Research

Ecography 37: 191–203, 2014
doi: 10.1111/j.1600-0587.2013.00205.x

Subject Editor: Robert P. Anderson. Accepted 21 June 2013

Where is positional uncertainty a problem for species distribution modelling?

Babak Naimi, Nicholas A. S. Hamm, Thomas A. Groen, Andrew K. Skidmore and Albertus G. Toxopeus

B. Naimi (naimi@itc.nl), N. A. S. Hamm, T. A. Groen, A. K. Skidmore and A. G. Toxopeus, Faculty of Geo-Information Science and Earth Observation (ITC), Univ. of Twente, PO Box 217, NL-7500 AE, Enschede, the Netherlands. BN also at: Dept of Environment and Energy, Science and Research Branch, Islamic Azad Univ., Tehran, Iran.

Species data held in museum and herbaria, survey data and opportunistically observed data are a substantial information resource. A key challenge in using these data is the uncertainty about where an observation is located. This is important when the data are used for species distribution modelling (SDM), because the coordinates are used to extract the environmental variables and thus, positional error may lead to inaccurate estimation of the species–environment relationship. The magnitude of this effect is related to the level of spatial autocorrelation in the environmental variables. Using local spatial association can be relevant because it can lead to the identification of the specific occurrence records that cause the largest drop in SDM accuracy. Therefore, in this study, we tested whether the SDM predictions are more affected by positional uncertainty originating from locations that have lower local spatial association in their predictors. We performed this experiment for Spain and the Netherlands, using simulated datasets derived from well known species distribution models (SDMs). We used the K statistic to quantify the local spatial association in the predictors at each species occurrence location. A probabilistic approach using Monte Carlo simulations was employed to introduce the error in the species locations. The results revealed that positional uncertainty in species occurrence data at locations with low local spatial association in predictors reduced the prediction accuracy of the SDMs. We propose that local spatial association is a way to identify the species occurrence records that require treatment for positional uncertainty. We also developed and present a tool in the R environment to target observations that are likely to create error in the output from SDMs as a result of positional uncertainty.

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 (Skidmore et al. 1996, Araújo and New 2006, Elith et al. 2006, Franklin 2010). 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. 2004, 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 (Wieczorek et al. 2004, Guo et al. 2008), however, this reduces the sample size, which in turn is one of the factors that reduces model accuracy (Hernandez et al. 2006, Graham et al. 2008). 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 (up to 5 km) 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. 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. Naimi et al. (2011) conducted a comprehensive set of analyses to assess the interaction between spatial autocorrelation in predictors and positional uncertainty in species occurrence. Using artificial data they analyzed the influence of five positional uncertainty scenarios on the prediction accuracy of seven frequently applied SDMs. They 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 (Naimi et al. 2011). 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) to explore whether observations with positional uncertainty are likely to be creating error in the output from SDMs.

Material and methods

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 10×10 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 yr 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; Supplementary material Appendix 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 values (greater than 10) one by one using a stepwise procedure. We repeated this procedure until all strongly correlated variables (i.e. with $\text{VIF} > 10$) were excluded.

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 (Hirzel et al. 2001, Austin et al. 2006, Jiménez-Valverde et al. 2009, 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. 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 × 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 (Zaniewski et al. 2002, Engler et al. 2004), 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 2006). The higher accuracy of ensemble modelling to predict habitat suitability has been shown by several studies (Marmion et al. 2009, Le Lay et al. 2010). 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 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.

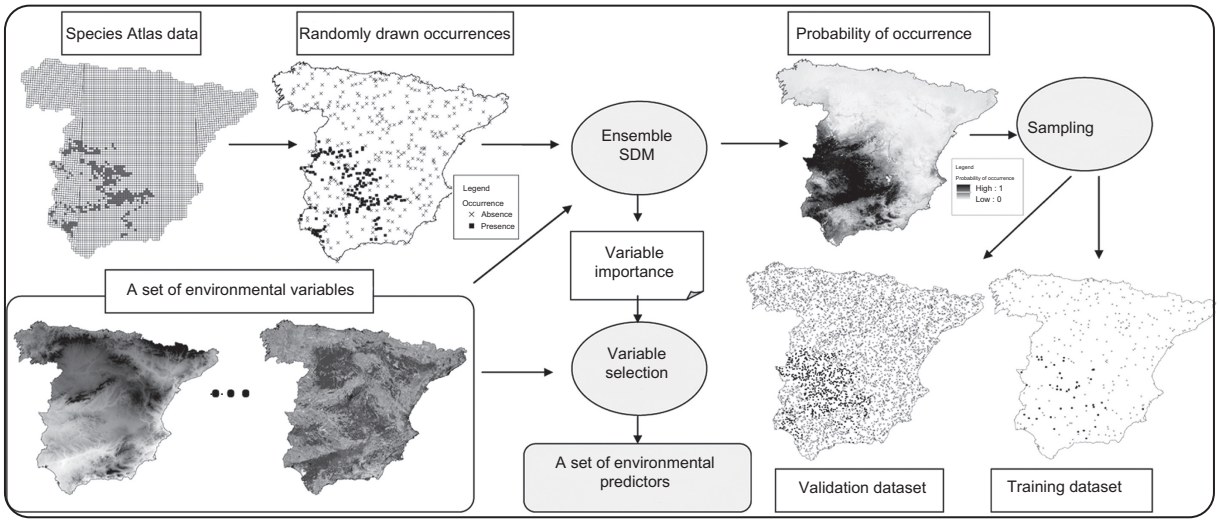

Figure 1. Flow diagram showing the procedure of generating semi-artificial datasets.

Species distribution modelling

Six commonly used SDMs that require presence–absence or presence-only records of species occurrences were selected (Table 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

Table 1. The details and settings of model implementation.

Model	Acronym	Data	Specifics and settings	Reference for more explanation
Generalized linear models	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 ver. 2.13.1. For simplicity we refer to this as 'GLM'.	(McCullagh and Nelder 1989, Austin 2002, R Development Core Team)
Generalized additive models	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 ver. 1.04.1.	(Hastie and Tibshirani 1990, Austin 2002, Hastie 2011)
Boosted regression trees	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 ver. 2.0-4. The function models binary data using a logit link function that uses a cross-validation to choose how many trees to add, stopping before it is too overfit. The optimal number of trees with the learning rate of 0.001 and tree complexity of 3 was used for each species.	(Friedman 2001, Elith et al. 2008)
Random forests	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 ver. 4.6-2 by setting the number of trees to 3000 to ensure that there is enough trees and every input row gets predicted several times. The default settings were used for the rest of parameters.	(Breiman 2001, Liaw and Wiener 2002, Cutler et al. 2007)
Support vector machine	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 ver. 0.9-14 to fit the SVM regression using ANOVA RBF kernel which typically performs well in regression problems.	(Vapnik 1995, Karatzoglou et al. 2006)
Maximum entropy	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.

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).

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 (ε) 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):

$$E_i^* = E_i + \varepsilon_{E_i}$$

$$N_i^* = N_i + \varepsilon_{N_i}$$

where i refers to each individual species occurrence, E and N refer to the true easting and northing and the asteric (*) indicates the perturbed location. 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. 2). The standard deviation of the model performances were also used to assess the stability of the resulting models.

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 (Ord and Getis 1995, 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 (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; Supplementary material Apendix 1, Eq. 2) 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

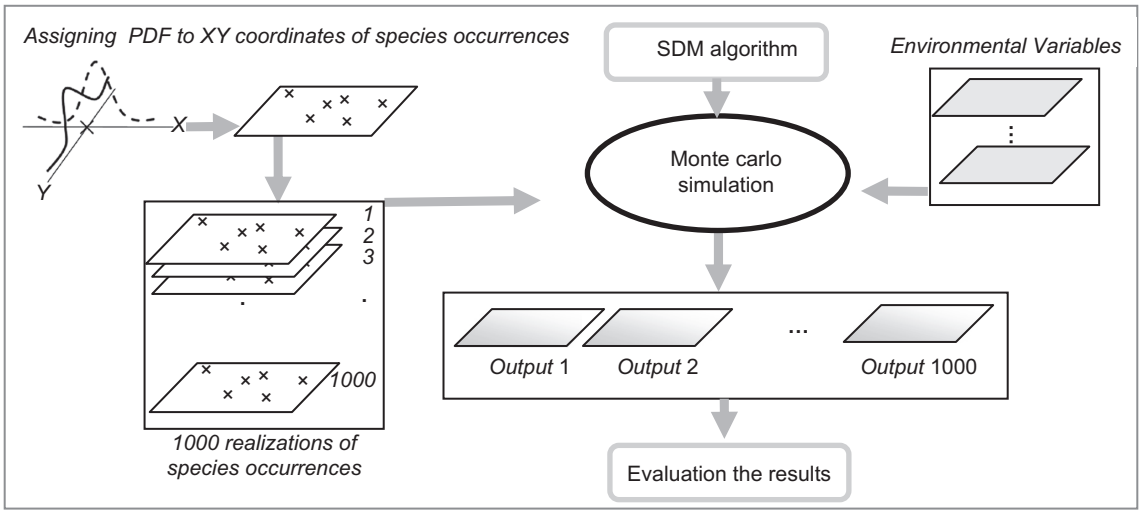

Figure 2. Flowchart showing the procedure of positional uncertainty assessment. PDF, probability density function; SDM, species distribution model.

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 (1 – correlation), as discussed in the section on generating an artificial dataset.

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.

Analysis and implementations in R

We implemented a package (usdm) in R (R Development Core Team) 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 (the instruction for installing the package and some examples are provided in Supplementary material Appendix 1). 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 ver. 2.13.1 (R Development Core Team) using different add-on packages (Table 1). Maxent was run by using the Maxent software ver. 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.

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 2).

The summary of the K statistic values at the locations of the species occurrences for the five case studies (Table 3) show that the level of local spatial association is, for both areas, high. The local spatial association in Spain was sub-

Table 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*)	*esp1*	*esp2*	*esp3*	*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		

Table 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

stantially greater than in the Netherlands (Fig. 3). 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.

Comparing the K statistics for different case studies (Table 3) and the summary statistics for AUC values from the Monte Carlo simulation for different species (Table 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 here (Fig. 4–5) and the results for the other species are provided in Supplementary material Appendix 3, Fig. A3–A5.

Comparing the scenarios for the models in *nl1* (Fig. 4 and Supplementary material Appendix 3, Fig. A1) 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 (Fig. 5 and Supplementary material Appendix 3, Fig. A2) 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 (Table 5) 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.

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 Table 4 and Fig. 4–5 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

Table 4. Summary statistics for the performance measures of the SDMs for all species and different scenarios.

Species	Model	Unperturbed data	The sample size scenarios											
			Equal (size = 100)				0.2% of the total pixels				1% of the total pixels			
			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 oeconomus (*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
Coronella 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

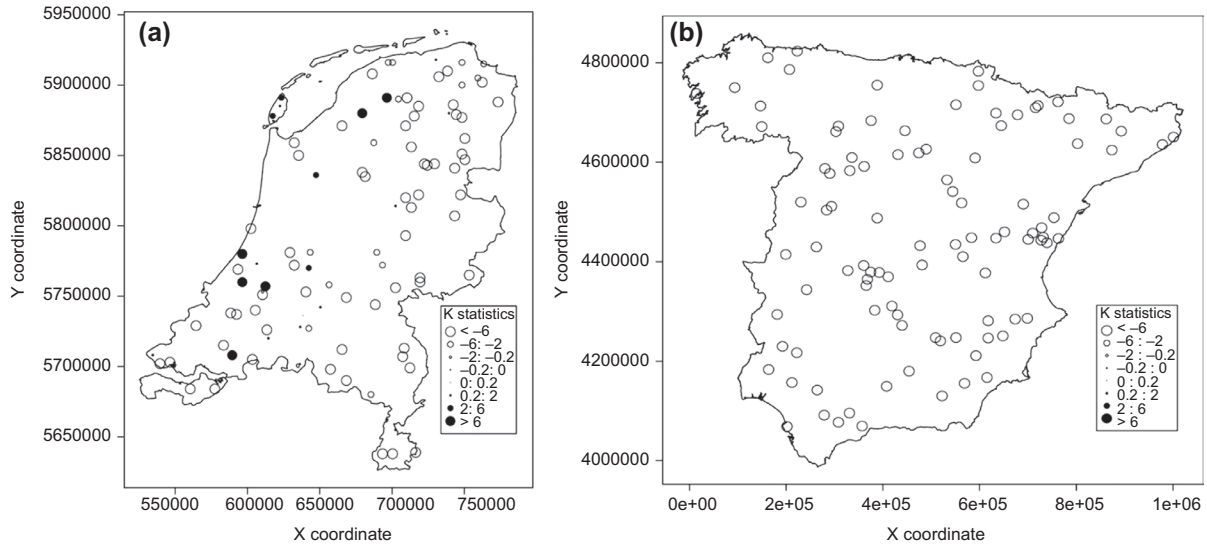

Figure 3. The level of local spatial association at the location of species occurrences, lower K indicates higher local spatial association for the case studies with the lowest (a) and highest (b) local spatial autocorrelation in predictors at species sample locations (i.e. *nl1* and *es1*, respectively).

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

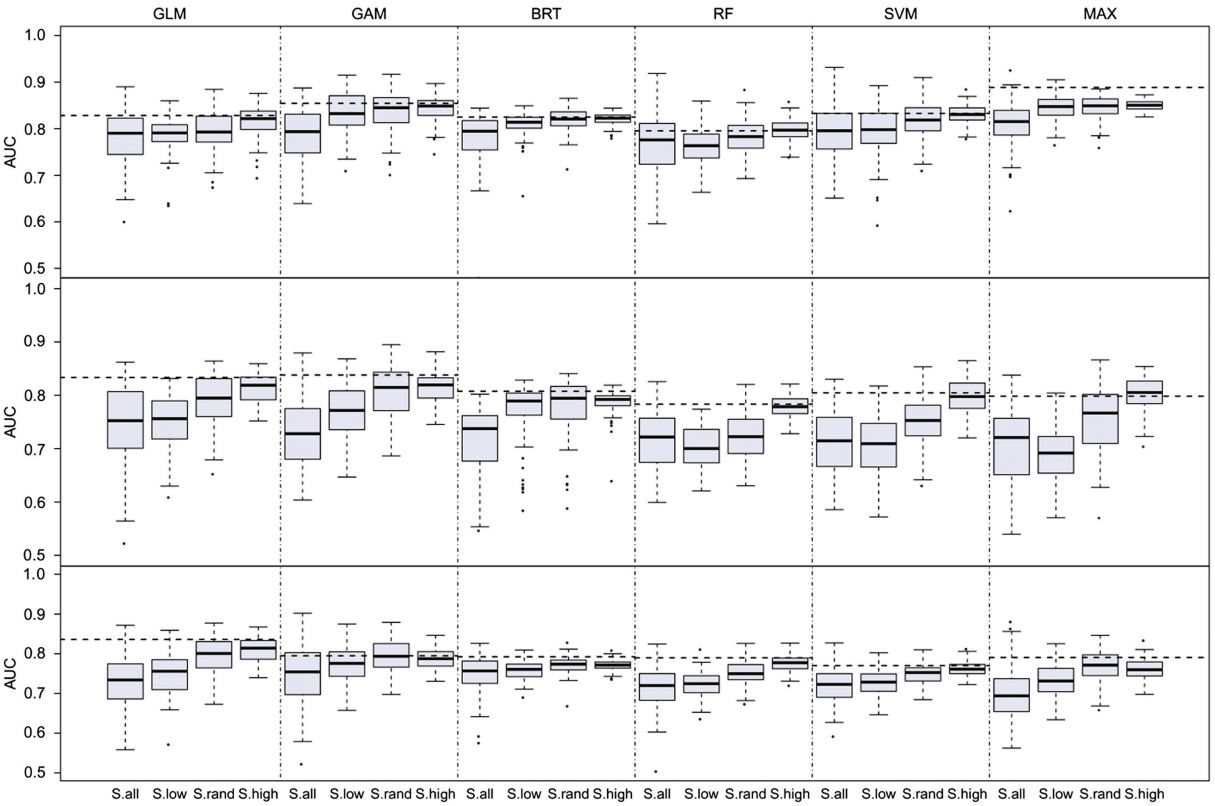

Figure 4. The interaction of the local spatial association and positional uncertainty for the case study with the lowest local spatial association in predictors at species sample locations (i.e. *nl1*). Boxes represent variation of the model accuracy (AUC) over the Monte Carlo simulation for different 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; 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.

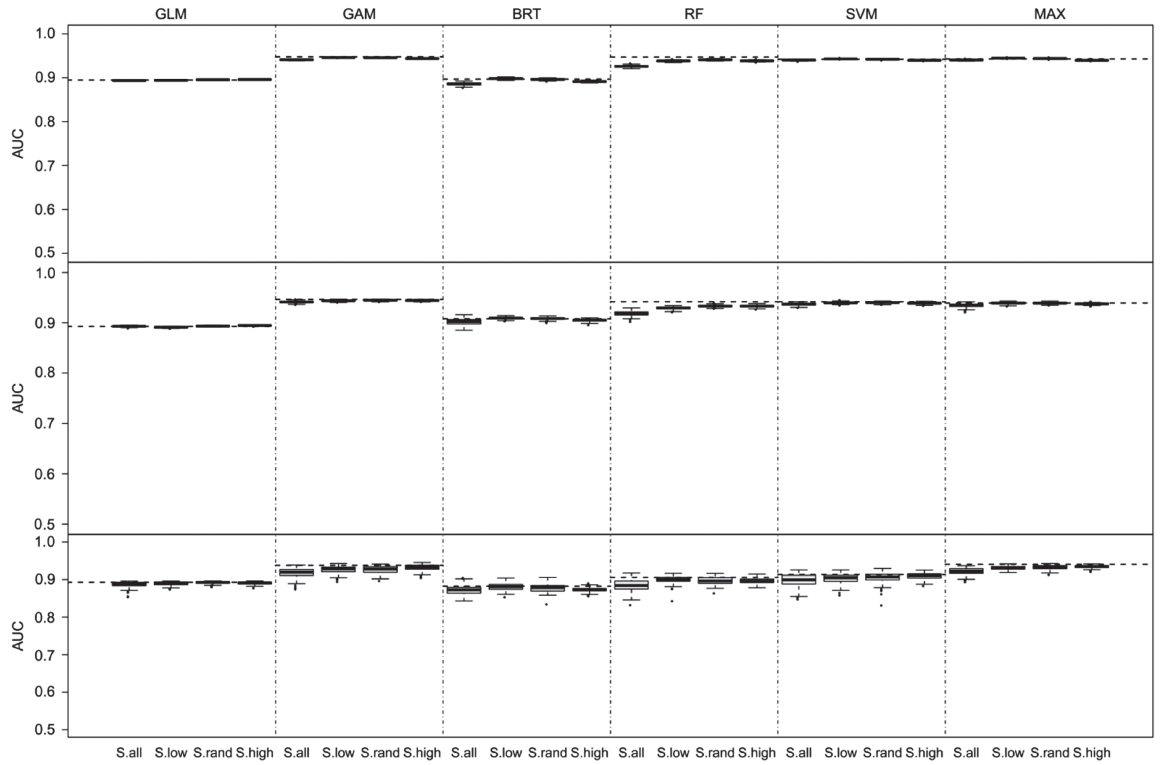

Figure 5. The interaction of the local spatial association and positional uncertainty for the case study with the highest local spatial association at species sample locations (i.e. *es1*). The different components of the graph are described in Fig. 4.

may be used as another solution to modify sampling schemes (Niamir et al. 2011) for the locations that are targetted.

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

Table 5. The level of significance for AUC mean comparison between different scenarios (S.all, S.low, S.rand, and S.high) and different sample size (x-axis) for different SDMs.

			GLM			GAM			BRT		
			S.low	S.rand	S.high	S.low	S.rand	S.high	S.low	S.rand	S.high
es1	equal	S.all	0.003	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		$p<0.001$	0.067		0.868	$p<0.001$		0.001	0.212
		S.rand			0.348			$p<0.001$			$p<0.001$
	0.2 %	S.all	0.320	1	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		0.320	$p<0.001$		$p<0.001$	0.029		0.012	$p<0.001$
		S.rand			0.113			0.04			$p<0.001$
	1%	S.all	1	$p<0.001$	$p<0.001$	0.559	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$
		S.rand			1			$p<0.001$			$p<0.001$
nl1	equal	S.all	0.052	$p<0.001$	$p<0.001$	0.003	$p<0.001$	$p<0.001$	0.061	$p<0.001$	$p<0.001$
		S.low		$p<0.001$	$p<0.001$		$p<0.001$	0.005		$p<0.001$	$p<0.001$
		S.rand			0.004			0.189			0.848
	0.2 %	S.all	0.747	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$		0.595	0.017
		S.rand			$p<0.001$			0.15			0.098
	1%	S.all	0.568	0.074	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		0.093	$p<0.001$		0.675	0.101		0.012	$p<0.001$
		S.rand			$p<0.001$			0.239			0.251

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; Manel et al. 2001, Hernandez et al. 2006). 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, Meynard and Quinn 2007, Jiménez-Valverde et al. 2009, 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 2006) 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)

Table 5. Continued.

			RF			SVM			MAX		
			S.low	S.rand	S.high	S.low	S.rand	S.high	S.low	S.rand	S.high
es1	equal	S.all	$p<0.001$	$p<0.001$	$p<0.001$	0.011	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		0.018	0.06		0.338	$p<0.001$		0.019	$p<0.001$
		S.rand			0.481			$p<0.001$			0.100
	0.2 %	S.all	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		$p<0.001$	$p<0.001$		0.32	0.178		0.979	0.066
		S.rand			0.905			0.045			0.066
	1%	S.all	$p<0.001$	$p<0.001$	$p<0.001$	1	1	1	$p<0.001$	$p<0.001$	$p<0.001$
		S.low		0.045	0.412		1	1		0.221	0.178
		S.rand			0.158			1			0.979
nl1	equal	S.all	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		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$
		S.rand			$p<0.001$			$p<0.001$			0.056
	0.2 %	S.all	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		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$
		S.rand			$p<0.001$			$p<0.001$			$p<0.001$
	1%	S.all	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		$p<0.001$	$p<0.001$		$p<0.001$	$p<0.001$		0.512	0.048
		S.rand			0.001			$p<0.001$			0.164

is necessary when the aim is to explore model sensitivity to a property in environmental data (e.g. as in Naimi et al. 2011); 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 (e.g. as in Santika and Hutchinson 2009) or to test a certain ecological theory.

Finally, we developed a package in the R environment for statistical computing (R Development Core Team) 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 (Supplementary material Appendix 2 for details on the package and examples). 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.

Conclusion

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 to examine 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.

Acknowledgements – We would like to thank the editor for valuable comments on an earlier draft of this manuscript. We are grateful to support of the European Union, Erasmus Mundus programme External Cooperation Windows (2007 /1139/001-001 MUN ECW). We thank Dutch mammal society (VZZ) for sharing the species occurrence data collection in the Netherlands.

References

Anselin, L. 1995. Local indicators of spatial association – LISA. – Geogr. Anal. 27: 93–115.

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Araújo, M. B. and New, M. 2006. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Austin, M. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Model. 157: 101–118.

Austin, M. P. et al. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. – Ecol. Model. 199: 197–216.

Beven, K. and Kirkby, M. 1979. A physically based, variable contributing area model of basin hydrology. – Hydrol. Sci. J. 24: 43–69.

Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.

Buermann, W. et al. 2008. Predicting species distributions across the Amazonian and Andean regions using remote sensing data. – J. Biogeogr. 35: 1160–1176.

Bystriakova, N. et al. 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. – Syst. Biodivers. 10: 305–315.

Chapman, A. D. 2005. Uses of primary species occurrence data, version 1.0. – Report for the Global Biodiversity Information Facility, Copenhagen.

Chatterjee, S. and Hadi, A. S. 2006. Regression analysis by example. – Wiley.

Cliff, A. D. and Ord, J. K. 1981. Spatial processes: models and applications. – Pion.

Cressie, N. 1993. Statistics for spatial data. – Wiley.

Cutler, D. R. et al. 2007. Random forests for classification in ecology. – Ecology 88: 2783–2792.

de Cabrera, T. 2007. *Microtus cabrerae*. – In: Marti, R. and Del Moral, J. C. (eds), Atlas de las aves reproductoras de Españia. SECEM-SECEMU, pp. 429–431.

Dormann, C. F. et al. 2012. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. – Ecography 35: 1–20.

Duckworth, W. D. et al. 1993. Preserving natural science collections: chronicle of our environmental heritage. – National Inst. for the Conservation of Cultural Property.

Elith, J. and Graham, C. H. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Elith, J. et al. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – J. Appl. Ecol. 41: 263–274.

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. – Ecol. Model. 160: 115–130.

Feeley, K. J. and Silman, M. R. 2010. Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. – J. Biogeogr. 37: 733–740.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. – Cambridge Univ. Press.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. – Ann. Stat. 29: 1189–1232.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. – J. Am. Stat. Assoc. 32: 675–701.
Getis, A. 2010. Spatial autocorrelation. – In: Fischer, M. M. and Getis, A. (eds), Handbook of applied spatial analysis: software tools, methods and applications. Springer, pp. 255–278.
Getis, A. and Ord, J. K. 1992. The analysis of spatial association by use of distance statistics. – Geogr. Anal. 24: 189–206.
Getis, A. and Ord, J. K. 1996. Local spatial statistics: an overview. – In: Longley, P. and Batty, M. (eds), Spatial analysis: modelling in a GIS environment. Wiley, pp. 261–277.
Goodchild, M. F. 1986. Spatial autocorrelation. – Geo Books.
Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.
Graham, C. H. et al. 2008. The influence of spatial errors in species occurrence data used in distribution models. – J. Appl. Ecol. 45: 239–247.
Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. – Ecology 84: 2809–2815.
Guisan, A. et al. 2007. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? – Ecol. Monogr. 77: 615–630.
Guo, Q. et al. 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. – Int. J. Geogr. Inform. Sci. 22: 1067–1090.
Hamm, N. et al. 2004. On the effect of positional uncertainty in field measurements on the atmospheric correction of remotely sensed imagery. – In: Sanchez-Vila, X. et al. (eds), geoENV IV – geostatistics for environmental applications. Springer, pp. 91–102.
Hamm, N. A. S. et al. 2012. A per-pixel, non-stationary mixed model for empirical line atmospheric correction in remote sensing. – Remote Sens. Environ. 124: 666–678.
Hastie, T. 2011. GAM: generalized additive models.
Hastie, T. and Tibshirani, R. 1990. Generalised additive models. – Chapman and Hall.
Hernandez, P. A. et al. 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: Longley, P. et al. (eds), Geographical information systems. Wiley, pp. 207–217.
Hijmans, R. and van Etten, J. 2011. Raster: geographic analysis and modeling with raster data.
Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Model. 145: 111–121.
Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Community Ecol. 10: 196–205.
Karatzoglou, A. et al. 2006. Support vector machines in R. – J. Stat. Softw. 15: 1–28.
Le Lay, G. et al. 2010. Prospective sampling based on model ensembles improves the detection of rare species. – Ecography 33: 1015–1027.
Legendre, P. 1993. Spatial autocorrelation – trouble or new paradigm. – Ecology 74: 1659–1673.
Leitão, P. J. et al. 2011. Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. – Int. J. Geogr. Inform. Sci. 25: 439–453.
Liaw, A. and Wiener, M. 2002. Classification and regression by random forest. – R news 2: 18–22.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.
Manel, S. et al. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. – J. Appl. Ecol. 38: 921–931.
Marmion, M. et al. 2009. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. – Ecol. Model. 220: 3512–3520.
Marquardt, D. W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. – Technometrics 12: 591–612.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. – Ecography 30: 135–151.
Meridional, C. l. 2007. *Coronella girondica*. – In: Pleguezuelos, J. M. et al. (eds), Atlas y libro rojo de los anfibios y reptiles de España. DGCN-AHE, pp. 275–277.
Meynard, C. N. and Quinn, J. F. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. – J. Biogeogr. 34: 1455–1469.
Naimi, B. et al. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. – J. Biogeogr. 38: 1497–1509.
NASA Land Processes Distributed Active Archive Center 2011. Modis/Terra. 2002–2010. – LP DAAC.
Negro, P. 2007. *Dryocopus martius*. – In: Marti, R. and Del Moral, J. C. (eds), Atlas de las aves reproductoras de Españia. Dirección General de Conservación de la Naturaleza – Sociedad Española de Ornitología, pp. 354–355.
Niamir, A. et al. 2011. Finessing atlas data for species distribution models. – Divers. Distrib. 17: 1173–1185.
Ord, J. K. and Getis, A. 1995. Local spatial autocorrelation statistics – distributional issues and an application. – Geogr. Anal. 27: 286–306.
Ord, J. K. and Getis, A. 2001. Testing for local spatial autocorrelation in the presence of global autocorrelation. – J. Reg. Sci. 41: 411–432.
Osborne, P. E. and Leitão, P. J. 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. – Divers. Distrib. 15: 671–681.
Peterson, A. T. et al. 2011. Ecological niches and geographic distributions: E-book. – Princeton Univ.
Phillips, S. J. and Dudik, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Santika, T. and Hutchinson, M. F. 2009. The effect of species response form on species distribution model prediction and inference. – Ecol. Model. 220: 2365–2379.
Skidmore, A. K. et al. 1996. Classification of kangaroo habitat distribution using three GIS models. – Int. J. Geogr. Inform. Syst. 10: 441–454.
Thuiller, W. et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.
Vapnik, V. 1995. The nature of statistical learning theory. – Springer.
Wieczorek, J. et al. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. – Int. J. Geogr. Inform. Sci. 18: 745–767.
Zaniewski, A. E. et al. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. – Ecol. Model. 157: 261–280.

Supplementary material (Appendix ECOG-00205 at < www.oikosoffice.lu.se/appendix). Appendix 1–3.