**US Army Corps
of Engineers®**
Engineer Research and
Development Center

ERDC
INNOVATIVE SOLUTIONS
for a safer, better world

Testing Maximum Entropy Analysis to Define Population Distributions

Robert C. Lozar, Scott A. Tweddale, Charles R. Ehlschlaeger,
Carey L. Baxter, and Jeffrey A. Burkhalter

August 2018

The U.S. Army Engineer Research and Development Center (ERDC) solves the nation's toughest engineering and environmental challenges. ERDC develops innovative solutions in civil and military engineering, geospatial sciences, water resources, and environmental sciences for the Army, the Department of Defense, civilian agencies, and our nation's public good. Find out more at www.erdcl.usace.army.mil.

To search for other technical reports published by ERDC, visit the ERDC online library at <http://acwc.sdp.sirsi.net/client/default>.

Testing Maximum Entropy Analysis to Define Population Distributions

Robert C. Lozar, Scott A. Tweddale, Charles R. Ehlschlaeger,
Carey L. Baxter, and Jeffrey A. Burkhalter

*U.S. Army Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
2902 Newmark Dr.
Champaign, IL 61824*

Final Report

Approved for public release; distribution is unlimited.

Prepared for Headquarters, U.S. Army Corps of Engineers
Washington, DC 20314-1000

Under Project "Phase Zero Assessment of Urban Security Threats"

Abstract

The Urban Security research project attempts to characterize sub-national populations and urban dynamics that impact indicators of urban security threats. This work used the MaxEnt (Maximum Entropy) program to define the probable locations of human habitation for Chittagong Division, Bangladesh. This represents the likely first application of maximum entropy analysis to the delineation of human populations. The model was parameterized with a large number of known house locations and several predictor geographic information system (GIS) data layers that might influence the geographic distribution of human habitation. Analyses and tests done using MaxEnt indicated that the MaxEnt model effectively delineated the probability distribution of human habitation based on a series of inputs based on a series of inputs from three data sources: Landsat NaturalVue TM (Thematic Mapper) imagery, Shuttle Radar Topography Mission (SRTM) elevation, and OpenStreetMap roads. Results show that, by using remotely-sensed data and “presence only” samples of human habitation, it is possible to estimate the habitation distribution probability based on sound statistical procedures used in MaxEnt.

DISCLAIMER: The contents of this report are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products. All product names and trademarks cited are the property of their respective owners. The findings of this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

DESTROY THIS REPORT WHEN NO LONGER NEEDED. DO NOT RETURN IT TO THE ORIGINATOR.

Contents

Abstract	ii
Tables and Figures.....	iv
Preface	v
Unit Conversion Factors	vi
Acronyms and Abbreviations	vii
1 Introduction.....	1
1.1 Background	1
1.2 Objective.....	2
1.3 Approach	2
1.4 Scope.....	3
2 Maximum Entropy Analysis with MaxEnt	4
3 Data Input to MaxEnt	6
3.1 Predictor layers	6
3.2 Sample locations of human habitation (presence sites).....	8
4 Modeling Human Population Distribution with MaxEnt	11
4.1 Testing alternatives.....	11
4.2 Maximum entropy analysis household probability result	13
4.2.1 Complete study area results.....	13
4.2.2 Evaluation of human habitation model.....	16
4.2.3 Analysis of variable contributions	19
5 Conclusions and Recommendation	21
References	22
Report Documentation Page (SF 298)	24

Tables and Figures

Tables

1	Twenty-three “predictor” GIS layers input to the MaxEnt model	7
2	MaxEnt Variable Contribution and Permutation table for test area	12
3	MaxEnt Variable Contribution and Permutation table for Chittagong Division, Bangladesh	13
4	Threshold tests and training omission rates for a variety of model objectives.....	18

Figures

1	Example of streams (blue) delineation from elevation data using GIS analysis to create a water accumulation layer (yellow [fewer number of contributing cells] to blue [greater number of contributing cells]). Red lines are larger drainage basin boundaries	8
2	Example of streams (blue) derived from elevation data from which distance from streams (gray scale) was generated	9
3	Example of panchromatic image from which housing locations were determined (dots)	9
4	MaxEnt test area output. Yellow dots are sample points for presence of buildings. Black lines are roads from which the distance from roads variable layer was generated. Darker shades of red indicate higher probability of human presence	11
5	MaxEnt complete study area output for subset area. Darker shades of red indicate higher probability of human presence.....	14
6	MaxEnt complete study area output for Chittagong Division, Bangladesh. Darker shades of red indicate higher probability of human presence. Dark red lines clearly delineate the road network.....	15
7	Omission on training samples and predicted area as a function of the cumulative threshold.....	17
8	ROC for MaxEnt model for Chittagong Division, Bangladesh	17
9	Jackknife test results	20

Preface

This study was conducted for the Engineering Research and Development Center (ERDC) under the project titled “Urban Security,” which was part of “Phase Zero Assessment of Urban Security Threats.” The technical monitor was Alan Anderson, Technical Director, CEERD-CZT.

The work was performed by the Cultural Resources Branch (CFN) of the Installations Division (CN), Construction Engineering Research Laboratory (CERL). At the time of publication, Michael Hargrave was Chief, CEERD-CNF and Michelle Hanson was Chief, CEERD-CN. The Deputy Director of ERDC-CERL was Dr. Kirankumar V. Topudurti and the Director was Dr. Lance D. Hansen.

COL Ivan P. Beckman was Commander of ERDC, and Dr. David W. Pittman was the Director.

Unit Conversion Factors

Multiply	By	To Obtain
acres	4,046.873	square meters
cubic feet	0.02831685	cubic meters
cubic inches	0.00001638706	cubic meters
degrees Fahrenheit	$(5/9) \times (^{\circ}\text{F} - 32)$	degrees Celsius
feet	0.3048	meters
gallons (U.S. liquid)	0.003785412	cubic meters
inches	0.0254	meters
miles (U.S. statute)	1.609347	kilometers
pounds (mass)	0.4535924	kilograms
square feet	0.09290304	square meters
square miles	2,589,998	square meters
tons (2,000 pounds, mass)	907.1847	kilograms
yards	0.9144	meters

Acronyms and Abbreviations

Term	Definition
ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
AUC	Area Under the ROC Curve
CERL	Construction Engineering Research Laboratory
CSIRO	Australia's Commonwealth Scientific and Industrial Research Organization
DoD	U.S. Department of Defense
ERDC	U.S. Army Engineer Research and Development Center
ERDC-CERL	Engineer Research and Development Center, Construction Engineering Research Laboratory
ESRI	Environmental Systems Research Institute, Inc.
GIS	Geographic Information System
NSN	National Supply Number
OMB	Office of Management and Budget
ROC	Receiver Operating Curve
SAR	Same As Report
SF	Standard Form
SOM	Self-Organizing Map
SRTM	Shuttle Radar Topography Mission
TM	Thematic Mapper
TR	Technical Report

THIS PAGE INTENTIONALLY LEFT BLANK

1 Introduction

1.1 Background

The Urban Security research project attempts to characterize sub-national populations and urban dynamics that impact indicators of urban security threats. The research goal of this project is to develop analytic techniques that use existing data to characterize indicators for long-term monitoring. The outputs of the developed techniques can be used to develop maps of household and cultural attributes at the neighborhood scale that reflect multiple population attributes that identify specific subgroups. The methodology extends traditional survey analysis methods to create cartographic products useful in agent-based modeling and geographic information analysis. The conceptual approach of the methodology is to locate potential household locations with a conflation model (Cobb et al. 1998), which is also known as “data fusion” (Wald 1999). Conflation is the process by which multiple data layers are combined to generate a product containing the best aspects of each layer. In this research, five types of data are conflated: choropleth population enumeration estimates, survey responses or microdata, a household or population density map, maps of survey response constraints, and samples of ground truth information containing specific survey responses. Each of the data inputs imparts specific types of precise information that other data types do not contain. Monte Carlo methods are employed to simulate survey cases’ locations in a study area. The distribution of point patterns requires household density surfaces that probabilistically located the households.

Since most of the basic data about the occurrence of human occupation relates to the location of structures in which people reside, this work concentrated on means of defining the distribution of human habitation based on “presence only” techniques. There are several techniques in the “presence only” genre for modeling of species geographic distribution: (1) BIOCLIM (Busby 1986, Nix and Busby 1986), which predicts suitable conditions in a “bioclimatic envelope” of observed presence values in each environmental dimension; (2) DOMAIN (Carpenter, Gillison, and Winter 1993), which uses a similarity metric, for which a predicted suitability index is given by computing the minimum distance in environmental space to any presence record; (3) GARP, which has been applied to several presence only studies (Anderson, Lew, and Peterson 2003; Joseph and Stockwell 2000, Peterson

and Kluza 2003, Peterson and Robins 2003, Peterson and Shaw 2003), and (4) MaxEnt (Phillips, Anderson, and Schapire 2006), which is a more recent software program that uses “presence only” techniques, and which was selected for used in this research because it provides the advantages of:

- presence only predictability
- statistical objectivity
- a generative approach that gives better predictions when the amount of training data is small
- statistical analysis of model performance
- intuitive logistic output that represents probability of presence.

This research used the MaxEnt software to apply the maximum entropy analysis technique. In image processing, entropy is a measure of image information content. MaxEnt is designed to determine the maximum information content expressed by the data submitted to it. MaxEnt defines the geographic distribution of species based on a multivariate approach. The technique develops an unbiased probability distribution on the basis of partial knowledge. In addition, it offers a suite of statistical analysis tools that are highly useful in evaluating the quality of the resulting range map and ancillary outputs. This research, which represents the likely first application of maximum entropy analysis to the delineation of human populations, focused on predicting the geographic distribution of humans based on human response to the availability of resources.

1.2 Objective

The objective of this research was to test and evaluate the maximum entropy analysis modeling approach for predicting the geographic distribution of human habitation for the Chittagong Division of Bangladesh.

1.3 Approach

The objective of this work was met in the following steps:

1. Many geospatially explicit data (i.e., geographic information system (GIS) data layers) necessary to parameterize the MaxEnt model were compiled. More specifically, GIS layers for various themes that are likely to impact where humans live (e.g., distance to roads, landcover, etc.) were input to the MaxEnt model as predictor layers.
2. Locations where humans are known to live were also input to the MaxEnt software.

3. Based on these inputs, MaxEnt produced a spatially explicit output map where every pixel value represented the probability of human habitation, conditional on the predictor variables provided as input.
4. MaxEnt also identified the importance of the contribution of each predictor layer to the overall estimate of probability of human habitation.
5. Finally, MaxEnt provided a measure of overall model performance.

1.4 Scope

The maximum entropy analysis approach was used only in the study area described. However, since the data layers used were those commonly prepared for operational planning and design, it is anticipated that the techniques developed in this work will be useful in all locations where the U.S. Department of Defense (DoD) collects and stores information.

2 Maximum Entropy Analysis with MaxEnt

Shannon (1948) described entropy as “a measure of how much ‘choice’ is involved in the selection of an event.” E.T. Jaynes (1957) suggested that the best approximation for an unknown probability distribution is to ensure that the approximation satisfies any constraints on the unknown distribution, and that subject to those constraints, the distribution should have maximum entropy.

The MaxEnt (MAXimum ENTropy analysis) program is a downloadable tool (Shapire 2017) developed to define the probable geographic distribution of species based on a multivariate approach. MaxEnt applies a technique called “maximum entropy” analysis. The maximum entropy approach is theoretically derived from a “First Principle” of physics (which starts directly at the level of established science and does not make assumptions such as empirical model and fitting parameters). Specifically, the Second Law of Thermodynamics states that, in systems without outside influences, processes move in a direction that maximizes entropy. As applied to the distribution of a species, the hypothesis is that the sum of the species population behavior will tend to follow this constraint and thus, result in a habitat usage that reflects maximum entropy. Since there may be outside influences not included among the input data for a species, the distribution is likely to be larger than the observed delineation. Thus, the maximum entropy approach can be expected to generate a “potential distribution” for each species based on the inputs available.

The maximum entropy analysis technique develops unbiased probability distribution on the basis of partial knowledge (Phillips, Anderson, and Schapire 2006). It uses a “presence only” approach (called an unconditional model) rather than including data that reflect “known absence sighting” (a conditional model). MaxEnt also provides a suite of statistical analysis tools that are highly useful in evaluating the quality of the resulting probable geographic distribution map and ancillary outputs. MaxEnt “software [is] based on the maximum entropy approach for species habitat modeling. This software takes as input a set of layers or environmental variables (such as elevation, precipitation, etc.), as well as a set of georeferenced occurrence locations, and produces a model of the range of the given species” (Phillips and Dudík 2008). Since humans respond to availability of resources just as other species do, this research used MaxEnt to predict the probable geographic distribution of humans.

The idea behind MaxEnt is to estimate a target probability distribution by finding the probability distribution of maximum entropy (i.e., that is most spread out, or closest to uniform surface), subject to a set of constraints that represent the incomplete information provided about the target distribution. This is accomplished via a deterministic sequential-update algorithm (Dudík, Phillips, and Schapire 2004). The process iteratively adjusts a weight so as to minimize the resulting regularized log loss. The algorithm is guaranteed to converge to the MaxEnt probability distribution. The default output is logistic, thereby providing an estimate between 0 and 1 of probability of presence in each pixel of the predicted geographic distribution map.

3 Data Input to MaxEnt

Since this project did not have access to the field to collect data, it was limited to those items that could be accessed from sources available more generally. MaxEnt uses two types of inputs:

1. GIS layers for various themes that are likely to impact where humans live (predictor layers)
2. Sample locations of human habitation (presence sites).

3.1 Predictor layers

Predictor GIS layers for all of the various themes that might influence the geographic distribution of human habitation are required as input to MaxEnt. In this project, 23 layers were initially provided as predictor layers, including individual bands of satellite imagery, unsupervised classifications of satellite imagery, satellite image pattern and texture measures, elevation and its derivatives, hydrography, and distance to primary roads (Table 1).

All of these predictor layer inputs, with the exception of roads, were derived from two remotely-sensed data sources: Thematic Mapper (TM) imagery (from the NaturalVue natural color worldwide seamless mosaic, collected circa 2000), and global elevation data (from the Shuttle Radar Topography Mission (SRTM)). Roads were derived from OpenStreetMap. All data were standardized to a 15-meter resolution (i.e., that of the NaturalVue image; the elevation data were resampled from 30 meters to match the 15-meter NaturalVue TM image).

An investigation to determine the grossest resolution that would be acceptable for this tasking was carried out. Imagery resolution in the 5- to 8-meter range would be useful to identify important human habitation patterns such as linear roads or building right angle corners. A 15-meter resolution is too gross to identify these indicators of human activity. Some 5-meter imagery was available, but it only covered a small part of the study area. Unfortunately, 15-meter imagery was the only type available that covered the entire study area.

Table 1. Twenty-three “predictor” GIS layers input to the MaxEnt model.

MaxEnt Input Layers		
TM NaturalVue (resampled to 15-meter)		
	Red	
	Green	
	Blue	
Unsupervised Land Cover		
		SOM (Self-Organizing Map) - 60 unique classes
		K-means - 35 unique classes
	Pattern	
		Diversity
		Dominance
		Center vs neighbors
Texture Haralick		
		Mean
		Variance
		Difference of Entropies
		Difference of Variances
		Information Measures of Correlation IC1
		Information Measures of Correlation IC2
Elevation (15-meter)		
	Elevation	
	Slope	
	Slope length	
	slope_deg_curve	
	slope_deg_tancurve	
Hydrology		
		Distance from streams
		Accumulation
		Drainage Direction (=Aspect)
GIS Layers from other		
	Distance from roads	

The NaturalVue Imagery was classified using two different unsupervised classification methods: Self-organizing map (SOM), and k-means classification. Three different pattern measures and six different texture measures were also derived from the satellite imagery and input to MaxEnt as predictors. From the elevation data, several derivatives of slope were extracted and used as input to MaxEnt as predictors, as well as hydrological data derived from elevation. Distance to primary roads was also input to MaxEnt.

All 23 “predictor” GIS data layers were processed so that they had the same geographic bounds and cell size. They were provided to MaxEnt as American Standard Code for Information Interchange (ASCII) raster grids (in the Environmental Systems Research Institute, Inc. [ESRI] *.asc format). They were all identified within MaxEnt as continuous data layers, with the exception of the two image classifications, which were identified as categorical. Figures 1 and 2 show examples of “predictor” GIS data layers that were derived from elevation data.

3.2 Sample locations of human habitation (presence sites)

Locations of human habitation, or residences, were identified as “presence” sites by digitizing points in high resolution panchromatic imagery at 1:1300 scale. For rural areas, at least 10-20 points were identified for each rural image that was available within the study area. This resulted in 1168 small town/rural points. Living quarter houses were desired. They had the character of having either a gable or hip roof with straight edges and a size appropriate for a living unit (rather than a shed) and shadows in the correct direction (Figure 3). Images in Google Earth were referenced as a secondary source of information when the type of building could not be determined in the panchromatic imagery.

Figure 1. Example of streams (blue) delineation from elevation data using GIS analysis to create a water accumulation layer (yellow [fewer number of contributing cells] to blue [greater number of contributing cells]). Red lines are larger drainage basin boundaries.

Figure 2. Example of streams (blue) derived from elevation data from which distance from streams (gray scale) was generated.

Figure 3. Example of panchromatic image from which housing locations were determined (dots).

Sampling for urban areas differed from that done in rural areas. Urban sampling was done on locations already identified as populated/urban. In urban areas, it was difficult to distinguish apartment buildings from offices. Context was used to distinguish between commercial and residential structures. Building dimensions were also used as a clue. Long, rectangular buildings were considered likely warehouses, while residences tended to be “squarer.” This resulted in 1921 small urban/suburban points.

Since the “presence” sights were identified from high detail panchromatic images and “predictor” GIS data layers were derived from separate, 15-meter source datasets (either satellite images or elevation) the “presence” data and the “predictor” layers were independent of each other.

4 Modeling Human Population Distribution with MaxEnt

4.1 Testing alternatives

To identify the importance of the contribution of each predictor layer to the overall estimate of probability of human habitation, MaxEnt was run for a subset of the entire study area. The test area was run with 23 predictor GIS layers and all “presence” locations. Outputs from MaxEnt for this subset area were a Probability of Human Presence image (Figure 4) and the Variable Contribution and Permutation table (Table 2).

Figure 4. MaxEnt test area output. Yellow dots are sample points for presence of buildings. Black lines are roads from which the distance from roads variable layer was generated. Darker shades of red indicate higher probability of human presence.

Table 2. MaxEnt Variable Contribution and Permutation table for test area.

Variable	Percent Contribution	Permutation Importance
Distance from roads	51.3	42.2
Haralick_texture_Variance	13.2	1.2
NaturalVue TM Band 1	7.9	34.5
Distance from streams	5.1	3.4
Haralick_texture_Difference_of_Entropies	4.2	1.3
Unsupervised landcover classification (SOM)	2.8	1.7
NaturalVue TM Band 3	2.3	3.8
Unsupervised landcover classification (k-means)	2.3	0
Slope_deg_curve	1.9	1.1
Haralick_texture_measures of correlation(IC2)	1.6	0
NaturalVue TM Band 2	1.5	8.4
Slope_deg_tancurve	1.1	0.7
Slope	1.1	0.1
Drainage direction (Aspect)	0.7	0.1
Haralick_texture_Difference_of_Variances	0.7	0.2
Haralick_texture_Mean	0.6	0.5
Pattern_dominance	0.4	0.1
Elevation	0.3	0.1
Pattern_center_vs_neighbors	0.3	0.2
Pattern_diversity	0.3	0
Slope length	0.2	0.1
Haralick_texture_measures of correlation IC1	0.2	0.1
Accumulation	0.1	0.1

Visual inspection of the geographic distribution probability map indicates that distance to roads is an important predictor of human habitation for the study area (Figure 4). This observation is confirmed in the Variable Contribution and Permutation table, where distance to roads contributes 51.3% to the overall model, and where the next highest contributor is a variance texture measure at 13.2% (Table 2). A second measure of predictor variable contribution is permutation importance. Again, distance to roads was clearly the most important variable with a permutation importance of 42.2%, followed by Band 1 of the NaturalVue image with a permutation importance of 34.5%.

To reduce MaxEnt run time when running the model for the entire Chittagong Division of Bangladesh, those predictor GIS layers with minimal contribution to the model were eliminated. More specifically, predictor layers

with a “percent contribution” value greater than 1.5% and a “permutation importance” value greater than 1.2% were retained as predictor layer inputs for analysis of the entire study area. The eight predictor layers that met this criteria are:

- Distance from roads
- Haralick_texture_Variance
- NaturalVue TM Band 1
- NaturalVue TM Band 2
- NaturalVue TM Band 3
- Distance from streams
- Haralick_texture_Difference_of_Entropies
- Unsupervised landcover classification (SOM).

4.2 Maximum entropy analysis household probability result

4.2.1 Complete study area results

A final MaxEnt analysis was run for the entire Chittagong Division of Bangladesh. A total of 3089 “presence” locations were input to the model. A total of seven predictor GIS layers were also input to the model. For technical reasons, distance from streams had to be removed from the analysis. MaxEnt requires that all input layers be in the same format, and there was a formatting problem with the distance to streams layer. Table 3 lists and Figures 5 and 6 show Final MaxEnt results.

Table 3. MaxEnt Variable Contribution and Permutation table for Chittagong Division, Bangladesh.

Variable	Percent Contribution	Permutation Importance
Distance from roads	84.8	60.8
Haralick_texture_Difference_of_Entropies	5.6	3.3
Haralick_texture_Variance	5.4	2.8
NaturalVue TM Band 1	2.5	22.7
NaturalVue TM Band 3	1.2	7.4
NaturalVue TM Band 2	0.3	2.6
Unsupervised landcover classification (SOM)	0.2	0.3

Figure 5. MaxEnt complete study area output for subset area. Darker shades of red indicate higher probability of human presence.

Figure 6. MaxEnt complete study area output for Chittagong Division, Bangladesh. Darker shades of red indicate higher probability of human presence. Dark red lines clearly delineate the road network.

Similar to results for the test area, MaxEnt results for the entire Chittagong Division of Bangladesh indicate that distance from roads is extremely important (85% Percent contribution to the model and 61% Permutation importance). The greater contribution of distance from roads for the entire study area as compared to the test area may result from the fact that, within the complete study area, more of the “presence” points, particularly the urban points, were located near the major road networks. The two textural measures contribute approximately 5% to the model, followed by NaturalVue bands 1 and 3.

4.2.2 Evaluation of human habitation model

As mentioned, one of the advantages of the MaxEnt modeling technique is that it generates several statistical evaluations of the quality of its output. Several of these statistics that were generated for the human habitation model for all of the Chittagong Division, Bangladesh are presented below. Figure 7 shows the omission rate and predicted area as a function of the cumulative threshold. The lack of correctness, or the omission rate (the fraction of the training localities that fall into pixels not predicted as suitable), is calculated on the training records (blue line). If the model performed well, the omission rate should be close to the predicted omission (black line). In this model, the lines are very similar, indicating that the model performed well.

Figure 8 shows the receiver operating curve (ROC) for the MaxEnt model for Chittagong Division, Bangladesh. A ROC is a plot of sensitivity, or 1-Omission Rate, plotted against 1-Specificity (Fractional Predicted Area). The name is derived from its original use in radar technology. The black line shown in the ROC curve represents what one would expect if the model resulted in one that has no discriminatory power. The size of the area between the black line and the red lines in the ROC curve reflects the ability of a model to discriminate between presence and non-presence of human habitation across the range of potential thresholds. More specifically, the Area Under the ROC Curve (AUC) provides a single measure of model performance, independent of any particular choice of threshold. The AUC can be interpreted as the probability that a random positive instance and a random negative instance are correctly identified by the classifier (Fielding and Bell 1997). The larger the AUC, the better the model has performed.

Figure 7. Omission on training samples and predicted area as a function of the cumulative threshold.

Figure 8. ROC for MaxEnt model for Chittagong Division, Bangladesh.

In Figure 8, the black line shows the line one would expect if this model was no better than random (AUC = 0.5). The red (training) line shows the “fit” of the model to the training data (AUC = 0.764). The further towards the top left of the graph that the red line is, the better the model is at predicting the presence of human habitation. With an AUC of 0.764, this MaxEnt model of the geographic distribution of human habitation predicts the distribution fairly well. In general, AUC values between 0.7 and 0.9 indicate a reasonable discrimination ability appropriate for many uses (Swets 1988, Pearce and Ferrier 2000). Likewise, Elith (2002) recommends that AUC values above 0.75 are considered potentially useful.

The MaxEnt program automatically calculates Fractional Predicted Area and Training Omission Rate for some common Cumulative and Logistic thresholds and model objectives. Table 4 shows some Fractional Predicted Areas and Training Omission Rates for Chittagong Division MaxEnt model using a variety of model objectives, under a series of differing threshold levels. Each threshold (Cumulative and Logistic) was used to obtain a binary (presence vs. absence) prediction for human habitation. Output column headings are:

- Thresholds (Cumulative and Logistic) used for the model objective
- Description = model objective
- Fractional predicted area = proportion of total area predicted to contain human habitation
- Training omission rate = rate of failure to predict an occurrence of human habitation where it is known to occur with the training data.

Table 4. Threshold tests and training omission rates for a variety of model objectives.

Cumulative Threshold	Logistic Threshold	Description	Fractional Predicted Area	Training Omission Rate
1.000	0.104	Fixed cumulative value 1	0.902	0.003
5.000	0.179	Fixed cumulative value 5	0.749	0.036
10.000	0.223	Fixed cumulative value 10	0.620	0.076
0.416	0.069	Minimum training presence	0.942	0.000
12.784	0.243	10 percentile training presence	0.560	0.100
30.627	0.396	Equal training sensitivity and specificity	0.297	0.297
32.002	0.414	Maximum training sensitivity plus specificity	0.284	0.308
1.234	0.111	Balance training omission, predicted area and threshold value	0.889	0.004
8.690	0.213	Equate entropy of thresholded and original distributions	0.651	0.065

For most model objectives tested, the training omission rates were low; only two were as high as the 0.3 level, indicating that the obtained results were satisfactory.

4.2.3 Analysis of variable contributions

An alternate estimate of which “predictor” variables are most important in the MaxEnt model can be determined by a “Jackknife” analysis. For a jackknife analysis, each “predictor” layer is excluded in turn so a model is created with the remaining “predictor” variables. Then a model is created using each “predictor” variable in isolation. Finally, a model is created using all “predictor” variables, as before. The results of the jackknife analysis are presented as bar charts.

Figure 9 shows the results of the jackknife analysis for the MaxEnt model of human habitation for Chittagong Division, Bangladesh. The red line at the bottom of the graph shows the complete human habitation model using all “predictor” variables for the model. The environmental variable with highest gain when used in isolation (the blue bars) is distance from roads, which therefore appears to have the most useful information by itself. This is followed by the two texture layers: variance followed by difference entropy. These results are in agreement with the values provided in Table 3, where distance from roads had the greatest influence on model results, followed by the two texture measures. In contrast, NaturalVue Band 2 makes almost no contribution in gain. It could be dropped from the analysis with little consequence. The input variable that decreases the gain the most when it is omitted (the green bars) is distance from roads, which therefore appears to have the most information that is not present in the other variables. In general, the “without variable” values are similar for the remaining variables, so none of the remaining variables contain a substantial amount of useful information that is not already contained in the other variables. In summary, the jackknife analysis confirmed that distance from roads is primarily important and cannot be dropped from consideration.

Figure 9. Jackknife test results.

5 Conclusions and Recommendation

This work used the MaxEnt (Maximum Entropy) program to define the probable locations of human habitation for Chittagong Division, Bangladesh. The analyses and tests done using MaxEnt indicate that the MaxEnt model effectively delineated the probability distribution of human habitation based on a series of inputs derived from two data sources: Landsat NaturalVue TM imagery and SRTM elevation. Results show that, by using remotely-sensed data and “presence only” samples of human habitation, it is possible to estimate the habitation distribution probability based on sound statistical procedures used in MaxEnt. MaxEnt also provides ample output to evaluate the performance of model output.

Within the Chittagong Division of Bangladesh, distance from roads (representing 85% of the importance to the model) was the single largest predictor of the probability of human habitation. Other predictor inputs that influenced the MaxEnt model to a significantly lesser degree were two texture measures and NaturalVue TM Bands 1 and 3. Imagery resolution in the 5- to 8-meter range would be useful to detect important human habitation patterns such as linear roads or building right angle corners. 15-meter resolution is too gross to detect these indicators of human activity.

It is recommended that future similar studies use imagery in the 5-meter range, which would provide improved information about human presence and likely eliminate the need to hand digitize the all-important road networks.

References

- Anderson, Robert P., Daniel Lew, and A. Townsend Peterson. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* 162(3):211-232.
- Busby, John R. 1986. *Bioclimate Prediction System (BIOCLIM)*. User's manual version 2.
- Carpenter, G., A. N. Gillison, and J. Winter. 1993. DOMAIN: A flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* 2(6):667-680.
- Dudík, M., S. J. Phillips, and R. E. Schapire. 2004. Performance guarantees for regularized maximum entropy density estimation. *Proceedings of the 17th Annual Conference on Computational Learning Theory*. New York: ACM Press, pp 655-662.
- Elith, J. 2002. Quantitative methods for modeling species habitat: Comparative performance and an application to Australian plants. In: S. Ferson and M. Burgman (eds.), *Quantitative Methods for Conservation Biology*. Springer, pp 39-58
- 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(1):38-49.
- Jaynes, E. T. 1957. Information theory and statistical mechanics. *Phys. Rev.* 106(4):620-630.
- Joseph, Leo, and David Stockwell. 2000. Temperature-based models of the migration of Swainson's Flycatcher (*Myiarchus swainsoni*) across South America: A new use for museum specimens of migratory birds. *Proceedings of the Academy of Natural Sciences of Philadelphia*, pp 293-300.
- Nix, H. A., and J. Busby. 1986. *BIOCLIM, a Bioclimatic Analysis and Prediction System*. Annual report CSIRO. Canberra, Australia: CSIRO Division of Water and Land Resources.
- Pearce, J., and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*. 133:225-245.
- Peterson, A. Townsend, and Daniel A. Kluza. 2003. New distributional modelling approaches for gap analysis. *Animal Conservation* 6(1):47-54.
- Peterson, A. Townsend, and C. Richard Robins. 2003. Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology* 17(4):1161-1165.
- Peterson, A. Townsend, and Jeffrey 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(9):919-931.

- Phillips, Steven J., and Miroslav Dudík. 2008. Modeling of species distributions with MaxEnt: New extensions and a comprehensive evaluation. *Ecography* 31(2):161-175. doi: 10.1111/j.0906-7590.2008.5203.x.
- Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231-259.
- Shannon, C. E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27:379-423, 623-656.
- Shapire, Robert. 2017. *MaxEnt Software for Modeling Species Niches and Distributions*. Web page (download site for MaxEnt software and documentation), http://biodiversityinformatics.amnh.org/open_source/maxent/
- Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240(4857):1285-1293.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 08/21/2018		2. REPORT TYPE Final		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Testing Maximum Entropy Analysis to Define Population Distributions				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT	
6. AUTHOR(S) Robert C. Lozar, Scott A. Tweddale, Charles R. Ehlschlaeger, Carey L. Baxter, and Jeffrey A. Burkhalter				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Engineer Research and Development Center (ERDC) Construction Engineering Research Laboratory (CERL) PO Box 9005, Champaign, IL 61826-9005				8. PERFORMING ORGANIZATION REPORT NUMBER ERDC/CERL TR-18-22	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Engineer Research and Development Center (ERDC) Construction Engineering Research Laboratory (CERL) PO Box 9005 Champaign, IL 61826-9005				10. SPONSOR/MONITOR'S ACRONYM(S) CEERD-CZT	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The Urban Security research project attempts to characterize sub-national populations and urban dynamics that impact indicators of urban security threats. This work used the MaxEnt (Maximum Entropy) program to define the probable locations of human habitation for Chittagong Division, Bangladesh. This represents the likely first application of maximum entropy analysis to the delineation of human populations. The model was parameterized with a large number of known house locations and several predictor geographic information system (GIS) data layers that might influence the geographic distribution of human habitation. Analyses and tests done using MaxEnt indicated that the MaxEnt model effectively delineated the probability distribution of human habitation based on a series of inputs based on a series of inputs from three data sources: Landsat NaturalVue TM imagery, SRTM elevation, and OpenStreetMap roads. Results show that, by using remotely-sensed data and “presence only” samples of human habitation, it is possible to estimate the habitation distribution probability based on sound statistical procedures used in MaxEnt.					
15. SUBJECT TERMS Cities and towns, Sociology, Military, Human beings--Geographical distribution, Geospatial data—Analysis, Geographic information systems, Maximum Entropy Analysis, MaxEnt					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT SAR	18. NUMBER OF PAGES 34	19a. NAME OF RESPONSIBLE PERSON
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code)