# ON UNCERTAINTY IN SPECIES DISTRIBUTION MODELLING

**Babak Naimi**

Examining committee:

| | |
|---|---|
| Prof. dr. Tom A. Veldkamp | University of Twente, ITC |
| Prof. dr. Alfred Stein | University of Twente, ITC |
| Prof. dr. Peter M. Atkinson | University of Southampton, UK |
| Prof. dr. Carsten F. Dormann | University of Freiburg, Germany |

ITC dissertation number 267
ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands

ISBN 978-90-365-3840-4
DOI 10.3990/1.9789036538404

Cover designed by Job Duim
Printed by ITC Printing Department
Copyright © 2015 by Babak Naimi


UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

# ON UNCERTAINTY IN SPECIES DISTRIBUTION MODELLING

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, February 18, 2015 at 16:45 hrs

by
**Babak Naimi**

born on 24 August 1975.
in Gorgan, Iran

This thesis is approved by

**Prof. dr. Andrew K. Skidmore**, promoter
**Dr. Thomas A. Groen**, co-promoter
**Dr. Nicholas A. S. Hamm**, co-promoter

# Acknowledgements

The time for conducting t this PhD research at ITC has been inspiring and enjoyable. This dissertation could not have been finished without the support of family members, friends, colleagues all of whom I made deeply grateful to.

First and foremost, I would like to express my sincerest appreciation to my supervisors, Prof. Andrew Skidmore, Dr. Thomas Groen, Dr. Nicholas Hamm and Dr. Bert Toxopeus. Andrew, I'll never forget the first day we met in 2009. Thanks for your trust and giving me this wonderful opportunity to join your group and for all your suggestions, constructive criticism, and incredible patience to guide me through my research. I have learned a lot from you. Thomas, you are a wonderful supervisor! I had a great freedom to plan and execute my ideas in research without any pressure. This made me to identify my own strength and drawbacks, and particularly boosted my self-confidence. It was great working with you Thomas, and my heartfelt thanks to you. I hope that there will be many more collaborations between us in the future. Nick, thank you for all nice and tough scientific discussions we had. Your frequent constructive criticism and endless remarks were always resulted in fruitful discussions with healthy outcomes. I appreciated your way of correcting the manuscripts and your availability to listen and help whenever needed. Bert, I always enjoyed talking with you. You are very encouraging, friendly and supportive person.

In my daily work I have been blessed with a friendly, cheerful and international group of fellows, you guys have been true gems throughout these years. Without the friends in Enschede, life would have been boring and dull. First, I would like to thank my amazing friends Aidin Niamir, Sam Khosravifard, Sanaz Salati, Farhang Sargordi, and Nima Moin. Aidin, it was wonderful to share my office with you. The interesting discussions I had with you on various topics were constructive and a relaxing time from research. Sanaz, thanks for your company in this long journey and for all those enjoyable moments, and good laughs we had. Sam, I enjoyed all the chats and beautiful moments we shared together. Nima, thanks for your advices and being by my side through my worst time. Farhang, thanks for

being such a great person and nice friend. I never forget our “Harekat zadanha & yaroo shodanha”!

During my PhD, I had great scientific discussions and collaborations with Dr. Alexey Voinov, Dr. Tatiana Filatova and Dr. David Rossiter. Alexey, you are a great person, and it has been an honour to collaborate with you. I learned a lot from your constructive comments and advices on all mattes I discussed with you. Tatiana, you are a nice friend. I enjoyed the discussions we had on your interesting scientific works. David, your positive words about me and my skills have always been encouraging.

I extend true thanks to Loes Colenbrander, who took care of my graduate program from very beginning to the graduation ceremony, to Esther Hondebrink for arranging all meetings in NRM department, to Theresa van den Boogaard, and Bettine Geerdink who dealt with all the bureaucracy that comes along with an international PhD student, to Marion Pierik who did her best to put down all the financial anxieties, to the staff of ITC library for providing me all necessary research materials, journal papers and books. I would like to express my sincere appreciation to other NRS members and ITC staff for their care and facilitating various issues, which made my life easier at ITC. I would like also to acknowledge the Erasmus Mundus program, and ITC research fund for granting me scholarship.

I am heartily grateful to my PhD friends Elnaz, Parinaz, Mitra, Si, Razieh, Marshal, Fatemeh, Sabrina, Amjad, Khan, Salim, Juan and the other PhD fellows. I wish you success in your study and career.

I would like to also express my appreciation and love to my wonderful parents. Thanks for your invaluable support, encouragement and endless love. I also would like to thank my sister Mitra and brothers, Reza, Amir and Ramin, for unconditional love and support.

My strongest support came from my wife, Shirin Taheri. Shirin, you have always been the best friend through my ups and downs. Your understanding, constant and great support and devotion before and during this PhD research always made me stronger.

# Table of Contents

# List of figures

# List of tables

# *Chapter 1*

# General Introduction

# 1. General introduction

## *1-1. Introduction*

Species distribution models (SDMs) are important tools for many applications in ecology, evolution, conservation planning and understanding of environmental impacts (*e.g.*, climate change). A key component in such predictive models is characterization of species distribution in ecological space based on quantifying species-environment relationships that can be useful in understanding of the potential distribution in geographic space(Dormann 2011 ; Peterson 2006).

There are several terminologies including habitat suitability models, habitat (or species) distribution models, resource selection functions, ecological niche models, which all address similar issues with different tools in many studies (Hirzel and Le Lay 2008). A striking characteristic of these models is their reliance on the ecological niche theory (Guisan and Thuiller 2005 ; Guisan and Zimmermann 2000). Guisan and Zimmerman (2000) provided a review of the state-of-the-art in pre-2000 period of species distribution modelling. They defined SDMs as "empirical models relating field observations to environmental predictor variables, based on statistically or theoretically derived response surfaces". Species data can be presence-only, presence-absence or abundance observations based on random or stratified field sampling, or observations obtained opportunistically, such as those in natural history collections (Guisan and Thuiller 2005). Environmental predictors reflect three main types of influences on species (Austin 2002 ; Guisan and Thuiller 2005): (i) limiting factors, defined as factors controlling species eco-physiologically (*e.g.*, temperature, water, soil); (ii) disturbances, defined as all types of perturbations affecting environmental systems; (iii) resources, defined as all compounds that can be assimilated by organisms (*e.g.*, energy). Six steps for the procedure of species distribution modelling, are discussed by Guisan and Thuiller (2005) as: (i) conceptualization; (ii) data preparation; (iii) model fitting; (iv) model

evaluation, (v) spatial prediction, and (vi) assessment of model applicability.

Many models are now available for predictive species distribution modelling. These models vary in how they treat the distribution of the response variable (species), selecting relevant predictors, defining functional response for each predictor, weighting variable contributions, incorporating interactions, and finally predicting geographical patterns of species occurrences. These methods have been employed in many studies as a tool to model geographical distribution of species.

## *1-2. Problem statement*

Despite the wide use of SDMs, important challenges about the applicability and validity of these models exist concerning errors and uncertainty. More reliable and robust models are essential for environmental management and for assessing the impacts of changing environmental conditions (Guisan et al. 2006). Therefore, it is essential to study the effect of data and model uncertainty (Araújo and Guisan 2006).

Uncertainty is incomplete knowledge or lack of confidence about possible outcomes and/or probabilities of these outcomes (Refsgaard et al. 2007 ; Regan et al. 2002). Reasons for this lack of confidence might include incomplete (e.g., due to insufficient sampling or bias in data), inaccurate (e.g., due to measurement errors), or unreliable information. In terms of causes of uncertainty, we can broadly specify two types of uncertainty (Li and Wu 2006)including data uncertainty (uncertainty due to data quality, availability, and inherent variability), and model uncertainty (due to model structure, and model parameterization).

Although there is increasing attention to some important aspects of error in recent SDM studies, there is little appreciation for the fact that there are many different dimensions of uncertainty. Moreover, there is a lack of understanding about their different characteristics, relative magnitudes, and available means of dealing with them (Walker et al. 2003). It has been

argued that comprehensive research is needed to acquire further knowledge and understanding of different types of uncertainty (*e.g.*, knowledge, data, models, and applicability) inherent in SDMS and how they affect the applicability of SDMs predictions (Araújo and Guisan 2006 ; Guisan et al. 2006).

Some degree of uncertainty is unavoidable in modelling because there are always errors associated with the stochastic nature of ecological processes, system complexity caused by spatial heterogeneity, unreliability and unavailability of data, and imperfection of models (Li and Wu 2006 ; Regan et al. 2002). There is little consensus in how to define uncertainty, what its characteristics are, and how we should relate these characteristics to the appropriate treatment or management of uncertainty. However, many sources of uncertainty in modelling can be quantified and reduced, and different sources of uncertainty be ranked with respect to their relative contributions to error in model output (Li and Wu 2006).

## *1-3. Research objectives*

The research presented in this thesis aims to investigate the impact of uncertainty on predictive species distribution modelling, and to explore possible solutions to deal with them. In particular, the following objectives are addressed in this thesis:

(i) to investigate the impact of different sources of uncertainties (*i.e.*, positional uncertainty in data and model uncertainty) on predictive species distribution models;
(ii) to discuss and explore solutions to quantify and visualize these uncertainties.

## *1-4. Thesis outline*

This thesis contributes to the understanding the impact of data and model uncertainty on species distribution models and provide some solutions for these problems.

Chapter 1 presents a general introduction to the thesis and brief research background, the statement of problem, research objectives as well as the thesis outline.

Chapter 2 investigates the impact of positional uncertainty in species data and explores if examining (global) spatial autocorrelation in predictors can be a solution to understand the models' robustness to these kind of uncertainty.

Chapter 3 extends the chapter 2 by using local indicators of spatial association to investigate where the positional uncertainty affects the models more.

Chapter 4 addresses the problem of modelling (local) spatial association and introduces a new method to measure local spatial associations.

Chapter 5 explores and visualizes model uncertainty in species distribution modelling.

Chapter 6 provides a synthesis based on the previous chapters, and suggestions for the future works.

# *Chapter 2*

# Positional Uncertainty and (Global) Spatial Autocorrelation

This chapter is based on:

Naimi, B., Skidmore, A. K., Groen, T. A., Hamm, N. A. S. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, 38, 1497-1509.

# 2. Positional Uncertainty and (Global) Spatial Autocorrelation

## *2-1. Introduction*

Species distribution models (SDMs) infer the ecological requirements of species as well as predicting their potential geographic distribution. These models have become important in a range of applications including regional biodiversity assessment, conservation biology, evolutionary biology, epidemiology, wildlife management, conservation planning etc. (Elith et al. 2006 ; Elith and Leathwick 2009 ; Guisan and Thuiller 2005 ; Guisan and Zimmermann 2000 ; Segurado and Araújo 2004 ; Skidmore and … 1996). Species distribution modelling is usually based on statistical relationships between species occurrences and corresponding environmental variables. A key component in this process is estimation of species distribution in ecological space, which can be useful to predict their potential distribution in geographic space (Peterson 2006). Many models are now available to explore this relationship, although these techniques may differ in their ability to summarize useful relationships between response and predictor variables (Segurado and Araújo 2004). These models vary in the kind of species data they use (*e.g.*, presence/absence versus presence only), the form of their output (a continuous or a binary prediction), the type of relationship they assume (from simple linear to complex nonlinear), how they estimate the distribution of the species (using parametric versus nonparametric approaches), how they select relevant predictor variables, whether variable contributions are weighted, and whether they allow for interactions of explanatory variables (Austin 2007 ; Elith et al. 2006 ; Guisan and Zimmermann 2000). Different models may yield different results, even when calibrated with the same response and predictor variables (Araújo and Guisan 2006).

Many SDMs have been developed using presence/absence or presence-only species occurrence data. The great majority of these data, especially in the

form of presence-only from museum or herbarium collections, are increasingly available over the Internet (Elith et al. 2006 ; Naimi et al. 2011). One of the problems with these data is that there is a potential uncertainty about where the occurrence was located. This uncertainty is caused by a variety of factors including positional error of the location, failure to specify the cartographic projection as well as georeferencing error (Graham et al. 2008 ; Graham et al. 2004a ; Rowe 2005).

Recent studies have addressed the impact of positional error in species occurrences on SDM accuracy. Graham *et al.* (2008) explored whether positional error in species occurrence data affected SDM performance, with a focus on comparison of models. They introduced a random error (up to 5 km) to the location of presence-only species data and evaluated how it influenced the prediction accuracy of 10 different models. They concluded that SDMs are, in general, robust to positional errors. Johnson & Gillingham (2008) assessed the sensitivity of a logistic-regression SDM to 20 levels of errors (from 50 to 1000 m) in presence/absence locations of a species, sampling bias, error in environmental data (misclassification of land classes) and model order. Their results showed that the species positional error made the greatest contribution to the reduction in prediction accuracy (Johnson and Gillingham 2008). Osborne & Leitão (2009) explored the impact of extreme and typical positional errors in both species and environmental data on the performance and ecological interpretation of three different SDMs. Concurring with Graham *et al.* (2008) their results showed that species positional errors had a small effect on the predictions from many models. Osborne & Leitão (2009) also opened an interesting issue, namely that the impact of positional errors on SDMs may be understood by examining spatial autocorrelation in predictor variables. They examined the relationship between spatial autocorrelation (quantified by Moran's *I*) in predictors and the consistency of the contribution of variables to models for four scenarios of positional error. They proposed that, if high spatial autocorrelation reduces the impact of positional error, it should show a link between autocorrelation of a variable

and the consistency of that variable's contribution to a model. They found a weak but significant relationship, $R^2 = 0.32$, and concluded that the degree to which spatial autocorrelation confers resilience to positional error is more complex than they first expected and needs further study.

Spatial autocorrelation is a statistical property of most ecological variables (Legendre 1993) and represents the relationship between values of the given variable at different geographical separations. It is hypothesized that, in species distribution modelling, errors in species location will matter less if nearby locations have similar environmental characteristics to the true location (Osborne and Leitão 2009). Therefore, SDM robustness to species positional uncertainty is expected to be affected positively by spatial autocorrelation in environmental variables.

This study aims to assess and test the interaction between spatial autocorrelation in predictors with positional uncertainty in species occurrences. Different scenarios were designed to assess the propagation of positional uncertainty through SDMs. This experiment was conducted by implementing a series of commonly used species distribution models.

## *2-2. Materials and methods*

The study was implemented in four main stages, as follows.

a) Designing scenarios and simulating artificial datasets. Each scenario included two environmental variables and the distribution of a virtual species.
b) Introducing error at the locations of species occurrence and generating realizations of uncertainty in the positions.
c) Positional uncertainty propagation using Monte Carlo simulation for seven SDMs.
d) Evaluating the results for prediction performance and assessing their interaction with spatial autocorrelation in predictors.

## 2-2-1. Simulating artificial datasets

To explore the interaction between spatial autocorrelation and positional uncertainty, both were varied in a controlled way over a range of values. To achieve this, environmental variables (predictors) and species occurrences (response variable) were simulated. The variogram was used to control the spatial autocorrelation in the predictors. The variogram is commonly used to model the spatial structure in a single variable. Formally it is defined as half the expected squared difference (half the variance of the difference) in the variable value at a specific geographical separation. The variogram parameters are (Webster and Oliver 2007): (1) the sill, which is the total variance and represents the variability in the absence of spatial correlation; (2) the range, which is the distance at which the variogram approaches the sill; and (3) the nugget effect, which is a combination of spatially unstructured variance (*e.g.,* attribute error) and spatially structured variance at distances shorter than the minimum measurement separation. The sill minus nugget is known as partial sill or structural variance (Fig. 2-1).

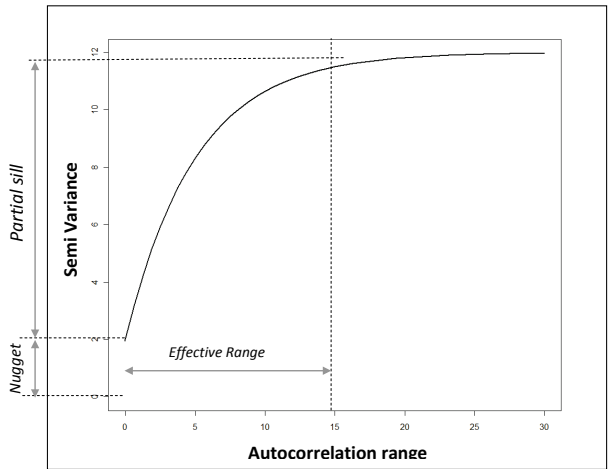



**Figure 2-1:** The exponential variogram model for range = 15, nugget = 2 and partial sill = 10

Datasets were generated covering a range of spatial autocorrelation in predictor variables. Each dataset included two artificial environmental gradients and the distribution of one virtual species. Unconditional simulation was used to construct regular grids of 150 × 150 cells for each environmental gradient. Unconditional simulation is a geostatistical technique that generates a realization of a spatially correlated variable, where the spatial correlation is defined by a variogram (Dungan 1999). Conditional simulation generates a realization using a defined variogram and measurements from the field. Various computational algorithms are available for implementing conditional and unconditional simulation, but all have the objectives of generating a surface with the appropriate correlation structure, as defined by the variogram. For this research the circulant-embedding algorithm (Dietrich and Newsam 1993) implemented in the RandomFields package v. 1.3.41 in the R programming environment (Schlather 2009) was used. An exponential variogram model with a sill of 10, and a nugget of 0 was used for all datasets. The variogram models were assigned different values for the range parameter to control the extent of spatial autocorrelation in the predictor variables. In total, 30 levels of range size, from 1 to 30 grid cells, were used, giving a transition from minimum spatial autocorrelation (range = 1) to relatively large scale spatial autocorrelation (range = 30). Additionally, a white-noise surface (range = 0) was simulated giving a total of 31 scenarios of spatial autocorrelation.

The distribution of the virtual species was simulated based on assumed species response curves, representing the probability of presence along an environmental gradient. They were constructed with a Gaussian and a decreasing linear function in response to the first and second predictor, respectively (Fig. 2-2). These functions were used to create an individual measure of habitat suitability for each predictor. The suitability index based on a Gaussian response curve was modelled using *Eq.* 2-1 ( ter Braak and Looman 1986):

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 = P_{\max} - \frac{(x-\mu)^2}{2t^2} \quad (2\text{-}1)$$

where $\beta_0$, $\beta_1$, and $\beta_2$ are regression parameters; $x$ is the environmental variable; $p(x)$ is the individual suitability index, $\mu$ is the species' optimum, which is the value of the environmental variable where $p(x)$ is at a maximum ($P_{\max}$); $t$ is the species' tolerance (a measure of ecological amplitude). The parameters $P_{\max}$, $\mu$, and $t$ are defined as follows ( ter Braak and Looman 1986):

$$\begin{aligned} \mu &= \beta_1/2\beta_2 \\ t &= 1/\sqrt{(-2\beta_2)} \\ P_{\max} &= 1/\{1 + e^{-\beta_0 - \beta_1\mu + \beta_2\mu^2}\} \end{aligned} \quad (2\text{-}2)$$

Parameter values of $P_{\max} = 1$, $\mu = 50$, and $t = 1.1$ were used for the Gaussian response in this study.

*Eq*. 2-3 was used to model the habitat suitability based on a linear response curve:

$$p(x) = \begin{cases} 1, & x < x_{low} \\ 1 - \left(\frac{x - x_{low}}{x_{high} - x_{low}}\right), & x_{low} \leq x \leq x_{high} \\ 0, & x > x_{high} \end{cases} \quad (2\text{-}3)$$

where $x$ represents the environmental variable; and $x_{low}$ and $x_{high}$ represent the range between where the species starts to experience negative effects from the environmental variable, causing a decline in its probability of occurrence. Parameter values of $x_{low} = 45$ and $x_{high} = 55$ were used for the linear response in this study.

Finally, the habitat suitability (Hirzel et al. 2001) for both environmental variables were combined at a grid cell resolution using *Eq.* 2-4:

$$HS = \sum_{i=1}^{2} w_i x_i \quad with \quad \sum_{i=1}^{2} w_i = 1 \qquad (2\text{-}4)$$

where $HS$ represents the habitat suitability score for the virtual species, $w_i$ denotes the weight or importance of variable $i$, and $x_i$ denotes the habitat suitability index based on environmental variable $i$. The same weight (0.5) was applied for both variables. $HS$ can be interpreted as probability of species occurrence. The habitat suitability values were used to realize presence and absence of the virtual species by applying a threshold of 0.5. One hundred and fifty randomly selected sites were used to train the SDMs and another 150 sites were randomly selected to evaluate the predictive performance of the models. The prevalence (the proportion of presence sample points) for both training and test data was 0.5. To provide sample points for methods that require presence-only data, the presence records in presence/absence samples were used. Fig. 2-2 illustrates schematically the procedure to simulate the artificial datasets.

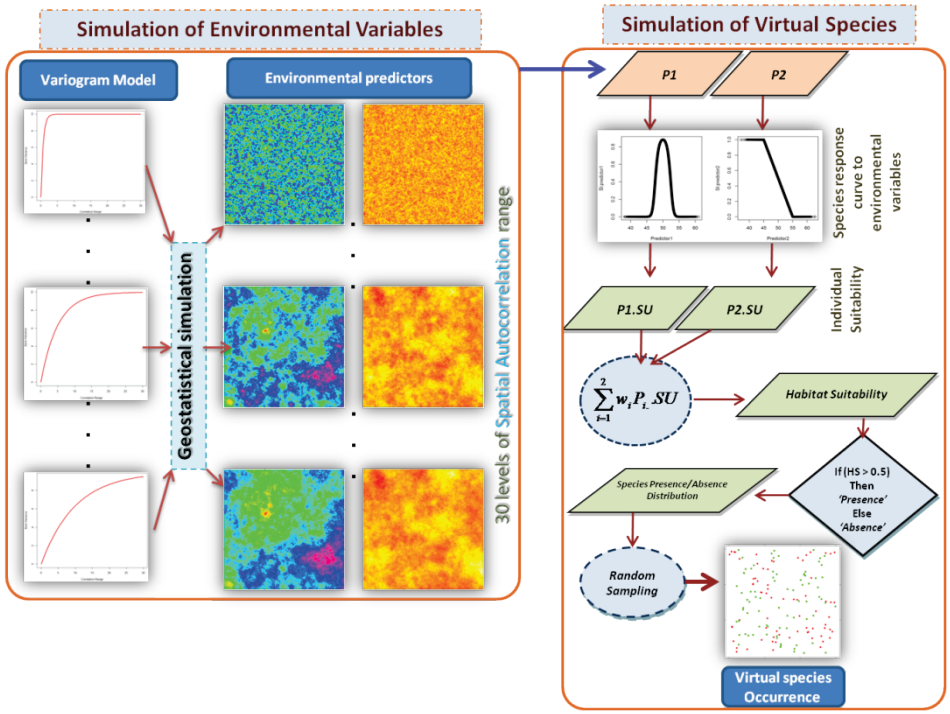



**Figure 2-2:** Flow diagram showing the procedure for simulating datasets

## 2-2-2. Species distribution modelling

To develop SDMs, several commonly implemented models that use presence/absence or presence-only records of species occurrences were selected. The presence/absence models were generalized linear models (GLM; McCullugh & Nelder, 1989), generalized additive models (GAM; Hastie & Tibshirani, 1990), boosted regression trees (BRT; Friedman, 2001), multivariate adaptive regression splines (MARS; Friedman, 1991) and random forests (RF; Breiman, 2001) and the presence-only models were maximum entropy (Maxent; Phillips et al., 2006), and genetic algorithm for rule-set production (GARP; Stockwell & Noble, 1992).

The GLM, GAM, MARS, BRT and RF models were implemented in the R development environment v. 2.8.1 (R Development Core Team 2008) using the bioclimatic niche modelling (biomod) v. 1.1-6.1 package

(Thuiller et al. 2009). This enables SDMs to be run simultaneously and incorporates several features for evaluating and examining species-environment relationships. To run the GARP model, the openModeller framework v. 1.0.9 was used. This framework was developed to perform the most common tasks related to species distribution modelling (De Souza Muñoz et al. 2009). All versions of GARP are available in this framework (see below). Maxent was run by using the Maxent software v. 3.3.1 that was developed and introduced by Philips et al. (2006). The specifics of each model are summarized as follows:

- *Generalized linear models (GLM)*: This method uses parametric functions to link the response variable to a linear, quadratic, and/or cubic combination of explanatory variables (Austin 2002 ; Mccullagh and Nelder 1989). Here, a GLM ordinary polynomial with an automatic stepwise model selection based on Akaike information criterion (AIC) was used. For simplicity we refer to this as the 'GLM' approach.
- *Generalized additive models (GAM)*: This method uses nonparametric and data-defined, smoother to fit, nonlinear functions (Austin 2002 ; Hastie and Tibshirani 1990). Here, a GAM with a cubic spline smoother and an automated stepwise process was used.
- *Boosted regression trees (BRT)*: This method (Elith et al. 2008 ; Friedman 2001) fits complex nonlinear relationships by combining two algorithms of regression trees (relate a response to their predictors by recursive binary splits) and boosting (an additive method to combine many single models to improve the performance). The recommended default settings (maximum number of trees = 3000, learning rate = 0.001) were used.
- *Multivariate adaptive regression spline (MARS):* This method is similar to GAM but uses a piece-wise linear basis function (Friedman 1991 ; Leathwick et al. 2005).
- *Random forests (RF):* This method selects many bootstrap samples from the data and generates and fits a large number of regression trees to

each of these subsamples. Each tree is used to predict the out-of-bag observations (i.e. those that were not selected as bootstrap samples). The classification given by considering each tree as a 'vote', and the predicted class of an observation is determined by the majority vote among all trees (Breiman 2001 ; Cutler et al. 2007). Models presented here had 500 trees with one variable randomly selected from the 2 candidates at each split.

- *Genetic algorithm for rule-set production (GARP):* GARP uses a genetic algorithm with an iterative process to produce a set of conditional rules in the form of 'if-then' statements that describe the ecological niches of the species under study (Anderson et al. 2003 ; Peterson et al. 2007 ; Stockwell and Noble 1992). The openModeller-GARP (De Souza Muñoz et al. 2009) followed by the 'best subset' procedure was used in this study. The 'best subset' procedure was originally developed to sift through the model-to-model variation generated by the random-walk nature of the GARP algorithm.

Spatial autocorrelation in the model residuals can be a problem for regression-based techniques where random sampling and independent residuals are important assumptions (Dormann et al. 2007). The Moran's I test and correlogram was used to estimate the correlation in the GLM residuals as a function of geographic distance (Schabenberger and Gotway 2005). The technique for linearly recovered errors (residuals) (LRE), described by Schabenberger & Gotway (2005, p. 315) was followed.

### 2-2-3. Model evaluation

The predicted distributions of both presence-only and presence/absence SDMs were evaluated for their performance using independent presence/absence data. It is important to use more than one metric to assess model performance because each quantifies different aspects of predictive performance (Elith and Graham 2009). Two methods were therefore used to measure the predictive performance of models: area under the curve (AUC) of a receiver operating characteristic (ROC) plot,

and Cohen's Kappa. A ROC curve plots sensitivity values (true positive fraction) on the y-axis against '1 – specifity' values (false positive fraction) for all thresholds on the x-axis (Fielding and Bell 1997). AUC is a threshold-independent metric and provides a single measure of model performance. AUC scores vary from 0 to 1. AUC values less than 0.5 indicate discrimination worse than chance, a score of 0.5 implies random predictive discrimination and a score of 1 indicates perfect discrimination.

Kappa is a proportional agreement between predictions and observations after removing the agreement expected to occur by chance (Cohen 1960). The Kappa ranges from –1 to +1, where +1 indicates perfect agreement, a value of 0 implies agreement expected by chance, and a value less than 0 indicates agreement worse than chance. This statistic is calculated from a confusion matrix which is a cross-tabulation of observed and predicted values. The calculation, therefore, is dependent on a threshold to reclassify predicted probabilities into binary values (presence/absence). In this study, a constant threshold of 0.5 was applied to all cases. This threshold is the same as that adopted when simulating the virtual species (see section simulating artificial datasets). The evaluation statistics were calculated using the PresenceAbsence package v. 1.1.3 (Freeman and Moisen 2008) implemented in R (R Development Core Team 2008).

### 2-2-4. propagation of positional uncertainty

Positional uncertainty in species occurrence leads to a shift in the point's position in the $X$ and $Y$directions (Heuvelink et al. 2007). A probabilistic approach was used to introduce a positional error ($\varepsilon$) in species occurrences that had no directional bias. Taking $\varepsilon \sim N(0,1)$ gives a normally distributed unbiased error with a standard deviation equal to 1 map unit (here 1 grid cell). This was added to the Easting and Northing of each location (Hamm et al. 2004):

$$x_i = Easting + \varepsilon_i$$
$$y_i = Northing + \varepsilon_i \qquad (2\text{-}5)$$

where $i$ refers to each individual species occurrence. Different realizations of the sample were simulated and used to explore the effect of positional uncertainty (Hamm et al. 2004). These were termed the 'perturbed' datasets. Five scenarios with increasing standard deviations, from 1 to 5 grid cells, were applied to explore the performance of the models over a range of positional uncertainties.

For each level of positional error, 1000 realizations of perturbed occurrence points were simulated. These realizations were used to train the models. The idea was to compute the result of the model repeatedly using varied input values (Heuvelink 1999) and then to assess the accuracy of each. This so-called Monte Carlo simulation allowed the assessment of uncertainty. The conceptual framework to run the Monte Carlo simulation is illustrated in Fig. 2-3.

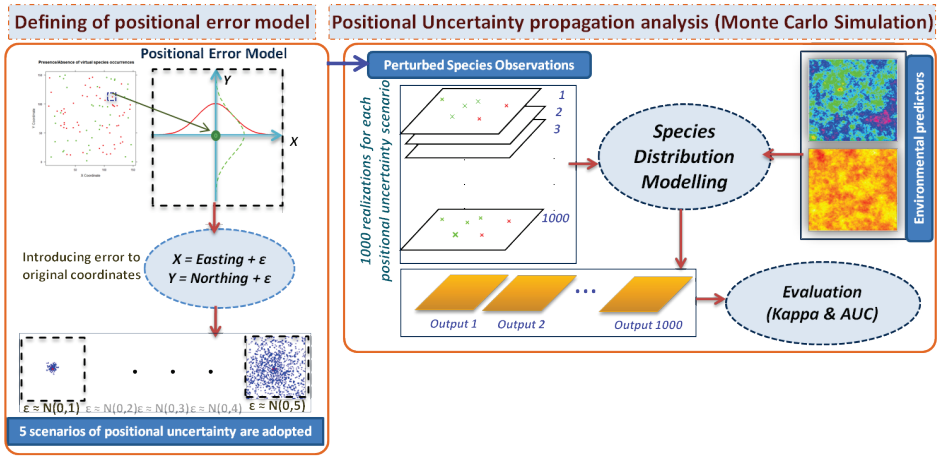



**Figure 2-3:** Conceptual framework of species positional error propagation analysis

## 2-2-5. Interaction of positional uncertainty and spatial autocorrelation

To assess the interaction of positional uncertainty and spatial autocorrelation, all combinations of the scenarios were considered for each SDM algorithm. Therefore, in total 155 scenarios (5 levels of positional

uncertainties × 31 spatial autocorrelation ranges) were applied to 7 SDM algorithms. The performance of 1000 model-runs was then calculated and compared with the performance of models with unperturbed data at different ranges of spatial autocorrelation at each level of positional error. To assess whether spatial autocorrelation range in predictors reduces the impact of positional uncertainty, a two-way Friedman's test (Friedman 1937) with spatial autocorrelation range and positional uncertainty scenarios as factors was applied for each model.

## *2-3. Results*

The Moran's *I* correlograms for the GLM residuals, showed that the residuals were not spatially autocorrelated (Fig. 2-4).

Three simulated environmental gradients (as an example of the generated dataset) with different spatial autocorrelation range (5, 15 and 25) together with their corresponding variogram are illustrated in Fig. 2-5. They exemplify continuous environmental predictors (*e.g.*, temperature, digital elevation model). The distribution of Kappa and AUC values as a result of the Monte Carlo simulation, and the response of this distribution to changes in spatial autocorrelation range of the environmental variables (Figs 2-6, 2-7 and 2-8) were generally consistent in the trend they showed. A similar interaction effect of positional uncertainty and spatial autocorrelation range in predictors on model performance was detected for both AUC and Kappa.

The mean Kappa and AUC for the presence/absence models fitted with the original data were 0.92 and 0.99, respectively, and for the presence-only models were 0.74 and 0.96, respectively. The effect of positional error on model performance was strongly influenced by the spatial autocorrelation (variogram) range. The level of accuracy depends on the spatial autocorrelation range and the level of error. Comparing the graphs suggests that the models in general behave consistently over the range of autocorrelation and level of error.

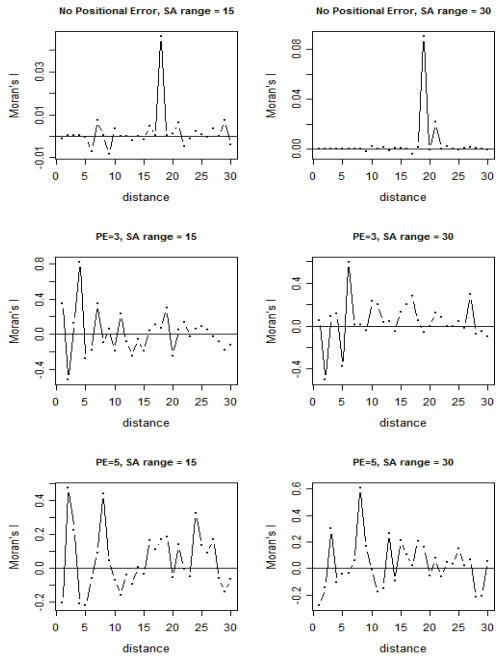



**Figure 2-4:** Moran's *I* correlogram for the GLM residuals for some scenarios: two levels of spatial autocorrelation range (SA range = 15 & 30) and three levels of positional error in occurrence data (No positional error & PE=3 & PE=5)

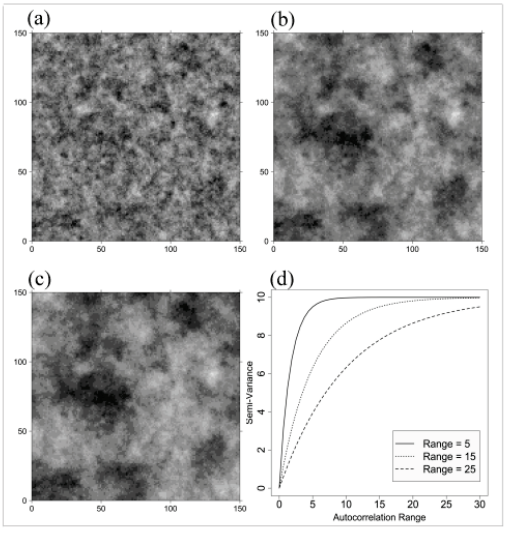



**Figure 2-5:** Three examples of simulated environmental gradients with spatial autocorrelation range of (a) 5, (b) 15, (c) 25 and (d) corresponding variogram models

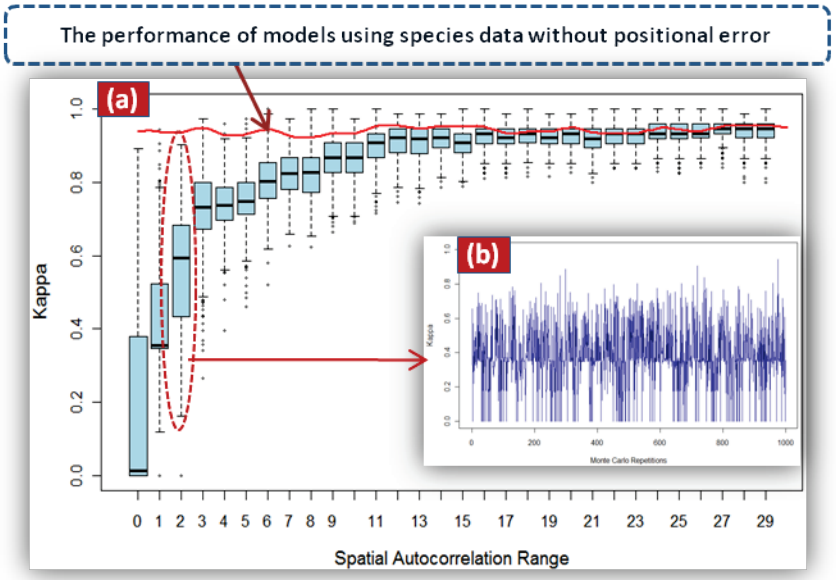



**Figure 2-6:** Variation of model accuracy (Kappa) over the Monte Carlo simulation for different ranges of spatial autocorrelation (x-axis) under the positional uncertainty; (a) For the generalized linear model (GLM) under the positional uncertainty with error $\varepsilon \sim \boldsymbol{N(0,1)}$, the vertical dashed line represents the limit of spatial autocorrelation range that divided the results into two categories to compare the scenarios; (b) Each box represents the results for 1000 Monte Carlo runs

As a normal probability density function (PDF) was used to introduce error in the locations, 99.7% of the perturbed points are expected to be within a distance equal to three times the standard deviation away from the original point. It was therefore expected that the performance of the SDMs would become stable when the range of the spatial autocorrelation of the predictors is larger than three times the standard deviation of the spatial error. This is because the value of the covariate at the perturbed location would then still be similar to that at the true location. This was used to group the results presented in each graph into two categories and to compare the performance of the models between these two categories, and

**Figure 2-7:** Variation of the model accuracy (Kappa) for five positional uncertainty scenarios with increasing standard deviation of error, from one to five grid cells, and seven species distribution models. Each row corresponds to the level of positional error; the different components of the graph are described in Fig. 2-5

over the different positional uncertainty scenarios. The first category included the simulations using predictors with an effective autocorrelation range equal to or lower than three times the standard deviation of the imposed positional error, and the second category included the simulations using predictors with a larger spatial autocorrelation range. Performance of the models was expected to be lower compared to model performance with the unperturbed location dataset when in the first category, and equal to performance of the unperturbed location dataset when in the second category. The summary statistics for the performance measures of the SDMs were then calculated for each category (Table 2-1). The Kappa and AUC values from the first category showed that the models were highly influenced by the positional error. The decline in the Kappa values, in the first category, was in a range between a minimum of 41% and a maximum of 65% in comparison with the Kappa for the model using the unperturbed data. The drop in AUC values was in a range between a minimum of 18% and a maximum of 32% in comparison with the AUC for the model using unperturbed data. Comparing the Kappa and AUC values for the second group showed that the drop was smaller than in the first group, ranging between a minimum of 0.1% and a maximum of 42% for the Kappa, and a minimum of 0% and a maximum of 17% for the AUC, depending on the level of positional error and the model used.

The decline in the performance with an increase of positional error was not equal for all models. For the GLM, GAM and BRT algorithms, the Kappa values for the second category dropped with a similar amount of 21% for the scenarios where the highest level of positional error was introduced (an increase of almost 4% for each level of error). In the same situation, the AUC dropped 5% for both the GLM and GAM, and 8% for the BRT. For the MARS and RF, the decline in Kappa was 27% and 42%, respectively, for the scenarios where the highest level of positional error was introduced (almost 5% for the MARS and 8% for the RF for each level of error), and this decline in the AUC was 11% and 17%, respectively. For presence-only algorithms, Maxent and GARP, the decline in the Kappa was 24% and

31%, respectively, and in the AUC was 13% and 11%, respectively. Of the models employed for this study, RF was the most sensitive algorithm to species positional error. The results, furthermore, indicated that, in many cases, the RF, GARP and Maxent performed less well than the GAM, GLM and BRT models.

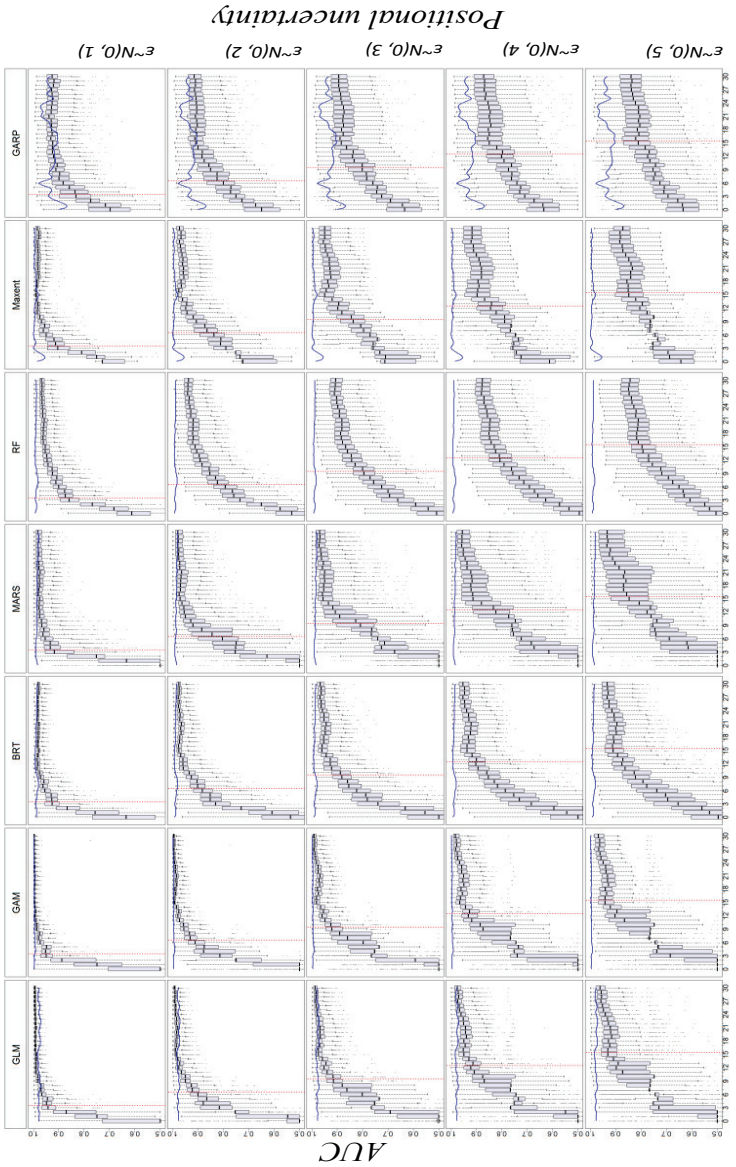



**Figure 2-8:** Variation in the model accuracy (AUC) over the Monte Carlo simulation through different ranges of spatial autocorrelation (x-axis) for all scenarios and modelling techniques; each row corresponds to the level of positional error (positional uncertainty scenario); the different components of the graph are described in Fig. 2-5

The results from the Friedman test with two factors (two categories of spatial autocorrelation range in predictors as the first factor and five scenarios of positional uncertainty as the second factor) indicated that the difference between the spatial autocorrelation range groups for the measures of model performance (Kappa and AUC) for all modelling techniques were significant ($P < 0.001$). The differences between positional uncertainty groups were also significant ($P < 0.001$). The interaction plots based on the Friedman test (Fig. 2-9) show the difference between two categories of spatial autocorrelation range through different levels of positional error as well as the trend of decline in performance as described above.

In order to provide a framework to link the positional uncertainty to model robustness, the results of all scenarios for each SDM were interpolated into a contour plot (Fig. 2-10). This graph shows the interaction of positional uncertainty and spatial autocorrelation range in environmental variables. By considering the level of positional uncertainty (y-axis) and the spatial autocorrelation range in the environmental variables (x-axis), an accuracy measure is provided which can be compared with the accuracy measure for the same range of spatial autocorrelation but with no positional error in species occurrences. The difference between these two values can be interpreted as an expected decline in performance, and shows the potential impact of positional uncertainty, given a level of spatial autocorrelation in the explanatory variables.

**Table 2-1:** Mean and standard deviations [mean | SD] of Kappa (a) and AUC (b) for the model outputs resulting from the Monte Carlo simulation for different positional error scenarios (PE1–PE5) and the Kappa for the model using unperturbed data. C1 and C2 specify the categories of Monte Carlo simulations where the spatial autocorrelation range was less than three times the standard deviation in positional error (C1) and those of which the spatial autocorrelation range was more than three times the standard deviation in positional error (C2)

(a)

| Algorithm | Unperturbed Data | The positional error scenarios | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PE1 | | PE2 | | PE3 | | PE4 | | PE5 | |
| | | C1 | C2 | C1 | C2 | C1 | C2 | C1 | C2 | C1 | G2 |
| GLM | 0.941 | 0.48 \| 0.27 | 0.90 \| 0.07 | 0.44 \| 0.24 | 0.86 \| 0.09 | 0.44 \| 0.23 | 0.82 \| 0.09 | 0.44 \| 0.23 | 0.78 \| 0.10 | 0.45 \| 0.23 | 0.74 \| 0.11 |
| GAM | 0.956 | 0.47 \| 0.29 | 0.92 \| 0.07 | 0.43 \| 0.27 | 0.88 \| 0.08 | 0.42 \| 0.26 | 0.84 \| 0.09 | 0.43 \| 0.26 | 0.81 \| 0.09 | 0.44 \| 0.26 | 0.76 \| 0.11 |
| BRT | 0.901 | 0.53 \| 0.22 | 0.85 \| 0.06 | 0.47 \| 0.21 | 0.81 \| 0.08 | 0.45 \| 0.21 | 0.78 \| 0.08 | 0.46 \| 0.22 | 0.75 \| 0.08 | 0.47 \| 0.22 | 0.71 \| 0.09 |
| MARS | 0.903 | 0.36 \| 0.29 | 0.84 \| 0.08 | 0.32 \| 0.25 | 0.80 \| 0.12 | 0.32 \| 0.24 | 0.76 \| 0.13 | 0.33 \| 0.24 | 0.72 \| 0.14 | 0.33 \| 0.24 | 0.66 \| 0.16 |
| RF | 0.913 | 0.39 \| 0.17 | 0.77 \| 0.09 | 0.34 \| 0.16 | 0.68 \| 0.10 | 0.32 \| 0.16 | 0.62 \| 0.10 | 0.32 \| 0.16 | 0.58 \| 0.10 | 0.32 \| 0.16 | 0.53 \| 0.10 |
| MAXENT | 0.711 | 0.40 \| 0.24 | 0.71 \| 0.07 | 0.36 \| 0.24 | 0.68 \| 0.10 | 0.35 \| 0.24 | 0.65 \| 0.11 | 0.34 \| 0.23 | 0.59 \| 0.13 | 0.34 \| 0.23 | 0.54 \| 0.13 |
| GARP | 0.774 | 0.39 \| 0.2 | 0.68 \| 0.10 | 0.36 \| 0.21 | 0.63 \| 0.11 | 0.35 \| 0.21 | 0.59 \| 0.11 | 0.35 \| 0.21 | 0.56 \| 0.11 | 0.34 \| 0.21 | 0.53 \| 0.11 |

(b)

| Algorithm | Unperturbed Data | The positional error scenarios | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PE1 | | PE2 | | PE3 | | PE4 | | PE5 | |
| | | C1 | C2 | C1 | C2 | C1 | C2 | C1 | C2 | C1 | G2 |
| GLM | 0.98 | 0.78 \| 0.17 | 0.98 \| 0.02 | 0.75 \| 0.16 | 0.97 \| 0.03 | 0.74 \| 0.16 | 0.96 \| 0.03 | 0.75 \| 0.16 | 0.95 \| 0.04 | 0.75 \| 0.16 | 0.93 \| 0.06 |
| GAM | 0.99 | 0.78 \| 0.18 | 0.99 \| 0.02 | 0.75 \| 0.17 | 0.98 \| 0.03 | 0.74 \| 0.16 | 0.97 \| 0.03 | 0.74 \| 0.16 | 0.96 \| 0.04 | 0.75 \| 0.17 | 0.94 \| 0.05 |
| BRT | 0.99 | 0.79 \| 0.16 | 0.97 \| 0.02 | 0.75 \| 0.17 | 0.96 \| 0.03 | 0.73 \| 0.17 | 0.94 \| 0.04 | 0.74 \| 0.17 | 0.93 \| 0.04 | 0.74 \| 0.17 | 0.91 \| 0.05 |
| MARS | 0.98 | 0.71 \| 0.17 | 0.97 \| 0.03 | 0.68 \| 0.15 | 0.95 \| 0.06 | 0.67 \| 0.14 | 0.93 \| 0.07 | 0.68 \| 0.14 | 0.91 \| 0.08 | 0.68 \| 0.14 | 0.88 \| 0.09 |
| RF | 0.99 | 0.73 \| 0.12 | 0.94 \| 0.04 | 0.68 \| 0.12 | 0.90 \| 0.05 | 0.67 \| 0.12 | 0.87 \| 0.05 | 0.67 \| 0.13 | 0.85 \| 0.05 | 0.67 \| 0.12 | 0.82 \| 0.06 |
| MAXENT | 0.99 | 0.79 \| 0.10 | 0.97 \| 0.03 | 0.77 \| 0.10 | 0.95 \| 0.04 | 0.76 \| 0.10 | 0.92 \| 0.05 | 0.76 \| 0.10 | 0.89 \| 0.06 | 0.76 \| 0.09 | 0.86 \| 0.06 |
| GARP | 0.93 | 0.76 \| 0.11 | 0.94 \| 0.04 | 0.74 \| 0.10 | 0.90 \| 0.05 | 0.74 \| 0.10 | 0.87 \| 0.05 | 0.74 \| 0.10 | 0.85 \| 0.05 | 0.73 \| 0.09 | 0.82 \| 0.06 |

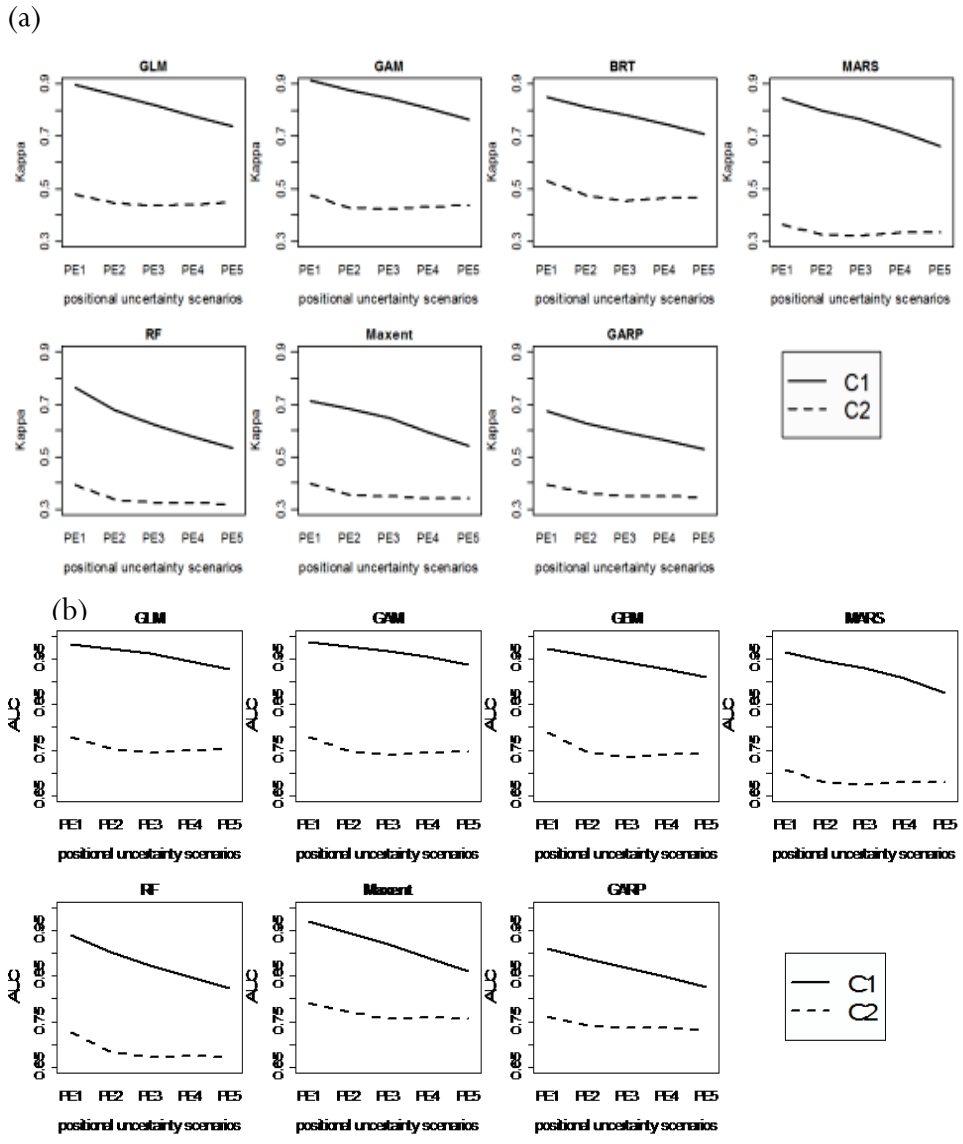



**Figure 2-9**: Interaction plots based on the Friedman test – difference of Kappa mean (a) and AUC mean (b) between two categories of spatial autocorrelation range (C1 and C2) through different levels of positional error (PE1–PE5). C1 and C2 specify the categories of MC simulations where the spatial autocorrelation range was less than three times the standard deviation in positional error (C1) and those of which the spatial autocorrelation range was more than three times the standard deviation in positional (C2)

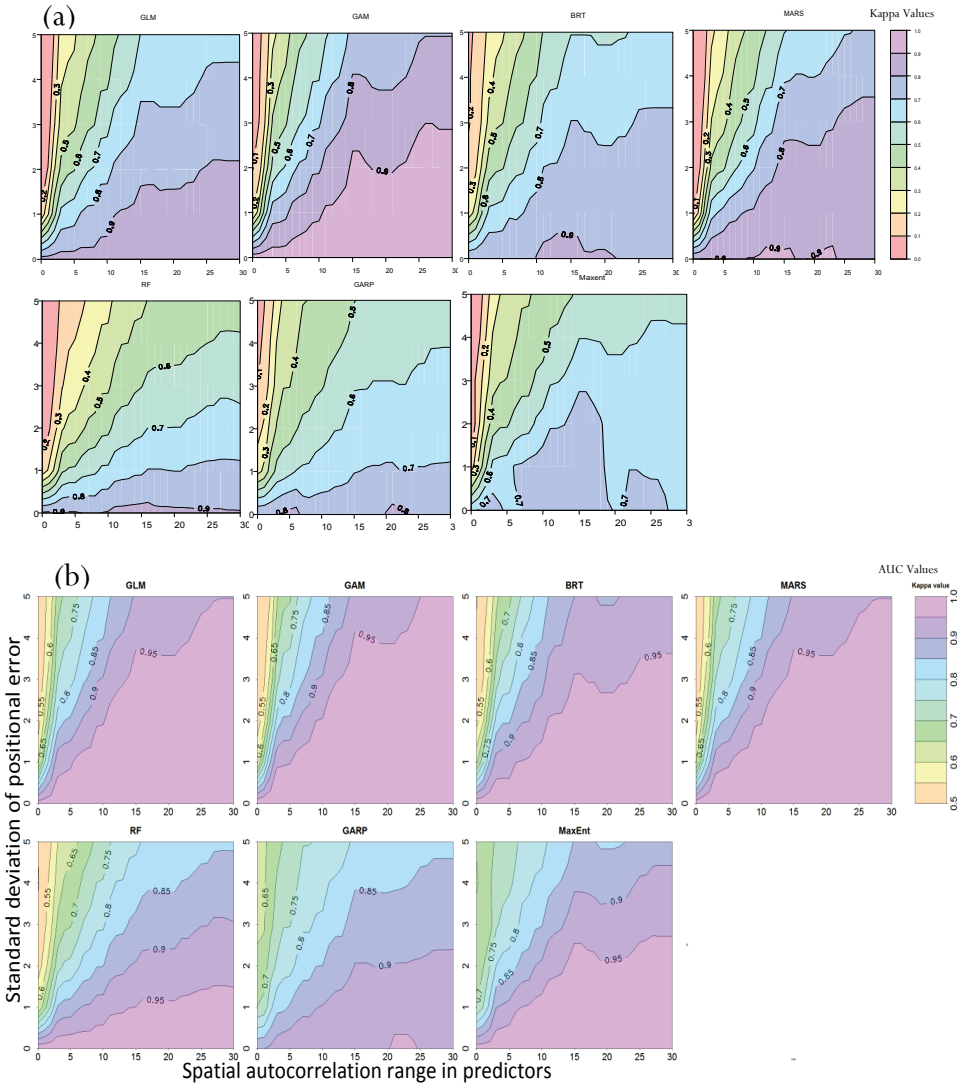



**Figure 2-10**: Interaction of species occurrence positional error and spatial autocorrelation range in predictors; x-axis represents the range of spatial autocorrelation in predictors, and the y-axis represents the standard deviation of positional error in species occurrences; considering the level of positional error and the spatial autocorrelation range of the predictors, the measure of accuracy (Kappa and AUC in (a) and (b), respectively) could be compared to the measure in the same range of spatial autocorrelation but without positional error (y = 0) in species occurrences

## 2-4. Discussion

This study linked the SDM robustness to positional uncertainty in species occurrences and spatial autocorrelation range in predictors. The results show that the impact of positional uncertainty can be assessed by examining spatial autocorrelation in the predictors. Graham et al. (2008) and Osborn & Leitão (2009) found that positional error had a small effect on predictive performance of models as judged by AUC and, consequently, useful predictions can be made even when species data are subjected to some positional error. Based on the results of the current study, this is not always true, and this depends on the range of spatial autocorrelation in the predictors relative to the level of positional error.

It was expected that larger spatial autocorrelation ranges in the predictors would diminish the impact of positional uncertainty on prediction accuracy. Comparing the Kappa and AUC values in the second category of spatial autocorrelation range over the positional uncertainty scenarios showed that, in all scenarios, positional error led to a drop in performance of the models, but most of them would still be regarded as acceptable and useful models (Manel et al. 2001). The drop was, however, greater for the extreme positional error. Although spatial autocorrelation range can reduce the effect of positional uncertainty, it cannot completely compensate for it.

For all scenarios the interaction between spatial autocorrelation range and positional uncertainty showed a consistent trend. The prediction accuracy of the models using predictors with a spatial autocorrelation range greater than the standard deviation of positional error, were high because the perturbed points should generate similar correlations between environmental predictors and presence/absence compared to when they would not be perturbed. It was shown that the performance of the SDMs became stable when the range of the spatial autocorrelation of the predictors is greater than three times the standard deviation of the spatial error (second category), because more than 99% of the points remained

relatively unaffected in their correlation with the environmental predictors by the perturbation. The graphs demonstrate this trend and the results of the Friedman test support this argument.

The results indicated there is a difference in the performance and behaviour of the models, especially between the two groups of presence/absence and presence-only algorithms. The models using the presence/absence data used more information than the presence-only models and, therefore, this information allowed the models to be better calibrated (Elith and Graham 2009). This may explain why presence/absence models performed better as judged by the Kappa and AUC. The variation in the performance measures decreased when the range of spatial autocorrelation in predictors increased or when the level of positional uncertainty decreased. This is not the same for all models. For instance, comparing the standard deviation of the Kappa measures in both categories for the RF (Table 2) showed that the variation is lower only in the first error scenario and, for the other scenarios, remained approximately the same. This variation in accuracy for the RF model, when judged by AUC (see Table 2-1), was changed slightly from 0.04 to 0.06 for the second category through the scenarios. The behaviour of Maxent showed that it might be sensitive to spatial autocorrelation in predictors as the variation is not consistent through different levels of spatial autocorrelation. These results showed how the models differed under the simulated conditions, and encourage further study using other conditions, such as different patterns of spatial autocorrelation in predictors or cross-correlated autocorrelated predictors.

The two most commonly used methods were selected for evaluating the SDMs (*i.e.*, AUC and Kappa) (Elith and Leathwick 2009). AUC is a threshold-independent measure (Fielding and Bell 1997) which has been used in most recent papers as evaluation method, because the major drawback of the Kappa statistic is related to the selection of a threshold, which is affected by prevalence and bias in data. Biased training, because of either low or high prevalence, affects the optimal value for the cut-off threshold (Jiménez-Valverde et al. 2009 ; Manel et al. 2001). The selection

of the threshold, however, was not an issue in this study since a constant threshold of 0.5 was applied to generate the artificial datasets and the prevalence was controlled at 0.5 to avoid bias. For all the models, except Maxent, the Kappa and AUC statistics were consistent. Based on the AUC (Table 2-1) Maxent provided high accuracy (the AUC measures ranged between 0.86-0.97) in the second category of results, whereas it gave only moderate accuracy based on Kappa (the Kappa measures ranged between 0.54 and 0.71).

Different levels of unknown complexity existing in real data make completely accurate inference in species distribution modelling impossible. Furthermore, statistical methods employed in SDMs are also difficult to evaluate because the 'truth' is unknown (Austin et al. 2006). As a result, simulated data have increasingly been used as a tool for developing and evaluating SDMs (Austin et al. 2006 ; Beale et al. 2010 ; Elith and Graham 2009 ; Jiménez-Valverde et al. 2009 ; Meynard and Quinn 2007 ; Santika and Hutchinson 2009). In this research, the simulated datasets for both species and predictors provided full control over the training data as well as providing reference data for accuracy assessment. Clearly they do not cover all possible complications that are likely to be found in real data. For example, the degree of spatial autocorrelation might vary locally over a real study area, whereas a homogenous area was simulated in this study. In a heterogeneous area the global spatial autocorrelation measures may not properly model its effect on positional uncertainty. Under this circumstance, measuring local spatial autocorrelation in the environmental predictors might be a promising solution to find the locations with high degree of influence. Also, the spatial autocorrelation range might be different for the different predictor variables. In this study, the range was the same for both the simulated predictors. Further study is required to analyse the effect of positional uncertainty when there is variation in spatial autocorrelation range between the contributing environmental variables. It is expected that the variables with higher contribution to the SDM will have a bigger effect on reducing the impact of positional uncertainty.

Under such circumstances, it is advisable that the minimum spatial autocorrelation range amongst all predictors, or the predictors with the highest contributions to the model, be considered when using the contour plots (Fig. 2-10).

Spatial autocorrelation has been widely cited as a problem for regression based techniques (for a review see Dorman et al., 2007). In particular, this arises when care has not been taken to address the assumption that the residuals should be independent and identically distributed. In this study, a random sampling design was employed. Furthermore, analysis of the Moran's I for the model residuals, showed that the residuals were not spatially correlated. Furthermore, it should be noted that there is no conflict between the potential problems of autocorrelation and the arguments of this paper, which regard it as an opportunity.

Finally, the range of spatial autocorrelation in the environmental predictors can be linked to positional uncertainty since both can be expressed in distance units. Of key importance here is the interplay between the level of positional error in the response, the spatial resolution of the predictor variable (*e.g.*, pixel size in a raster grid) and the range of spatial autocorrelation of the predictor. For a given pixel size, an increase in the positional error will increase the probability that the response will be fall within an incorrect grid cell and, hence be associated with incorrect predictor attribute values (Hamm et al. 2003). The results presented in this paper showed that the extent to which this reduces prediction accuracy depends primarily on the range of spatial autocorrelation of the predictors and, secondarily, on the SDM employed. Importantly the autocorrelation was modelled at the same resolution as the predictor variable. It should be noted that autocorrelation range is likely to change with spatial resolution (Atkinson 1993). Hence a user who wishes to apply the approach employed in this paper should take care that the autocorrelation is assessed at the resolution of their predictor variable and not at a finer resolution (Graham et al. 2008).

## *2-5. Conclusions*

This study has explored how different degrees of positional uncertainty in species occurrence data influence the prediction accuracy of SDMs for varying levels in the spatial autocorrelation range of predictors. Spatial autocorrelation in predictors reduced the impact of positional uncertainty on prediction accuracy of the SDMs when the range of spatial autocorrelation was greater than three times of the standard deviation of positional error. In such circumstances, although positional error led to a decline in predictive performance they yielded the maximum achievable performance, given the level of uncertainty in the presence.

It is argued that examining the spatial autocorrelation in predictors to find the effective autocorrelation range can give insight into whether predictions are likely to be affected by the uncertainty in the sample locations. This is especially important when the output of these models is to be used in conservation and planning as it provides an evidence to use them with a clear level of confidence.

# *Chapter 3*

# Positional Uncertainty and (Local) Spatial Autocorrelation

*This chapter is based on:*

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K. & Toxopeus, A. G. 2014. Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37, 191-203.

# 3. Positional Uncertainty and (local) Spatial Autocorrelation

## *3-1. Introduction*

Species distribution models (SDMs) based on presence-absence or presence-only data have been used widely in biogeography to characterize the ecological niche of species and to predict the geographical distribution of their habitat (Araújo and New 2007 ; Elith et al. 2006 ; Franklin 2010 ; Skidmore et al. 1996). This approach has been employed for numerous applications, including conservation planning, wildlife management as well as predicting the impact of future scenarios such as climate change or habitat fragmentation, on species occurrence and biodiversity (Franklin 2010 ; Peterson et al. 2011).

Species data held in museum and herbaria, survey data and opportunistically observed data, provide a vast information resource (Chapman 2005). It is estimated that there are more than 2.5 billion specimen collections worldwide in museums, herbaria and other institutions (Duckworth et al. 1993). Increasingly these data are made available through Internet portals.

A key challenge in using these data is the uncertainty about where an observation is located. The error in position is caused by a variety of factors, including inaccuracy in location (for example due to incorrect map reading or a GPS set to the incorrect datum) and georeferencing error (Graham et al. 2004b ; Wieczorek et al. 2004). In particular, the majority of species data that were collected before the popularization of GPS technology, were recorded as textual descriptions, often based on names and places that can change over time (Wieczorek et al. 2004). When these records were digitized, geographic coordinates were often inferred and may be substantially (several kilometres) incorrect in their position (Feeley and Silman 2010). This problem, so called positional uncertainty, becomes important when the data are used to develop a species distribution model

(SDM). Coordinates are used to extract the co-located environmental variables and thus, positional error will transfer to inaccurate characterizations of the species-environment relationship (Feeley and Silman 2010).

Some techniques have been developed to estimate and document the positional uncertainty in occurrence data and remove highly uncertain observations prior to analysis (Guo et al. 2008), however, this reduces the sample size, which in turn is one of the factors that reduces model accuracy (Graham et al. 2008 ; Hernandez et al. 2006). Having error in data does not automatically have to be a reason to discard the data (Chapman 2005). In this case, it is important to know whether and where the error is problematic. For example, Graham et al. (2008) compared different models to see if they were affected by an introduced random error to the location of occurrence data. Although they concluded that the SDMs are, in general, robust to positional errors, Osborne and Leitão (2009) and Naimi et al. (2011) argued that this is not always true, and it is related to the level of spatial autocorrelation in the predictor variables (see chapter 2). Spatial autocorrelation is a property of most ecological variables (Legendre 1993) and represents the relationship between nearby spatial units (Getis 2010). Positional uncertainty matters less in developing species-environment relationships if nearby locations have similar attribute values to the original location. In chapter 2, we conducted a comprehensive set of analyses to assess the interaction between spatial autocorrelation in predictors and positional uncertainty in species occurrence. Using artificial data we analysed the influence of five positional uncertainty scenarios on the prediction accuracy of seven frequently applied SDMs. We concluded that the magnitude of the spatial autocorrelation range relative to the magnitude of the positional uncertainty can give insight into whether SDMs are affected by the uncertainty in the sample locations.

Most indices that measure spatial autocorrelation, such as Moran's *I*, Geary's *c* (Cliff and Ord 1981) and the variogram (Cressie 1993) are global

in nature and assume stationarity of the spatial process. This assumption is often not met (Anselin 1995), and the degree of spatial autocorrelation can vary across a study area (Hamm et al. 2012). We propose that, under such circumstances, adopting a stationary global spatial autocorrelation measure is inappropriate for modelling the effect of positional uncertainty. A possible solution would be to adopt a non-stationary model (Hamm et al. 2012) that can address local heterogeneity in the data. Another possibility is to use local indicators of spatial association (LISA) (Anselin 1995 ; Getis and Ord 1996). LISAs give a measure of correlation between a single location and its neighbours up to a specified distance (Getis and Ord 1996). It has been shown that spatial autocorrelation in predictors can be linked to SDM robustness (chapter 2). For this purpose, using LISAs may be more insightful because it may lead to identification of the specific occurrence records that cause the largest drop in SDM performance.

We designed an experiment to assess the propagation of positional uncertainty in occurrence locations based on the local spatial association among the predictors. We used a Geary type statistic, called the K statistic (Getis and Ord 1996), to quantify local spatial association for each predictor at the location of species occurrences. We tested the hypothesis that the SDMs' predictions are more affected by positional uncertainty originating from locations that have lower local spatial association in their predictors. We performed this experiment in Spain and the Netherlands using artificial datasets derived from well-known SDMs. Further, we developed a tool in the R environment (R Development Core Team 2011) to explore whether observations with positional uncertainty are likely to be creating error in the output from SDMs.

## *3-2. Materials and methods*

### 3.2.1. Data sources

Spain and the Netherlands were selected for this study. The overall landscape structure is rather heterogeneous and more influenced by anthropogenic activities in the Netherlands compared to Spain. Therefore

we expected different levels of local spatial association between these two areas. This gives the possibility to test if the lower local spatial association impacts the predictions of the SDMs. Three species were selected randomly from all available species data in three classes of vertebrates in Spain (one species for each class): *Microtus cabrerae* [hearafter, *es1*] (De Cabrera 2007), *Dryocopus martius* [hearafter, *es2*] (Negro 2007), and *Coronella girondica* [hearafter, *es3*] (Meridional 2007) from mammals, birds and reptiles, respectively based on the Spanish vertebrate presence-absence atlas data which includes 5220 grid cells with a spatial resolution of 10 x 10 km. Two mammals' species, *Microtus oeconomus* [hearafter, *nl1*] and *Neomys fodiens* [hearafter, *nl2*], were selected in the Netherlands from the field data surveyed between 2000 and 2009 by the Dutch mammal society. The sample sizes for these two species were 1601 and 991 presence-only records, respectively. We used 20 environmental variables including 4 topographic variables [elevation, slope, southness, and topographic wetness index (Beven and Kirkby 1979)], and 16 seasonal means of satellite image products including the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI) and day and night time land surface temperature (LST). Satellite products were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite image archive (Nasa Land Processes Distributed Active Archive Center 2011). The images are from period 2000-2009. This matches with the collection period of the occurrence data in the Netherlands. It was also assumed that the computed seasonal means based on these 10 years images are representative of a longer period. This assumption makes them appropriate to be used with the occurrence data in Spain which were collected over a longer period. It has been shown that remotely sensed data can contribute significantly to defining habitat characteristics even within an area with similar climatic conditions (Buermann et al. 2008).

It is likely that some of the predictor variables are correlated. Strong correlation between two or more predictor variables is collinearity (Graham 2003) and can cause instability in parameter estimation in

regression-type models (Dormann et al. 2012). We used the variance inflation factor ($VIF$) to detect collinearity (Marquardt 1970). $VIF$ is based on the square of the multiple correlation coefficient (${R^2}_P$) resulting from regressing the predictor variable $x_P$ against all other predictor variables (*Eq.* 3-1). A $VIF$ greater than 10 is a signal that the model has a collinearity problem (Chatterjee and Hadi 2006). We excluded the variables with large $VIF$ valus (greater than 10) one by one using a stepwise procedure. We repeated this procedure until all strongly correlated variables (*i.e.*, with $VIF > 10$) were excluded.

$$VIF_p = \frac{1}{1 - R^2}, p = 1, \dots, n \qquad (3\text{-}1)$$

where $VIF_p$ is the $VIF$ associated with the $p^{th}$ predictor, $n$ is the number of predictor variables. If $x_P$ has a strong linear relationship with at least one other variable, $R_p^2$ would be close to 1, and $VIF_p$ would be large.

### 3-2-2. Generating a realistic artificial dataset

Predictions made from SDMs are difficult to evaluate because the 'truth' is unknown (Austin et al. 2006). In recent years, simulated data, also known as artificial data or virtual species, have been used as a tool to conduct controlled experiments in SDM studies (Austin et al. 2006 ; Hirzel et al. 2001 ; Jiménez-Valverde et al. 2009 ; Naimi et al. 2014 ; Naimi et al. 2011). There is, however, a risk that virtual species do not correctly simulate reality (Hirzel et al. 2001). To reduce this risk we used real species occurrence and landscape data to generate a distribution for each of the five animal species (Fig. 3-1). These distributions were then treated as the "true" distribution of these species.

For each species in Spain, a sample of species occurrence points (presence-absence) was drawn randomly from the atlas map, based solely on the criterion of having not more than one point in each 10 x 10 km atlas grid cell. We chose a sample size of 10% of the total number of atlas grid cells

(522 in this case). For the two species in the Netherlands, only presence records were available. We generated pseudo-absence data (with the same size as the presence) using random sampling, weighted by the environmental distance (Engler et al. 2004 ; Zaniewski et al. 2002), i.e. absences are more likely in environmental conditions dissimilar to environmental conditions at the presence locations (Lobo et al. 2010). The environmental distances between locations were calculated using the Mahalanobis Distance (Farber and Kadmon 2003).

We used an ensemble approach to develop SDMs for each species and predict its habitat over the study area. The idea behind the ensemble modelling approach is that a combined multiple-model prediction is more accurate than at least half of the original models (Araújo and New 2007). The higher accuracy of ensemble modelling to predict habitat suitability has been shown by several studies (Le Lay et al. 2010 ; Marmion et al. 2009a). We used five SDM techniques that use presence-absence data: generalized linear models (GLM; McCullagh and Nelder, 1989), generalized additive models (GAM; Hastie and Tibshirani, 1990), boosted regression trees (BRT; Friedman, 2001), random forests (RF; Breiman, 2001), and support vector machine (SVM; Vapnik, 1995). A 5-fold cross-validation for each model was applied. There are different approaches to combine an ensemble of model predictions (for a review, see Araújo and New, 2006). We used committee averaging (a simple unweighted average of the predictions) to generate a single prediction from the outputs of the SDMs. The predicted distribution probability was then used as the reference suitability.

For each species, we selected a final set of environmental variables that showed a significant contribution (defined by an importance greater than 0.05) for at least one model in the ensemble. The final set was used to simulate the habitat suitability using the ensemble modelling approach and in the rest of the study as predictor variables for the species (Table 3-2). To estimate the importance of each variable we used a randomization procedure that is implemented in the BIOMOD R-package (Thuiller et al. 2009). It is a model-independent procedure that uses Pearson's correlation

coefficient between the predicted values and predictions where the variable under investigation was randomly permutated. If the contribution of a variable to the model is high, it is expected that the prediction is more affected by a permutation and therefore the correlation is lower. Therefore, '1 – correlation' can be considered as a measure of variable importance. We repeated this procedure 30 times for each variable and each SDM.

For the experiment, we drew two presence-absence realizations for each species, one to train the SDMs and one to validate the results. To simulate a realistic sampling procedure, we designed a sampling scheme with a random uniform distribution over space. We then used the ensemble predicted suitability value in each grid cell as the success rate for each sample point to contain the species (Elith and Graham 2009). For example a cell with a suitability of 0.7 has a 70% probability of being occupied by the species. To see if our approach is sensitive to sample size, we evaluated three sample sizes: 'equal', '0.2%', and '1%'. The sample size for 'equal' had the same size (= 100) for both study areas. Since the study areas are not equal in size, the two other sample sizes were considered in order to keep the sample density the same in both areas. The sample size for the '0.2%' and '1%' scenarios were equal to 0.2% and 1% of the total grid cells in the study area, respectively.

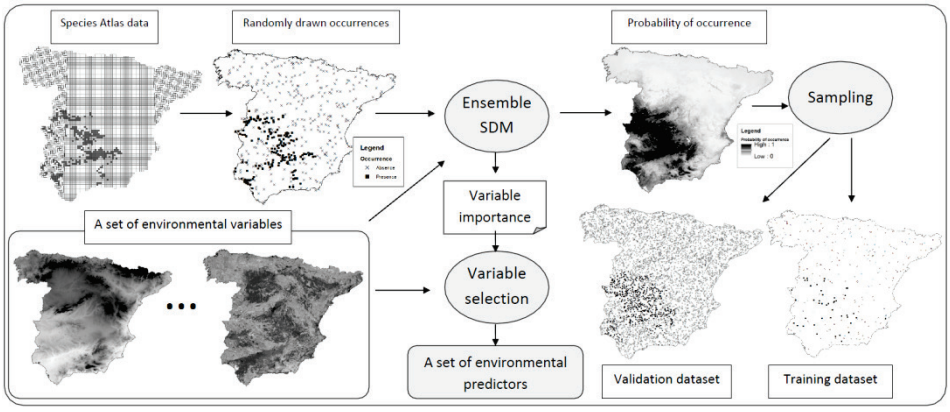



**Figure 3-1:** Flow diagram showing the procedure of generating semi-artificial datasets

### 3-2-3. Species distribution modelling

Six commonly used SDMs that require presence-absence or presence-only records of species occurrences were selected (Table 3-1). For this purpose, we made an implicit assumption that the species are in equilibrium with the environmental variables (i.e. there are no dispersal limitations or biotic interactions). The predicted distributions for both presence-only and presence-absence SDMs were evaluated for their accuracy using a separate presence-absence validation dataset. We used the area under curve (AUC) of a receiver operating characteristic (ROC) to measure the predictive performance of the SDMs. An ROC curve plots true positive rates (TPR) on the y-axis against false positive rates (FPR) on the x-axis for all thresholds (Fielding and Bell 1997). An AUC value of 0.5 implies random predictive discrimination and a value of less than 0.5 indicate discrimination worse than chance. For a SDM to be a good discriminator, this measure should be close to 1. AUC is a commonly used statistic for evaluating the accuracy of SDMs, although its usefulness has been questioned by some authors (see for example, Lobo et al., 2008).

### 3-2-4. Positional uncertainty assessment

We used Monte Carlo simulation to assess the effect of positional uncertainty in occurrence data on the SDMs' performance. A probabilistic approach was used to introduce a positional error ($\varepsilon$) with no directional bias in species occurrence. Taking $\varepsilon \sim N(0, 5000)$ gives a normally distributed unbiased error with a standard deviation of 5000 m. Concurring with Graham et al. (2008), we assumed that this is representative of the error associated with museum data. This was added to the easting and northing of each location (Hamm et al. 2004):

$$x_i = Easting + \varepsilon_i$$
$$y_i = Northing + \varepsilon_i \quad (3\text{-}2)$$

where $i$ refers to each individual species occurrence, $x$ and $y$ refer to the perturbed location by adding the error to the true easting and northing. Different realizations with introduced positional error were generated for each species and used to explore the effect of positional uncertainty. These were termed the 'perturbed' datasets. In total 1000 realizations of perturbed datasets were generated for each species. We used these realizations to train the models. The accuracy of each model, trained by each of these 1000 datasets, was assessed against the true (unperturbed) species occurrence data (Heuvelink 1999). This Monte Carlo simulation allowed us to assess the impact of positional uncertainty on model performance (Fig. 3-2). The standard deviation of the model performances were also used to assess the stability of the resulting models.

**Table 3-1:** The details and settings of model implementation

| Model | Data | Specifics and settings | Reference |
|---|---|---|---|
| GLM | PA | Uses parametric functions to link the response variable to a linear, quadratic, and/or cubic combination of explanatory variables. We used a GLM linear with logit link function, implemented in R stats package v 2.13.1. For simplicity we refer to this as 'GLM'. | (Austin 2002 ; Mccullagh and Nelder 1989) |
| GAM | PA | Uses nonparametric and data-defined, smoother to fit, nonlinear functions. Here, we fitted GAM with a cubic spline smoother using the R gam package v 1.04.1. | (Austin 2002 ; Hastie 2011 ; Hastie and Tibshirani 1990) |
| BRT | PA | Fits complex nonlinear relationships by combining two algorithms of regression trees (relate a response to their predictors by recursive binary splits) and boosting (an additive method to combine many single models to improve the performance). We used custom code published by Elith et al. (2008). This code used the R gbm package v 2.0-4. (Settings: the optimal number of trees; learning rate = 0.001; tree complexity = 3). | (Elith et al. 2008 ; Friedman 2001) |
| RF | PA | Selects many bootstrap samples from the data and generates and fits a large number of regression trees to each of these subsamples. Each tree is used to predict the out-of-bag observations (*i.e.*, those that were not selected as bootstrap samples). The classification given by considering each tree as a 'vote', and the predicted class of an observation is determined by the majority vote among all trees. We used the R randomForest package v. 4.6-2 (settings: number of trees = 3000, the default settings for the rest of parameters). | (Breiman 2001, Cutler et al. 2007, Liaw and Wiener 2002) |
| SVM | PA | SVM apply a simple linear model of data into a high-dimensional (hyperplane) feature space defined by a kernel function. We used the R kernlab package v 0.9-14 to fit the SVM regression using ANOVA RBF kernel which typically performs well in regression problems. | (Karatzoglou et al. 2006, Vapnik 1995) |
| Maxent | PO | Uses a maximum entropy density estimation algorithm to approximate the true distribution of species as a probability distribution which respects a set of constraints where the mean of each environmental variable is required to be close to the empirical average over the presence sites. The default settings were applied. | (Phillips et al. 2006, Phillips and Dudik 2008) |

PA, presence-absence; PO, presence-only.

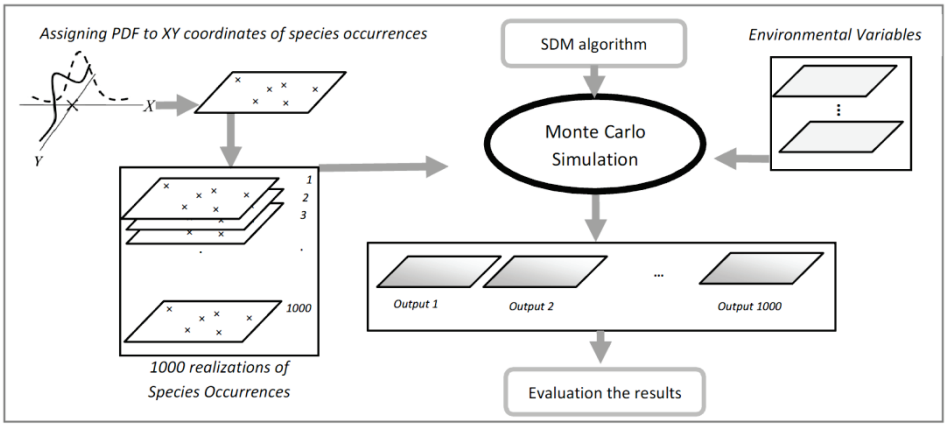



**Figure 3-2:** Flowchart showing the procedure of positional uncertainty assessment. PDF, probability density function; SDM, species distribution model

### 3-2-5. Local spatial association

Spatial autocorrelation is concerned with the degree to which variable $y_s$, measured at location $s$, is similar to $y_{s+h}$, measured at a specific geographic distance ($h$) from $s$ (Goodchild 1986). Measures of spatial autocorrelation are either global or local, as reviewed by Getis (2010). Global measures provide a statistic for the entire field under the assumption that the mean, and covariance do not vary over the study area. Local indicator of spatial association (LISA) statistics (Anselin 1995), including local Moran's *I* and local Geary's *c*, decompose the single global measure into the contribution of each individual grid cell thus revealing which grid cells have the most impact on the global measure. Getis and Ord (1992) developed local statistics, including $G_i$ and $G_i^*$, which indicate local clustering of high and low values to detect pockets of spatial association that may not be evident when using global statistics. These local statistics are useful for identifying hotspots, but they are influenced by the presence of global spatial autocorrelation and must be interpreted according to the degree of global spatial autocorrelation in the data (Getis and Ord 1992 ; Ord and Getis 1995 ; Ord and Getis 2001).

Getis and Ord (1996) proposed the K statistic, which measures deviations from the observed values at a reference site i, that is $(x_i - x_j)$ rather than

deviation from the global mean ($x_i - \bar{x}$) as is used in local Geary (See Anselin, 1995). This allows identification of the local association without assuming global stationarity. Hence, this statistic is most appropriate for this study as it quantifies the local dissimilarity of environmental variables for each location in an absolute rather than a relative way. We used the K statistic in standardized form (Getis and Ord 1996; *Eq.* 3-3) for each environmental variable at each grid cell. Values of the K statistic less than 0 imply that local similarity is greater than expected (*i.e.*, high spatial association), and values greater than 0 imply that local similarity is less than expected (*i.e.*, low spatial association).

The K statistic was calculated for each environmental variable at the location of species occurrences within a local distance of 15 km. This is equal to three times the standard deviation of the error introduced to species location (*i.e*, 5 km). Hence, 99.7% of the perturbed points are expected to be within this distance from the original point. The K statistics of the selected environmental variables for each species were aggregated to a single K statistic by weighted averaging using the contribution (importance) of the environmental variables to the SDMs. The weighting was according to the variable importance, as discussed in the section on generating an artificial dataset.

$$z(K_i) = \frac{\langle K_i - \mathrm{E}(K_i) \rangle}{\sqrt{Var(K_i)}} \qquad (3\text{-}3)$$

where $K_i$ is the unstandardized K statistic at grid cell $i$:

$$K_i = \sum_j \omega_{ij} |x_i - x_j| \qquad (3\text{-}4)$$

$\omega_{ij}$ is a binary weight which specifies if the cell $j$ is within a specified distance from cell $i$; and are the attribute values at cell $i$ and $j$, respectively. The mean and variance of the K statistic for all grid cells were calculated using *Eq.* 3-5 and 3-6:

$$E(K_i) = W_i D_i \qquad (3\text{-}5)$$

where:

$$W_i = \sum_j \boldsymbol{\omega_{ij}}$$

$$D_i = \frac{\sum_j |z_i - z_j|}{n-1}$$

$$z_i = x_i - \bar{x} \quad \text{and} \quad z_i = x_i - \bar{x}$$

and $n$ is the number of cells.

$$Var(K_i) = \frac{W_i \times [(n-1) - W_i] \times (t_i^2 - D_i^2)}{n-2} \qquad (3\text{-}6)$$

where: $t_i^2 = \sum_j (z_i - z_j)^2/(n-1)$

### 3-2-6. Data analysis

For each species, four scenarios were applied to explore the effect of positional uncertainty on the performance of the models. In the first scenario (S.all), we introduced the positional error in all occurrence sample locations. This allows full assessment of the impact of positional uncertainty for each species. In the second and third scenarios (S.low and S.high), we introduced positional error to only half of all occurrence locations and the other points maintained their original correct coordinates in all 1000 realizations in the perturbed dataset. We selected these two partitions based on the median of the local spatial association (K statistic) at the occurrence locations. The first partition included the half of the occurrence points with the lower local spatial association in the environmental variables, and the second partition included the half of occurrence points but with the higher local spatial association. In the last scenario (S.rand), the positional error was introduced to a randomly selected sample of half of all locations to provide a control for comparison.

We tested whether the level of local spatial association in predictors at the occurrence locations influences the impact of positional uncertainty on SDM prediction. For this hypothesis, the model performances (AUC) of Monte Carlo simulation runs were compared across different scenarios using a one-way Friedman test (Friedman 1937). We also used a multiple ANOVA to test the interaction effect of sample sizes, SDM and the spatial association scenarios on model accuracy to test the robustness of the results to the data generation process.

### 3-2-7. Analysis and implementation in R

We implemented a package (usdm) in R (R Development Core Team 2013) which includes functions to quantify and visualize the local spatial association, defined as the K statistic, for a set of environmental variables (predictors) at species occurrence locations. This package uses the basic functionality provided in the R raster package (Hijmans and Van Etten 2011) to manipulate environmental variables as raster objects.

We also implemented the SDMs in the R environment v.2.13.1 (R Development Core Team 2013) using different add-on packages (Table 3-1). Maxent was run by using the Maxent software v. 3.3.3, developed by Phillips et al. (2006). In order to run Maxent in the Monte Carlo simulations, together with other techniques within the R environment, we implemented an R function that accessed Maxent in command-mode.

## *3-3. Results*

Absolute correlation values between environmental variables ranged from 0.009 to 1.000 (mean = 0.385, median = 0.358) in Spain and from 0.002 to 1.000 (mean = 0.362, median = 0.355) in the Netherlands. Our procedure to exclude the highly collinear predictor variables (*i.e.*, with $VIF > 10$) led to removing twelve and six predictors from the 20 predictors in Spain and the Netherlands, respectively (Table 3-2).

**Table 3-2:** The selected set of predictors and their variance inflation factor (VIF) and variable importance for each species in Spain (*es*) and the Netherlands (*nl*); the predictors that had ***VIF* > 10** in both areas have been excluded from the Table; gray represents the predictors that have not been selected due to collinearity (see VIF columns) or no variable importance (see the columns for the five case studies); *es1*, *es2*, *es3*, *nl1*, and *nl2* are the abbreviations for the case studies in Spain (*es*) and in the Netherlands (*nl*)

| **Environmental variables** | VIF (*es*) | VIF (*nl*) | *es1* | *es2* | *es3* | *nl1* | *nl2* |
|---|---|---|---|---|---|---|---|
| Seasonal NDVI_1 | | 8.48 | | | | 0.21 | 0.43 |
| Seasonal NDVI_2 | | 8.73 | | | | 0.38 | |
| Seasonal NDVI_3 | | 5.86 | | | | 0.20 | |
| Seasonal NDVI_4 | 6.39 | | 0.23 | 0.32 | 0.17 | | |
| Seasonal EVI_1 | 7.28 | | | | 0.09 | | |
| Seasonal EVI_4 | | 5.62 | | | | 0.14 | |
| Seasonal LST_day_1 | 1.92 | 1.52 | 0.03 | 0.01 | 0.05 | | |
| Seasonal LST_day_2 | | 1.23 | | | | | |
| Seasonal LST_day_3 | | 1.21 | | | | | |
| Seasonal LST_day_4 | | 1.27 | | | | 0.07 | |
| Seasonal LST_night_1 | | 1.82 | | | | | |
| Seasonal LST_night_4 | 4.93 | 5.07 | 0.09 | | 0.05 | | 0.57 |
| Elevation | 5.36 | | 0.65 | | 0.14 | | |
| Slope | 1.83 | | | 0.67 | 0.24 | | |
| Southness | 1.03 | | | | 0.16 | | |
| TWI | 1.78 | | | | 0.10 | | |

The summary of the K statistic values at the locations of the species occurrences for the five case studies (Table 3-3) show that the level of local spatial association is, for both areas, high. The local spatial association in Spain was substantially greater than in the Netherlands. The positive maximum values of the K statistics in *es2*, *es3*, *nl1* and *nl2* show that there are some species occurrence locations where the local spatial association was low. These values show that *es1* (*i.e.*, *Microtus cabrerae* in Spain) and *nl1*

(*i.e.*, *Microtus oeconomus* in the Netherlands) are, respectively, the case studies with the highest and lowest local spatial association in the predictors at the species occurrence locations, compared to all other case studies.

**Table 3-3:** The summary of the K statistics at species occurrence locations for five case studies

| Case studies | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| *es1* | -50.74 | -32.93 | -29.36 | -28.39 | -24.31 | -3.68 |
| *es2* | -42.35 | -20.18 | -15.42 | -15.66 | -11.12 | 11.21 |
| *es3* | -32.59 | -20.85 | -17.14 | -17.22 | -14.03 | 4.52 |
| *nl1* | -17.05 | -9.47 | -6.45 | -5.51 | -1.97 | 15.03 |
| *nl2* | -21.12 | -12.90 | -9.90 | -7.61 | -2.95 | 14.71 |

Comparing the K statistics for different case studies (Table 3-3) and the summary statistics for AUC values from the Monte Carlo simulation for different species (Table 3-4) shows that when the local spatial association in the predictors was low (*i.e.*, the K statistic was high), the model accuracy was more influenced by the positional uncertainty. The mean decline in the AUC values between all scenarios, for the models in *es1*, was 0.5% and for the models in *nl1* was 5.1% in comparison with the AUC for the models using the unperturbed data.

The influence of the positional uncertainty on the model accuracy consistently changed between the five case studies according to the level of local spatial association in the predictors. The results for the two species which had the lowest and the highest local spatial association at the sample locations (*i.e.*, *nl1* and *es1*, respectively) are presented in Figs 3-3 and 3-4; and the results for the other species are provided in Figs from 3-5 to 3-7.

**Table 3-4:** Summary statistics for the performance measures of the SDMs for all species and different scenarios

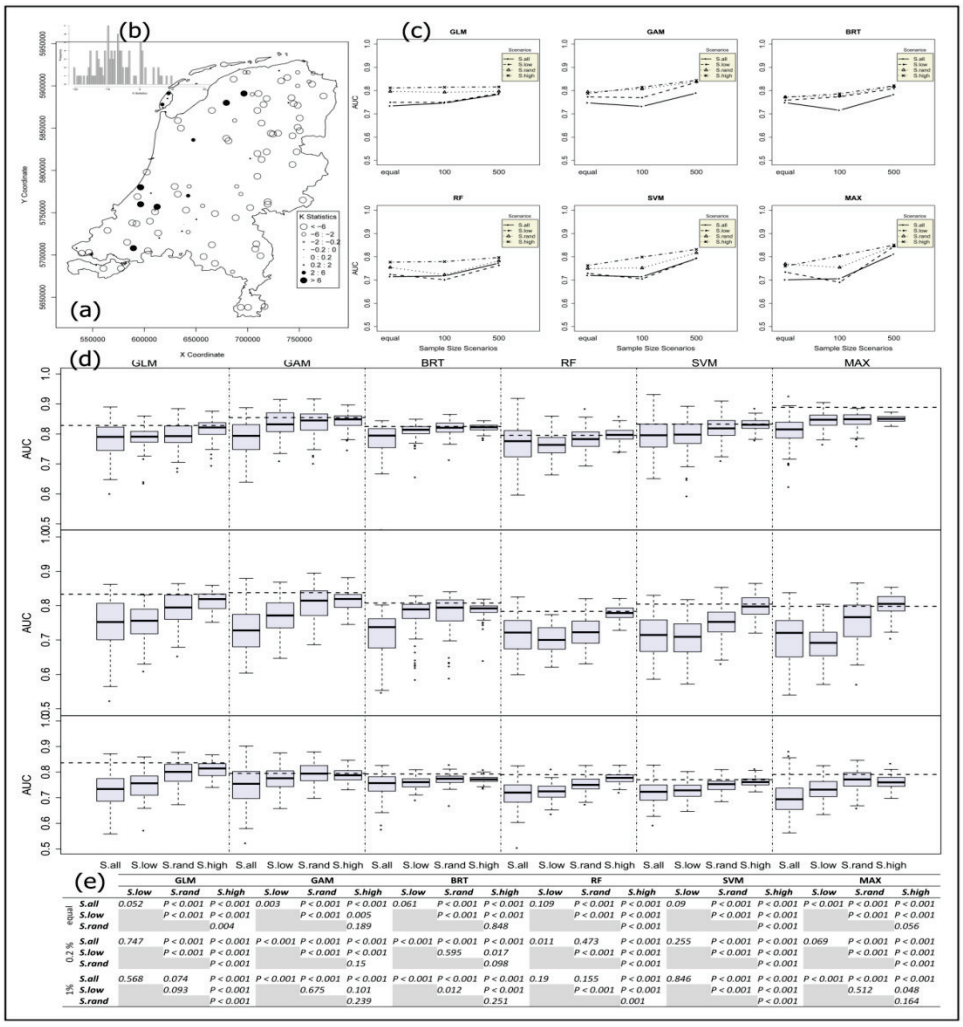
| | | | The sample size scenarios | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Unperturbed data | Equal (size = 100) | | | | 0.2% of the total pixels | | | | 1% of the total pixels | | | |
| Species | Model | | S.all | S.low | S.rand | S.high | S.all | S.low | S.rand | S.high | S.all | S.low | S.rand | S.high |
| *Neomys fodiens (nl1)* | GLM | 0.83 | 0.73\|0.06 | 0.75\|0.05 | 0.8\|0.04 | 0.81\|0.03 | 0.75\|0.07 | 0.75\|0.05 | 0.79\|0.05 | 0.81\|0.03 | 0.78\|0.06 | 0.79\|0.03 | 0.8\|0.04 | 0.82\|0.03 |
| | GAM | 0.86 | 0.75\|0.07 | 0.77\|0.05 | 0.79\|0.04 | 0.79\|0.03 | 0.73\|0.07 | 0.77\|0.05 | 0.81\|0.05 | 0.81\|0.03 | 0.79\|0.06 | 0.84\|0.04 | 0.84\|0.04 | 0.84\|0.03 |
| | BRT | 0.83 | 0.75\|0.05 | 0.76\|0.02 | 0.77\|0.02 | 0.77\|0.01 | 0.72\|0.07 | 0.77\|0.05 | 0.78\|0.05 | 0.79\|0.02 | 0.78\|0.05 | 0.81\|0.03 | 0.82\|0.02 | 0.82\|0.01 |
| | RF | 0.80 | 0.71\|0.05 | 0.72\|0.03 | 0.75\|0.03 | 0.78\|0.02 | 0.72\|0.05 | 0.70\|0.04 | 0.72\|0.04 | 0.78\|0.02 | 0.77\|0.06 | 0.76\|0.04 | 0.78\|0.04 | 0.80\|0.02 |
| | SVM | 0.83 | 0.72\|0.04 | 0.73\|0.03 | 0.75\|0.03 | 0.76\|0.02 | 0.71\|0.06 | 0.70\|0.06 | 0.75\|0.05 | 0.80\|0.03 | 0.79\|0.05 | 0.79\|0.05 | 0.82\|0.04 | 0.83\|0.02 |
| | MAX | 0.89 | 0.70\|0.06 | 0.73\|0.05 | 0.77\|0.04 | 0.76\|0.03 | 0.71\|0.07 | 0.69\|0.05 | 0.75\|0.06 | 0.80\|0.03 | 0.81\|0.05 | 0.84\|0.03 | 0.85\|0.02 | 0.85\|0.01 |
| *Microtus aeconomus (nl2)* | GLM | 0.77 | 0.72\|0.03 | 0.74\|0.02 | 0.75\|0.01 | 0.76\|0.01 | 0.75\|0.02 | 0.76\|0.01 | 0.76\|0.01 | 0.76\|0.01 | 0.76\|0.02 | 0.76\|0.01 | 0.76\|0.01 | 0.77\|0.01 |
| | GAM | 0.79 | 0.72\|0.05 | 0.72\|0.03 | 0.72\|0.02 | 0.74\|0.02 | 0.75\|0.02 | 0.76\|0.02 | 0.75\|0.02 | 0.76\|0.01 | 0.76\|0.02 | 0.76\|0.02 | 0.77\|0.02 | 0.78\|0.01 |
| | BRT | 0.80 | 0.78\|0.03 | 0.78\|0.03 | 0.77\|0.02 | 0.79\|0.01 | 0.78\|0.03 | 0.8\|0.02 | 0.79\|0.02 | 0.79\|0.02 | 0.77\|0.02 | 0.77\|0.02 | 0.78\|0.02 | 0.78\|0.01 |
| | RF | 0.75 | 0.70\|0.06 | 0.69\|0.04 | 0.71\|0.03 | 0.71\|0.03 | 0.72\|0.04 | 0.74\|0.03 | 0.72\|0.03 | 0.74\|0.02 | 0.74\|0.04 | 0.73\|0.03 | 0.73\|0.04 | 0.75\|0.02 |
| | SVM | 0.79 | 0.71\|0.08 | 0.70\|0.05 | 0.70\|0.04 | 0.70\|0.03 | 0.73\|0.05 | 0.75\|0.03 | 0.72\|0.04 | 0.75\|0.02 | 0.74\|0.04 | 0.72\|0.03 | 0.73\|0.04 | 0.75\|0.02 |
| | MAX | 0.75 | 0.71\|0.02 | 0.72\|0.02 | 0.73\|0.02 | 0.73\|0.01 | 0.73\|0.03 | 0.73\|0.02 | 0.73\|0.02 | 0.73\|0.01 | 0.74\|0.02 | 0.74\|0.01 | 0.74\|0.01 | 0.74\|0.01 |
| *Microtus cabrerae (es1)* | GLM | 0.90 | 0.89\|0.01 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.89\|0.00 | 0.90\|0.00 | 0.90\|0.00 |
| | GAM | 0.95 | 0.92\|0.01 | 0.93\|0.01 | 0.93\|0.01 | 0.93\|0.01 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.95\|0.00 | 0.95\|0.00 | 0.95\|0.00 |
| | BRT | 0.90 | 0.87\|0.01 | 0.87\|0.01 | 0.88\|0.01 | 0.88\|0.01 | 0.90\|0.01 | 0.91\|0.00 | 0.91\|0.00 | 0.91\|0.00 | 0.89\|0.00 | 0.90\|0.00 | 0.90\|0.00 | 0.90\|0.00 |
| | RF | 0.95 | 0.88\|0.02 | 0.90\|0.01 | 0.90\|0.01 | 0.90\|0.01 | 0.92\|0.00 | 0.93\|0.00 | 0.93\|0.00 | 0.93\|0.00 | 0.93\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 |
| | SVM | 0.94 | 0.90\|0.02 | 0.90\|0.01 | 0.90\|0.02 | 0.91\|0.01 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 |
| | MAX | 0.94 | 0.92\|0.01 | 0.93\|0.00 | 0.93\|0.01 | 0.93\|0.00 | 0.93\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 | 0.94\|0.00 |
| *Dryocopus martius (es2)* | GLM | 0.96 | 0.96\|0.01 | 0.96\|0.00 | 0.96\|0.00 | 0.96\|0.00 | 0.96\|0.00 | 0.96\|0.00 | 0.96\|0.00 | 0.96\|0.00 | 0.95\|0.00 | 0.96\|0.00 | 0.96\|0.01 | 0.96\|0.00 |
| | GAM | 0.96 | 0.95\|0.02 | 0.96\|0.01 | 0.96\|0.01 | 0.96\|0.01 | 0.95\|0.02 | 0.95\|0.01 | 0.96\|0.01 | 0.96\|0.01 | 0.95\|0.01 | 0.94\|0.01 | 0.95\|0.01 | 0.96\|0.01 |
| | BRT | 0.96 | 0.97\|0.01 | 0.96\|0.01 | 0.97\|0.01 | 0.97\|0.00 | 0.96\|0.01 | 0.95\|0.01 | 0.96\|0.01 | 0.96\|0.00 | 0.96\|0.01 | 0.96\|0.01 | 0.96\|0.01 | 0.96\|0.01 |
| | RF | 0.96 | 0.95\|0.02 | 0.95\|0.01 | 0.95\|0.01 | 0.96\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.96\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.96\|0.01 | 0.96\|0.01 |
| | SVM | 0.96 | 0.94\|0.02 | 0.95\|0.01 | 0.95\|0.01 | 0.94\|0.01 | 0.94\|0.02 | 0.95\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.94\|0.02 | 0.94\|0.02 | 0.95\|0.02 | 0.95\|0.01 |
| | MAX | 0.95 | 0.96\|0.01 | 0.96\|0.00 | 0.96\|0.01 | 0.96\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.95\|0.01 | 0.94\|0.01 | 0.93\|0.01 |
| *Caronella girondica (es3)* | GLM | 0.80 | 0.69\|0.03 | 0.68\|0.02 | 0.72\|0.02 | 0.74\|0.02 | 0.71\|0.04 | 0.74\|0.02 | 0.76\|0.02 | 0.76\|0.02 | 0.75\|0.02 | 0.76\|0.02 | 0.77\|0.02 | 0.78\|0.01 |
| | GAM | 0.76 | 0.72\|0.03 | 0.77\|0.02 | 0.77\|0.02 | 0.78\|0.02 | 0.69\|0.04 | 0.70\|0.03 | 0.74\|0.03 | 0.76\|0.02 | 0.73\|0.03 | 0.73\|0.02 | 0.74\|0.03 | 0.77\|0.02 |
| | BRT | 0.81 | 0.70\|0.03 | 0.74\|0.02 | 0.75\|0.02 | 0.75\|0.01 | 0.70\|0.03 | 0.71\|0.02 | 0.74\|0.02 | 0.75\|0.02 | 0.74\|0.02 | 0.76\|0.02 | 0.78\|0.02 | 0.79\|0.01 |
| | SVM | 0.80 | 0.73\|0.03 | 0.77\|0.02 | 0.77\|0.02 | 0.78\|0.01 | 0.71\|0.03 | 0.72\|0.02 | 0.74\|0.02 | 0.76\|0.02 | 0.74\|0.02 | 0.75\|0.02 | 0.77\|0.02 | 0.79\|0.01 |
| | RF | 0.80 | 0.72\|0.03 | 0.77\|0.02 | 0.77\|0.02 | 0.79\|0.02 | 0.72\|0.03 | 0.73\|0.02 | 0.75\|0.02 | 0.77\|0.01 | 0.75\|0.02 | 0.76\|0.02 | 0.78\|0.02 | 0.80\|0.01 |
| | MAX | 0.81 | 0.71\|0.03 | 0.78\|0.02 | 0.78\|0.02 | 0.79\|0.01 | 0.72\|0.04 | 0.75\|0.02 | 0.77\|0.02 | 0.77\|0.02 | 0.74\|0.03 | 0.74\|0.02 | 0.77\|0.02 | 0.79\|0.02 |

Comparing the scenarios for the models in *nl1* (Figs 3-4) shows that the mean decline in AUC values, for the S.low scenario, was 6.7% (range: 1.9% to 13.6%), and for the S.high scenario, 1.6% (range: 0% to 4.4%) compared to the AUC for the models using the unperturbed data. For this case study, consistent with the S.low and the S.high scenarios, the mean decline in the AUC values for the S.rand scenario was in between these two values with 3.6% (range: 0.1% to 7.7%), and for the S.all scenario,

8.4% (range: 2.9% to 12.6%). For this case study, the standard deviation of AUC within scenarios decreased when the local spatial association increased. For instance, the standard deviation in S.low, ranged between 0.025 and 0.058 (median = 0.045) and in S.high, ranged between 0.010 and 0.035 (median = 0.024).

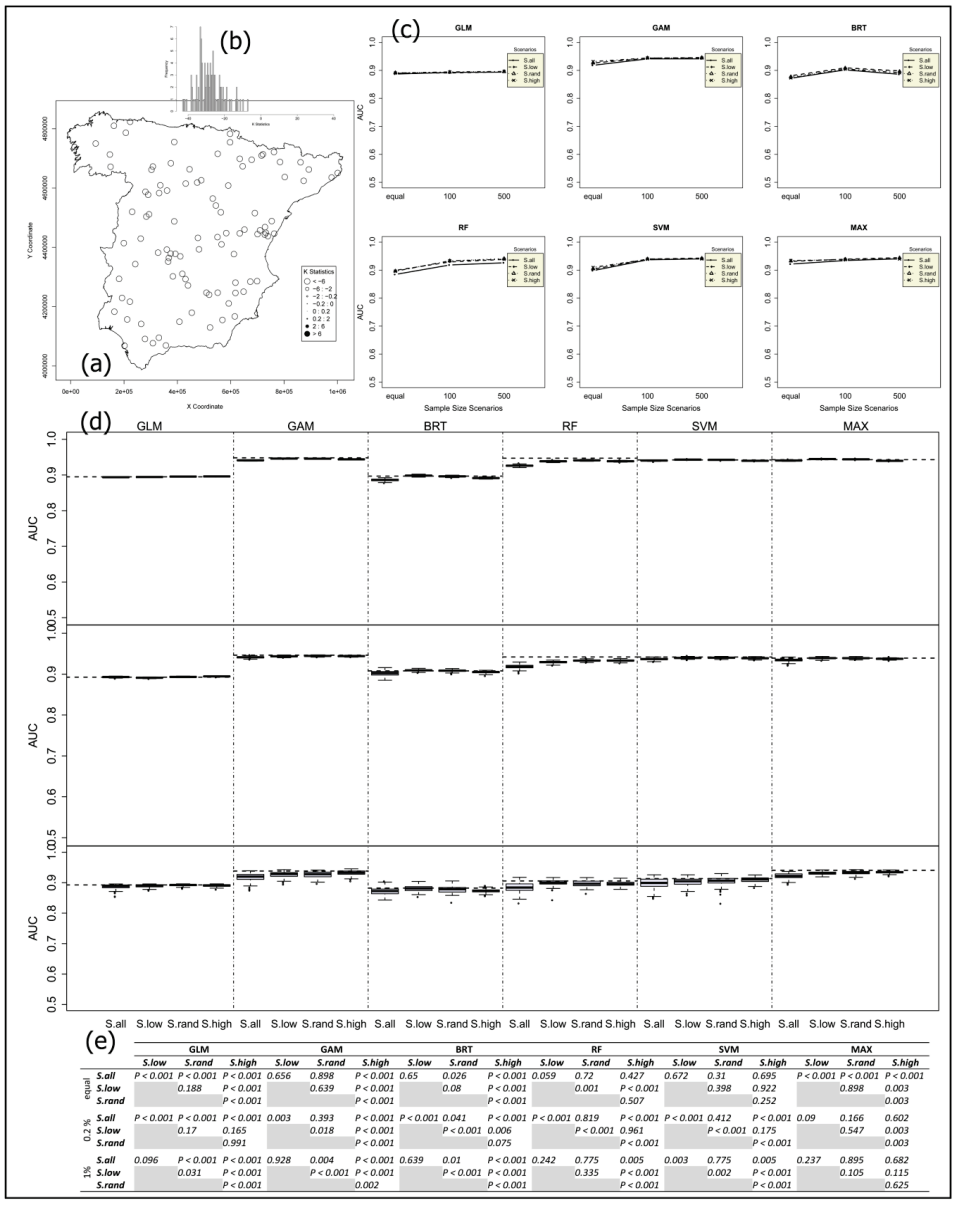
Comparing the scenarios for the models in *es1* (Figs 3-4) shows that the magnitude of the mean decline in AUC values was low. This decline, in the S.low, S.rand and S.high scenarios was 0.4% (range: 0% to 1.4%, 0% to 1.3% and 0% to 1.1%, respectively), and in S.all scenario was 1.1% (range: 0% to 2.5%) compared to the AUC for the models using unperturbed data. For this case study, the standard deviation, in the S.low, ranged between 0.000 and 0.013 (median = 0.002) and in the S.high, between 0.000 and 0.009 (median = 0.002).

For most scenarios, the AUC values slightly, but significantly, increased when the sample size was increased. The mean AUC for all models using the 'equal', '0.2%' and '1%' sample size scenarios, were 0.820, 0.824 and 0.840, respectively. The results of multiple ANOVA (sub-figure (e) in Figs 3-3, 3-4, 3-5, 3-6 and 3-7) revealed that the mean AUC for the models within the same spatial association scenario (*i.e.*, S.low, S.high and S.rand) but between sample size scenarios (*i.e.*, 'equal', '0.2%' and '1%') are not significantly different (except in *nl2*), suggesting that the behaviour of the models through the Monte Carlo simulation is generally the same. Comparing the standard deviation of the AUC values for all models that were implemented using perturbed data showed that the SVM was the most sensitive model to species positional error. However the differences in sensitivity between all models were small.



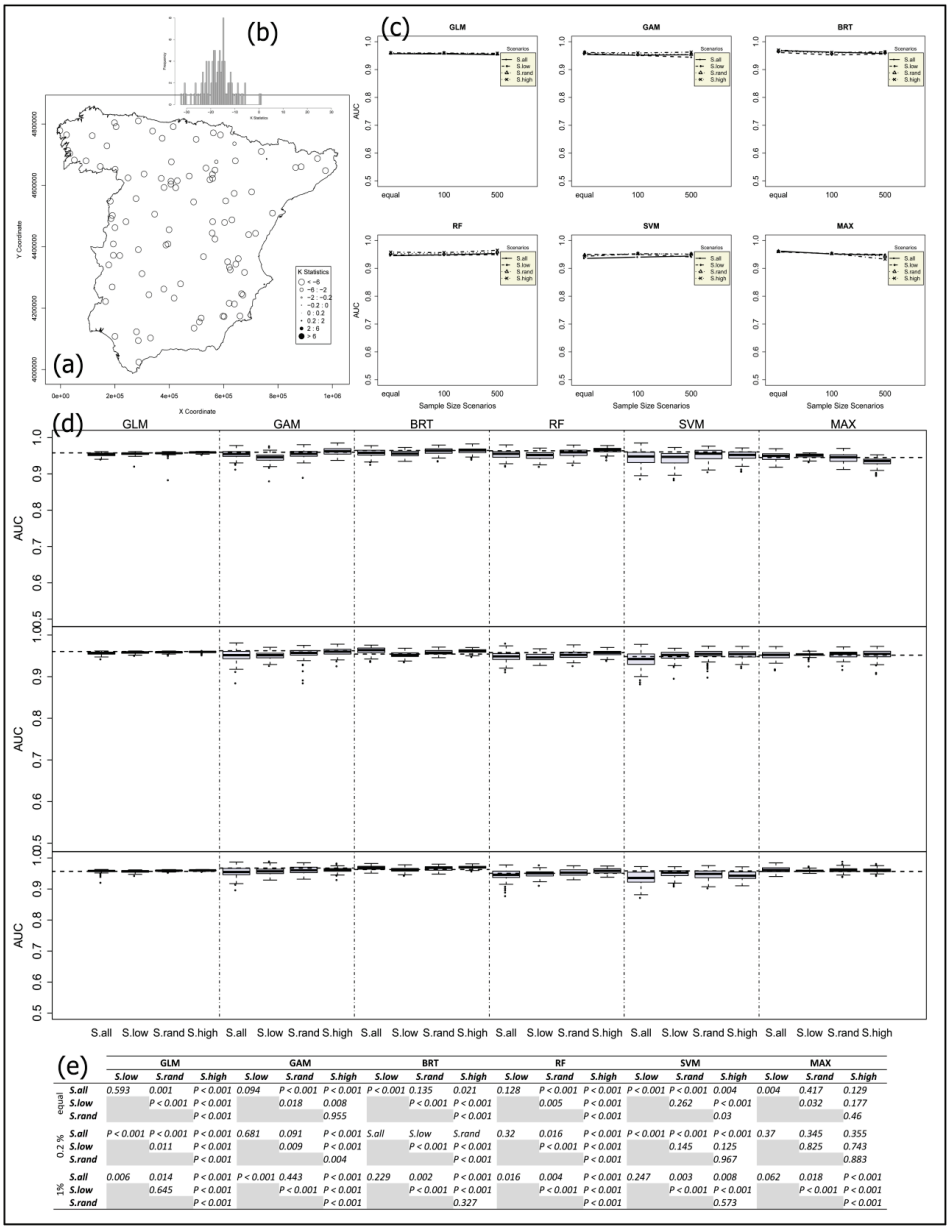

| | | GLM | | | GAM | | | BRT | | | RF | | | SVM | | | MAX | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high |
| equal | S.all | 0.052 | P < 0.001 | P < 0.001 | 0.003 | P < 0.001 | P < 0.001 | 0.061 | P < 0.001 | P < 0.001 | 0.109 | P < 0.001 | P < 0.001 | 0.09 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 |
| | S.low | | P < 0.001 | P < 0.001 | | P < 0.001 | 0.005 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 |
| | S.rand | | | 0.004 | | | 0.189 | | | 0.848 | | | P < 0.001 | | | P < 0.001 | | | 0.056 |
| 0.2 % | S.all | 0.747 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | 0.011 | 0.473 | P < 0.001 | 0.255 | P < 0.001 | P < 0.001 | 0.069 | P < 0.001 | P < 0.001 |
| | S.low | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | 0.595 | 0.017 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 |
| | S.rand | | | P < 0.001 | | | 0.15 | | | 0.098 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 |
| 1% | S.all | 0.568 | 0.074 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | 0.19 | 0.155 | P < 0.001 | 0.846 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 |
| | S.low | | 0.093 | P < 0.001 | | 0.675 | 0.101 | | 0.012 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | 0.512 | 0.048 |
| | S.rand | | | P < 0.001 | | | 0.239 | | | 0.251 | | | 0.001 | | | P < 0.001 | | | 0.164 |

**Figure 3-3:** The interaction of the local spatial association and positional uncertainty for the case study with the lowest local spatial association in predictors at species sample locations (i.e. *nl1*). (a) The level of local spatial association at the location of species occurrences, lower K indicates higher local spatial association. (b) Histogram of the K statistics. (c) Interaction plots based on the Friedman test – difference of AUC mean between three scenarios (S.all, S.low, S.rand, and S.high) and different sample size (x-axis) for different SDMs; S.all represents the scenario for which the positional error was introduced in all species sample locations; in the S.low and S.high scenarios, the positional error was introduced to half of all occurrences where the value of K statistics were lower and higher than median of the K statistics, respectively. In the S.rand, the positional error was introduced to the half of randomly selected occurrences (d) Variation of the model accuracy (AUC) over the Monte Carlo simulation for different scenarios of the impact of positional uncertainty on SDMs prediction based on the local spatial association (S.all, S.low, S.rand and S.high on x-axis) and six SDMs with increasing sample size; each box represents the results for 1000 Monte Carlo runs. (e) The level of significance for AUC mean comparison between different scenarios



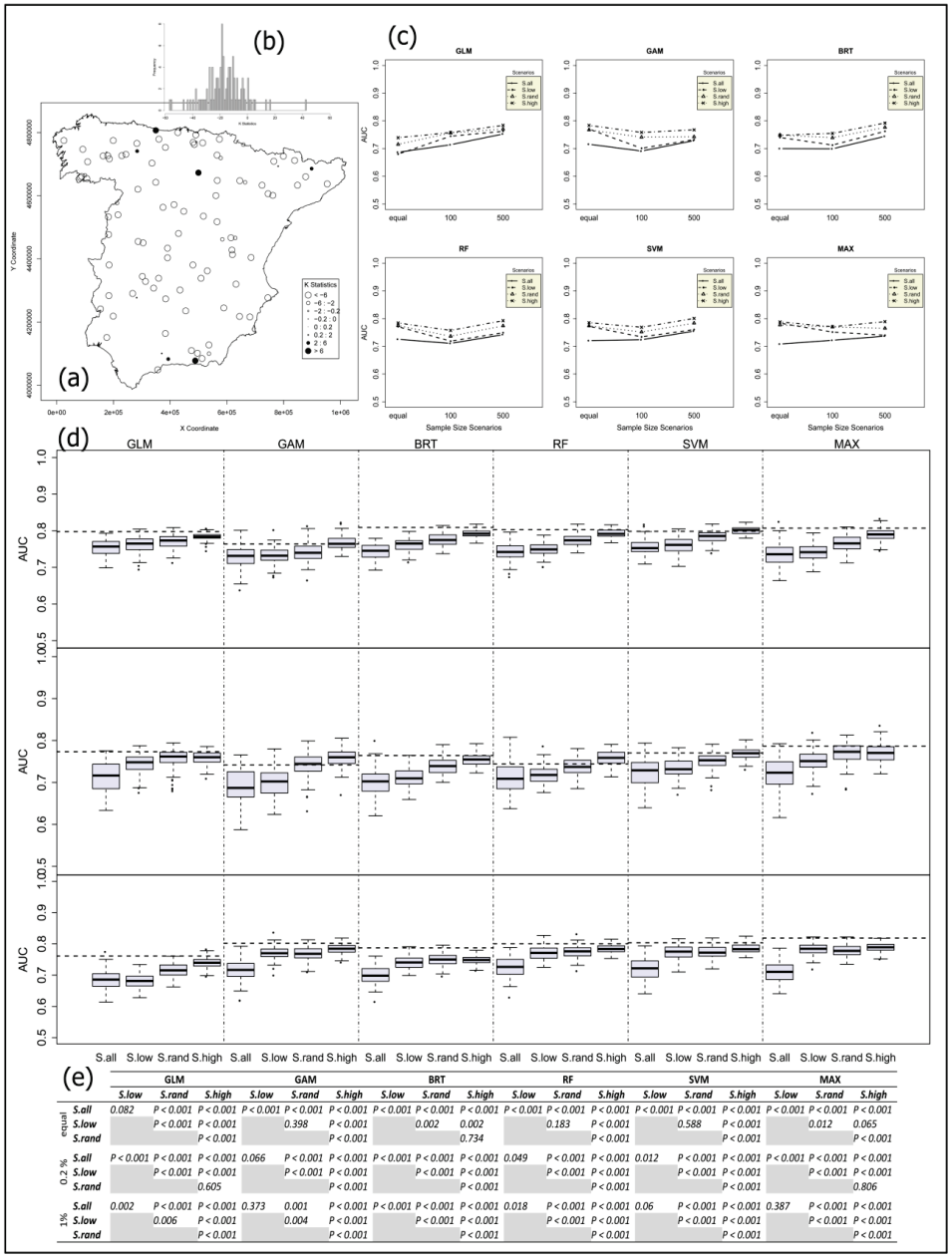

| | | GLM | | | GAM | | | BRT | | | RF | | | SVM | | | MAX | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high |
| equal | S.all | P < 0.001 | P < 0.001 | P < 0.001 | 0.656 | 0.898 | P < 0.001 | 0.65 | 0.026 | P < 0.001 | 0.059 | 0.72 | 0.427 | 0.672 | 0.31 | 0.695 | P < 0.001 | P < 0.001 | P < 0.001 |
| | S.low | | 0.188 | P < 0.001 | | 0.639 | P < 0.001 | | 0.08 | P < 0.001 | | 0.001 | P < 0.001 | | 0.398 | 0.922 | | 0.898 | 0.003 |
| | S.rand | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | 0.507 | | | 0.252 | | | 0.003 |
| 0.2 % | S.all | P < 0.001 | P < 0.001 | P < 0.001 | 0.003 | 0.393 | P < 0.001 | P < 0.001 | 0.041 | P < 0.001 | P < 0.001 | 0.819 | P < 0.001 | P < 0.001 | 0.412 | P < 0.001 | 0.09 | 0.166 | 0.602 |
| | S.low | | 0.17 | 0.165 | | 0.018 | P < 0.001 | | P < 0.001 | 0.006 | | P < 0.001 | 0.961 | | P < 0.001 | 0.175 | | 0.547 | 0.003 |
| | S.rand | | | 0.991 | | | P < 0.001 | | | 0.075 | | | P < 0.001 | | | P < 0.001 | | | 0.003 |
| 1% | S.all | 0.096 | P < 0.001 | P < 0.001 | 0.928 | 0.004 | P < 0.001 | 0.639 | 0.01 | P < 0.001 | 0.242 | 0.775 | 0.005 | 0.003 | 0.775 | 0.005 | 0.237 | 0.895 | 0.682 |
| | S.low | | 0.031 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | 0.335 | P < 0.001 | | 0.002 | P < 0.001 | | 0.105 | 0.115 |
| | S.rand | | | P < 0.001 | | | 0.002 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | 0.625 |

**Figure 3-4:** The interaction of the local spatial association and positional uncertainty for the case study with the highest local spatial association at species sample locations (*i.e.*, *es1*). The different sub-figures are described in Fig. 3-3



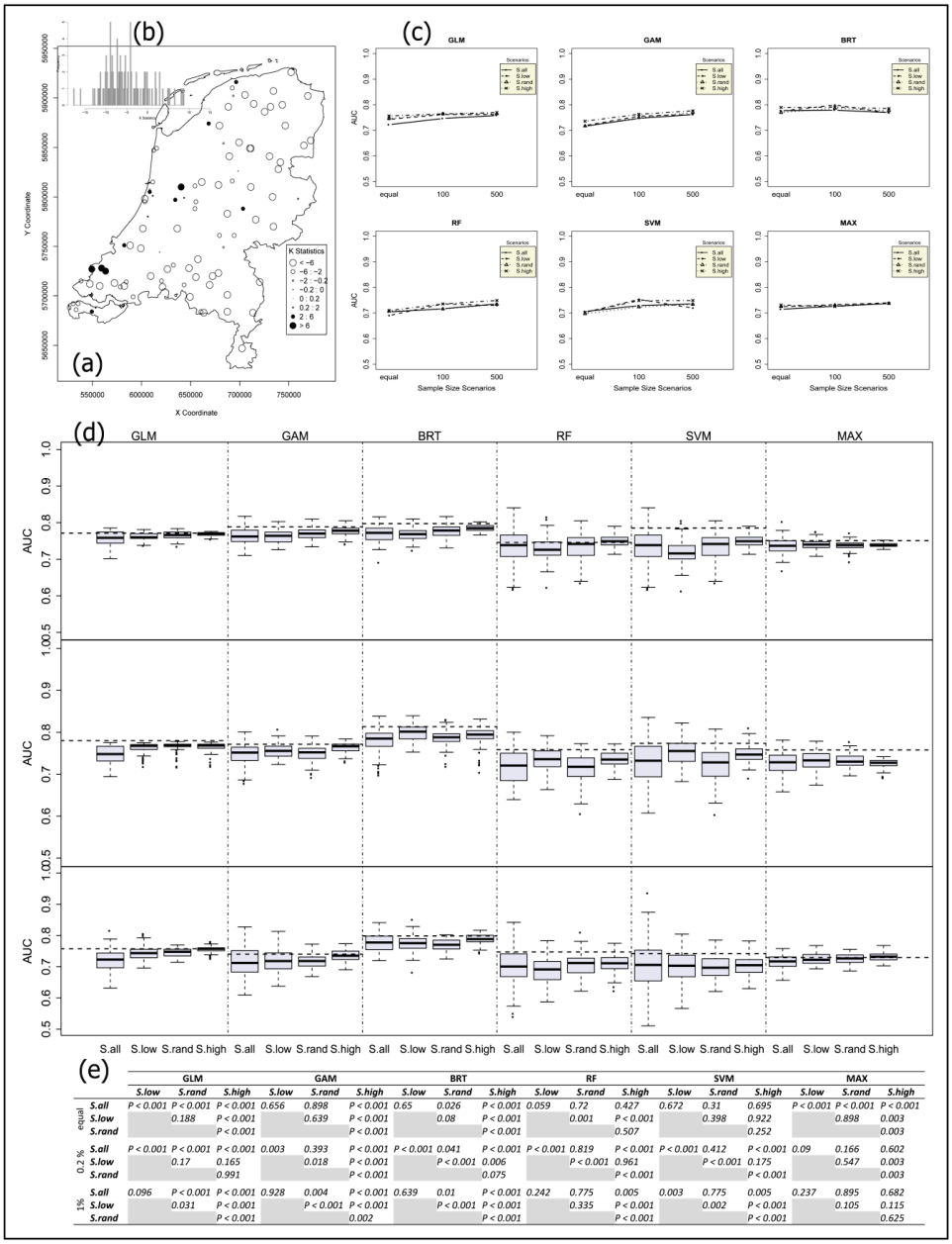

| | | GLM | | | GAM | | | BRT | | | RF | | | SVM | | | MAX | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* |
| equal | ***S.all*** | 0.593 | 0.001 | P < 0.001 | 0.094 | P < 0.001 | P < 0.001 | P < 0.001 | 0.135 | 0.021 | 0.128 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | 0.004 | 0.004 | 0.417 | 0.129 |
| | ***S.low*** | | P < 0.001 | P < 0.001 | | 0.018 | 0.008 | | P < 0.001 | P < 0.001 | | 0.005 | P < 0.001 | | 0.262 | P < 0.001 | | 0.032 | 0.177 |
| | ***S.rand*** | | | P < 0.001 | | | 0.955 | | | P < 0.001 | | | P < 0.001 | | | 0.03 | | | 0.46 |
| 0.2 % | ***S.all*** | P < 0.001 | P < 0.001 | P < 0.001 | 0.681 | 0.091 | P < 0.001 | S.all | S.low | S.rand | 0.32 | 0.016 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | 0.37 | 0.345 | 0.355 |
| | ***S.low*** | | 0.011 | P < 0.001 | | 0.009 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | 0.145 | 0.125 | | 0.825 | 0.743 |
| | ***S.rand*** | | | P < 0.001 | | | 0.004 | | | P < 0.001 | | | P < 0.001 | | | 0.967 | | | 0.883 |
| 1% | ***S.all*** | 0.006 | 0.014 | P < 0.001 | P < 0.001 | 0.443 | P < 0.001 | 0.229 | 0.002 | P < 0.001 | 0.016 | 0.004 | P < 0.001 | 0.247 | 0.003 | 0.008 | 0.062 | 0.018 | P < 0.001 |
| | ***S.low*** | | 0.645 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 |
| | ***S.rand*** | | | P < 0.001 | | | P < 0.001 | | | 0.327 | | | P < 0.001 | | | 0.573 | | | P < 0.001 |

**Figure 3-5:** The interaction of the local spatial association and positional uncertainty for the case study of es2. The different sub-figures are described in Fig. 3-3




| | | GLM | | | GAM | | | BRT | | | RF | | | SVM | | | MAX | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high | S.low | S.rand | S.high |
| equal | S.all | 0.082 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 |
| | S.low | | P < 0.001 | P < 0.001 | | 0.398 | P < 0.001 | | 0.002 | 0.002 | | 0.183 | P < 0.001 | | 0.588 | P < 0.001 | | 0.012 | 0.065 |
| | S.rand | | | P < 0.001 | | | P < 0.001 | | | 0.734 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 |
| 0.2 % | S.all | P < 0.001 | P < 0.001 | P < 0.001 | 0.066 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | 0.049 | P < 0.001 | P < 0.001 | 0.012 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 |
| | S.low | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 |
| | S.rand | | | 0.605 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | 0.806 |
| 1% | S.all | 0.002 | P < 0.001 | P < 0.001 | 0.373 | 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | P < 0.001 | 0.018 | P < 0.001 | P < 0.001 | 0.06 | P < 0.001 | P < 0.001 | 0.387 | P < 0.001 | P < 0.001 |
| | S.low | | 0.006 | P < 0.001 | | 0.004 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 | | P < 0.001 | P < 0.001 |
| | S.rand | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 | | | P < 0.001 |

**Figure 3-6:** The interaction of the local spatial association and positional uncertainty for the case study of *es3*. The different sub-figures are described in Fig. 3-3




| | | GLM | | | GAM | | | BRT | | | RF | | | SVM | | | MAX | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* | *S.low* | *S.rand* | *S.high* |
| equal | ***S.all*** | *P < 0.001* | *P < 0.001* | *P < 0.001* | *0.656* | *0.898* | *P < 0.001* | *0.65* | *0.026* | *P < 0.001* | *0.059* | *0.72* | *0.427* | *0.672* | *0.31* | *0.695* | *P < 0.001* | *P < 0.001* | *P < 0.001* |
| | ***S.low*** | | *0.188* | *P < 0.001* | | *0.639* | *P < 0.001* | | *0.08* | *P < 0.001* | | *0.001* | *P < 0.001* | | *0.398* | *0.922* | | *0.898* | *0.003* |
| | ***S.rand*** | | | *P < 0.001* | | | *P < 0.001* | | | *P < 0.001* | | | *0.507* | | | *0.252* | | | *0.003* |
| 0.2 % | ***S.all*** | *P < 0.001* | *P < 0.001* | *P < 0.001* | *0.003* | *0.393* | *P < 0.001* | *P < 0.001* | *0.041* | *P < 0.001* | *P < 0.001* | *0.819* | *P < 0.001* | *P < 0.001* | *0.412* | *P < 0.001* | *0.09* | *0.166* | *0.602* |
| | ***S.low*** | | *0.17* | *0.165* | | *0.018* | *P < 0.001* | | *P < 0.001* | *0.006* | | *P < 0.001* | *0.961* | | *P < 0.001* | *0.175* | | *0.547* | *0.003* |
| | ***S.rand*** | | | *0.991* | | | *P < 0.001* | | | *0.075* | | | *P < 0.001* | | | *P < 0.001* | | | *0.003* |
| 1% | ***S.all*** | *0.096* | *P < 0.001* | *P < 0.001* | *0.928* | *0.004* | *P < 0.001* | *0.639* | *0.01* | *P < 0.001* | *0.242* | *0.775* | *0.005* | *0.003* | *0.775* | *0.005* | *0.237* | *0.895* | *0.682* |
| | ***S.low*** | | *0.031* | *P < 0.001* | | *P < 0.001* | *P < 0.001* | | *P < 0.001* | *P < 0.001* | | *0.335* | *P < 0.001* | | *0.002* | *P < 0.001* | | *0.105* | *0.115* |
| | ***S.rand*** | | | *P < 0.001* | | | *0.002* | | | *P < 0.001* | | | *P < 0.001* | | | *P < 0.001* | | | *0.625* |

**Figure 3-7:** The interaction of the local spatial association and positional uncertainty for the case study of *nl2*. The different sub-figures are described in Fig. 3-3

## 3-4. Discussion

This study shows that the positional uncertainty in species occurrence data matters more in locations with low local spatial association in the predictors. To show this effect, we first explored whether positional uncertainties in species occurrences decreases the accuracy of SDMs, and second, examined which occurrence locations have more impact on the prediction of the SDMs. Our approach is formalized based on examining local spatial association in predictors. The results presented in Figs from 3-3 to 3-7 suggest that the most accurate models are least sensitive to positional uncertainty. A formal evaluation of this would require additional experimentation and we leave that for future research.

The link between global spatial autocorrelation in predictors and robustness of SDMs has been already demonstrated by Osborne and Leitão (2009) and Naimi et al. (2011). The methodology presented in this study extends this by using the local spatial association in predictors. This offers several advantages. First, examining global spatial autocorrelation in predictors provides insight into whether predictions are likely to be affected by the uncertainty in the sample locations, while our approach leads to identification of local areas with a high degree of influence. Of key importance is that an appropriate strategy can be considered to overcome the problem. For example, one may simply exclude observations that are located in areas with low local spatial association and at the same time have high positional uncertainty. This, however, is not a good option because it may bias the sample data. The interpretation of SDMs built with biased data should be made with explicit awareness of the potential problems of those biases (Leitão et al. 2011). In particular, the vast majority of data available for species distribution modelling are often incomplete and biased in relation to the true spatial distributions of species (Araújo and Guisan 2006 ; Bystriakova et al. 2012). Setting up a completely new survey to generate a new dataset using an appropriate sampling design is appealing but unfeasible in many circumstances (Araújo and Guisan 2006); however

our approach can help to target locations that could be selected for additional field sampling. A limited survey then may be designed for these areas to provide or modify the sample locations which are located at the problematic area. Expert knowledge may be used as another solution to modify sampling schemes for the locations that are targeted.

The second advantage of using local spatial association is that it does not rely on the assumption of global stationarity. Using the global spatial autocorrelation measure, as proposed in Naimi et al. (2011), is valid when the stationarity can be assumed. When this assumption is not met, the presented method provides an alternative.

In this study, an uncommon indicator of local spatial association, the K statistic, was used. Although there exist more commonly used indicators for local spatial association, (such as local Moran's *I* and local Geary's *c*), these are influenced by the presence of global spatial autocorrelation (Getis and Ord 1996) and must therefore be interpreted according to the degree of global spatial autocorrelation in the data. This makes them less suitable to assess local spatial association between points, when averaging over different layers with different degrees of global spatial autocorrelation.

We selected five species, occurring in two different countries to test our approach to cover a broad spectrum of situations with respect to species' characteristics. However, our case studies do not cover all possible combinations of situations that may be found in species-environment relationship studies. Species may vary considerably in commonness, range size, habitat preferences, and population trend (Guisan et al. 2007 ; Mcpherson and Jetz 2007). It has been shown that these characteristics influence significantly the accuracy of SDMs (Mcpherson and Jetz 2007). Generally, models for species that have broad geographic ranges and high environmental tolerances (*i.e.*, generalists) tend to be less accurate than those for species with smaller geographic range and limited environmental tolerances, *i.e.*, specialists (Hernandez et al. 2006 ; Manel et al. 2001). Further systematic exploration is required to test whether, and which of

the mentioned ecological characteristics matters when our approach is applied. We expect that our approach should work more effectively for species with the limited environmental tolerances, because the SDMs for such species are more sensitive to positional uncertainty (*i.e.*, it is more likely that positional error leads to associations with locations of incorrect environmental attributes).

We developed a new approach for simulating artificial data. In recent years, the number of studies that use simulated datasets has increased. This is because it provides advantages of having an assumed "truth", while avoiding the influence of unknown underlying complexity on the evaluation of the models. In order to simulate species distribution, assumed species response curves to environmental gradients are commonly used (Austin et al. 2006 ; Jiménez-Valverde et al. 2009 ; Meynard and Quinn 2007 ; Naimi et al. 2014 ; Naimi et al. 2011). Using such simplifications to simulate a virtual species does not cover all complications that are likely to be found in real data. Hence there is a risk that modelling with virtual species does not correctly simulate reality (Hirzel et al. 2001). Simulating realistic artificial data should be consistent with relevant ecological processes but is limited by our understanding of such processes (Austin et al. 2006). In this study, we used a multiple-model approach to link the real species data to real environmental gradients. Different models in this approach differ in the procedure to derive response surfaces, and therefore the resulting niche shape is not in favour of one particular response curve. The combined multiple-model prediction is likely to be more accurate than a single model (Araújo and New 2007) and was considered as a simulated distribution and a reference for each species in this study. Furthermore, we used a particular procedure to select the most relevant environmental variables for each species among the possible variables. This is, however, a procedure to approximately find true predictors. In real situations true predictors are generally unknown (Hirzel et al. 2001). We assert that our approach for simulating species distribution is more realistic than the one where the habitat is simulated using a priori imposed response curves.

Instead of using a response curve based on a priori assumption on ecological theory, we generated empirically derived response curves that fall within ecological ranges that are present in the real species data. However, other approaches of simulating artificial dataset may be more appropriate in some situations. For example, using simulated environmental data (rather than real data) is necessary when the aim is to explore model sensitivity to a property in environmental data (*e.g.*, as in chapter 2); or using assumed species response curves to environmental gradients for generating virtual species when the aim is to understand the effect of response shape on model or to test a certain ecological theory.

Finally, we developed a package in the R environment for statistical computing (R Development Core Team 2013) to calculate the level of local spatial association in the predictors at the location of species occurrences. Given species sample locations, an estimate of the positional uncertainty, as well as a set of predictors, a function in the package calculates the K statistic for each predictor at each sample location. The K statistics of the predictors are then aggregated. Mapping the results allows us to target the locations that are likely to affect the predictions from the SDMs. The function takes the importance of predictor variables as the weights into account when aggregating the K statistics. To use the function, we recommend that a pre-analysis of an SDM is required for calculating variables' importance and excluding unimportant and/or collinear variables (as is demonstrated in this study). We would like to emphasize that our tool can be used for any study that uses SDMs but is hampered by concerns about positional uncertainty.

## *3-5. Conclusions*

A key challenge in using a great majority of available species occurrence records in museum and herbaria for species distribution modelling is positional uncertainty. In this study, we proposed a method to test whether and where this uncertainty is problematic for SDMs. We have shown that the impact of positional uncertainty in species occurrence data on the

predictions of the species distribution modelling is related to the level of local spatial association in the predictors. Our results indicate that the species occurrence locations where local spatial associations in the predictors was lower, affect the SDMs significantly more than the locations with higher local spatial association in the predictors. We suggest examining a local indicator of spatial association in predictors at species occurrence locations when species data are subjected to positional uncertainty. This can give insight into whether the positional uncertainty in the sample locations affects the prediction accuracy of SDMs, and to detect which sample locations are likely to affect the predictions. This can also be used as a basis to target the observations where species occurrence are observed but need treatment of the positional uncertainty. We propose the use of the local K statistic for this purpose.

# *Chapter 4*

# A New Local Indicator of Spatial Association

*This chapter is based on:*
Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K. & Toxopeus, A. G. (under review) ELSA: entropy-based local indicator of spatial association. *Geographical Analysis,*

# 4. A new Local Indicator of Spatial Association

## *4-1. Introduction*

Spatial analysis is concerned with exploration and identification of associations over geographical space. Such associations quantify the degree to which a value of a variable measured at one location is dependent on the values of the same variable measured at a specific geographic distance from that location (Cliff and Ord 1981 ; Goodchild 1986). If such dependency exists in a dataset, the variable is said to exhibit spatial autocorrelation (Sokal and Oden 1978). Several statistics have been developed to quantify spatial autocorrelation both globally and locally. Global measures provide a statistic for the entire field under the assumption of spatial stationarity, *i.e.* the mean and covariance do not vary over space. This assumption is often unrealistic. Recent advances have addressed non-stationarity in the mean through spatially varying coefficient modelling (Finley 2011 ; Gelfand et al. 2003). Heteroskedasticity in the variance can be addressed through a weighting function (Hamm et al. 2012 ; Lark 2009) and further efforts have been directed at modelling non-stationary covariance functions (Haskard and Lark 2009 ; Paciorek and Schervish 2006). Nevertheless, these models are often difficult to implement and there is a lack of standard software tools. Further, at some point all these models must make an assumption of stationarity and all provide a global measure. This may be of limited relevance when local areas are of interest.

Recent researches on spatial data analysis developed a number of local spatial statistics (Anselin 1995 ; Boots 2003 ; Getis and Ord 1992 ; Ord and Getis 1995). In contrast to global measures these allow exploration of local patterns in spatial association. They rely less on the assumption of global stationarity. Seminal papers of this type were published by Getis and Ord (1992), Anselin (1995), and Ord and Getis (1995). Anselin (1995) introduced a set of statistics, called local indicators of spatial association (LISA), including local Moran's I and local Geary's c, that decompose a single global measure into the contribution of each individual location.

Thus these statistics reveal which observations have most impact on the global measure. Getis and Ord (1992) and Ord and Getis (1995) also defined two local statistics, Gi and Gi*, which are somewhat different from Anselin's LISAs, indicating local clustering of high and low values. These allow detection of pockets of spatial association that may not be evident when using global statistics. These methods, however, can be used for continuous or interval variables only. Despite numerous situations where qualitative (nominal/categorical) variables are encountered, only a few attempts have been devoted to develop methods to explore the spatial pattern for categorical data. Examples include early work using joint count statistics (Dacey 1968 ; Moran 1948), and more recent works by Boots (2003) and Ruiz et al. (2010). To our knowledge, there is no statistic of local spatial association that can be used for both continuous and categorical data at the same time.

In this paper, we propose a new local indicator of spatial association, called the entropy-based local indicator of spatial association (ELSA), for exploratory analysis of both continuous and categorical spatial data at the same time. Entropy has its root in thermodynamic and information theory. Entropy based approaches have been applied in many disciplines, as a measure of complexity in physics (López-Ruiz et al. 1995), as a measure of diversity or structural complexity in ecology (Anand and Orloci 1996 ; Ricotta and Anand 2006), and as a measure of information content or uncertainty in information theory (Yeung 2008). This concept has also been used by geographers, economists and social scientists to describe spatial phenomena (Batty 1974 ; Batty 1976 ; Heikkila and Hu 2006). Recently, some attempts have been conducted to use the concept of entropy as a measure of spatial contiguity for qualitative data (Ruiz et al. 2010), or to detect spatially varying multivariate relationships (Guo 2010). Matilla-García et al. (2011) highlighted the potential role that entropy-based measures might play in detecting spatial structure. They used the spatial symbolic entropy for detecting the order of contiguity (spatial lag) of a spatial dependent process (Matilla-García and Marín 2011).

Although entropy-based approaches have been widely used for categorical data (e.g. soil type, classified data), there is a challenge in using entropy for exploratory analysis of continuous data. For continuous data the probability density function is often unknown, which is necessary to calculate entropy (Guo 2010). A common solution to this problem in most practical applications is to construct a contingency table by binning (discretizing) a continuous variable into a finite number of classes (Guo 2003 ; Journel and Deutsch 1993), which then are used as categorical data in the entropy measure. This, however, raises two new challenges. Firstly, binning continuous variables causes information loss in data. Secondly, dissimilarity between binned data is not the same as it is in categorical data. For example, when using a categorical land use map in spatial analysis, no difference in the level of dissimilarity between pairs of classes is assumed. If a continuous variable binned into, for example five categories (C1, … , C5), then C1 is more similar to C2 than to C5. In this paper, we illustrate how ELSA addresses these challenges.

The main objectives of our study are: (1) to introduce a new statistic for measuring local spatial association (ELSA) that can be used for both continuous and categorical spatial data; (2) to explore the application of ELSA for calculating local spatial association and comparing this with other indicators; (3) to demonstrate the usability of ELSA to detect global patterns as well, and calculate a variogram-like global spatial structure, named 'entrogram'.

## *4-2. Local spatial statistics*

In this section, we provide a brief review on local spatial statistics that can be used for both exploratory and inferential spatial data analysis. These statistics evaluate how the strength of spatial association varies with locations within the study area. We review four local statistics for quantitative data: Moran's Moran's $I_i$, Geary's $c_i$, $G_i$ and $G_i^*$, and two local statistics for binary categorical data: local composition and joint

count. The latter are also referred to as local indicators for categorical data; LICD.

In this paper our standardized notation is as follows: $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ is the set of observed attribute values at $n$ locations. The locations themselves are denoted $\mathbf{u} = (\mathbf{u_1}, \mathbf{u_2}, \dots, \mathbf{u_n})'$, such that $x_i$ is the attribute observed at location $\mathbf{u_i}$. The subscripts $i$ and $j$ are indices referring to specific observations.

### 4-2-1. Moran's $I_i$ and Geary's $c_i$

Anselin (1995) defined local indicators of spatial association (LISA) with a motivation of decomposing global statistics such as Moran's $I$ and Geary's $c$ into their local components to explore the locations that are major contributors to the global autocorrelation. In the absence of global spatial autocorrelation, these statistics can be used to test if local spatial clustering of similar values around the observation is significantly different from the global mean.

Local Moran's $I$ at site $i$ is:

$$I_i = \frac{(x_i - \overline{x})}{s^2} \sum\nolimits_j w_{ij} \left(x_j - \overline{x}\right) \qquad (4\text{-}1)$$

where $I_i$ is Moran's $I$ for site $i$, $w_{ij}$ is a spatial weight for an observation $j$ (often binary, i.e. 1 for neighbouring locations and 0 elsewhere), $\overline{x}$ and $s^2$ are the global sample mean and variance of the observations, respectively, and $x_i$ and $x_j$ are the observed values at site $i$ and site $j$ (*i.e.*, the neighbourhood of site $i$), respectively. Positive values of $I_i$ indicate a cluster of similar values around site $i$, and negative values indicate that the neighbour values are dissimilar to site $i$.

Local Geary's $c$ statistic at site $i$ is:

$$c_i = \frac{1}{s^2} \sum\nolimits_j w_{ij} \left(x_i - x_j\right)^2 \qquad (4\text{-}2)$$

with the same variables as Moran's $I$. This statistic quantifies a standardized squared distance between the values at site $i$ and the neighbour locations $j$. High values of $c_i$ indicate substantial differences between site $i$ and its neighbours, and low values of $c_i$ indicate that site $i$ is similar to its neighbours.

### 4-2-2. Local $G_i$ and $G_i^*$

Getis & Ord (1992) and Ord & Getis (1995) introduced another approach to estimate local spatial autocorrelation. These two statistics ($G_i$ and $G_i^*$) identify spatial clusters of high or low values around site $i$, by comparing local averages to global averages (Getis and Ord 1992 ; Ord and Getis 1995).

$G_i$ is calculated as:

$$G_i = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}, j \neq i; \qquad (4\text{-}3)$$

where $x_j$ is the set of observed values, and $w_{ij}$ is a weight for an observation $j$. $\sum_j x_j$ is the sum of the values at all locations except site $i$. If spatial association exists, it will show spatial clustering of high or low values of $x$ (Getis and Ord 1996). High or low values of $G_i$ indicate clusters of high or low values, respectively.

The $G_i^*$ statistic is the same as $G_i$, but does not exclude the attribute value at site $i$ in its calculation (*i.e.* $j$ may equal $i$).

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j} \qquad (4\text{-}4)$$

Although $G_i$ and $G_i^*$, are quite similar in practice, the latter is preferred because the similarity of $x_i$ to its neighbours is quantified (Getis and Ord 1996).

These statistics provide different information about the spatial structure. Local Moran's $I$ measures deviations of locations from the global mean. Local Geary's $c$ summarizes differences between a site and the values of its neighbouring sites. The Getis and Ord statistics ($G_i$ and $G_i^*$) tend to agree more with $I_i$ than $c_i$ (Sokal et al. 1998). These local statistics are influenced by the presence of global spatial autocorrelation and must be interpreted according to the degree of global spatial autocorrelation in the data (Ord and Getis 1995 ; Ord and Getis 2001).

### 4-2-3. Local indicators for categorical data (LICD)

LICD have been designed for exploring patterns in binary categorical data. These measures use conventional landscape pattern indices (O'neill et al. 1988) and are separated into two components: composition (proportion of one class) and configuration (characteristics of the spatial distribution of classes). The purpose of LICD is to identify the nature and spatial extent of local neighbourhoods that are distinctive or unusual compared to a priori expectation (Boots 2003 ; Boots 2006). For example, the LICD compositional measure tests whether the composition of any local neighbourhood is significantly different from the global composition. The join counts statistic, as a spatial configuration measure, is used locally by expressing the number of black/white, black/black or white/white joins in a local area as a proportion of all joins in that area, and then tests whether this proportion differs significantly from the proportion in the entire area (for more detail, see Boots, 2003 and Boots, 2006).

## *4-3. Entropy*

In information theory, the concept of complexity is closely related to predictability. It is the amount of information required to achieve an optimal prediction (Boschetti 2008). The entropy measure is also known as the information content. The Shannon entropy has been defined as an average amount of information to eliminate uncertainty, given by a finite number of events:

$$H = -\sum_{k=1}^{m} p_k \log_2 p_k \qquad (4\text{-}5)$$

where $H$ measures the entropy of a system with a finite number of $m$ possible events, and $p_k$ represents the probability of event $k$. $H$ is at a maximum when all events occur in equal abundance and can be quantified by $\log_2 m$. This measure can be standardized by dividing by $\log_2 m$, providing a measure of relative entropy ranging between 0 and 1. The function is dimensionless and depends only upon the number of events, not upon any other invariant property of the system to which it is applied (Batty 1976).

## *4-4. ELSA*

The Entropy based Local indicator of Spatial Association, ELSA, extends the above described entropy measure using a term that summarizes the attribute distance between a location and its neighborhood locations over a given geographical distance.

As previously $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ are $n$ observations related to a spatial process at locations $\mathbf{u} = (\mathbf{u_1}, \mathbf{u_2}, \dots, \mathbf{u_n})'$. Further, denote by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m)$ the set of possible categories that $x_i$ can take. For a categorical variable, it is usually assumed that pairs of categories are equally dissimilar. There are, however, situations where the level of dissimilarity varies for different pairs of categories. For example, 'dense forest' and 'sparse forest' in a land use map are more similar than a pair of either of these two classes 'dense forest' and 'lake' (for more details, see the Section "ELSA for categorical data"). Likewise, when the values of a continuous variable are binned (discretized) into categories, the level of dissimilarity varies between categories. By ranking the binned values, the difference between rank numbers can be interpreted as the level of dissimilarity. These levels of dissimilarity between categories are taken into account in the calculation of ELSA.

ELSA (*E* statistic) at site $i$ is defined as:

$$E_i = -\frac{\sum_j \omega_{ij} d_{ij}}{\max\{d\} \sum_j \omega_{ij}} \times \frac{\sum_{k=1}^{m_\omega} p_k \log_2(p_k)}{\log_2 m_i}, j \neq i$$

$$m_i = \begin{cases} m \text{ if } \sum_j \omega_{ij} > m \\ \sum_j \omega_{ij}, \text{otherwise} \end{cases} \quad (4\text{-}6)$$

$$d_{ij} = |c_i - c_j|$$

where $\omega_{ij}$ is a binary weight which specifies whether the site $j$ is within a specified distance from site $i$. Next, $d_{ij}$ describes the dissimilarity between $x_i$ and $x_j$, which is calculated as the absolute difference of the ranks assigned to the categories at sites $i$ and $j$ (*i.e.*, $c_i$ and $c_j$), and $max\{d\}$ is the maximum possible dissimilarity between any pair of observations in the entire map. This is discussed for continuous and categorical variables in the upcoming sections. There are $m$ categories in the entire map, $p_k$ is the probability of $k$*th* category from the $m_\omega$ categories within the local distance from site $i$, and $m_i$ is the maximum possible number of categories within the local distance from site $i$. This means that if the number of observations within the local distance from site $i$, including site $i$, is greater than the number of categories in the entire map ($\sum_j \omega_{ij} > m$), then $m_i$ is equal to the number of categories, otherwise it is equal to the number of observations ($\sum_j \omega_{ij}$) within the local distance from site $i$.

The first term on the right hand side of *Eq.* 4-6 is a coefficient that summarizes the attribute distance between site $i$ and the neighboring sites. This coefficient is bounded between 0 and 1. Low values indicate high similarity of site $i$ to neighboring sites, and high values indicate low similarity with neighbouring sites.

The second term on the right hand side of the $E_i$ statistic is the Shannon entropy (*Eq.* 4-5), normalized by $\log_2(m_i)$. This term ranges between 0

and 1. By normalizing, values are invariant to the number of categories present in a dataset. In other words, datasets with different numbers of categories are comparable. For normalizing, $m_i$ is defined in relation with the global number of categories in the entire map. This term quantifies diversity of the categories within the local distance from site $i$, but it is not sensitive to the level of dissimilarities between pairs of observations.

### 4-4-1. ELSA for continuous data

A key step to calculate ELSA for continuous data is that the variable should be first categorized (binned or grouped) into a number of categories; a procedure that may cause information loss. Inspired by Morrison (1972), we estimate the optimum number of categories that minimizes the information loss. This optimum is the minimum number of categories that is able to reproduce the spatial data statistically (i.e., the amount of information stored in the data is not affected through categorization). To find the optimum number of categories, our procedure uses Spearman's rank correlation coefficient, $\rho$, as a measure of information between the continuous variable and the categorized variable. If the amount of information is not affected through categorizing, the observed correlation should be equal to one. Any loss of information would result in the observed correlation to be less than one. Therefore the magnitude of the difference $1-\rho$ provides a measure of information loss (Quester and Dion 1997). The procedure of selecting the optimum number involves the following steps:

- The categorization procedure is repeated iteratively taking different number of categories, $m = 1, 2, \dots, \max\{m\}$
- The procedure assigns a rank number (between 1 and $m$, where $m$ is the total number of categories) to each category.
- The $\rho$ coefficient between the continuous values and the assigned ranks is calculated for each iteration.

- The minimum number of categories for which $\rho$ is greater than 0.99 is selected as the optimum number of categories, $m$.

It is assumed that the information loss due to the categorization is not substantial when the optimum number is used. Fig. 4-1, illustrates an example of using this procedure for a continuous variable.

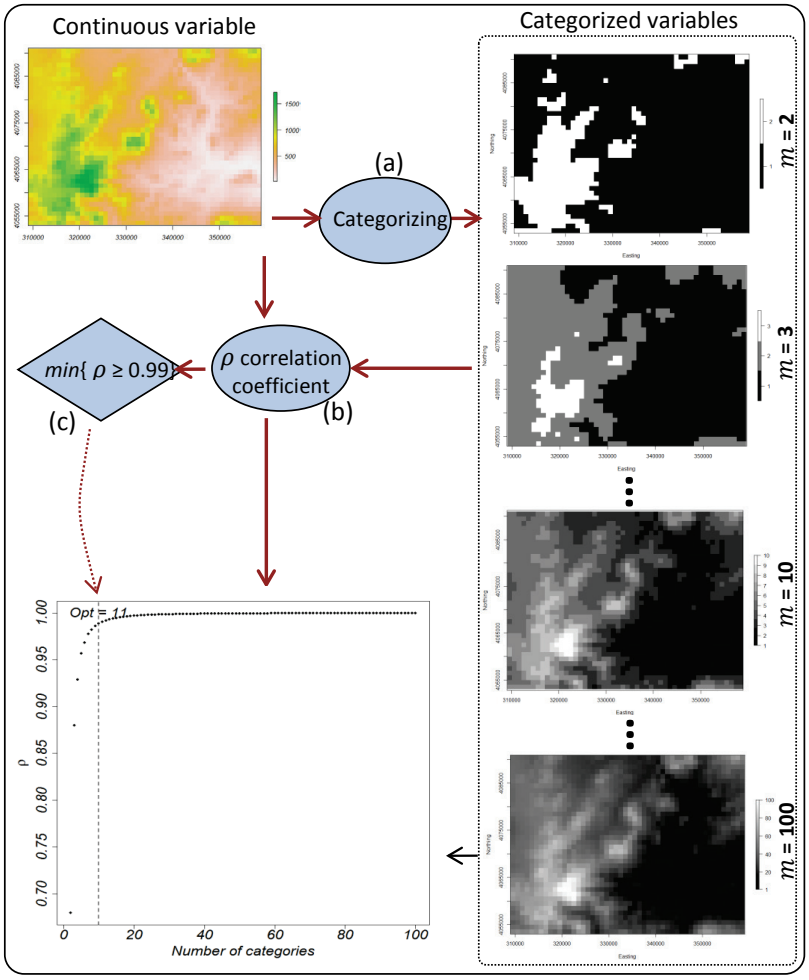



**Figure 4-1:** A flow diagram showing the procedure of finding the optimum number of categories for categorizing continuous spatial data; (a) an iterative categorization procedure taking different number of categories; (b) calculating the $\rho$ correlation coefficient between the continuous and each categorical variable; (c) taking the minimum number of categories for which $\rho$ is greater than 0.99 as the optimum number

To calculate ELSA for the continuous map using the Eq. 4-6, the original continuous data $\mathbf{x}$ are then mapped into the ranked categories: $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m)'$ with ranks $c_1, c_2, \dots, c_m$, where $c_1 = 1$ and $c_m = m$. The maximum level of dissimilarity in the entire map is therefore: $max(d) = c_m - 1$ and the dissimilarity between the categories at two locations $\mathbf{u_i}$ and $\mathbf{u_j}$ is $d_{ij} = |c_i - c_j|$.

### 4-4-2. ELSA for categorical data

Categorical variables are typically conceptualized as having no inherent ordering (Ahlqvist and Shortridge 2010), which means that all pairs of categories are equally dissimilar. When using this simplification of class differences and denoting the dissimilarity as $d$ (as in *Eq.* 4-6):

$max(d) = 1$, $d_{ij} = 1$ if $x_i \neq x_j$ and $d_{ij} = 0$ if $x_i = x_j$.

The assumption that all categories are equally dissimilar is often an oversimplification. Several studies have been conducted to estimate a measure of dissimilarity between categories (Romme 1982 ; Uuemaa et al. 2008). Categories may also be classified into a hierarchical structure. For example, the United Nations Food and Agricultural Organization's (FAO) land cover classification system arranges classes hierarchically (Di Gregorio and Jansen 2009). Such hierarchies can be used to describe the level of dissimilarity. Consider a categorical map with four categories: 'mixed forest' ($\alpha_1$), 'coniferous forest' ($\alpha_2$), 'olive groves' ($\alpha_3$) and 'vineyards' ($\alpha_4$). These can be grouped under primary categories such that 'mixed forest' and 'coniferous forest' belong to the primary category 'forests' (denoted $\beta_1$) and 'olive groves' and 'vineyards' belong to 'agricultural areas' (denoted $\beta_2$). To calculate $E_i$ for this categorical map, $c_i$ (the rank number of for site $i$) is set to 1 (always), and $c_j$ is set to 1 (if sites $i$ and $j$ are the same), or 2 (if the categories in sites $i$ and $j$ are different, but belong to the same primary category), or 3 (if the categories in sites $i$ and $j$ are different and also belong to different primary categories). Consequently, the level of dissimilarity, $d_{ij}$, between two different

subcategories of the same primary category is set as $d_{ij} = 1$ (*e.g.*, $d(\text{coniferous forest, mixed forest}) = d(\alpha_1, \alpha_2) = 1)$ , and between two different subcategories from different primary categories is set as $d_{ij} = 2$ (e.g., $d\big(\text{coniferous forest, vineyards}\big) = d(\alpha_1, \alpha_4) = 2$).

To develop this algorithm, we specify the dissimilarity as:

1) If $\alpha_i = \alpha_j$ then $c_i = 1$ & $c_j = 1$therefore $d_{ij} = 0$
2) If $(\alpha_i \neq \alpha_j)$ & $(\beta_i = \beta_j)$ then $c_i = 1$ & $c_j = 2$ therefore $d_{ij} = 1$
3) If $(\alpha_i \neq \alpha_j)$ & $(\beta_i \neq \beta_j)$ then $c_i = 1$ & $c_j = 3$ therefore $d_{ij} = 2$

The level of dissimilarity between pairs of categories can be illustrated in a matrix (Table 4-1). This makes ELSA flexible enough to handle situations where the level of dissimilarity can be specified for pairs of categories or when categories are hierarchically ordered.

**Table 4-1:** The level of dissimilarity between pairs of categories in an exemplified land use map with four (sub-)categories: 'mixed forest' ($\boldsymbol{\alpha_1}$), 'Coniferous forest' ($\boldsymbol{\alpha_2}$), 'olive groves' ($\boldsymbol{\alpha_3}$), 'vineyards' ($\boldsymbol{\alpha_4}$); the first two belong to the main category of 'For ests' and the last two are related to 'Agricultural areas'

| Categories | $\boldsymbol{\alpha_1}$ | $\boldsymbol{\alpha_2}$ | $\boldsymbol{\alpha_3}$ | $\boldsymbol{\alpha_4}$ |
|---|---|---|---|---|
| $\boldsymbol{\alpha_1}$ | 0 | 1 | 2 | 2 |
| $\boldsymbol{\alpha_1}$ | | 0 | 2 | 2 |
| $\boldsymbol{\alpha_1}$ | | | 0 | 1 |
| $\boldsymbol{\alpha_1}$ | | | | 0 |

The above situation can be extended to the case where there are more levels in the hierarchy. The maximum level of dissimilarity is then equal to the degree of hierarchy. In Fig. 4-2, the level of dissimilarity is presented schematically for a hierarchical system that contains 3 levels.

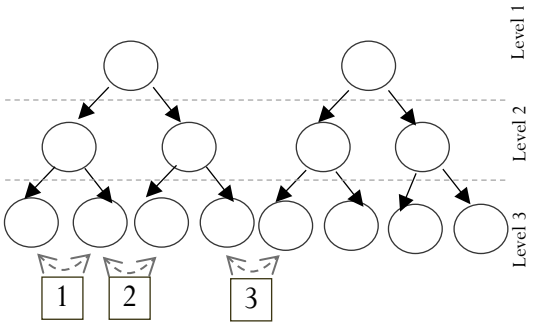



**Figure 4-2:** A hierarchical way of presenting the classes in an exemplified categorical map with 3 levels of categories; the numbers in the boxes indicate d (dissimilarity) of the relevant pairs of classes

## *4-5. Application of ELSA to assess local spatial association*

In this section, we illustrate the use of ELSA by means of several examples using real and synthetic data sets, covering both continuous and categorical variables.

### 4-5-1. Experiment with continuous data

A digital elevation model (DEM) from southern Spain was used as an example to assess the application of ELSA for continuous variables. For comparison, we also quantified other commonly used local indicators of spatial association; $G_i$ and $G_i^*$, local Moran's $I$ and local Geary's $c$. These statistics were compared with ELSA to explore to what extent these measures are related. A Spearman statistic was used to test and estimate a rank-based measure of association test between ELSA and each local statistic.

The DEM variable is a raster layer including 2769 grid cells of 1 x 1 km (39 columns x 71 rows). In this experiment, the LISAs were calculated at each cell within a local distance of 5 km.

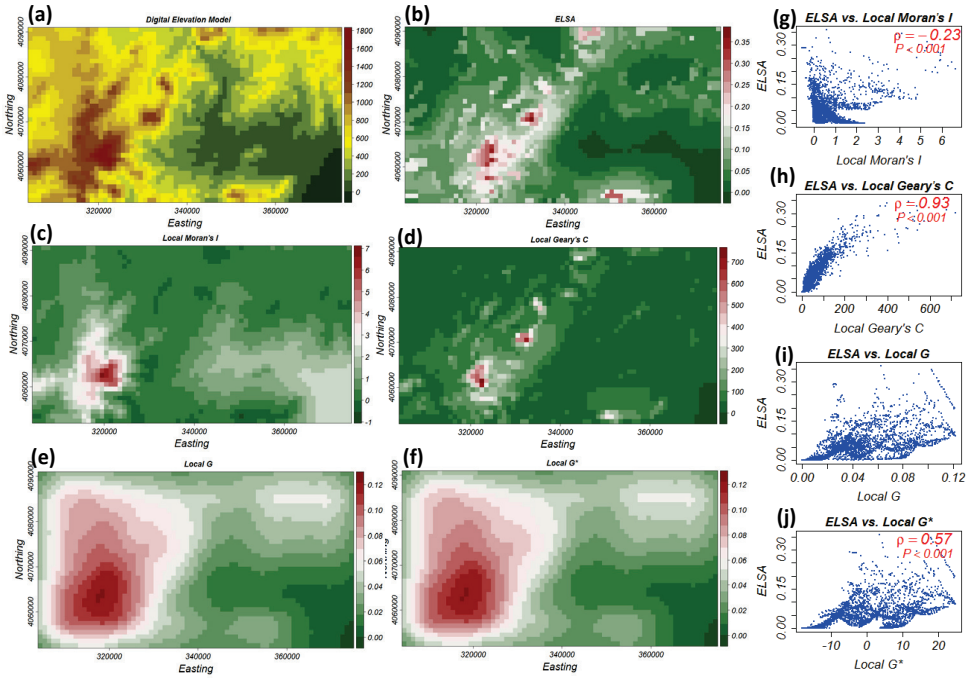



**Figure 4-3:** Local indicators of spatial association for a digital elevation model in southern Spain (a), calculated with ELSA (b), local Moran's $\boldsymbol{I}$ (c), local Geary's $\boldsymbol{c}$ (d), $\boldsymbol{G_i}$ (e) and $\boldsymbol{G_i^*}$ (f); scatter plots between ELSA on y axis and local Geary's $c$ (g), local Moran's $\boldsymbol{I}$ (h), $\boldsymbol{G_i}$ (i) and $\boldsymbol{G_i^*}$ (j) statistics on x axis

The results from the scatter plots between ELSA and other local statistics as well as their corresponding correlation tests (Fig. 4-3-g, h, i, and j) indicate that the correlation coefficient ranging from -0.22 to 0.92. From Fig. 4-3, it can be seen that ELSA has a strong relationship with the local Geary's $c$ statistic, although there is evidence of non-linearity in the relationship.

### 4-5-2. Experiment with categorical data with the same level of dissimilarity between classes

A land cover map in a raster layer including 2769 grid cells of 1 x 1 km (same as the DEM variable in the previous experiment) from southern Spain was used for this experiment (Fig. 4-4-a). The map consists of 6 land cover classes. It is assumed that the level of dissimilarity between pairs of classes is equal. In this experiment, the ELSA was calculated at each cell

within a local distance of 5 km (Fig. 4-4-b). As it was expected, the ELSA statistic becomes 0 for homogenous areas with only one class (mostly at the western part of the area), and when the area becomes more heterogeneous, the ELSA value becomes higher as it is obvious at centre and eastern parts of the area.

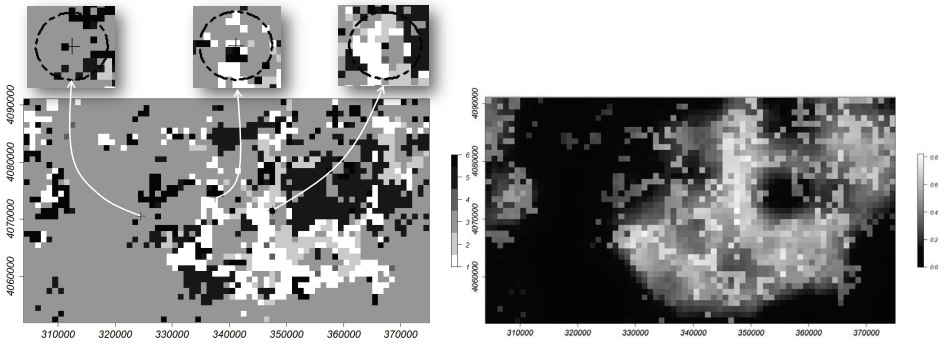



**Figure 4-4:** ELSA for land cover data in the south of Spain; (a) land cover map including six classes with the same level of dissimilarity between pairs, three cells within their five km neighbourhoods are specified as A, B, C; (b) ELSA statistic for the land cover map, the values of ELSA for the three cells are represented on top

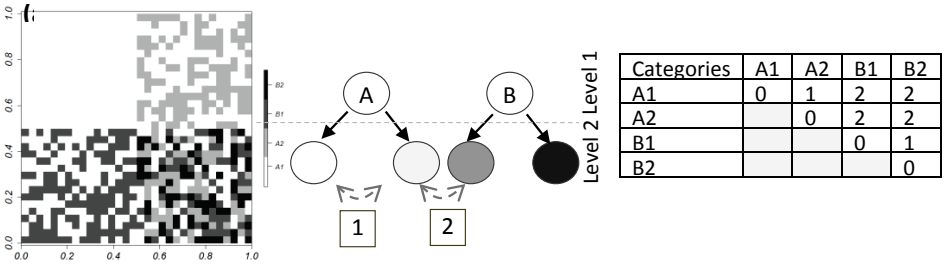
The three specified locations in Fig. 4-4 (*i.e.*, A, B, and C) and the ELSA values at these locations show how this statistic changes when the landscape changes. By looking at these locations on the land cover map, it can be recognized that the level of heterogeneity changes from low to high from the location of A to the location of C. The calculated values for the ELSA statistics at these locations range from 0.102 to 0.823, indicating that the level of spatial association at these locations change from high to low, respectively.

### 4-5-3. Experiment with categorical data with non-equal level of dissimilarity between classes

There are numerous situations where the level of dissimilarity between pairs of classes in a categorical variable is not equal. To show how to deal

with these situations, we illustrate two experiments using both synthetic and real data.

- *Synthetic data*: We generated a synthetic raster categorical map consisting of 1024 grid cells (32 rows × 32 columns) and four classes (*i.e.*, A1, A2, B1, and B2). The first and second two classes belong to the main categories of A and B, respectively. So, A1 is more similar to A2 than to B1 or B2. The level of dissimilarity between different pairs from the same main category (*i.e.*, A1-A2 or B1-B2) is set to 1, and from different main categories (*e.g.*, A1-B1, A1-B2, etc.) is set to 2 (Fig. 4-5).




| Categories | A1 | A2 | B1 | B2 |
|---|---|---|---|---|
| A1 | 0 | 1 | 2 | 2 |
| A2 | | 0 | 2 | 2 |
| B1 | | | 0 | 1 |
| B2 | | | | 0 |

**Figure 4-5:** Synthetic land cover map including four classes which distributions are controlled into four equal zones (a); level of dissimilarity between pairs of classes in a hierarchical view (b) and table view (c)

The region is divided into four zones with different combinations of categories. This shows how ELSA changes under these controlled situations. Zone 1 includes only one class (*i.e.*, A1), Zone 2 includes a random distribution of A1 and A2 (*i.e.*, from the same primary category), Zone 3 includes a random distribution of A1 and B1 (*i.e.*, from two different primary categories), and Zone 4 includes random distribution of all four categories.

The maximum level of dissimilarity ($max\{d\}$) for this experiment is 2. We calculated ELSA at each grid cell using a 3 × 3 window as the local neighbourhood (including diagonal neighbours, *i.e.*, queen's case). We also calculated the mean ELSA for each zone to provide a base for comparison (Fig. 4-6). The results show that the two extreme zones have the minimum (Zone 1) and maximum (Zone 4) mean ELSA. The other two zones both

follow the same structure (a random distribution of two classes, but with a different attribute distances). Since Zone 2 consists of two more similar classes than Zone 3, it is expected that the ELSA statistic for the Zone 2 should be less than Zone 3. This is backed up by the empirical results (Fig. 4-6-b).

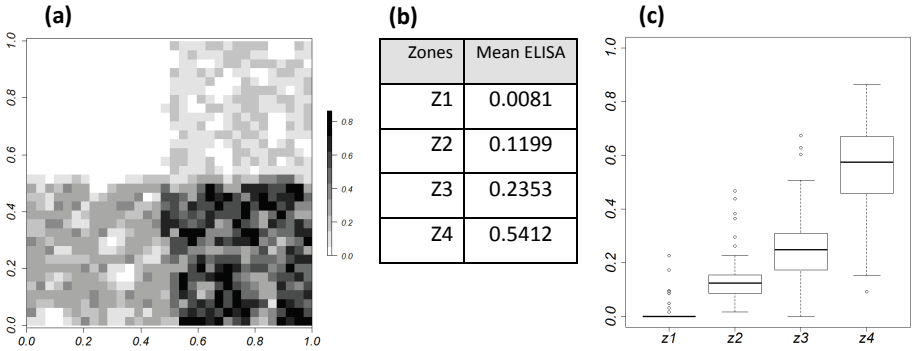



**Figure 4-6:** The ELSA map (a) and the Mean ELSA statistic in four zones of the region (b); the region is divided into four zones including Z1 (upper-left), Z2 (upper-right), Z3 (lower-right) and Z4 (lower-left); the boxplot (c) represents the distribution of ELSA values at grid cells over different zones

- *Real data*: We used the CORINE [Coordination of Information on the Environment of the European Environmental Agency (EEA 2007)] 2006 land cover map from central Spain including 392336 grid cells of 250 x 250 m. The land cover classes in the map were described using a hierarchical scheme with three levels. The first level indicates the primary land cover class (*e.g.*, agricultural area), which are subdivided to more specific types of land cover classes at the second and third levels (*e.g.*, permanent crops and vineyards at the second and third levels, respectively). A three-digit code is used for each land cover, specifying the class at the three levels from left to right (*e.g.*, 221 and 223 are 'vineyards' and 'olive groves' respectively, *i.e.*, class 1 and 3 specified at the third level, respectively, but both belong to the same classes of 'agricultural areas' [class 2] at level 1, and 'Permanent crops' [class 2] at level 2). A list of the classes at the three levels for the codes is provided in Table 4-2.

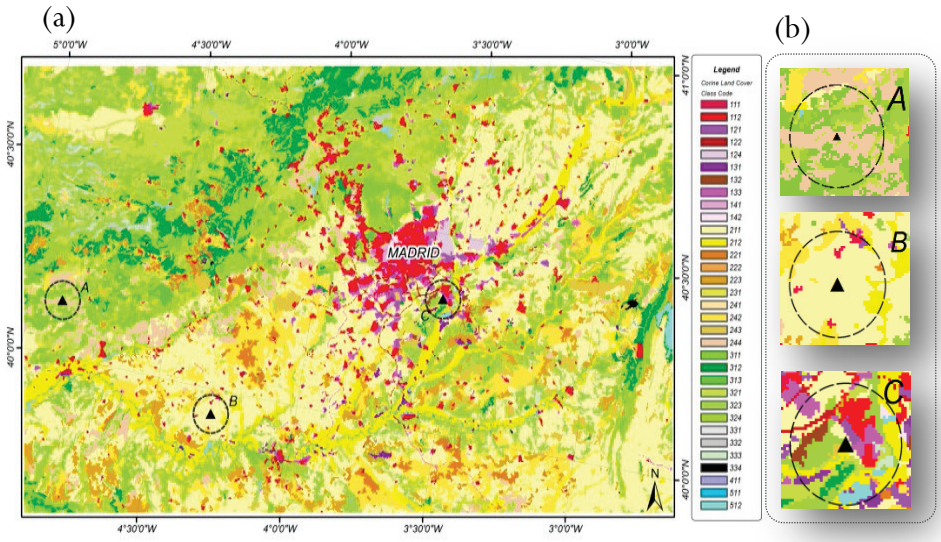



**Figure 4-7** CORINE Land cover map from the central Spain; a three-digit code is used to define each land cover class (a); three randomly selected points around which the circle specifies a 5 km of their neighbours (b)

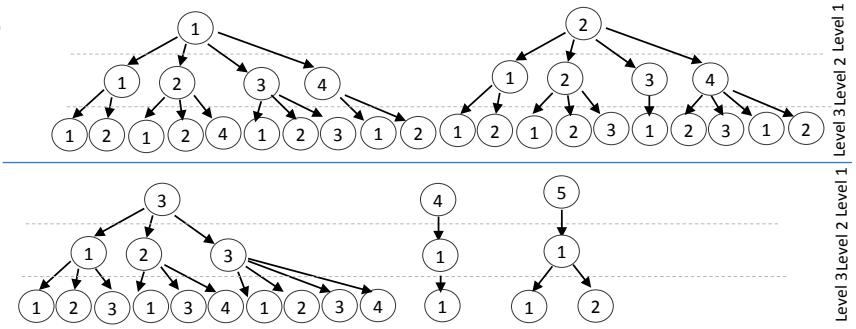
**Table 4-2:** The CORINE land cover class definitions

| Corine | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| 111 | Artificial surfaces | Urban fabric | Continuous urban fabric |
| 112 | Artificial surfaces | Urban fabric | Discontinuous urban fabric |
| 121 | Artificial surfaces | Industrial, commercial and transport | Industrial or commercial |
| 122 | Artificial surfaces | Industrial, commercial and transport | Road and rail networks |
| 124 | Artificial surfaces | Industrial, commercial and transport | Airports |
| 131 | Artificial surfaces | Mine, dump and construction sites | Mineral extraction sites |
| 132 | Artificial surfaces | Mine, dump and construction sites | Dump sites |
| 133 | Artificial surfaces | Mine, dump and construction sites | Construction sites |
| 141 | Artificial surfaces | Artificial, non-agricultural vegetated | Green urban areas |
| 142 | Artificial surfaces | Artificial, non-agricultural vegetated | Sport and leisure facilities |
| 211 | Agricultural areas | Arable land | Non-irrigated arable land |
| 212 | Agricultural areas | Arable land | Permanently irrigated land |
| 221 | Agricultural areas | Permanent crops | Vineyards |
| 222 | Agricultural areas | Permanent crops | Fruit trees and berry plantations |
| 223 | Agricultural areas | Permanent crops | Olive groves |
| 231 | Agricultural areas | Pastures | Pastures |
| 241 | Agricultural areas | Heterogeneous agricultural areas | Annual crops/permanent crops |
| 242 | Agricultural areas | Heterogeneous agricultural areas | Complex cultivation patterns |
| 243 | Agricultural areas | Heterogeneous agricultural areas | Land occupied by agriculture |
| 244 | Agricultural areas | Heterogeneous agricultural areas | Agro-forestry areas |
| 311 | Forest and semi natural | Forests | Broad-leaved forest |
| 312 | Forest and semi natural | Forests | Coniferous forest |
| 313 | Forest and semi natural | Forests | Mixed forest |
| 321 | Forest and semi natural | Scrub and/or herbaceous vegetation | Natural grasslands |
| 323 | Forest and semi natural | Scrub and/or herbaceous vegetation | Sclerophyllous vegetation |
| 324 | Forest and semi natural | Scrub and/or herbaceous vegetation | Transitional woodland-shrub |
| 331 | Forest and semi natural | Open spaces with little/no vegetation | Beaches, dunes, sands |
| 332 | Forest and semi natural | Open spaces with little/no vegetation | Bare rocks |
| 333 | Forest and semi natural | Open spaces with little/no vegetation | Sparsely vegetated areas |
| 334 | Forest and semi natural | Open spaces with little/no vegetation | Burnt areas |
| 411 | Wetlands | Inland wetlands | Inland marshes |
| 511 | Water bodies | Inland waters | Water courses |
| 512 | Water bodies | Inland waters | Water bodies |

The categories were hierarchically ordered at the three levels based on the three-digit codes. The attribute distance (level of dissimilarity) between each pair of categories was calculated based on their position on the hierarchical scheme. The maximum level of dissimilarity is 3 (*e.g.*, between

class 132 and 211). The dissimilarity between pairs of classes is illustrated in Fig. 4.9.

(a)




(b)

| | 111 | 112 | 121 | 122 | 124 | 131 | 132 | 133 | 141 | 142 | 211 | 212 | 221 | 222 | 223 | 231 | 241 | 242 | 243 | 244 | 311 | 312 | 313 | 321 | 323 | 324 | 331 | 332 | 333 | 334 | 411 | 511 | 512 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 111 | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 112 | | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 121 | | | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 122 | | | | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 124 | | | | | 0 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 131 | | | | | | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 132 | | | | | | | 0 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 133 | | | | | | | | 0 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 141 | | | | | | | | | 0 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 142 | | | | | | | | | | 0 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 211 | | | | | | | | | | | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 212 | | | | | | | | | | | | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 221 | | | | | | | | | | | | | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 222 | | | | | | | | | | | | | | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 223 | | | | | | | | | | | | | | | 0 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 231 | | | | | | | | | | | | | | | | 0 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 241 | | | | | | | | | | | | | | | | | 0 | 1 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 242 | | | | | | | | | | | | | | | | | | 0 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 243 | | | | | | | | | | | | | | | | | | | 0 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 244 | | | | | | | | | | | | | | | | | | | | 0 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 311 | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 312 | | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 313 | | | | | | | | | | | | | | | | | | | | | | | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 321 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 323 | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 324 | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 331 | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 1 | 1 | 3 | 3 | 3 |
| 332 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 1 | 3 | 3 | 3 |
| 333 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 1 | 3 | 3 | 3 |
| 334 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 3 | 3 | 3 |
| 411 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 3 | 3 |
| 511 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | 1 |
| 512 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 |

**Figure 4-8:** Hierarchical scheme of different classes (a) and the level of dissimilarity between pairs of classes (b) in the CORINE land cover map

We calculated ELSA at each grid cell within a local distance of 5 km (Fig. 4-9). A visual comparison of the three specified locations on the land cover map (Fig. 4-7) show that the local association among the three locations is expected to be minimum and maximum at the location B and C, respectively. The values of the ELSA map at these locations are consistent with the visual interpretation.

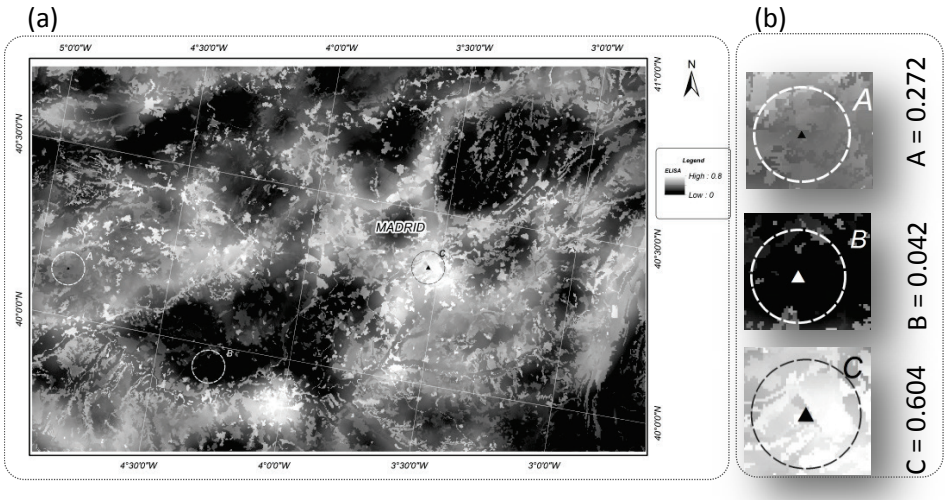



**Figure 4-9:** ELSA map calculated based on the CORINE land cover map in Fig. 4-7 (a); the ELSA value at the three specified locations (b)

## *4-6. Application of ELSA to assess global spatial structure*

In this section we explore if global spatial structure can be assessed by employing ELSA into a procedure of generating a sample variogram-like diagram, called 'entrogram'. The idea is that by assessing ELSA with increasing local distances (lags) and putting the averaged values against these distances in a diagram, we can explore spatial structure in the field. The sample variogram is a well-known approach for exploring spatial structure in continuous variables or binary categorical variables (*i.e.* indicator variogram) (Journel 1983). The semantic sample variogram has been developed for multinomial categorical variables (Ahlqvist and Shortridge 2006). If ELSA can be used for this purpose, then the spatial

structure for both continuous and categorical variables can be explored with the same technique. The entrogram for lag distance $h$ is calculated as the average of the ELSA statistics at different sites within the distance equal to the lag size. The entrogram is calculated as follows:

$$E(h) = \frac{\sum_{i=1}^{n_h} E_i(h)}{n_h} \qquad (4\text{-}5)$$

where $E(h)$ is the value of the entrogram for distance class $h$, $E_i(h)$ is the ELSA statistic at site $i$ within local distance $h$, $n_h$ is the total number of sites within the distance $h$ for which the ELSA statistic is calculated.

In the following experiments we explored the behaviour of entrogram for both continuous and categorical data.

### 4-6-1. Entrogram for continuous data

We generated a series of continuous variables with different ranges of spatial autocorrelation. We used an unconditional simulation to construct regular grids of $200 \times 200$ cells for each variable. Unconditional simulation is a geostatistical technique that generates a realization of a spatially correlated variable, where the spatial correlation is defined by a variogram (Dungan 1999). The circulant-embedding algorithm (Dietrich and Newsam 1993) implemented in the RandomFields package v. 1.3.41 (Schlather 2009) in the R programming environment was used to conduct the unconditional simulation. An exponential variogram model with an arbitrary sill of 10 and a nugget of 0 was used for all datasets. The exponential variogram model is parameterized by a scale parameter, $\phi$, which controls the range of spatial autocorrelation (the range is the maximum lag separation where two points are expected to be correlated). Under the exponential variogram the range is approximated as $3\phi$. We assigned $\phi$ different values to control the range of spatial autocorrelation. Six levels of $\phi = 1$, 5, 10, 15, 20, and 25 grid cells were used, giving a transition from minimum spatial autocorrelation ($\phi = 1$) to relatively

large-scale spatial autocorrelation ($\phi = 25$). Additionally, a white-noise surface ($\phi = 0$) was simulated, giving a total of 7 variables with different levels of spatial autocorrelation. We then quantified and visualized the empirical variogram and entrogram for each surface (Fig. 4-10). The visual comparisons of the graphs showed that the entrogram visualizes a spatial structure similar to the variogram.

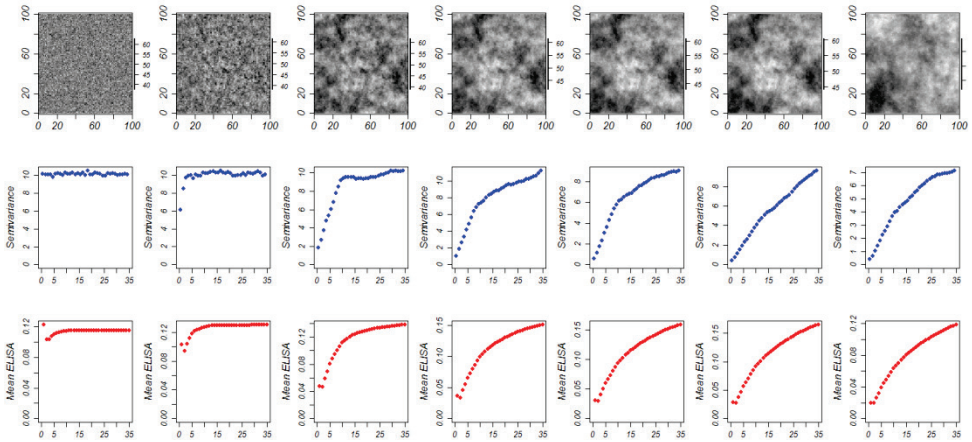



**Figure 4-10:** Comparing variogram and entrogram; the first row displays the 7 simulated continuous fields with different levels of spatial autocorrelation ($\phi = 0$, 1, 5, 10, 15, 20, and 25 from left to right), the second row displays the corresponding variograms and the third row displays the entrograms

### 4-6-2. Entrogram for categorical data

We generated four synthetic categorical maps, on a 20 x 20 raster grid, to explore the capability of the entrogram for calculating the spatial structure of categorical maps. The first two maps are binary, where classes are spatially structured with a maximum degree of spatial clustering, and are randomly distributed in the first and second map, respectively. These binary maps provide the opportunity to compare the entrogram with the existing method of exploring the spatial structure for binary categorical variable (*i.e.*, the indicator variogram). We quantified both the entrogram, using our developed R package (*i.e.*, ELSA), and the indicator variogram using the gstat package v. 1.0-18 (Pebesma 2004) in the R developing

environment (R Development Core Team 2013). For both, we used a lag size equal to one grid cell and the cutoff values (number of lags) equal to 20 grid cells. The graphs are then visually interpreted (Fig. 4-11).

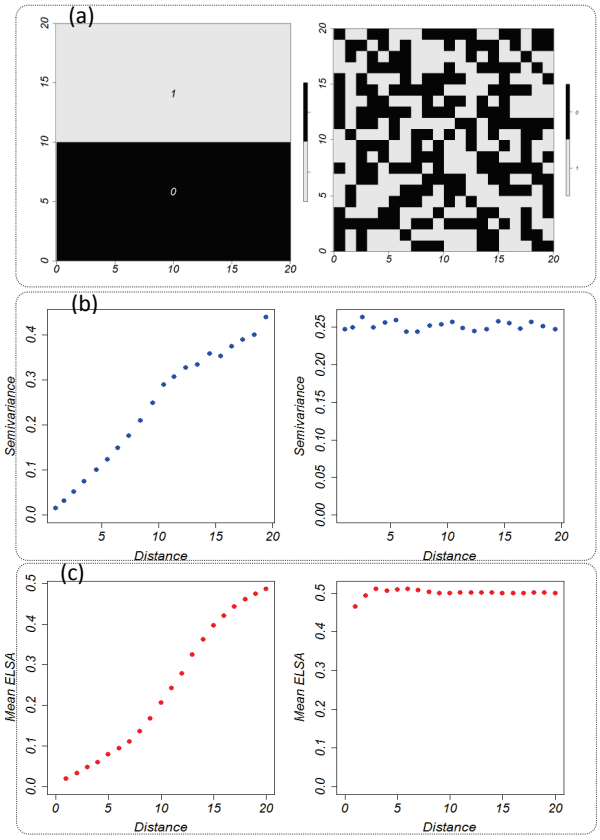



**Figure 4-11:** Comparing variogram and entrogram for two binary categorical maps (a); (b) and (c) represent the corresponding variograms and entrograms, respectively

The second two categorical maps include four classes. The classes in the first map are spatially structured, giving a maximum degree of spatial clustering, while in the second map they are randomly distributed. This dataset provides the opportunity to illustrate the capability of the entrogram for exploring the spatial structure of multinomial data. We quantified the entrogram with a lag distance equal to one grid cell and the cutoff value (number of lags) equal to 20 grid cells. The visual

interpretation of the graphs (Fig. 4-12) for the spatially clustered map shows that the mean ELSA value is low within the lower distances and is increased when the lag distance is increased (as it was expected). For the randomly distributed classes, on the other hand, this value remains at the maximum level which show there is no spatial structure at the field.

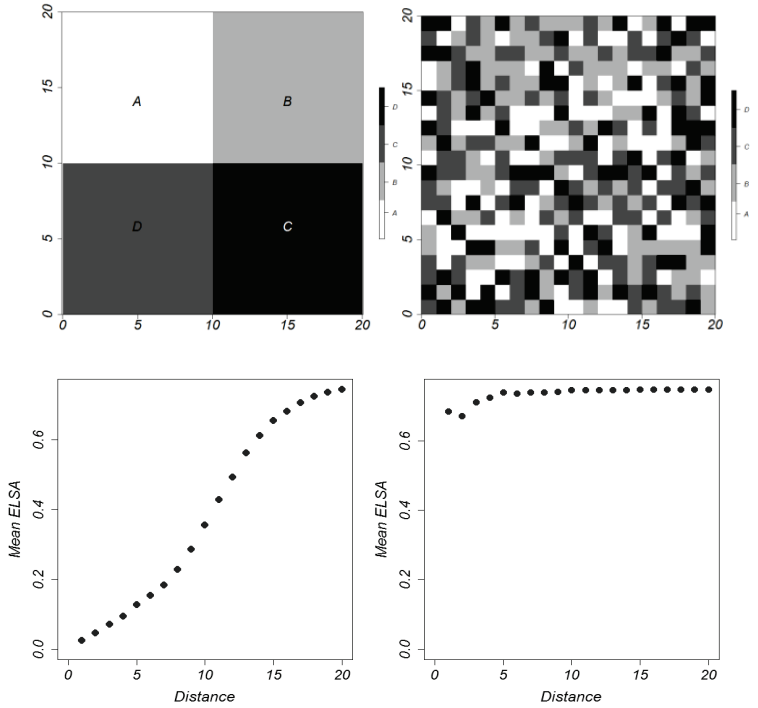



**Figure 4-12:** Two categorical maps including four classes (first row) and their corresponding entrograms (second row)

## *4-7. Discussion*

ELSA allows for exploring the local spatial association for both continuous and categorical variables. This provides the opportunity of using one statistic for a study where both types of variables are used. There are many situations in which both types of variables need to be evaluated for their spatial structure. For instance, Naimi et al. (2011, 2014) tested whether spatial association in environmental layers, which were used to predict recorded occurrences of species, can be used to understand the effect of

positional uncertainty of these recordings on the accuracy of prediction. Those studies were limited to only using continuous environmental variables while categorical variables also are widely used in this kind of modelling. The ELSA statistic measures the local spatial associations within the same range (between 0 and 1) for both types of data, making the outputs are comparable.

The ELSA calculation for categorical data supports different levels of dissimilarities. This allows evaluation of the dissimilarity between nominal categories in a graded fashion. Several authors in the field of landscape ecology have considered graded differences for measuring landscape patchiness (Degraaf and Yamasaki 2002 ; Desrochers et al. 2003). Use of this approach allows transformation from a nominal to an ordered or numerical scale, and provides a foundation for handling attribute distances for categorical data. This is important when quantifying spatial structure for categorical data, since without this aspect all categories are considered equally dissimilar. This approach, however, relies on the subjective evaluation of the class similarity (Ahlqvist and Shortridge 2010), and requires additional data and a guiding theory. Multi-criteria decision making, the analytic hierarchy process, and conjoint analysis are well-known frameworks for evaluating class similarities in categorical maps (Ahlqvist and Shortridge 2010 ; Schwering 2008). In this study, we used and introduced a more general and simple rule to define the class dissimilarities based on a hierarchical scheme of classes. Given additional knowledge about the meaning of class definitions, a more formal evaluation of the categorical dissimilarities can be used to calculate the ELSA statistic for categorical maps.

To measure spatial association, a majority of studies were devoted to the development of statistics for continuous data. In this study, we addressed four of these commonly used statistics and explored how ELSA related to them. Although all of these statistics have been widely used as measures of local spatial associations, they measure different properties. Therefore, the appropriate technique for identification of the spatial association should

correspond to the nature of the question concerning dependence/independence (Getis and Ord 1996). Our results confirmed these differences and revealed that the ELSA is more related to local Geary's *c*. Local Geary's *c* (like in a variogram) is based on differences between pairs of observations. Analogous to variance, entropy is a measure of dissimilarity and diversity and their equivalence has been explored and discussed in several studies e.g. (Ebrahimi et al. 1999 ; Lindley 1956).

It should be noted that ELSA is not a LISA in the terminology defined in Anselin (1995), since its individual components are not related to a global statistic of spatial association. Therefore, ELSA cannot be used as a diagnostic of local instability in the presence of global spatial association. However, this provides the advantage that ELSA is not sensitive to the presence of global spatial association while the LISA statistics are (Getis and Ord 1996).

Despite the early work in geographical analysis (Batty 1974), entropy has been mainly used in the fields of physics and information theory. The most relevant works in these disciplines introduced several extensions of the entropy measure that apply to calculating the structural complexity or patterns of two dimensional dynamical systems (Feldman and Crutchfield 2003 ; Robinson et al. 2011). In recent years, there were several efforts to develop some entropy-based methods for detecting (global) spatial association and patterns of complexity for univariate data (Matilla-García and Marín 2011 ; Matilla-García et al. 2012 ; Pham 2010 ; Ruiz et al. 2010) or to discover different forms of local multivariate relationships (Guo 2010). Even though ELSA uses the entropy measure (compared to the previous contributions), it can be considered as a novel approach that offers some unique features (as described in the manuscript). The method, however, can currently be used only for exploratory purposes and cannot be used for inferential spatial data analysis. For example, it cannot be tested if a location is significantly clustered using a generated p-value. Further study is required to develop an appropriate diagnostic statistical test for

ELSA. Relevant research exists that may provide guidance on this (Guo 2010 ; Matilla-García et al. 2012 ; Matilla-García and Ruiz Marín 2008).

Together with ELSA, this chapter introduced the entrogram, an approach for exploring the global spatial structure within the entire area (like a variogram). Such explorations are applied in many fields, such as landscape ecology, geography, and soil science. The entrogram uses the ELSA, and therefore, can be used for both continuous and categorical data. The indicator variogram is a known technique to measure the spatial variability of classes in categorical variables. However, it has been argued that the binary treatment of categorical variables in this technique is an unnecessary oversimplification, and that it should be replaced by ordered measures based on semantic similarity evaluations (Ahlqvist and Shortridge 2010). Semantic variograms (Ahlqvist and Shortridge 2006) were developed based on this concern and provide the capability to consider semantic distance between categories in the calculation. We showed that the entrogram is capable of this as well.

## *4-8. Conclusions*

This paper has focused on the development of an entropy-based statistic (ELSA) for the quantification of local spatial association. The ELSA statistic presented in this paper showed to be a robust and reliable method for exploratory spatial data analysis. The method provides the advantage of using one statistic for both continuous and categorical data to measure the degree of spatial association. This makes comparisons between spatial structures of both types of data possible. Also, this statistic provides the ability to incorporate both spatial and attribute aspects of spatial association into the statistic for both continuous and categorical data. By introducing the 'entrogram' we demonstrated that ELSA can also be used to measure global spatial structure of both continuous and categorical data.

# *Chapter 5*

# Model Uncertainty in Species Distribution Modelling

*This chapter is based on:*

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K., Toxopeus, A. G., Araujo, B. A. (in preparation) From geographical distributions to ecological niches and back.

# 5. Model Uncertainty in Species Distribution Modelling

## *5-1. Introduction*

In the pursuit of more robust species distribution models (SDMs) researchers have gradually constructed more complicated approaches. Challenges with available data (*e.g.,* incomplete records, positional errors), and with models (*e.g.*, their sensitivity to spatial autocorrelation) as well as complexities in species-environment relationships inspired needs to develop new approaches for the modelling of species distribution. Earliest efforts used some simple envelope methods to define a hyper-rectangle that bounds species observations in a multidimensional environmental space which describes a species' range in relation to its' environment (Box 1981). In recent years, focus on machine learning approaches resulted in developing a variety of techniques that are expected to model more complex non-linear species-environment relationships (Elith and Graham 2009). These techniques vary in the statistical formulation they use, and may differ in their ability to summarize useful relationships between response and predictor variables. This causes an important source of uncertainty (*i.e.*, model uncertainty), meaning they may yield different results, even when calibrated with the same response and predictor variables (Araújo and Guisan 2006 ; Pearson et al. 2006). This wide array of methods and still existing challenges imply that knowledge is required about models' behaviour and performance. The criteria and advice that should enable an informed choice of methods are currently scattered throughout the literature, and are incomplete (Elith and Graham 2009).

Comparison between models, trying to identify the most appropriate method(s) for modelling the response variable and find the best predictions, have been evaluated and discussed in several studies (Brotons et al. 2004 ; Elith et al. 2006 ; Manel et al. 1999 ; Meynard and Quinn 2007 ; Munoz and Felicisimo 2004 ; Ortega-Huerta and Peterson 2008 ;

Segurado and Araújo 2004). These studies usually count on discriminatory metrics to evaluate the SDMs which provide a global measure (*i.e.*, a single summary measure for the entire area) to show how well model predictions discriminate occupied and unoccupied sites in an evaluation dataset. Among these metrics, the area under the curve (AUC) of a receiver operating characteristic (ROC) plot, is a widely used statistic, largely because it gives a threshold and prevalence (proportion of presence locations) independent metric. This statistic has been questioned in recent years because (Lobo et al. 2008): (i) it ignores the probability values; (ii) it includes the regions of the ROC space into the summary statistic in which one would rarely operate; (iii) it weights omission and commission errors equally; (iv)it does not give information about the spatial distribution of errors; and (v) the total extent to which the models are applied highly influences the rate of well-predicted absences and AUC scores. These reasons may also be valid for the other commonly used statistics to evaluate SDMs. Moreover, relying on a single and summary measure to assess a model performance may lead to misleading conclusions (Jiménez-Valverde et al. 2013). Using niche similarity metrics (Warren et al. 2008) may be broadly useful to evaluate SDMs and compare models when the true niche and suitability of a habitat is known (Warren and Seifert 2011). Although the need to develop robust methods for assessing the quality of SDMs are acknowledged (Araújo and Guisan 2006 ; Hijmans 2012 ; Jiménez-Valverde 2012), the progress toward adoption of a comprehensive toolbox of evaluation metrics is slow (Elith and Leathwick 2009).

In this chapter, we compared several models using a comprehensive set of evaluation metrics. These metrics cover different aspects of a model performance including discrimination capacity in geographical space, similarity of the predicted and the true niches in environmental space, and consistency/inconsistency between different models. This chapter seeks to understand models' behaviour and accuracy. This can be used to more precisely identify the most appropriate models and understand model uncertainty in SDMs. Moreover, this chapter provides insight into whether

the commonly used evaluation metrics can consistently quantify the performance of SDMs. We used a series of synthetic data (in which the 'true' environmental niche is known) to avoid effects arising from different levels of unknown complexity existing in real data.

## *5-2. Materials and methods*

### 5-2-1. Generating virtual species

We simulated 10 virtual species and mapped them onto Spain. The virtual species respond to two variables: normalized difference vegetation index (NDVI), obtained from the moderate resolution imaging spectroradiometer (MODIS) satellite image archive (Nasa Land Processes Distributed Active Archive Center 2011), and precipitation seasonality, obtained from bioclimatic variable (Hijmans et al. 2008), both with a spatial resolution of 10×10 km. We assume that the species are in equilibrium with the environment. A number of functional links between species and environment (*i.e.*, species response curve) were a-priori determined to simulate species with different niche characteristics (shape). Species response may take a wide variety of shapes including a straight line, symmetrical bell shaped or a skewed unimodal shape (Austin 2002). These shapes represent the probability of presence along an environmental gradient. The response curves were constructed with Gaussian, linear and beta functions. These functions were used to create an individual measure of habitat suitability for each environmental variables or interaction of the two variables. For each species, the final habitat suitability, which is considered as the probability of occurrence, was generated by averaging or multiplying the habitat suitabilities for the both environmental variables (Fig. 5-1).

The habitat suitability maps were used to draw a sample (size=100) including presence-absence realizations to train the SDMs for each species. This sample size was chosen to avoid the effects of a very few or a very large sizes on the behaviour of the models. It has been shown that some models perform better with a few species observations (Pearson et al.

2007). It has been shown that most SDMs reach near maximum or maximum accuracy at size>50 or size=100 (Hernandez et al. 2006 ; Stockwell and Peterson 2002). For this purpose, we used a sampling scheme with a random uniform distribution over space. The habitat suitability value in each grid cell was then used as the success rate for each sample point to contain the species (Elith and Graham 2009 ; Naimi et al. 2014). For example a cell with a suitability of 0.7 has a 70% probability of being occupied by the species. For each species, 100 realizations of presence-absence points were simulated.

### 5-2-2. Species distribution modelling

To develop SDMs, several commonly implemented models that use presence/absence or presence-only records of species occurrences were selected. The presence-absence models were generalized linear models (GLM; McCullugh & Nelder, 1989), generalized additive models (GAM; Hastie & Tibshirani, 1990), boosted regression trees (BRT; Friedman, 2001), random forests (RF; Breiman, 2001), and two neural network algorithms inclusing multi-layer perceptron (nnetMLP; (Rosenblatt 1958)) and radial basic function (nnetRBF) (Hudak 1992); and the presence-only models were maximum entropy (Maxent; Phillips et al., 2006), maxlike (Royle et al. 2012) and profile based methods including Bioclim (Busby 1991), Domain (Carpenter et al. 1993), and Mahalanobis (Farber and Kadmon 2003). We also employed an ensemble (multi-model or consensus) approach (Araújo and New 2007) which is based on combining the predictions of several single models. There are various methods used in the ensemble approach to combine the predictions of single models (for a review see Garcia et al., 2012). In this study we used one that combines the predicted values using a weighted averaging procedure (Araújo and New 2007 ; Garcia et al. 2012). This method takes a weighted average over the predictions by the single models for which the predictive performance measure (*i.e.*, TSS) is considered as a weight for each single model. We separately used this method to combine the predictions by the

presence-absence models (hereafter EnsemblePA), presence-only models (hereafter EnsemblePO) and all models (hereafter Ensemble).

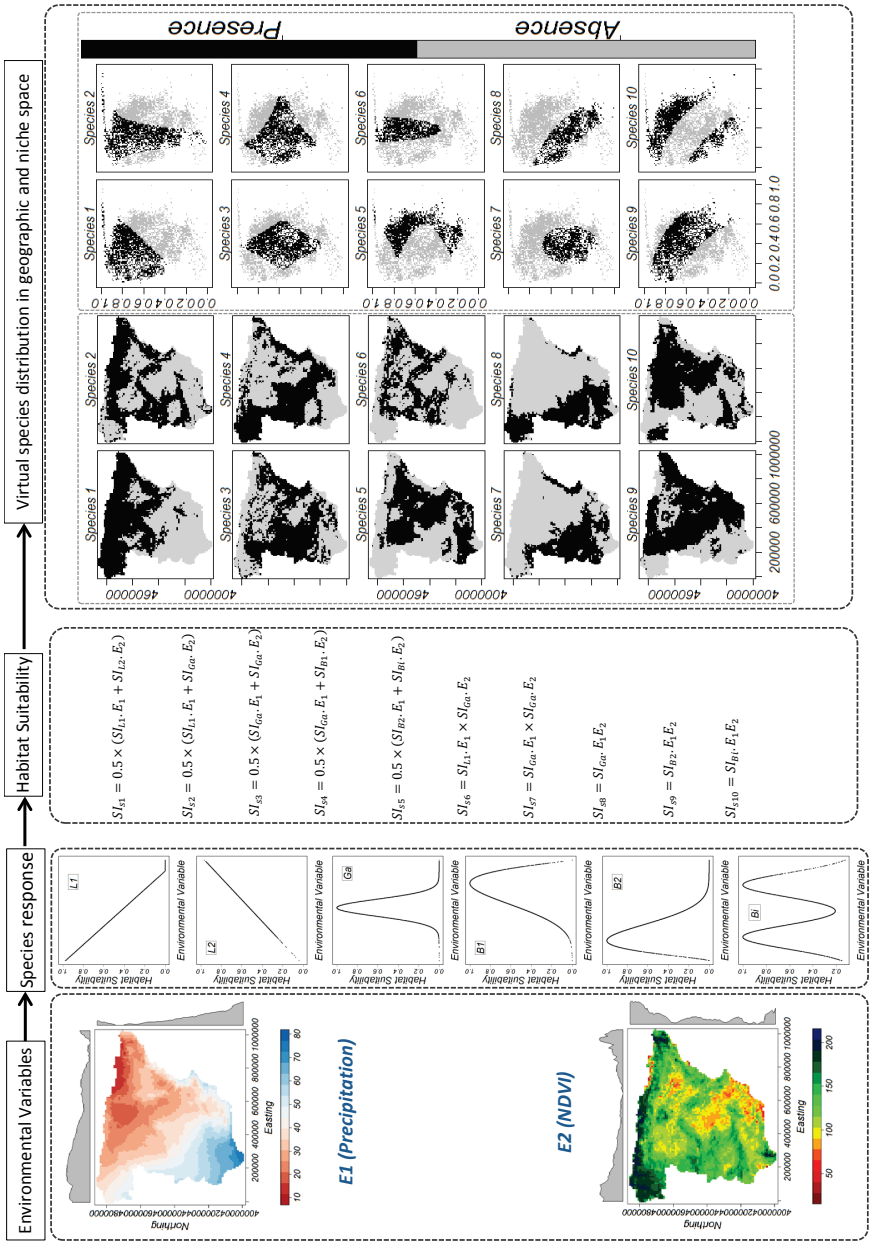



**Figure 5-1:** Flow diagram of generating virtual species

The GLM, GAM, BRT, SVM, RF, nnetMLP, nnetRBF, maxlike, profile based and ensemble models were implemented in the R development environment v. 3.0.2 (R Development Core Team 2013). Maxent was run by using the Maxent software v. 3.3.1 that was developed and introduced by Philips et al. (2006). The details of the models are summarized in chapter 2 and 3, and the cited references.

### 5-2-3. Model evaluation

We used a variety of metrics to evaluate the performance of the SDMs globally and also at grid cell (pixel) level in the both geographic and environmental space. These metrics quantified the discrimination capacity of the models in different ways using the known and the predicted records in the form of either the probability or presence/absence. These metrics include the commonly used discriminatory statistics, a map comparison measure, and a niche similarity measure. For the metrics that gives the level of agreement at a grid cell level, the values are summarized either globally or locally within suitable/unsuitable/marginal areas of the habitat for each species (Fig. 5-2). The later leads to evaluate whether a model performs the same over different parts of the study area with different degrees of suitability, and therefore gives a more precise estimation of the model performance comparing to the abstract global statistics (*e.g.*, AUC).

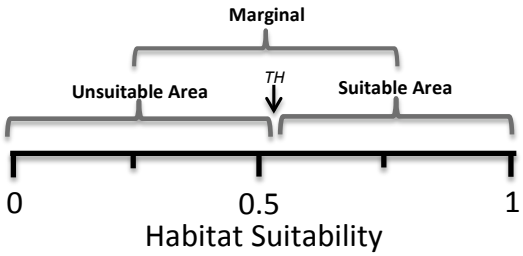



**Figure 5-2:** Defining the suitable/unsuitable/marginal area of the habitat

We used the known species data as the validation datasets (in the form of either probability or presence-absence) over the entire area. For metrics including map comparison and niche similarity, the whole data (entire map) were used in the evaluation procedure. For the other metrics, we

drew a sample including 500 presence-absence records realized from the probability map for each species. For this procedure, we used the same probabilistic approach as for realizing the training dataset (see the generating virtual species section). Using such approach avoids an over-optimistic estimate of discrimination ability (Meynard and Kaplan 2013). We believe that this is a necessary step in generating a virtual species that allows taking the stochasticity of species detection into account.

- *Discriminatory metrics* - Discriminatory metrics measure the ability of a model to correctly discriminate between occupied and unoccupied sites in a validation dataset (Pearce and Ferrier 2000). We used four commonly used discrimination metrics including Youden index (Youden 1950), area under the curve (AUC) of a receiver operating characteristic (ROC) plot, sensitivity (true positive rate) and specificity (true negative rate). An ROC was created to calculate the AUC for each SDM. An ROC plots sensitivity on the y-axis against "1 – specificity" (false positive rates) on the x-axis for all thresholds (Fielding and Bell 1997). Youden index is referred to as true skill statistics (hereafter TSS) in SDM studies (Allouche et al. 2006) and is quantified as "sensitivity+specificity-1". Both AUC and TSS are prevalence independent measures. AUC and TSS range from 0 to 1 and from -1 to 1, respectively. A value of 1 for both statistics indicates perfect discrimination. An AUC value of 0.5 and a TSS value of 0 imply random predictive discrimination while values less than 0.5 for AUC and less than 0 for TSS indicate discrimination worse than chance. Although AUC is a threshold-independent metrics, meaning that the predicted probabilities of occurrence are used to quantify the metric, TSS works with presence and absence values. Therefore, the predicted probabilities were converted to presence and absence by applying an optimum threshold at which sensitivity is equal to specificity (*i.e.*, when $min|\text{sensitivity} - \text{specificity}|$ is satisfied) (Jiménez-Valverde 2014). We used this threshold everywhere when we needed to transform the probabilities to the presence-absences.

- *Map comparison metric* –We used a map comparison approach to evaluate the spatial similarity (agreement) of two observed and predicted presence-

absence patterns. Although this is the first time that a map comparison method is used in an SDM study, these methods have been widely used in some other disciplines such as remote sensing and land use dynamic studies. Among several available map comparison methods, we selected hierarchical fuzzy pattern matching (hereafter, FuzzyMatch) (Power et al. 2001), which uses fuzzy set theory to capture the complexity of the spatial patterns at both a local and global level and provides a more robust alternative to traditional approaches (which were mostly based on pixel-by-pixel comparison). Comparison at the local level (*i.e.*, local matching) determines the degree of containment of each unique polygon (*i.e.*, a presence or absence patch or group of pixels) in terms of fuzzy areal intersections. The local agreement values are calculated from a fuzzy logical Max-Min compositional algorithm (see Power et al., 2001 for a full description). A global similarity value is derived by the fuzzy summation of the local agreements. We implemented the method as a new function in R (R Development Core Team 2013) and used it to compare each predicted presence-absence map (in the form of raster) with the know presence-absence map for each species. The output of the function is a new raster map, where the pixels contain the fuzzy measure of local agreement as well as a global measure of overall agreement. The values range between 0 and 1 indicate no and perfect match between the observed and the predicted presence-absence pattern, respectively.

- *Niche similarity metric* – There are several indices to measure the niche similarity (also called niche equivalency or niche overlap) between two species (Hurlbert 1978 ; Warren et al. 2008) which quantify the degree that two species overlap (similar) in their utilization of niche. We used one of these metrics, *I* to measure the degree that the predicted niche is similar to the true niche. In contrast with the other metrics that use presence-absence data, this metric uses the probability of occurrence. This metric ranges from 0 (where the predicted habitat shows no similarity to the truth) to 1 (where the two habitats are completely similar).

*I* can be measured as:

$$I = 1 - \frac{1}{2}\sqrt{\sum_{i=1}^{N}\left(\sqrt{P_{x_i}} - \sqrt{P_{y_i}}\right)^2} \qquad (5\text{-}1)$$

where $P_{x_i}$ and $P_{y_i}$ denote the standardized probability of occurrence for species x and y (in our case, the known and the predicted probability) at cell $i$, respectively; and $N$ is the total number of cells in the geographic space (in our case, total number of cells in the environmental space; see the following description).

This method has been used to measure the niche similarity based on deriving the cell value in space, *i.e.*, two niches are compared in geographic space. This method, however, may suffer from a weighting in favour of common environments (Dormann et al. 2010). Alternatively, the niche overlap can be calculated in environmental space (Dormann et al. 2010 ; Graham et al. 2004b) which automatically avoids the problem of commonness of environments. It has also been criticized that the estimation of niche similarity may be biased because in environmental space certain environmental variable value combinations would be used which are not found in the study region (Warren et al. 2008). To avoid the both problems, we calculated the metric in environmental space where we masked out the environmental value combinations that were not found in the geographic region.

### 5-2-4- Model uncertainty

We evaluated the inconsistency in predictions at the cell (pixel) level as an indicator of model uncertainty. To do so, we used a standardized Shannon entropy statistic to measure the amount of inconsistency in the presence-absence predictions by different models and different simulations at each pixel. The Shannon entropy has been defined as an average amount of information to eliminate uncertainty, given by a finite number of events:

$$H = -\sum_{i=1}^{2} p_i \log_2 p_i \Big/ log_2(2) \qquad (5\text{-}1)$$

where $H$ measures the entropy of a system with two possible events (*i.e.*, presence or absence) and $p_i$ presents the probability of event $i$. $H$ ranges between 0 and 1. A value of 0 indicates that all predictions are consistently equal, and a value of 1 indicates a maximum inconsistency between the predictions. The inconsistency is at maximum when all events occur in equal abundance, *i.e.* half of the predictions are presence and the other half are absence. We measured the uncertainty between and within the models at the cell level, and then summarized the values in the three suitability regions (*i.e.*, suitable, marginal and unsuitable). The uncertainty within a model gives the inconsistency in predictions of a certain model (*e.g.*, GLM) in our simulations (measured from the 100 realized predictions for each species) for all species. Since the numbers of cells at each suitability group were different for different species, we drew a sample with a size of 100 from each suitability region for each species. Then the inconsistency values for the 10 species were put together, gave 1000 values at each suitability region (3000 values in total) for each model.

The uncertainties between the models gives the inconsistency of predictions by different models (also known as model-based uncertainty) (Pearson et al. 2006). Given the predictions from 100 realizations of different models, the inconsistency between the models at each cell was calculated. We grouped the models into presence-absence and presence-only methods, and separately measured the model uncertainty for each group.

## *5-3- Results*

In Fig. 5-3, we illustrated the predicted habitat suitability obtained using four different SDMs (as an example) together with the truth distribution for one of the species (*i.e.*, species 7) in both the geographic and environmental space.

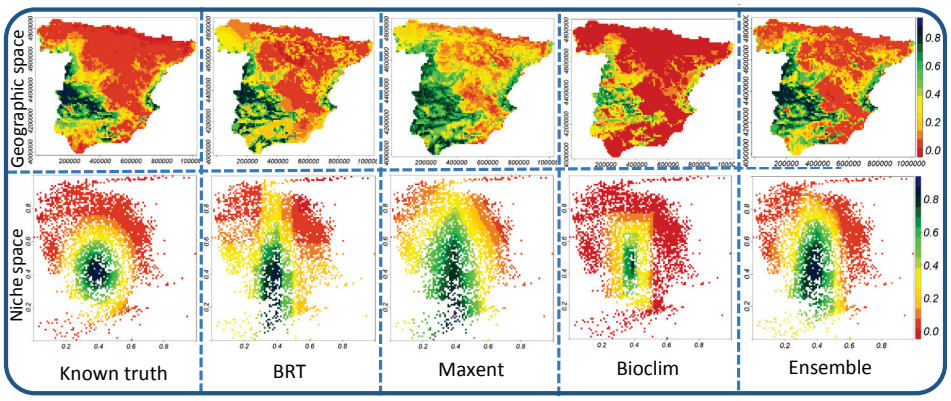



**Figure 5-3:** Maps for one species represents the known and the predicted habitat suitability (by four SDMs) in both the geographic and environmental space; colour scheme represents the probability of occurrence; in the figures in the lower panel, x and y axis represent precipitation NDVI, respectively

### 5-3-1. Comparison of models using global measures

Variation of the performance measures including AUC, TSS, Sensitivity, Specificity, *I* and FuzzyMatch statistics over all species and for each SDM were illustrated using boxplots (Fig. 5-4). The mean AUC values across the models ranged between a minimum of 0.67 (for Bioclim) and a maximum of 0.82 (for Ensemble). The same trend was derived from TSS for which the mean values across the models ranged between 0.27 (for Bioclim) and 0.47 (for Ensemble). The difference between the mean AUC (as well as the mean TSS) of the presence-absence models (including Maxent from the presence-only models) were not significant ($P > 0.1$), whereas this statistic for the profile models (*i.e.*, Bioclim, Domain, and Mahalanobis) were significantly lower ($P < 0.001$) than the other presence-only (and presence-absence) models. The test showed that the mean AUC and TSS values of the Ensemble approach were greater than the corresponding statistics of all the other models ($P < 0.001$).

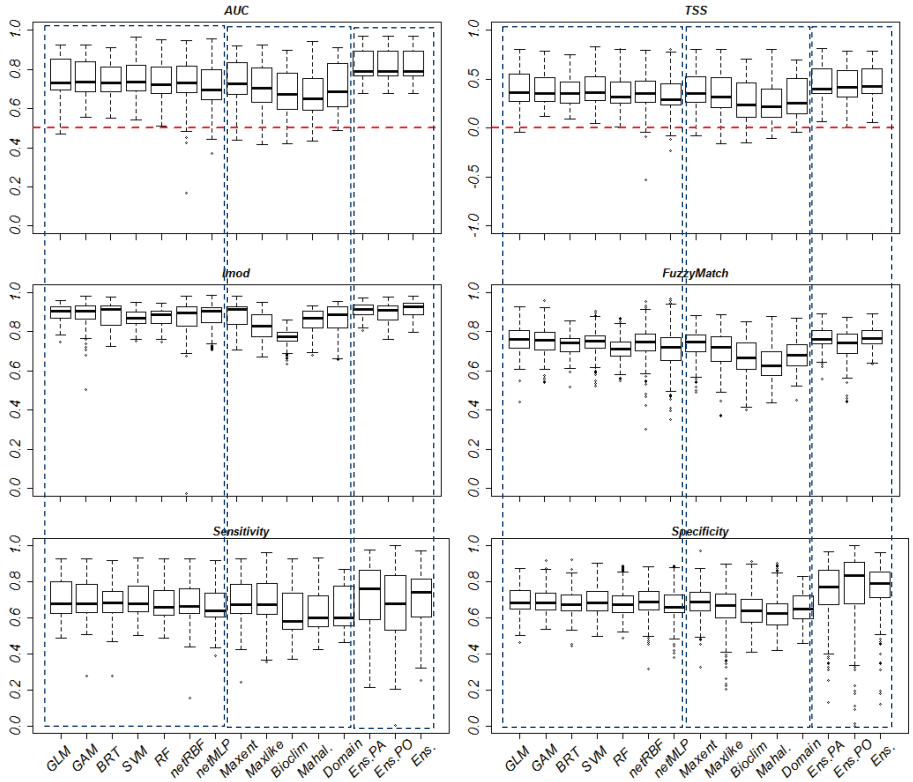



**Figure 5-4:** Boxplots of overall performance measures for each 15 models; the blue boxes separate the presence-absence, presence-only and ensemble models (from left to right); the dashed red line represent the measure at which the model performs not better than random

The other metrics generally showed almost the same trend between the models. We also measured the standard deviation (SD) of the metrics over the simulations as another base to compare the models. Obviously, lower SD indicates less variability in the performance and is preferable. We summarized all indicators into a single measure of performance for each model by taking their mean over all the simulations. To do so, we rescaled the AUC values to make it comparable with the other metrics using equation $2 \times AUC - 1$. We then generated a scatterplot by putting the summarized single measure of performance and the mean of SD values for each model on the x and y axes, respectively (Fig. 5-5). This provides a base to visually compare the models by taking all the metrics into

consideration. The graph clearly shows that the ensemble approaches strongly outperform the other models. A group of presence-absence models including BRT, GLM, GAM, RF, SVM and NNetRBF as well as Maxent from the presence-only models performed almost the same based on the model accuracy (performance in our graph) but different based on SD. Between these models, BRT performed the best as showed the minimum SD. Bioclim and Mahalanobis from the profile based models were the worst as shows the poor accuracy and the high SD.

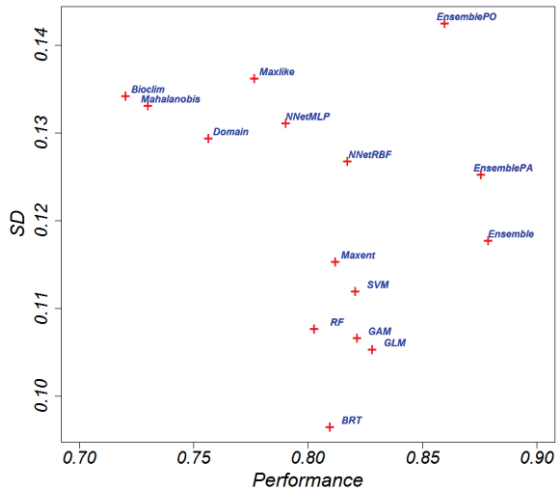



**Figure 5-5:** Scatterplot of overall performance measures based on AUC, TSS, *I*, FuzzyMatch, sensitivity and specificity vs the mean standard deviation of the performance measures across all species; it is desirable to have low SD and high performance for consistent and good accuracy

### 5-3-2. Comparison of models using local measures

Both *I* and FuzzyMatch metrics measured the level of agreement between the true and the predicted habitat suitability at the cell level. The mean of the cell values over each suitability region was calculated as the performance of a model for the local area. The variations of the local measures for each model across all species were illustrated in Fig. 5-6.

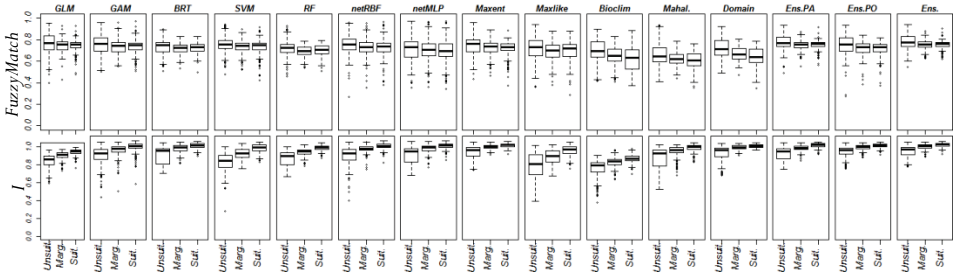



**Figure 5-6:** Boxplots of the performance measures including FuzzyMatch and I metrics across all species at each suitability area (i.e., unsitable, marginal and suitable) and for each model

The mean FuzzyMatch metric across the local areas showed no significant difference for the presence-absence as well as EnsPA and Ens. models ($P > 0.1$), but was slightly greater in the unsuitable area compared to the other areas for the presence-only and EnsPO models. The mean value of this metric over all the models for the unsuitable, marginal, and suitable areas were 0.74, 0.71 and 0.70 respectively. The min and the max values were 0.61 and 0.78, related to the Mahalanobis model in the suitable area, and the Ensemble model in the unsuitable area, respectively. The variability (standard deviation) of the metric was greater in the unsuitable area compared to the other areas for all the models.

In contrast to FuzzyMatch, the *I* metric showed significantly greater in the suitable area than the two other areas, and also in the marginal area than the unsuitable area for all the models ($P < 0.001$). The mean value of this metric over all the models for the unsuitable, marginal, and suitable areas were 0.83, 0.89 and 0.93 respectively. The min and the max values were 0.71 and 0.95, related to the Bioclim model in the unsuitable area, and the Ensemble model in the suitable area, respectively. As with the FuzzyMatch, the variability of the *I* metric was greater in the unsuitable area comparing to the other areas for all the models.

### 5-3-3. Model uncertainty

The model uncertainty or the inconsistency of different models in their predictions was measured at the cell level by grouping the models into the presence-absence and presence-only (Fig. 5-7).

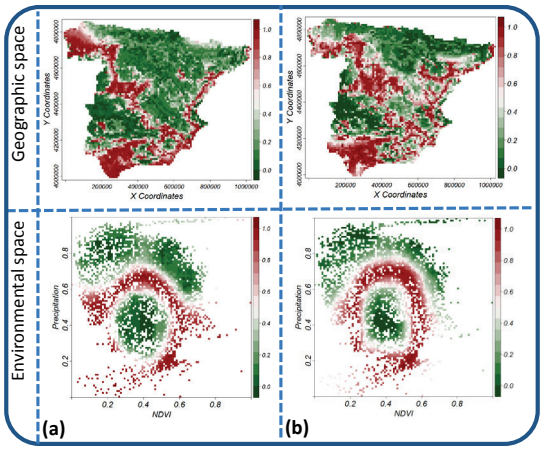



**Figure 5-7:** Model uncertainty maps which show the inconsistency of models over the geographic and environmental space for one of the species (i.e., species7); (a) presence-absence models; (b) presence-only models

The overall variability of the inconsistencies for all species across the suitability regions (Fig. 5-8) showed that the inconsistencies are greater in the marginal area compared to the suitable and unsuitable areas. Both groups of models performed more consistent in the suitable area compared to the two other areas. The level of inconsistency was higher for the presence-only models compared to the presence-absence models.

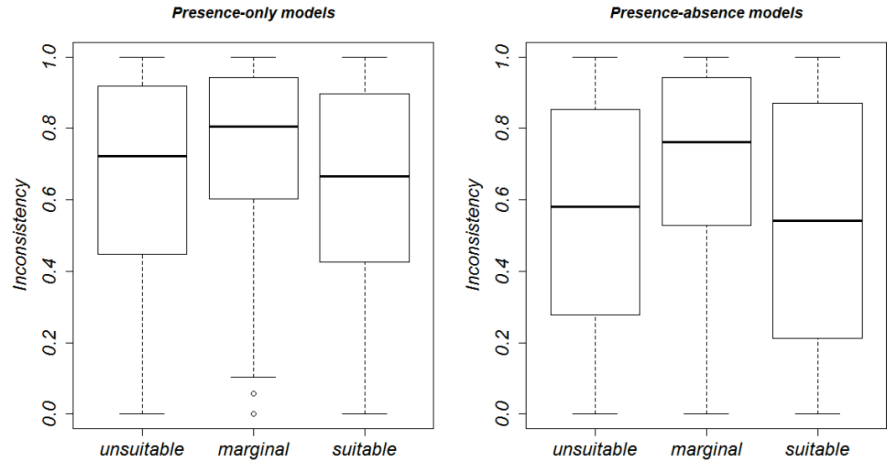



**Figure 5-8:** The level of inconsistency over the presence-only (left) and the presence-absence (right) SDMs, separated across the suitability regions

Within model uncertainty (i.e., the inconsistency between the predictions over the multiple runs of each model) provided a base to compare the models. The lowest and the highest inconsistency values were related to the ensemble (i.e., Ens.) and the Mahalanobis models with a mean inconsistency of 0.66 and 0.91, respectively. Among the group of the presence-absence models, GLM was the most consistent model (with a mean inconsistency of 0.67), and RF was the most inconsistent model (with a mean inconsistency of 0.82). Among the group of the presence-only models, the lowest inconsistency was related to the Maxent model (with a mean of 0.74). The mean inconsistency of the ensemble models based on the presence-absence and the presence-only models (i.e., EnsPA and EnsPO) were equal to the lowest inconsistency values of the models in the corresponding group (i.e., 0.67 and 0.74, respectively).

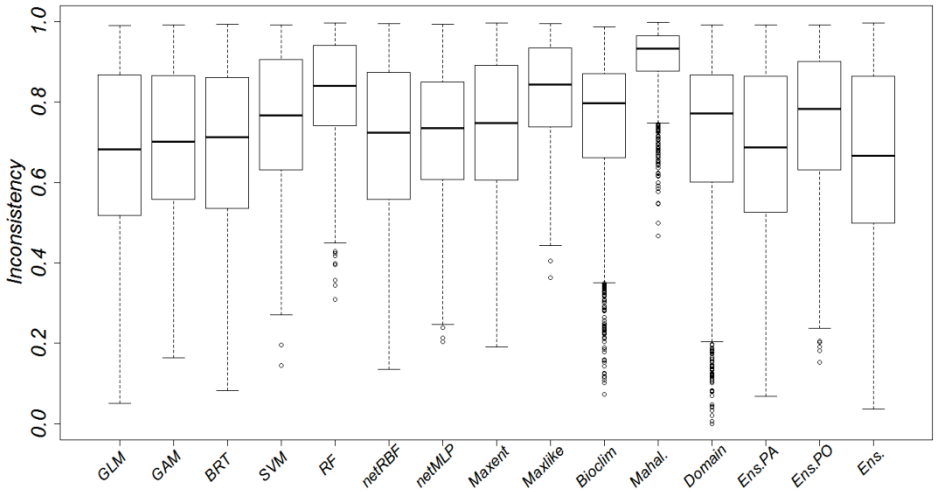



**Figure 5-9:** Within model uncertainty across all species over the 100 runs for each species

## *5-4- Discussion*

The contribution of this study is unique as the performance together with the model uncertainty of the commonly used models were explored using a set of alternative evaluation methods through a controlled way. In similar studies (*e.g.*, Elith et al., 2006), the summarized global statistics (such as AUC) were commonly used as a base to compare the models. Although our extended set of evaluations are not against such methods (as we found the consistent results between alternative evaluation methods at the global level), we showed that the alternative evaluation methods gain deeper insights into the behaviour and characteristics of a model. Using these methods opens a door to explore where a model performs well (or badly) and why. These insights can be useful to choose a method for an intended application, and even more useful for development of new SDM methods.

This study delivered a clear message in favour of the superiority of the ensemble methods over any single model for species distribution modelling. Our results showed that the ensemble of SDM predictions not only reduces the uncertainty associated with the predictions but also significantly increase the predictive accuracy. It has been widely

appreciated that the ensemble (consensus) methods can be used to make robust decision making in the face of uncertainty (Araújo and New 2007 ; Araújo et al. 2005 ; Buisson et al. 2010). Several studies also evaluated the ensemble methods and argued that they may increase the accuracy of species distribution forecasts (Gritti et al. 2013 ; Marmion et al. 2009b). Our results strongly support both of these arguments and suggest that the ensemble method can be considered as a preferable model to fit, predict or forecast species distributions either with presence-absence or presence-only data.

In this study, we tested 12 single and one ensemble (i.e., weighted averaging) models. There are, however, many other alternatives available and also new methods will continue to be developed. Our findings confirm that the ensemble approach is a promising area to focus. We suggest that using our extended set of evaluation tools can be helpful in attempts intending to develop new methods or to address critical underlying issues in the existing methods. Our precise evaluation over space may also be useful to extend the current ensemble methods (*e.g.*, the weighted averaging as we used in this study) toward more precise methods (for example using spatially varying weights rather than global weights in the procedure of weighted averaging), and would be worthwhile for further studies.

It is difficult to measure a model quality when the model may be affected by issues in data such as spatial sampling bias, errors in species identification, uncertainty in species location, incomplete sample, etc. and consequently may lead to a misleading comparison. In this research, we used simulated species rather than real one which provided the opportunity of avoiding such misleading effects and proceeding with the study in a controlled way. However, the simulated data do not cover all possible complications that are likely to be found in real data (Naimi et al. 2011). We simulated 10 species with different niche shapes, from simple (*e.g.*, species1) to more complex (*e.g.*, species10), to cover at least part of the complications that may be expected in real species-environment

relationships. Clearly, the models may differ in their potentials to detect and characterize the true niche for any of these species. Although exploring such differences (*i.e.*, evaluating how the models perform for each type of species with a specific niche shape) is interesting and worthwhile, we did not go through such details because the aim of the research was to generally compare the models in a realistic and controlled way. Therefore, we assumed that the overall behaviour of a model for our 10 different species (as examples of possibilities) can be inferred as the model behaviour in real situations.

Most of widely used evaluation metrics need to apply a threshold transformation of predicted probabilities into presence-absences. Together with these binary (*i.e.*, threshold-dependent) metrics, we also employed a metric (*i.e.*, *I*) that directly use the probabilities in the evaluation procedure. These kinds of metrics provide the opportunity of avoiding any information loss through the probability transformation which may provide new insights into the behaviour of a model. However, it should be noted that we need to know the true habitat suitability (or probability of occurrence) to be able to use these metrics. Hence, we worked with simulated data. There are also some other metrics which directly use predicted probabilities in the evaluation procedure (*i.e.*, when the observed values are presence-absences; Lawson et al., 2014). These kinds of metrics can be considered as a tool to measure a calibration or reliability of a model. A calibration metric measures the degree to which the proportion of observed positive cases (empirically estimated probabilities) equals to the model estimated probabilities in any given validation dataset (Jiménez-Valverde et al. 2013). Although the necessity of using the calibration metrics together with discrimination metrics have been emphasized (Pearce and Ferrier 2000), we did not use such metrics in our study as they remain untested for SDMs and moreover, we doubted whether they can be used in comparison studies. In the few SDM studies that did pay attention to calibration rather than focusing on the discrimination ability (as with majority of studies), the measurement were usually limited to a calibration

plot to only visually check whether a model is calibrated (Jiménez-Valverde et al. 2013).

Our results showed a contradiction between the FuzzyMatch and the *I* metrics at the local level (*i.e.*, the models performed the best in the suitable area based on the mean *I* value while the mean FuzzyMatch values showed a greater value in the unsuitable areas). A major difference between these two metrics is related to the type of data they use in the procedure, *i.e.*, the *I* metric uses the probability while FuzzyMatch uses the presence-absence. Based on this result, we know that the predicted values in the suitable and then in the marginal areas are more similar to the true values of habitat suitability than in the unsuitable area. Moreover, the opposite result from the FuzzyMatch could be related to the selection of a threshold.

## *5-5- Conclusions*

In this study, we comprehensively evaluated several SDMs using a set of evaluation metrics. The results indicate that the ensemble-based model strongly outperforms the other models either when accuracy or uncertainty in the predictions is a matter. We suggest that using ensemble approach would be worthwhile even when presence-only data are available. The set of evaluation metrics we used in this study can be useful for further investigations of model behaviours and provide deeper insights into the causes of varying model performances.

# *Chapter 6*

# Synthesis

# 6. Synthesis

## 6-1. Introduction

The overall goal of this thesis was to gain insight into the impact of errors and uncertainties on the performance of species distribution modelling. Of the many sources of uncertainty that may affect the modelled species distributions, locational errors in species data and variability in (or inconsistency within) different SDMs' predictions are the major sources. This research focused on these two important sources of uncertainty and explored their potential impact on SDMs and the possible solutions to overcome them. Furthermore, a major focus of this thesis was dedicated to the role of spatial autocorrelation (a property of spatial data). Measuring spatial autocorrelation was used as an opportunity to understand the impact of positional uncertainty both locally (at each location) and globally (overall impact). A new method was also developed to overcome an existing gap in measuring spatial autocorrelation.

In this final chapter, the most important results from the thesis are summarized together in order to better understand the key aspects of this thesis and to highlight the inter-relationships between them.

## 6-2. Summary of results and their inter-relationships

1) *Are the species distribution models robust to positional uncertainty?*

Massive information resources of species data (more than 2.5 billion specimen collections worldwide) held in museums, herbaria and other institutions can potentially play an important role in our understanding of ecological and evolutionary processes (Graham et al. 2004a) and provide a base of the knowledge for many applications such as conservation planning. Problems with quality of these data is a major source of uncertainty and may limit their use in a diverse array of applications (Graham et al. 2004a). Increasingly these data are made available through Internet portals. Among different potential problems that may affect the quality of these data, the

uncertainty about where an observation is located (positional uncertainty) is the most obvious one. The reason is that the majority of these data were collected as textual descriptions before the popularization of GPS technology. When these records were digitized, geographic coordinates were often inferred and may be several kilometres incorrect in their positions. Therefore, an important question is that to what extent an SDM is affected by this source of uncertainty. The chapters 2 and 3 contributed to answer this question and provided a numerical approach to explore whether and where the SDMs are affected by the positional uncertainty.

Our findings showed that SDMs are robust to positional uncertainty only when the level of uncertainty is less than the range of spatial autocorrelation in predictors. Based on this result, we introduced an approach (chapter 2) to understand the potential impact of positional uncertainty in species occurrences on the predictions of SDMs by examining the spatial autocorrelation range in predictor variables. We provided a graph for each model as a practical solution to link model robustness to positional uncertainty. This graph (Fig. 6-1) can be used in any SDM study to explore the potential decline in accuracy as a consequence of positional uncertainty.

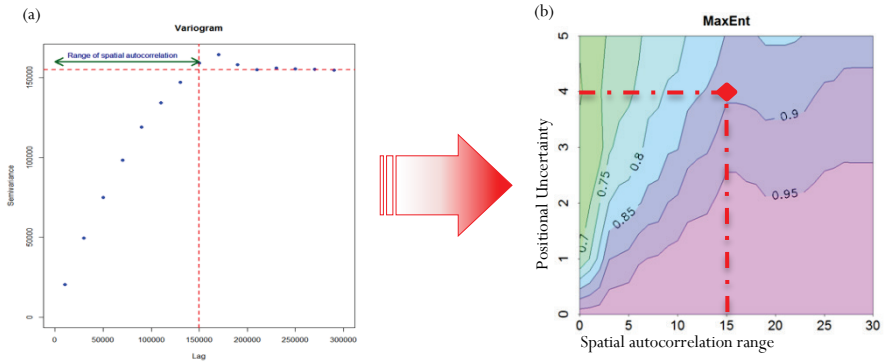



**Figure 6-1:** A solution to understand the impact of positional uncertainty on SDM; (a) examining spatial autocorrelation in a predictor using variogram to find out the autocorrelation range; (b) crossing the level of positional uncertainty and the spatial autocorrelation range gives the expected discrimination capacity of the model (i.e., AUC) that should be compared with the accuracy at the same autocorrelation range on x-axis but with no positional uncertainty (i.e., y=0) to understand the decline in the performance

The approach introduced in chapter 2, used global spatial autocorrelation to understand whether positional uncertainty impacts SDM. Chapter 3 introduced an approach for exploring where positional uncertainty caused a problem in geographic space. The approach was developed based on extending the main idea in chapter 2 by employing local spatial autocorrelation to explore whether the errors in locations affect SDMs negatively. For this purpose, using local spatial autocorrelation statistics are more insightful because we showed that it led to identification of the specific occurrence records that cause the largest drop in SDM prediction accuracy. Of key importance is that an appropriate strategy can be considered to overcome the problem. For example, our approach can help to target locations (i.e., those with low local spatial autocorrelation) that could be selected for additional field sampling. A limited survey then may be designed for these areas to provide or modify the sample locations which are located at the problematic area.

2) *Measuring local spatial autocorrelation for both continuous and categorical data*

The approach presented in chapters 2 and 3 takes the spatial autocorrelation as an opportunity to understand the impact of positional uncertainty in species data. There is, however, a limitation related to the available methods for measuring spatial autocorrelation. Although several methods exist for this purpose, these methods can only be used for continuous or interval variables. For SDM studies as well as many other applications, there are numerous situations where categorical variables are encountered. In the absence of a method to measure the level of spatial autocorrelation in categorical variables, our approach in the chapter 2 and 3 is not suitable. This concern provided the main motivation for developing a new statistic to measure local spatial association in both continuous and categorical data (chapter 4). This statistic may be used not only for SDM studies (in relation to what we presented in chapter 2 and 3), also may be used in many other disciplines where local spatial autocorrelation needs to be quantified (*e.g.*, characterizing land cover structure, disease clustering, analysing economic data).

3) *Model uncertainty in SDMs' predictions and possible solutions*

In chapter 4, we measured within and between model inconsistencies in their predictions using several commonly used modelling approaches. This can be considered as another important source of uncertainty (*i.e.*, model uncertainty) for SDM studies. The comprehensive explorations we used in chapter 4 using simulated data provided the possibility of evaluating how different models perform and whether they generate inconsistent predictions. We showed that the ensemble (consensus or multi-model) prediction is a possible solution to reduce model uncertainty. Moreover, we showed that this approach outperforms the other SDMs and increases the accuracy of the predictions. Therefore, our findings in chapter 4 suggest that the ensemble approach should be used as an alternative to any single-model SDM.

## 6-3. General discussion

SDMs are useful if they are robust (Guisan and Thuiller 2005). Measuring the accuracy can show how well a model performs but cannot determine to what extent the model is reliable or robust. This can be understood by quantifying the level of uncertainty in the model predictions.

SDMs have been widely used to inform management decisions (Araújo and Peterson 2012). By identifying the sources of uncertainty, advice and management actions will be better informed than if based on false certainty. For this purpose, quantifying the level of uncertainty in model predictions is as important as the predictions themselves (Beale and Lennon 2012). This information, however, is usually missing from the studies, and its' importance is rarely emphasized. This might be because uncertainty is a difficult issue (Guisan et al. 2006) and its' quantification is complicated and computationally expensive. Despite an emphasis on the need to develop an integrated framework for assessing uncertainties and error propagation analysis throughout the modelling process (Guisan et al. 2006), to our knowledge, no such framework has been developed so far. The approach we introduced in this thesis (chapter 2 and 3) can be considered as an alternative to such frameworks. Of key important of this approach is that researchers are able to understand whether a model is robust to an uncertainty (positional uncertainty in our study) without the need for tracing the uncertainty propagation through complicated and computationally expensive procedures.

Uncertainty in SDM predictions originates from many different sources but falls broadly into two main groups: data, and model. Due to numerous sources of uncertainty, the models and their results should only be applied with a thorough understanding of the limitations involved (Heikkinen et al. 2006). Many studies tried to broaden our understanding of the wide range of methodological issues (*e.g.*, multicollinearity, sample size, spatial autocorrelation) that may affect the usefulness of a model. In this thesis, we introduced a set of evaluation tools which provide more information on the

behaviour of the models compared to the traditional methods. We argued that using such tools is necessary especially for studies with a focus on methodological aspects of modelling (*e.g.*, in order to explore technical issues or improve the accuracy of models).

Although many sources of uncertainty (*e.g.*, collinearity, biased sampling, missing predictors, autocorrelation) have been previously identified (Heikkinen et al. 2006) and occasionally measured (Buisson et al. 2010), there is no study where the effect of all known sources of uncertainty together has been measured (Beale and Lennon 2012). An integrated framework to assess multi-source uncertainty and measure multiplicative effects of difference sources of uncertainty on SDMs is a missing but necessary tool. Moreover, the sensitivity of various SDMs to multi-source may be different, and therefore, it would be worthwhile to take this into account in a comparison study (like chapter 5 of this study). In other words, as well as comparing SDMs under perfect situations (e.g., when artificial data are generated in a controlled way to avoid the effect of any errors), it would be also worthwhile to study SDMs and compare them under imperfect situations (e.g., when they are subjected to multi-source uncertainty as in many realistic situations).

Finally, we know that SDMs currently contain many assumptions and the uncertainty (Wiens et al. 2009). Despite the ways of addressing some assumptions and reducing the uncertainty (such as the ensemble approach for reducing uncertainty; chapter 5), knowing the level of uncertainty in SDMs outputs is important not only for managers to understand and manage the risk of actions, but also for scientists to focus their efforts in advancing niche-based modelling.

## 6-4. Future research avenues

Following recommendations are suggested for the future works on the topic:

1) Evaluate how SDMs perform in a data poor setting using extensive evaluation tools (as introduced in chapter 5). This would be an extension to our evaluations of SDMs in chapter 5 when the situation with data is imperfect (*e.g.*, when data are subjected to common issues including geographical bias, low sample size, spatial autocorrelation, etc.).
2) Evaluate how different functional shapes (shapes of species response to environmental variables) affect various SDMs, and how these functions are better rooted in ecological theories. This would help to evaluate the models with respect to ecological theory and set a trade-off between goodness of fit (statistically) and interpretability (ecologically).
3) Exploring the effect of multi-source uncertainty on the performance of SDMs.
4) Developing an integrated framework for quantifying the level of uncertainty (single and multiple sources) as well as new statistics that identify and measure uncertainty together with accuracy.

# References


ter Braak, C.J.F.& C.W.N. Looman (1986). "Weighted Averaging, Logistic-Regression and the Gaussian Response Model." *Vegetatio,* 65, 3-11.

Ahlqvist, O. & A. Shortridge (2006). "Characterizing Land Cover Structure with Semantic Variograms." In *Progress in Spatial Data Handling,* 401-415, edited by A. Riedl, W. Kainz & G. Elmes. Springer Berlin Heidelberg.

Ahlqvist, O. & A. Shortridge (2010). "Spatial and Semantic Dimensions of Landscape Heterogeneity." *Landscape Ecology,* 25, 573-590.

Allouche, O., A. Tsoar & R. Kadmon (2006). "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (Tss)." *Journal of Applied Ecology,* 43, 1223-1232.

Anand, M. & L. Orloci (1996). "Complexity in Plant Communities: The Notion and Quantification." *Journal of Theoretical Biology,* 179, 179-186.

Anderson, R.P., D. Lew & A.T. Peterson (2003). "Evaluating Predictive Models of Species' Distributions: Criteria for Selecting Optimal Models." *Ecological Modelling,* 162, 211-232.

Anselin, L. (1995). "Local Indicators of Spatial Association—Lisa." *Geographical Analysis,* 27, 93-115.

Araújo, M.B. & A. Guisan (2006). "Five (or So) Challenges for Species Distribution Modelling." *Journal of Biogeography,* 33, 1677-1688.

Araújo, M.B. & M. New (2007). "Ensemble Forecasting of Species Distributions." *Trends in Ecology &Evolution,* 22, 42-47.

Araújo, M.B. & A.T. Peterson (2012). "Uses and Misuses of Bioclimatic Envelope Modeling." *Ecology,* 93, 1527-1539.

Araújo, M.B., R.J. Whittaker, R.J. Ladle & M. Erhard (2005). "Reducing Uncertainty in Projections of Extinction Risk from Climate Change." *Global Ecology and Biogeography,* 14, 529-538.

Atkinson, P.M. (1993). "The Effect of Spatial-Resolution on the Experimental Variogram of Airborne Mss Imagery." *International Journal of Remote Sensing,* 14, 1005-1011.

Austin, M. (2002). "Spatial Prediction of Species Distribution: An Interface between Ecological Theory and Statistical Modelling." *Ecological Modelling,* 157, 101-118.

Austin, M. (2007). "Species Distribution Models and Ecological Theory: A Critical Assessment and Some Possible New Approaches." *Ecological Modelling,* 200, 1-19.

Austin, M.P., L. Belbin, J.A. Meyers, M.D. Doherty & M. Luoto (2006). "Evaluation of Statistical Models Used for Predicting Plant Species Distributions: Role of Artificial Data and Theory." *Ecological Modelling,* 199, 197-216.

Batty, M. (1974). "Spatial Entropy." *Geographical Analysis,* 6, 1-31.

Batty, M. (1976). "Entropy in Spatial Aggregation." *Geographical Analysis,* 8, 1-21.

Beale, C.M. & J.J. Lennon (2012). "Incorporating Uncertainty in Predictive Species Distribution Modelling." *Philosophical Transactions of the Royal Society B: Biological Sciences,* 367, 247-258.

Beale, C.M., J.J. Lennon, J.M. Yearsley, M.J. Brewer & D.A. Elston (2010). "Regression Analysis of Spatial Data." *Ecology Letters,* 13, 246-264.

Beven, K. & M. Kirkby (1979). "A Physically Based, Variable Contributing Area Model of Basin Hydrology/Un Modèle À Base Physique De Zone D'appel Variable De L'hydrologie Du Bassin Versant." *Hydrological Sciences Journal,* 24, 43-69.

Boots, B. (2003). "Developing Local Measures of Spatial Association for Categorical Data." *Journal of Geographical Systems,* 5, 139-160.

Boots, B. (2006). "Local Configuration Measures for Categorical Spatial Data: Binary Regular Lattices." *Journal of Geographical Systems,* 8, 1-24.

Boschetti, F. (2008). "Mapping the Complexity of Ecological Models." *Ecological Complexity,* 5, 37-47.

Box, E. (1981). "Predicting Physiognomic Vegetation Types with Climate Variables." *Vegetatio,* 45, 127-139.

Breiman, L. (2001). "Random Forests." *Machine Learning,* 45, 5-32.

Brotons, L., W. Thuiller, M.B. Araújo & A. H. Hirzel (2004). "Presence-Absence Versus Presence-Only Modelling Methods for Predicting Bird Habitat Suitability." *Ecography,* 27, 437-448.

Buermann, W., S. Saatchi, T.B. Smith, B.R. Zutta, J.A. Chaves, B. Milá & C.H. Graham (2008). "Predicting Species Distributions across the Amazonian and Andean Regions Using Remote Sensing Data." *Journal of Biogeography,* 35, 1160-1176.

Buisson, L., W. Thuiller, N. Casajus, S. Lek & G. Grenouillet (2010). "Uncertainty in Ensemble Forecasting of Species Distribution." *Global Change Biology,* 16, 1145-1157.

Busby, J.R. (1991). "Bioclim-a Bioclimate Analysis and Prediction System." *Plant Protection Quarterly (Australia)*.

Bystriakova, N., M. Peregrym, R.H.J. Erkens, O. Bezsmertna & H. Schneider (2012). "Sampling Bias in Geographic and Environmental Space and Its Effect on the Predictive Power of Species Distribution Models." *Systematics and Biodiversity,* 10, 305-315.

Carpenter, G., A. Gillison & J. Winter (1993). "Domain: A Flexible Modelling Procedure for Mapping Potential Distributions of Plants and Animals." *Biodiversity &Conservation,* 2, 667-680.

Chapman, A.D. (2005). *Uses of Primary Species Occurrence Data,* Copenhagen, Global Biodiversity Information Facility.

Chatterjee, S. & A.S. Hadi (2006). *Regression Analysis by Example,* New York, John Wiley & Sons.

Cliff, A.D. & J.K. Ord (1981). *Spatial Processes: Models and Applications,* London, Pion.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and psychological measurement,* 20, 37.

Cressie, N. (1993). *Statistics for Spatial Data,* New York, John Wiley.

Cutler, D.R., T.C. Edwards, K.H. Beard, A. Cutler & K.T. Hess (2007). "Random Forests for Classification in Ecology." *Ecology,* 88, 2783-2792.

Dacey, M.F. (1968). "A Review on Measures of Contiguity for Two and K-Color Maps." In *Spatial Analysis: A Reader in Statistical Geography,* 479-495, edited by B. J. L. Berry & D. F. Marble. New Jersey, Prentice-Hall Englewood Cliff.

De Cabrera, T. (2007). "Microtus Cabrerae." In *Atlas De Las Aves Reproductoras De Españia,* 429-431, edited by R. Marti & J. C. Del Moral. Madrid, SECEM-SECEMU.

De Souza Muñoz, M., R. De Giovanni, M. De Siqueira, T. Sutton, P. Brewer, R. Pereira, D. Canhos & V. Canhos (2009). "Openmodeller: A Generic Approach to Species' Potential Distribution Modelling." *GeoInformatica*, 1-25.

Degraaf, R.M. & M. Yamasaki (2002). "Effects of Edge Contrast on Redback Salamander Distribution in Even-Aged Northern Hardwoods." *Forest Science,* 48, 351-363.

Desrochers, A., I.K. Hanski & V. Selonen (2003). "Siberian Flying Squirrel Responses to High-and Low-Contrast Forest Edges." *Landscape Ecology,* 18, 543-552.

Di Gregorio, A. & L.J. Jansen (2009). *Land Cover Classification System: Lccs: Classification Concepts and User Manual,* Rome, Food and Agriculture Organization of the United Nations.

Dietrich, C.R. & G.N. Newsam (1993). "A Fast and Exact Method for Multidimensional Gaussian Stochastic Simulations." *Water Resources Research,* 29, 2861-2869.

Dormann, C.F. (2011). "Modelling Species' Distributions." In *Modelling Complex Ecological Dynamics,* 179-196, edited., Springer.

Dormann, C.F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J.R.G. Marquéz, B. Gruber, B. Lafourcade, P.J. Leitão, T. Münkemüller, C. Mcclean, P.E. Osborne, B. Reineking, B. Schröder, A.K. Skidmore, D. Zurell & S. Lautenbach (2012). "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance." *Ecography,* 35, 001-020.

Dormann, C.F., B. Gruber, M. Winter & D. Herrmann (2010). "Evolution of Climate Niches in European Mammals?" *Biology Letters,* 6, 229-232.

Dormann, C.F., J.M. Mcpherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R.G. Davies, A. Hirzel, W. Jetz, W. D. Kissling, I. Kühn, R. Ohlemüller, P.R. Peres-Neto, B. Reineking, B. Schröder, F.M. Schurr & R. Wilson (2007). "Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review." *Ecography,* 30, 609-628.

Duckworth, W.D., H.H. Genoways & C.L. Rose (1993). *Preserving Natural Science Collections: Chronicle of Our Environmental Heritage,* Washington, D.C., National Institute for the Conservation of Cultural Property.

Dungan, J. (1999). "Conditional Simulation: An Alternative to Estimation for Achieving Mapping Objectives." In *Spatial Statistics for Remote*

*Sensing,* 135-152, edited by A. Stein, F. Meer & B. Gorte. Springer Netherlands.

Ebrahimi, N., E. Maasoumi & E.S. Soofi (1999). "Measuring Informativeness of Data by Entropy and Variance." In *Advances in Econometrics, Income Distribution and Scientific Methodology,* 61-77, edited., Springer.

Elith, J. & C.H. Graham (2009). "Do They? How Do They? Why Do They Differ? On Finding Reasons for Differing Performances of Species Distribution Models." *Ecography,* 32, 66-77.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.M. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberon, S. Williams, M.S. Wisz & N.E. Zimmermann (2006). "Novel Methods Improve Prediction of Species' Distributions from Occurrence Data." *Ecography,* 29, 129-151.

Elith, J. & J.R. Leathwick (2009). "Species Distribution Models: Ecological Explanation and Prediction across Space and Time." *Annual Review of Ecology Evolution and Systematics,* 40, 677-697.

Elith, J., J.R. Leathwick & T. Hastie (2008). "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology,* 77, 802-813.

Engler, R., A. Guisan & L. Rechsteiner (2004). "An Improved Approach for Predicting the Distribution of Rare and Endangered Species from Occurrence and Pseudo-Absence Data." *Journal of Applied Ecology,* 41, 263-274.

Farber, O. & R. Kadmon (2003). "Assessment of Alternative Approaches for Bioclimatic Modeling with Special Emphasis on the Mahalanobis Distance." *Ecological Modelling,* 160, 115-130.

Feeley, K.J. & M.R. Silman (2010). "Modelling the Responses of Andean and Amazonian Plant Species to Climate Change: The Effects of Georeferencing Errors and the Importance of Data Filtering." *Journal of Biogeography,* 37, 733-740.

Feldman, D.P. & J.P. Crutchfield (2003). "Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy." *Physical Review E,* 67, 051104.

Fielding, A.H. & J.F. Bell (1997). "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models." *Environmental Conservation,* 24, 38-49.

Finley, A.O. (2011). "Comparing Spatially-Varying Coefficients Models for Analysis of Ecological Data with Non-Stationary and Anisotropic Residual Dependence." *Methods in Ecology and Evolution,* 2, 143-154.

Fisher, P.F. (1999). "Models of Uncertainty in Spatial Data." *Geographical information systems,* 1, 191-205.

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction,* Cambridge, UK, Cambridge University Press.

Freeman, E.A. & G.G. Moisen (2008). "A Comparison of the Performance of Threshold Criteria for Binary Classification in Terms of Predicted Prevalence and Kappa." *Ecological Modelling,* 217, 48-58.

Friedman, J.H. (1991). "Multivariate Adaptive Regression Splines." *Annals of Statistics,* 19, 1-67.

Friedman, J.H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics,* 29, 1189-1232.

Friedman, M. (1937). "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association,* 32, 675-701.

Garcia, R.A., N.D. Burgess, M. Cabeza, C. Rahbek & M. B. Araújo (2012). "Exploring Consensus in 21st Century Projections of Climatically Suitable Areas for African Vertebrates." *Global Change Biology,* 18, 1253-1269.

Gelfand, A.E., H.-J. Kim, C.F. Sirmans & S. Banerjee (2003). "Spatial Modeling with Spatially Varying Coefficient Processes." *Journal of the American Statistical Association,* 98, 387-396.

Getis, A. (2010). "Spatial Autocorrelation." In *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications,* 255-278, edited by M.M. Fischer & A. Getis. Berlin Heidelberg, Springer Verlag.

Getis, A. & J.K. Ord (1992). "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis,* 24, 189-206.

Getis, A. & J.K. Ord (1996). "Local Spatial Statistics: An Overview." In *Spatial Analysis: Modelling in a Gis Environment,* 261-277, edited by P. Longley & M. Batty. New York, John Wiley.

Goodchild, M.F. (1986). *Spatial Autocorrelation,* Norwich, Geo Books.

Graham, C.H., J. Elith, R.J. Hijmans, A. Guisan, A.T. Peterson & B.A. Loiselle (2008). "The Influence of Spatial Errors in Species Occurrence Data Used in Distribution Models." *Journal of Applied Ecology,* 45, 239-247.

Graham, C.H., S. Ferrier, F. Huettman, C. Moritz & A.T. Peterson (2004a). "New Developments in Museum-Based Informatics and Applications in Biodiversity Analysis." *Trends in Ecology & Evolution,* 19, 497-503.

Graham, C.H., S.R. Ron, J.C. Santos, C.J. Schneider & C. Moritz (2004b). "Integrating Phylogenetics and Environmental Niche Models to Explore Speciation Mechanisms in Dendrobatid Frogs." *Evolution,* 58, 1781-1793.

Graham, M.H. (2003). "Confronting Multicollinearity in Ecological Multiple Regression." *Ecology,* 84, 2809-2815.

Gritti, E.S., A. Duputié, F. Massol & I. Chuine (2013). "Estimating Consensus and Associated Uncertainty between Inherently Different Species Distribution Models." *Methods in Ecology and Evolution,* 4, 442-452.

Guisan, A., C.H. Graham, J. Elith & F. Huettmann (2007) Sensitivity of Predictive Species Distribution Models to Change in Grain Size. In, 332-340, edited., Blackwell Publishing.

Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J.M.C. Overton, R. Aspinall & T. Hastie (2006). "Making Better Biogeographical Predictions of Species' Distributions." *Journal of Applied Ecology,* 43, 386-392.

Guisan, A. & W. Thuiller (2005). "Predicting Species Distribution: Offering More Than Simple Habitat Models." *Ecology Letters,* 8, 993-1009.

Guisan, A. & N.E. Zimmermann (2000). "Predictive Habitat Distribution Models in Ecology." *Ecological Modelling,* 135, 147-186.

Guo, D. (2003). "Coordinating Computational and Visual Approaches for Interactive Feature Selection and Multivariate Clustering." *Information Visualization,* 2, 232-246.

Guo, D.S. (2010). "Local Entropy Map: A Nonparametric Approach to Detecting Spatially Varying Multivariate Relationships." *International Journal of Geographical Information Science,* 24, 1367-1389.

Guo, Q., Y. Liu & J. Wieczorek (2008). "Georeferencing Locality Descriptions and Computing Associated Uncertainty Using a Probabilistic Approach." *International Journal of Geographical Information Science,* 22, 1067-1090.

Hamm, N.A.S., P.M. Atkinson & E. Milton (2004) On the Effect of Positional Uncertainty in Field Measurements on the Atmospheric Correction of Remotely Sensed Imagery. In *geoENV IV—Geostatistics for Environmental Applications,* 91-102, edited by X. Sanchez-Vila, J. Carrera & J. J. Gomez-Hernandez. Barcelona, Spain, Springer Netherlands.

Hamm, N., P.M. Atkinson & E.J. Milton (2003) The Combined Effect of Spatial Resolution and Measurement Uncertainty on the Accuracy of Empirical Atmospheric Correction. In *Igarss 2003: Ieee International Geoscience and Remote Sensing Symposium, Vols I - Vii, Proceedings - Learning from Earth's Shapes and Sizes,* 2082-2084, edited. New York, Ieee.

Hamm, N.A.S., P.M. Atkinson & E. J. Milton (2012). "A Per-Pixel, Non-Stationary Mixed Model for Empirical Line Atmospheric Correction in Remote Sensing." *Remote Sensing of Environment,* 124, 666-678.

Haskard, K.A. & R.M. Lark (2009). "Modelling Non-Stationary Variance of Soil Properties by Tempering an Empirical Spectrum." *Geoderma,* 153, 18-28.

Hastie, T. (2011) Gam: Generalized Additive Models. In, edited. R package version 1.04.1 ed.

Hastie, T. & R. Tibshirani (1990). *Generalised Additive Models,* London, Chapman & Hall.

Heikkila, E.J. & L. Hu (2006). "Adjusting Spatial-Entropy Measures for Scale and Resolution Effects." *Environment and Planning B: Planning and Design,* 33, 845-861.

Heikkinen, R. K., M. Luoto, M.B. Araújo, R. Virkkala, W. Thuiller & M. T. Sykes (2006). "Methods and Uncertainties in Bioclimatic Envelope Modelling under Climate Change." *Progress in Physical Geography,* 30, 751-777.

Hernandez, P.A., C.H. Graham, L.L. Master & D.L. Albert (2006). "The Effect of Sample Size and Species Characteristics on Performance of

Different Species Distribution Modeling Methods." *Ecography,* 29, 773-785.

Heuvelink, G.B.M. (1999). "Propagation of Error in Spatial Modelling with Gis." In *Geographical Information Systems,* 207-217, edited by P. Longley, M. Goodchild, D. Maguire & D. Rhind. John Wiley & Sons, Inc.

Heuvelink, G.B.M., J.D. Brown & E.E. Van Loon (2007). "A Probabilistic Framework for Representing and Simulating Uncertain Environmental Variables." *International Journal of Geographical Information Science,* 21, 497-513.

Hijmans, R., S. Cameron, J. Parra, P. Jones, A. Jarvis & K. Richardson (2008) Worldclim Version 1.4. In, edited.

Hijmans, R. & J. Van Etten (2011) Raster: Geographic Analysis and Modeling with Raster Data. In, edited. R package version 1.8.12 ed.

Hijmans, R.J. (2012). "Cross-Validation of Species Distribution Models: Removing Spatial Sorting Bias and Calibration with a Null Model." *Ecology,* 93, 679-688.

Hirzel, A.H., V. Helfer & F. Metral (2001). "Assessing Habitat-Suitability Models with a Virtual Species." *Ecological Modelling,* 145, 111-121.

Hirzel, A.H. & G. Le Lay (2008). "Habitat Suitability Modelling and Niche Theory." *Journal of Applied Ecology,* 45, 1372-1381.

Hudak, M.J. (1992). "Rce Classifiers: Theory and Practice." *Cybernetics and System,* 23, 483-515.

Hurlbert, S.H. (1978). "The Measurement of Niche Overlap and Some Relatives." *Ecology*, 67-77.

Jiménez-Valverde, A. (2012). "Insights into the Area under the Receiver Operating Characteristic Curve (Auc) as a Discrimination Measure in Species Distribution Modelling." *Global Ecology and Biogeography,* 21, 498-507.

Jiménez-Valverde, A. (2014). "Threshold-Dependence as a Desirable Attribute for Discrimination Assessment: Implications for the Evaluation of Species Distribution Models." *Biodiversity and Conservation*, 1-17.

Jiménez-Valverde, A., P. Acevedo, A.M. Barbosa, J.M. Lobo & R. Real (2013). "Discrimination Capacity in Species Distribution Models

Depends on the Representativeness of the Environmental Domain." *Global Ecology and Biogeography,* 22, 508-516.

Jiménez-Valverde, A., J.M. Lobo & J. Hortal (2009). "The Effect of Prevalence and Its Interaction with Sample Size on the Reliability of Species Distribution Models." *Community Ecology,* 10, 196-205.

Johnson, C.J. & M.P. Gillingham (2008). "Sensitivity of Species-Distribution Models to Error, Bias, and Model Design: An Application to Resource Selection Functions for Woodland Caribou." *Ecological Modelling,* 213, 143-155.

Journel, A.G. (1983). "Nonparametric Estimation of Spatial Distributions." *Journal of the International Association for Mathematical Geology,* 15, 445-468.

Journel, A.G. & C.V. Deutsch (1993). "Entropy and Spatial Disorder." *Mathematical Geology,* 25, 329-355.

Lark, R.M. (2009). "Kriging a Soil Variable with a Simple Nonstationary Variance Model." *Journal of Agricultural, Biological, and Environmental Statistics,* 14, 301-321.

Le Lay, G., R. Engler, E. Franc & A. Guisan (2010). "Prospective Sampling Based on Model Ensembles Improves the Detection of Rare Species." *Ecography,* 33, 1015-1027.

Leathwick, J.R., D. Rowe, J. Richardson, J. Elith & T. Hastie (2005). "Using Multivariate Adaptive Regression Splines to Predict the Distributions of New Zealand's Freshwater Diadromous Fish." *Freshwater Biology,* 50, 2034-2052.

Legendre, P. (1993). "Spatial Autocorrelation - Trouble or New Paradigm." *Ecology,* 74, 1659-1673.

Leitão, P.J., F. Moreira & T. J. Osborn (2011). "Effects of Geographical Data Sampling Bias on Habitat Models of Species Distributions: A Case Study with Steppe Birds in Southern Portugal." *International Journal of Geographical Information Science,* 25, 439-453.

Li, H. & J. Wu (2006). "Uncertainty Analysis in Ecological Studies: An Overview." In *Scaling and Uncertainty Analysis in Ecology: Methods and Application,* 45-66, edited by J. Wu, H. Li & O. L. Loucks. The Netherlands, Springer.

Lindley, D.V. (1956). "On a Measure of the Information Provided by an Experiment." *The Annals of Mathematical Statistics,* 27, 986-1005.

Lobo, J.M., A. Jimenez-Valverde & J. Hortal (2010). "The Uncertain Nature of Absences and Their Importance in Species Distribution Modelling." *Ecography,* 33, 103-114.

Lobo, J.M., A. Jimenez-Valverde & R. Real (2008). "Auc: A Misleading Measure of the Performance of Predictive Distribution Models." *Global Ecology and Biogeography,* 17, 145-151.

López-Ruiz, R., H.L. Mancini & X. Calbet (1995). "A Statistical Measure of Complexity." *Physics Letters A,* 209, 321-326.

Manel, S., J.-M. Dias & S.J. Ormerod (1999). "Comparing Discriminant Analysis, Neural Networks and Logistic Regression for Predicting Species Distributions: A Case Study with a Himalayan River Bird." *Ecological Modelling,* 120, 337-347.

Manel, S., H.C. Williams & S.J. Ormerod (2001). "Evaluating Presence-Absence Models in Ecology: The Need to Account for Prevalence." *Journal of Applied Ecology,* 38, 921-931.

Marmion, M., J. Hjort, W. Thuiller & M. Luoto (2009a). "Statistical Consensus Methods for Improving Predictive Geomorphology Maps." *Computers & Geosciences,* 35, 615-625.

Marmion, M., M. Parviainen, M. Luoto, R.K. Heikkinen & W. Thuiller (2009b). "Evaluation of Consensus Methods in Predictive Species Distribution Modelling." *Diversity and Distributions,* 15, 59-69.

Marquardt, D.W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." *Technometrics,* 12, 591-612.

Matilla-García, M. & M.R. Marín (2011). "Spatial Symbolic Entropy: A Tool for Detecting the Order of Contiguity." *Geographical Analysis,* 43, 228-239.

Matilla-García, M., J. R. Ruiz & M.R. Marín (2012). "Detecting the Order of Spatial Dependence Via Symbolic Analysis." *International Journal of Geographical Information Science,* 26, 1015-1029.

Matilla-García, M. & M. Ruiz Marín (2008). "A Non-Parametric Independence Test Using Permutation Entropy." *Journal of Econometrics,* 144, 139-155.

Mccullagh, P. & J.A. Nelder (1989). *Generalized Linear Models,* Second edition ed. London, Chapman & Hall.

Mcpherson, J.M. & W. Jetz (2007). "Effects of Species' Ecology on the Accuracy of Distribution Models." *Ecography,* 30, 135-151.

Meridional, C.L. (2007). "Coronella Girondica." In *Atlas Y Libro Rojo De Los Anfibios Y Reptiles De España,* 275-277, edited by J. M. Pleguezuelos, R. Márquez & M. Lizana. Madrid, DGCN-AHE.

Meynard, C.N. & D.M. Kaplan (2013). "Using Virtual Species to Study Species Distributions and Model Performance." *Journal of Biogeography,* 40, 1-8.

Meynard, C.N. & J.F. Quinn (2007). "Predicting Species Distributions: A Critical Comparison of the Most Common Statistical Models Using Artificial Species." *Journal of Biogeography,* 34, 1455-1469.

Moran, P. a. P. (1948). "The Interpretation of Statistical Maps." *Journal of the Royal Statistical Society. Series B (Methodological),* 10, 243-251.

Munoz, J. & A.M. Felicisimo (2004). "Comparison of Statistical Methods Commonly Used in Predictive Modelling." *Journal of Vegetation Science,* 15, 285-292.

Naimi, B., N.A.S. Hamm, T.A. Groen, A.K. Skidmore & A.G. Toxopeus (2014). "Where Is Positional Uncertainty a Problem for Species Distribution Modelling?" *Ecography,* 37, 191-203.

Naimi, B., A.K. Skidmore, T.A. Groen & N.A.S. Hamm (2011). "Spatial Autocorrelation in Predictors Reduces the Impact of Positional Uncertainty in Occurrence Data on Species Distribution Modelling." *Journal of Biogeography,* 38, 1497-1509.

Nasa Land Processes Distributed Active Archive Center (2011) Modis/Terra. In *2002-2010*, edited. 2011 ed., LP DAAC.

Negro, P. (2007). "Dryocopus Martius." In *Atlas De Las Aves Reproductoras De Españia,* 354-355, edited by R. Marti & J. C. Del Moral. Madrid, Dirección General de Conservación de la Naturaleza – Sociedad Española de Ornitología.

O'neill, R.V., J.R. Krummel, R.H. Gardner, G. Sugihara, B. Jackson, D. L. Deangelis, B.T. Milne, M.G. Turner, B. Zygmunt, S.W. Christensen, V.H. Dale & R.L. Graham (1988). "Indices of Landscape Pattern." *Landscape Ecology,* 1, 153-162.

Ord, J.K. & A. Getis (1995). "Local Spatial Autocorrelation Statistics - Distributional Issues and an Application." *Geographical Analysis,* 27, 286-306.

Ord, J.K. & A. Getis (2001). "Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation." *Journal of Regional Science,* 41, 411-432.

Ortega-Huerta, M.A. & A.T. Peterson (2008). "Modeling Ecological Niches and Predicting Geographic Distributions: A Test of Six Presence-Only Methods." *Revista Mexicana De Biodiversidad,* 79, 205-216.

Osborne, P.E. & P.J. Leitão (2009). "Effects of Species and Habitat Positional Errors on the Performance and Interpretation of Species Distribution Models." *Diversity and Distributions,* 15, 671-681.

Paciorek, C.J. & M.J. Schervish (2006). "Spatial Modelling Using a New Class of Nonstationary Covariance Functions." *Environmetrics,* 17, 483-506.

Pearce, J. & S. Ferrier (2000). "Evaluating the Predictive Performance of Habitat Models Developed Using Logistic Regression." *Ecological Modelling,* 133, 225-245.

Pearson, R.G., C.J. Raxworthy, M. Nakamura & A.T. Peterson (2007). "Predicting Species Distributions from Small Numbers of Occurrence Records: A Test Case Using Cryptic Geckos in Madagascar." *Journal of Biogeography,* 34, 102-117.

Pearson, R.G., W. Thuiller, M.B. Araújo, E. Martinez-Meyer, L. Brotons, C. Mcclean, L. Miles, P. Segurado, T.P. Dawson & D.C. Lees (2006) Model-Based Uncertainty in Species Range Prediction. In, 1704-1711, edited., Blackwell Publishing.

Pebesma, E.J. (2004). "Multivariable Geostatistics in S: The Gstat Package." *Computers & Geosciences,* 30, 683-691.

Peterson, A.T. (2006). "Uses and Requirements of Ecological Niche Models and Related Distributional Models." *Biodiversity Informatics,* 3, 59-72.

Peterson, A.T., M. Papes & M. Eaton (2007). "Transferability and Model Evaluation in Ecological Niche Modeling: A Comparison of Garp and Maxent." *Ecography,* 30, 550-560.

Peterson, A T., J. Soberón, R.G. Pearson, R.P. Anderson, E. Martínez-Meyer, M. Nakamura & M.B. Araújo (2011). *Ecological Niches and Geographic Distributions : E-Book,* Princeton, Princeton University.

Pham, T.D. (2010). "Geoentropy: A Measure of Complexity and Similarity." *Pattern Recognition,* 43, 887-896.

Power, C., A. Simms & R. White (2001). "Hierarchical Fuzzy Pattern Matching for the Regional Comparison of Land Use Maps." *International Journal of Geographical Information Science,* 15, 77-100.

Quester, P. & E. Dion (1997). "Scaling Numerical Variables and Information Loss: An Appraisal of Morrison's Work." *MARKETING BULLETIN-DEPARTMENT OF MARKETING MASSEY UNIVERSITY,* 8, 59-65.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing,* Vienna.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing,* Vienna, Austria.

Refsgaard, J.C., J.P. Van Der Sluijs, A.L.HÖJberg & P.A. Vanrolleghem (2007). "Uncertainty in the Environmental Modelling Process - a Framework and Guidance." *Environmental Modelling & Software,* 22, 1543-1556.

Regan, H.M., M. Colyvan & M.A. Burgman (2002). "A Taxonomy and Treatment of Uncertainty for Ecology and Conservation Biology." *Ecological Applications,* 12, 618-628.

Ricotta, C. & M. Anand (2006). "Spatial Complexity of Ecological Communities: Bridging the Gap between Probabilistic and Non-Probabilistic Uncertainty Measures." *Ecological Modelling,* 197, 59-66.

Robinson, M.D., D.P. Feldman & S.R. Mckay (2011). "Local Entropy and Structure in a Two-Dimensional Frustrated System." *Chaos: An Interdisciplinary Journal of Nonlinear Science,* 21, 037114-037114-11.

Romme, W.H. (1982). "Fire and Landscape Diversity in Subalpine Forests of Yellowstone National Park." *Ecological Monographs,* 52, 199-221.

Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological review,* 65, 386.

Rowe, R.J. (2005). "Elevational Gradient Analyses and the Use of Historical Museum Specimens: A Cautionary Tale." *Journal of Biogeography,* 32, 1883-1897.

Royle, J.A., R.B. Chandler, C. Yackulic & J.D. Nichols (2012). "Likelihood Analysis of Species Occurrence Probability from Presence-Only Data for Modelling Species Distributions." *Methods in Ecology and Evolution,* 3, 545-554.

Ruiz, M., F. López & A. Páez (2010). "Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics." *Journal of Geographical Systems,* 12, 281-309.

Santika, T. & M.F. Hutchinson (2009). "The Effect of Species Response Form on Species Distribution Model Prediction and Inference." *Ecological Modelling,* 220, 2365-2379.

Schabenberger, O. & C.A. Gotway (2005). *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC Press.

Schlather, M. (2009) Randomfields: Simulation and Analysis of Random Fields. In*,* R package version 1.3.41, edited. 1.3.41 ed.

Schwering, A. (2008). "Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey." *Transactions in GIS,* 12, 5-29.

Segurado, P. & M.B. Araújo (2004). "An Evaluation of Methods for Modelling Species Distributions." *Journal of Biogeography,* 31, 1555-1568.

Skidmore, A.K. & … (1996). "Operational Gis Expert System for Mapping Forest Soils." *PE&RS = Photogrammetric Engineering and Remote Sensing,* 62.

Skidmore, A.K., A. Gauld & P. Walker (1996). "Classification of Kangaroo Habitat Distribution Using Three Gis Models." *International Journal of Geographical Information Systems,* 10, 441-454.

Sokal, R. R. & N. L. Oden (1978). "Spatial Autocorrelation in Biology: 1. Methodology." *Biological Journal of the Linnean Society,* 10, 199-228.

Sokal, R.R., N.L. Oden & B.A. Thomson (1998). "Local Spatial Autocorrelation in a Biological Model." *Geographical Analysis,* 30, 331-354.

Stockwell, D.R.B. & I.R. Noble (1992). "Induction of Sets of Rules from Animal Distribution Data: A Robust and Informative Method of Data Analysis." *Mathematics and Computers in Simulation,* 33, 385-390.

Stockwell, D.R.B. & A.T. Peterson (2002). "Effects of Sample Size on Accuracy of Species Distribution Models." *Ecological Modelling,* 148, 1-13.

Thuiller, W., B. Lafourcade, R. Engler & M.B. Araújo (2009). "Biomod - a Platform for Ensemble Forecasting of Species Distributions." *Ecography,* 32, 369-373.

Uuemaa, E., J. Roosaare, A. Kanal & Ü. Mander (2008). "Spatial Correlograms of Soil Cover as an Indicator of Landscape Heterogeneity." *Ecological Indicators,* 8, 783-794.

Walker, W.E., J. Harremoes, J. Rotmans, Van Der Sluijs, J.P., B.A. Van Asselt Marjolein, P. Janssen & M.P.K. Von Krauss (2003). "Defining Uncertainty, a Conceptual Basis for Uncertainty Management in Model-Based Decision Support." *Integrated Assessment,* 4, 5-17.

Warren, D.L., R.E. Glor & M. Turelli (2008). "Environmental Niche Equivalency Versus Conservatism: Quantitative Approaches to Niche Evolution." *Evolution,* 62, 2868-2883.

Warren, D.L. & S.N. Seifert (2011). "Ecological Niche Modeling in Maxent: The Importance of Model Complexity and the Performance of Model Selection Criteria." *Ecological Applications,* 21, 335-342.

Webster, R. & M.A. Oliver (2007). *Geostatistics for Environmental Scientists,* Chichester, John Wiley & Sons, Ltd.

Wieczorek, J., Q.G. Guo & R.J. Hijmans (2004). "The Point-Radius Method for Georeferencing Locality Descriptions and Calculating Associated Uncertainty." *International Journal of Geographical Information Science,* 18, 745-767.

Wiens, J.A., D. Stralberg, D. Jongsomjit, C.A. Howell & M.A. Snyder (2009). "Niches, Models, and Climate Change: Assessing the Assumptions and Uncertainties." *Proceedings of the National Academy of Sciences,* 106, 19729-19736.

Yeung, R.W. (2008). *Information Theory and Network Coding*, Springer.

Youden, W.J. (1950). "Index for Rating Diagnostic Tests." *Cancer,* 3, 32-35.

Zaniewski, A.E., A. Lehmann & J.M.C. Overton (2002). "Predicting Species Spatial Distributions Using Presence-Only Data: A Case Study of Native New Zealand Ferns." *Ecological Modelling,* 157, 261-280.

## Summary

Species distribution models (SDMs), also known as bioclimatic envelope models (BEM), Ecological niche models (ENM), and habitat suitability models (HSM) are widely used to infer the ecological requirements of species and to predict their geographic distributions. These models have become important in a range of applications including regional biodiversity assessment, conservation biology, evolutionary biology, epidemiology, wildlife management, conservation planning, forecasting the effects of climate change on species distributions and on phylogenetic diversity, etc.

Despite the wide use of SDMs, an important challenge about the applicability and validity of these models is that they are subject to uncertainty; an issue which is usually missing from the studies or only partially considered. By identifying the sources of uncertainty and quantifying their level in model predictions, advice and management actions will be better informed than if based on false certainty. It is also important for scientists who focus in advancing niche-based modelling, as it helps to see where the future improvements can be made. Although there is increasing attention to some important aspects of error in recent SDM studies, there is little appreciation for the fact that there are many different dimensions of uncertainty. Moreover, there is a lack of understanding about their different characteristics, relative magnitudes, and available means of dealing with them.

The overall goal of this thesis was to gain insight into the impact of errors and uncertainties on the performance of species distribution modelling. Of the many sources of uncertainty that may affect the modelled species distributions, locational errors in species data and variability in (or inconsistency within) different SDMs' predictions are the major sources. This research focused on these two important sources of uncertainty and explored their potential impact on SDMs and the possible solutions to overcome them. Furthermore, a major focus of this thesis was dedicated to the role of spatial autocorrelation (a property of spatial data).

Our findings showed that SDMs are robust to positional uncertainty only when the level of uncertainty is less than the range of spatial autocorrelation in predictors. Based on this result, we introduced an approach to understand the potential impact of positional uncertainty in species occurrences on SDMs by examining the spatial autocorrelation range in predictor variables. We also showed that using local spatial autocorrelation statistics leads to identification of the specific occurrence records that are problematic as a consequence of positional uncertainty. Of key importance is that an appropriate strategy can be considered to overcome the problem.

We also explored the impact of model uncertainty by measuring within and between model inconsistencies in their predictions using several commonly used modelling approaches. We showed that the ensemble (consensus or multi-model) prediction is a possible solution to reduce model uncertainty. Moreover, we showed that this approach outperforms the other SDMs and increases the accuracy of the predictions. Therefore, our findings suggest that the ensemble approach should be used as an alternative to any single-model SDM.

Our approach to use the spatial autocorrelation as an opportunity to understand the impact of positional uncertainty in species data revealed a limitation in the available spatial autocorrelation statistics. Although several methods exist for this purpose, most of these methods can only be used for continuous or interval variables. For SDM studies as well as many other applications, there are numerous situations where categorical variables are encountered. This concern provided the main motivation for developing a new statistic (ELSA) to measure local spatial association in both continuous and categorical data. This statistic may be used not only for SDM studies, but also in many other disciplines where local spatial autocorrelation needs to be quantified (e.g., characterizing land cover structure, disease clustering, analysing economic data).

# Samenvatting

Soortdistributiemodellen (SDMs), ook wel bekend als bioclimatic envelope modellen (BEM), ecologische niche modellen (ENM) of habitatgeschiktheidsmodellen (HSM), worden veel gebruikt om de ecologische behoeften van soorten te bepalen en om hun geografische verspreiding vast te stellen. Deze modellen zijn belangrijk in tal van toepassingen, zoals biodiversiteitsschattingen, bescherming van soorten, evolutie, epidemiologie, wildbeheer of het inschatten van klimaatseffecten op soortdistributies.

Ondanks het wijdverbreide gebruik van SDMs, blijft onzekerheid van de model uitkomsten een struikelblok in de kwaliteit en dus toepassing van dit soort modellen. Dit aspect wordt in veel studies genegeerd, of slechts ten dele overwogen. Door de bronnen van die onzekerheid te identificeren en de uiteindelijke onzekerheid in de modeluitkomsten te kwantificeren, kunnen advies en beheersmaatregelen die zijn gebaseerd op deze uitkomsten beter worden dan wanneer ze op valse zekerheden zijn gebaseerd. Ook voor wetenschappers die zich bezighouden met het verder ontwikkelen van niche modellen kan deze informatie helpen om te bepalen waar verbeteringen in deze technieken nodig zijn. Hoewel er in toenemende mate aandacht is voor een aantal belangrijke aspecten van fouten in SDM onderzoeken, is er weinig aandacht voor het feit dat er verschillende dimensies van onzekerheid zijn. Bovenal, is er een gebrek aan begrip van de verschillende eigenschappen van onzekerheden, de relatieve ordes van grootte van onzekerheden en de beschikbare hulpmiddelen om met deze onzekerheden om te gaan.

Het hoofddoel van dit proefschrift was om inzicht te krijgen in de invloed van fouten en onzekerheden op de kwaliteit van soortdistributiemodellen. Van de vele bronnen van onzekerheid op de uiteindelijk gemodelleerde verspreiding van een soort zijn positionele fouten en variatie tussen (en binnen) model technieken de grootste. Deze studie richt zich op deze twee bronnen en onderzocht hun invloed op SDMs en wat mogelijke oplossingen zijn. Verder is er veel aandacht besteed in dit proefschrift aan de rol van ruimtelijke autocorrelatie, wat een eigenschap is van ruimtelijke gegevens.

De resultaten laten zien dat SDMs weinig invloed laten zien van positionele fouten wanneer de grootte van deze positionele fout minder is dan de

"range" in ruimtelijke autocorrelatie. Op basis van deze informatie is een benadering geïntroduceerd om de mogelijke invloed van fouten in positie op SDM resultaten in te schatten als functie van ruimtelijke autocorrelatie. Daarnaast heeft het onderzoek laten zien dat we specifieke punten waar soorten zijn waargenomen kunnen identificeren als problematisch door hun positionele fout te vergelijken met lokale ruimtelijke autocorrelatie. Belangrijk is dat hiermee een duidelijke benadering kan worden toegepast om met positionele fouten om te gaan.

De invloed van model-onzekerheid voor verschillende veel gebruikte modeltechnieken is ook onderzocht door verschillen binnen en tussen modellen met elkaar te vergelijken. Dit proefschrift laat zien dat een ensemble van model resultaten (op basis van consensus or door te combineren) een mogelijke oplossing is om modelonzekerheid te reduceren. Bovendien overtreft deze benadering andere SDM technieken, en verhoogt het de nauwkeurigheid van de uitkomsten. Deze resultaten suggereren dat een ensemble benadering beter gebruikt kan worden dan enkele model benaderingen.

De benadering om ruimtelijke autocorrelatie te gebruiken om positionele onzekerheden te begrijpen liep tegen een tekortkoming van bestaande ruimtelijke autocorrelatie statistieken op. Hoewel er al verschillende methoden bestaan om autocorrelatie te berekenen, kunnen de meeste methodes alleen op continue waardes, of interval waardes worden toegepast. Voor SDMs, maar ook voor vele andere toepassingen, worden vaak ook categorische variabelen gebruikt. Daarom werd in dit proefschrift een nieuwe statistiek gepresenteerd (ELSA) die lokale ruimtelijke autocorrelatie kan berekenen voor zowel continue als categorische gegevens. Deze statistiek is niet alleen nuttig voor SDM onderzoek, maar kan ook gebruikt worden in andere disciplines waar autocorrelatie van belang is (b.v. voor het bepalen van landgebruiksstructuren, clustering van ziekten, het analyseren van economische gegevens).

# ITC Dissertation List

http://www.itc.nl/research/phd/phd_graduates.aspx