MODELING AND ANALYSIS OF MOSQUITO AND ENVIRONMENTAL DATA TO PREDICT THE RISK OF JAPANESE ENCEPHALITIS

Penny Masuoka[1], Terry A. Klein[2], Heung-Chul Kim[3], David M. Claborn[4], Nicole Achee[1], Richard Andre[1], Judith Chamberlin[1], Kevin Taylor[1], Jennifer Small[5], Assaf Anyamba[5], Michael Sardelis[6], John Grieco[1]

[1]Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences
4301 Jones Bridge Road, Bethesda, MD 20814
pmasuoka@usuhs.mil

[2]US Army MEDDAC-Korea, Unit 15244, Box 754, APO AP 96205-5244

[3]5th Medical Detachment, 168th Multifunctional Medical Battalion, Unit 15247, Box 754, APO AP 96205-5247

[4]Department of Nursing/Center for Homeland Security, Missouri State University
901 South National Avenue, Springfield, MO 65897

[5]Hydrological and Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, Code 614, Greenbelt, MD 20771

[6]National Center for Medical Intelligence, 1607 Porter Street, Fort Detrick, MD 21702

ABSTRACT

Culex tritaeniorhynchus is the primary vector of Japanese encephalitis virus (JEV) throughout much of the tropical and temperate climates of Asia. Several recent papers have used ecological niche modeling programs, e.g., Maxent and GARP, to predict the distribution of disease vectors (e.g. Peterson and Shaw 2003, Moffett et al. 2007). In this on-going study, we used the Maxent program to model the distribution of *Cx. tritaeniorhynchus* in the Republic of Korea. Using mosquito collection data, temperature, precipitation, elevation, land cover, and SPOT normalized difference vegetation index (NDVI), models were created for each month for a period of five years. Output maps from the models matched several known ecological characteristics of this species' distribution. The output maps show the highest probabilities of mosquito occurrence in August and September, which correlates to the observed mosquito population density peaks. The model demonstrated low probabilities for forest covered mountains, which corresponds to findings in the literature that *Cx. tritaeniorhynchus* is infrequently found above 1,000 meters. The modeling effort demonstrated several limitations in the data set, including a low number of collection sites that did not cover the full range of environmental conditions within the study area. Additional collection sites would improve the models and allow for improved testing of the results. Future goals of this project include developing real-time predictions based on NDVI data and expanding the prediction to a larger geographical area.

INTRODUCTION

Japanese encephalitis virus (JEV) is widely distributed throughout tropical Asia, including e.g., Indonesia, Borneo and the Philippines, as well as temperate countries, e.g., Japan and Korea (Straus and Straus, 2002). Over 35,000 cases resulting in about 10,000 deaths are reported annually. Survivors of the JEV infections can suffer long-term neurological effects requiring continual hospital/home care, with many deaths resulting from complications several years following infection.

JEV is reportedly spread by several species of mosquitoes, but the importance of the primary vector, *Cx. tritaeniorhynchu*s, is unquestioned. This mosquito prefers zoonotic hosts, primarily large water birds, cattle and pigs. The vertebrate reservoirs for JEV are large water birds and swine which often live in close proximity to human populations, develop very high viremias and serve as amplifying hosts. Early identification and prevention of JEV infections depends, in part, on vector surveillance and identification of circulating virus in the mosquitoes. Because mosquito populations are affected by factors such as elevation, precipitation, temperature, and land cover (including rice fields), these factors can be used to develop predictive models of likely areas of vector presence.

Several recent papers have used ecological niche modeling programs to predict the distribution patterns of diseases vectors (e.g. Peterson and Shaw 2003, Moffett et al. 2007). In this on-going study, we used the Maxent program to model the distribution of *Cx. tritaeniorhynchus* in the Republic of Korea (ROK).

METHODS

Mosquito Data

Adult mosquitoes were collected annually from May through October, 2002 to 2007, by the 5th Medical Detachment, 168 Multifunctional Battalion (Yongsan Army Garrison, Korea). The mosquitoes were collected using New Jersey light traps at selected U.S. military installations and training sites in the ROK (Figure 1). Because JEV has only been isolated from *Culex tritaeniorhychus* in South Korea, we restricted the data analysis to that mosquito species. The mosquitoes were totaled for each military base by month, and the numbers per trap-night were calculated. Because the number of mosquitoes varied greatly, we rescaled the data with a log transformation using the equation

$\ln(x+1) \times 10$, where x is the number of mosquitoes per trap-night.

Figure 1. Map showing the adult mosquito collection sites in South Korea.

Environmental Data

A number of environmental datasets were obtained for this project (Figure 2). All the environmental layers were coregistered, resampled to 1-km spatial resolution, and subset to cover the same area.

Climate data were obtained from the WorldClim Version 1.4 data set (http://www.worldclim.org). Gridded WorldClim precipitation and temperature data were based on monthly measurements from weather stations all over the world and averaged from years 1950 to 2000 to form a single 50-year 'climate surface' for each month. For a complete description of how the WorldClim data set is compiled, see Hijmans (2005). Elevation data derived from the Shuttle Radar Topography Mission (www2.jpl.nasa.gov/srtm/) were downloaded from the WorldClim site.

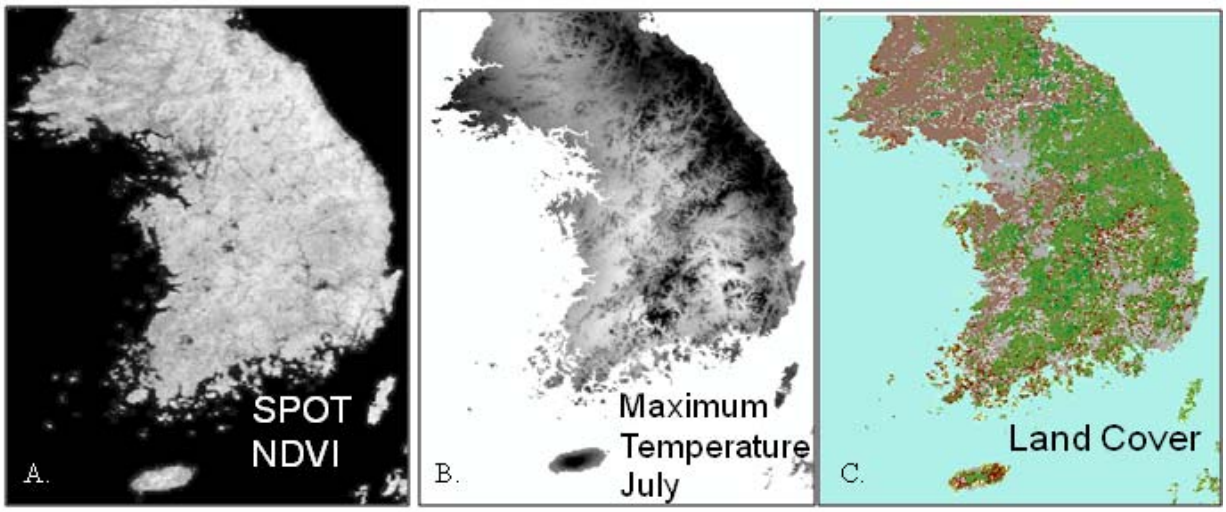

Figure 2. Examples of environmental layers used in the Maxent models: (A) normalized difference vegetation index data, (B) WorldClim 50-year averaged maximum temperature and, (C) Boston University's land cover derived from MODIS data.

Land cover data were obtained from two sources: the U.S. Geological Survey's (USGS) Global Land Cover Characteristics Database (http://edcsns17.cr.usgs.gov/glcc/) and Boston University's land cover website (http://www-modis.bu.edu/landcover/). The USGS data set was derived from the 1-km Advanced Very High Resolution Radiometer (AVHRR) data for a 12-month period (April 1992-March 1993) using the seasonal vegetation changes to aid in the land cover characterization process. The Global Ecosystems classification scheme, which contains 96 land cover classes, including a rice crop class, was used in the analysis.

Boston University's land cover data set was derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data. Of several possible classification schemes, the International Geosphere Biosphere Programme (IGBP) classification scheme, which contains 17 land cover classes, was used for analysis. In addition to using the original format, the image was also processed using a 5 x 5 majority filter that replaces the center of a 5 by 5 pixel area with the value of the most frequently occurring class within the box. The filtering allows the model to take into account the area surrounding the mosquito collection sites and not just the one pixel (1 square kilometer) at the collection site. In capture and release studies, female *Cx. tritaeniorhynchus* mosquitoes were collected up to 8.4 km from the release site (Wada et al., 1969) but have a usual flight range of <1.0 km. A 5 x 5 filter window includes pixels within 2.5 km of the center of each military base.

Normalized difference vegetation index data (NDVI) were used in the model as a proxy for ecological conditions. NDVI data were derived from the Systeme Probatoire pour l'Observation de la Terre (SPOT) VEGETATION data. Two NDVI products were used: monthly composite NDVI for each month for 2007 and long-term (May 1998- April 2008) monthly means.

Modeling Methods

The maximum entropy method, Maxent (Phillips et al., 2004, Phillips et al., 2006), was used to model the distribution of *Cx. tritaeniorhynchus*. Maxent takes a set of gridded environmental layers (such as land cover and precipitation), and a text file of known species' locations, and produces a map that predicts the potential distribution of a species.

Although other programs are available for ecological niche modeling, Maxent was selected because it has been shown to be one of the top performing modeling programs (Elith et al., 2006). Additionally, Maxent uses presence-only data, rather than presence/absence data. Absence data were not recorded in the Korea mosquito collections. Maxent has been shown to produce representative models using small numbers of occurrence points (Hernandez et al., 2006). [Although many mosquitoes were collected in Korea at the individual sites (at some sites >1,000), relatively few collection sites were used (no more than 29).] Finally, Maxent has been well documented and is readily available for download from Princeton University (http://www.cs.princeton.edu/~schapire/maxent/).

Maxent allows the user to select a percentage of randomly selected occurrence points to be used for testing the accuracy of the model (testing points) and the remaining to be used for building the model (training points). Twenty-five percent of the points were reserved for testing in this analysis. To evaluate the contribution of each environmental variable to the model, Maxent performs a jackknife of the training gain and produces plots of the prediction changes as each environmental variable changes.

RESULTS

Species Distribution Maps

Running the model using the number of mosquitoes per trap-night resulted in clustering of high probabilities of occurrence around individual collection sites. When the data were rescaled using ln(x+1) x 10, clustering was reduced, but still present. Finally, running the model using the rescaled data with no duplicate presence records resulted in the least clustering around individual collection sites. Figure 3 compares the results of using the rescaled data with and without using duplicate presence records.

Output Maxent maps, using rescaled data and no duplicate presence records, for 2007 are shown in Figure 4. High probability of *Cx. tritaeniorhynchus* occurrence is shown in red, and low probability of occurrence is shown in blue. There were only enough mosquitoes collected in July, August and September to run Maxent. The models show an increase in the probability of occurrence from July into August and September, with low probability areas corresponding to higher elevations.

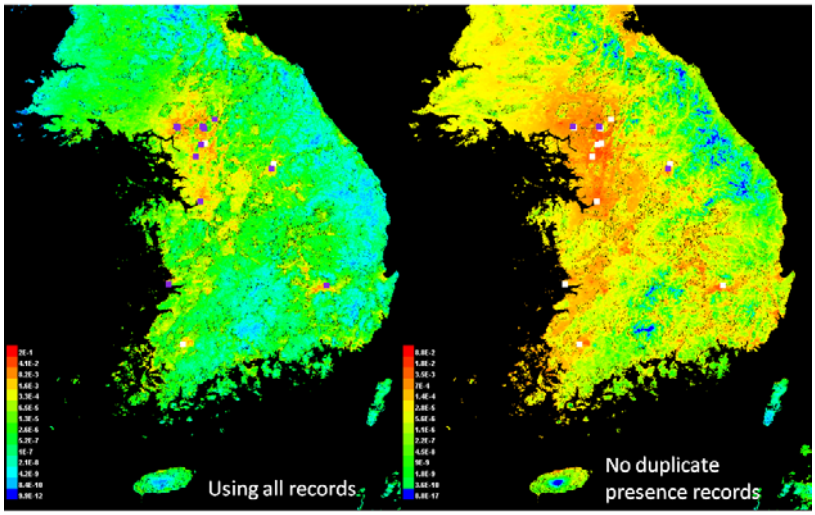

Figure 3. Output maps from Maxent for August 2007 showing the difference in predicted probability using all presence records versus no duplicate presence records. Both maps were created from data that had been rescaled using a log transformation.

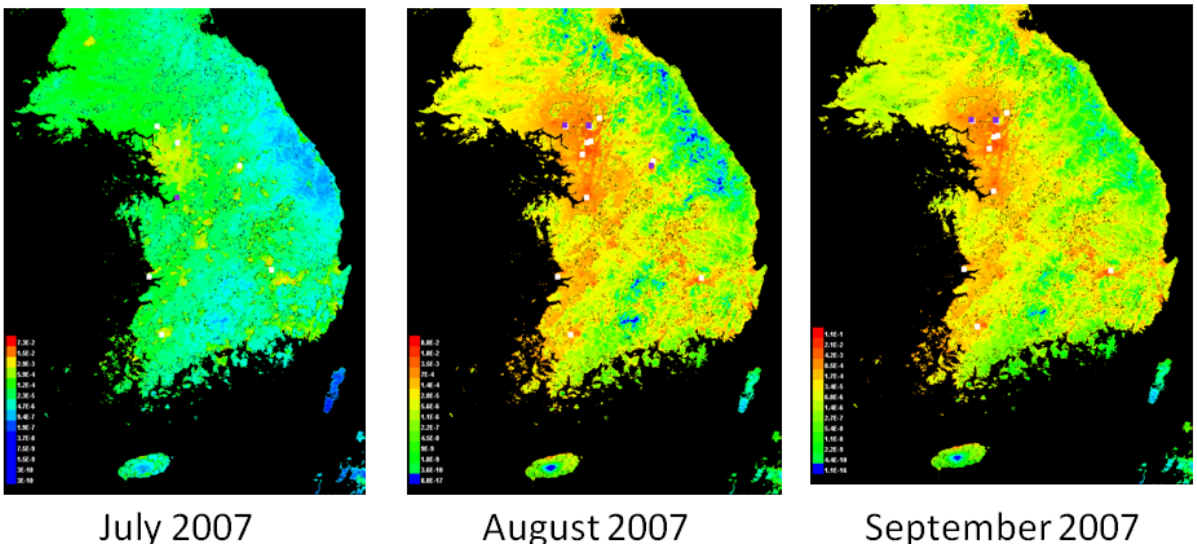

Figure 4. Output maps from Maxent showing the probability of distribution from high (red) to low probability (blue) for 2007.

Contributions of the Variables to the Model

Figure 5 shows the results of the jackknife of the regularized training gain for August 2007. The aqua bars show that when the model was run without each individual variable, there was little to no effect on the model. The blue bars show the training gain achieved by running the model using only that one variable. Land cover contributed the most to this model. Elevation, NDVI monthly ten year means, NDVI monthly means for 2007, and the minimum and maximum temperatures averaged over 50 years were important. For the minimum and maximum temperatures, the months of May through August provided the most training gain in the model.

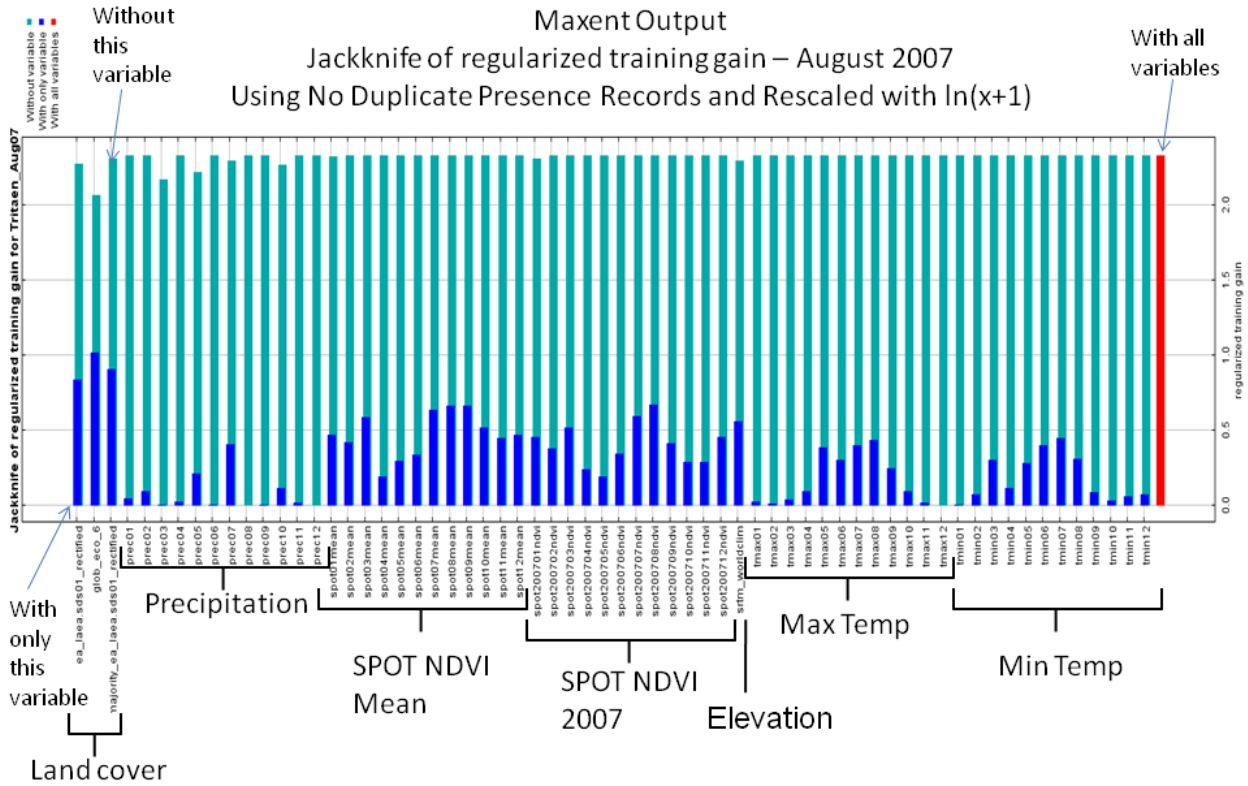

Figure 5. Graph showing the results of the jackknife of the regularized training gain for the August 2007 data set. The red bar shows the training gain achieved by using all variables in the model. The aqua bars show the training gain achieved by using all variables except one. The blue bars show the gain from a model built using only one variable.

Maxent also measures the response of individual environmental variables in the model. Maxent predicted probability of occurrence at elevations between 0 and 400 meters, with the lowest elevations being the most important. For land cover, the urban and savanna (trees with grass) land cover classes were the most important. However, on the majority filtered image, croplands and urban land cover were more important.

DISCUSSION

The output distribution prediction maps from Maxent matched several known ecological characteristics for the distribution of *Cx. tritaeniorhynchus* throughout its range in South Korea. The output maps indicated the highest probabilities of mosquito occurrence in August and September, which corresponds to the observed monthly mosquito population peaks.

The model demonstrated low probabilities for higher elevations, which matches the findings in the literature that *Cx. tritaeniorhynchus* is infrequently collected above elevations greater than 1,000 meters (Pandey et al., 2003 and Peiris et al., 1993). However, the model shows that in South Korea, the mosquitoes have little probability of occurring above 400 meters in elevation, most likely as a result of the forested hillsides and mountains that have very limited breeding habitat (mostly intermittent to permanent streams). Examining the locations of the collection sites, no collections were made above 310 meters. Future exploration of a more wide-ranging study area is a priority of future analyses.

Adult mosquito data used in this modeling effort were originally collected as a way to monitor mosquito

populations on U.S. military installations and training sites and not for JEV risk modeling purposes. As a result, many mosquito collections were made at a limited number of locations. While this is an ideal way to monitor mosquito densities for vector control, the collection method resulted in oversampling of some areas for modeling purposes. However, the results showed that by using no duplicate presence records, the clustering that resulted from oversampling can be reduced.

Another problem with the current mosquito sampling method is that all collections were made on U.S. military bases, which are representative of the urban land cover class. The important factor for presence or absence of mosquitoes is the mosquito breeding habitat near the bases and not the urban land cover. Using the majority filter to determine the majority land cover at each pixel location resulted in croplands (primarily rice) becoming more important in the model. One of the future goals for this project is to develop a better GIS method for incorporating the surrounding amount of rice-land cover in the model.

CONCLUSIONS

The modeling effort used in the current study demonstrated several limitations in the data sets including a low number of collection sites that did not cover the entire range of environmental conditions in the study area (in particular, elevation and land cover). Additional collection sites might improve the models and would allow for better testing of the results. In the coming years, some of these issues will be addressed by increasing the number and distribution of collection sites in an attempt to improve the model.

The current modeling effort demonstrates the importance of NDVI, land cover, and 50-year average monthly temperature in model development. Future goals of this project include developing real-time predictions incorporating baseline land cover data, elevation and time series NDVI data and expanding the prediction to a larger geographical area.

ACKNOWLEDGMENT

This research was funded by the Department of Defense Global Emerging Infections Surveillance and Response System (DoD-GEIS).

REFERENCES

Elith, J., C.H. Graham, R.P. Anderson, M.Dudik, S. Ferrier, A. Guisan, R.J. Hijamans, F. Huettmann, J.R. Lethwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, J. Sobéron, S. Williams, M.S. Wisz, and N.E. Zimmermann, 2006. Novel methods improve prediction of species' distributions from occurrence data, *Ecography*, 29:129-151.

Hernandez, P., C. Graham, L. Master and D. Albert, 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods, *Ecography*, 29: 773-785.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas, *International Journal of Climatology*, 25: 1965-1978.

Moffett, A., N. Shackelford and S. Sarkar, 2007. Malaria in Africa: Vector species' niche models and relative rick maps, *PLoSOne*, 2(9): e824. doi:10.1371/journal.pone.0000824.

Pandey, B., A. Yamamoto, K. Morita, Y. Kurosawa, S. Rai, S. Adhikari, P. Kandel, and I. Kurane, 2003. Serodiagnosis of Japanese encephalitis among Nepalese patients by particle agglutination assay, *Epi. and Infection*, 131: 881-885.

Peiris, J.S.M., F.P. Amerasinghe, C.K. Arunagiri, L.P. Perera, P.P. Karunaratne, C.B. Ratnayake, T.A. Kulatilaka, and M.R.N. Abeysinghe, 1993. Japanese encephalitis in Sri Lanka: Comparison of vector and virus ecology in different agro-climatic areas, *Trans. Royal Soc. Trop. Med. and Hygiene*, 87:541-548.

Peterson, A.T. and J.J. Shaw, 2003. *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: ecological niche models, predicted geographic distribuitons, and climate change effects, *International Journal of Parasitology*, 33: 19-931.

Phillips, S.J., M. Dudík, and R.E. Schapire, 2004. A maximum entropy approach to species distribution modeling,

In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 655-662.
Phillips, S.J., R.P. Anderson, and R.E. Schapire, 2006. Maximum entropy modeling of species geographic distributions, *Ecological Modelling*, 190:231-259.
Straus, J.H. and E.G. Straus, 2002. *Viruses and Human Disease*, Academic Press: San Diego, 383 pp.
Wada, Y., Kawai, S., Oda, T., Miyagi, I., Suenaga, O., Nishigaki, J., and N. Omori, 1969. Dispersal experiment of *Culex tritaeniorhynchus* in Nagasaki area (preliminary report), *Tropical Medicine*, 11(1): 37-44.