ECOGRAPHY

Research

Disentangling drivers of spatial autocorrelation in species distribution models

Konrad P. Mielke, Tom Claassen, Michela Busana, Tom Heskes, Mark A. J. Huijbregts, Kees Koffijberg and Aafke M. Schipper

K. P. Mielke (https://orcid.org/0000-0003-1593-428X) ✉ (k.mielke@science.ru.nl), T. Claassen and T. Heskes, Dept of Data Science, Inst. for Computing and Information Sciences, Radboud Univ. Nijmegen, Nijmegen, the Netherlands. – M. Busana, M. A. J. Huijbregts, A. M. Schipper and KPM, Dept of Environmental Science, Inst. for Water and Wetland Research, Radboud Univ. Nijmegen, Nijmegen, the Netherlands. AMS also at: PBL Netherlands Environmental Assessment Agency, The Hague, the Netherlands. – K. Koffijberg, Sovon Dutch Centre for Field Ornithology, Nijmegen, the Netherlands.

Ecography
43: 1741–1751, 2020
doi: 10.1111/ecog.05134

Subject Editor: Carsten Dormann
Editor-in-Chief: Miguel Araújo
Accepted 16 July 2020

Species distribution models (SDMs) are frequently used to understand the influence of site properties on species occurrence. For robust model inference, SDMs need to account for the spatial autocorrelation of virtually all species occurrence data. Current methods do not routinely distinguish between extrinsic and intrinsic drivers of spatial autocorrelation, although these may have different implications for conservation.

Here, we present and test a method that disentangles extrinsic and intrinsic drivers of spatial autocorrelation using repeated observations of a species. We focus on unknown habitat characteristics and conspecific interactions as extrinsic and intrinsic drivers, respectively. We model the former with spatially correlated random effects and the latter with an autocovariate, such that the spatially correlated random effects are constant across the repeated observations whereas the autocovariate may change. We tested the performance of our model on virtual species data and applied it to observations of the corncrake *Crex crex* in the Netherlands.

Applying our model to virtual species data revealed that it was well able to distinguish between the two different drivers of spatial autocorrelation, outperforming models with no or a single component for spatial autocorrelation. This finding was independent of the direction of the conspecific interactions (i.e. conspecific attraction versus competitive exclusion). The simulations confirmed that the ability of our model to disentangle both drivers of autocorrelation depends on repeated observations. In the case study, we discovered that the corncrake has a stronger response to habitat characteristics compared to a model that did not include spatially correlated random effects, whereas conspecific interactions appeared to be less important. This implies that future conservation efforts should primarily focus on maximizing habitat availability.

Our study shows how to systematically disentangle extrinsic and intrinsic drivers of spatial autocorrelation. The method we propose can help to correctly identify the main drivers of species distributions.

Keywords: autologistic regression, conspecific interaction, longitudinal measurements, spatial autocorrelation, spatially correlated random effects, species distribution model

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Species distribution modelling is increasingly used in various fields, including macro-ecology, biogeography, wildlife management and conservation planning (Guisan and Thuiller 2005, Franklin 2010). A species distribution model (SDM) is a quantitative relationship between the occurrence of a species and a set of environmental characteristics. Important applications of SDMs are predicting current and future ranges of species and identifying and understanding site properties that influence species distributions (Elith and Leathwick 2009, Araújo et al. 2019).

Modelling approaches employed for species distribution modelling need to account for the fact that ecological data are typically characterized by spatial autocorrelation (Dormann et al. 2007). When this spatial structure is not adequately accounted for in the setup of an SDM, model residuals typically exhibit either positive or negative spatial autocorrelation (Cliff and Ord 1970, Schröder and Seppelt 2006). Residual spatial autocorrelation can be caused by various factors or processes not accounted for in the modelling, which may include extrinsic processes, such as unknown habitat characteristics (Legendre 1993), intrinsic processes, such as conspecific interactions (Wintle and Bardos 2006), or a combination of the two. Failure to adequately account for spatial autocorrelation in the modelling of species distributions is known to lead to unreliable predictions of species ranges (Campomizzi et al. 2008, Wisz et al. 2013, Guélat and Kéry 2018), wrong conclusions about preferred site properties (Kühn 2007) and underestimated uncertainty of model parameters (Teng et al. 2018). For example, conspecific attraction may lead to clustered distributions of wildlife within the available habitat, which in turn may result in biased estimates of habitat preferences as well as increased type 1 errors, i.e. concluding a habitat factor is relevant when it is not (Dormann 2007).

Disentangling extrinsic and intrinsic drivers of spatial autocorrelation is challenging, as they may result in similar patterns of positive or negative clustering of individuals (Verhoef et al. 2018) that cannot be explained by the known habitat characteristics. Hence, spatial autocorrelation is commonly tackled with a single overarching method (Dormann et al. 2007). Popular methods used to account for spatial autocorrelation in SDMs include generalized linear models (GLMs) with an autocovariate and generalized linear mixed models (GLMMs) with spatially correlated random effects (Miller 2014). Combining the different drivers of spatial autocorrelation, however, ignores the fact that they may have different implications for conservation and modelling design (Teng et al. 2018). For example, if conspecific attraction is a major driver of the clustered distribution of a species, conservation measures should focus on the preservation of sufficiently large contiguous habitat for that species (Schipper et al. 2011).

In this paper, we present a new modelling approach that is designed to disentangle extrinsic and intrinsic drivers of spatial autocorrelation in binary species distribution models. We focus on unknown habitat characteristics as an example of extrinsically caused spatial autocorrelation and conspecific interaction, which is often ignored in SDMs (Campomizzi et al. 2008), as an example of intrinsically caused spatial autocorrelation. We propose a mixed effect logistic regression model that includes 1) an autocovariate to handle spatial autocorrelation caused by conspecific interactions and 2) spatially correlated random effects designed to handle spatial autocorrelation caused by unknown habitat characteristics. Our method relies on repeated observations of a species. We assume that potential changes in the locations of individuals between sampling events, e.g. corresponding to observations in different years, allow the model to identify the extent to which the spatial autocorrelation is caused by the environment, i.e. the selected habitat is similar but distances between conspecifics vary across sampling events, or by conspecific interactions, i.e. the selected habitat varies but distances between conspecifics are similar across sampling events (Fig. 1).

We tested the performance of our model using simulated records of a virtual species and compared it to the performance of three other models, two of which are commonly used to address spatial autocorrelation. We also investigated whether the performance of our model depends on the direction of the conspecific interactions, i.e. conspecific attraction versus conspecific exclusion. Furthermore, we verified our premise that repeated observations of a species are required for our model to disentangle the two drivers of spatial autocorrelation.

We then applied our model, as well as the three alternative approaches, to disentangle the drivers of spatial autocorrelation in the distribution of the corncrake *Crex crex* in the Netherlands. Populations of this migratory bird have been strongly declining in most western European countries (Koffijberg et al. 2016). The corncrake is a species with specific habitat requirements, but males are also known to form clusters when advertising for females (Schäffer and Koffijberg 2004). Knowledge of both habitat preferences and conspecific interactions of the species is considered vital to designing informed conservation measures for this species (Schipper et al. 2011).

Material and methods

Models

Model approach

We model the occurrence of a species as a function of various environmental site properties, using observations of the species from multiple sampling events, e.g. observations in different years. We include two components to account for spatial autocorrelation: spatially correlated random effects $\hat{\varepsilon}_i$ (Paulitz et al. 2003, Rhodes et al. 2009) with i denoting the site, and an additional explanatory variable $c_{t,i}$ commonly known as autocovariate (Crase et al. 2014, Wang et al. 2018), with t denoting the sampling event. In our approach, the key

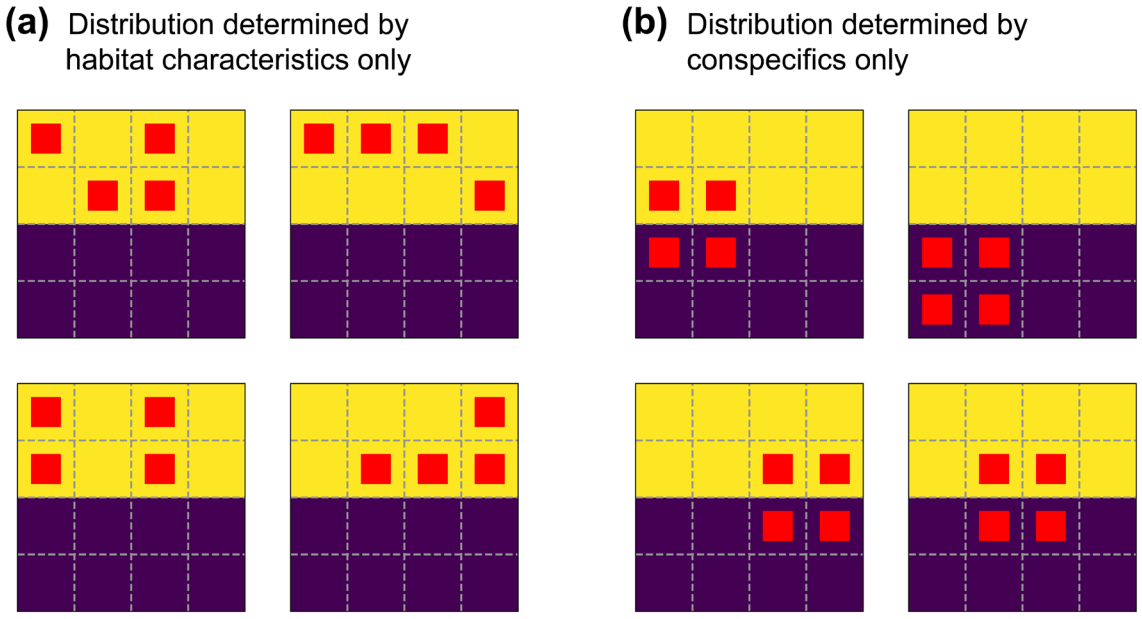

Figure 1. Conceptual representation of how repeated sampling events (e.g. observations of the same species in different years) help to disentangle different drivers of spatial autocorrelation. The figure shows individuals of a hypothetical species (in red) in an environment consisting of two different habitat types (dark blue and yellow) in two extreme scenarios: (a) the species is always in the same habitat yet with varying distances among individuals (indicating that site selection is only determined by habitat characteristics) and (b) the species is observed in different habitats yet with constant distance among individuals (indicating that site selection is only determined by the presence of conspecifics).

difference between the two components of spatial autocorrelation is that $\hat{\varepsilon}_i$ does not depend on the observations of a single sampling event, but $c_{t,i}$ does. Thus, changes in the locations of the individuals between sampling can be used to identify the extent to which the autocorrelation is caused by the environment or by conspecific interactions (Fig. 1).

We combined the two different concepts used to account for spatial autocorrelation into a single model:

$$\log\left(\frac{p_{t,i}}{1-p_{t,i}}\right) = \beta_0 + \beta_{\text{auto}} c_{t,i} + \sum_{k=1}^{K} \beta_k x_{k,i} + \tilde{\varepsilon}_t + \hat{\varepsilon}_i + \varepsilon_{t,i} \quad (1)$$

with the following assumptions:

$$y_{t,i} \sim \text{Bernoulli}\left(p_{t,i}\right) \quad (2)$$

$$c_{t,i} = \frac{1}{\sum_{j \neq i} y_{t,j}} \sum_{j \neq i} y_{t,j} \exp\left(-\lambda d_{i,j}\right) \quad (3)$$

$$\varepsilon \sim N\left(0, \sigma^2_{\text{res}}\right) \quad (4)$$

$$\tilde{\varepsilon} \sim N\left(0, \sigma^2_{\text{time}}\right) \quad (5)$$

$$\hat{\varepsilon} \sim N\left(0, \sigma^2_{\text{space}} W\right) \quad (6)$$

$$w_{i,j} = \exp\left(-\frac{d_{i,j}}{\rho}\right) \quad (7)$$

We refer to our model as an autologistic spatial error model. It is a mixed effect logistic regression model that is designed to model the probability of occurrence of a species, $p_{t,i}$, in sampling event t at site i. The response variable $y_{t,i}$ follows a Bernoulli distribution of the probabilities $p_{t,i}$. The fixed part of our model is composed of an intercept β_0, the linear combination of environmental site properties $\sum_{k=1}^{K} \beta_k x_{k,i}$, with k running over all site properties K, and the autocovariate part $\beta_{\text{auto}} c_{t,i}$. The autocovariate is calculated from the data prior to model fitting following Eq. 3. Here, $y_{t,j}$ is the response variable, $d_{i,j}$ is the distance between sites i and j, and λ is a range parameter that determines how sharply the influence of neighbouring observations declines with distance. In addition, our model incorporates a random intercept $\varepsilon_{t,i}$ for each observation (Eq. 4) and a random intercept $\tilde{\varepsilon}_t$ for different sampling events (Eq. 5), which both follow a normal distribution. The random intercept $\tilde{\varepsilon}_t$ is included to account for differences in the population size between sampling events. Further, our model includes spatially correlated random effects $\hat{\varepsilon}$ (Eq. 6). W is a symmetric weight matrix, with the entries $w_{i,j}$ being defined in Eq. 7. For the

spatial correlations, we chose a Matern correlation structure (Guttorp and Gneiting 2006) with $\nu = \frac{1}{2}$, which simplifies to an exponential correlation structure (Eq. 7) with a range parameter ρ.

The free parameters of our model are the intercept β_0, the slopes β_{auto} and β_k, the standard deviations σ_{space}, σ_{time} and σ_{res}, and the parameter ρ. We treat λ as a hyperparameter and perform a grid search to find the best value (Claesen and De Moor 2015).

Alternative model specifications

To evaluate the performance of our model, we compared it to three models with only one or no component for spatial autocorrelation. The models are defined by the following formulas:

$$\text{Alternative 1 (Baseline):}\quad \log\left(\frac{p_{t,i}}{1-p_{t,i}}\right) = \beta_0 + \sum_{k=1}^{K}\beta_k x_{k,i} + \tilde{\varepsilon}_t + \varepsilon_{t,i} \tag{8}$$

$$\text{Alternative 2 (Autologistic):}\quad \log\left(\frac{p_{t,i}}{1-p_{t,i}}\right) = \beta_0 + \beta_{auto} c_{t,i} + \sum_{k=1}^{K}\beta_k x_{k,i} + \tilde{\varepsilon}_t + \varepsilon_{t,i} \tag{9}$$

$$\text{Alternative 3 (Spatial error):}\quad \log\left(\frac{p_{t,i}}{1-p_{t,i}}\right) = \beta_0 + \sum_{k=1}^{K}\beta_k x_{k,i} + \tilde{\varepsilon}_t + \hat{\varepsilon}_i + \varepsilon_{t,i} \tag{10}$$

These models are based on the assumptions defined in Eq. 3–7. Alternative 1 ('Baseline model', Eq. 8) is a logistic regression model that does not include any component to account for spatial autocorrelation. Alternative 2 ('Autologistic model', Eq. 9) incorporates an autocovariate as an additional variable, but no spatially correlated random effects. Alternative 3 ('Spatial error model', Eq. 10) includes spatially correlated random effects, but no autocovariate.

Data

Virtual species data

To test the performance of the models we generated repeated observations of a virtual species. We simulated gridded virtual landscapes of 50 × 50 cells (i.e. sites) characterized by four site properties. Three site properties represented habitat characteristics expressed on a continuous suitability gradient from zero (minimum suitability) to one (maximum suitability). Two of these variables exhibited a spatial structure whereas the third was spatially randomly distributed. The fourth site property was a spatially structured binary variable (zero or one), implemented to represent sites that are completely hostile to the species (i.e. suitability of zero). For a visual representation of the virtual landscape simulations, Supplementary material Appendix 1 Fig. S1.

We then generated the species records, using a two-step approach. In the first step, we sequentially placed a random number of individuals in the landscape. As the first individuals selected their sites almost exclusively based on habitat characteristics, even in scenarios in which individuals were interacting, we implemented a second step in which we repeatedly picked one of the individuals at random and let it choose a site again. In both steps, individuals picked site i with probability p_i which was determined by the dynamic site suitability q_i with

$$p_i = \frac{\exp(q_i)}{\sum_j \exp(q_j)} \tag{11}$$

$$q_i = \beta_{auto} c_i + \sum_{k=1}^{K}\beta_k x_{k,i} \tag{12}$$

$$c_i = \sum_{j \neq i} y_j \exp(-\lambda d_{i,j}) \tag{13}$$

where $x_{k,i}$ is the value of site property k at site i, $d_{i,j}$ is the distance between sites i and j and y_j is the occupation status of site j. β_k are the slopes of the environmental site properties for which we used values of $\beta_1 = 3$, $\beta_2 = 3$ and $\beta_3 = 2$. The slope β_{auto} is associated with conspecific interactions and determines whether they are positive (conspecific attraction, $\beta_{auto} = 1.5$), neutral ($\beta_{auto} = 0$) or negative (competitive exclusion, $\beta_{auto} = -2.5$). We assigned a larger (absolute) value to the competitive exclusion slope because the importance of a variable in the selection process of an individual is a product of the standard deviation of that variable and the associated parameter. The standard deviation of c_i is intrinsically higher for conspecific attraction than for competitive exclusion because the former leads to clustering and the latter has the opposite effect. We chose all parameter values to ensure roughly equal variable importance. For the range parameter in Eq. 13, we used $\lambda = 1$.

In the simulations, each grid cell could be selected by one individual only. We allowed individuals to resettle 10 times on average, as we found that this was enough to ensure a stable distribution (Supplementary material Appendix 1 Fig. S2). For each of the three types of interactions, we simulated 30 datasets with each dataset consisting of 10 independent sampling events. We resampled the number of individuals for each sampling event (ranging between 30 and 150) to reflect the fact that numbers of individuals may fluctuate between surveys. Per sampling event, we randomly selected a subset of 1000 cells without an individual as absences. We kept the virtual landscape constant within the datasets but varied it between datasets. We used the same landscapes for all three types of interactions. To simulate the datasets, we used the

programming language Python (Van Rossum and Drake 2011), ver. 3.6.9.

Case study: real-world data

As our case study, we used records of the corncrake *Crex crex* in the floodplains of the Rhine River in the Netherlands (Supplementary material Appendix 1 Fig. S3), which provide the species with an important breeding habitat. Sovon Dutch Centre for Field Ornithology has been conducting systematic simultaneous surveys in the floodplain areas twice per breeding season since 2001 (Koffijberg and Schoppers 2009). In these surveys, the entire study area is scrutinized for the presence of corncrakes. Presence records refer to singing males, which are indicative of breeding sites, and are obtained at night when the singing activity is highest. Males sing more or less continuously between 11:00 pm and 3:00 am at stable singing sites, and their songs can be heard over considerable distances (500–1000 m), ensuring a very high probability of detection (Stowe et al. 1993, Wettstein et al. 2001, Sklíba and Fuchs 2004). We selected observations from 2001 through 2007 using records from the second simultaneous survey only, as carried out by mid-June, to avoid potential pseudo-replication due to possible correlations in bird records between both surveys. We preferred data from the second survey because the first survey is conducted shortly after the corncrakes arrive from their wintering grounds, with limited time for interactions and resettling. For each year, we included observations of four days (Friday–Monday) centred around the survey weekend. This yielded 143 observations in total. Per survey year, we randomly selected 1000 pseudo-absences across the surveyed floodplains (Barbet-Massin et al. 2012).

We characterized the habitat of each site based on its vegetation characteristics and elevation. We used elevation as a proxy for food availability as the number of invertebrates found in river floodplains is highly correlated with elevation (Schipper et al. 2008). We retrieved information on vegetation types from an ecotope map of the Netherlands (Rijkswaterstaat 1998, Houkes 2008). Because of the large number of vegetation types (ecotopes) relative to the number of observations, we classified each ecotope as either a suitable or an unsuitable corncrake habitat, based on information on habitat requirements provided by Schäffer and Koffijberg (2004). For elevation, we used a 25 m resolution elevation map (Schellekens et al. 2014, Straatsma et al. 2019). For each record (presence or pseudo-absence), we then calculated the proportion of suitable habitat area (%) and the mean elevation (m a.s.l.) in a circular zone with a radius of 250 m surrounding the given position, as a proxy for the home range (Supplementary material Appendix 1 Fig. S4). We chose a radius of 250 m for the home ranges, based on information on home range size provided by Koffijberg et al. (2007). We standardized all site properties before model fitting. For three records, the assumed home range did not overlap with the ecotope map and we therefore removed them, leaving a total of 140 records for our analysis.

Model application and evaluation

We first applied our model as well as the three alternative models to the 90 simulated datasets. To fit the models, we excluded one of the two spatially structured site properties (x_1). This introduced clusters of individuals in the species data that did not originate from the site properties known to the model, which was necessary to test whether the models were able to disentangle (unknown) habitat characteristics and conspecific interactions as drivers of spatial autocorrelation.

To find the best value of λ (Eq. 3), we performed a grid search in which we tested five different values, centred around the true value of $\lambda = 1$ which we used in the simulations: 1.25^{-2}, 1.25^{-1}, 1.25^{0}, 1.25^{1} and 1.25^{2}. We investigated the goodness of fit of the models by cross-validation (Witten et al. 2016). To that end, we trained models with the data of all but two sampling events and validated them with the data of the sampling events that were left out. In the left-out sampling events, we calculated the area under the receiver operating curve (ROC; Fawcett 2004) as

$$\mathrm{AUC}_t = \frac{1}{mn}\sum_{i=1}\sum_{j=1} 1_{y_{t,i} > y_{t,j}} \quad (14)$$

with $y_{t,i}$ and $y_{t,j}$ being the model responses for the sites i and j in sampling event t and $1_{y_{t,i} > y_{t,j}}$ being the indicator function: it is 1 if and only if $y_i > y_j$ and 0 otherwise. With i running over all presence sites and j running over all absence sites, the AUC_t gives the fraction of absence sites that are ranked lower than presence sites. As the number of presences per sampling event varied considerably, we weighted the AUC_t by the number of presences of the left-out sampling events to calculate a mean AUC. We selected the λ that resulted in the model with the highest AUC, and standardized β_{auto} to the standard deviation of the autocovariate in the simulation to ensure comparability of the slope used in the simulation and the estimated slopes. This is needed for the autocovariate but not for the environmental variables, because different values of λ only affect the former (Eq. 3). We further checked for remaining autocorrelation in the model residuals using Moran's I (Moran 1950).

For the case study, we performed a seven-fold cross-validation to identify the λ value resulting in the best model performance. We trained models with the data of all but one year and validated them with the data of the year we left out, reasoning that correlations between years can be considered to be negligible due to the high mortality rates of corncrakes (Green 2004). As we did not know the true value of λ, we tested a larger range of possible values for λ, ranging from 0.1 to 10.0. The models were evaluated as explained above.

We further investigated the performance of our model approach when fitted with data of a single sampling event, in order to verify whether our approach indeed needs repeated measurements. For both the simulation study and the case study, we used only one sampling event (year) for model

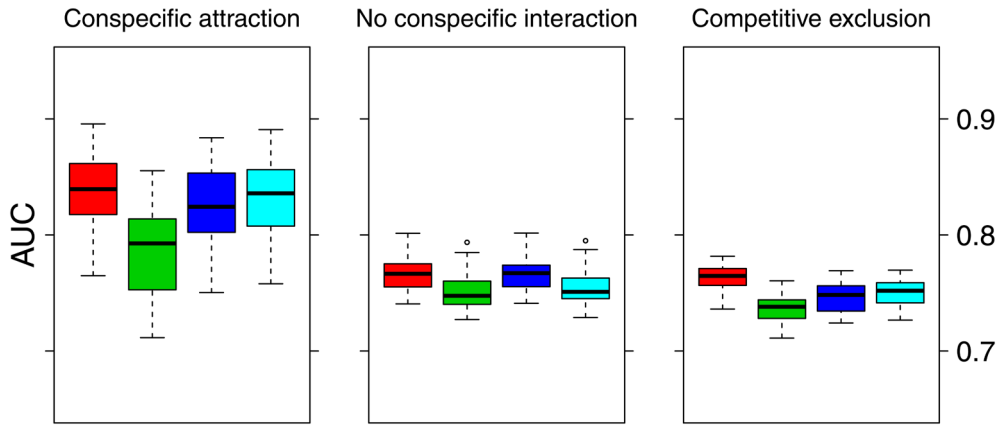

Figure 2. Results of the virtual species study. Boxplots of the AUC obtained for the autologistic spatial error model (in red), the baseline model (in green), the autologistic model (in blue) and the spatial error model (in turquoise) on sampling events that have been left out in the model fitting. From left to right, the figures show the results for the scenarios of conspecific attraction (left), no conspecific interaction (centre) and competitive exclusion (right).

fitting and the rest of the data for model validation. We separately performed the single-event analysis for each year. In this analysis, we put in the true parameter value (simulation study) and the optimal parameter value (case study) for λ upfront. We computed the AUC following Eq. 14 and averaged across the years as outlined above.

We fitted all models using the Integrated Nested Laplace Approximation (INLA; Rue et al. 2009, Martins et al. 2013). INLA is a Bayesian framework in which model parameters are optimized using the Laplace Approximation, i.e. approximating distributions with normal distributions. INLA works on a large range of models, including the spatial models that we are interested in here (Lindgren et al. 2011). We performed all model fitting and evaluation in R (R Core Team), ver. 3.6.3, including the r-INLA package (Lindgren et al. 2015), ver. 19.09.03, for model fitting, the pgirmess package (Giraudoux 2018), ver. 1.6.9, for calculating residual autocorrelation and the pROC package (Robin et al. 2011), ver. 1.15.3, for calculating AUC values.

Results

Virtual species data

Compared to the alternative models, our autologistic spatial error model generally had the best performance in terms of AUC (Fig. 2) and resulted in the most accurate slope estimates (Fig. 3). We obtained accurate slope estimates for the environmental variables in all three conspecific interaction scenarios. Slope estimates for the conspecific interactions were accurate for the competitive exclusion and no conspecific interaction scenarios, and less biased compared to the alternative model in the scenario of conspecific attraction. The autologistic model overestimated the conspecific interaction slope in all three scenarios. For the baseline model and the spatial error model, the spatially correlated variable slopes were overestimated in the scenario of conspecific interaction and underestimated in the scenario of competitive exclusion. The spatially uncorrelated variable slope was underestimated by the same models in the scenario of conspecific attraction.

Using Moran's I, we did not detect any remaining spatial autocorrelation in the residuals of our model. There was also no remaining autocorrelation in the residuals of the spatial error model, while the models without spatially correlated random effects (the baseline model and the autologistic model) were unable to completely remove spatial autocorrelation (Supplementary material Appendix 1 Table S1).

When applied to single sampling events, our model underestimated the slope of conspecific interactions in all three scenarios. Further, the environmental variable slopes were overestimated in the scenario of no conspecific interaction and associated with high uncertainty in the other two scenarios (Supplementary material Appendix 1 Table S2). This finding confirms that repeated observations of a species in the same environment are required for our model approach to disentangle extrinsic and intrinsic drivers of spatial autocorrelation.

Corncrake case study

In the case study application, the AUC of our autologistic spatial error model (0.89 ± 0.01) was slightly higher than that of the autologistic model (0.88 ± 0.01) and significantly higher than the AUCs of the spatial error model (0.83 ± 0.03) and the baseline model (0.73 ± 0.03) (Fig. 4). Our model classified the proportion of suitable habitat in the home range as the most important predictor of the presence of corncrakes. Conspecific interactions and elevation were classified as equally important. While the ranking for the environmental variables was the same for the other models, the autologistic model classified conspecific interactions as more important

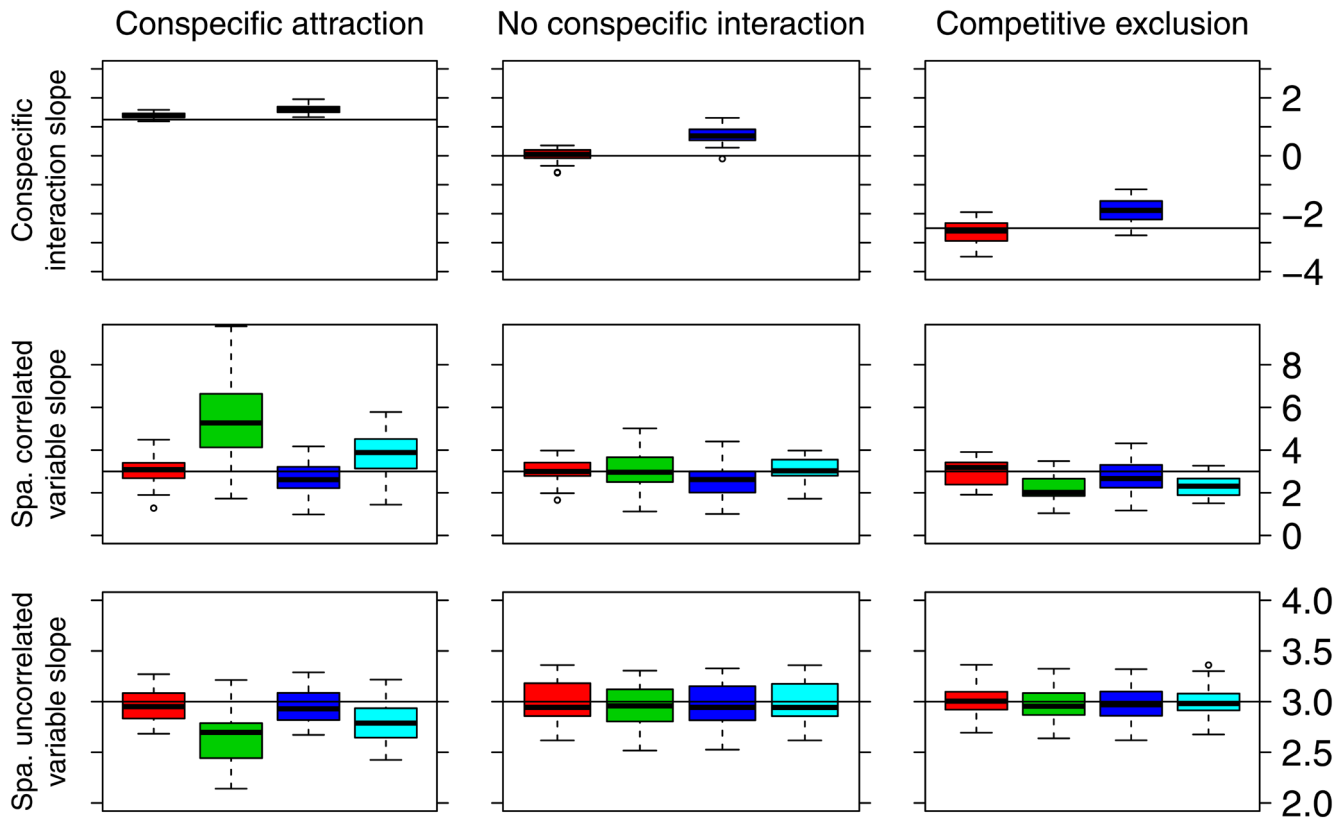

Figure 3. Results of the virtual species study. Boxplots of the slope estimates for the conspecific interaction (top), the spatially correlated habitat property (centre) and the spatially uncorrelated habitat property (bottom) obtained with the autologistic spatial error model (in red), the baseline model (in green), the autologistic model (in blue) and the spatial error model (in turquoise). Note that the baseline model and the spatial error model do not give estimates of the conspecific interaction slope. From left to right, the figures show the results for the scenarios of conspecific attraction (left), no conspecific interaction (centre) and competitive exclusion (right). The true slopes are represented by the horizontal lines.

than elevation (Fig. 5). The application of our model to single years of data did not provide a clear ranking, as the slopes for all three explanatory variables were associated with large estimated standard deviations. For elevation, the 1σ confidence interval overlapped with zero (Supplementary material Appendix 1 Table S4).

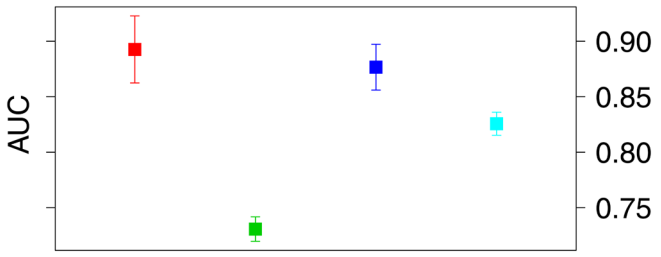

Figure 4. Results of the corncrake case study. AUCs obtained for the autologistic spatial error model (in red), the baseline model (in green), the autologistic model (in blue) and the spatial error model (in turquoise) on data of years left out in the model fitting. Each data point corresponds to the weighted mean AUC across all folds of the cross-validation and error bars show the weighted standard deviation.

Discussion

The performance of the models

In this study, we investigated the performance of a new SDM approach designed to account for spatial autocorrelation caused by both extrinsic and intrinsic factors, as represented by unknown habitat characteristics and conspecific interaction, respectively. We showed that it is important to explicitly account for these different drivers of spatial autocorrelation, as this clearly improved model performance. The goodness of fit, measured in terms of AUC, increased, while the bias of slope estimates decreased in comparison to models that included no or only one component to account for spatial autocorrelation.

The autologistic model overestimated the slope of the conspecific interaction across all scenarios. As there is only one overarching term for spatial autocorrelation in this model, conspecific interactions and unknown habitat characteristics as drivers of spatial autocorrelation were merged. In our setting, this led to an inflation of the conspecific interaction slope (spatial confounding; Hanks et al. 2015). We also found that the autologistic model underestimated the importance of the spatially correlated environmental predictor. As

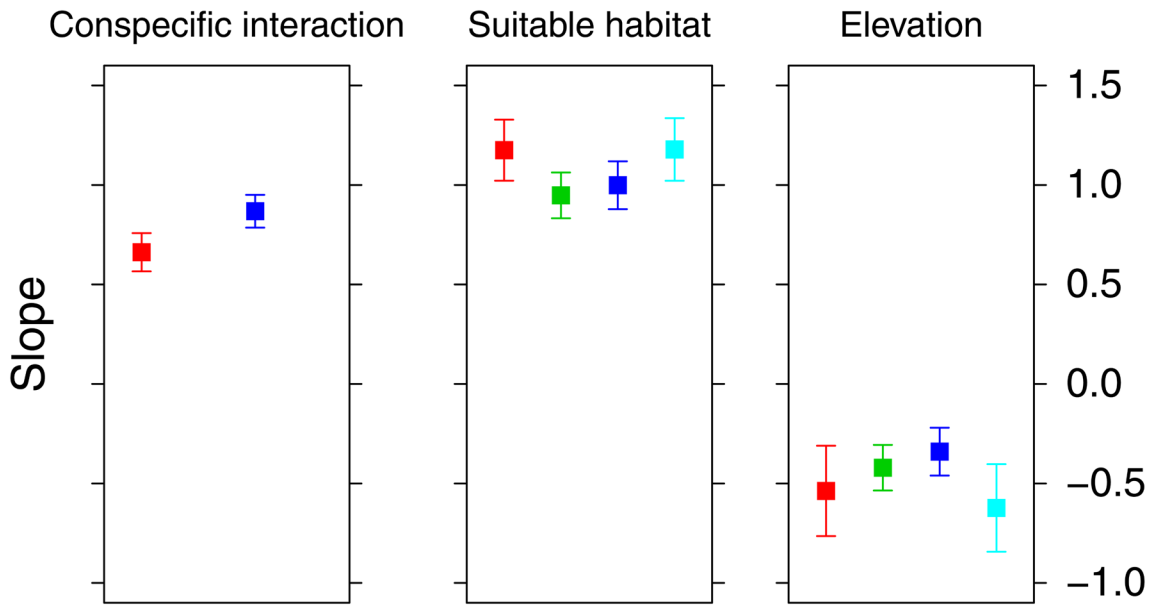

Figure 5. Results of the corncrake case study. Slope estimates obtained with the autologistic spatial error model (in red), the baseline model (in green), the autologistic model (in blue) and the spatial error model (in turquoise). From left to right, the figures show the estimates for the conspecific interaction (left), the amount of suitable habitat (centre) and the elevation (right). Each data point corresponds to the averaged mean estimate across all folds of the cross-validation and error bars show the average estimate of the standard deviation. Note that the baseline model and the spatial error model do not give estimates of the conspecific interaction slope.

our virtual species data included multiple environmental site properties, individuals were more likely to choose sites that were suboptimal when considering a single site property. Thus, excluding one of the spatially correlated site properties from the model fitting, resulted in an underestimation of the importance of the other spatially correlated site property in the autologistic model. For a similar reason, i.e. having only one term to account for spatial autocorrelation, the spatial error model overestimated the importance of the spatially correlated environmental property in the scenario of conspecific attraction and underestimated the importance in the scenario of competitive exclusion, showing that it incorrectly attributed (part of) the conspecific interactions to the environmental properties.

Our results show that the autologistic model provided biased slope estimates even in the absence of conspecific interaction, while the random error model was unbiased in that scenario. This result is in line with the findings of Dormann et al. (2007) and Crase et al. (2012) who focused on extrinsic drivers of spatial autocorrelation only, excluding conspecific interactions. Our study thus extends their conclusions to a setting in which different drivers of spatial autocorrelation are at play. Importantly, our results also showed that slopes were not biased homogenously (Fig. 2), implying that the baseline model, the autologistic model and the spatial error model may give flawed rankings of site property importance.

Case study results

Compared to the three alternative models, our autologistic spatial error model showed a better fit to the case study data. The slope estimates of our model indicate that the amount of suitable habitat is the most important predictor of the presence of corncrakes within the floodplains of the Rhine River in the Netherlands. Although this result was shared by all four models, our model showed that the amount of suitable habitat is more important than suggested by both the baseline model and the autologistic model, emphasizing that future conservation efforts should focus on maximizing habitat availability. On the other hand, our model estimated a lower importance of conspecific interaction compared to the autologistic model. This finding suggests that there are indeed hidden site properties that partly drive the clustering of the corncrakes.

The superior performance of our model in terms of AUC suggests three relevant implications. First, despite its complexity and thus the danger of overfitting (Lever et al. 2016), our model did not overfit the data, as indicated by the high performance when tested on temporally independent data (i.e. from a different year) that had been left out from the model training. Second, as mentioned above, the results of our model indicate that there are additional habitat factors that drive spatial autocorrelation of the corncrake, as including spatially correlated random effects resulted in improved model performance. Third, conspecific interactions are more important to the corncrake than unknown site properties, as the performance of the autologistic model was much better than that of the spatial error model (Fig. 4).

Although identifying additional habitat factors is beyond the scope of this study, we hypothesize that floodplain management – in particular, mowing – could be one of these. Parts of the grasslands and meadows in the Rhine River

floodplains are mown before corncrakes arrive, which renders these potential breeding habitats unsuitable (Green et al. 1997). As a result, corncrakes may cluster in the remaining habitats where mowing takes place later. As mowing was not included in our environmental site properties, the spatially correlated random effect term may have captured this additional clustering.

Applicability

In this study, we build on autologistic regression because autocovariate models are 1) easy to interpret, with explicit slope estimates for both the autocovariate and the environmental site properties, which are not provided by other methods (Dormann 2007), and 2) very popular in the research community (Ramakers et al. 2014, Gallien et al. 2015, Carman and Jenkins 2016). We confirmed our premise that repeated species observations are needed to disentangle intrinsic and extrinsic drivers of spatial autocorrelation, as models fitted with data of single sampling events led to biased slopes with large uncertainty. This reflects that a single sampling event does not provide enough information for the model to adequately ascribe spatial autocorrelation to either the autocovariate or the spatial error term. While we applied our approach to occurrence data, the method is easily generalized to other types of data (e.g. count data). Furthermore, the key concept to disentangle drivers of spatial autocorrelation by using longitudinal data of multiple sampling events in the same environment can be transferred to many methods, including more advanced forms of Bayesian modelling (Kéry and Royle 2016) and dynamic range models (Soriano-Redondo et al. 2019). Finally, our concept is not limited to a single species, but could also be used in joint species distribution models (Pollock et al. 2014, Lany et al. 2019).

Note that we assumed that the unknown habitat characteristics do not change between sampling events. This assumption may not be justified in some cases, e.g. for a species that clusters due to an unmeasured weather-related phenomenon but is observed in different seasons of the year. Likewise, difficulties may arise if observations are carried out with huge time lags in between. Our approach will also be unable to distinguish conspecific interaction and other drivers of spatial autocorrelation that vary between sampling events, e.g. varying sampling effort. This is especially important for presence-only data. In our case study, the data was gathered comprehensively and systematically, minimizing potential observation biases. However, in settings in which this is not the case, it is important that the data sample is unbiased, see, e.g. Warton et al. (2013).

We argue that disentangling drivers of spatial autocorrelation is relevant particularly because any species distribution modelling study is likely to miss potentially relevant environmental site properties. Our approach can pick up such signals via the spatially correlated random effects. We further showed that the slope of the conspecific interaction term converged towards zero when individuals do not interact (scenario of no conspecific interaction). Thus, our model will help to diagnose whether conspecific interactions influence the distribution of a species. We, therefore, advocate the application of our methodology regardless, as it explicitly accounts for potential interactions as well as habitat factors as drivers of spatial autocorrelation and yields accurate estimates of both.

Data availability statement

The code for the generation of the virtual species data is available from Zenodo, doi: 10.5281/zenodo.3776179. The code for the data analysis is available from Zenodo, doi: 10.5281/zenodo.3776176. The ecotope map used in the corncrake case study is archived at the Dutch Nationaal Georegister (<www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/78d31ab4-4116-45b7-bcf5-e14960916b0f>) and accessible under the identifier 09ffe157-93b7-4599-ba6b-17001128a1fd. The corncrake data will be available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.rv15dv45w> (Mielke et al. 2020). The elevation data is not publicly accessible. To acquire access to the data, contact Menno Straatsma (m.w.straatsma@uu.nl).

Acknowledgements – We would like to thank our two anonymous reviewers for their valuable feedback, which helped us to improve the quality of this article. We would also like to thank Remon Koopman for his help in obtaining an ecotope layer and Menno Straatsma for his help in obtaining an elevation layer of the case study area.
Funding – This research was partially financed by the Netherlands Organisation for Scientific Research (NWO), under project number 617.001.451.
Author contributions – TC, KPM and AMS conceived the ideas and designed methodology; KK collected the corncrake field observations for the case study; KPM analysed the results with help from TC, AMS, MB, TH and MAJH; KPM, MB and AMS led the writing of the manuscript. All authors contributed critically to the drafts and gave their final approval for publication.

References

Araújo, M. B. et al. 2019. Standards for distribution models in biodiversity assessments. – Sci. Adv. 5: eaat4858.
Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.
Campomizzi, A. J. et al. 2008 Conspecific attraction is a missing component in wildlife habitat modeling. – J. Wildl. Manage. 72: 331–336.
Carman, K. and Jenkins, D. G. 2016. Comparing diversity to flower-bee interaction networks reveals unsuccessful foraging of native bees in disturbed habitats. – Biol. Conserv. 202: 110–118.
Claesen, M. and De Moor, B. 2015. Hyperparameter search in machine learning. – arXiv preprint arXiv: 1502.02127.
Cliff, A. D. and Ord, K. 1970. Spatial autocorrelation: a review of existing and new measures with applications. – Econ. Geogr. 46: 269–292.
Crase, B. et al. 2012. A new method for dealing with residual spatial autocorrelation in species distribution models. – Ecography 35: 879–888.

Crase, B. et al. 2014. Incorporating spatial autocorrelation into species distribution models alters forecasts of climate-mediated range shifts. – Global Change Biol. 20: 2566–2579.
Dormann, C. F. 2007. Assessing the validity of autologistic regression. – Ecol. Model. 207: 234–242.
Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.
Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.
Fawcett, T. 2004. ROC graphs: notes and practical considerations for researchers. – Mach. Learn. 31: 1–38.
Franklin, J. 2010. Species distribution modeling. – In: Franklin, J. (ed.), Mapping species distributions. Cambridge Univ. Press, pp. 3–20.
Gallien, L. et al. 2015. Contrasting the effects of environment, dispersal and biotic interactions to explain the distribution of invasive plants in alpine communities. – Biol. Invas. 17: 1407–1423.
Giraudoux, P. 2018. pgirmess: spatial analysis and data mining for field ecologists. – R package ver. 1.6.9, <https://CRAN.R-project.org/package=pgirmess>.
Green, R. E. 2004. A new method for estimating the adult survival rate of the corncrake *Crex crex* and comparison with estimates from ring-recovery and ring-recapture data. – Ibis 146: 501–508.
Green, R. E. et al. 1997. A simulation model of the effect of mowing of agricultural grassland on the breeding success of the corncrake (*Crex crex*). – J. Zool. 243: 81–115.
Guélat, J. and Kéry, M. 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. – Methods Ecol. Evol. 9: 1614–1625.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Guttorp, P. and Gneiting, T. 2006. Studies in the history of probability and statistics XLIX on the matern correlation family. – Biometrika 93: 989–995.
Hanks, E. M. et al. 2015. Restricted spatial regression in practice: geostatistical models, confounding and robustness under model misspecification. – Environmetrics 26: 243–254.
Houkes, G. 2008. Toelichting omzetting ecotopenkartering 1e cyclus. – RWS-DID.
Kéry, M. and Royle, J. A. 2016. Applied hierarchical modeling in ecology. – Elsevier.
Koffijberg, K. and Schoppers, J. 2009. Kwartelkoningen in Nederland in 2008 en evaluatie van het Beschermingsplan Kwartelkoning. – SOVON, Beek-Ubbergen.
Koffijberg, K. et al. 2007. Territorial behaviour and habitat use of corncrakes *Crex crex* in the Netherlands revealed by radio-tracking. – Limosa 80: 167–171.
Koffijberg, K. et al. 2016. Recent population status and trends of corncrakes *Crex crex* in Europe. – Vogelwelt 136: 75–87.
Kühn, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. – Divers. Distrib. 13: 66–69.
Lany, N. K. et al. 2019. Complementary strengths of spatially-explicit and multi-species distribution models. – Ecography 43: 456–466.
Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – Ecology 74: 1659–1673.
Lever, J. et al. 2016. Model selection and overfitting. – Nat. Methods 13: 703–704.
Lindgren, F. et al. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. – J. R. Stat. Soc. B 73: 423–498.
Lindgren, F. et al. 2015. Bayesian spatial modelling with R-INLA. – J. Stat. Softw. 63: 1–25.
Martins, T. G. et al. 2013. Bayesian computing with INLA: new features. – Comput. Stat. Data Anal. 67: 68–83.
Mielke, K. P. et al. 2020. Data from: Disentangling drivers of spatial autocorrelation in species distribution models. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.rv15dv45w>.
Miller, J. A. 2014. Virtual species distribution models: uing simulated data to evaluate aspects of model performance. – Prog. Phys. Geogr. 38: 117–128.
Moran, P. A. 1950. Notes on continuous stochastic phenomena. – Biometrika 37: 17–23.
Paulitz, T. et al. 2003. Spatial distribution of *Rhizoctonia oryzae* and rhizoctonia root rot in direct-seeded cereals. – Can. J. Plant Pathol. 25: 295–303.
Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). – Methods Ecol. Evol. 5: 397–406.
Ramakers, J. J. et al. 2014. Surviving on the edge: a conservation-oriented habitat analysis and forest edge manipulation for the hazel dormouse in the Netherlands. – Eur. J. Wildl. Res. 60: 927–931.
Rhodes, J. R. et al. 2009. GLMM applied on the spatial distribution of koalas in a fragmented landscape. – In: Zuur, A. F. et al. (eds), Mixed effects models and extensions in ecology with R. Springer, pp. 469–492.
Rijkswaterstaat. 1998. Toelichting bij de ecotopenkartering Rijktakken-Oost 1997. – RIZA.
Robin, X. et al. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. – BMC Bioinform. 12: 77.
Rue, H. et al. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. – J. R. Stat. Soc. B 71: 319–392.
Schäffer, N. and Koffijberg, K. 2004. *Crex crex* corncrake. – BWP Update 6: 55–76.
Schellekens, J. et al. 2014. Rapid setup of hydrological and hydraulic models using OpenStreetMap and the SRTM derived digital elevation model. – Environ. Model. Softw. 61: 98–105.
Schipper, A. M. et al. 2008. Spatial distribution and internal metal concentrations of terrestrial arthropods in a moderately contaminated lowland floodplain along the Rhine River. – Environ. Pollut. 151: 17–26.
Schipper, A. M. et al. 2011. The distribution of a threatened migratory bird species in a patchy landscape: a multi-scale analysis. – Landscape Ecol. 26: 397–410.
Schröder, B. and Seppelt, R. 2006. Analysis of pattern–process interactions based on landscape models – overview, general concepts and methodological issues. – Ecol. Model. 199: 505–516.
Sklíba, J. and Fuchs, R. 2004. Male corncrakes *Crex crex* extend their home ranges by visiting the territories of neighbouring males. – Bird Study 51: 113–118.
Soriano-Redondo, A. et al. 2019. Understanding species distribution in dynamic populations: a new approach using spatio-temporal point process models. – Ecography 42: 1092–1102.
Stowe, T. J. et al. 1993. The decline of the corncrake *Crex crex* in Britain and Ireland in relation to habitat. – J. Appl. Ecol. 30: 53–62.

Straatsma, M. W. et al. 2019. Towards multi-objective optimization of large-scale fluvial landscaping measures. – Nat. Hazards Earth Syst. Sci. 19: 1167–1187.
Teng, S. N. et al. 2018. Effects of intrinsic sources of spatial autocorrelation on spatial regression modelling. – Methods Ecol. Evol. 9: 363–372.
Van Rossum, G. and Drake, F. L. 2011. The python language reference manual. – Network Theory.
Verhoef, J. M. et al. 2018. Spatial autoregressive models for statistical inference from ecological data. – Ecol. Monogr. 88: 36–59.
Wang, F. et al. 2018. Incorporating biotic interactions reveals potential climate tolerance of giant pandas. – Conserv. Lett. 11: e12592.
Warton, D. I. et al. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. – PLoS One 8: e79168.
Wettstein, W. et al. 2001. Habitat selection of corncrakes (*Crex crex* L.) in Szatmár-Bereg (Hungary) and implications for further monitoring. – Ornis Hungarica 11: 9–18.
Wintle, B. and Bardos, D. 2006. Modeling species–habitat relationships with spatially autocorrelated observation data. – Ecol. Appl. 16: 1945–1958.
Wisz, M. S. et al. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. – Biol. Rev. 88: 15–30.
Witten, I. H. et al. 2016. Data mining: practical machine learning tools and techniques. – Morgan Kaufmann.

Supplementary material (available online as Appendix ecog-05134 at <www.ecography.org/appendix/ecog-05134>). Appendix 1.