

An Integrated Bayesian Modelling Approach for Predicting Mosquito Larval Habitats

LI LI and LING BIAN

Department of Geography, State University of New York at Buffalo
Amherst, New York 14261
e-mail: lli7@buffalo.edu; lbian@buffalo.edu

GUIYUN YAN

Program in Public Health, College of Health Sciences
University of California, Irvine, CA 92697
e-mail: guiyuny@uci.edu

Malaria is the leading cause of death in Kenya highlands. Malaria is a vector-borne disease and mosquito is the vector that transmits the parasites from infected people to others. Controlling mosquito larval habitats is important in eradicating malaria. Modeling the spatially distributed mosquito larval habitats is challenging. This study explores a mosquito habitat modeling approach, which integrates a Bayesian modeling method with Ecological Niche Factor Analysis (ENFA). ENFA transforms original environmental variables into niche factors. These factors are extracted to maximize the variations between the global space, which represents the general conditions of an area, and the focal space, which represents habitats. The integrated method is compared with the Bayesian method without ENFA. Both methods accurately predicted the habitats approximately 70-80%. Although the two methods have similar prediction accuracies, the integrated approach produces more ecologically justified spatial distribution of larval habitats. The map derived from this estimation has important implications for malaria control and eradication.

Keywords: Mosquito habitats; Bayesian modeling; Ecological Niche Factor Analysis

1. Introduction

Malaria induced by *Plasmodium falciparum* parasites is the largest cause of mortality in Kenya highlands (Zhou et al., 2004). Malaria is a vector-borne disease and mosquito is the vector that transmits the parasites from infected people to others. *Anopheles gambiae* is the primary mosquito species for the transmission (Githeko and Ndegwa, 2001). The number of malaria cases is related to the mosquito population, which is influenced by the availability of the mosquito habitats (Keating, 2003). The mosquito larvae are of relative low mobility compared with the flying adults. The identification and control of the *An. gambiae* larval habitats is critical in reducing malaria cases. Modelling the spatially distributed larval habitats is especially challenging, because of the associated uncertainties (Gu and Novak, 2005).

Bayesian statistical inference is recognized as a suitable method to deal with uncertainties in habitat modelling (Aspinall, 1992; Calcerrada and Luque, 2006). This method can be used to update a prior probability of the presence of a species using environmental variables, in order to estimate a posterior probability of the presence (Aspinall, 1992). This Bayesian approach has been used to map spatial distribution of the habitats of a number of wildlife (Aspinall, 1991; 1993).

Conditional independence is a fundamental assumption of Bayes' theorem. Violating of this assumption will cause error in predicting the posterior probability. This assumption is problematic for environmental variables, since many are often dependent to each other. Some studies simply assumed or tested the independence between the environmental variables (Calcerrada and Luque, 2006). The common treatment for the conditional dependence is to disregard one variable or transform the correlated variables (Calcerrada and Luque, 2006). This may result in losing useful information. Few effective solutions for the dependence problem have been identified.

Based on Hutchinson's ecological niche concept, Ecological Niche Factor Analysis (ENFA) (Hirzel et al., 2002) transform the multiple correlated environmental variables into uncorrelated niche factors. These uncorrelated factors satisfy the conditional independence assumption of Bayes' theorem, offering a plausible mechanism to transform variables for the Bayesian modelling. Unlike other transformation methods, such as principle component analysis, ENFA transformation aims at the contrasting of favourable condition of habitats and general conditions of an area.

The objective of this study is to integrate the ENFA transformation with Bayesian habitat modelling in order to improve the prediction of the spatial distribution of *An. gambiae* larval habitats in the Kenya highlands. Results of this integration are compared with a Bayesian modelling without the ENFA transformation, to evaluate the effectiveness of this integration.

2. Study Area and Data

This study focuses on a 4x4 km area, centered at 0°10' N and longitude 34°45' W (Fig. 1). This area is located in Iguhu Village, Kakamega District of Kenya. Yala River drains this area into Lake Victoria to the west (Zhou et al., 2004). The climate of the region is subtropical, characterized by an annual precipitation of 40-70 inches and two rainy and two dry seasons (Kaplan et al., 1976). The first rainy season beginning at late April is predominantly longer than the other. The mean annual daily temperature is 20.3°C. The terrain is typical of Africa highlands, consisting of a mosaic of hills and small basins, with elevation ranging from 1420-1540m. The land use is mostly farmlands interleaved by patches of forests, pastures, shrubs, and swamps (Minakawa et al., 2002b). Data for this study include the mosquito larval information and environmental variables. The larvae data identify the locations where larvae were observed. A total of 373 locations were obtained from mass field surveys conducted in the long rainy season (May) in 2004 and 385 locations in May

2005. Data for seven variables are used to represent the environmental conditions of a larval habitat. These variables include elevation, wetness index, curvature, heat load index, land use, distance to stream and Normalized Difference Vegetation Index (NDVI) (Table 1). A contour map of 1:50,000 scale and 20 meter contour interval is used to construct a digital elevation model (DEM). The DEM is used to calculate elevation, wetness index, curvature and heat load index. A one meter resolution IKONOS image acquired in April 2002, is used to derive the land use types, streams and NDVI data.

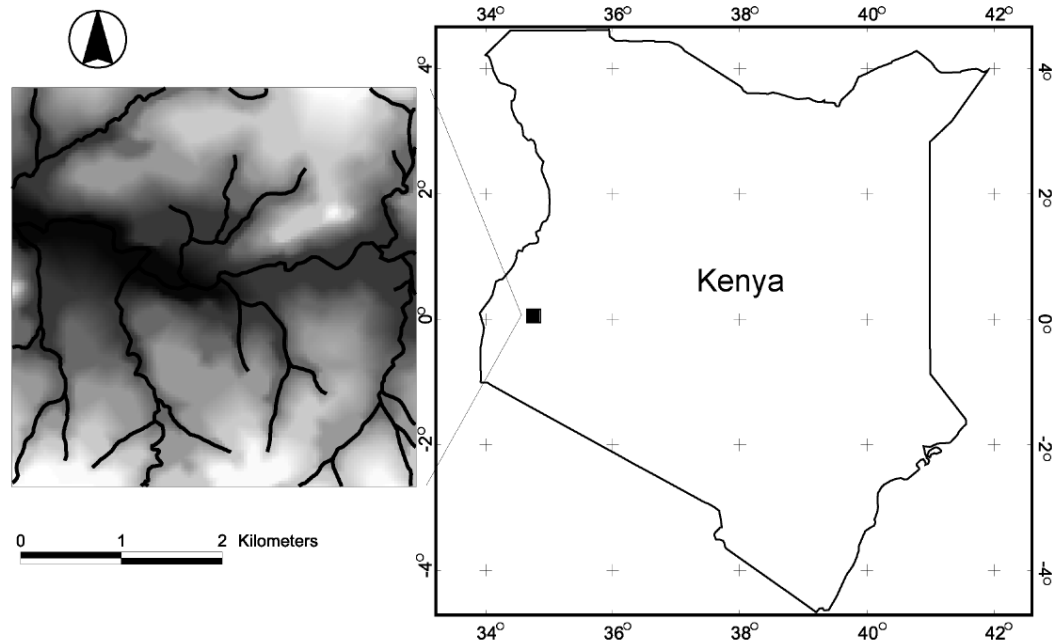


Fig. 1. The 4 × 4 km study area in western Kenya (Bian et al., 2006).

Elevation and heat load index are used to represent climatic conditions. Elevation is in general positively related to precipitation and negatively related to temperature, whereas heat load index is a substitute for solar radiation (McCune and Keon, 2002). Since larvae occur only in water, their dependence on aquatic environment is one of the dominant characteristics of their habitats. Wetness index, curvature, and distance to streams are used to describe the aquatic environment. The wetness index $\ln(A/\tan B)$ is widely used to represent soil moisture (Beven and Kirkby, 1979), where A is the draining area and B is the slope of a location. Curvature is a measure of the convexity or concavity of the land surface, an indicator of possible accumulation of ephemeral waters. Distance to streams designates the water availability. Land use types are related to the impact of human activities on the larval habitats. Five land use types were identified: forest, shrub, swamp, farm and pasture. NDVI is used to represent the vegetation condition, which has been used in several malaria related studies (Lindsay et al., 1998; Srivastava et al., 2005).

Table 1. Environmental variables, associated source data and code.

Variables	Data Source	Code
Elevation	DEM	Elev
Heat load index	DEM	Heat
Wetness index	DEM	Wet
Curvature	DEM	Cur
Distance to streams	IKONOS Image	Dist
Land use	IKONOS Image	Land
NDVI(Normalized Difference Vegetation Index)	IKONOS Image	NDVI

All seven environmental variables are prepared in raster maps, and the point data of larval locations are also converted to the raster format. All eight variables are in UTM coordinates.

3. Method

Firstly we tested the conditional independence using contingency matrix between each pairs of environmental variables. ENFA is conducted to transform the original environmental variables into niche factors. Bayesian habitat modelling is used to update the prior probability of the presence by these niche factors to produce a posterior probability. This posterior probability is mapped to predict the spatial distribution of the suitable habitats. In addition, Bayesian habitat modelling is also applied to the original environmental variables to derive a posterior probability map for comparison purpose. Lastly, these two maps are validated by three sets of data. The difference in these two maps is examined and discussed within the ecological context.

3.1 Conditional independence test

Conditional independence between the environmental variables is tested by using a contingency matrix between all pairs of variables (Bonham-Carter, 1989). Two variables are evaluated in the matrix to detect the dependence between them conditioning on the presence of habitat. The rows of the contingency matrix are values of one variable, and the columns of the matrix are values of the other variable. Each cell of the matrix records the number of co-occurrence of value i of one variable and value j of the other variable, in the presence of habitat. Chi-square test is then applied to evaluate whether the co-occurrence of these values is significantly higher than chance. A statistically significant high co-occurrence indicates a conditional dependence between the two variables.

3.2 Ecological Niche Factor Analysis

Like Principal Component Analyses (PCA), ENFA transforms original variables into orthogonal factors. Each factor is the linear combination of the original variables. Each component captures the

maximum variance remaining in the data, while in ENFA, in contrast, the first axis is chosen to maximize the variation between a “global” space and a “focal” space. The global space represents the collective condition of an entire area and is defined by the original variables. The focal space represents the most favourable habitat condition, and is a part of the global space. The first factor, thus, represents “marginality” of the niche. Each succeeding factor maximizes the ratio between the variance of the global space and that of the focal space remaining in the data (Hirzel et al., 2002). These factors represent “specialization” of the niche. In other words, how specialized the habitats are.

Coefficients associated with original variables in the linear combination indicate the contribution of these original variables to a factor. Those variables that contribute the most to the first factor are the most important in distinguishing the optimal habitat conditions from the general conditions in a study area. The ENFA method has been used in many habitat studies (Hirzel et al., 2002; Hirzel and Arlettaz, 2003). In addition to the orthogonal transformation, the factors carry distinct ecological meanings (Hirzel et al., 2004; Chefaoui et al., 2005).

For this study, the ENFA analysis is used to transform the seven environmental variables, presented in raster maps, into several niche factors also presented in raster maps. The full ranges of these variables define the global space. Their ranges associated with the presence of larval habitats define the focal space. As aforementioned, the first niche factor is extracted based on the marginality of the focal space, and the remaining factors are based on the specialization of the focal space. For detailed treatments on statistical context of the method please see Hirzel et al. (2002). In this study, the ENFA transformation procedure is implemented using a GIS package Biomapper (Hirzel et al. 2002). Because categorical variables cannot be used in Biomapper, the land use variable is converted into five Boolean maps, one for each category. Totally 11 maps (six original variables and five Boolean variables) are processed by the ENFA procedure.

3. 3 Bayesian habitat modeling integrated with ENFA transformation (BE)

Bayesian modelling has been used to estimate probability of the presence of habitats. In this process, a prior probability of the presence is updated according to environmental conditions, represented by various variables (Aspinall, 1992). In this study, the priori habitat probability is set as the percentage of observed habitat locations in the study area. The niche factors are used to define the environmental conditions. The value of each niche factor is re-classified into several classes according to the variation of factor values. The frequency of each factor class conditioned on the presence of the habitats is used to update the prior probability and subsequently produce the habitat probability, as shown in Equation 1.

$$P(H / E) = \frac{P(H) * \prod_{i=1}^n P(E_i | H)}{\prod_{i=1}^n P(E_i)} \quad \text{(Equation 1, Aspinall, 1992)}$$

where P(H) is the prior probability of habitats. P(E_i) is the frequency of each factor class i. P(E_i|H) is the frequency of each factor class conditioned on the presence of the habitats. This conditional frequency is then used to update the P(H). The probability for the presence of habitats is estimated for each pixel in the study area. A habitat suitability map is produced for *An. gambiae* larval based on these probabilities. This procedure is implemented in a software package called Bayesian modeling V1.0 (Aspinall, 1992).

3.4 Bayesian habitat modeling without the ENFA transformation (BA)

To evaluate the effect of the ENFA transformation, the Bayesian habitat modelling procedure is also operated on the original environmental variables. This procedure also produced a habitat suitability map. The major difference between Bayesian habitat modelling integrated with ENFA (BE) and Bayesian habitat modelling without ENFA (BA) is the input variables. The habitat suitability map produced by BE is compared with the map produced by BA. Both habitat suitability maps are reclassified into two categories, presence and absence, with a cut off point 0.5. Those locations with probability ≥ 0.5 are classified as presence of habitats, and probability < 0.5 as absence.

3.5 Validation

For validation, the 373 habitat locations collected in May 2004 are divided into two parts. 80% of these locations are used for the Bayesian model development and the remaining 20% are used for validation. The 385 habitat locations gathered in May 2005 are also used to evaluate the temporal robustness of the model.

4. Results and Discussion

4.1 Conditional independence test and Validation

The integrated Bayesian model (BE) predicted 28% of the study area as suitable habitats, and Bayesian model without the ENFA transformation (BA) predicted 36% of the study area as suitable habitats. The estimated suitability is compared against the data used for model development (80% of the data). 75% and 76% are correctly estimated by BE and BA, respectively. Both models are further validated using the 20% data that are not involved in the model development. The validation accuracy is 81% and 80% for the BE and BA, respectively. The validation using the data collected in May 2005 shows 70% accuracy for BE and 67% for BA. These three sets of validation show similar prediction accuracy between BE and BA. These results are satisfactory, although the significant improvement by using BE over BA is not observed.

The lack of improvement of BE could be caused by a relatively weak conditional dependence between the original environmental variables. Table 2 displays the conditional dependence between the seven environmental variables. The “*” sign indicates correlation at the significance level of 0.05. Half of the pairs are significantly independent to each other, while 11 pairs show conditional

Table 2. The probability table shows the conditional independence test.

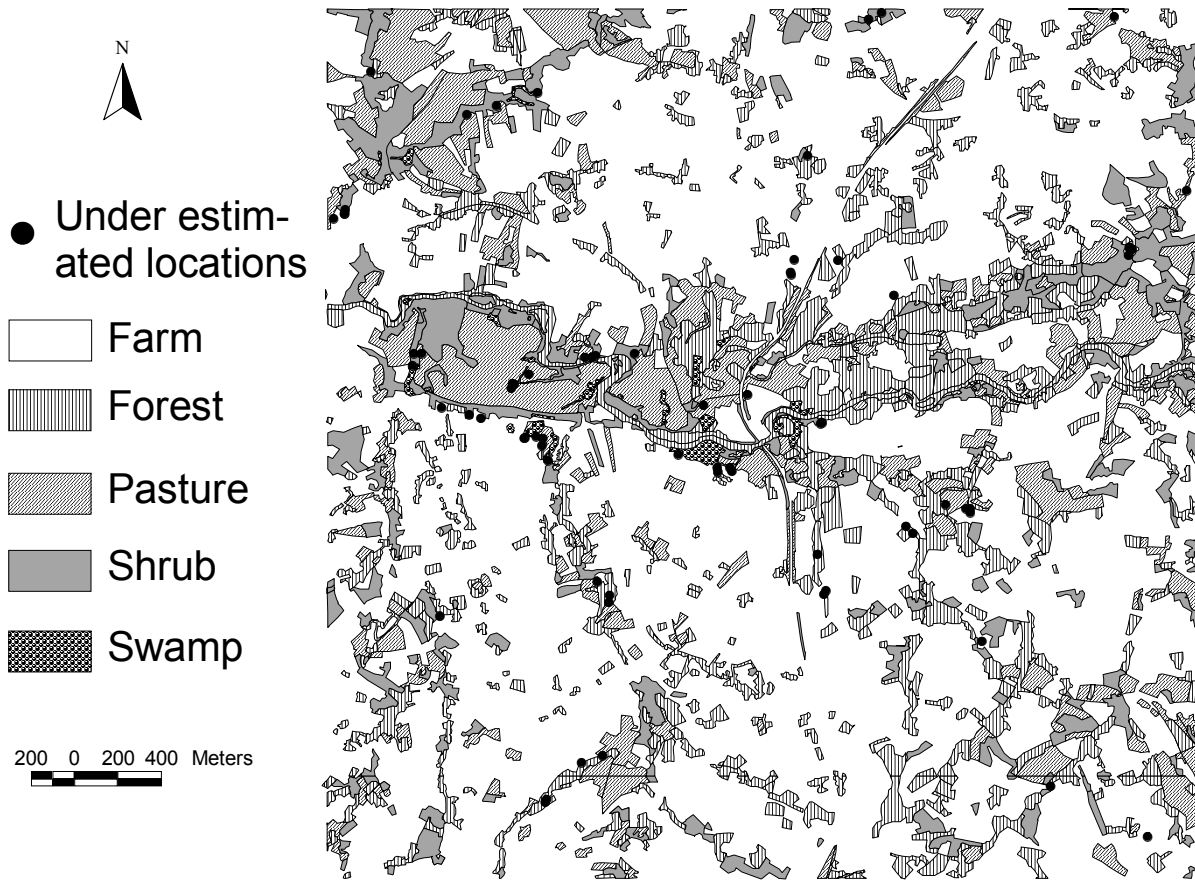
	Elev	Heat	Wet	Cur	Dist	Land	NDVI
Elev	/	0.4648	*	0.9987	*	*	0.6849
Heat	0.4648	/	*	*	0.9989	*	*
Wet	*	*	/	*	0.9536	*	0.9985
Cur	0.9987	*	*	/	1	0.537	0.9873
Dist	*	0.9989	0.9536	1	/	*	0.9999
Land	*	*	*	0.537	*	/	*
NDVI	0.6849	*	0.9985	0.9873	0.9999	*	/

Note: * indicates a conditional dependence between the variables.

correlations. Five out of these 11 pairs involves land use. Removing this variable can significantly reduce the amount of conditional dependence between variables. Therefore, BE cannot take the full advantage of the orthogonal transformation offered by ENFA.

Although there is no obvious difference in prediction accuracy, BE and BA present different spatial patterns of suitable habitats areas. Fig. 2(a) and 2(b) display the incorrectly predicted locations by BA and BE, respectively.

(a)



(b)

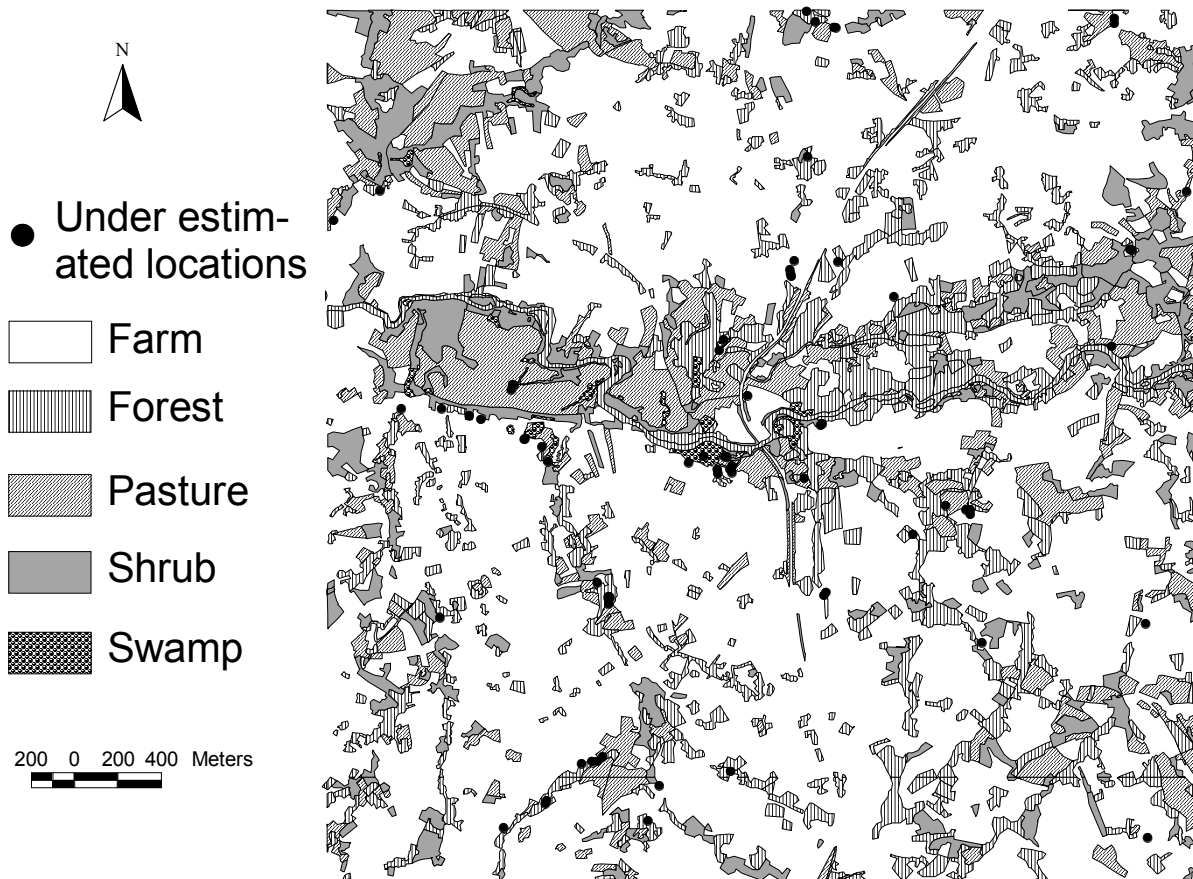


Fig. 2. Land use map of the study area showing locations of underestimated presence sites (black dots) by the (a) Bayesian without ENFA transformation method and (b) Bayesian with ENFA transforms.

Bayesian without ENFA transformation tends to incorrectly predict locations associated with shrubs, where as few such locations are wrongly predicted by Bayesian with ENFA transformation. Most of the shrubs in the study area are strips along streams, thus providing aquatic environments, an ideal habitat, for the larvae. The shrubs occupy only 8% of the total area but contribute 29% of the observed habitat locations. Most of the incorrectly predicted locations of BE are associated with farmlands and forests. These two land use types have been found to be negatively correlated with larval habitats (Bian et al., 2006). The larval habitats observed in farmlands and forests are most likely transitional (Zhou et al., 2004). Although with similar accuracy, BE predictions are more ecological meaningful in terms of corresponding to the actual habitats.

4.2 Ecological Niche Factor Analysis (ENFA)

Table 3 displays the marginality coefficient and specialization coefficients of the niche factors. These coefficients indicate the contribution of the original variables to each of the factors. The

percentage of total variance explained by each factor is displayed for the first five factors that accounted for 85% of the total variance in the data.

Table 3. The marginality and specialization coefficients for each variable.

	Factor 1 (41%)#	Factor 2 (20%)*	Factor 3 (10%)*	Factor 4 (8%)*	Factor 5 (6%)*
Wet	0.59	0.04	-0.21	0.45	-0.05
Dist	-0.53	0.48	-0.55	0.2	0.23
Farm	-0.34	0.02	0.03	-0.13	-0.6
Shrub	0.31	0	-0.15	-0.02	-0.02
Elev	-0.27	0.03	0.66	0.6	0
Heat	-0.23	-0.8	-0.41	0.45	0.02
Cur	-0.13	-0.36	0.18	-0.41	0.38
Pasture	0.12	0.03	-0.04	0.02	0.07
NDVI	0.07	0.05	0.04	0.1	0.14
Swamp	0.06	0	0.04	0.04	-0.02
Forest	-0.01	0.05	-0.01	0	0.64

Note: # indicates a marginality factor and * indicates a specialization factor.

Factor 1 accounts for the highest variance and more importantly, this factor is the one that distinguishes habitat conditions from the general conditions of the study area. The marginality coefficients that associate the original variables with this factor are most important. The positive marginality coefficients indicate a positive correlation between habitats and these variables. Habitats are usually found at the area where the variable value is higher than its average value in the area. The negative marginality coefficients indicate a negative correlation, and habitat is usually found at the area where the variable value is below its average value. For the remaining factors, the specialization coefficients define the niche breadth. Higher coefficient of the specialization for one variable shows that the distribution of habitats is specially restricted by this variable.

Table 3 shows that two original variables, the wetness index and distance to streams, contribute the most to the first factor. While the wetness index is positively related, the distance to the streams is negatively related. This confirms the mosquito larvae's dependence on the aquatic environment. The next pair of the original variables that make notable contribution to the first factor are farmlands and shrubs. Their negative and positive contributions, respectively, are explained earlier.

The highest absolute value of specialization coefficient is 0.8, which associates the original variable, heat load index, to factor 2. This indicates the importance of solar radiation. This confirms the belief that *An. gambiae* larvae prefer sunlit water bodies (Minakawa et al., 1999, 2002a).

5. Conclusions

Although incorporating the ENFA does not improve the prediction accuracy, the integrated approach produces more ecologically justified predictions. This method shows its potential in estimating spatial distribution of *An. gambiae* habitats. The map derived from this estimation has important implications for malaria control and eradication. The plans to control larval habitats can

be directed to the most critical areas revealed by this map. Methods used in this study can be easily applied to large areas to facilitate malaria prevention effort.

Acknowledgements

This research was supported in part by the National Institute of Health under Award No. #R01AI50243. The authors also appreciate the travel award provided by UCGIS.

References

- Aspinall, R. 1991. Use of an inductive modelling procedure based on Bayes theorem for analysis of pattern in spatial data. *Computer Modelling in the Environmental Sciences*. Clarendon Press, Oxford, pp. 325–339.
- Aspinall, R. 1992. An Inductive Modeling Procedure Based on Bayes Theorem for Analysis of Pattern in Spatial Data. *International Journal of Geographical Information Systems*, 6:105-121.
- Aspinall, R. and Veitch, N. 1993. Habitat mapping from satellite imagery and wildlife survey data using a bayesian modeling procedure in GIS. *Photogrammetric Engineering and Remote Sensing*, 59(4): 537-543.
- Beven, K. and Kirkby, M. 1979. A Physically Based Variable Contributing Area Model of Basin Hydrology. *Hydrological Sciences Bulletin*, 24(1):43-69.
- Bian, L., Li, L., and Yan, G. 2006. Combining global and local estimates for spatial distribution of mosquito larval habitats. *GIScience and Remote Sensing*, 43(1): 95-108.
- Bonham-Carter, G.F., Agterberg, F.P. and Wright, D.F. 1989. Weights of evidence modelling: a new approach to mapping mineral potential. *Statistical Applications in the Earth Sciences*, Agterberg, pp. 171-183.
- Calcerrada, R. and Luque, S. 2006. Habitat quality assessment using Weights-of-Evidence based GIS modelling: The case of *Picoides tridactylus* as species indicator of the biodiversity value of the Finnish forest. *Ecological Modelling, In Press*.
- Chefaoui, R., Hortal, J. and Lobo, J.M. 2005 Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian Copris species. *Biological Conservation*, 122: 327-338.
- Githeko, A. and Ndegwa, W. 2001. Predicting malaria epidemics in the Kenyan highlands using climate data. *Global change & human health*, 1: 54-63.
- Gu, W. and Novak, R.J. 2005. Habitat-based modeling of impacts of mosquito larval interventions on entomological inoculation rates, incidence, and prevalence of malaria. *American Journal of Tropical Medicine and Hygiene*, 73:546-552.
- Hirzel, A.H., Hausser, J., Chessel, D. and Perrin, N. 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83:2027-2036.
- Hirzel, A.H. and Arlettaz, R. 2003. Modelling habitat suitability for complex species distributions by the environmental-distance geometric mean. *Environmental Management*, 32: 614-623.
- Hirzel, A.H., Posse B., Oggier, P.A., Crettenand Y., Glenz C. and Arlettaz, R. 2004. Ecological requirements of a reintroduced species, with implications for release policy: the Bearded vulture recolonizing the Alps. *Journal of Applied Ecology*, 41:1103-1116.
- Keating, J., Macintyre, K., Mbogo, C., Githeko, A., Regens, J.L., Swalm, C., Ndenga, B., Steinberg, L.J., Kibe, L., Githure, J.I. and Beier, J.C. 2003. A geographic sampling strategy for studying relationships

- between human activity and malaria vectors in urban Africa. *American Journal of Tropical Medicine and Hygiene*, 68:357-365.
- Kaplan, I., Matthews, J., Rossi, E., Townsend, C. and Walpole, N. 1976. Area Handbook for Kenya. Second Ed., U.S. Government Printing Office, Washington, D.C.
- Lindsay, S.W., Parson, L., and Thomas, C.J. 1998. Mapping the ranges and relative abundance of the two principal African malaria vectors, *Anopheles gambiae sensu stricto* and *An. Arabiensis*, using climate data. *Proceedings of Royal Society B*, 265: 847-854
- McCune, B., and Keon, D. 2002. Equations for potential annual direct incident radiation and heat load index. *Journal of Vegetation Science*, 13:603-606.
- Minakawa, N., Mutero, C.M., Githure, J. I., Beier, J.C. and Yan, G. 1999. Spatial Distribution and Habitat Characterization of Anopheline Mosquito Larvae in Western Kenya. *American Journal of Tropical Medicine and Hygiene*, 61(6):1010-1016.
- Minakawa, N., Sonye, G., Mogi, M., Githeko, A. and Yan, G. 2002a. The Effects of Climatic Factors on the Distribution and Abundance of Malaria Vectors in Kenya. *Journal of Medical Entomology*, 39(6):833-841.
- Minakawa, N., Seda, P. and Yan, G. 2002b. Influence of host and larval habitat distribution on the abundance of African malaria vectors in western Kenya. *American Journal of Tropical Medicine and Hygiene*, 67: 32-38.
- Srivastava, A., Nagpal, B.N., Saxena, R. and Dev, V. 2005. Prediction of *Anopheles minimus* habitat in India – a tool for malaria management. *International Journal of Geographical Information Science*, 19(1): 91-97.
- Zhou, G., Minakawa, N., Githeko, A.K. and Yan, G. 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.