Published in final edited form as:
Ecol Inform. 2011 September ; 6(5): 257–269. doi:10.1016/j.ecoinf.2011.04.004.

Assessing the Application of a Geographic Presence-Only Model for Land Suitability Mapping

Benjamin W. Heumann[1,2,*], **Stephen J. Walsh**[1,2], and **Phillip M. McDaniel**[2]

[1]Department of Geography, University of North Carolina at Chapel Hill. CB# 3220 Chapel Hill, NC 27599-3220, USA

[2]Carolina Population Center, University of North Carolina at Chapel Hill, CB# 8120 Chapel Hill, NC 27516-2524, USA

Abstract

Recent advances in ecological modeling have focused on novel methods for characterizing the environment that use presence-only data and machine-learning algorithms to predict the likelihood of species occurrence. These novel methods may have great potential for land suitability applications in the developing world where detailed land cover information is often unavailable or incomplete. This paper assesses the adaptation and application of the presence-only geographic species distribution model, MaxEnt, for agricultural crop suitability mapping in a rural Thailand where lowland paddy rice and upland field crops predominant. To assess this modeling approach, three independent crop presence datasets were used including a social-demographic survey of farm households, a remote sensing classification of land use/land cover, and ground control points, used for geodetic and thematic reference that vary in their geographic distribution and sample size. Disparate environmental data were integrated to characterize environmental settings across Nang Rong District, a region of approximately 1,300 sq. km in size. Results indicate that the MaxEnt model is capable of modeling crop suitability for upland and lowland crops, including rice varieties, although model results varied between datasets due to the high sensitivity of the model to the distribution of observed crop locations in geographic and environmental space. Accuracy assessments indicate that model outcomes were influenced by the sample size and the distribution of sample points in geographic and environmental space. The need for further research into accuracy assessments of presence-only models lacking true absence data is discussed. We conclude that the Maxent model can provide good estimates of crop suitability, but many areas need to be carefully scrutinized including geographic distribution of input data and assessment methods to ensure realistic modeling results.

Keywords

presence-only; land suitability; agriculture; Thailand; Maxent

[*]corresponding author's contact information: benjamin.heumann@mail.mcgill.ca.

1. Introduction

Understanding land suitability is important for a host of scientific questions and policy applications. Whether for studies involving the spread of invasive flora and fauna, patterns of rare and endangered species, agricultural productivity, or land use dynamics in human-managed landscapes, land suitability analysis has gained increasing favor among social, natural, and spatial scientists, in part, due to recent advances in geographic information sciences, statistical modeling, and spatial analysis. Similarly, niche-based modeling of species distributions and habitat suitability have rapidly developed in part due to the greater availability of geo-spatial data and the need to model ecological phenomena across the landscape, particularly, in response to land use dynamics, climate change, species invasion, and conservation biology. With the development of niche-based models, a suite of new applications have been developed to predict geographic outcomes from environmental data.

Niche-based models link environmental and geographic space by linking species locations with environmental conditions and then geographically-projecting where the species are likely to be found based on suitable environmental conditions (Guisan and Zimmerman 2000; Austin 2002; Dudik et al. 2004; Guisan and Thuiller 2005; Peterson 2003; Hirzel and Le Lay 2008). There are a variety of models and techniques that have been used to assess and predict habitat patterns and species distributions including climatic-envelop models, statistical models, such as Generalized Linear Models and Generalized Additive Models, and machine-learning algorithms such as Genetic Algorithm for Rule-set Production - GARP (Stockwell and Peters 1999) as well as Maximum Entropy - MaxEnt (Phillips et al. 2006, 2009). While numerous models exist (see Elith et al. 2006 for a review), there is a consensus that machine-learning algorithms are consistently among the best performing models (Elith et al. 2006; Hernandez et al. 2006; Ortega-Huerta and Peterson 2008). Such models optimize their performance based on iterative calculations that use defined geographic positions of target species and their environmental conditions to calculate habitat suitability or the likelihood of species presence (Dudik et al. 2004; Phillips et al. 2006; Stockwell and Peters 1999). The development of these models has resulted in new methods that may be applicable to other geographic applications such as land suitability analysis. Of interest here are the presence-only models, such as MaxEnt, that do not require true-absence data, which is often difficult to define or collect, particularly, in geographically inaccessible areas as well as in many parts of the under-developed and developing worlds.

This paper describes and assesses a new application of presence-only modeling – agricultural crop suitability mapping. The application of a machine-learning, algorithm-based model designed to estimate the likelihood of occurrence, based on presence-only data, has great potential for use, particularly, where extensive land use information is often difficult to obtain. Crop suitability research has a strong tradition of integrating expert and traditional knowledge with environmental data using analog techniques including soil surveys and aerial photo-interpretation, and more recently, a geographic information system (GIS) approach to integrate local site conditions and regional settings of ecological parameters (Bandyopadhyay et al. 2009; Joel et al. 2009). Multi-criteria evaluations (Pereira and Duckstein 1993; Ceballos-Silva and Lopez-Blanco 2003; Rahman and Saha 2008), fuzzy classification techniques (Groenemans et al. 1997; Burrough et al. 1992; Joss et al. 2008; Kurtener et al. 2008), and simulation models (Littleboy et al. 1996; Manna et al. 2009) are commonly used to measure the suitability of extant distributions and to predict suitability across the landscape. These techniques require either a detailed *a-priori* knowledge of the required conditions (i.e., multi-criteria evaluation and simulations) or absence data (i.e., multi-criteria evaluation and fuzzy classification).

Novel presence-only, machine learning algorithms are potentially well-suited to the type and quality of existing data often used in land suitability analysis. In such studies, the landscape is generally sampled by the placement of field plots, often used to calibrate and validate a remote sensing image classification. In such studies, the field data are not necessarily collected with modeling in mind. In this study, we focus on an agricultural landscape in Northeastern Thailand and the application of the MaxEnt model to assess crop suitability in a former agricultural frontier. The study area is well positioned to serve our project needs. Because of our long history of work in the region, we have assembled a rich longitudinal, demographic database of household characteristics, location of nuclear villages of survey and non-survey villages, demarcation of village territories indicated by household use patterns, and land use/land cover information for household field plots for the year 2000. Through previous efforts, and in support of this paper, we have developed an array of GIS coverages derived to assess geographic accessibility and resource endowments, as well as an extensive satellite image time-series that has been classified into land use/land cover types and validated in the field for contemporary imagery and through the interpretation of aerial photography for historical periods (Walsh et al. 2005).

1.2 Research Aims

The aim of this paper is to assess the application of presence-only data for crop suitability modeling, using the MaxEnt, model for upland cassava and lowland paddy rice varieties in a human-managed landscape in Northeastern Thailand. First, we briefly review the assumptions of our modeling approach. Second, we determine crop suitability using the geographic presence-only model, MaxEnt, for the dominant lowland and upland crops in the study area. Third, using independent data sources, we explore the difference between model outputs based on the sample size and distribution of input data. Model assessment is conducted using the Receiver Operating Characteristic (ROC) and Cumulative Distribution Functions (CDF) plots within and between crop datasets.

1.3 Study Area Nang Rong District, Northeastern Thailand

Nang Rong district, located in Buriram province, occupies approximately 1300 km^2 in Northeast Thailand (Figure 1 – Note that the location of actual study villages has been omitted for confidentiality reasons related to Human Subjects research). The district is geographically positioned in the southwest portion of the Khorat Plateau, a wide and shallow basin that is underlain by Cretaceous sandstones, shales, and siltstones, intruded in places by Tertiary basalts. The dominant occupation in the region is farming and the majority of farm households own on average three hectares of land (Ghassemi et al. 1995). Per capita income in the Northeast is the lowest in the country, largely because of low and unstable agricultural productivity resulting from inconsistent monsoonal rains and generally poor soils and inadequate drainage (Parnwell 1992). Nearly one-third of the region is unsuitable for successful cropping due to steep topography or soil laterization (Parnwell 1988), and only one-third of the land area is suitable for the cultivation of rice, though yields are only moderate. In the upland settings, where cassava and sugarcane dominate, the soils are susceptible to erosion. Throughout the region, the soils are generally infertile with high levels of salinity or acidity (Ghassemi et al. 1995). The exception is the alluvial soils of the lowlands that support relatively high yields of rice, limited, however, by flooding caused by extreme monsoonal rains. Over 80-percent of the average annual precipitation that affects the Northeast occurs as unevenly distributed torrential rains that occur between April and November (Kaida and Surarerks 1984). For the remainder of the year, soil moisture deficits are common and droughts and floods are a persistent threat to agriculture (Fukui 1993). Topography and agricultural suitability are explicitly linked, particularly, to the biophysical characteristics of the region s landforms, i.e., hills, uplands, high-, medium-, low-terraces, and floodplains. Figure 2 shows the idealized cropping systems, agro-geomorphic settings,

land use, and resource limitations and gradients between the uplands and lowlands in support of agriculture.

1.4 Social-Ecological Context of Nang Rong District, Northeastern Thailand

The people of the Northeast live in nuclear villages in which household dwelling units are clustered in space and associated agricultural lands are arrayed in a general circular pattern surrounding the village centers. Land parcels of associated farm households are generally discontinuous in space. Villages were historically located in close proximity to rivers to support the dominant crop, rain-fed paddy rice, for subsistence agriculture, but now villages include upland area and practice a mixed subsistence and commercial cultivation. For lowland villages, rice paddies dominate the landscape. The highlands are dominated by forest remnants resulting from deforestation that began in the district in the late 1960s and early 1970s. Through agricultural extensification, cassava, and to a lesser degree sugarcane, corn and jute, were cultivated by Thai farmers who began a transition from subsistence agriculture to a mix of subsistence and commercial agriculture by cultivating forested uplands (Walsh et al. 1999). The upland field crops, primarily cassava, have very different resource requirements than lowland paddy rice. In addition, the upland field crops can be left in the field for a period of time without significant yield declines, thereby, affording the opportunity to focus labor, time, and energy on lowland paddy rice during planting and harvesting periods. Today, the isolated parcels of upland field crops have coalesced into extensively cultivated fields that are suitable for mechanized agriculture. Between the uplands and the lowlands, high-, medium-, and low-terraces exist that extend the field crops “down” into the high-middle terraces and the lowland paddy rice “up” into the low-middle terraces, depending upon environmental conditions, geographic accessibility from nuclear villages, and crop prices. The terraces are considered transitional areas between the upland field crops and the lowland paddy rice.

1.5 Assumptions, Simplifications, and Limitations

The MaxEnt model is a niche-based model that assumes the distribution of observations, i.e., presence data, represents the realized niche. Although the niche concept (Hutchison 1957) is central to ecology and biogeography, considerable debate remains about definitions of the niche and how the niche concept is applied to geography and space (Pulliam 2000; Soberon 2007; Hirzel and Le Lay 2008; Araujo and Guisan 2006; Peterson 2003; Jimenez-Valverde et al. 2008). This is particularly important given the rapid increase and broad availability of species data and the methods for deriving species distributions and habitat models. At the very outset, niche-based modeling requires a clear statement of how the niche concept is defined and applied (Guisan and Thuiller 2005; Austin 2007). In a human managed ecosystem, such as Northeastern Thailand, the niche concept must be re-defined. In such environments, the realized niche for crops is described by the geographic position of their cultivation, based on the modified environment, for instance, irrigation practices or the geographic distance from the village to the agricultural plots of households, as well as socio-economic conditions such as market prices and labor availability.

In our use of the MaxEnt model, we have made the following assumptions and simplifications, given our project objectives and the available data: (1) agricultural crops are grown more often in better suited locations relative to other crops. Rain-fed paddy rice is best suited to lowland settings that are accessible to sufficient water and alluvial soils, whereas cassava grows best on drier, upland settings. Thus, the MaxEnt species distribution model is appropriate for this analysis as best” and “worse” geographic settings are used to describe the primary rice vs. cassava suitability; (2) human management is either systematic (i.e., use of upper terraces for cassava and alluvial areas and low terraces for rice based on geomorphic setting and elevation, and fertilizer use based on soil constraints) or random

(i.e., household application of pesticides and access to household labor) across space. As such, human management is captured as a co-variant of existing environmental conditions or it is viewed as a stochastic process; (3) land use decision-making by households is a process that is influenced by environmental settings and conditions, socio-economic factors, and demographic considerations. Our model does not explicitly consider the socio-economic and demographic aspects of crop choice, but like other factors of human management, we assume that these choices are either systematic or random across the study area; (4) since human-management factors, such as commodity prices, as well as environmental factors, such as the timing and amount of monsoonal rains, can vary over time, our model applies only to conditions in the year 2000, a normal to wet year in Northeastern Thailand and average crop prices on the local and global markets; and (5) Just as ecological niches have regional variations of soils and climate (Murphy and Lovett-Doust 2007), socio-ecological niches also have regional variations of culture and context and thus our model results are not directly transferable beyond our study region, however, the applicability of the general approach to modeling crop suitability using the MaxEnt model in human managed landscapes is suggested through the subsequent project findings.

2. Methods

2.1 Crop Location Data

Three sources of crop data are used in this study (Table 1): (1) “Form 6 Social Survey” (Form 6) data collected in year 2000 through a household questionnaire that asked farmers what they planted on specific “plangs,” i.e., land parcels used by farming households, (2) a “Remote Sensing Classification” (Remote Sensing)” of Landsat TM data, and (3) “Field Data from Geodetic Ground Control Points” (GCP), over 100 points distributed at strategic locations, e.g., road intersections and tributary branches of rivers and creeks, throughout the district and collected using differentially-corrected, GPS technology. All data were collected in 1999–2000.

2.1.1 Household Social Survey - “Form 6”—Household questionnaire “Form 6” is the portion of the social survey that deals with land use decisions of households on land parcels that are geographically described on cadastral maps. In Thailand, these household land parcels are commonly known as *plangs.* The questionnaire was administered to all households in 92 study villages in Nang Rong District in year 2000 (Rindfuss et al. 2003); nearly 10,000 households were surveyed, land parcels were link to the households that used them in that year, and nuclear village centroids were defined by collecting GPS coordinates for all survey villages. Included in the questionnaire were details about land use, such as the type (i.e., rice or cassava) and variety of crops grown (i.e., jasmine, sticky, or heavy rice) in 2000 on a specific *plang* by a specific household of known social and demographic characteristics. In addition, farming practices as well as the names of the users of neighboring plangs were also collected as part of the social survey. In Nang Rong District, farmers' dwelling units are separate from the *plangs* that are used for the cultivation of crops: generally, farmers living in nuclear village settlements do not live on the land that they farm, but rather, their land parcels are geographically arrayed in a discontinuous fashion around the village, with some parcels up to 5-km (or more) away from the village centroid. *Plangs* used by households were geo-located using existing cadastral maps and integrated with other data assembled within a geographic information system. In addition, community participatory mapping was used that involved large-format, contemporary, panchromatic, aerial photography and GIS overlays of roads and other landmark features for recognition by discussion groups, composed of village headman, hunters, long-term farmers, and other key informants who had an extensive knowledge about the land and the use of *plangs* by households in his/her village and nearby-villages.

Although households were surveyed about the cultivation of cassava on their corresponding *plangs* on Form 6, cassava was divided into two types based on its location and function. “Plantation” cassava is grown in large upland fields as a cash crop, often far from villages, on marginal rice land, and harvested through mechanization. “Village” cassava is grown in small garden plots in or near villages and often in the lowlands that are dominated by rice cultivation. These forms of cassava were separated in our analysis to test whether certain conditions exist for “plantation” versus “village” cassava. Areally and economically, “plantation” cassava plots dominate the landscape and they substantially affect household wealth and assets in villages that can access upland areas. In broad upland areas, individual household plots coalesce into extensive cassava uplands that are suitable for mechanized agriculture. “Village” cassava are smaller plots maintained near the household s dwelling unit and used for food for local family consumption, service local markets, and as feed for livestock that are locally maintained.

2.2.2 Remote Sensing Land Cover Classification—Relying upon an assembled Landsat TM classification and an image time-series, a classified image that most suitably matched the date of the social survey was selected for inclusion in this analysis. The remote sensing classifications were derived through a multi-phase hybrid classification designed to examine generalized land use/land cover in Nang Rong District for all image dates of an assembled Landsat MSS and TM time-series that extends from 1972 to 2006 and includes nearly 40 images. The classification approach was designed to be repeatable across all images in the time-series, rely upon image characteristics derived through statistical measures, limit the reliance on *in-situ* data because of the antecedent nature of the image time-series and the difficulty of validating historical land use/land cover classification, and integrate acquired aerial photography with expert knowledge of the landscape by local informants for strategic dates in the image time-series that corresponded to important environmental events (e.g., significant floods or droughts) and noteworthy development issues in the District (e.g., construction of new roads or the improvement of gravel roads to all-weather roads) (Messina et al. 2001; Walsh et al. 2005).

The classification procedure followed an approach called cluster busting that is designed to produce consistent land cover classifications over time for change detection and monitoring (Jensen et al. 1993). The process began by performing an unsupervised classification that relied upon the ISODATA decision-rule to define 100 “naturally” occurring spectral classes that were subsequently reduced to approximately 30 classes through the interpretation of the transformed divergence statistics, generated as output from the classification process. Iterative testing was conducted to estimate the number of classes for the initial unsupervised classification by increasing and decreasing the estimated number of “naturally” occurring spectral classes by 25% from the initial test group of 75-classes. Test classifications were conducted for a high of 200 “naturally” occurring spectral classes to a low of 50 “naturally” occurring spectral classes, thereby, arriving at 100 spectral classes for the initial unsupervised classification. The classifications, generated to determine the number of classes for the unsupervised approach, were conducted for a 1994 Landsat TM image of high quality, i.e., relatively free from clouds/shadows that closely matched the date and areal coverage of the 1:50,000 scale, vertical, panchromatic aerial photography of the District. After each unsupervised classification, it was compared to the land use/land cover information for three study villages that we had routinely visited in the field, possessed good air-photos, and represented a diversity of landscape conditions and land use/land cover types that we sought to characterize. Informal assessments were conducted that involved the generation of error matrices of 150-random points, i.e., observed (classification) vs. expected (interpretation of aerial photography), to guide us in the selection of 100 spectral classes for the initial unsupervised classification and for the classification of subsequent images in the Landsat time-series, including the image used in this analysis.

Following the unsupervised classification, a supervised classification was applied using the maximum likelihood classifier to relate unclassified pixels to the 30 spectral classes (i.e., the training data) defined through the unsupervised classification. The approach allows for the generalization of classes to a few key land use/land cover types as well as the expansion of cover types to additional classes as details warrant. The land use/land cover classes that were generated included high density forest, low/medium density forest, scrub/shrub, upland field crops, lowland rice, bare/barren, water, built-up/urban, and wetlands. For this study, we assume that upland crops are cassava, by far, the dominant upland field crop in Nang Rong.

Accuracy of the classification was assessed using independent ground control points. Table 2 shows the classification error matrix. The overall accuracy is 88.7% (kappa = 0.833). The largest errors were commission errors of forest and omission errors of built land. The upland class (cassava) has larger errors than the lowland rice class, although the error was only 15% or less. However, this level of error should be noted when interpreting results based on these data.

2.2.3 Field Data—To support the remote sensing project objectives, specifically geometric ortho-rectification, geodetic control points (GCPs) were collect in Nang Rong in February 2000. The GCPs were collected to provide geometrically-correct Landsat imagery using a systematic spacing of sample locations positioned at road intersections across Nang Rong. The land use/land cover occurring at those locations were recorded in eight cardinal directions in the field and photographed in each direction to assist with the geo-location of GCPs in the imagery. We have adapted the network of over 100-GCPs as crop presence data, based on the observed land use/land cover at each referenced location. The GCPs are independent of the remote sensing land use/land cover classification, as they were used only for geometric correction and were not included as part of the classification training data. Analysis of historical precipitation records show that the timing and amount of rainfall during the monsoons from 1999 – 2001 were average to conditions over the preceding 50-years.

2.2 Environmental Data

2.2.1 DEM—The digital elevation model (DEM) was derived from the digitization of all 10-meter topographic contour lines on the 1:50,000 scale Thai military base maps. As a supplement to the contours, a total of 1,373 spot elevation points, collected from the topographic maps, were used to obtain a finer level of detail. ArcInfo s Topogrid was used to integrate the contour lines with spot elevations and hydrographic enforcement in the construction of the DEMs. All topographic sinks were removed from the dataset.

2.2.2 Soils—The soils data are derived from two 1:100,000 scale Thai military base maps. A 1974 soil survey map, created using aerial photography and field verification, contains detailed soil characteristics, but the data do not cover the entire study area. A 1992 topographic/geomorphic map covers the entire study area, but contains less descriptive information and less spatial detail than the 1974 soil survey. A combined dataset was created primarily using the 1974 soils map. The 1992 topographic maps were compared to the soils map to extend the areal extent of soils data. In the areas in which the 1992 topographic/ geomorphic map are used, the spatial accuracy is less than that of the 1974 map, but suitably comparable for this application.

2.2.3 Solar Radiation—Solar radiation load was calculated from the DEM using the Area Solar Radiation Tool in ArcGIS Spatial Analyst. Solar radiation was calculated for the growing season (i.e., June through January), using the default settings for all other parameters.

2.2.4 Other Variables—Several other variables were tested for model selection, but they were excluded due to either a high co-variance with existing variables or a low model contribution (<5%). These variables included slope angle, slope aspect, topographic wetness index, and soil characteristics such as pH, Phosphorous, Nitrogen, saturation, and texture.

2.3 Modeling Techniques

2.3.1 Description of the Maximum Entropy Model (MaxEnt)—MaxEnt is a presence-only model that uses randomly generated background data to distinguish the pattern of species occurrence from the random environment. Both categorical and continuous data are used. Data are fit using linear and non-linear functions and different functions can be hinged together. Unlike GARP (Genetic Algorithm for Rule-set Production) and BRT (Boosted Regression Trees), MaxEnt is a deterministic model when provided with consistent input parameters. MaxEnt has been found to be consistently among the best methods for niche-based geographic species distribution modeling, and performs especially well for small data sets (Elith et al. 2006; Hernandez et al. 2007). MaxEnt also provides useful model assessment tools such as jack-knife environmental parameter contribution, species-environment curves (with and without other environmental parameters), the “Area Under the Curve” of the Receiver Operating Characteristic (AUC-ROC) as a metric of model performance, and a user-friendly graphical user interface and command line functions.

2.3.2 Modeling Procedure—In this study, MaxEnt is used to model crop suitability. A variety of conditions were modeled and compared. All models had the following characteristics: (1) 1,000 runs were made using a random seed to partition crop locations into training and testing data sets to generate confidence intervals of model output; (2) environmental parameters included continuous surfaces including elevation, categorical soil type, and solar radiation; (3) all models partitioned the crop presence data using a random 50/50% split for training and calibration; and (4) model parameters were set to defaults, following extensive experimentation.

To assess and understand the performance of the MaxEnt model, the following experiments were tested: (1) crop suitability of rice and the main varieties in the study area and cassava (Plantation and Village), using only the Form 6 data, and (2) differences in model selection and output between different crop presence datasets – the “Form 6” social survey, the “remote sensing” Landsat classification, and the “GCP” field data for rice and cassava to better understand model accuracy and sensitivity to input data.

2.3.3 Analysis—Model output was compiled in the computing environment MATLAB. The MaxEnt model outputs a map of occurrence probabilities, and tables of model selection (e.g., variable contribution to the model) and the AUC-ROC for the training and testing data sets. The mean and the 95-percentile range of the 1,000 runs for habitat suitability are mapped. Variable contribution and AUC are displayed as standard box-plots, with the central mark indicating the mean, box edges at the 25th and 75th percentiles, and whiskers illustrating the data limits (~99.7 percentile), without consideration of outliers. Additionally, suitability is compared with a sample of input data and between models, using empirical cumulative distribution function plots.

3. Results

3.1 Crop Varieties - Rice

Figure 3 illustrates model results for rice suitability. The suitability for jasmine rice (Figure 3B) was medium to high across most of the lowlands with the highest suitability scores

occurring in a patch located in the northeast corner of the study area. The distribution of suitability for heavy rice (Figure 3C) was similar to that of jasmine rice, but with a distinct pattern along the lower portion of the floodplain. In contrast, sticky rice had a distinctly different suitability distribution (Figure 3D). Overall, variability in suitability scores among model runs was low, although individual soil polygons have moderate variability (Figure 4A–D).

The contribution of each environmental variable is illustrated in Figure 5(A–C). Elevation was the dominant variable contributing to the suitability of jasmine rice and heavy rice. Heavy rice was the most affected by solar radiation. Sticky rice differs from jasmine and heavy rice as soil type and elevation have a similar contribution to the suitability scores and patterns. Model performance was good, AUC-ROC was well above random (ROC = 0.5), and consistent for all models between training and testing data (Figure 6A–B). The pattern of model performance was interesting in that the major components of the composite rice model, jasmine rice and heavy rice, performed significantly worse than the composite model.

3.2 Crop Varieties – Cassava

Cassava was most suitable in the uplands that are located in the south and southwest portions of Nang Rong District (Figure 3E–F). Although the composite cassava suitability was dominated by the plantation cassava, the village cassava (Figure 3G) did have a subtle impact on suitability in the lowlands in the north portion of the study area. Similar to the rice models, elevation was the dominant environmental parameter in terms of model contribution (Fig. 5D). Soils were more important than solar radiation (Figure 5E & F). Model performance, as indicated by the AUC-ROC value, was good (> 0.8). Plantation cassava has a higher AUC-ROC than the composite or village cassava, likely due to the smaller environmental envelop and higher sample size (Figure 6C & D).

Variation of suitability was generally lower for the composite and plantation cassava models, although there is moderate variability (95% range of 30) in several patches in the southwest portion of the study area (Figure 4E–G). Village cassava was much more variable with 95% ranges, nearly 40 throughout and several patches around 60. This pattern was consistent with model contribution (Figure 5D–F). The composite and plantation cassava models have lower variability between each run, while the village cassava models are highly variable. For instance, the contribution of elevation varied between ~80% and 30% for the village cassava models compared to 70% and 55% for the composite cassava models. The AUC-ROC values were also more variable for the village cassava models, although model performance remains acceptable even at the lowest outliers.

3.3 Crop Varieties: Comparisons of Suitability

Figure 7 compares the suitability of each variety of rice and composite cassava across the other crop type locations from Form 6. Each Cumulative Distribution Function (CDF) plot illustrates the cumulative proportion of crop locations from Form 6 on the y-axis with suitability equal to or lower than the x-axis. CDF curves that pass through the lower right quadrant indicate proportionately higher suitability, while curves that pass through the upper left quadrant indicate proportionately lower suitability. Figure 7A shows the CDF of jasmine rice suitability for jasmine rice, sticky rice, heavy rice, and composite cassava plots. The CDF curves for all varieties of rice show moderate suitability (e.g., 20% of locations have a suitability between 40 and 50, and 50% of locations have a suitability between 50 and 60). The CDF curve of composite cassava locations show poor suitability, with 70% of locations showing suitability scores less than 40. These patterns are similar for heavy rice (Figure 7B).

The CDF curves diverge for sticky rice suitability (Figure 7C). Sticky rice locations have a greater proportion of higher suitability scores, followed by jasmine rice and heavy rice. Composite cassava plots are less suitable for sticky rice than for heavy rice varieties, with 90% of locations with a suitability score of 40 or less. Note that the top 10% of locations where sticky rice is grown has a suitability score greater than 85, while the bottom 40% of locations have a suitability score of 50 or less. The CDF curves for composite cassava show the inverse pattern of the rice varieties (Figure 7D). As expected, the plots where cassava is grown are more suitable than where rice is grown. In sum, the difference between jasmine rice and heavy rice is minimal, while sticky rice and cassava have distinctly different growing environments.

3.4 Data Sets - Rice

Modeled crop suitability varied greatly between the rice data sets (Figure 8A–C). The Form 6 model (Figure 8A) illustrates a structured pattern that followed elevation and soil, while the remote sensing model (Figure 8B) shows a homogenous, moderate suitability for all areas, except for the upland locations. The GCP model (Figure 8C) showed distinct suitability patterns based on soil type. The differences correspond to the level of contribution to the model of each environmental variable (Figure 5G–I). For the rice model, using the Form 6 survey data, the contribution to the model is primarily elevation (~60%), followed by solar radiation (~25%), and soil type (~15%). While elevation is the primary contributor to the remote sensing-based model, soils and solar radiation contribute very little. Conversely, soil type is the dominant factor for the GCP data set. The variability in crop suitability scores are similarly low for the large, widespread Form 6 and remote sensing data sets (Figure 9A and B, respectively), but the smaller GCP data set (Figure 9C) has considerable variability in several areas.

3.5 Data Sets - Cassava

Figure 8(D–F) maps the suitability of cassava using the Form 6, remote sensing, and GCP data sets, respectively. All three models predict higher suitability scores for cassava in the uplands over the lowlands, but the spatial pattern of the suitability scores differs. The Form 6 model (Figure 8D) shows higher suitability scores in a large patch positioned in the southwest portion of the study area as well as flanking the slopes of the higher volcanic hills in the south of the district. The remote sensing data set (Figure 8E) shows the highest suitability scores at the top of these hills and moderate suitability scores around the lowland villages. The GCP data set (Figure 8F) shows a similar pattern as compared to the Form 6 and remote sensing models, but with distinct polygons being represented from the soils data set. The variability of cassava suitability (Figure 9D–F) between the data sets is similar to the rice data sets. Overall, elevation is the dominant suitability factor, although for the remote sensing data set, soil type is also a major contributor.

3.6 Data Sets - Model Evaluation

Figure 6(D–G) illustrates the AUC-ROC for each of the models. The Form 6 models have the highest and most consistent AUC-ROC values. For the rice models (Figure 6E–F), the GCP model performs better than the remote sensing model. Interestingly, the variability of AUC-ROC does not substantially vary between models, despite larger differences in the sample size between the GCP and other models. For the cassava models (Figure 6G,H), Form 6 has a higher AUC on average, although for individual runs, the remote sensing model may have performed better as indicated by the box whiskers. However, some runs are also worse than a random model (AUC < 0.5). This high variably illustrates the importance of training and testing data.

Figure 10 illustrates the CDF of suitability scores predicted for crop presence locations from Form 6, remote sensing, and GCP data sets. Overall, the distribution of rice was good, with a strong delineation between locations with suitability scores less than and greater than 50. Conversely, the cassava curves show that many of the Form 6 and remote sensing locations have low predicted suitability. Interestingly, while the Form 6 and remote sensing data sets have similar distributions for both rice and cassava, the Form 6 model actually predicts suitability best for the GCP data.

4. Discussion

4.1 Crop Varieties

Nang Rong initially developed as an agricultural frontier where rain-fed, lowland paddy rice predominated on alluvial soils and in areas adjacent to perennial rivers and streams. In the uplands, relatively far from reliable sources of water, dry dipterocarp forests dominated until a demand for high calorie animal feed in Europe transformed the uplands of the Northeast through deforestation to support the cultivation of cassava, and to a lesser degree sugarcane. The topographic gradient from the cassava uplands to the paddy rice lowlands is punctuated by a series of high to low terraces that serve as a transitional landscape subject to resource, climate, and price variations. This transition is evident from our results. Elevation is the most important environmental factor in almost every model. Rice, including each variety, is most suitable in the lowlands and cassava most suitable in the uplands. But in the middle-high terraces, the difference between rice suitability and cassava diminishes. This suggests that either rice or cassava can be grown in these locations depending on a variety of factors that vary spatially and temporally, given environment and socio-economic variation.

The Form 6 data set is the largest, most detailed, and most accurate crop data set available for the study area. Model results from the Form 6 data show that while there is a relatively clear pattern of suitability scores between cassava and rice, the differences between the rice varieties are less distinct. It is important to note that the MaxEnt model predicts the likelihood of occurrence for a single crop (i.e., potential suitability) at a moment in time, even though the distribution of a given crop is relative to the distribution of other crops.

While interpretation of high and low crop suitability scores is clear, scores around 50%, widely observed for the rice varieties, are more difficult to interpret. There are several potential interpretations. First, the most straight forward interpretation is that in areas of medium suitability farmers select between crops, not due to environmental conditions, but rather due to socio-economic factors (e.g., farming tradition). Another interpretation calls into question the ability of the model to discern between high and low suitability, indicating possible missing data in the model. For example, missing data can lead to ambiguous results. Given the widespread occurrence of these crops, the inability of the model to detect high suitability in the lowlands may be due to the overwhelming presence of these rice varieties across the observed environmental space, with the exception of higher, upland elevations. Previous studies using MaxEnt have shown that the model performs exceptionally well for rare or cryptic species (e.g., Hernandez et al. 2006; Ortega-Huerta and Peterson 2008). It is likely that if the study area were expanded to include substantial areas in which rice cultivation does not occur, the model may better distinguish between high and low suitability and the associated AUC-ROC scores would be higher.

4.2 Sample Size

Sample size has a clear impact on the performance of the models, in terms of variability of suitability scores and patterns between model runs as well as AUC-ROC scores. The models that were generated with more observations tend to have lower variability between model

runs, i.e., suitability scores, variable contribution, and AUC-ROC. Overall, this trend is consistent with the findings of Trani (2002) and Stockwell and Peterson (2002). However, the models run with smaller data sets (n ~ 100) are highly variable in their suitability scores and AUC_ROC compared to the larger data sets (n > 1000). Stockwell and Peterson (2002), Hernandez et al. (2006), Papes and Gaubert (2007), and Williams et al. (2009), however, report consistent models with a sample size far less than 100. The difference is likely due to the distribution of the study organism and whether the species is a specialist or a generalist. Specialist organisms and particularly rare species have a limited environmental envelop in which they occur, and thus a smaller sample size is likely to capture that envelop. For a widespread generalist species, a larger sample size is required to capture all the possible environmental conditions in which a species may occur. The use of 1,000 random permutations to partition the data in our models for training and evaluation suggests how smaller sample sizes of widespread organisms can produce variable model results.

4.4 Sampling Distribution

The areal distribution of observation points has a clear impact on model output, as briefly discussed in Section 4.1. The effects of areal distribution relate to environmental and geographic space. For instance, the remote sensing dataset served as an alternate approach to compare Form 6 data. Since the remote sensing classification was district-wide, a geographically-based, random sample could be selected. The social survey data (Form 6), however, is highly clustered and limited in its areal extent as a consequence of the location of the study villages. Results from the Form 6 rice and cassava models demonstrate two issues associated with their areal distribution. First, the CDF curves (Figure 10) illustrate an over-fitting of the model. The cumulative distribution curves have similar form for the model data and the independent test data, indicating that the environmental space modeled is comparable. However, the predicted suitability of the test data is noticeably less than the training data. The agreement of the remote sensing and field data sets strengthen this argument. Second, the cassava model demonstrates a problem with omitted environmental space. Our study villages in the Form 6 data excluded the uplands in the southwest corner of the Nang Rong District that are dominated by the cultivation of cassava, an upland field crop. The effect of this omission in training data is clear from the comparison of the suitability maps produced from the Form 6 and the remote sensing data (Figure 8). This is further illustrated in the CDF curves (Figure 10) that are distinctly different from each other, and the Form 6 model estimates indicate lower suitability scores for the majority of locations of the remote sensing and field datasets . It should be noted that the latest version of Maxent allows for sampling bias to be accounted for in the selection of random background values (Phillips et al. 2009). We explored this function for the Form 6 cassava data, but we did not find any substantial improvement. As Phillips et al. (2009) indicate, the greatest improvement is found for highly (areally) constrained species; all of our crop types are widespread throughout this agricultural district, although the sampling distribution is not.

4.5 Assessment Challenges of Fuzzy Suitability Analysis using Presence-only Data

The creation of a fuzzy suitability score using presence-only methods is very appealing for a number of reasons. However, one area that presents challenges is model assessment. The AUC-ROC is an assessment tool in the MaxEnt modeling software. AUC-ROC is calculated based on the same data used to optimize the species-environment relationships derived from the machine learning algorithm. The AUC-ROC assessment uses data consistent with that of the model and provides a comparison of model performance between runs and between the training and testing datasets. Lobo et al. (2008) note that the AUC-ROC approach serves as a measure of the model s ability to distinguish between absolute absences and presences. However, in the case of Maxent, the “absences” are randomly selected “background” values. Thus, the AUC-ROC actually measures the ability of the model to distinguish between

presences and the random background. Several papers have investigated this issue in the species distribution modeling community drawing attention to how AUC-ROC is calculated, how pseudo absences are created, and how commission and omission errors are created (Peterson et al. 2008, Phillips et al. 2009, Lobo et al. 2010, Lobo and Tongelli, 2011). Despite these issues, Manel et al. (2001) found that the AUC-ROC and the kappa-statistic are highly correlated and the lack of true absence data required for a kappa statistic and the fuzzy model output, the AUC-ROC serves as a reasonable modeling assessment tool in this application. Several alternate approaches have been suggested. Peterson et al. (2008) compare the AUC-ROC between two competing models at a given predicted area threshold, rather than comparing a model of the null AUC = 0.5 over the entire predicted area threshold. Philips et al. (2009) suggest altering the selection of pseudo-absence points based on sampling bias, whereas Lobo et al. (2010) caution against unrealistic absence points.

In this study, our data and goals are not identical to species distribution modeling and thus our model interpretation and assessment are not the same. For example, there are no true absences in our data; a given crop can be grown in any location, but given environmental constraints, the probability of it being grown is less. Thus, we cannot directly compare presence and absences. Our interest in the output of the model is the continuous suitability of a given crop and not a binary presence/absence based on a threshold value. As such, the suggestions of Peterson et al. (2008) do not apply. We did explore the effect of sampling bias on the Form 6 data as suggested by Phillips et al. (2009) but did not find that this had a substantial impact (less than 2% difference in the AUC for cassava, the more constrained crop). As Phillips et al. (2009) note, sampling bias has the greatest effect on constrained species; our crop presences are widespread across the study area.

We do not have any true absence data and in fact argue that such data does not exist by definition for our study area; rather crop suitability ranges across a continuous gradient. For this analysis, the conceptual framework of the potential niche is more appropriate than a remote sensing classification. Given that the validity of pseudo absence data has many issues, we chose to compare model results between our target dataset, with two alternative datasets to better understand the strengths and weaknesses of the model output. Additionally, we relied not only on the AUC-ROC, but also the CDF of suitability across the input data locations. Our comparison illustrated that the Form 6 data was overall comparable to the remote sensing data with the exception of the omitted area of high crop suitability in the southwest region where we did not have any study villages. In defense of the use AUC-ROC assessment method with Maxent, the data used to calculate the AUC-ROC are consistent with the data used to train and optimize Maxent s machine learning algorithm and the AUC-ROC is calculated for both the training and testing data. Thus, while the use of random background points may not be ideal absence data, it provides a consistent comparison with how the model functions as well as tests for consistency between the training and testing data, especially for problems with model overfitting

The continuous fuzzy model output and lack of true absence data makes model assessment using existing tools a serious challenge. While we have relied on a comparison of different dataset, each with their own data limitations, to better understand the quality of our results, this does not address the methodological need for a better robust assessment tool. We sincerely hope that this issue is addressed in the near future as presence-only modeling is a very promising technique.

5. Conclusions

Novel methods of presence-only modeling have great potential for crop suitability modeling where true absences do not exist because any given crop could be grown in most or all

locations in a study area. Our results demonstrate that MaxEnt predicts crop suitability reasonably well, given the real-world challenges and limitations of our data sets. However, there are a number of weaknesses apparent as well. First, the model is sensitive to both sample size and distribution, despite the wide use of the MaxEnt model that relies on incomplete and areally restricted data sets in many ecological applications. Recent paper suggest methods to overcome these issues, thought further development is needed with a focus on potentiall whidespread occurrence species. Second, model assessment is not straight forward. While the AUC-ROC provides an indication of model performance between model runs and training and testing datasets, AUC-ROC measures the ability of the model to distinguish a pattern different than random, rather than a "goodness of fit." To validate the model beyond the AUC-ROC, an assessment of the results at locations from independent data works well. The strengths of MaxEnt outweigh the weaknesses, particularly, if they are addressed in the design of the experiment. For example, independent testing data can be used to validate the model, and process-based models can be integrated to estimate crop yield.

From our analysis, we suggest the following lessons for future modeling studies in human-managed ecosystems: (1) randomly partition the data and run the model numerous times (e.g., ~1,000) to understand variability and uncertainty in the data; (2) validate the model using appropriate data and test for the research question, i.e., do not rely on AUC-ROC alone; (3) use independent data sets, where available, to understand the limitations of each data set, especially in the absence of "true" validation data; (4) ensure that the observations are appropriately distributed in environmental space and (5) social factors may play a significant role in the distribution of crops in human-managed landscapes that unexplained by environmental conditions alone. Future areas of research should investigate robust model assessment of fuzzy classifications using presence-only data and how social factors affect land suitability using presence-only methods.

Acknowledgments

This research was supported by grants from the National Science Foundation – HSD: "Marginality in a Marginal Environment: An Agent-Based Approach to Population-Environment Relationships" (B. Entwisle, PI), NSF 0728822; and the National Institutes of Health – NICHD "Modeling Household Dynamics and Land Use" (B. Entwisle, PI), 1 R21 HD051176. Thanks to Peter Mucha, George Malanson, Megan Rua, Josh Gray, and four anonymous reviewers for comments that greatly improved this manuscript.

References

Araujo MB, Guisan A. Five (or so) challenges for species distribution modelling. Journal of Biogeography. 2006; 33:1677–1688.

Austin M. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling. 2007; 200:1–19.

Austin MP. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling. 2002; 157 PII S0304-3800(02)00205-3.

Bandyopadhyay S, Jaiswal RK, Hegde VS, Jayaraman V. Assessment of land suitability potentials for agriculture using a remote sensing and GIS based approach. International Journal of Remote Sensing. 2009; 30:879–895.

Burrough PA, Macmillan RA, Vandeursen W. Fuzzy classification methods for determining land suitability from soil-profile observations and topography. Journal of Soil Science. 1992; 43:193–210.

Ceballos-Silva A, Lopez-Blanco J. Delineation of suitable areas for crops using a Multi-Criteria Evaluation approach and land use/cover mapping: a case study in Central Mexico. Agricultural Systems. 2003; 77:117–136.

Dudik, M.; Phillips, SJ.; Schapire, RE. Performance guarantees for regularized maximum entropy density estimation. In: ShaweTaylor, J.; Singer, Y., editors. Learning Theory, Proceedings. Berlin: Springer-Verlag Berlin; 2004. p. 472-486.

Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberon J, Williams S, Wisz MS, Zimmermann NE. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 2006; 29:129–151.

Fukui, H. Food and Population in a Northeast Thai Village. Honolulu: University of Hawaii Press; 1993.

Ghassemi, F.; Jakeman, AJ.; Nix, HA. Salinisation of land and water resources: Human causes, extent, management and case studies. Englewood Cliffs, New Jersey: University of New South Wales Press Ltd; 1995.

Graham CH, Hijmans RJ. A comparison of methods for mapping species ranges and species richness. Global Ecology and Biogeography. 2006; 15:578–587.

Groenemans R, VanRanst E, Kerre E. Fuzzy relational calculus in land evaluation. Geoderma. 1997; 77:283–298.

Guisan A, Thuiller W. Predicting species distribution: offering more than simple habitat models. Ecology Letters. 2005; 8:993–1009.

Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecological Modelling. 2000; 135:147–186.

Hernandez PA, Graham CH, Master LL, Albert DL. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography. 2006; 29:773–785.

Hirzel AH, Lay GL. Habitat suitability modelling and niche theory. Journal of Applied Ecology. 2008; 45:1372–1381.

Jelinski DE, Wu JG. The modifiable areal unit problem and implications for landscape ecology. Landscape Ecology. 1996; 11:129–140.

Jimenez-Valverde A, Lobo JM, Hortal J. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions. 2008; 14:885–890.

Joel A, Wesstrom I, Messing I. Mapping suitability of controlled drainage using spatial information of topography, land use and soil type, and validation using detailed mapping, questionnaire and field survey. Hydrology Research. 2009; 40:406–419.

Joss BN, Hall RJ, Sidders DM, Keddy TJ. Fuzzy-logic modeling of land suitability for hybrid poplar across the Prairie Provinces of Canada. Environmental Monitoring and Assessment. 2008; 141:79–96. [PubMed: 17674133]

Kaida, Y.; Surarerk, V. Climate and Agricultural Land Use in Thailand. In: MM; Yoshino, editors. Climate and Agricultural Land Use in Monsoon Asia. Tokyo: University of Tokyo Press; 1984. p. 231-254.

Kurtener D, Torbert HA, Krueger E. Evaluation of agricultural land suitability: application of fuzzy indicators. Computational Science and Its Applications - ICCSA 2008. 2008:475–90.

Jensen JR, Cowen DJ, Althausen JD, Narumalani S, Weatherbee O. An Evaluation of the coastwatch change detection protocol in South-Carolina. Photogrammetric Engineering and Remote Sensing. 1993; 59(6):1039–1046.

Littleboy M, Smith DM, Bryant MJ. Simulation modelling to determine suitability of agricultural land. Ecological Modelling. 1996; 86:219–225.

Lobo JM, Jimenez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography. 2008; 17:145–151.

Lobo JM, Jimenez-Valverde A, Hortal J. The uncertain nature of absences and their importance in species distribution modelling. Ecography. 2010; 33:103–114.

Lobo JM, Tognelli MF. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. Journal for Nature Conservation. 2011; 19:1–7.

Manna P, Basile A, Bonfante A, De Mascellis R, Terribile F. Comparative land evaluation approaches: an itinerary from FAO framework to simulation modelling. Geoderma. 2009; 150:367–378.

Manel S, Williams HC, Ormerod SJ. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology. 2001; 38:921–931.

Messina JP, Walsh SJ. 2.5 D Morphogenesis: modeling landuse and landcover dynamics in the Ecuadorian Amazon. Plant Ecology. 2001; 156:75.

Murphy HT, Lovett-Doust J. Accounting for regional niche variation in habitat suitability models. Oikos. 2007; 116:99–110.

Ortega-Huerta MA, Peterson AT. Modeling ecological niches and predicting geographic distributions: a test of six presence-only methods. Revista Mexicana De Biodiversidad. 2008; 79:205–216.

Papes M, Gaubert P. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. Diversity and Distributions. 2007; 13:890–902.

Parnwell MJG. Rural poverty, development and the environment - the case of Northeast Thailand. Journal of Biogeography. 1988; 15:199–208.

Parnwell MJG. Southeast-Asia - a region in transition - a thematic human-geography of The Asean region. Progress in Human Geography. 1992; 16:651–652.

Pereira JMC, Duckstein L. A multiple criteria decision-making approach to GIS-based land suitability evaluation. International Journal of Geographical Information Systems. 1993; 7:407–424.

Peterson AT. Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology. 2003; 78:419. [PubMed: 14737826]

Peterson AT, Papes M, Soberon J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological Modelling. 2008; 213:63–72.

Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecological Modelling. 2006; 190:231–259.

Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications. 2009; 19:181–197. [PubMed: 19323182]

Pulliam HR. On the relationship between niche and distribution. Ecology Letters. 2000; 3:349–361.

Rahman MR, Saha SK. Remote sensing, spatial multi criteria evaluation (SMCE) and analytical hierarchy process (AHP) in optimal cropping pattern planning for a flood prone area. Journal of Spatial Science. 2008; 53:161–177.

Rindfuss, RR.; Walsh, SJ.; Entwisle, B.; Sawangdee, Y.; Vogler, JB. Household-parcel linkages in Nang Rong, Thailand: challenges of large samples. In: Rindfuss, RR.; Fox, J.; Walsh, SJ.; Mishra, V., editors. People and the Environment: Approaches for Linking Household and Community Surveys to Remote Sensing and GIS. Boston, MA: Kluwer Academic Publishers; 2003. p. 131-172.

Soberon J. Grinnellian and Eltonian niches and geographic distributions of species. Ecology Letters. 2007; 10:1115–1123. [PubMed: 17850335]

Stockwell D, Peters D. The GARP modeling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science. 1999; 13:143–158.

Stockwell DRB, Peterson AT. Effects of sample size on accuracy of species distribution models. Ecological Modeling. 2002; 148(1):1–13.

Walsh, SJ.; Rindfuss, RR.; Prasartkul, P.; Entwisle, B.; Chamratrithrong, A. Population Change and Landscape Dynamics: The Nang Rong, Thailand Studies. In: Entwisle, B.; Stern, PC., editors. Research Directions: Population, Land Use, and Environment. Washington, DC: National Academies of Science, Committee on the Human Dimensions of Global Change, National Research Council; 2005. p. 135-162.

Walsh SJ, Welsh WF, Evans TP, Entwisle B, Rindfuss RR. Scale dependent relationships between population and environment in Northeastern Thailand. Photogrammetric Engineering and Remote Sensing. 1999; 65(1):97–105.

Williams JN, Seo CW. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions. 2009; 15(4):565–576.

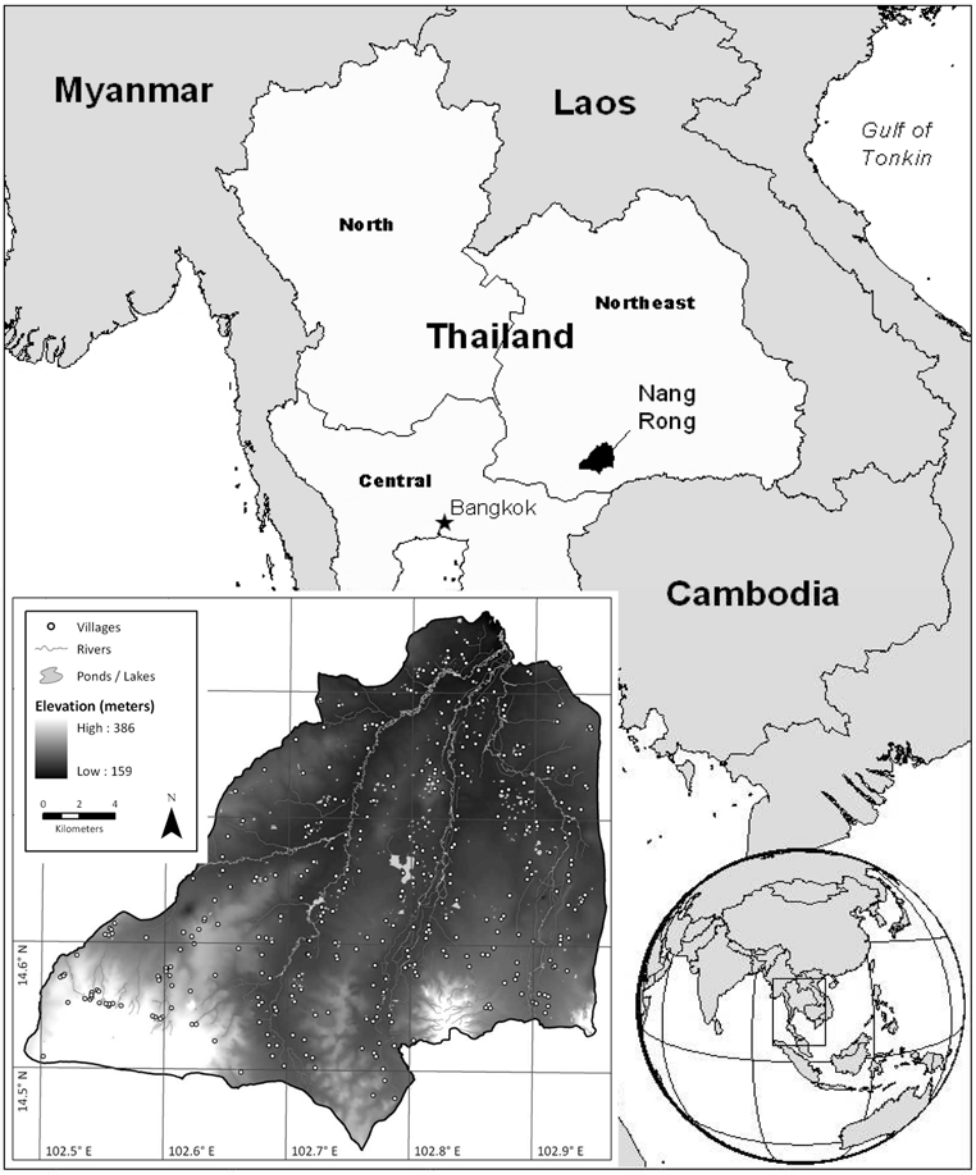

Figure 1. Study Area: Nang Rong District, Thailand

Note that the location of actual study villages has been omitted for confidentiality reasons related to Human Subjects research.

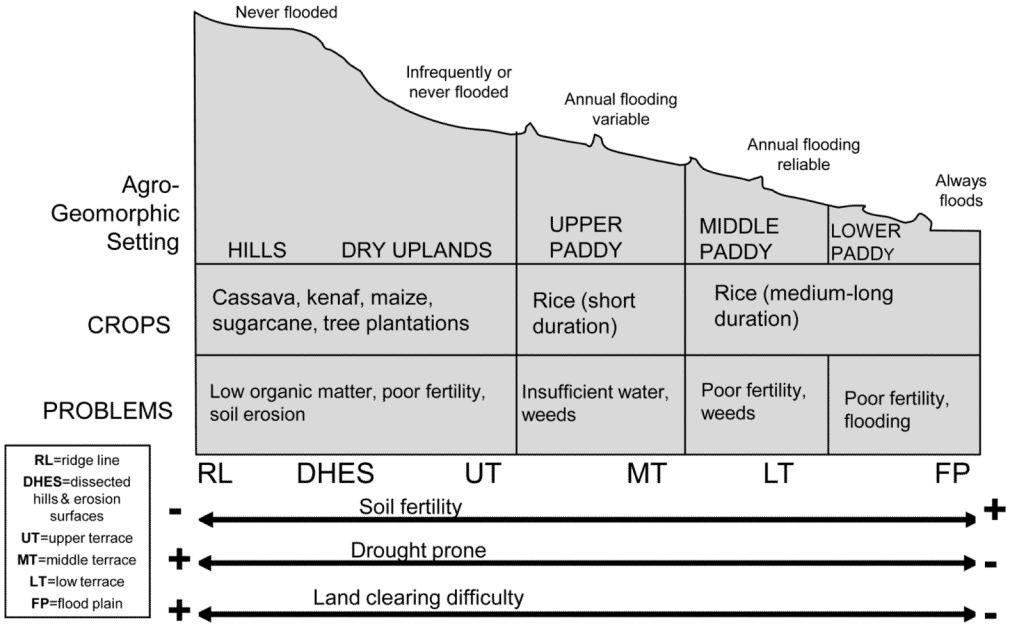

Figure 2.
Idealized cropping systems: gradients & constraints (after Rigg 1991).

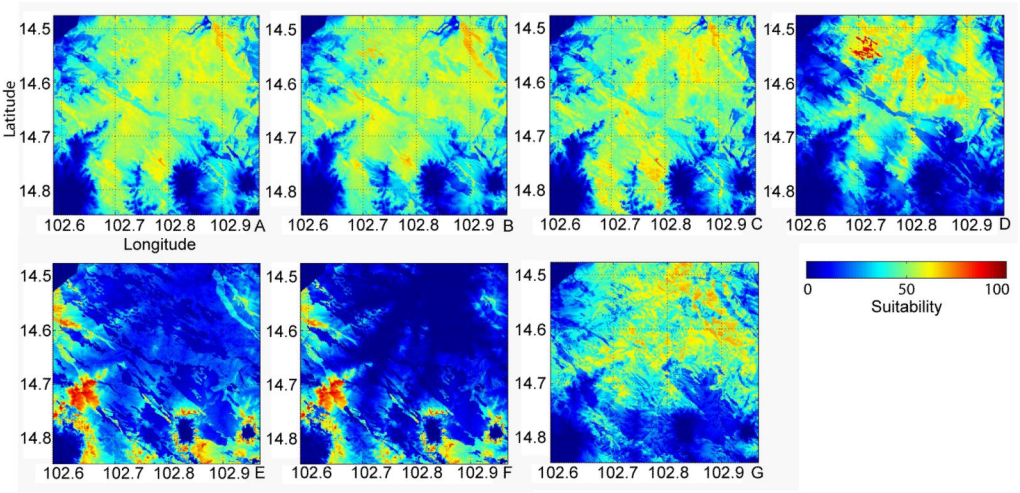

Figure 3.
Mean crop suitability of Form 6 models: all rice (A), jasmine rice(B), other rice (C), sticky rice (D), all cassava (E), plantation cassava(F), village cassava(G).

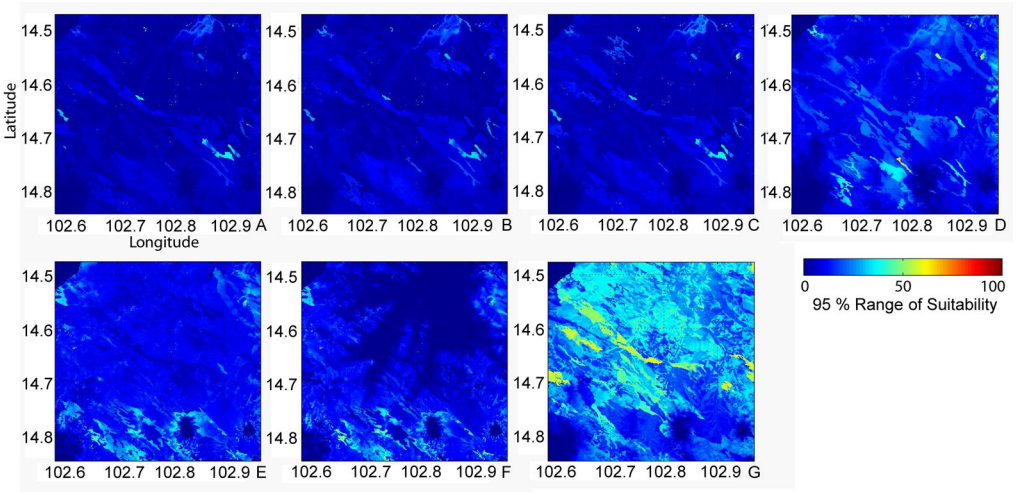

Figure 4.

95% range of suitability of Form 6 models: all rice (A), jasmine rice(B), other rice (C), sticky rice (D), all cassava (E), plantation cassava(F), village cassava(G).

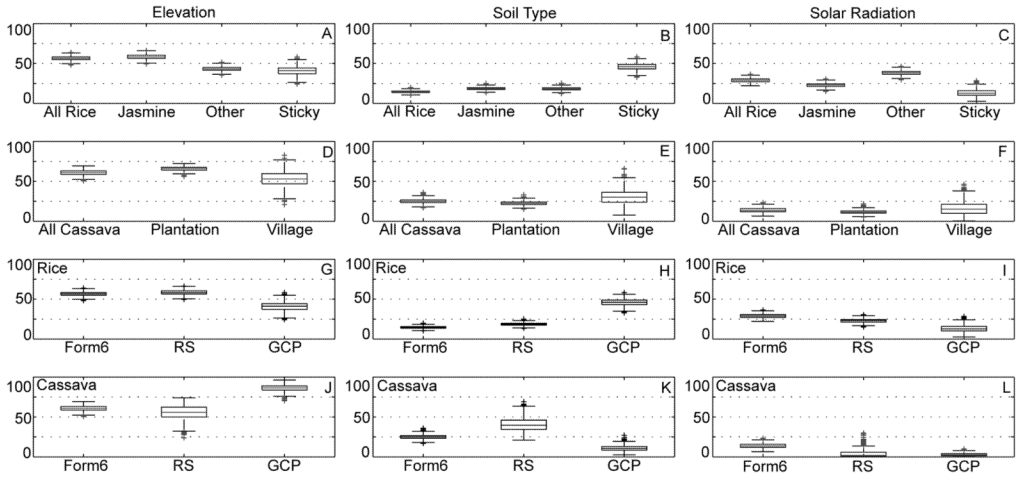

Figure 5.

Comparison of the percent of total model contribution of elevation, soil type, and solar radiation for Form 6 rice varieties (A,B,C), Form 6 cassava varieties (D,E,F), rice datasets (G,H,I), and cassava datasets (J,K,L).

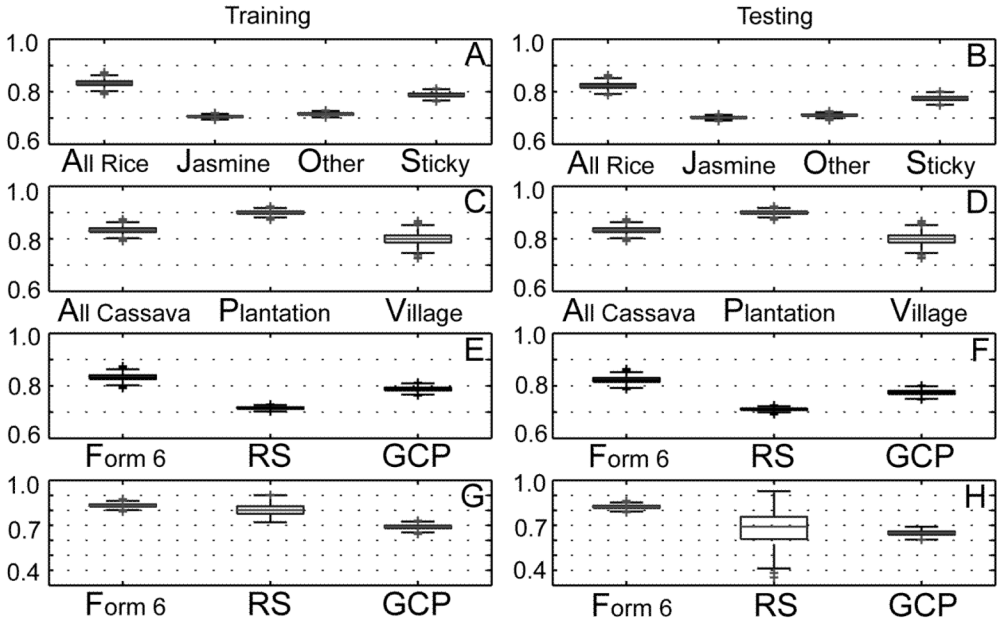

Figure 6.
Comparison of training and testing AUC scores among Form 6 rice varieties (A,B), Form 6 cassava varieties (C,D), rice datasets (E,F), and cassava datasets (G,H).

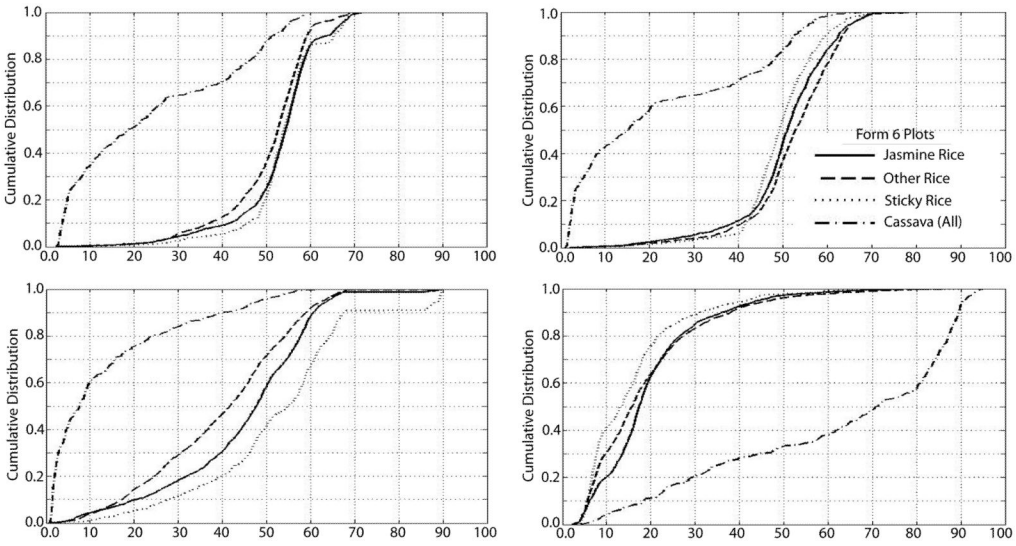

Figure 7.
Cumulative distribution function (CDF) plots of modeled suitability for jasmine rice (A), other rice (B), sticky rice (C), and all cassava (D) at Form 6 plot locations of the other crop varieties.

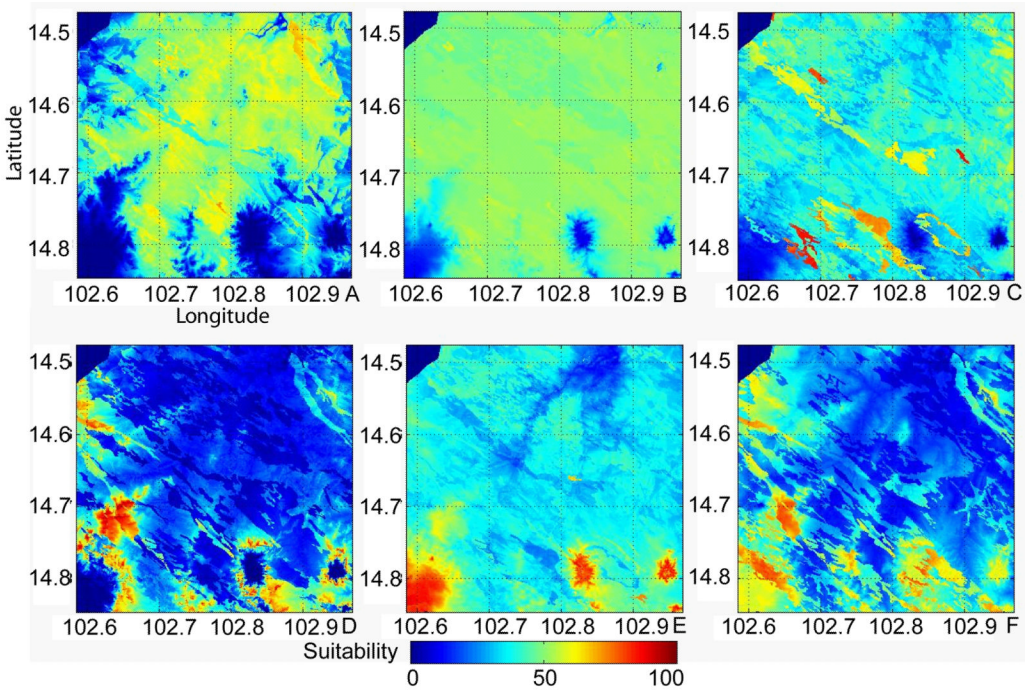

Figure 8.

Mean crop suitability of all rice from the Form 6 (A), remote sensing (B), and field (C) datasets and all cassava from Form 6 (D), remote sensing (E), and field (F) datasets.

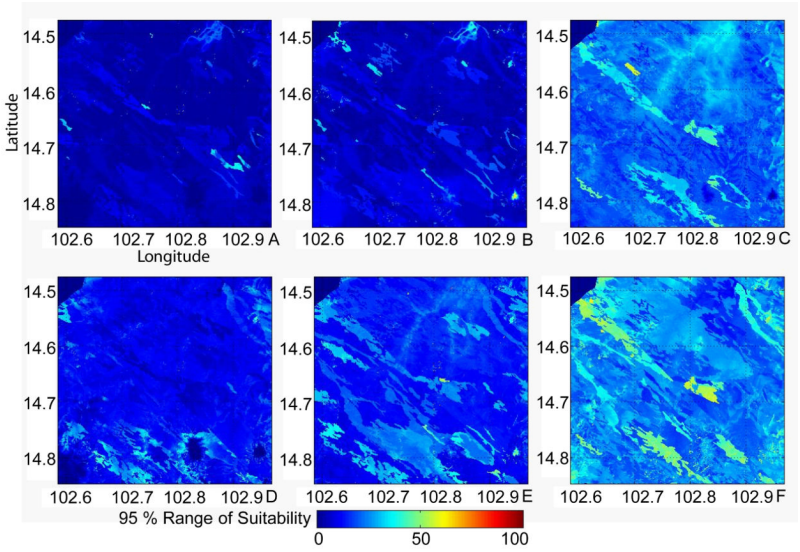

Figure 9.
95% range of suitability for all rice from the Form 6 (A), remote sensing (B), and field (C) datasets and all cassava from Form 6 (D), remote sensing (E), and field (F) datasets.

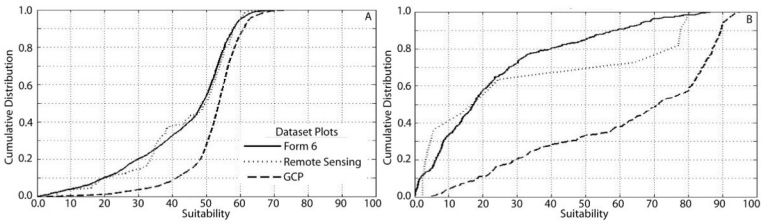

Figure 10.
Cumulative distribution function (CDF) plots of modeled suitability of Form 6 all rice (A) and Form 6 all cassava (B) at Form 6, remote sensing (RS), and field data (GCP) locations.

Table 1

Data Source	Crop Type	n	Nearest Neighbors Ratio
Form 6	Rice (All)	7739*	0.24
Remote Sensing	Rice (All)	3221	0.95
GCP	Rice (All)	90	0.78
Form 6	Jasmine Rice	5894	0.5
Form 6	“Other” Rice	4520	0.5
Form 6	Sticky Rice	646	0.49
Form 6	Cassava (All)	486	0.41
Remote Sensing	Cassava (All)	471	0.87
GCP	Cassava (All)	16	0.71
Form 6	Plantation Cassava	428	0.52
Form 6	Village Cassava	58	0.51

Note: Nearest Neighbor Ratio (mean of 5 nearest neighbors used) = mean distance nearest neighbors / expected mean distance nearest neighbors from a random distribution

Table 2

		Field Data						
		Rice	**Upland**	**Forest**	**Built**	**Water**	**Wetland**	**Comission Error**
Classification	Rice	41	1	0	0	0	0	0.02
	Upland	3	23	1	0	0	0	0.15
	Forest	1	2	16	3	0	0	0.27
	Built	0	0	0	6	0	0	0.00
	Water	0	0	0	0	8	0	0.00
	Wetland	0	0	0	0	1	0	N/A
Omission Error		0.09	0.12	0.06	0.33	0.11	N/A	

Ecol Inform. Author manuscript; available in PMC 2012 September 1.