



Improving lives in Africa

FACILITATED BY IFPRI

Modeling the Distribution and Probability of Aflatoxin Occurrence Using Environmental Data

WORKING PAPER 2 • OCTOBER 2010

Penny Masuoka, Judith Chamberlin, and Maribel Elias

CONTENTS

Contents.....	i
List of Tables	ii
List of Figures	ii
Introduction	1
Background	1
Methods.....	2
Environmental data	2
Climate data.....	2
Vegetation indices.....	2
Elevation data	3
Land cover data.....	3
Other data layers	3
Aflatoxin data.....	3
Modeling.....	3
Input from Other Groups that We Will Need	5
Aflatoxin location data.....	5
Other data needed.....	9
Map of known maize and groundnut crops.....	9
Weather data	9
Harvest dates	9
Questions, Issues, and Limitations to the Study	9
Other Mapping Contributions.....	9
Conclusion.....	9
References	10

LIST OF TABLES

Table 1. Suggested format for collecting aflatoxin data.....	8
--	---

LIST OF FIGURES

Figure 1. Example raster environmental data sets to be used in the model.....	2
Figure 2. Example of output map from Maxent showing the predicted probability of the <i>Culex tritaeniorhynchus</i> mosquito occurring in South Korea.....	4
Figure 3. Example of a Maxent program AUC used for evaluating a <i>Culex tritaeniorhynchus</i> mosquito model.	4
Figure 4. Results of jackknife procedure for <i>Culex tritaeniorhynchus</i> mosquito.....	5
Figure 5. Example of good sampling plan for field collection of aflatoxin data with sampling sites shown as black dots.....	6
Figure 6. Examples of inadequate sampling of aflatoxin to create a good species distribution model.	7

INTRODUCTION

Aflatoxins, which are produced by *Aspergillus* molds, depend on environmental conditions in order to thrive. These toxins occur under drought conditions, when plants are more vulnerable to colonization by *Aspergillus* (Sanders et al. 1993), and post-harvest storage conditions that involve high humidity.

In recent years, there have been numerous studies of on the use of ecological niche models to determine species distribution.¹ These models use species presence records recorded as point locations with environmental layers to predict the distribution of the species. In this paper, we discuss our plan to use the Maxent program (Phillips, Dudik, and Schapire 2004; Phillips, Anderson, and Schapire 2006) to create an ecological niche model for aflatoxin.

BACKGROUND

Ecological niche models have been developed for a variety of applications, including:

- guiding population surveys in conservation biology (Guisan et al. 2006),
- predicting the effects of global change on a species (Pearson and Dawson 2003; Pearson, Dawson, and Liu 2004),
- finding a suitable habitat for the reintroduction of species (Pearce and Lindenmayer 1998), and
- modeling risk for malaria, Marburg virus, leishmaniasis, and other diseases (Peterson et al. 2006; Peterson and Shaw 2003; and Moffett, Shackelford, and Sarkar 2007).

Increased use of ecological niche modeling research can be attributed in part to the recent development of global environmental datasets including satellite data. For example, the normalized difference vegetation index (NDVI), a measure of the amount of healthy green vegetation on the ground, has been widely used in creating disease risk models. NDVI data, routinely created from several different satellite datasets, have been frequently used as a substitute measure for precipitation data in Africa and other dry regions where rainfall promotes rapid growth of green vegetation.

Using NDVI satellite data from the Advanced Very High Resolution Radiometer (AVHRR), Boken et al. (2008) studied the specific relationship between environmental factors and aflatoxins in Mali and reported: “A moderate relationship was found between aflatoxin amounts and NDVI [levels] averaged for the first 10 days in July, that is, during the early part of the reproductive phase. In addition, the total precipitation and average maximum temperature during the reproductive phase were found to be linked to the post-harvest amounts of aflatoxin in peanut in Mali, Africa.” Boken et al. showed that a statistical association exists between NDVI and aflatoxins; the next step is to model the geographical distribution.

In this project, we will attempt to model the geographic extent of aflatoxins using NDVI, land cover,

¹Richard Pearson’s informative tutorial on species distribution modeling is available here: http://biodiversityinformatics.amnh.org/files/SpeciesDistModelingSYN_1-16-08.pdf.

elevation data, and historical weather data along with known locations of aflatoxins.

METHODS

Environmental data

Environmental datasets at 1 kilometer (km) resolution will be collected and assembled in a GIS. The types of environmental datasets that will be included are described below, but other datasets may be added if available. (See Figure 1 for examples of environmental layers created for Kenya.)

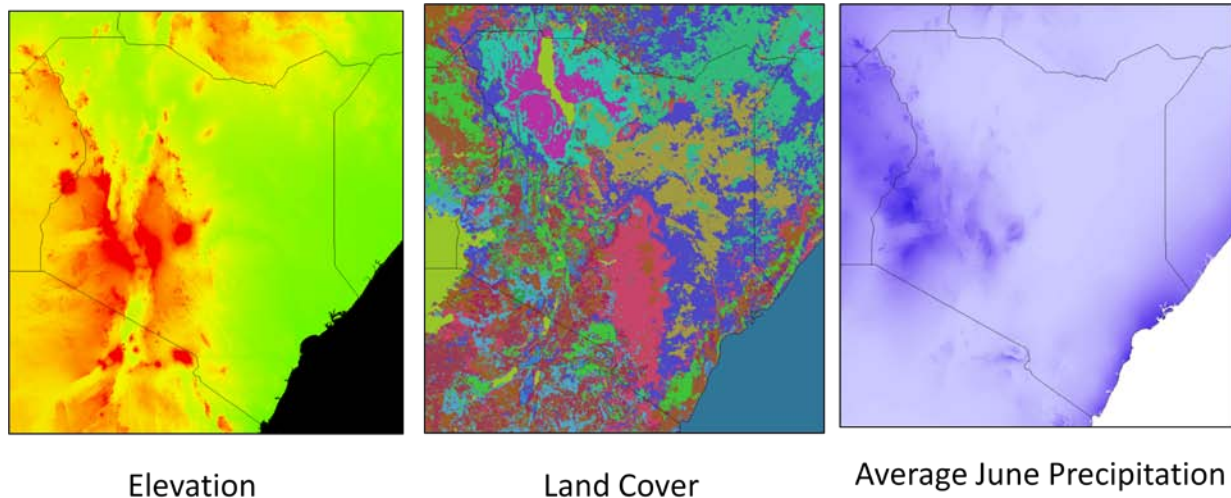


Figure 1. Example of raster environmental datasets to be used in the model

Climate data

Climate data used for modeling will consist of monthly precipitation, mean minimum temperature, and mean maximum temperature from the WorldClim (www.worldclim.org) version 1.4 dataset. WorldClim was created by interpolating mean monthly climate data from ground weather stations from 1950 to 2000. The resulting data layers are in raster format and are available in several resolutions; we will use the 1-km resolution dataset, which is the highest resolution available. For a complete description of how the WorldClim dataset is compiled, see Hijmans (2005).

Vegetation Indices

Normalized difference vegetation index values are calculated from the red and near infrared (NIR) bands of an image using this formula: $NDVI = (NIR - red) / (NIR + red)$. For modeling, we will use low NDVI values prior to harvest as an indicator of drought and high NDVI values after harvest as an indication of humidity during drying and storage.

Vegetation Indices will be obtained from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), which includes NDVI, enhanced vegetation index (EVI), and several visible bands. We will examine EVI briefly to determine whether it can produce a better prediction for the aflatoxins model. However, since EVI was developed for humid regions with high biomass, we anticipate that NDVI will be a better indicator of weather conditions in Africa.

To validate the NDVI, we will attempt to acquire current weather data for the area.

Elevation data

Elevation data will be obtained from WorldClim along with the climate data (see above). The elevation data on the WorldClim website were derived from the Shuttle Radar Topography Mission dataset to match the format, scale, and projection of the climate data layers.

Land cover data

We will download land cover maps from Boston University (www-modis.bu.edu/landcover/) and the US Geological Survey (USGS) (<http://edc2.usgs.gov/glcc/>). The Boston University land cover data are derived from MODIS satellite data while the USGS data are from AVHRR satellite data. Boston University land cover data are more recent than the USGS land cover, but the USGS data may have a different classification scheme that would be desirable to use.

Other data layers

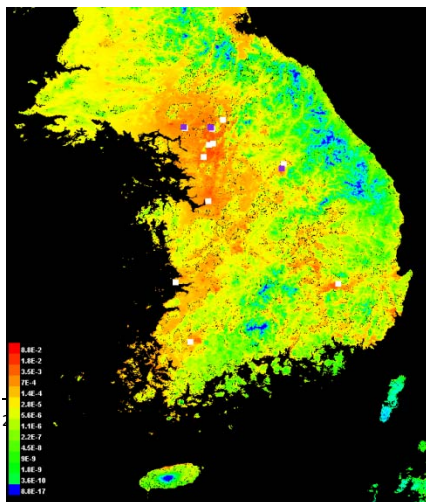
Additional data layers may be added to the model if they are available. The only restriction is that the data layers must be raster and converted to the same extent and projection as the other data layers. For example, it would be possible to add income by district to the model if those data could be converted to a format such as a tiff image.

Aflatoxin data

Aflatoxin data will be obtained from other teams on the project. The specific sampling needs for the model effort are discussed below.

Modeling

The maximum entropy method (Maxent²) will be used to model the aflatoxin geographic distribution for Kenya and Mali. Maxent uses input—a set of raster environmental layers (such as NDVI and elevation) and a text file of known locations of a species—to produce probability maps that predict the potential range of a species. (See Figure 2 for an example of a Maxent model built for *Culex tritaeniorhynchus* in South Korea.) Maxent has been shown to be a high-performing model-building program (Elith et al. 2006) that can create useful models using small numbers of known locations (Hernandez et al. 2006).



<http://www.cs.princeton.edu/~schapire/maxent.>

Figure 2. Example of output map from Maxent showing the predicted probability of the *Culex tritaeniorhynchus* mosquito occurring in South Korea

Note: Red areas have the highest probability, blue the lowest.

Maxent allows a percentage of presence locations to be withheld for testing the accuracy of the model (testing points), and the remaining points are used to build the model (training points). Maxent calculates several measures to test the usefulness of the model. One measure of accuracy is the Receiver Operating Characteristic (ROC) with a measurement of the Area under the Curve (AUC) (Figure 3). ROC and AUC methods have been described by Swets (1988) and Fielding and Bell (1997) and are widely employed in evaluating models (Phillips et al. 2006).

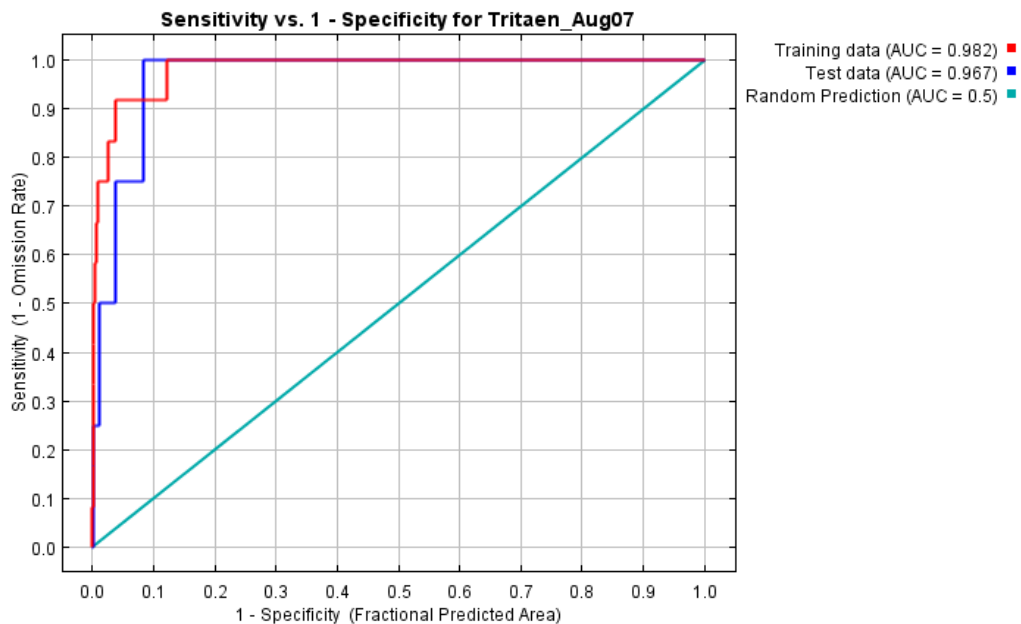


Figure 3. Example of a Maxent program AUC used for evaluating a *Culex tritaeniorhynchus* mosquito model

Note: Curve for the training data used to build the model is shown in red. Curve for the testing data is shown in blue. The aqua line represents the area under the curve that would be produced if the model were totally random wherein. An AUC of 1.0 would represent a perfect model.)

Maxent estimates the importance of the input environmental variables in the model by using a jackknife procedure, which produces multiple models by (1) using each individual variable to create a model by itself and (2) using all variables but dropping one variable at a time to determine the effect on the model. Figure 4 shows the results of the jackknife procedure for the same data model shown in Figure 2.

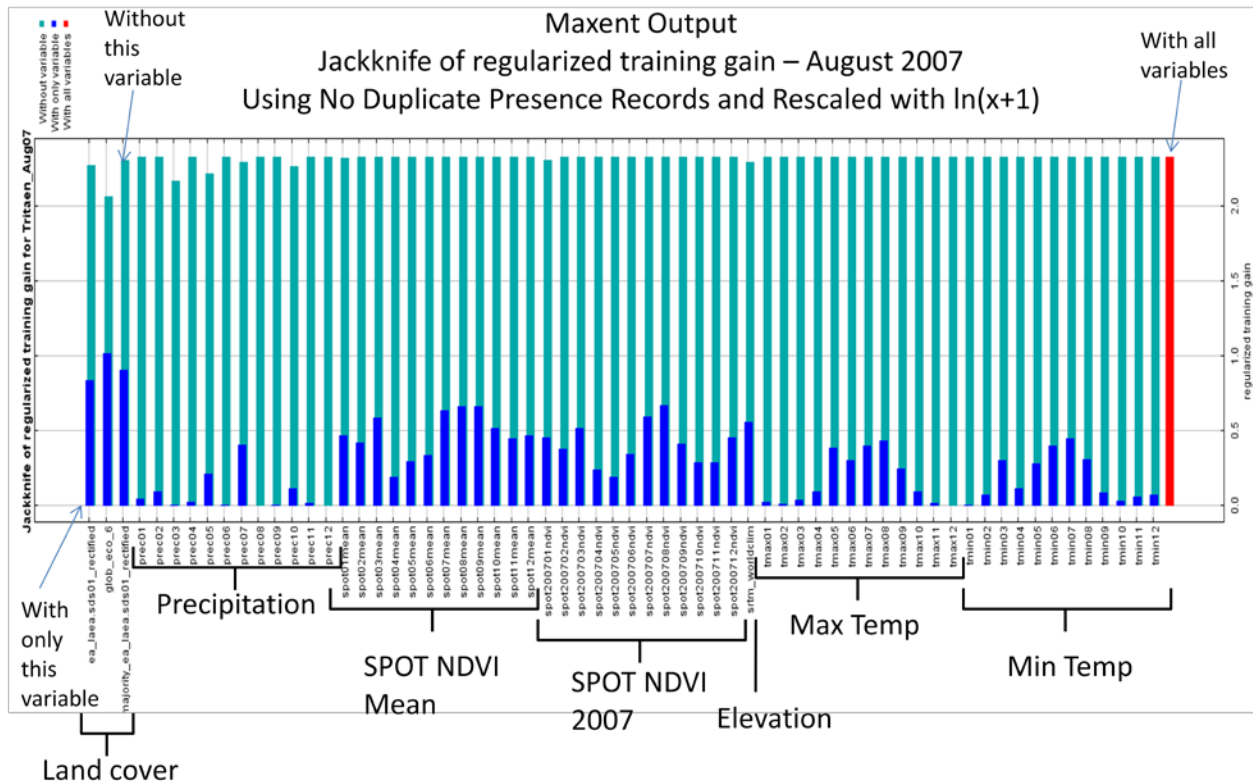


Figure 4. Results of jackknife procedure for *Culex tritaeniorhynchus* mosquito

Note: Red bar shows the model produced using all input variables. Aqua bars show the result of using all variables except one. Where the aqua bars are the same height as the red bar, that particular variable could have been dropped from the model with no significant change to the results. Blue bars show the result of using only one variable in the model. In this particular model, land cover is the most important variable because the blue bars are bigger for the land cover layers.

INPUT FROM OTHER GROUPS THAT WE WILL NEED

To create our models and risk maps, we need the data describe below.

Aflatoxin location data

Aflatoxin location data are a critical requirement for producing the predictive models. We need a minimum of about 30 aflatoxin-positive samples (although 50 or more would be better) for each country and the latitude/longitude of where samples were grown in the field. Latitude/longitude should be accurate within 500 meters (better if possible). Note that a standard unit from a global positioning system (GPS) has an accuracy of about 10 meters.

If aflatoxins are collected from grain in a storage bin, researchers will need to record a GPS point for where the grain was harvested. It is better to go to the actual field where the grain was harvested to record that GPS point, but, if that is not possible, estimates of the distance to and direction of the field from the storage bin are required—for example, “The maize was grown approximately 100 meters to the southeast of latitude X/longitude Y.” We may have to drop samples from the model if they are not located accurately enough, so researchers must make every effort to obtain accurate location

information.

Sampling should be done over a wide geographical area (throughout the whole country, if possible) and within as many different environmental conditions—including levels of elevation, sizes of farm, climates, and so forth—as possible. Samples that lie within 3 kilometers of each other count as one sample. Figures 5 and 6 show some examples of good and bad sampling plans. While the bad sampling plans (often with clustered or low numbers of samples) may be adequate to answer other research questions in the overall project, they will result in poor models.

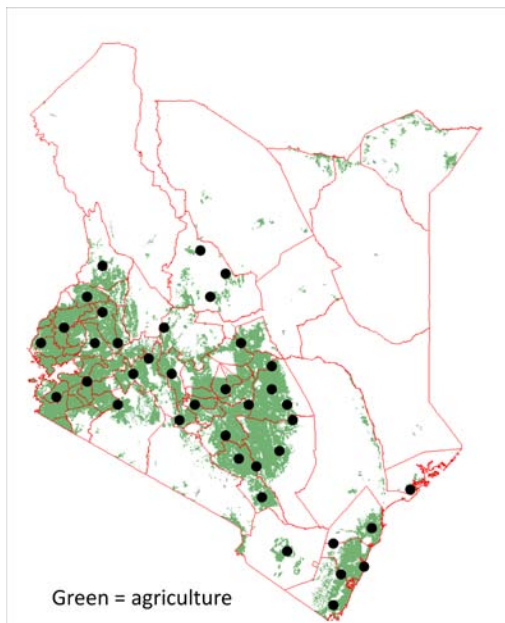


Figure 5. Example of good sampling of aflatoxins data for use in species distribution modeling, with sampling sites shown as black dots

Note: Samples cover a wide geographic area and range of environmental variables (such as different elevations and climate regions). There are more than 30 samples that are positive for aflatoxins.

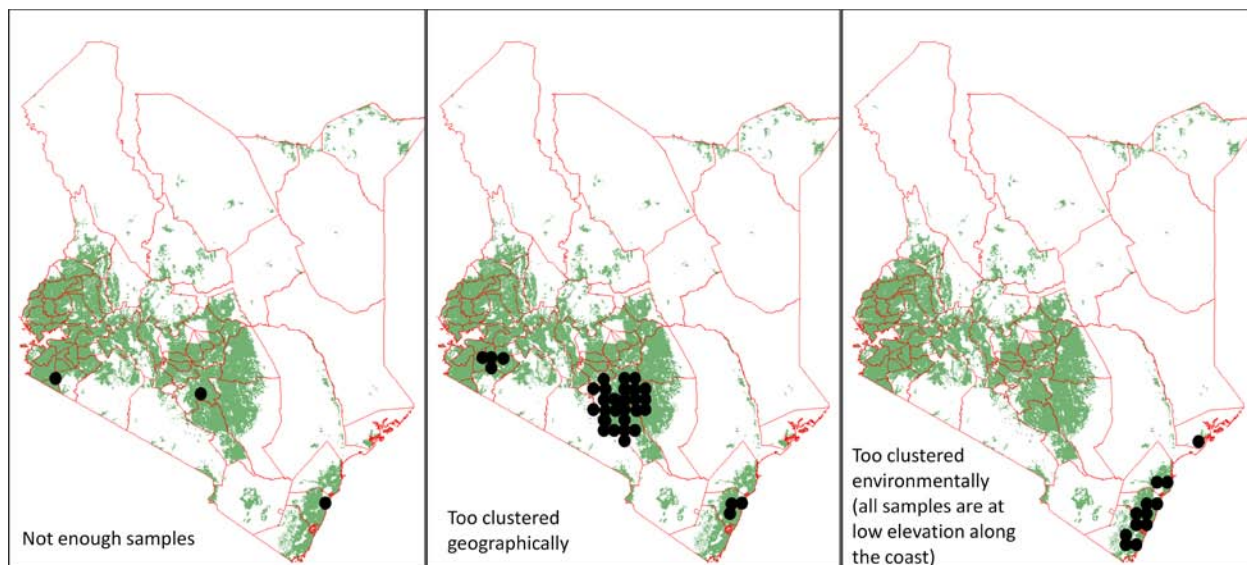


Figure 6. Examples of inadequate sampling of aflatoxins data for use in species distribution modeling, with sampling sites shown as black dots

Negative samples will be important to validate the model, so every sample collected will include a recorded latitude and longitude point. We will use WGS 84 as the datum and latitude/longitude as the coordinate system wherever possible; if this system is not possible, the alternative system of datum/projection/coordinate used by the GPS will be recorded.

Researchers will also record other data that might be useful, including surrounding crops, the irrigation system used (if any), the field size, and so forth.

Table 1 shows a suggested format for collecting aflatoxin data that would be useful for the model project. The important features of the format are that latitude and longitude points are available for each sample and that other relevant data are also collected for each site.

Table 1. Suggested format for collecting aflatoxin data

Sample No.	Longitude (decimal degrees, WGS84)	Latitude* (decimal degrees, WGS84)	Date Collected	Date of planting	Date of harvest	Sample collected pre-harvest or post-harvest	Amount Aflatoxin	Size of field	Type crop (groundnuts or maize)	Surrounding area %	Irrigated
1	39.674512	- 5.452109	7/7/2009	4/10/2009	8/15/2009	pre	Units?	1 acre	corn	100% corn	No
2	39.538263	- 5.298375	9/10/2009	4/1/2009	8/5/2009	post		.25 acre	groundnuts	50% urban, 25% ag., 25% grass	Yes
3									corn	30% water, 10% peanuts, 60% shrub	Hand watered

*Specific percent of area surrounding the field that is urban/village; forest; shrub; grasslands; or used to grow corn, peanuts, or other crops.

Other data needed

In addition to the aflatoxin location data, the following datasets would greatly assist in model development and validation.

Map of known maize and groundnut crops

Although we have a general land cover map derived from satellite data, the map has limitations including fairly low resolution (1 km pixels) and generalization of the land cover categories (with all types of agriculture lumped into one class). It would be helpful to have a better idea of where specifically maize and groundnuts are growing since these are the crops being sampled for aflatoxins.

Weather data

Weather data for our period of study would be extremely useful for this project to help validate the NDVI values that will be used in the model. Currently, we have data averaged for 1950 to 2000.

Harvest dates

Harvest dates would be extremely useful for the modeling effort and will improve our understanding of the environmental conditions that contribute to aflatoxin development.

QUESTIONS, ISSUES, AND LIMITATIONS TO THE STUDY

An obvious issue with this part of the project is that it is highly dependent on obtaining data on the occurrence and location of aflatoxins from the field. Thus, if the field collections take place during a low-aflatoxin year, we may be unable to create the models.

Another limitation is that the study assumes field conditions are a primary cause of the development of aflatoxins. We realize, however, that storage methods are also a contributing factor and that local variations of storage methods may affect the modeling accuracy. If the field teams can obtain information on storage methods, this would be useful.

Since aflatoxins are produced by two species—*Aspergillus flavus* and *Aspergillus parasiticus*—another concern arises in whether or not an attempt will be made to distinguish the resulting aflatoxins found in field samples. The two species may have different environmental requirements that may complicate the modeling.

Finally, for simplicity, we talk above about needing “positive” samples of aflatoxins. We do realize, however, that a continuous range of aflatoxin amounts are likely to be found and that a small, acceptable aflatoxin level may be present in many samples. Unfortunately, the modeling programs we are using are based on presence-vs.-absence data, so we will need to have the team decide some threshold value for what will be considered a “positive” sample.

OTHER MAPPING CONTRIBUTIONS

The GIS/modeling team could contribute to other parts of the project by creating maps for the other teams. For example, given the data, we could map agricultural practices, storage methods, and market locations.

CONCLUSION

This study will use environmental data and known locations (that is, points of latitude and longitude) of aflatoxin occurrence to model and make probability maps of the distribution of aflatoxins. The resulting maps may be useful for determining which areas should be tested for aflatoxins and which areas and farmers could benefit from demonstrations of aflatoxin remediation methods. The maps will also allow us to estimate the amount of land and proportion of crops that might be affected by aflatoxins. The risk maps can be updated yearly as new data become available.

REFERENCES

- Boken, V. K., G. Hoogenboom, J. H. Williams, B. Diarra, S. Dione, and G. L. Easson, 2007. Monitoring peanut contamination in Mali (Africa) using AVHRR satellite data and a crop simulation model. *International Journal of Remote Sensing* 29:1, 117–29.
- Elith, J. C., H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–51.
- Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38–49.
- Guisan, A., O. Broennimann, R. Engler, M. Vust, N. G. Yoccoz, A. Lehman, and N. E. Zimmermann. 2006. Using niche-based models to improve the sampling of rare species. *Conservation Biology* 20: 501–11.
- Hastings, D. A., and P. K. Dunbar, 1998. [Development and assessment of the Global Land One-km Base Elevation Digital Elevation Model \(GLOBE\)](#). *ISPRS Archives* 32: 218–221.
- 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: 773–85.
- 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: 1965–78.
- Jeschke, J., and D. Strayer. 2008. Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences* 1134: 1–24.
- Kapel, C. M. O. 2000. Host diversity and biological characteristics of the *Trichinella* genotypes and their effect on transmission. *Veterinary Parasitology* 93: 263–78.
- Moffett, A., N. Shackelford, and S. Sarkar. 2007. Malaria in Africa: Vector species' niche models and relative risk maps. *PLoS ONE* 2(9): e824.
- Papes, M., and P. Baubert. 2007. Modeling ecological niches from low numbers of occurrences: Assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions* 13: 890–902.
- Pearce, J., and D. B. Lindenmayer. 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology* 6: 238–43.
- Pearson, R. G. and T. P. Dawson, 2003. Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography* 12: 361–71.
- Pearson, R. G., T. P. Dawson, and C. Liu. 2004. Modeling species distributions in Britain: A hierarchical integration of climate and land-cover data. *Ecography* 27: 285–98.
- Pearson, R. G., T. P. Dawson, C. J. Raxworthy, M. Nakamura, and A. T. Peterson. 2007. Predicting species' distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34: 102–17.

Peterson, A. T., and J. J. Shaw. 2003. *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. [*International Journal for Parasitology* 33: 19–931.](#)

Peterson, A. T., J. J. Shaw, R. R. Lash, D. S. Carroll, and K. M. Johnson. 2006. Geographic potential for outbreaks of Marburg hemorrhagic fever. [*American Journal of Tropical Medicine & Hygiene* 75: 9–15.](#)

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. [*Ecological Modeling* 190: 231–59.](#)

Phillips, S. J., M. Dudik, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. In [*Proceedings of the 21st international conference on machine learning*](#), ed. Carla Brodley. New York: AMC Press.

Sanders, T. H., R. J. Cole, P. D. Blankenship, and J. W. Donner. 1993. Aflatoxin contamination of peanuts from plants drought stressed in pod or root zones. [*Peanut Science* 20: 5-8.](#)

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. [*Science* 240: 1285–93.](#)

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

2033 K Street, NW • Washington, DC 20006-1002 USA

T +1 (202) 862-5600 • F +1 (202) 467-4439

Skype: ifprihomeoffice • ifpri@cgiar.org

AUTHORS: Penny Masuoka, Assistant Professor, Uniformed Services University of the Health Sciences (USU), 4301 Jones Bridge Road, Bethesda, MD, United States, 20814

Judith Chamberlin, Adjunct Assistant Professor, Uniformed Services University of the Health Sciences (USU), 4301 Jones Bridge Road, Bethesda, MD, United States, 20814

Maribel Elias, Research Analyst, International Food Policy Research Institute (IFPRI), 2033 K Street NW, Washington, DC, 20006

For more information about the Aflacontrol Project, please visit www.ifpri.org/afla/afla.asp.

Aflacontrol is a collaborative research project funded by the Bill & Melinda Gates Foundation and managed by the International Food Policy Research Institute.

Project Partners: International Maize and Wheat Improvement Center (CIMMYT); International Crops Research Institute for Semi-Arid Tropics (ICRISAT); ACIDI-VOCA; University of Pittsburgh; United States Uniformed Services University of the Health Sciences (USU); Institut d’Economie Rurale (IER); and Kenya Agricultural Research Institute (KARI)

Copyright © 2010 International Food Policy Research Institute. All rights reserved. To obtain permission to republish, contact ifpri-copyright@cgiar.org.

This working paper has been prepared as an output for the Aflacontrol Project (facilitated by IFPRI) and has not been peer reviewed. Any opinions stated herein are those of the author(s) and do not necessarily reflect the policies or opinions of IFPRI, its partners, donors or collaborators, or of the cosponsoring or supporting organizations.