a few studies on the use of RF in the study of malaria incidence causal factors. Rather than directly examining the variable importance of malaria incidence, most have attempted to compare the performance of the RF prediction method with that of other machine-learning methods (Cianci *et al.*, 2015; Harvey *et al.*, 2021; McCann *et al.*, 2014; Rhodes et al., 2022; Wang *et al.*, 2019). Exceptions include the studies of Cohen *et al.* (2013), Kapwata and Gebreslasie, (2016), and Georganos et al., (2020), who used global RF (GRF) to determine the main causative factor of malaria incidence. The RF algorithm, however, has the major disadvantage of interpreting the relationships between the response and explanatory variables (Georganos *et al.*, 2019) and it is one of the most accurate classification models except for regression (Sullivan, 2017). A recently developed nonlinear, non-parametric geographically weighted RF (GWRF) has been developed and used to solve the GWR and RF limitations (Georganos *et al.*, 2019; Maiti *et al.*, 2021; Quiñones *et al.*, 2021).

In this study, the GWRF was applied to examine the relationships between malaria incidence and the underlying factors to shed light on the spatial variations in nonlinear relationships between variables. The GWRF is promising with regard to data-mining due to its capacity to analyze various types of variables and assess their importance without prior model specification (Georganos *et al.*, 2020; Maiti *et al.*, 2021). To date, its applicability to this topic has been largely unexplored. To our knowledge, few researchers have adopted the GWRF method to determine, *e.g.*, the causative factors of population modelling (Georganos *et al.*, 2019), diabetes (Quiñones *et al.*, 2021) and COVID-19 (Maiti *et al.*, 2021). The goal of this study was not to predict malaria incidence, but rather to test the GWRF in order to model and map the contribution of individual factors to malaria incidence using remotely sensed environmental data. To exhibit the GWRF's robustness, it was compared with the GWR and GRF models.

This study is expected to contribute to the existing literature by providing a recent application of the GWRF method to examine the main factors causing malaria incidence. Second, it will showcase the use of GWRF in a 'scary' context where tough feature engineering is applied to get valuable inputs for better predictive performance. On top of that, the spatial downscaling approach applied can potentially address the problem of low model goodness-of-fit where there is a limited number of sample values used. Understanding the relative importance of explanatory variables could significantly assist prediction of malaria incidence and therefore contribute to the improvement of clinical and intervention strategies for malaria elimination. This is fundamental for better

understanding the malaria pattern in Rwanda and the need to find the best possible decisions.

Materials and Methods

Study area

A population density of 394 persons per km^2 makes Rwanda one of the most populated countries in Africa. The large majority of Rwandans live in rural areas (NISR, 2018). The climate is conditioned by the topography: the further west, the lower the altitude resulting in warmer temperatures and lower levels of precipitation in that part of the country (Gasana *et al.*, 1996), which is therefore favoured by mosquitoes. Rwanda significantly lowered the incidence of malaria between 2005 and 2011 through the scaling up of interventions, however from 2012 to 2017, there was an increase in the number of cases. Malaria incidence in Rwanda is characterized by spatial variability, manifested by clustered patterns of malaria cases. As a result, effective malaria elimination necessitates a spatial perspective with a geographical component.

Mapping of malaria prevalence

In Rwanda, most malaria cases are reported and treated at the health centre level. Patients tend to attend the nearest health facility, which is not only the strongest factor in malaria treatment and health seeking behaviour, but also the only factor that can be affected by the patient. This choice implies that travel distance (or travel time) has the highest influence on the outcome. The health catchment (HC), also known as a service area, is the polygon surrounding a health facility that includes the majority of people who use its services (Macharia *et al.*, 2022). A HC serves as the core building block for estimating a reliable population denominator for disease mapping, appropriate healthcare planning and resource allocation within a population. Consequently, the methods employed to define the HC significantly impact the model's accuracy and interpretability. From the literature, straight-line distances, also called Euclidean (Pattnaik *et al.*, 2021; Stresman *et al.*, 2014), the Thiessen polygon (Kundrick *et al.*, 2018) and its derivatives were used to delineate the HC. The Euclidean approach, however, constrains this definition because it does not consider potential, physical barriers that may impede malaria patients from reaching the nearest health facility (Macharia *et al.*, 2022). On the other hand, the creation of catchment areas overcomes such barriers by being more effective in displaying the incidence data. The use of catchment areas also accounts for physical barriers and depicts how a patient might move across a landscape (Bizimana & Nduwayezu, 2021). Under the assumption that walking is the most common transport mode in rural areas in Rwanda, a cost-allocation model was created by taking into consideration physical barriers such as rivers, lakes, flooded areas, water bodies and topography.

We used both the annual average malaria incidence and a point layer of health centres acquired from the Rwandan Ministry of Health to delineate the HCs. A cost allocation analysis was then performed based on geographic coordinates as input source. To estimate malaria incidence in each polygon, we joined the data from delineated HCs with the health facility points. We then mapped and visualized the malaria incidence as a response variable for each HC (Figures 1 and 2).

Preparation of predictor variables

The predictors were selected based on their probable association with malaria incidence based on literature review and data availability. We used malaria incidence, vector polygons, elevation, population density, rainfall, normalized difference vegetation index (NDVI), land surface temperature (LST), air temperature, relative humidity and evapotranspiration raster data. Details on the data used are explained in Table 1, Figure 3, and the Discussion section.

All variables were raster images with the same spatial reference. First, using the Raster to Point geoprocessing tool in ArcGIS Pro 3.0 (ESRI, Redlands, CA, USA), we created point values by converting the initial input raster image to point features. In this case, a point was created for each cell of the input raster image. Second, we used the Extract Multi Values to Points geoprocessing tool to extract cell values from these points for other input rasters. A new field containing the cell values for each input raster was appended to the initial input raster image converted to a point feature class. The input rasters were not resampled; instead, the cell values were extracted from all input rasters in their original resolution and spatial reference. Third, to generate a single value representing each HC, we then joined the HC polygons with all the generated input raster variable points from another layer using the

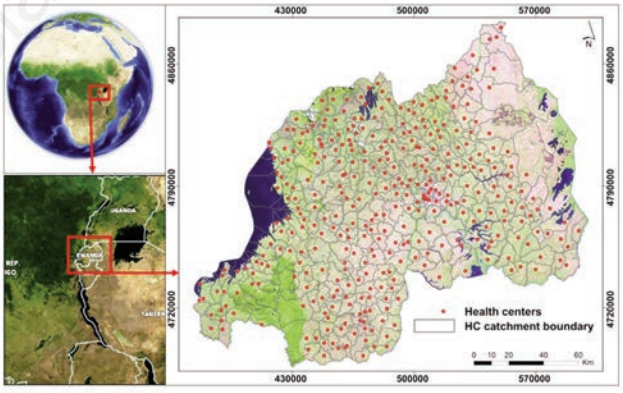

Figure 1. Location of the study area with catchment boundaries of the malaria health centres.

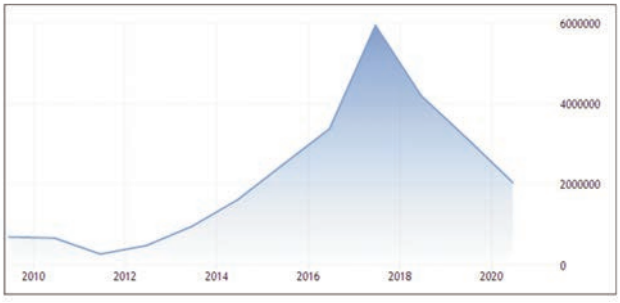

Figure 2. Malaria incidence between 2010 and 2020 in Rwanda. Source: U.S. President's Malaria Initiative [Rwanda] Malaria Operational Plan FY 2022 (https://www.pmi.gov).

Join Data tool based on the spatial location option. We summarized our attribute values by taking the average value. Later, to ensure that each covariate would contribute equally to the analysis, all variables were scaled using the z-score standardization method in the R environment. To confirm that our model would fit the data well, we computed the global Moran's *I* test to detect spatial residual autocorrelation and the local Moran's *I* to track potential spatial residual clustering (Anselin, 1995).

Modelling the spatial relationships

Gaussian areal Kriging-based model

Kriging interpolation is defined as the reaggregation of data from one set of polygons (the source polygons) to another set of polygons (the target polygons) (Comber & Zeng, 2019; Lam, 1983; Zeng & Comber, 2020). This interpolation fits for Gaussian data (Krivoruchko *et al.*, 2011), binomials (Flowerdew *et al.*, 1991) and Poisson data (Goovaerts 2006). All these three models differ only in their interpretations of the prediction surfaces and the reaggregated predictions. The Gaussian area Kriging model, which was used in this research, is applied when actual measurement locations are not provided (Krivoruchko *et al.*, 2011). This method is preferred when measurements are collected in relatively large polygons (Rosenshein, 2010). This process involves visiting each polygon centroid and identifying all centroids within a predetermined radius of that centroid. The mean inter-centroid distance is then computed and used to determine where the kernel levels out, *i.e.* cells at distances from the centroid greater than the mean inter-centroid distance; they are given zero weights in the calculations (Krivoruchko *et al.*, 2011). As a result, the size of the kernel is determined by the mean inter-centroid for each centroid. To select the best model, we tried different models in Table 2. Further details on each of the methods can be found in the work of (Krivoruchko *et al.*, 2011; Rosenshein, 2010). We computed twice for each cross validation: first with default parameters and again with hyperparameters turned. For measuring how well the data fits a model, we used evaluation statistics or error measures, including the mean, the root mean square (RMS), the mean square and the root mean square standardized (RMSS). We focused on root mean square standardized because it is the most widely used metric for choosing the optimal model (Ohmer *et al.*, 2017). We used the best

Table 1. Variables used.

Variable	Format	Year	Expression/ resolution	Source	Reference
Country boundary	Vector	2021	Vector	National Land Authority	Rwanda
Positive cases/HC	Vector	2016	Geographical coordinates	Ministry of Health	Rwanda
DEM	Raster	2013	Metres (90 m)	National Land Authority	Rudasingwa and Cho 2020; Hasyim *et al*., 2018
Population density	Raster	2016	Number/km^2	https://www.worldpop.org /geodata/listing?id=76	Murindahabi *et al*. 2021; Bizimana *et al*., 2016, Bizimana *et al*. 2015
Rainfall	Raster	2016	Millimetres (0.05°×0.05°)	https://data.chc.ucsb.edu/ products/CHIRPS-2.0/EAC_monthly/tifs/.	Loevinsohn 1994; Midekisa *et al*., 2015
NDVI	Raster	2016	-1.0 to +1.0 (0.01 m)	https://scihub.copernicus. eu/dhus/#/home	Murindahabi *et al*. 2021; McMahon *et al*., 2021; Kibret *et al*., 2019
LST	Raster	2016	Degrees Celsius (90 m)	https://scihub.copernicus. eu/dhus/#/home	Murindahabi *et al*., 2021; Rudasingwa and Cho, 2020; Loevinsohn, 1994
Air temperature	Raster	2016	Degrees Celsius (90 m)	Rwanda Meteorology Agency	Rulisa *et al*., 2013; Loevinsohn, 1994
Relative humidity	Vector	2016	Percentages (90 m)	Rwanda Meteorology Agency	Chirebvu *et al*., 2016; Sewe *et al*., 2016
Evapotranspiration	Vector	2016	Millimetres (90 m)	Rwanda Meteorology Agency	Chirebvu *et al*., 2016; Loevinsohn, 1994

DEM, digital elevation model; NDVI, the normalized difference vegetation index; LST, land surface temperature.

Table 2. Cross-validation statistics of default and adjusted model results.

Indices	Mean		RMS		MS		RMSS	
Model	Default	Adjusted	Default	Adjusted	Default	Adjusted	Default	Adjusted
Spherical	-60.993	-140.759	7,151.348	7,010.523	0.001	-0.036	3.610	3.566
Exponential	-46.961	-117.259	7,060.010	6,908.593	0.000	-0.016	1.798	1.767
Circular	-58.510	-140.478	7,146.924	7,009.579	0.002	-0.038	3.831	3.791
Tetraspherical	-62.594	-140.887	7,153.816	7,010.759	0.000	-0.035	3.464	3.425
Gaussian	-3,123.799	-6,493.449	27,804.628	38,631.592	-37.735	-36.120	411.012	275.017
K-Bessel	-246.037	-378.039	20,051.147	18,985.190	-13.405	-13.037	188.520	202.043
Rational quadratic	-303.260	-338.545	8,492.594	8,254.878	-0.104	-0.090	5.053	5.072
Stable	-3,123.799	-6,493.449	27,804.628	38,631.592	-37.735	-36.120	411.012	275.017

RMS, root mean square; RMSS, root mean square standardized: MS, mean square.

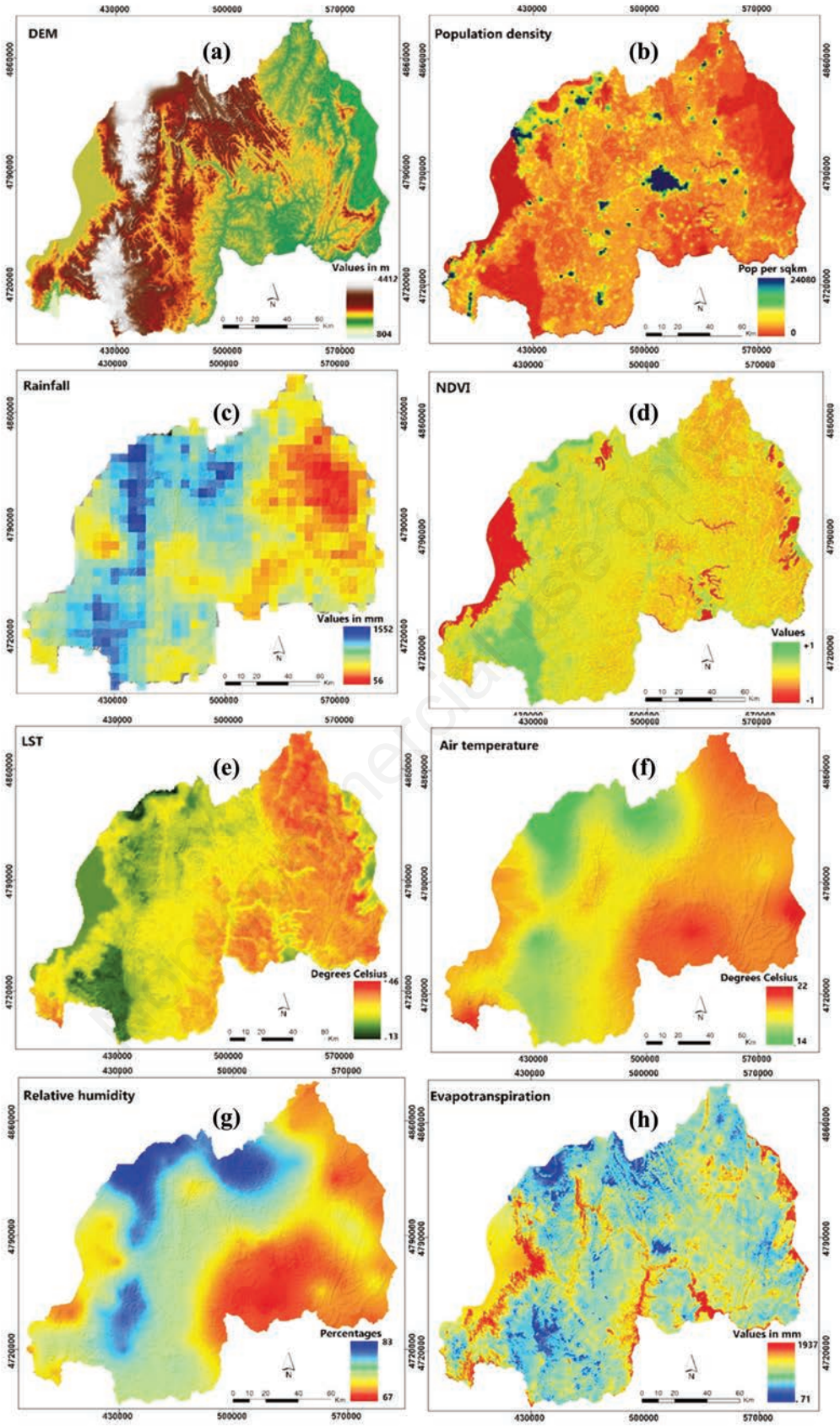

Figure 3. Model predictor input variables.

model to disaggregate malaria incidence from the malaria catchment to the administrative cell level to compute new values for predictor variables.

Geographically weighted regression (GWR)

Various global models, such as ordinary least squares (OLS), are in existence for understanding the relationships between disease incidence and its risk factors. The OLS model is always applied to analyze the relationships between a set of predictor variables and a response variable. It estimates the global statistic that assumes a stationary and constant relationship over space, so the estimated parameters are the same for the entire study area (Brunsdon *et al.*, 1996):

$$y_i = \beta_0 + \sum \beta_i x_i + e_i \quad \text{(Eq. 1)}$$

where y_i is the value of the dependent variable at location i, β_0 the intercept; β_i the coefficient that shows the magnitude of change in the response variable y with 1 the unit change in the predictor variable x_i; and e_i the error term.

Regrettably, Eq 1 does not take into consideration the spatial heterogeneity in the variables under investigation (Anselin & Sergio, 2014). The fundamental premise underlying all non-spatial models is that the spatial relationships between variables are the same across space. Hence, the spatial dynamics of the explanatory variables used cannot be explained by these models. Real-world phenomena such as disease prevalence, human interaction and physical environmental indicators, on the other hand, vary in space even at the micro-level. To uncover local dynamics and enrich reality, such strict spatial stationarity must be relaxed. To address this problem, a GWR model was developed to examine a non-stationary relationship between predictor variables and response variables (Fotheringham *et al.*, 2002; Kalogirou 2003). This technique takes non-stationary variables into consideration and models the local relationships between predictors and the outcome of interest based on the following formula:

$$y_i = \beta_0(u_i, v_i) + \Sigma_j \beta_j(u_i, v_i) x_{ij} + e_i \quad \text{(Eq. 2)}$$

where y_i is the dependent variable at location i; $\beta_0 (u_i, v_i)$ the intercept; x_{ij} the j^{th} predictor variable; $\beta_j (u_i, v_i) x_{ij}$ the j^{th} coefficient; and e_i the error term.

Due to spatially varying parameters in weighted analysis regression, each explanatory variable in the GWR model has different regression parameters (Kalogirou, 2003). The weights were calculated from a weighting scheme known as the kernel (Fotheringham *et al.*, 2002; Kalogirou, 2003). The local model, based on the combination of a geographically weighted estimator matrix, kernel, and bandwidth (Fotheringham *et al.*, 2002; Kalogirou, 2003; Peng *et al.,* 2019), was established and calibrated to distinguish the spatial association among nearby HCs. We applied a fixed Gaussian kernel function for the weighting scheme and determined the optimal bandwidth size using the golden search method (Fotheringham *et al.*, 2002), considering the score with the lowest Akaike's Information Criterion (AIC) value.

Global random forest (GRF)

The RF algorithm is a non-parametric method of statistical learning that took root through the so called bagging paradigm (Breiman, 2001; Grömping, 2009). Bagging predictors are part of a method constructing multiple versions of a predictor and combining them into a single aggregated component (Breiman, 1996). RF is a group of unpruned regression trees generated from a random sample of training data emanating from the bagging method (Grekousis *et al.*, 2022).

The general definition of the RF approach given by (Breiman, 2001) is the following: Let $(h., \Theta_1),\ldots,(h., \Theta_q)$ be a collection of tree predictors, with $\Theta_1,\ldots,\Theta_q$ *i.i.d.* random variables independent of $\mathcal{L}_n$; the random forest predictor $\hat{h}_{RF}$ is obtained by aggregating this collection of random trees, which is done as follows:

$$\hat{h}_R(x) = \frac{1}{q}\sum_{\ell=1}^{q} \hat{h}(x, \Theta_\ell) \quad \text{(Eq. 3)}$$

which gives the average of individual tree predictions in regression

$$\hat{h}_{RF}(x) = \underset{1\le c\le C}{\arg\max} \sum_{\ell=1}^{q} 1_{\hat{h}(x,\Theta_\ell)=C} \quad \text{(Eq. 4)}$$

which signifies the majority vote among individual tree predictions in classification

The output of the RF gives a direct estimate of the prediction error (Breiman, 1996), also known as the "out of bag (OOB)" error. The main idea behind this error estimator is to use samples that were not selected as test data (Maiti *et al.*, 2021), which is expressed as "variable importance" signifies average impurity reduction of regression forests (Breiman, 2001; Grömping, 2009). OOB error estimations and variable importance rankings are two key features of the RF method (Breiman, 1996; Genuer *et al.*, 2010). The remaining one-third of samples (known as OOB samples) are used for error monitoring in an internal cross-validation technique that calculates the number of correct predictions (Breiman 2001). The average error of all OOB forecasts is then used to calculate the overall OOB score. The set of observations not used for building the current tree, the OOB sample, is first used to estimate the prediction error and then to evaluate the variable importance (Genuer *et al.*, 2010).

Using RF, the predictive power of variables can be measured using a variety of methods (e.g., the permutation feature, the Gini index, the accuracy decrease), but we employed an increase in mean square error (IncMSE%) to calculate the relevance of each variable (Genuer *et al.*, 2010). We utilized the partial dependence plots to characterize the nonlinear relationships between the malaria incidence and its predictors. A partial dependence (PD) plot shows the functional relationship between a number of input predictors and the expected target responses and reveals how the predictions partially depend on the values of the input variables of interest. The PD plot can show whether the relationship between the response and a predictor variable is linear, monotonic or more complex (Molnar, 2022).

Geographically weighted random forest (GWRF)

Georganos *et al.* (2019) proposed GWRF as a disaggregation of the RF in the geographical space in the form of local sub-models. It is a predictive method based on the concept of spatially varying coefficient models, in which a global process is decomposed into many local sub-models. GWRF is in fact a local GRF version that enables the investigation of spatial non-stationarity in the relationship between the response variable and a set of predictors.

While the vast majority of RF problems can be solved using a unique (or global) model, this method generates multiple spatially weighted (or local) RF models (Maiti et al. 2021). The equation for a typical GWRF model is:

$$Y_i = a(u_i, v_i)x_i + e \quad \text{(Eq. 5)}$$

where $a\ (u_i,\ v_i)x_{ij}$ is the prediction of the RF model calibrated for location i, and $u_i,\ v_i$ are the coordinates of the centroid of the spatial unit i.

In essence, either the number of nearest neighbours (adoptive kernel) or a distance threshold value (bandwidth-fixed kernel) is used to build the neighbourhood or kernel. We adopted the adaptive kernel bandwidth search approach to implement this model. The optimal bandwidth (BW) was determined using the minimized OOB (Fotheringham *et al.*, 2015, Georganos & Kalogirou, 2022) and we used a random grid search to determine the GWRF's optimal hyperparameters, i.e. the number of trees (ntree), number of predictor variables at each tree (mtry) and minimum size at each terminal node (node size). We used 10-fold cross-validation to select the most suitable hyperparameter combinations setting the number of trees to 500, the number of variables randomly sampled as candidates at each split to 2 and the bandwidth to 162 observations. We used the permutation feature importance approach to evaluate the contribution of the predictor variables in the models. Similar to the GRF we employed an increase in the mean square error (IncMSE%) and the RSS (residual sum of squares) to determine the importance of each variable. The model coefficient of determination (R^2) and the variable importance of each predictor variable were mapped to investigate the spatial variation and its effect on malaria incidence adopting *"SpatialML"*, a developed R package, to implement this model according to Georganos *et al.*, (2019).

Results

Disaggregated malaria incidence

Table 2 shows the cross-validation values for the default and adjusted models. It can be observed from Figure 4 that the latter perform better than the default models. There is, nevertheless, a slight difference between the values of the two models. In particular, the stable, K-Bessel, and Gaussian models do not fit the data as well as the default scatterplot curves. Although the exponential default curve looks better, the model can still be improved as it tends to predict negative values. Specifically, the adjusted exponential model gives the best fit, which was confirmed by the cross-validation where the adjusted exponential model had the lowest RMSS value (1.767). This is good because, values close to 1 indicate that the model fits the data and can be trusted. Once a proper prediction surface has been obtained, the surface can be used to predict malaria incidence. To understand the relationships at a fine scale, this model was chosen to disaggregate the malaria incidence at the cell administrative level. The map in Figure 5 illustrates the spatial distribution of the predicted malaria incidence using the adjusted exponential model. It shows the prediction surface for the malaria incidence (a), which was used to disaggregate malaria at the HC (b) and the cell administrative level (c). A cell is an administrative entity in Rwanda.

GWR, GRF and GWRF

As seen in Table 3, the goodness-of-fit varied between 0.47 and 0.79 for the different regression models (based on eight variables) investigated. Although the R^2 does not measure the level of model complexity, it tells us which model has the best goodness-of-fit, *i.e.* the higher the R^2, the better the fit with the observed data, *e.g.*, GWRF model explains approximately 79% of the variation in the response variable. In other words, the model predicts roughly 79% of the predicted malaria incidence. This shows how accurate the GWRF is at analyzing the correlation between the risk factors investigated and the malaria endemicity in most of the study areas, especially in the eastern and southern parts of the country.

From the GWRF model, we computed the average local effect and the proportion of the local variable importance of each explanatory variable on malaria incidence as per Figure 6 and Table 4. The statistics of the local coefficient of determination (R^2) of the GWRF, with the environmental variables, such as rainfall, LST, evapotranspiration and NDVI showing a strong relationship with the spatial distribution of malaria incidence. Contrary to expectation, DEM, air temperature, relative humidity and population density revealed only a moderate correlation with the malaria endemicity. The variable importance in Figure 6 represents the correlation between the predictor variables and the response variable for GRF and GWRF. As can be seen, rainfall and LST come out as the two most important variables for malaria incidence.

Figure 7 depicts the spatial distribution of the local R^2 of the GWRF model. It can be noted that the majority of cells, particularly those in the south-western, north-western, north-eastern, western and the central areas all have high local R^2 values. This indicates the proposed GWRF model yields an accurate result in the majority of the cells. However, R^2 values were lower in a few cells in the Southeast.

The Figure 7 indicates that the mean square error (IncMSE) would increase by the percentage displayed if a variable were excluded from the model. With this imbalanced and heterogeneous pattern, the GWRF model provides a detailed spatial distribution of the local importance of the all eight variables. The results show that elevation, evapotranspiration and LST predominantly affect the eastern part of the country, whereas population density and relative humidity define the malaria endemicity in the city of Kigali. However, rainfall, NDVI and temperature influenced the malaria endemicity to the greatest extent in the southern part of the country, the same area that was affected by several risk factors (Figure 8). Referring to Figure 9, applying the GWRF model would elim-

Table 3. Performance of three methods employed for the study of malaria incidence.

Level/method	GWR	RF	GWRF
Health catchment level (R^2)	0.32	0.60	0.64
Administrative cell level (R^2)	0.47	0.76	0.79

GWR, geographical weighted regression; RF, random forest; GWRF, geographically weighted random forest; R^2, coefficient of determination.

inate relative humidity, air temperature, and DEM since they were subjected to multi-collinearity, according to its Pearson's correlation coefficient. Interestingly, due to their capability to perform even in high-dimensional settings with complex interactions, these variables were kept in the model and exhibited the strongest relationship with malaria incidence.

Figure 10 reports the PD plots of the total effect of every input variable on the predicted number of malaria cases. The plots combine the main effects of each of the features and their interaction effects to increase malaria incidence. The figure illustrates the inverse association between malaria incidence and rainfall, DEM, relative humidity, and population density. This indicates that the lower the values, the higher the rates of malaria incidence. The higher the air temperature, NDVI, ET, and LST, however, the more malaria cases are recorded.

Discussion

Malaria endemicity in Rwanda is not explained by a single factor but by the combination of different causative factors (Bizimana *et al.*, 2016; Murindahabi *et al.*, 2022). Our findings indicate how the GWRF methodology may be used to model malaria incidence and highlight how the importance of different predictor variables varies over space, with outcomes broadly consistent with previous GWRF models, such as Georganos *et al.* (2020); Quiñones *et al.* (2021) and Maiti *et al.* (2021). However, our findings also reveal a sometimes unbalanced and heterogeneous contribution of the variables considered.

We found rainfall having a negative relationship with malaria incidence, exhibiting a high value of relative contribution in all regions, which is supported by previous work (Colón-González *et al.*, 2016). Rainfall has an ambiguous association with malaria incidence, since moderate amounts provides suitable humid conditions for survival the mosquito vector and also supports egg deposition (Ayanlade *et al.*, 2020), while stagnant pools and open containers with water create ideal vector breeding even during drought (Patz *et al.*, 2003). Heavy rainfall can also have a dual effect, on the one hand forcing mosquitoes to seek refuge in houses increasing the likelihood of vector-human contact, while on the other flushing them out of their aquatic habitat and killing them. Thus, leading to immature vector populations suffering high losses (Cohen *et al.*, 2013).

The land surface temperature has a positive relationship that contributes significantly to the malaria incidence. The positive relationship between land surface temperature and air temperature in this study is logical, as areas with higher LST and air temperature have higher malaria endemicity. This contradicts the lower pattern found in the south-eastern regions of the study area. However, there might be other factors that could be the cause of the malaria incidence in these areas, as proved by Murindahabi *et al.*, (2021). These findings are consistent with previous research. Temperature affects the development rates of mosquitoes and

Table 4. Results of GWRF model with regard to mean decrease accuracy and mean decrease Gini.

%IncMSE Variable	%IncMSE	%	Min	Max	Mean	SD
DEM	29933103	35.59427	2.6595540	34.73443	15.53268	4.359357
Population density	9846901	27.25940	0.8798735	21.83710	10.65293	3.517268
Rainfall	43073223	65.10787	1.6966413	40.94450	20.73048	6.858403
NDVI	16803411	39.36799	1.1588439	22.92965	12.20980	4.031928
LST	44345236	53.36050	2.5162553	27.97506	15.05863	4.243806
Temperature	26491704	31.96840	3.3089317	26.63110	15.64650	3.683335
Relative humidity	29565100	37.14612	3.4132375	32.21923	16.23759	4.216917
Evapotranspiration	15744360	40.33078	1.5956731	24.20495	12.891850	4.604404
IncNodePurity Variable	**IncNodePurity**	**%**	**Min**	**Max**	**Mean**	**SD**
DEM	33730041487	33.97913	1383165.6	5979688415	1356612493	1245518377
Population density	14528763648	15.17221	873506.8	4115033016	795624247	755483114
Rainfall	46585952795	46.22394	1713292.8	7301245129	1826299438	1652785813
NDVI	20142208536	19.88212	933883.2	3995473248	996649402	985164646
LST	36212430303	15.17221	1143877.6	8077355224	1512263984	1671286619
Temperature	28206655736	26.55502	1492961.6	8100629407	1429253678	1443953227
Relative humidity	26184522035	26.88809	1482848.9	6864203394	1289797397	1264673007
Evapotranspiration	19022222429	18.77565	1440003.3	7189294829	1111886205	13098345150
OOB (RSS): 46407530807						
OOB (R^2): 0.7990488						
Predicted RSS: 9072806255						
Predicted R^2: 0.9607135						

%IncMSE, mean decrease accuracy; IncNodePurity, mean decrease Gini; DEM, digital elevation model; NDVI, the normalized difference vegetation index; LST, land surface temperature; SD, standard deviation; R^2=coefficient of determination; RSS, residual sum of squares; OOB, out-of-bag.

malaria parasites (Githeko, 2007). However, the influence of temperature on malaria transmission is not always linear. The non-linear response of the malaria parasite to temperature means that even a slight warming may drive large increases in malaria transmission if other conditions are suitable (Alonso *et al.*, 2011).

Mordecai *et al.* (2013) showed that optimal malaria transmission is at temperature of 25^{0}C, and the transmission substantially decreases when the temperature exceeds 28^{0}C. An increased temperature near the minimum threshold temperature for transmission may result in increased mosquito, duration of the incubation period, and replication of the malaria parasites (Lindsay & Birley, 1996). Temperature may also modify malaria carrying mosquitoes

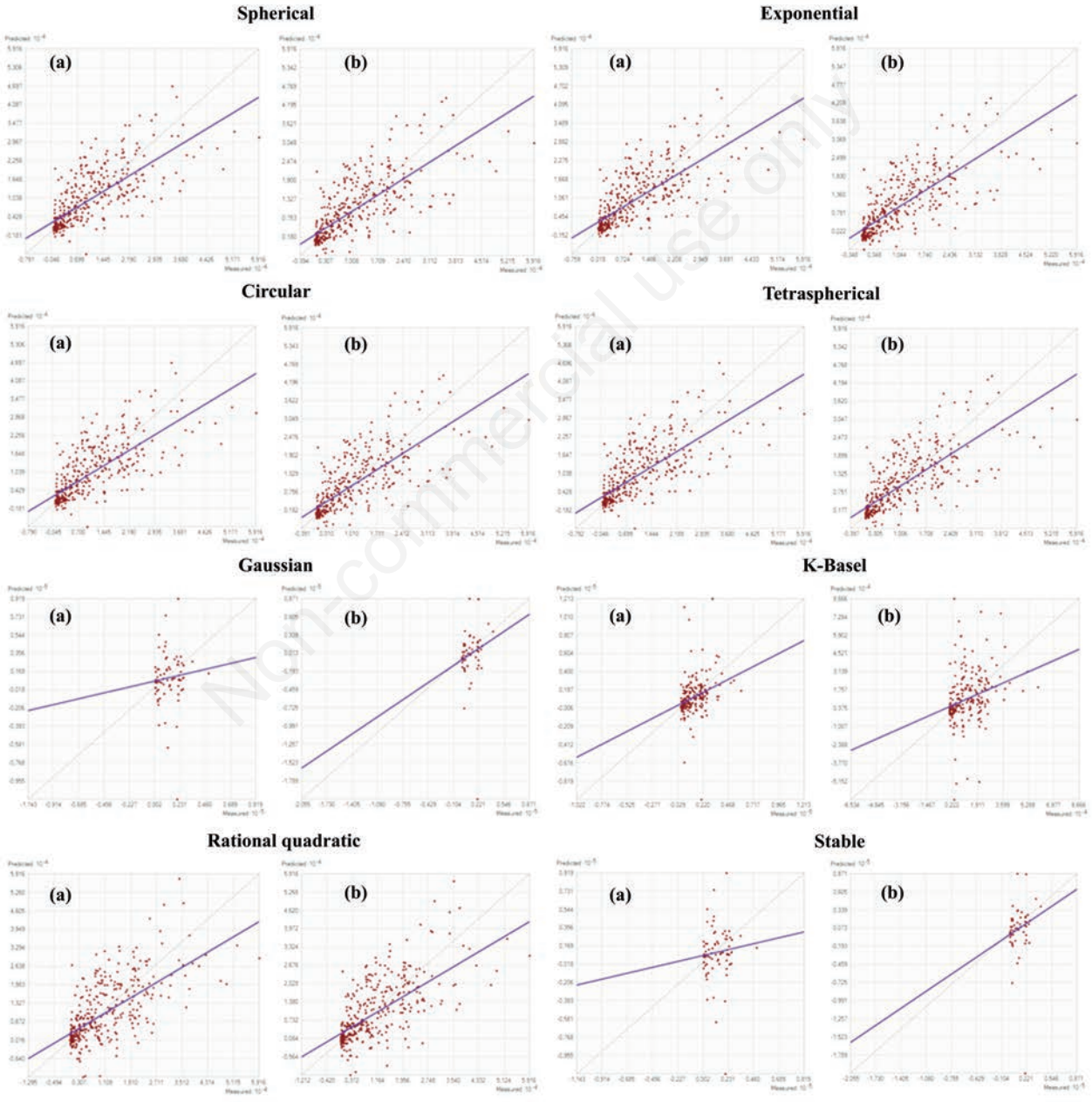

Figure 4. Model scatterplots. a) default values; b) adjusted values.

by changing biting rates and vector's dynamics. A shift in temperature regime can alter the length of the malaria transmission season and change the geographical distribution of the disease (Gubler *et al.*, 2001). Given that current public efforts are targeted at malaria control, the direct effects of warming temperatures are likely to reduce the suitability for malaria transmission (Mordecai *et al.*, 2020). Malaria emergence in the East African highlands was frequently associated with mosquito vectors shifting habitats to adapt to warming temperatures (Chaves & Koenraadt, 2010). Similar studies in Rwanda revealed an epidemiological effect of climate warming near the altitude limits of malaria transmission (Loevinsohn, 1994). As temperature increases, the children become more prone to the two malaria illness (Karekezi *et al.*, 2021). An increase in temperature reduces the time it takes for new generations of mosquitoes to emerge, as well as the parasite's incubation period in mosquitoes (Zhao *et al.*, 2014).

Elevation also has an inverse relationship with malaria incidence. This was consistent with previous studies. For instance, a spatial modeling study revealed that altitude significantly influences the number of malaria cases (Hasyim *et al.*, 2018). Many studies agree that malaria transmission does not occur at altitudes above 2,000 to 2,500 m (latitude dependent) (Bishop & Litch, 2000). The current upper height limit for malaria transmission in the African highlands is difficult to define precisely, and is likely to rise. In many countries, this boundary was thought to occur around 2000 m in Rwanda and Burundi (Meyus *et al.*, 1962), in Ethiopia (Melville *et al.*, 1945), and in Kenya (Garnham, 1945). Malaria epidemics have occasionally been reported at higher altitudes up to 2550 m (Garnham, 1945), but they are rare. In other parts of Africa, the upper limit is slightly lower: around 1700-1800 m in the Democratic Republic of Congo (Schwetz, 1942) and 1200 m in Zimbabwe (Taylor & Mutambu, 1986). Generally, areas higher than 1500 m have little or no malaria (Lindsay & Martens, 1988). The transmission of Plasmodium falciparum generally decreases with increasing elevation, in part because lower temperature slows the development of both parasites and mosquitoes. However, other aspects of the terrain, such as the shape of the land, may affect habitat suitability for Anopheles breeding and thus the

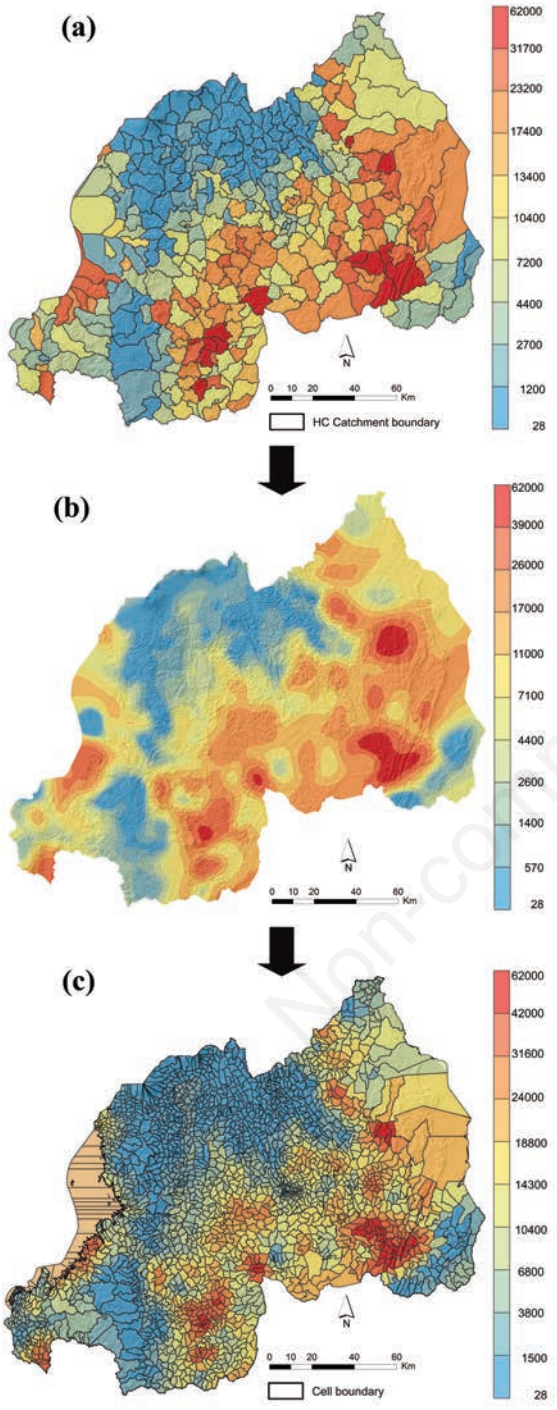

Figure 5. Malaria incidence per year in Rwanda. a) at the health catchment level; b) Gaussian exponential Kriging model; c) disaggregated incidence at the administrative cell level.

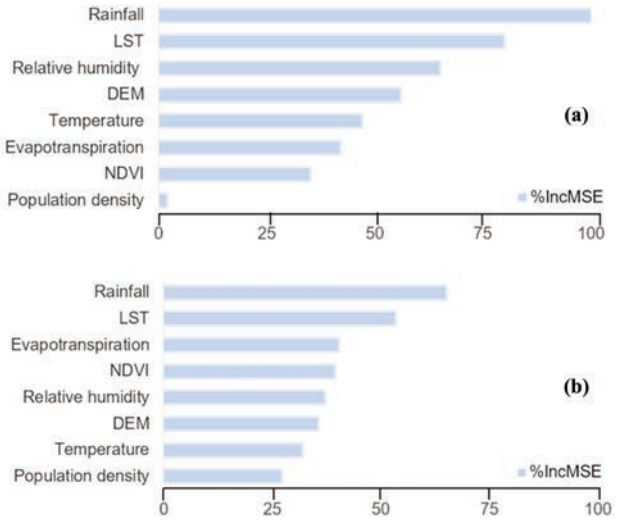

Figure 6. Outcome with regard to the association between malaria incidence and the variables investigated by two regression models as expressed by the mean decrease accuracy. Variable outcomes sorted in decreasing order from top to bottom (the higher the value of the variable importance, the stronger the correlation); (a) GRF (global random forest); (b) GWRF (geographically weighted random forest); %IncMS=mean decrease accuracy.

risk of malaria transmission. Understanding these local topographic effects may permit prediction of regions at high risk of malaria within the highlands at small spatial scales (Cohen *et al.*, 2008).

Humidity is also the most important climatic parameter that determines the number of malaria cases (Nyasa *et al.*, 2022). Relative humidity is the amount of water vapor in the air and is inversely proportional to temperature. It influences malaria transmission by impacting the activity and survival of mosquitoes (Kotepui & Kotepui, 2018). A study by Santos-Vega *et al.*, (2022) on the neglected role of relative humidity in the interannual variability of urban malaria indicated that relative humidity is a critical factor in the spread of urban malaria and potentially other vector-borne epidemics. This study concluded that climate change and a lack of hydrological planning in urban areas might jeopardize malaria elimination efforts. The ambient relative humidity positively influences the life cycle of malaria mosquitoes and results in very frequent biting, leading to a higher risk of malaria (Chirebvu *et al.*, 2016). A mean monthly relative humidity under 60% causes a shortened lifespan in malaria vector mosquitoes, which results in low malaria transmission rates (Pampana, 1969), and a relative humidity of less than 10% is fatal (Yamana & Eltahir, 2013).

Vegetation characteristics provide different opportunities for malaria vectors to thrive. For instance, the crop type determines the breeding and resting place for mosquitoes; the greening of vegetation determines the timing of habitat creation; and deforestation results in sunlit pools suitable for breeding (Beck *et al.*, 2000). The Normalized Difference Vegetation Index (NDVI) determines how much near-infrared light is reflected compared to visible red and helps to evaluate vegetation conditions or to differentiate bare soil from grass or forest (Drisya *et al.*, 2018). An investigation of local environmental variables linked to malaria transmission in Ethiopia revealed that the monthly NDVI (lagged by 1 and 2 months) is significantly correlated with malaria incidence (Kibret *et al.*, 2019). Several studies have shown a positive and significant correlation between NDVI and malaria in West, Central, and East Africa (Gaudart *et al.*, 2009). A study of the relationship between NDVI and malaria mortality in endemic regions of Western Kenya found that the effect of vegetation cover is very consistent in areas with higher risk of malaria mortality with NDVI less than 0.4 and negatively associated with malaria with NDVI greater than 0.4 (Sewe *et al.*, 2016).

The population density variable exhibits a relationship with malaria endemicity, as expected. The probability of malaria endemicity increases with the proportion of population density. Human population density impacts mosquito biting rates, which decrease as the population increases (Hay *et al.*, 2005). The sensitivity of the entomological infection rate (EIR) to population density reveals that as population density increases, the force of infection decreases (Tompkins & Ermert, 2013). Clear and significant differences in EIR exist between urban and rural populations. Thus, low population densities in rural areas and high population densities in urban areas can substantially influence malaria transmission (Tatem *et al.*, 2008). Given the challenges in classifying urban areas across the country, population density provides a reliable metric to adjust for the patterns of malaria risk in densely populated urban areas. Despite a reduction in malaria risk associated with increasing population density, the high-density settlement areas do not have zero risks of infection (Kabaria *et al.*, 2017). Conversely, high density and population pressure in the highlands result in limited land resources and increase human susceptibility to diseases (Bizimana *et al.*, 2015). This is relevant for the Rwanda highlands, where demographic pressure has significantly modified the local environment during the past decades (Bizimana *et al.*, 2016). Additionally, high population density and pressure have significantly influenced environmental degradation and declining land holdings (Clay & Johnson, 1992), which have pushed people to settle near unsuitable sites with more exposure to mosquito bites (Cotter *et al.*, 2013).

The models used provided substantial information that explains the relationship between the drivers and the malaria incidence in Rwanda. But it is also important to emphasize our contributions considering the applied GWRF, which can examine the spatial variations of the non-linear relationships between malaria and the underlying factors. This research is the first local-level

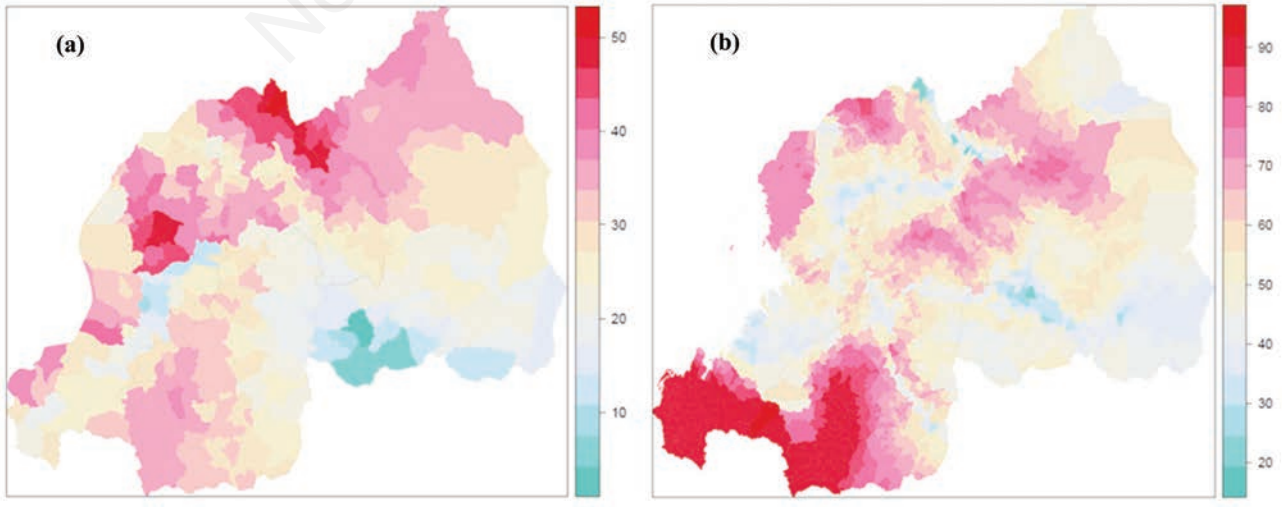

Figure 7. The distribution of the local coefficient of determination when applying the geographically weighted random forest model. a) at the health catchment level; b) at the administrative cell level.

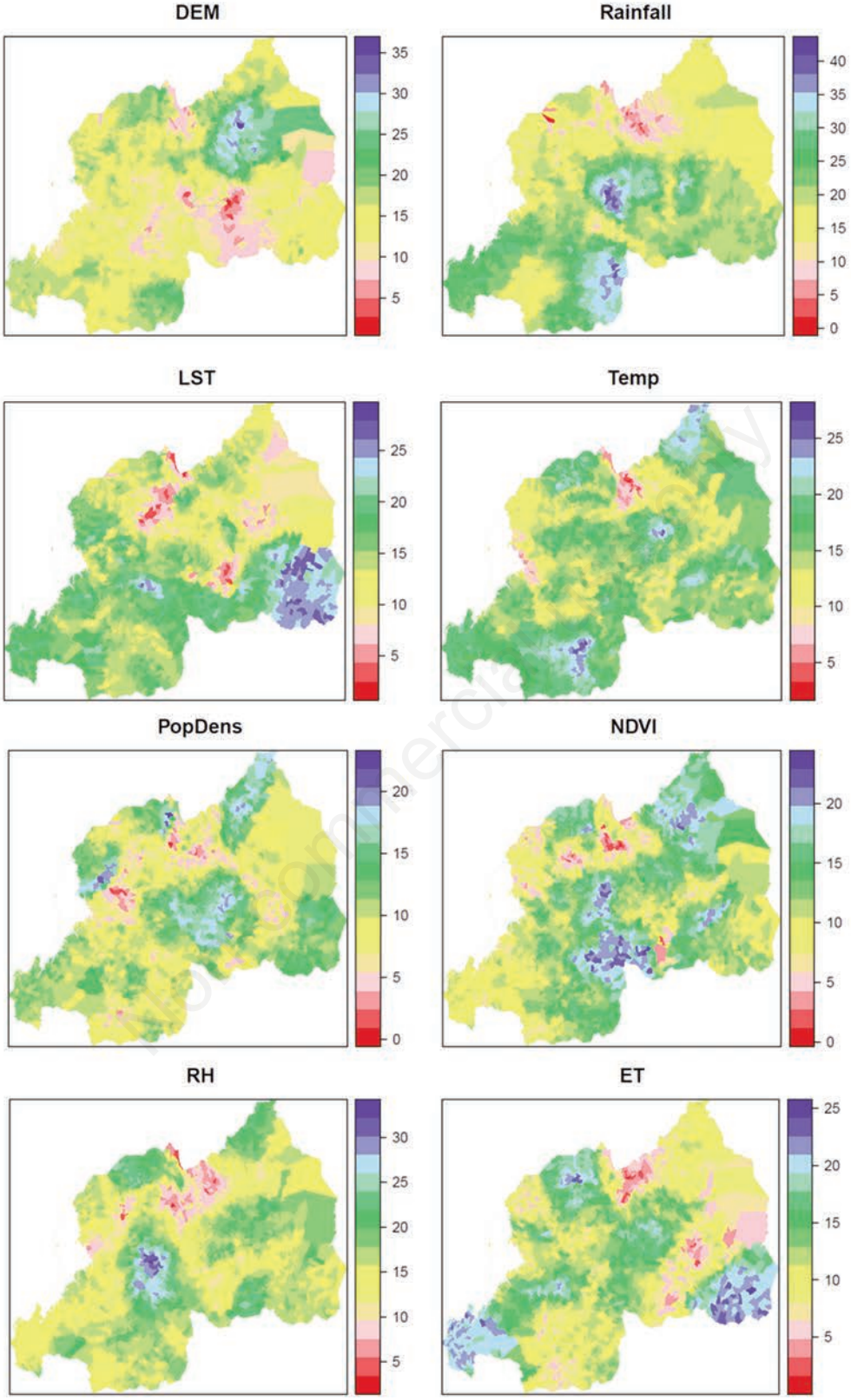

Figure 8. The spatial variation of the local feature importance. Spatial variation in %; local feature importance according to the mean decrease accuracy; DEM, digital elevation model; PopDens, population density: NDVI, normalized difference vegetation index; LST, land surface temperature; Temp, air temperature; RH, relative humidity; ET, evotranspiration.

malaria study implementing the GWRF model in entire Rwanda, which added to our understanding of how the relative importance of variables on the malaria incidence varied with spatial local scale and geographical location. In previous studies conducted in Rwanda, most researchers assumed a heterogeneous effect of predictor variables on malaria incidence (Hammerich *et al.*, 2002; Karekezi *et al.*, 2021; Loevinsohn, 1994). In addition, the evidence from this GWRF model supports the hypothesis that the influence of environmental predictors on malaria transmission is not always linear (Gasana *et al.*, 1996; Harvey *et al.*, 2021). The literature on malaria in Kenya (Cohen *et al.*, 2008), Uganda (Colón-González *et al.*, 2016), Swaziland (Cohen *et al.*, 2013), and Botswana (Chirebvu *et al.*, 2016), the environmental risk factors have a significant correlation with malaria cases that vary at the local level. From a machine learning perspective, the local R^2 depicted shows that the GWRF model still had higher performance in explaining the spatial variations of the non-linear relationships between malaria and the underlying factors when compared with earlier similar studies from eastern Africa (Georganos *et al.*, 2020), the USA (Quiñones *et al.*, 2021; Maiti *et al.*, 2021; Grekousis *et al.*, 2022), and European Union regions (Georganos & Kalogirou, 2022). The capability of the models to analyze spatial datasets at a local scale improves previous results in the study of drivers of malaria incidence in Rwanda by showing the spatial variability of the influence of the drivers. Other factors, however, such as socioeconomic, policy, and political intervention variables that are not used may increase the predictability of the model for further studies. This study employed some coarse resolution remote sensing-based products, such as rainfall variable. Although the machine learning used was capable of handling complex dimensional data, future research could test different downscaling techniques to rescale the coarse resolution input variables at a fine scale to compare the model's predictability.

Conclusions

This study used the GWR, the global random forest, and the GWRF to understand the spatial non-stationarity in the relationships between malaria incidence and ecological risk factors in Rwanda. The predictive ability of the GWRF model in spatial epidemiology has not been investigated for this type of case study characterized by a scarcity of data. The geographical random forest outperforms the GWR and the global random forest model in terms

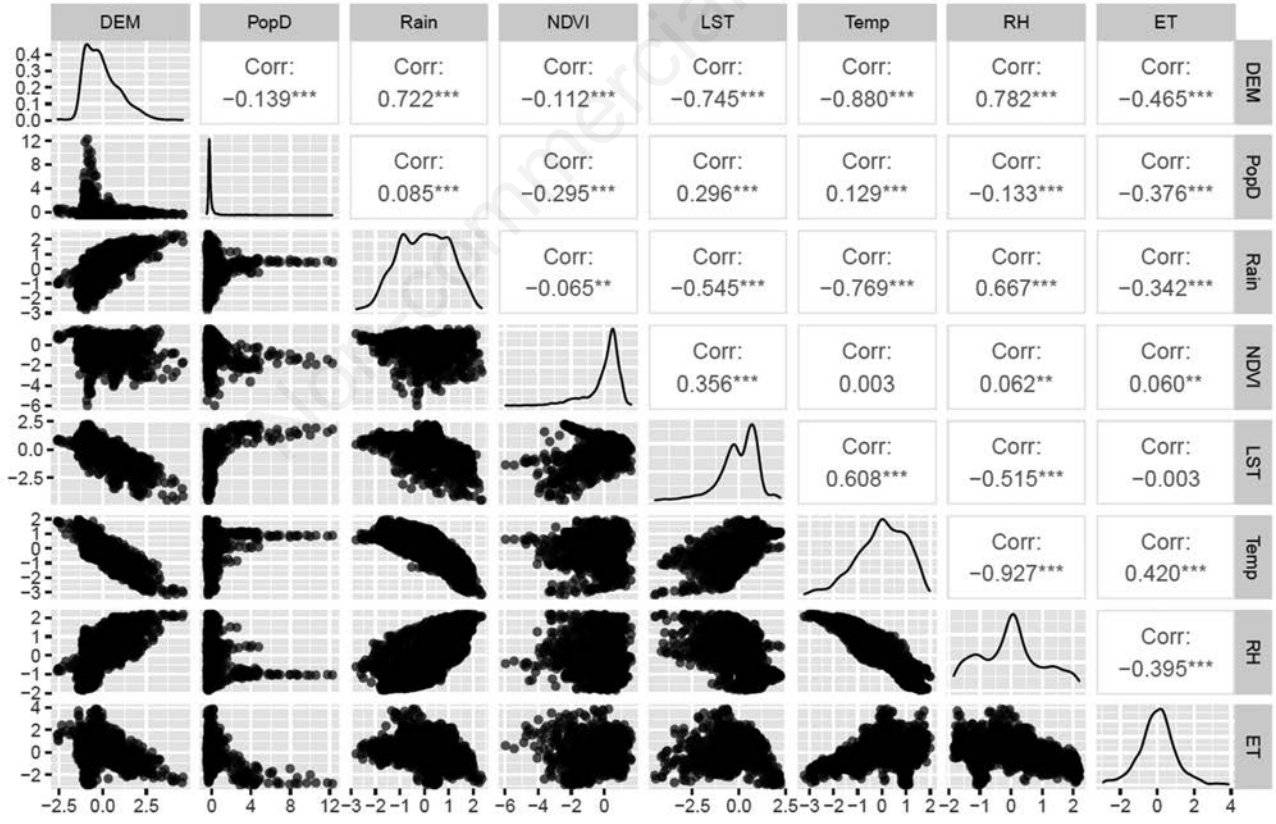

Figure 9. Pair-wise scatterplots of ecological variables for malaria incidence in Rwanda. The boxes along the diagonals display the density plot for each variable. The boxes in the lower left corner display the scatterplot between each variable. The boxes in the upper right corner display the Pearson correlation coefficient between each variable; DEM, digital elevation model; PopD, population density: Rain=rainfall; NDVI, normalized difference vegetation index; LST, land surface temperature; Temp, temperature; RH, relative humidity; ET, evotranspiration.

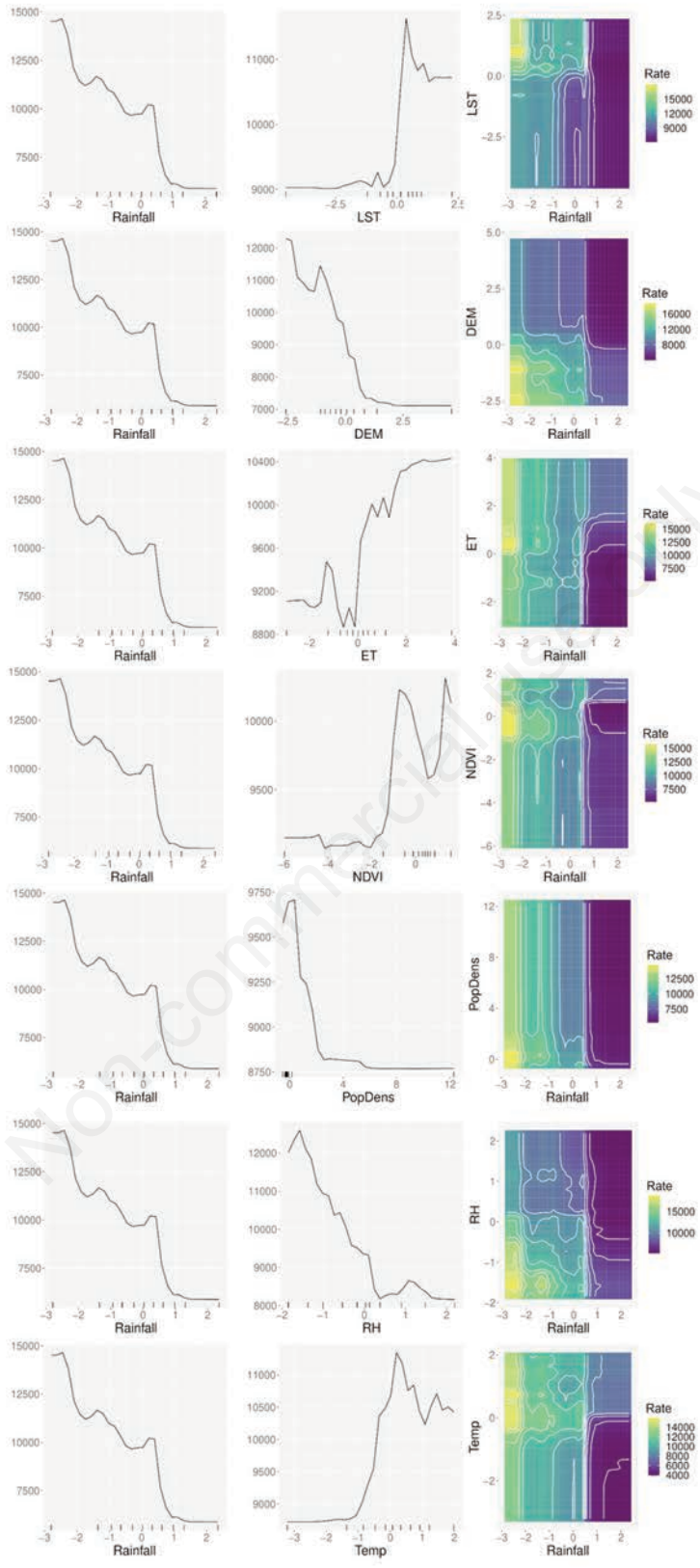

Figure 10. Partial dependence plots of the determinants of the global random forest. The y-axis represents the probability of malaria incidence, and the x-axis gives the probability of the Z-score range values of the predictor variable. Marks on the x-axis indicate the data distribution. The two-variable partial dependence plots for rainfall versus other predictors are shown in the last part of each row.

of the coefficients of determination and prediction accuracy. The key variables importance to the malaria incidence were rainfall, land surface temperature, evapotranspiration, and NDVI. This study adds to our understanding of how the relative importance of variables on the malaria incidence in Rwanda varied with spatial scale and geographical location. Future research should include more time series malaria incidence data as well as more socioeconomic factors to predict future malaria endemicity scenarios in Rwanda.

References

Alonso D, Bouma MJ, Pascual M, 2011. Epidemic malaria and warmer temperatures in recent decades in an East African highland. Proceedings of the Royal Society B 278:1661–9.

Anselin L, Sergio JR, 2014. Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL. GeoDa Press LLC. Available from: https://www.amazon.com/Modern-Spatial-Econometrics-Practice-GeoDaSpace/dp/0986342106

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. Geograph Anal 27:93–115.

Ayanlade A, Nwayor IJ, Sergi C, Ayanlade OS, Di Carlo P, Jeje OD, Jegede MO, 2020. Early warning climate indices for malaria and meningitis in tropical ecological zones. Sci Rep 10:1–13.

Beck LR, Lobitz BM, Wood BL, 2000. Remote sensing and human health: new sensors and new opportunities. Emerg Infect Dis 6:217–27.

Bishop RA, Litch JA, 2000. Malaria at high altitude. J Travel Med 7:157–8.

Bizimana JP, Kienberger S, Hagenlocher M, Twarabamenye E, 2016. Modelling homogeneous regions of social vulnerability to malaria in Rwanda. Geospatial Health 11:129–46.

Bizimana JP, Nduwayezu G, 2021. Spatio-temporal patterns of malaria incidence in Rwanda. Transactions in GIS 25:751–67.

Bizimana JP, Twarabamenye E, Kienberger S, 2015. Assessing the social vulnerability to malaria in Rwanda. Malaria J 14:1–21.

Breiman L, 1996a. Bagging predictors. Machine Learning 26:123–40.

Breiman L, 1996b. Out-of-bag estimation. "Citeseer; 1996."

Breiman L, 2001. Random Forests. Machine Learning 45:5–32.

Brunsdon C, Fotheringham AS, Charlton ME, 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Anal 28:281–98.

Chaves LF, Koenraadt CJM, 2010. Climate change and highland malaria: Fresh air for a hot debate. Quarterly Rev Biol 85:27–55.

Cheng L, Chen X, De Vos J, Lai X, Witlox F, 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav Soc 14:1–10.

Chirebvu E, Chimbari MJ, Ngwenya BN, Sartorius B, 2016. Clinical Malaria Transmission Trends and Its Association with Climatic Variables in Tubu Village, Botswana: A Retrospective Analysis. PLoS One 2016;11:e0139843.

Cianci D, Hartemink N, Ibáñez-Justicia A, 2015. Modelling the potential spatial distribution of mosquito species using three different techniques. Int J Health Geograph 14:1–10.

Clay DC, Johnson NE, 1992. Size of farm or size of family: Which comes first? Population Studies, 46:491–505.

Cohen JM, Dlamini S, Novotny JM, Kandula D, Kunene S, Tatem AJ, 2013. Rapid case-based mapping of seasonal malaria transmission risk for strategic elimination planning in Swaziland. Malaria J 12:61.

Cohen JM, Ernst KC, Lindblade KA, Vulule JM, John CC, Wilson ML, 2008. Topography-derived wetness indices are associated with household-level malaria risk in two communities in the western Kenyan highlands. Malaria J 7:40.

Colón-González FJ, Tompkins AM, Biondi R, Bizimana JP, Namanya DB, 2016. Assessing the effects of air temperature and rainfall on malaria incidence: an epidemiological study across Rwanda and Uganda. Geospatial Health 11:379.

Comber A, Zeng W, 2019. Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. Geography Compass 13:e12465.

Cotter C, Sturrock HJW, Hsiang MS, Liu J, Phillips AA, Hwang J, Gueye CS, Fullman N, Gosling RD, Feachem RGA, 2013. The changing epidemiology of malaria elimination: new strategies for new challenges. Lancet 382:900–11.

Drisya J, Kumar DS, Roshni T, 2018. Spatiotemporal Variability of Soil Moisture and Drought Estimation Using a Distributed Hydrological Model. In Integrating Disaster Science and Management Global Case Studies in Mitigation and Recovery, 2018, p. 451-460

Flowerdew R, Green M, Kehris E, 1991. Using areal interpolation methods in geographic information systems. Papers Regional Sci 70;303–315.

Fotheringham AS, Brunsdon C, Charlton M, 2002. Geographically weighted regression: the analysis of spatially varying relationships. Wiley.

Fotheringham AS, Crespo R, Yao J, 2015. Geographical and Temporal Weighted Regression (GTWR). Geographical Analysis 47:431–52.

Garnham PCC, 1945. Malaria epidemics at exceptionally high altitudes in kenya. Br Med J 2:45–47.

Gasana J, Cailas MD, Brenniman GR, Hallenbeck WH, 1996. Environmental variables involved in the endemicity of malaria in the valley of the Nyabarongo river in Rwanda. Epidemiology 7:78.

Gaudart J, Touré O, Dessay N, Dicko AL, Ranque S, Forest L, Demongeot J, Doumbo OK, 2009. Modelling malaria incidence with environmental dependency in a locality of Sudanese savannah area, Mali. Malaria J 8:9.

Genuer R, Poggi J-M, Tuleau-Malot C, 2010. Variable selection using Random Forests. Pattern Recognition Letters 31:2225–2236.

Georganos S, Brousse O, Dujardin S, Linard C, Casey D, Milliones M, Parmentier B, Van Lipzig NPM, Demuzere M, Grippa T, Vanhuysse S, Mboga N, Andreo V, Snow RW, Lennert M, 2020. Modelling and mapping the intra-urban spatial distribution of Plasmodium falciparum parasite rate using very-high-resolution satellite derived indicators. Int J Health Geograph 19:1–18.

Georganos S, Grippa T, Niang Gadiaga A, Linard C, Lennert M, Vanhuysse S, Mboga N, Wolff E, Kalogirou S, 2019. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int 36:121–36.

Georganos S, Kalogirou S, 2022. A Forest of Forests: A Spatially Weighted and Computationally Efficient Formulation of Geographical Random Forests. ISPRS Internat J Geo-Inform

11:471.
Githeko AK, 2007. Malaria, Climate Change and Possible Impacts on Populations in Africa. International Studies in Population book series (ISIP, volume 6).
Goovaerts P, 2006. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. Int J Health Geograph 5:52.
Grekousis G, Feng Z, Marakakis I, Lu Y, Wang R, 2022. Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach. Health Place 74:102744.
Grömping U, 2009. Variable importance assessment in regression: Linear regression versus random forest. American Statistician 63:308–19.
Gubler DJ, Reiter P, Ebi KL, Yap W, Nasci R, Patz JA, 2001. Climate variability and change in the United States: Potential impacts on vector- and Rodent-Borne diseases. Environ Health Perspect 109:223–233.
Habyarimana F, Ramroop S, 2020. Prevalence and Risk Factors Associated with Malaria among Children Aged Six Months to 14 Years Old in Rwanda. Int J Environ Res Public Health 17:1–13.
Hakizimana E, Karema C, Munyakanage D, Githure J, Mazarati JB, Tongren JE, Takken W, Binagwaho A, Koenraadt CJM, 2018. Spatio-temporal distribution of mosquitoes and risk of malaria infection in Rwanda. Acta Tropica 182:149–57.
Hammerich A, Campbell OMR, Chandramohan D, 2002. Unstable malaria transmission and maternal mortality experiences from Rwanda. Trop Med Int Health 7:573–6.
Harvey D, Valkenburg W, Amara A, 2021. Predicting malaria epidemics in Burkina Faso with machine learning. PLoS ONE 16:0253302.
Hasyim H, Nursafingi A, Haque U, Montag D, Groneberg DA, Dhimal M, Kuch U, Müller R, 2018. Spatial Modeling of Malaria Cases Associated with Environmental Factors in South Sumatra, Indonesia. Malaria J 17:1–15.
Hay SI, Guerra CA, Tatem AJ, Atkinson PM, Snow RW, 2005. Urbanization, malaria transmission and disease burden in Africa. Nature Reviews. Microbiology 3:81–90.
Kabaria CW, Gilbert M, Noor AM, Snow RW, Linard C, 2017. The impact of urbanization and population density on childhood Plasmodium falciparum parasite prevalence rates in Africa. Malaria J 16:1–10.
Kalogirou S, 2003. The statistical analysis and modelling of internal migration flows within England and Wales. Newcastle University.
Kapwata T, Gebreslasie MT, 2016. Random forest variable selection in spatial malaria transmission modelling in Mpumalanga Province, South Africa. Geospatial Health 11:251–262.
Karekezi P, Nzabakiriraho JD, Gayawan E, 2021. Modelling the Shared Risks of Malaria and Anemia in Rwanda. Electronic J https://doi.org/10.2139/SSRN.3986223
Kateera F, Mens PF, Hakizimana E, Ingabire CM, Muragijemariya L, Karinda P, Grobusch MP, Mutesa L, Van Vugt M, 2015. Malaria parasite carriage and risk determinants in a rural population: A malariometric survey in Rwanda. Malaria J 14:1–11.
Kibret S, Glenn Wilson G, Ryder D, Tekie H, Petros B, 2019. Environmental and meteorological factors linked to malaria transmission around large dams at three ecological settings in Ethiopia. Malaria J 18:1–16.
Kotepui M, Kotepui KU, 2018. Impact of weekly climatic variables on weekly malaria incidence throughout Thailand: A country-based six-year retrospective study. J Environ Public Health, 2018:8397815
Krivoruchko K, Gribov A, Krause E, 2011. Multivariate areal interpolation for continuous and count data. Procedia Environ Sci 3:14–19.
Kundrick A, Huang Z, Carran S, Kagoli M, Grais RF, Hurtado N, Ferrari M, 2018. Sub-national variation in measles vaccine coverage and outbreak risk: A case study from a 2010 outbreak in Malawi. BMC Public Health 18:1–10.
Lam NSN, 1983. Spatial Interpolation Methods: A Review. American Cartographer 10:129-149
Lindsay S, Martens W, 1988. Malaria in the African highlands: Past, present and future. Bulletin of the World Health Organization, 76. Available from: https://booksc.org/book/63786001/dbc423
Lindsay SW, Birley MH, 1996. Climate change and malaria transmission. Ann Trop Med Parasitol 90:573–88.
Loevinsohn ME, 1994. Climatic warming and increased malaria incidence in Rwanda. Lancet, 343:714–718.
Macharia PM, Odhiambo JN, Mumo E, Maina A, Giorgi E, Okiro EA, 2022. Approaches to defining health facility catchment areas in sub-Saharan Africa. MedRxiv, 2022.08.18.22278927.
Maiti A, Zhang Q, Sannigrahi S, Pramanik S, Chakraborti S, Cerda A, Pilla F, 2021. Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. Sustainable Cities Soc 68:102784.
McCann RS, Messina JP, MacFarlane DW, Bayoh MN, Vulule JM, Gimnig JE, Walker ED, 2014. Modeling larval malaria vector habitat locations using landscape features and cumulative precipitation measures. Int J Health Geograph 13:1–12.
McMahon A, Mihretie A, Ahmed AA, Lake M, Awoke W, Wimberly MC, 2021. Remote sensing of environmental risk factors for malaria in different geographic contexts. Int J Health Geograph 20:1–15.
Melville A, Wilson DB, Glasgow J, Hocking K, 1945. Malaria in Abyssinia. East Afr Med J 22:285–294.
Meyus H, Caubergh ML, H,1962. L' état actuel du problème du paludisme d'altitude au Ruanda-Urundi. Annales de La Société Belge de Médecine Tropicale 5:771–782.
Midekisa A, Beyene B, Mihretie A, Bayabil E, Wimberly MC, 2015. Seasonal Associations of Climatic Drivers and Malaria in the Highlands of Ethiopia. Parasites and Vectors.
Molnar C, 2022. Interpretable Machine Learning:A Guide for Making Black Box Models Explainable. Independently published.
Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, de Moor E, Mcnally A, Pawar S, Ryan SJ, Smith TC, Lafferty KD, 2013. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecology Letters 16:22–30.
Mordecai EA, Ryan SJ, Caldwell JM, Shah MM, LaBeaud AD, 2020. Climate change could shift disease burden from malaria to arboviruses in Africa. Lancet Plan Health 4:e416–e423.
Murindahabi MM, Hoseni A, Vreugdenhil LCC, Van Vliet AJH, Umupfasoni J, Mutabazi A, Hakizimana E, Poortvliet PM, Mutesa L, Takken W, Koenraadt CJM, 2021. Citizen science for monitoring the spatial and temporal dynamics of malaria vectors in relation to environmental risk factors in Ruhuha, Rwanda. Malaria J 20:453.

Murindahabi MM, Takken W, Hakizimana E, van Vliet AJH, Marijn Poortvliet P, Mutesa L, Koenraadt CJM, 2022. A hand-made trap for malaria mosquito surveillance by citizens in Rwanda. PloS One 17:e0266714

NISR, 2018. Rwandan Integrated Household Living Conditions Survey (EICV5) 2007/2018. Main Indicators Report.

Nyasa RB, Awatboh F, Kwenti TE, Titanji VPK, Lucy N, Ayamba M, 2022. The effect of climatic factors on the number of malaria cases in an inland and a coastal setting from 2011 to 2017 in the equatorial rain forest of Cameroon. BMC Infect Dis 22:1–11.

Ohmer M, Liesch T, Goeppert N, Goldscheider N, 2017. On the optimal selection of interpolation methods for groundwater contouring: an example of propagation of uncertainty regarding inter-aquifer exchange. Adv Water Resour 109:121–132.

Pampana E, 1969. A textbook of malaria eradication (2nd ed.).

Pattnaik A, Mohan D, Tsui A, Chipokosa S, Katengeza H, Ndawala J, Marx MA, 2021. The aggregate effect of implementation strength of family planning programs on modern contraceptive use at the health systems level in rural Malawi. PLoS ONE 16:e0232504.

Patz J, Githeko A, McCarty J, Hussein S, Confalonieri U, Wet NDe, 2003. Climate change and infectious diseases. In Climate change and human health: risks and responses; pp. 103–137.

Peng Y, Li W, Luo X, Li H, 2019. A Geographically and Temporally Weighted Regression Model for Spatial Downscaling of MODIS Land Surface Temperatures over Urban Heterogeneous Regions. IEEE Transact Geosci Remote Sensing 57:5012–27.

Quiñones S, Goyal A, Ahmed ZU, 2021. Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. Sci Rep 11:1–13.

Rhodes CG, Loaiza JR, Romero LM, Alvarado JMG, Delgado G, Salas OR, Rojas MR, Aguilar-Avendaño C, Maynes E, Cordero JAV, Mora AS, Rigg CA, Zardkoohi A, Prado M, Friberg, MD, Bergmann LR, Rodríguez RM, Hamer GL, Chaves LF, 2022. Anopheles albimanus (Diptera: Culicidae) Ensemble Distribution Modeling: Applications for Malaria Elimination. Insects 13:13030221

Rosenshein L, 2010. The Local Nature of a National Epidemic: Childhood Overweight and the Accessibility of Healthy Food. Masters dissertation. George Mason University, Fairfax, Virginia, USA.

Rudasingwa G, Cho Sil, 2020. Determinants of the persistence of malaria in Rwanda. Malaria J 19:1–9.

Rulisa S, Kateera F, Bizimana JP, Agaba S, Dukuzumuremyi J, Baas L, de Dieu Harelimana J, Mens PF, Boer KR, de Vries PJ, 2013. Malaria prevalence, spatial clustering and risk factors in a low endemic area of Eastern Rwanda: a cross sectional study. PloS One 8:e0069443

Santos-Vega M, Martinez PP, Vaishnav KG, Kohli V, Desai V, Bouma MJ, Pascual M, 2022. The neglected role of relative humidity in the interannual variability of urban malaria in Indian cities. Nature Communications 13:1–9.

Schwetz J, 1942. Recherches sur la limite altimetrique du paludisme dans le Congo Orientale et sur la cause de cette limite. Annales de La Societe Belge de Medecine Tropicale, 183–209.

Semakula M, Niragire F, Faes C, 2020. Bayesian spatio-temporal modeling of malaria risk in Rwanda. PLoS One 15:e0238504.

Sewe MO, Ahlm C, Rocklöv J, 2016. Remotely Sensed Environmental Conditions and Malaria Mortality in Three Malaria Endemic Regions in Western Kenya. PLoS ONE 11:e0154204.

Stresman GH, Stevenson JC, Owaga C, Marube E, Anyango C, Drakeley C, Bousema T, Cox J, 2014. Validation of three geolocation strategies for health-facility attendees for research and public health surveillance in a rural setting in western Kenya. Epidemiol Infection 142:1978.

Sullivan W, 2017. Machine Learning For Beginners: Algorithms, Decision Tree & Random Forest Introduction. https://1lib.sk/book/3428860/bec3d9

Tatem AJ, Guerra CA, Kabaria CW, Noor AM, Hay SI, 2008. Human population, urban settlement patterns and their impact on Plasmodium falciparum malaria endemicity. Malaria J 7:9

Taylor P, Mutambu SL, 1986. A review of the malaria situation in Zimbabwe with special reference to the period 1972-1981. Transactions of the Royal Society of Tropical Medicine and Hygiene, 80:12–19.

Tompkins AM, Ermert V, 2013. A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology. Malaria J 12:13

Wang M, Wang H, Wang J, Liu H, Lu R, Duan T, Gong X, Feng S, Liu Y, Cui Z, Li C, Ma J, 2019. A novel model for malaria prediction based on ensemble algorithms. PLoS ONE 14:e0226910.

Yamana TK, Eltahir EAB, 2013. Incorporating the effects of humidity in a mechanistic model of Anopheles gambiae mosquito population dynamics in the Sahel region of Africa. Parasites Vectors 6:1–10.

Zeng W, Comber A, 2020. Using household counts as ancillary information for areal interpolation of population: Comparing formal and informal, online data sources. Computers, Environment and Urban Systems 80:101440.

Zhao X, Chen F, Feng Z, Li X, Zhou XH, 2014. Characterizing the effect of temperature fluctuation on the incidence of malaria: An epidemiological study in south-west China using the varying coefficient distributed lag non-linear model. Malaria J 13:192.