Positional error propagation analysis in habitat distribution modelling

Babak Naimi
Faculty of Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands
Graduate School of the Environment and Energy Science and Research Branch, Islamic Azad University, Tehran, Iran
naimi@itc.nl

Andrew K. Skidmore, Nicholas A.S. Hamm, and **Thomas A. Groen**
Faculty of Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands

Abstract—**This study examines how robust habitat distribution models are to uncertainty in the position of species occurrence. An artificial species was simulated and mapped in southern Spain (Malaga) and error was introduced to the location of samples. Three commonly used habitat distribution modelling algorithms (GAM, BRT, and MaxEnt) were selected. The propagation of error into the predictions was then analyzed using Monte Carlo (MC) simulation. The models were evaluated for overall performance using the area under receiver operating characteristic curve (AUC). The Root Mean Square Error (RMSE) was also calculated to assess the accuracy of probabilities predicted at grid cells. The results indicate only a small decline in the performance of models with introduced error in species position. Visualizing of RMSEs at grid cells indicates that uncertainty varies with location.**

Keywords: Habitat distribution modeling; positional uncertainty; spatial error propagation

I. INTRODUCTION

Habitat distribution modelling relies on statistical relationships between observations of individuals of a species and environmental variables (Guisan and Thuiller, 2005, Guisan and Zimmermann, 2000). They can be useful for understanding the distribution in geographic space (Peterson, 2006). Despite the wide use of these models, they are subjected to error and uncertainty (e.g. bias in sampling design, location error, etc.) that raise challenges about applicability and validity of the results. This study is concerned with the positional error in species occurrence data. An implicit assumption in the habitat distribution models is that both the species occurrence and the associated environmental conditions are measured without positional errors (Osborne and Leitao, 2009), whilst this will not often be true. Uncertainty about the position of an observation of a species will decrease the accuracy of the prediction of these models.

In this paper a species distribution was simulated and used to analyze the impact of species location error on the predictions of probability of occurrence from habitat distribution models. Three commonly used habitat modelling methods were selected, of which two need presence/absence records of species, and one needs presence-only records.

II. METHOD

A. Simulated Dataset

A species was simulated and mapped in Malaga province in southern Spain. Three species response curves (to precipitation, southness, and land cover) were defined. A Gaussian (bell-shaped) and a linear response to precipitation and southness were considered. As land cover is a categorical variable, discrete suitability values were allocated to classes (Fig. 1). Following Elith and Graham (2009),to simulate the species occurrence location, the habitat suitability is given as a function of response to precipitation, southness, and land cover (Eq. 1).

$$HS = SI_p \times 0.5 \times (SI_s + SI_{lc}) \quad (1)$$

where HS is the habitat suitability of a grid cell; SI_p, SI_s, and SI_{lc} are suitability index of precipitation, southness, and land cover, respectively.

The precipitation variable was derived from the Worldclim dataset (Hijmans et al., 2008; a set of global climate grids with a spatial resolution of 1 km^2). The aspect variable was derived from the Shuttle Radar Topography Mission (SRTM) dataset with a resolution of 90 m. The aspect was then converted to southness. The Corine land cover map with a resolution of 250 m was used as the categorical predictor. All three environmental predictors were aggregated to 1 km resolution by taking the mean for the continuous variables and majority class for the categorical variable.

The suitability index (SI) for each predictor based on species response to them was calculated. The habitat suitability (HS) derived from Eq. 1, shows the suitability (was scaled to range between 0 and 1) of each grid cell for the species (Fig. 2). By adopting a threshold of 0.5, the habitat suitability scores were converted to binomial distribution as presence/absence pattern. Next 100 sites were selected randomly, as species occurrence sample points. These were used to train habitat distribution models. Another independent 100 sites were selected randomly which were used to evaluate the results. The absence records were excluded from the

presence/absence data and remaining points were used for the model requires presence data.

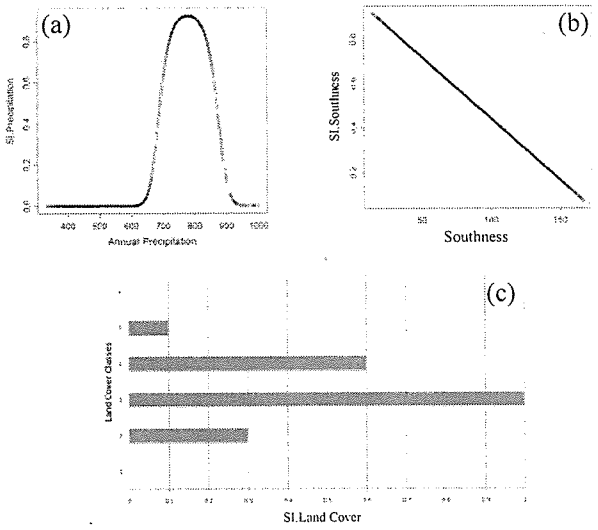

Figure 1. Species response to: (a) precipitation; (b) southness; (c) land cover

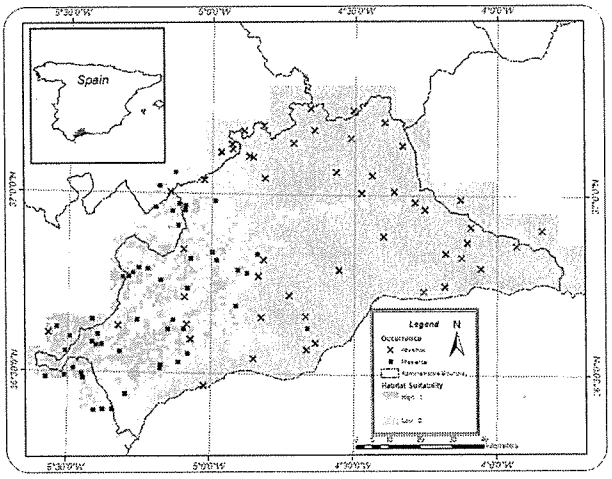

Figure 2. Simulated habitat suitability and species occurrence (presence/absence) points for the study area in southern of Spain (Malaga)

B. Habitat Distribution Modelling and Evaluation

To develop habitat distribution models, three commonly used algorithms were selected: Generalized Additive Models (GAM; Hastie and Tibshirani, 1990), and Boosted Regression Trees (BRT; Friedman, 2001) require presence/absence records; and Maximum Entropy (MaxEnt; Phillips et al., 2006) requires presence-only records.

GAM uses non-parametric and non-linear functions to link the mean of the response variable to a combination of explanatory variables. BRT fits complex non-linear relationships by combining two algorithms of regression trees and boosting. MaxEnt uses a maximum entropy density estimation algorithm to approximate the true distribution of species respecting the constraints (the mean of each environmental variable over the presence sites).

The independent presence/absence test data was used to evaluate the performance of habitat distribution models. Area under the curve (AUC) of a receiver operating characteristic (ROC) plot was selected as the evaluation method (Fielding and Bell, 1997). A ROC curve plots sensitivity values (true positive fraction) on the y-axis against their equivalent (1-specifity) values (false positive fraction) for all thresholds on the x-axis. AUC is a threshold-independent metric and provides a single measure of model performance. The AUC score varies from 0 to 1. An AUC value less than 0.5 indicates discrimination worse than chance, a score of 0.5 implies random predictive discrimination and a score of 1 indicates perfect discrimination.

C. Positional Error Propagation

The positional uncertainty of a species occurrence point leads to a shift in the point's position in the X- and Y-directions (Heuvelink *et al.*, 2007). The error in the coordinates of a species point location was assumed to be normally distributed. A normal probability density function (PDF), therefore, was assigned to the X and Y coordinates to define the error distribution of point locations. The PDF was characterized by considering the true X/Y, and an error with zero mean and standard deviation of 3000 m. 1000 realizations of presence/absence points within the corresponding PDF in each location were realized. The Data Uncertainty Engine (DUE) (Brown and Heuvelink, 2007), a prototype software tool for assessing uncertainties in environmental data, was used for generating realizations. These realizations were then used as input data to run the models (so called Monte Carlo simulations, MC). The outcomes of the simulations were evaluated for the performance to calculate the upper and lower boundary and variation of the prediction accuracy. To examine the stability of prediction (Bishop *et al.*, 2006) at grid cell level, the Root Mean Square Error (RMSE) was calculated for each grid cell:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P(x_0) - P(x_i))^2}{n}} \quad (2)$$

where $P(x_0)$ is the predicted probability of occurrence at the grid cell using data without positional error, $P(x_i)$ is the predicted probability using data with positional error for realization i, and n is the number of realizations.

III. Results

The AUC for the models used data without positional error are compared with the average of performances resulted from the models that had added positional error through MC simulation (Table 1). Although all models showed a decline in performance when species locations were subjected to error, they are still robust as judged by AUC. AUC values > 0.75 generally indicate reasonable discrimination for use in different applications (Graham *et al.*, 2008; Pearce and Ferrier, 2000). Fig. 3 illustrates the variation of AUC over 1000 iterations for each model. It shows that the MaxEnt

performed better according to AUC score and the variation of AUC through the MC simulation.

Table 2 contains the summary statistics of RMSE for all grid cells (i.e. across the whole region). The results indicate that there is a systematic variability in the RMSE – with larger values occurring in the west than in the east. The spatial pattern of RMSE calculated at each grid cell for each model is presented as a map (Fig. 4). These patterns indicate that the western parts of the study area are more sensitive to positional error in sample location compared to the other parts

TABLE I. MEAN AND STANDARD DEVIATION (SD) OF THE AUC FOR THE CONTROL MODEL (USED ORIGINAL DATA) AND THE MODELS USED PERTURBED DATA

Modelling algorithm	Control AUC	AUC based on MC simulation	
		Mean	SD
GAM	0.939	0.909	0.045
BRT	0.999	0.921	0.054
MaxEnt	0.986	0.938	0.032

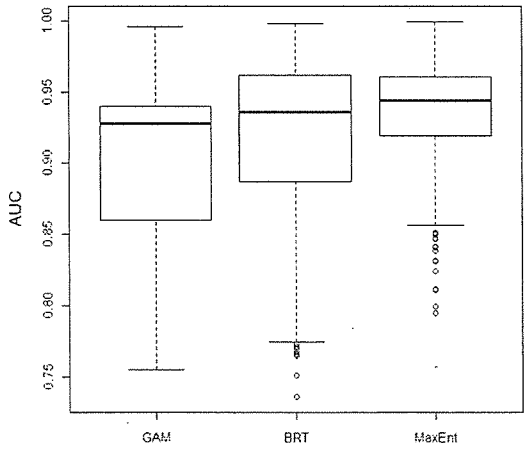

Figure 3. Variation of the AUC through MC simulation for the habitat distribution models

TABLE II. SUMMARY STATISTICS OF RMSE OF PROBABILITY FOR HABITAT DISTRIBUTION MODEL PREDICTIONS

Modeling algorithm	Mean	SD	Min	Max
GAM	0.226	0.261	0.0002	0.924
BRT	0.19	0.215	0.0002	0.803
MaxEnt	0.119	0.121	0.0002	0.5041

IV. DISCUSSION

Species occurrence data often include museum, natural history collections, herbaria etc. that are increasingly available through data portals on the Internet (Newbold, 2010). These data are subjected to error in position. In this paper a procedure for exploring and quantifying the uncertainty propagated by error in species locations is presented.

The results indicate the models are not very sensitive to positional error when judged overall by AUC. Comparing the predicted probability of occurrences at the level of a grid cell, however, we see that in space the accuracy of models differ and that uncertainty is higher in the west than in the east. It might be because of geographic heterogeneity in the west. Logically, error in species location will matter less if an area is more homogenous, because shifted location shares the same environmental features as the true location (Osborne and Leitao, 2009).

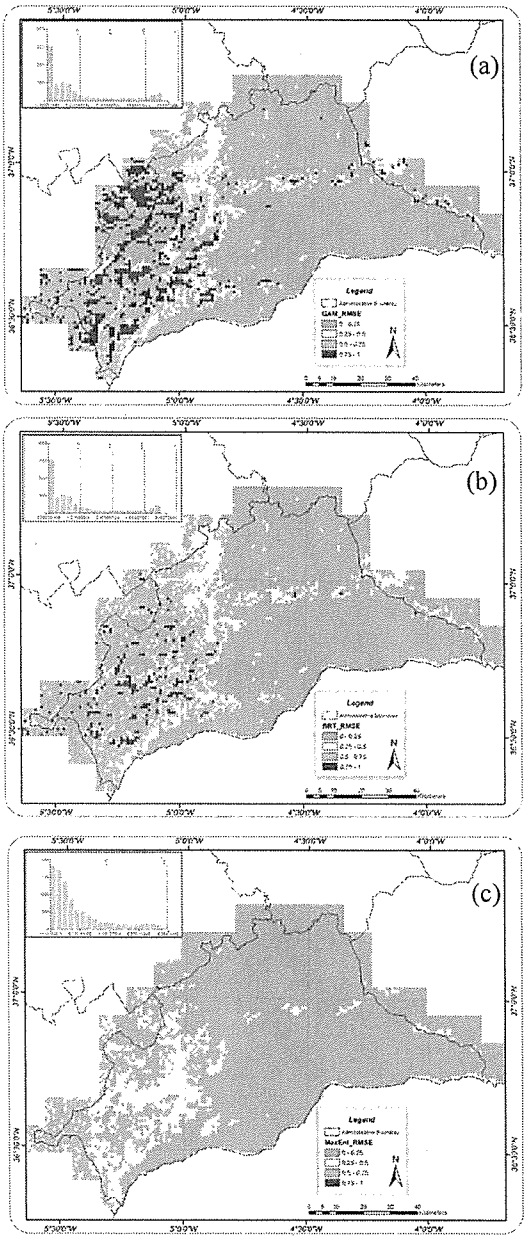

Figure 4. Spatial pattern of RMSE for predicted probability of occurrence at grid cell lever for the habitat distribution models: (a) GAM; (b) BRT; (c) MaxEnt;

ACKNOWLEDGEMENTS

The authors would like to acknowledge the European Union for the funding of the research as Erasmus Mundus program (2007/1139/001-001 MUN ECW), which this paper

is based. We thank Aidin Niamir for his support in preparing the predictor variables.

REFERENCES

Bishop, T. F. A., Minasny, B. and Mcbratney, A. B. (2006). Uncertainty analysis for soil-terrain models. *International Journal of Geographical Information Science*, 20, 117-134.

Brown, J. D. and Heuvelink, G. B. M. (2007). The Data Uncertainty Engine (DUE): A software tool for assessing and simulating uncertain environmental variables. *Computers and Geosciences*. 33, 172-190.

Elith, J. and Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*. 32, 66-77.

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*. 24, 38-49.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 29, 1189-1232.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Peterson, A. T. and Loiselle, B. A. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*. 45, 239-247.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*. 8, 993-1009.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*. 135, 147-186.

Hastie, T. and Tibshirani, R. (1990). *Generalised additive models*. London: Chapman and Hall.

Heuvelink, G. B. M., Brown, J. D. and Van Loon, E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*. 21, 497-513.

Hijmans, R., Cameron, S., Parra, J., Jones, P., Jarvis, A. and Richardson, K. (2008). *WorldClim version* 1.4.

Mccullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd ed., London: Chapman and Hall.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*. 34, 3-22.

Osborne, P. E. and Leitao, P. J. (2009). Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*. 15, 671-681.

Pearce, J. and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*. 133, 225-245.

Peterson, A. T. (2006). Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*. 3, 59-72.

Phillips, S. J., Anderson, R. P. and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*. 190, 231-259.