UC Irvine

UC Irvine Previously Published Works

Title

Combining global and local estimates for spatial distribution of mosquito larval habitats

Permalink

<https://escholarship.org/uc/item/6qd3x096>

Journal

GIScience and Remote Sensing, 43(2)

ISSN

1548-1603

Authors

Bian, L
Li, L
Yan, G

Publication Date

2006-04-01

DOI

10.2747/1548-1603.43.2.128

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at

<https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Combining Global and Local Estimates for Spatial Distribution of Mosquito Larval Habitats

Ling Bian¹ and Li Li

*Department of Geography, State University of New York at Buffalo
Amherst, New York 14261*

Guiyun Yan

*Program in Public Health, College of Health Sciences
University of California, Irvine, CA 92697-4050*

Abstract: Malaria, transmitted by mosquitoes, is the leading cause of mortality in many African countries. This study assesses the spatial distribution of mosquito larval habitats on the highlands of western Kenya. Environmental factors such as terrain, land use, and surface waters are evaluated for their influence on the habitats. Logistic regression is used to model the global trend in the variation of larval habitats, while kriging is used to model the spatially dependent local variation. Results show notable improvement in modeling accuracy by combining the regression and kriging estimates over using the regression estimate alone.

INTRODUCTION

Malaria epidemics are the leading cause of mortality in many African countries (Kleinschmidt et al., 2000; Hay et al., 2002; Zhou et al., 2004). Malaria is a vector-borne disease, and mosquitoes are the major vector that transmits virulent malarial parasites to human populations (Minakawa et al., 2002; Zhou et al., 2004). In Africa, malaria epidemics occur most often in lowland areas because climatic conditions are most favorable for malaria parasites and mosquito vectors. Because of regular exposure to malaria, human populations in these lowland areas develop a certain degree of immunity against the infection. Since 1988, a series of malaria outbreaks has occurred almost annually in the highland regions in East Africa that only had infrequent epidemics between the 1920s and the 1950s (Zhou et al., 2004). Unlike people who live in malaria-endemic areas, the highland populations generally lack immunity and are vulnerable to the infection and its severe consequences. Significant human mortality has been reported in these regions (Hay et al., 2002).

Climatic conditions are one of the factors crucial for the development and survival of malaria parasites and mosquito vectors. Precipitation and temperature particularly influence the spatial distribution of mosquito larval habitats and, subsequently,

¹lbian@geog.buffalo.edu

the mosquito population (Kleinschmidt et al., 2000; Minakawa et al., 2002; Zhou et al., 2004). Aquatic environments are another critical factor. Mosquito larvae prefer sunlit permanent water bodies with aquatic vegetation or ephemeral still waters (Minakawa et al., 1999, 2002).

Research on the reemergence of malaria in the highland regions has been primarily based on climatic variables (Kleinschmidt et al., 2000; Minakawa et al., 2002; Mbogo et al., 2003; Zhou et al., 2004). Spatially, most of these studies focus on site-specific or aggregated regional analysis. It is widely recognized that environmental factors other than climate are equally, if not more, important and should not be ignored. In addition, the spatial heterogeneity of these factors, including climate, is critical in identifying the spatial distribution of larval habitats (Kleinschmidt et al., 2000; Minakawa et al., 2002; Mbogo et al., 2003; Zhou et al., 2004). Few studies have explicitly taken these environmental factors and their spatial variation into consideration.

On the highlands, microclimates may vary considerably within a small area, due to varying terrain and land use, among other possible factors. Climatic data, on the other hand, are usually not available at a local spatial scale necessary to assess larval habitats. Spatial interpolation of sparse climatic data using generic methods cannot effectively depict these local variations (Zhou et al., 2004). Other environmental factors, such as terrain, land use, and surface waters need to be explored. This exploration is increasingly feasible as data recording these factors become available at the spatial scale appropriate for assessing local habitats.

Spatial assessment of habitats usually relies on data collected at point locations for the habitats and the associated explanatory environmental factors. Regression is a common approach to establishing the statistical relationship between them. The relationships obtained from these point locations are then extended to the entire study area. Like many statistical methods, regression is a global estimator, assuming that one relationship fits for all locations in an area. Local variations, especially those affected by surrounding locations, are usually considered as residuals of the model and ignored (Kleinschmidt et al., 2000; Miller, 2005). This approach leaves a portion of the total variation in the data unaccounted for.

This study attempts to assess spatial suitability of mosquito larval habitats on the highlands of western Kenya. Environmental factors, such as terrain, land use, and surface waters are evaluated for their influence on these habitats. A regression method is used to account for the global trend and a kriging method is used to explicitly account for the spatially dependent local variation in the distribution of larval habitats.

An understanding of the spatial distribution of mosquito larval habitats is valuable for forecasting potential areas of malaria outbreaks. Management plans can be devised to control larval habitats, especially in the most critical areas. Results from this study can be readily extended to large areas in the highland regions of Africa to eradicate and prevent malaria epidemics.

STUDY AREA AND DATA

The study area, located in Iguhu village, Kakamega district, in western Kenya, covers a 4×4 km area centered at latitude $0^{\circ}10'$ N and longitude $34^{\circ}45'$ W (Fig. 1). The region is characteristic of a tropical climate with a dry and a rainy season. The

Fig. 1. The 4×4 km study area in western Kenya.

mean annual daily temperature is 20.3°C and the mean annual precipitation of 1,980 mm. The terrain of the region is typical of an old plateau, consisting of high peaks, hills, valleys, and small basins. The altitude ranges between 1,420 and 1,540 m, demonstrating considerable relief within the 4×4 km study area. The Yala River and its tributaries flow through the landscape from east to west.

The region used to be heavily forested. In recent decades, a significant proportion of the forests, especially those near the village, were cleared and replaced by agriculture lands, due to the demands of a rapidly expanding population. The primary land use types in the study area include farmlands, forests, shrubs, pastures, swamps, streams, and roads. In this environment, *Anopheles gambiae* is the dominant species of the mosquito vector (Minakawa et al., 2002; Zhou et al., 2004).

Extensive data have been collected in this study area in the past several years. For the research presented in this paper, data describing four themes are used to support the intended analysis. These themes are: the presence of mosquito larvae, terrain, land use, and surface waters. The presence of larvae is treated as the dependent variable in the subsequent regression analysis. The latter three themes represent environmental factors and are used as independent variables in the analysis, in order to explain the variation in the presence of mosquito larvae.

Data for the presence of larvae were obtained from thorough field surveys in May 2003 and May 2004 during the rainy season. Mosquito larvae were found in 1,229 locations (hereafter called the “presence” locations). Coordinates of these locations were recorded using global positioning system devices. In addition to the 1,229 presence locations, another 1,229 locations are randomly sampled in the study area to represent the absence of larvae (hereafter called the “absence” locations). The 2,458

locations are prepared as point data and the presence is coded as 1 and absence as 0 for the analysis.

To represent terrain, three variables—elevation, aspect N-S, and aspect E-W—are obtained. Elevation, presented in meters, is used as a surrogate variable for climatic conditions, because temperature in general decreases, whereas precipitation increases, with elevation (Minikawa et al., 2002). Aspect is a substitute for solar radiation and associated moisture conditions (Hodgson and Gaile, 1999). Cosine values of aspect are used to represent the N-S variation and sine values for the E-W variation. For computing these terrain variables, a contour map of 1:50,000 scale and 20 m contour interval is initially used to construct a digital elevation model with a 30 m spatial resolution. This resolution is to capture the density of contour lines in the study area. Aspect is then extracted from the digital elevation model at the same spatial resolution.

The land use information is obtained from a GIS layer that was originally generated from an IKONOS image acquired in April 2002. The land use types, treated as nominal data, were identified through manual interpretation and subsequently digitized as polygons. Seven land use types were identified: forests, shrubs, swamps, farmlands, pastures, roads, and river. The polygon layer is converted into a grid with a 30 m spatial resolution. Note that both words “land use” and “land cover” can be used to describe some or all of these types depending on the perspective of the research. The term “land use” is deemed most appropriate for this study area that has experienced considerable human-induced changes.

Surface waters in the study area are represented by three variables. These are a wetness index, slope curvature, and distance to the nearest surface waters. The wetness index is the ratio of the slope angle of a cell to the area draining into the cell. This index is a modification of the widely used wetness index $\ln(A/\tan B)$ (Beven and Kirkby, 1979), where A is the draining area and B is the slope. The modification is intended to avoid dividing zero when computing wetness for flat areas, which occupy a proportion of the study area. This slope/area index represents topographically driven soil moisture conditions that decrease with the slope angle of a location and increase with the area draining into the location. The second surface water variable, slope curvature, is defined as the change rate of slope over a unit of distance (Band, 1996). This variable is used to detect concave terrain features for possible accumulation of ephemeral waters in the study area. These two variables are derived from the digital elevation model. The third surface water variable, distance to the nearest waters, measures the distance-dependent influence of surface waters. This variable is derived using the stream data from the field surveys. The stream data include a portion of the Yala River and its tributaries.

The aforementioned three sets of a total of seven environmental variables are prepared as spatial data layers. Of these, the wetness index is derived using a freeware, TauDEM (Tarboton, 2004), and subsequently brought into ArcGIS (Environmental Systems Research Institute) to join the other six variables that are prepared in ArcGIS. All of the spatial data are prepared in or transformed to UTM coordinates with a spatial resolution of 30 m. Values of the seven environmental layers are extracted at the 1,229 presence locations and 1,229 absence locations for subsequent statistical analysis.

ANALYSIS DESIGN

Spatial variation of environmental phenomena can be expressed as consisting of three components: a global trend, a spatially dependent local variation, and random noise (Burrough, 1986; Davis, 2002). Conventional statistical approaches, such as various regression methods, model the global trend by using a single formulation to fit the entire study area. Residuals of the fitting are expected to contain the random noise and variations that might have been explained by variables not included in the regression model. This conventional approach estimates the value of a variable at a location by focusing on the explanatory factors at the location, independent of surrounding locations. As a global estimator, conventional statistical models may not effectively account for local variations, especially those that are dependent on values at adjacent locations.

Geostatistical approaches, such as various kriging methods, are designed to model the spatially dependent local component. This approach describes the spatial dependence in the sample data through a semivariogram model. This model is then applied locally to account for the spatially dependent local variation. When estimating the value of a variable at a location, this approach does not look into explanatory factors at the location. Rather, it uses values of the variable at adjacent locations for the estimation.

Established methods exist, such as universal kriging, to simultaneously model the global trend and spatially dependent local component. It is common, however, to model the two components separately. This separate approach first models the global trend and then the spatially dependent local variation in the residuals. The two components are then combined to describe the spatial variation of the phenomena. This separate approach has been applied to modeling soils (Bishop and McBratney, 2001), vegetation (Miller, 2005), and epidemiology (Kleinschmidt et al., 2000).

This study explores the separate approach and performs the statistical analysis in three steps. The first uses logistic regression to establish a statistical relationship between the presence of larvae and the environmental factors at sampled locations. This step is designed to model the global trend in the presence of larvae. In the second step, a geostatistical method, ordinary kriging, is used to model the spatially dependent local variation in the residuals of the logistic model. The third step combines the two sets of estimates to predict the spatial variation of larvae habitats. A flow chart is displayed in Figure 2 to illustrate the processes involved in this study and the sequence in which these processes are executed.

Out of the concern that spatial dependence, common to most spatial data, may affect the accuracy of regression model, two data sets are used for the statistical analysis. One consists of the full 2,458 locations (hereafter called the “full” set). The other set is re-sampled from the full data set, so that the spatial dependence is minimized between locations (hereafter called the “minimized” set). Results of statistical analysis using the two data sets are later compared.

In order to create the minimized set, the full data set is re-sampled by eliminating data locations that fall within a certain distance from surrounding data locations. Moran's I analysis is used to identify the critical distance. For each independent variable layer, Moran's I coefficient is computed at spatial lags from 30 m to 1,980 m, with a 30 m increment. At the lag of 120 m, the Moran's I values for aspect N-S,

Fig. 2. A flow chart illustrating the processes involved in this study and the sequence in which these processes are executed.

aspect E-W, wetness, and slope curvature layers are reduced to 0.1 or lower. Moran's I analysis is also applied to the nominal data layer, land use, where the land use types are treated as indicator variables. Results show that at the 120 m lag, the Moran's I values for all land use types are reduced to 0.2 or lower. The distance of 120 m is selected to re-sample the full data set. Both elevation and distance to the nearest waters, as expected, show strong spatial dependence. Their Moran's I values are not considered for re-sampling. This is because these data layers cannot be re-sampled to a minimal level of spatial dependence, while still yielding adequate data locations for the intended analysis. The re-sampling produces a total of 618 locations (309 presence and 309 absence) for the minimized data set.

Descriptive statistical tests are applied to each independent variable in order to evaluate whether their values differ between larval presence and absence locations. Because the independent variables are selected empirically, such a screening procedure is preferred before model development (Pereira and Itami, 1991). A t -test and its non-parametric equivalent Mann-Whitney test are applied to the six continuous variables (i.e. elevation, aspect N-S, aspect E-W, wetness, slope curvature, and distance to the nearest waters). The t -test is for variables with a normal distribution and Mann-Whitney test otherwise. All six variables are statistically significant in distinguishing between the larval presence and absence locations. The larval presence locations appear to be associated with lower elevations, more north-facing and west-facing slopes, greater wetness, more concave features, and shorter distance to the nearest waters than absent locations. The nominal variable, land use, is also statistically significant, using a chi-square test, with the availability of each category in the study area adjusted.

For the purpose of model validation, two-thirds of the data are randomly selected to develop the logistic and kriging models, and the remaining third is used for validating the models. This data partition procedure results in a total of 1,638 locations (819 presence and 819 absence) for the full data set and 412 locations (206 presence and 206 absence) for the minimized data set for the model development. The remaining 820 locations from the full data set (410 presence and 410 absence) and 206 from the minimized set (103 presence and 103 absence) are used for model validation.

LOGISTIC MODELS

Logistic regression is commonly used to assess potential habitats for various organisms (Pereira and Itami, 1991; Mladenoff et al., 1995; Bian and West, 1997; Klienschmidt et al., 2000; Miller, 2005). Logistic regression relates a dichotomous dependent variable to a number of independent variables that can take numeric or nominal forms. Results of the regression predict the probability, ranging from 0 to 1, of the presence of habitats. The probability responds to the change in the independent variables as a logistic curve. That is, the probability changes more rapidly in response to the change in a certain range of the independent variable values than to changes of other values. This method is appropriate to model the observation that larval habitats tend to be sensitive to a certain range of environmental conditions.

A multiple logistic regression is applied to the full and the minimized data sets using a stepwise procedure. The intention is to further screen the independent variables according to the contribution of each to the overall goodness-of-fit. For the full data set, the logistic regression model includes four independent variables, distance to the nearest waters, elevation, land use, and wetness (hereafter called the "full" model). For the minimized data set, the logistic regression includes three variables, distance to the nearest waters, elevation, and land use (hereafter called the "minimized" model). Table 1 lists the estimated parameters for both the full and the minimized models. Both models suggest that larvae might prefer a close distance to the nearest waters, a low elevation, and the land use type of swamps. The fourth independent variable, wetness, included only in the full model, suggests that larvae may prefer high values of wetness.

The modeled probability of larval habitats varies continuously from 0 to 1. A threshold value of 0.5 is used as the cut-off point to classify those locations with probability ≥ 0.5 as suitable larval habitats, and probability < 0.5 as unsuitable. The estimated suitability is compared against the original larval presence and absence data (the two-thirds of the data sets) that are used to develop the logistic models. For the full data set, 88.2% of the original larval presence and 83.3% of the absence locations are correctly estimated, respectively (Table 2). For the minimized data set, 82.0% of presence and 78.6% of the absence locations are correct, respectively (Table 2).

Both models are further validated using the one-third of data (820 for the full set and 206 for the minimized set) that are not involved in the model development. The validation accuracy is 89% and 91.7% for the presence and absence locations, respectively, for the full logistic model, and 85.4% and 87.4% for presence and absence locations, respectively, for the minimized model (Table 2). In order to compare the validation directly between the two models using the same validation data set, the 820-location data set is also used to validate the minimized model, in addition to its

Table 1. Independent Variables Included in the Full Logistic Regression Model and the Minimized Regression Model, and Associated Coefficients (*B*) and Significance (Sig)^a

Variables	Full		Minimized	
	<i>B</i>	Sig.	<i>B</i>	Sig.
Distance	- 0.011	0.000	-0.008	0.000
Elevation	- 0.031	0.000	-0.025	0.000
Farmland	-20.162	0.000	-19.893	0.001
Forest	-19.019	0.000	-19.439	0.001
Pasture	-19.254	0.000	-20.788	0.001
Shrub	-18.793	0.000	-19.842	0.001
Wetness	0.000003	0.000		
Constant	66.413	0.993	58.032	0.998

^a“Distance” refers to the independent variable Distance to Nearest Waters. The coefficients (*B*) of the four land use types Farmland, Forest, Pasture, and Shrub, are referenced by the land use type Swamp.

Table 2. Accuracy (%) of Logistic Regression Models (Logi) and Combined Models (Com) for the Full and Minimized Models^a

	Model				Validation (820)				Validation (206)			
	Full		Minimized		Full		Minimized		Full		Minimized	
	P	A	P	A	P	A	P	A	P	A	P	A
Logi	88.2	83.3	82.0	78.6	89.0	91.7	90.2	86.6	84.5	93.2	85.4	87.4
Com					96.1	86.3	95.6	86.3	91.3	88.3	89.3	87.4

^aValidation (820) refers to validation using the 820 locations, and Validation (206) refers to validation using 206 locations. P = presence; A = absence.

use to validate the full model. Similarly, the 206-location set is also used to validate the full model, in addition to validating the minimized model. This treatment is also intended for the model comparison as discussed in a later section. Results of these two validations are listed in Table 2.

The two logistic models appear to be similar. Both include three identical independent variables. The estimation accuracy is higher for the full model, with a margin, than the minimized model. This is possibly because the full model includes one additional independent variable, although whose contribution to the overall goodness-of-fit is modest (0.006 out of 0.448 in Cox and Snell *R* square). In the validation results using either the 820- or 206-location data sets, the better performance of the full model is less apparent.

Incorrect estimations are examined for the four data sets. These include: (1) the 1,638-location data set estimated by the full model; (2) the 820-location set for the validation of the full model; (3) the 412-location set estimated by the minimized

model: and (4) the 206-location set for the validation of the minimized model. The underestimated locations, i.e. the actual presence locations with $p < 0.5$, show a certain spatial pattern. A significant portion of these locations seems to be distributed along linear forest or shrub strips, where streams most likely exist but are missing from the stream data layer available for this study. This error occurs in all four data sets. For the overestimated locations that are actual absence locations with $p \geq 0.5$, no obvious spatial patterns are observed.

The initial intention to develop two models, i.e., the full and the minimized model, is to examine the impact of spatial dependence on logistic modeling. Results of the two models do not seem to suggest a significant difference between them, in terms of the independent variables included, performance of the models, and spatial pattern of the mis-estimations. This is possibly because, of the three independent variables included in both models, two of them (elevation and distance to the nearest waters) contain prominent spatial dependence (see the section on Analysis Design). Despite this, the subsequent kriging analysis continues to treat these two models separately in order to evaluate whether the residuals of the two models differ.

KRIGING MODELS

Kriging methods, designed for spatial interpolation, are built on the assumption of spatial dependence (Isaaks and Srivastava, 1989). The variation of a variable between two locations is a function of the distance between them, not where they are. Kriging methods express this distance-dependent function using a semivariogram. According to the semivariogram model, kriging methods estimate the value at an unsampled location using values of adjacent samples and distances from the unsampled location to these samples. Kriging methods have been used in many health-related studies (Walter, 1994; Diggle et al., 1998; Klienschmidt et al., 2000; Sakai et al., 2004). Ordinary kriging is used in this study to model the spatially dependent local variation in the residuals.

The modeled probability values, estimated by the full and the minimized models, respectively, are subtracted from all 2,458 data locations to obtain two sets of residuals. One set of 2,458 residuals is for the full model and the other for the minimized model. For each set of residuals, those at the 1,638 locations (whose original values are used to develop the full logistic model) are used to establish the kriging models. Residuals at the remaining 820 locations are reserved for a later use (detailed in the section on Combined Models below). Semivariograms are constructed for the two sets of 1,638 residuals. Both semivariograms are fitted by spherical models, one common type of model for describing semivariograms.

Both semivariograms show a notable spatial dependence in the residuals. Of the three components of spatial variation (a global trend, a spatially dependent local variation, and random noise), the logistic models might have adequately accounted for the global trend, but leaving the other two components in their residuals. Explicitly modeling the spatially dependent local component, using kriging, helps reduce the remaining residuals to the level close to random noise, which is considered ideal in model performance (Davis, 2002). The two semivariograms depict a similar degree of spatial dependence, perhaps due to the similar performance of the two logistic models, as explained previously. The one additional independent variable included in

Table 3. Results of Cross Validation for Kriging the Residuals for the Full and Minimized Logistic Model, Respectively, Residuals Using the 1,638 Locations

	Cross-Validation (1,638)	
	Full	Minimized
Mean error	0.0075	0.0080
Root mean square error	0.2201	0.2206
Average standard error	0.1755	0.1802

the full model, wetness, has minimal spatial dependence (maximum Moran's $I = 0.25$ at lag = 30 m), thus having little impact on the spatial dependence in the residuals. Both semivariograms show the spatial dependence at a range of approximately 120 m. This reflects the spatial dependence in some of the independent variables, such as land use in both the full and minimized logistic models and wetness in the full model, as explained previously (see the section on Analysis Design).

Using the semivariograms, the ordinary kriging method is used to model the spatially dependent local variation in the residuals. This yields two models, a full kriging model and a minimized kriging model. The accuracy of the two models is evaluated using a cross validation procedure. Results are listed in Table 3. The low values of the error measurements indicate a reasonable model fit.

COMBINED MODELS

The full logistic model and minimized logistic model are each combined with its corresponding kriging model to produce two new models. One is the combined full model and the other is the combined minimized model. The combination is conducted by adding the kriging estimations to the logistic estimations, for the full and the minimized models, respectively. Note that kriging methods retain, rather than estimate, residual values at the 1,638 locations that are used to establish the kriging models. Adding these 1,638 residuals to the logistic estimations reverts to their original presence or absence. In order to evaluate the accuracy of the combined models, the 820-location data set is used for the validation. At these locations, the estimations of the spatially dependent local variation in the residuals are available and can be combined with the logistic estimations for the intended validation.

Using the same cutoff value of 0.5, the full combined model correctly estimates 96.1% and 86.3% of the presence and absence locations, respectively (Table 2). The minimized combined model correctly estimates 95.6% and 86.3% of the presence and absence locations, respectively (Table 2). Data at the 206 locations (used to validate the minimized logistic model) are also used for the validation. The combined models appear to improve the accuracy for the presence locations, but are less accurate for the absence locations. In this situation, the accuracy in presence is preferred to that in absence because the presence locations are from actual observations whereas the absence locations are arbitrarily generated.

The spatial pattern of combined models shows a remarkable improvement along the linear strips of forest or shrub where the actual presence is underestimated by the logistic models. This improvement occurs for both the combined full and combined minimized models. Figure 3 demonstrates this improvement by comparing the mis-estimated locations between the minimized logistic model and the combined minimized model. By adding the residuals estimated by the kriging models, the combined models virtually eliminate the estimation errors at these locations.

DISCUSSION AND CONCLUSIONS

This study estimates spatial distribution of larval habitats at a fine spatial scale, in order to explain the variation not accounted for by climatic variables varying at coarse scales. The two components in this spatial variation, a global trend and spatially dependent local variation in residuals, are estimated by logistic and kriging models, respectively.

The logistic regression models confirm the observed characteristics of larval habitats. The three independent variables included in the logistic models (distance to the nearest waters, elevation, and land use) represent three underlying environmental factors. The statistical significance of land use may contribute to the discussion on the role of land use change plays in the outbreaks of malaria. The clearing of forests is the most drastic environmental change in the study area. Forests are not known as a favorable larval habitat because of their shade and cool temperature. The logistic models confirm this belief. However, there is no statistical evidence that farmlands that replace forests are directly associated with larval habitats. The effect of land use change on malaria outbreaks lies in its impact on other aspects of environment. Studies suggest that the removal of forests can increase sunlight, air temperature, and water temperature (Minakawa et al., 1999; Lindblade et al., 2000; Zhou, et al., 2004). Combined with precipitation, these changes induced by the removal of forests can increase the spatial distribution of larval habitats and mosquito populations, consequently triggering major malaria outbreaks (Lindsay and Birley, 1996; Githeko et al., 2000).

Residuals are normally expected to contain random noise and unaccounted-for variables that may or may not have spatial dependence (Kleinschmidt et al., 2000; Miller, 2005). It is usually difficult to identify the unaccounted-for variables from numerous possible ones. Kriging, taking advantage of spatial dependence, can estimate the spatially dependent local variation in residuals based on other residuals at adjacent locations, without having to search for causal variables. Adding this spatially dependent local variation to the global trend improves the modeling accuracy of this study. This observation is consistent with other reports in the literature (Kleinschmidt et al., 2000; Miller 2005).

Spatial dependence is likely to be found in the residuals of a regression model when geographic data are involved. The modeling of this spatial dependence is expected to improve the overall model accuracy. The range of the spatial dependence (i.e., the range of a semivariogram), however, is constrained to the spatial resolution of the residual data. The average spacing between the 1,638 data points used in this study is equivalent to a 99 m resolution (while the actual spacing between a good proportion of the data points is much shorter than 99 m). This data set allows the detection of spatial dependence of a 120 m range present in the residuals. Spatial

Fig. 3. A comparison of mis-estimated locations between the minimized logistic model (A) and the combined minimized model (B). The results are from validations of the models at the 820 locations. Presence = correctly estimated presence locations ($p \geq 0.5$). Absence = correctly estimated absence locations ($p < 0.5$). Over Estimate = actual absence locations with $p \geq 0.5$. Under Estimate = actual presence locations with $p < 0.5$.

dependence with a range shorter than the spatial resolution of the data cannot be effectively modeled, as a phenomenon can only be modeled at the resolution that the data are collected at or coarser resolutions.

No systematic difference in accuracy is observed between the full models and minimized models. This is because elevation and distance to the nearest waters, two of the three independent variables included within both sets of models, are spatially dependent. The minimized models, contrary to the initial intention to minimize the spatial dependence, cannot be distinguished from the full models. Therefore, it is difficult to identify the impact of minimizing spatial dependence in this study area. Perhaps different sampling treatments or using simulated data can be more effective in reaching this goal. From a practical perspective, the similarity between the full and minimized models can be used to design a field survey in order to obtain an adequate data set using a lesser number of locations. This is especially useful for extended studies over a large area in highland Africa.

ACKNOWLEDGMENTS

This research was funded in part by the National Institute of Health under Award No. #R01AI50243. The authors acknowledge the assistance provided by the College of Arts and Sciences at the State University of New York at Buffalo. The insightful suggestions and comments provided by the anonymous reviewers are greatly appreciated.

REFERENCES

- Band, L. E., 1986, "Topographic Partition of Watersheds with Digital Elevation Models," *Water Resources Research*, 22(1):15-24.
- Beven, K. and M. Kirkby, 1979, "A Physically Based Variable Contributing Area Model of Basin Hydrology," *Hydrological Sciences Bulletin*, 24(1):43-69.
- Bian, L and E. West, 1997, "GIS Modeling of Elk Calving Habitat in a Prairie Environment with Statistics," *Photogrammetric Engineering and Remote Sensing*, 63(2):161-167.
- Bishop, T. E. A. and A. B. McBratney, 2001, "A Comparison of Prediction Methods for the Creation of Field Extent Soil Property Maps," *Geoderma*, 103(1-2):149-160.
- Burrough, P. A., 1986, *Principles of Geographical Information Systems for Land Resources Assessment*, Oxford, UK: Oxford University Press, 194 p.
- Davis, J. C., 2002, *Statistics and Data Analysis in Geology*, third edition, New York, NY: Wiley, 638 p.
- Diggle, P. J., Tawn, J. A., and R. A. Moyeed, 1998, "Model Based Geostatistics," *Journal of the Royal Statistical Society Series C*, 47(3):299-350.
- Githeko, A. K., Lindsay, S. W., Confalonieri, U. E., and J. A. Patz, 2000, "Climate Change and Vector Borne Diseases: A Regional Analysis," *Bulletin of the World Health Organization*, 78(9):1136-1147.
- Hay, S. I., Simba, M., Busolo, M., Noor, A. M., Guyatt, H. L., Ochola, S. A. and R. W. Snow, 2002, "Defining and Detecting Malaria Endemics in Highlands of Western Kenya," *Emerging Infectious Diseases*, 8(6):555-562.

- Hodgson, M. E. and G. L. Gaile, 1999, "A Cartographic Modeling Approach for Surface Orientation-Related Applications," *Photogrammetric Engineering and Remote Sensing*, 65(1):85-95.
- Isaaks, E. H. and R. M. Srivastava, 1989, *An Introduction to Applied Geostatistics*, New York, NY: Oxford University Press, 561 p.
- Klienschmidt, I., Bagayoko, M., Clarke, G. P. Y., Craig, M., and D. Le Sueur, 2000, "A Spatial Statistical Approach to Malaria Mapping," *International Journal of Epidemiology*, 29:355-361.
- Lindblade, K. A., Walker, E. D., Onapa, A. W., Katungu, J., and M. L. Wilson, 2000, "Land Use Change Alters Malaria Transmission Parameters by Modifying Temperature in a Highland Area of Uganda," *Tropical Medicine and International Health*, 5(4):263-274.
- Lindsay, S. W. and M. H. Birley, 1996, "Climate Change and Malaria Transmission," *The Annals of Tropical Medicine and Parasitology*, 90:573-588.
- Mbogo, C. M., Mwangangi, J. M., Nzovu, J., Gu, W., Yan, G., Gunter, J. T., Swalm, C., Keating, J., Regens, J. L., Shililu, J. I., Githure, J. I., and J. C. Beier, 2003, "Spatial and Temporal Heterogeneity of Anopheles Mosquitoes and *Plasmodium falciparum* Transmission along the Kenya Coast," *American Journal of Tropical Medicine and Hygiene*, 68(6):734-742.
- Miller, J., 2005, "Incorporating Spatial Dependence in Predictive Vegetation Models: Residuals Interpolation Methods," *The Professional Geographer*, 57(2):169-184.
- Minakawa, N., Muteru, C. M., Githure, J. I., Beier, J. C., and G. Yan, 1999, "Spatial Distribution and Habitat Characterization of Anopheline Mosquito Larvae in Western Kenya," *American Journal of Tropical Medicine and Hygiene*, 61(6):1010-1016.
- Minikawa, N., Sonye, G., Mogi, M., Githeko, A. and G. Yan, 2002, "The Effects of Climatic Factors on the Distribution and Abundance of Malaria Vectors in Kenya," *Journal of Medical Entomology*, 39(6):833-841.
- Mladenoff, D. J., Sickley, T. A., Haight, R. G. and A. P. Wydeven, 1995, "A Regional Landscape Analysis and Prediction of Favorable Gray Wolf Habitat in the Northern Great Lakes Region," *Conservation Biology*, 9(2):279-294.
- Pereira, J. M. C. and R. M. Itami, 1991, "GIS-based Habitat Modeling Using Logistic Regression: A Study of the Mt. Graham Red Squirrel," *Photogrammetric Engineering and Remote Sensing*, 57(11):1475-1486.
- Sakai, T., Suzuki, H., Sasaki, A., Saito, R., Tanabe, R. and K. Taniguchi, 2004, "Geographic and Temporal Trends in Influenza Like Illness, Japan, 1992-1999," *Emerging Infectious Diseases*, 10(10):1822-1826.
- Tarboton, D. G., 2004, "Terrain Analysis Using Digital Elevation Models (TauDEM)," [<http://www.engineering.usu.edu/dtarb/>].
- Walter, S. D., 1994, "A Simple Test for Spatial Pattern in Regional Health Data," *Statistics in Medicine*, 13:1037-1044.
- Zhou, G., Minakawa, N., Githeko, A. K., and G. Yan, 2004, "Association between Climate Variability and Malaria Epidemics in the East Africa Highlands," *Proceedings of the National Academy of Sciences of the United States of America*, 101(8):2375-2380.