Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging

Tomislav Hengl [a,*] Henk Sierdsema [b] Andreja Radović [c] Arta Dilo [d]

[a] *Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands*

[b] *SOVON Dutch Centre for Field Ornithology, Rijksstraatweg 178, 6573 DG Beek-Ubbergen, The Netherlands*

[c] *Institute of Ornithology, Croatian Academy of Sciences and Arts, Gundulićeva 24/II, 10000 Zagreb, Croatia*

[d] *GIS-technology section, the Research Institute for Housing, Urban and Mobility Studies, Technical University Delft, Jaffalaan9, 2628 BX Delft, The Netherlands*

Abstract

A computational framework to map species' distributions (realized density) using occurrence-only data and environmental predictors is presented and illustrated using a textbook example and two case studies: distribution of root vole (*Microtes oeconomus*) in the Netherlands, and distribution of white-tailed eagle nests (*Haliaeetus albicilla*) in Croatia. The framework combines strengths of point pattern analysis (kernel smoothing), Ecological Niche Factor Analysis (ENFA) and geostatistics (logistic regression-kriging), as implemented in the spatstat, adehabitat and gstat packages of the R environment for statistical computing. A procedure to generate pseudo-absences is proposed. It uses Habitat Suitability Index (HSI, derived through ENFA) and distance from observations as weight maps to allocate pseudo-absence points. This design ensures that the simulated pseudo-absence points fall further away from the occurrence points in both feature and geographical spaces. After the pseudo-absences have been produced, they are combined with occurrence locations and used to build regression-kriging prediction models. The output of prediction are either probability of species' occurrence or density measures. Addition of the pseudo-absence locations has proven effective — the adjusted R-square increased from 0.71 to 0.80 for root vole (562 records), and from 0.69 to 0.83 for white-tailed eagle (135 records) respectively; pseudo-absences improve spreading of the points in feature space and ensure consistent mapping over the whole area of interest. Results of cross validation (leave-one-out method) for these two species showed that the model explains 98% of the total variability for the root vole, and 94% of the total variability for the white-tailed eagle. The framework could be further extended to Generalized multivariate Linear Geostatistical Models and spatial prediction of multiple species. A copy of the R script and detailed instruction on how to run such analysis are available via contact author's website.

Key words: spatial prediction, pseudo-absence, R, adehabitat, gstat, spatstat

1 Introduction

A Species Distribution Model (SDM) can be defined as a statistical/analytical algorithm that predicts (either actual or potential) distribution of a species, given field observations and auxiliary maps, as well as expert knowledge. A special group of Species Distribution

[*] Tel.: +31-(0)20-5257379; fax: +39-(0)20-5257431. *E-mail addresses*: T.Hengl@uva.nl (T. Hengl), Henk.Sierdsema@sovon.nl (H. Sierdsema), anradovic@hazu.hr (Andreja Radović), A.Dilo@tudelft.nl (Arta Dilo).

Preprint of a draft for Ecological modelling *29 May 2009*

Models (SDMs) focuses on the so-called '*occurrence-only records*' — pure records of locations where a species occurred (Elith et al., 2006). The most frequently used techniques to generate species' distribution from occurrence-only records seem to be various kernel smoothing techniques, the Environmental-Niche Factor Analysis (ENFA) approach of Hirzel and Guisan (2002), the Genetic Algorithm for Rule-Set Prediction (GARP) approach of Stockwell and Peters (1999), and the Maximum entropy method (Maxent) introduced by Phillips et al. (2006). It has never been proven that any of these techniques outperforms its competitors. Zaniewski et al. (2002) evaluated performance of General Additive Models versus ENFA models and concluded that ENFA will likely be better in detecting the potential distribution hot-spots, especially if occurrence-only data is used. Tsoar et al. (2007) compared six occurrence-only methods for modeling species distribution (BIOCLIM, HABITAT, Mahalanobis distance method, DOMAIN, ENFA, and GARP), and concluded that GARP is significantly more accurate than BIOCLIM and ENFA; other techniques performed similarly. Jiménez-Valverde et al. (2008b) argue whether it is sensible to compare SDMs that conceptually aim at different aspects of spatial distribution at all — there is especially big difference between models predicting potential and realized distributions (although both are put under SDM).

So far, geostatistical techniques have not yet been used to generate (realized) species' distributions using occurrence-only data, mainly for two reasons: (1) absence locations are missing ('1's only), so that it is not possible to analyze the data using e.g. indicator geostatistics; and (2) the sampling is purposive and points are often clustered in both geographical and feature spaces, which typically causes difficulties during the model estimation. Spatial statisticians (e.g. Diggle (2003) and/or Bivand et al. (2008)) generally believe that geostatistical techniques are suited only for modeling of features that are inherently continuous (spatial fields); discrete objects (points, lines, polygons) should be analyzed using point pattern analysis and similar techniques. Bridging the gap between conceptually different techniques — point pattern analysis, niche analysis and geostatistics — is an open challenge.

Some early examples of using geostatistics with the species occurrence records can be found in the work of Legendre and Fortin (1989) and Gotway and Stroup (1997). Kleinschmidt et al. (2005) uses regression-kriging method, based on the Generalized mixed model, to predict the malaria incidence rates in South Africa. Miller (2005) uses a similar principle (predict the regression part, analyze and interpolate residuals, and add them back to predictions) to generate vegetation maps. Miller et al. (2007) further provide a review of predictive vegetation models that incorporate geographical aspect into analysis. Geostatistics is considered to be one of the four spatially-implicit group of techniques suited for species distribution modeling – the other three being: autoregressive models, geographically weighted regression and parameter estimation models (Miller et al., 2007). Pure interpolation techniques will often outperform niche based models (Bahn and McGill, 2007), although there is no reason not to combine them. *Hybrid* spatial/niche-analysis SDMs have been suggested also by Allouche et al. (2008). Pebesma et al. (2005) demonstrates that geostatistics is fit to be used with spatio-temporal species occurrence records. Analysis of spatial auto-correlation and its use in species distribution models is now a major research issue in ecology and biogeography (Guisan et al., 2006; Rangel et al., 2006; Miller et al., 2007).

Engler et al. (2004) suggested a hybrid approach to spatial modeling of occurrence-only records — a combination of Generalized Linear Model (GLM) and ENFA. In their approach, ENFA is used to generate the so-called '*pseudo-absence*' data, which are then added to the original presence-only data and used to improve the GLMs. In our opinion, such combination of factor analysis and GLMs is the most promising as it utilizes the best of the two techniques. In this paper, we extend the idea of Engler et al. (2004) by proposing a computational framework that further combines density estimation (kernel smoothing), niche-analysis (ENFA), and geostatistics (regression-kriging). We implement this framework in the R statistical computing environment, where various habitat analysis (adehabitat package), geostatistical (gstat package), and point pattern analysis (spatstat package) functions can be successfully combined. We decided to use a series of case studies, starting from a most simple to some real-life studies, to evaluate performance of our framework and then discuss its benefits and limitations.

2 Theory: combining kernel density estimation, ENFA and regression-kriging

The key inputs to a SDM is the inventory (population) of animals or plants consisting of a total of N individuals (a point pattern $\mathbf{X} = \{x_i\}_1^N$; where x_i is a spatial location of individual animal/plant), covering some area $B_{HR} \subset \mathbb{R}^2$ (where HR stands for home-range and $\mathbb{R}^2$ is the Euclidean space), and a list of environmental covariates/predictors $(q_1, q_2, \dots q_p)$ that can be used to explain spatial distribution of a target species. In principle, there are two distinct groups of statistical techniques that can be used to map the realized species' distribution: (a) the point pattern analysis techniques, such as kernel smoothing, which aim at predicting density of a point process; and (b) statistical, GLM-based, techniques that aim at predicting the probability distribution of occurrences. Both approaches are now explained in detail in the following sections.

2.1 Species' density estimation using kernel smoothing and covariates

Spatial density (λ; if unscaled, also known as "*spatial intensity*") of a point pattern for a given time interval is estimated as:

$$\mathbb{E}\left[N(\mathbf{X} \cap B)\right] = \int_B \lambda(x)dx \qquad (1)$$

In practice, it can be estimated using e.g. a kernel estimator (Diggle, 2003; Baddeley, 2008):

$$\lambda(x) = \sum_{i=1}^{n} \kappa \cdot (\|x - x_i\|) \cdot b(x) \qquad (2)$$

where $\lambda(x)$ is spatial density at location x, $\kappa(x)$ is the kernel (an arbitrary probability density), x_i is location of an occurrence record, $\|x - x_i\|$ is the distance (norm) between an arbitrary location and observation location, and $b(x)$ is a border correction to account for missing observations that occur when x is close to the border of the region. A common (isotropic) kernel estimator is based on a Gaussian function with mean zero and variance 1:

$$\widehat{\lambda}(x) = \frac{1}{H^2} \cdot \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\|x-x_i\|^2}{2}} \cdot b(x) \qquad (3)$$

The key parameter for kernel smoothing is the bandwidth (H) i.e. the smoothing parameter, which can be connected with the choice of variogram in geostatistics. The output of kernel smoothing is typically a map (image) consisting of M grid nodes, and showing spatial pattern of species' clustering.

Spatial density of a point pattern can also be modeled using a list of spatial covariates q's (in ecology, we call this environmental predictors), which need to be available over the whole area of interest B. For example, using a Poisson model (Baddeley, 2008):

$$log\lambda(x) = \log\beta_0 + \log q_1(x) + \ldots + \log q_p(x) \qquad (4)$$

where log transformation is used to account for the skewed distribution of both density values and covariates; p is the number of covariates. Models with covariates can be fitted to point patterns e.g. in the `spatstat` package (this actually fits the maximum pseudolikelihood to a point process; for more details see Baddeley (2008)). Such point pattern–covariates analysis is commonly run only to determine/test if the covariates are correlated with the feature of interest, to visualize the predicted trend function, and/or to inspect the spatial trends in residuals. Although statistically robust, point pattern–covariates models are typically not considered as a technique to improve prediction of species' distribution. Likewise, the model residuals are typically not used for interpolation purposes.

2.2 Predicting species' distribution using ENFA and GLM (pseudo-absences)

An alternative approach to spatial prediction of species' distribution using occurrence-only records and environmental covariates is the combination of ENFA and regression modeling. In general terms, predictions are based on fitting a GLM:

$$\mathbb{E}(\mathbf{P}) = \mu = g^{-1}(\mathbf{q} \cdot \beta) \qquad (5)$$

where $\mathrm{E}(\mathbf{P})$ is the expected probability of species occurrence ($P \in [0,1]$), $\mathbf{q} \cdot \beta$ is the linear regression model, and g is the link function. A common link function used for SDM with presence observations is the logit link function:

$$g(\mu) = \mu^{+} = \ln\left(\frac{\mu}{1-\mu}\right) \qquad (6)$$

so the Eq.(5) becomes logistic regression (Kutner et al., 2004).

The problem of running regression analysis with occurrence-only observations is that we work with 1's only, which obviously means that we can not fit any model to such data. To account for this problem, species distribution modelers (see e.g. Engler et al. (2004); Jiménez-Valverde et al. (2008a) and Chefaoui and Lobo (2008)) typically insert the so-called "*pseudo-absences*" — 0's simulated using a plausible model, such as ENFA, to depict areas where a species is not likely to occur. ENFA is a type of factor analysis that uses observed presences of a species to estimate which are the most favorable areas in the feature space, and then uses this information to predict the potential distribution of species for all locations (Hirzel and Guisan, 2002). The difference between ENFA and the Principal Component Analysis is that the ENFA factors have an ecological meaning. ENFA results in a Habitat Suitability Index (HSI$\in$ $[0 - 100\%]$) — by depicting the areas of low HSI, we can estimate were the species is very unlikely to occur, and then simulate a new point pattern that can be added to the occurrence locations to produce a '*complete*' occurrences+absences dataset. Once we have both 0's and 1's, we can fit a GLM as shown in Eq.(5) and generate predictions using geostatistical techniques as described in e.g. Gotway and Stroup (1997). Chefaoui and Lobo (2008) recently demonstrated that insertion of pseudo-absences greatly controls the success of species' distribution modeling by GLMs.

2.3 Predicting species' spatial density using ENFA and logistic regression-kriging

We now describe the technique that is advocated in this article, and that combines the two previously-described approaches. We make several additional steps that make the method somewhat more complicated, but also more suited for occurrence-only observations used in ecology. First, we assume that our input point pattern represents only a sample of the whole population ($\mathbf{X}_S = \{x_i\}_1^n$), so that the density estimation needs to be standardized to avoid biased estimates. Second, we assume that pseudo-absences can be generated using both information about the potential habitat (HSI) and geographical location of the occurrence-only records. Finally, we focus on mapping the actual count of individuals over the grid nodes (realized distribution), rather than mapping the probability of species' occurrence.

Spatial density values estimated by kernel smoothing are primarily controlled by the bandwidth size (Bivand et al., 2008). Obviously, the higher the bandwidth, the lower the values in the whole map; likewise, the higher the sample size (n/N), the higher the overall intensity, which eventually makes it difficult to physically interpret mapped values of spatial intensity. To account for this problem, we propose to use relative density ($\lambda_r : B \rightarrow [0, 1]$) expressed as the ratio between the local and maximum density at all locations:

$$\lambda_r(x) = \frac{\lambda(x)}{\max\left\{\lambda(x)|x \in B\right\}_1^M} \qquad (7)$$

An advantage of using the relative density is that the values are in the range $[0, 1]$, regardless of the bandwidth and sample size (n/N). Assuming that our sample $\mathbf{X}_S$ is representative and unbiased, it can be shown that $\lambda_r(x)$ is an unbiased estimator of the true spatial density (see e.g. Diggle (2003) or Baddeley (2008)). In other words, regardless of the sample size, by using relative intensity we will always be able to produce an unbiased estimator of the spatial pattern of density for the whole population (see further Fig. 1).

Furthermore, assuming that we actually know the size of the whole population (N), by using predicted relative density, we can also estimate the actual spatial density (number of individuals per grid node):

$$\lambda(x) = \lambda_r(x) \cdot \frac{N}{\sum\limits_{j=1}^{M} \lambda_r(x)}; \qquad \sum_{j=1}^{M} \lambda(x) = N \qquad (8)$$

which can be very handy if we wish to aggregate the species' distribution maps over some polygons of interest, e.g. to estimate the actual counts of individuals.

Our second concern is the insertion of pseudo-absences. Here, two questions arise: (1) how many pseudo-absences should we insert? and (b) where should we locate them? Intuitively, it makes sense to generate the same number of pseudo-absence locations as occurrences. This is also supported by the statistical theory of model-based designs, also known as "*D-designs*". For example, assuming a linear relationship between density and some predictor q, the optimal design that will minimize the prediction variance is to put half of observation at one extreme and other at other extreme. All D-designs are in fact symmetrical, and all advocate higher spreading in feature space (for more details about D-designs, see e.g. Montgomery (2005)), so this principle seems logical.

After the insertion of the pseudo-absences, the extended observations dataset is then:

$$\mathbf{X_f} = \left\{ \{x_i\}_1^n , \{x^*{}_i\}_1^{n*} \right\}; \qquad n = n^* \tag{9}$$

where $x^*{}_i$ are locations of the simulated pseudo-absences. This is not a point pattern any more because now also quantitative values — either relative densities ($\lambda_r(x_i)$) or indicator values — are attached to locations ($\mu(x_i) = 1$ and $\mu(x^*{}_i) = 0$).

The remaining issue is where/how to allocate the pseudo-absences? Assuming that a spreading of species in an area of interest is a function of the potential habitat and assuming that the occurrence locations on the HSI axis will commonly be skewed toward high values (see further Fig. 3 left; see also Chefaoui and Lobo (2008)), we can define the probability distribution (τ) to generate the pseudo-absence locations as e.g.:

$$\tau(x^*) = [100\% - \mathrm{HSI}(x)]^2 \tag{10}$$

where the square term is used to insure that there are progressively more pseudo-absences at the edge of low HSI. This way also the pseudo-absences will approximately follow Poisson distribution. In this paper we propose to extend this idea by considering location of occurrence points in geographical space also (see also an interesting discussion on the importance of geographic extent for generation of pseudo-absences by VanDerWal et al. (2009)). The Eq.(10) then modifies to:

$$\tau(x^*) = \left[\frac{d_R(x) + (100\% - \mathrm{HSI}(x))}{2} \right]^2 \tag{11}$$

where d_R is the normalized distance in the range $[0, 100\%]$, i.e. the distance from the observation points ($\mathbf{X}$) divided by the maximum distance. By using Eq.(11) to simulate the pseudo-absence locations, we will purposively locate them both geographically further away from the occurrence locations and in the areas of low HSI (unsuitable habitat).

After the insertion of pseudo-absences, we can attach to both occurrences-absences locations values of estimated relative density, and then correlate this with environmental predictors. This now becomes a standard geostatistical point dataset, representative of the area of interest, and with quantitative values attached to point locations (see further Fig. 2d).

Recall from Eq.(7) that we attach relative intensities to observation locations. Because these are bounded in the $[0, 1]$ range, we can use the logistic regression model to make predictions. Thus, the relative density at some new location (x_0) can be estimated using:

$$\widehat{\lambda}_r^+(x_0) = \left[1 + \exp\left(-\beta^{\mathbf{T}} \cdot \mathbf{q_0}\right)\right]^{-1} \tag{12}$$

where β is a vector of fitted regression coefficients, $\mathbf{q_0}$ is a vector of predictors (maps) at new location, and $\widehat{\lambda}_r^+(x_0)$ is the predicted logit-transformed value of the relative density. Assuming that the sampled intensities are continuous values in the range $\lambda_r \in (0, 1)$, the model in Eq.(12) is in fact a liner model, which allows us to extended it to a more general linear geostatistical model such as regression-kriging (also known as "*universal kriging*" or "*kriging with external drift*"). This means that the regression modeling is supplemented with the modeling of variograms for regression residuals, which can then be interpolated and added back to the regression estimate (Hengl, 2007):

$$\widehat{\lambda}_r^+(x_0) = \mathbf{q_0^T} \cdot \widehat{\beta}_{\texttt{GLS}} + \delta_{\mathbf{0}}^{\mathbf{T}} \cdot \left(\lambda_{\mathbf{r}}^+ - \mathbf{q} \cdot \widehat{\beta}_{\texttt{GLS}} \right) \tag{13}$$

where δ is the vector of fitted weights to interpolate the residuals using ordinary kriging. In simple terms, logistic regression-kriging consists of five steps:

(1) convert the relative intensities to logits using Eq.(6); if the input values are equal to 0/1, replace with the second smalles/highest value;

(2) fit a linear regression model using Eq.(12);

(3) fit a variogram for the residuals (logits);

(4) produce predictions by first predicting the regression-part, then interpolate the residuals using ordinary kriging; finally add the two predicted trend-part and residuals together (Eq.13)

(5) back-transform interpolated logits to the original $(0, 1)$ scale by:

$$\widehat{\lambda}_r(x_0) = \frac{e^{\widehat{\lambda}_r^+(x_0)}}{1 + e^{\widehat{\lambda}_r^+(x_0)}} \tag{14}$$

After we have mapped relative density over area of interest, we can also estimate the actual counts using the Eq.(8).

2.4 Species' Distribution Modeling using a textbook example

At this stage the above introduced theory might seem rather difficult to follow (especially because it links to different statistical theories such as ENFA, geostatistics, D-designs and point pattern analysis), hence we will also try to illustrate this theory using a real data set and prove our assumptions using a simple example. For readers requiring more detail, the complete R script used in this exercise with plots of outputs and interpretation of steps is available from the contact authors' homepage [1].

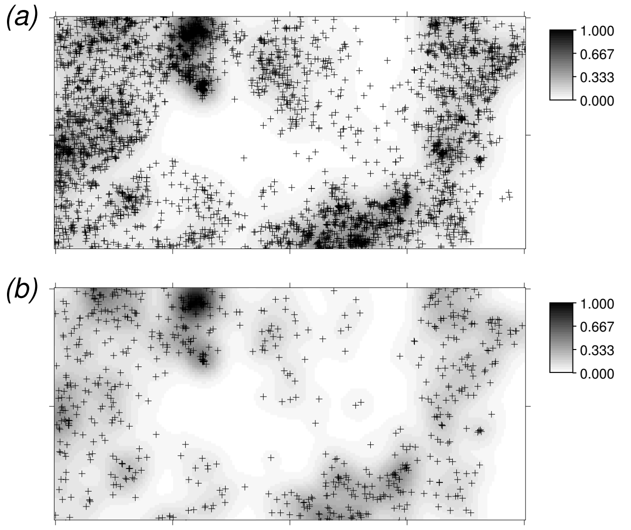

Fig. 1. Relative density estimated for the original `bei` data set (a), and its 20% sub-sample (b). In both cases the same bandwidth was used: H=23 m.

We use the `bei` dataset, distributed together with the spatstat package, and used in textbooks on point pattern analysis by Baddeley (2008) and many other authors. This data set consists of a point map showing locations of trees of the species *Beilschmiedia pendula Lauraceae* (in this case we deal with the whole population) and Digital Elevation Model (5 m resolution) as an auxiliary map, which can be used to improve mapping of the tree species. What makes this dataset especially suitable for such testing is the fact that the complete population of the trees has been visited/mapped for the area of interest (N is known, and so is B_{HR}). We will now implement all steps described in section 2.3 to predict spatial density of trees over the area of interest (M=20301 grid nodes). We will use a sample of 20% of the original population, and then validate the accuracy of our technique versus the whole population.

We start by estimating a suitable bandwidth size for kernel density estimation (Eq.3). For this, we use the method of Berman and Diggle (1989) (as described in Bivand et al. (2008, p.166–167)) that looks for the smallest Mean Square Error (MSE) of a kernel estimator. This only shows that we should not use bandwidths sizes smaller than 4 m (which is below resolution of our GIS); higher values seem plausible. We also consider the least squares cross validation method to select the bandwidth size using the method of Worton (1995), and as implemented in the adehabitat package. This does not converge, hence we need to set the bandwidth size using some *ad hoc* method (this is unfortunately a very common problem with many real point patterns). As a rule of thumb, we can start by estimating the smallest suitable range as the average size of block ($\sqrt{area(B)/N}$), and then set the bandwidth size at two times this value. There are 3605 trees (N) in the area of size 507,525 m^2, which means that we could use a bandwidth of 24 m (H).

We next derive a relative kernel density map (Eq.7), which is shown in Fig. 1a. If we randomly subset the original occurrence locations and then re-calculate the relative densities, we can notice that the spatial pattern of the two maps does not differ significantly, neither do their histograms. This supports our assumption that the relative density map (Eq. 7) can be indeed reproduced also from a representative sample (n=721).

We proceed with preparing the environmental predictors and testing their correlation with the density values. We can extend the original single auxiliary map (DEM) by adding some hydrological parameters: slope, topographic wetness index and altitude above channel network (all derived in SAGA GIS). The result of fitting a non-stationary point process with a log-linear density using the `ppm` method of spatstat shows that density is negatively correlated with wetness index, and positively correlated with all other predictors. A comparison between the Akaike Information Criterion (AIC) for a model without predictors and with predictors shows that there is a slight gain in using the covariates to predict the spatial density. Visually (Fig. 2a), we can see that the predicted trend seriously misses some hot-spots/clusters of points. This shows that using point pattern analysis techniques only to map (realized) species' distribution with covariates will be of limited use.

We proceed with ENFA. It shows that this species generally avoids the areas of low wetness index, i.e. it prefers ridges/dry positions (Fig. 2b; see also supplementary materials). This spatial correlation is now more distinct

[1] `http://spatial-analyst.net`

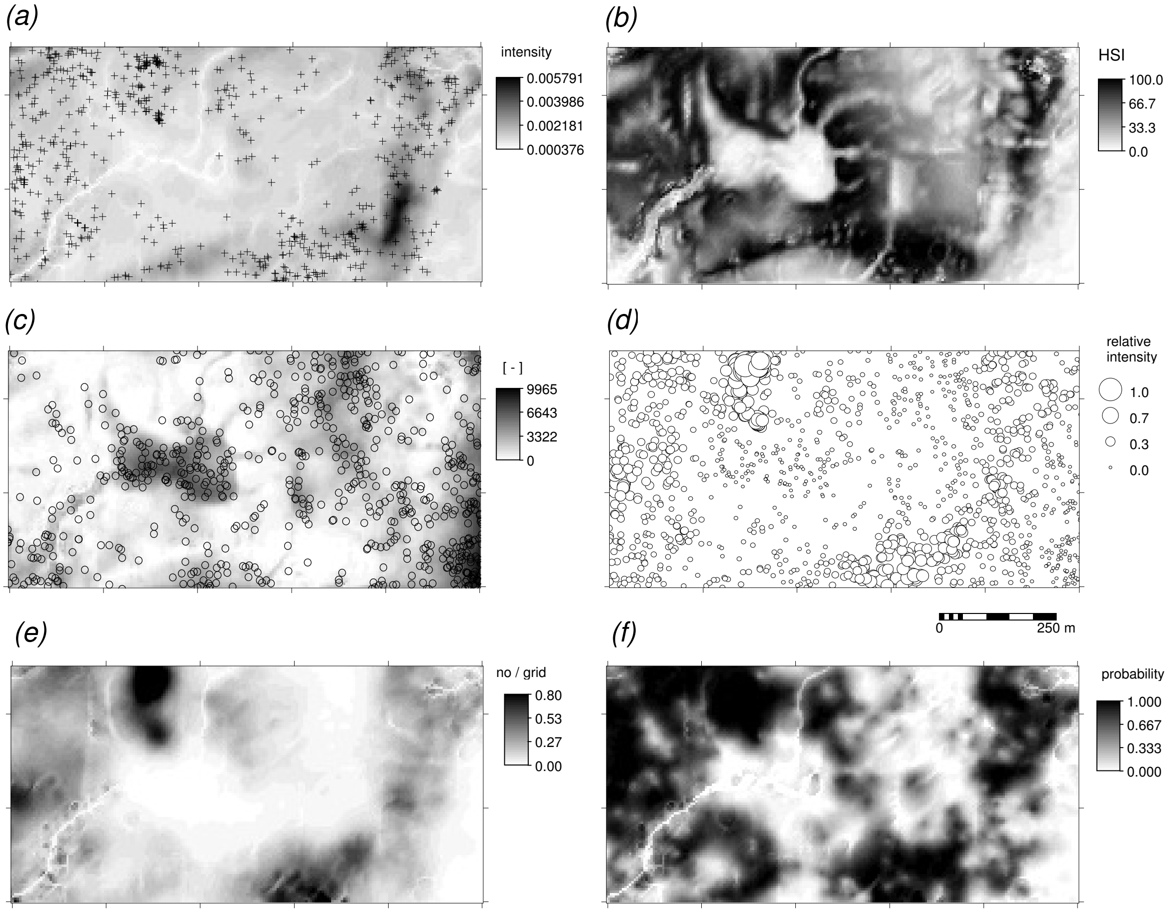

Fig. 2. Spatial prediction of the species distribution using the `bei` data set (20% sub-sample): (a) fitted trend model (`ppm`) using elevation, slope, topographic wetness index and altitude above channel network as environmental covariates; (b) Habitat Suitability Index derived using the same covariates; (c) the weight map and the randomly generated pseudo-absences using the Eq.(11); (d) input point map of relative intensities (includes the simulated pseudo-absences); (e) the final predictions of the overall density produced using regression-kriging (showing number of individuals per grid cell as estimated using Eq.8); and (f) predictions using a binomial GLM.

(compare with the trend model in Fig. 2a). This demonstrates the power of ENFA, which is in this case more suited for analysis of the occurrence-only locations than the regression analysis i.e the point pattern analysis.

By combining HSI and buffer map around the occurrence locations (Eq. 11), we are able to simulate the same amount of pseudo-absence locations (Fig. 2c). Note that the correlation between the HSI and density is now clearer, and the spreading of the points around the HSI feature space is symmetric (Fig. 3, right). Consequently, the model fitting is more successful: the adjusted R-square fitted using the four environmental predictors jumped from 0.07 to 0.28. This demonstrates the benefits of inserting the pseudo-absence locations. If we would randomly insert the pseudo-absences, the model would not improve (or would become even noisier).

We proceed with analyzing the point data set indicated in Fig. 2d using standard geostatistical tools. We can fit a variogram for the residuals, and then run the regresssion-kriging, as implemented in the `gstat` package. For a comparison, we also fit a variogram for the occurrence-absence data but using the residuals of the GLM modelling with binomial link function, i.e. 0/1 values (Fig. 4). As with any indicator variable, the variogram of the binomial GLM will show higher nugget and less distinct auto-correlation then the variogram for the density values. This is also because the residuals of the density values will still reflect kernel smoothing, especially if the predictors explain only a small part of variation in the density values.

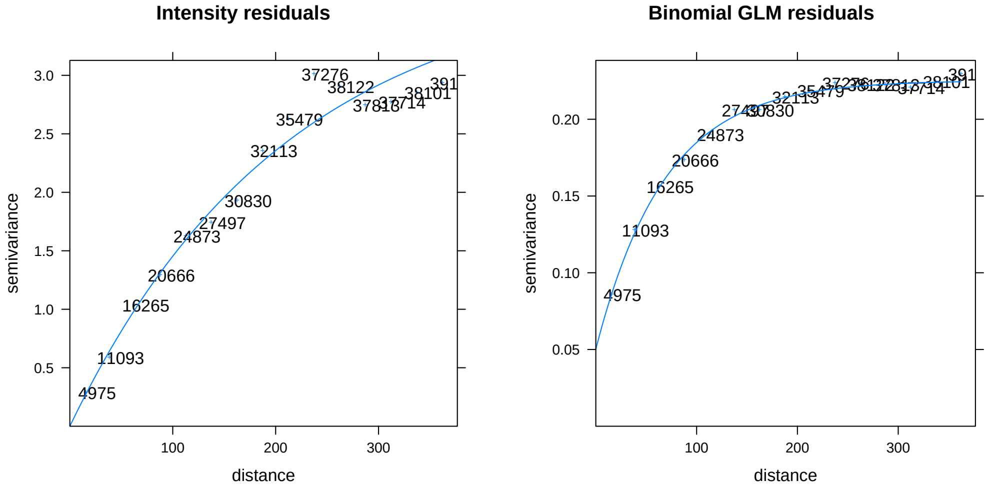

Fig. 4. Variogram models for residuals fitted in gstat using occurrence-absence locations: (left) density values (logits), and (right) probability values.

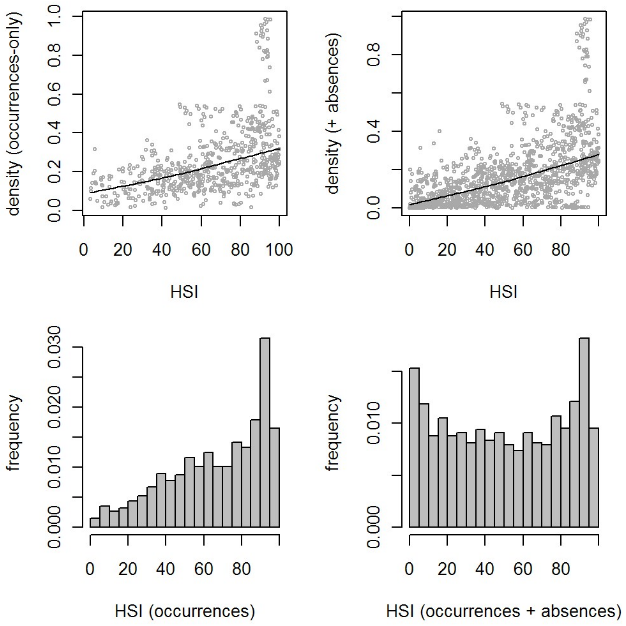

Fig. 3. Correlation plot HSI vs relative density with occurrence-only locations (left) and after the insertion of the pseudo-absence locations (right). Note that the pseudo-absences ensure equal spreading around the feature space (below).

The resulting map of density predicted using regression-kriging (Fig. 2e) is indeed a hybrid map that reflects kernel smoothing (hot spots) and environmental patterns, thus it is a map richer in contents than the pure density map estimated using kernel smoothing only (Figs. 1), or the Habitat Suitability Index (Fig. 2b). Note also that, although the GLM-kriging with bimodial link function (Fig. 2f) is statistically a more straight forward procedure (it is completely independent from point pattern analysis), it's output is limited in content because it also misses to represent the hot-spots/quantities of individuals. GLM-kriging in fact only shows the areas where a species' is likely to occur, without any estimation of how dense will the population be. Another advantage of using the occurrences+absences with attached density values is that we are able not only to generate predictions, but also to generate geostatistical simulations, map the model uncertainty, and run all other common geostatistical analysis steps.

In the last step of this exercises we want to validate the model performance using cross-validation and the original complete population. The ten-fold cross validation (as implemented in gstat) for the intensities interpolated regression-kriging shows that the model is highly precise — it explains over 99% of the variance in the training samples. Further comparison between the map shown in Fig. 2e and Fig. 1a shows that, with a 20% of samples and four environmental predictors, we are able to explain 96% of the pattern in the original density map (R-square=0.96). Fig. 5 indeed confirms that this estimator is unbiased.

One last point: although it seems from this exercise that we are recycling auxiliary maps and some analysis techniques (we use auxiliary maps both to generate the pseudo-absences and make predictions), we in fact use the HSI map to generate the pseudo-absences, and the original predictors to run predictions, which not necessarily need to reflect the same features. Relative densities, do not have to be directly correlated with the HSI,

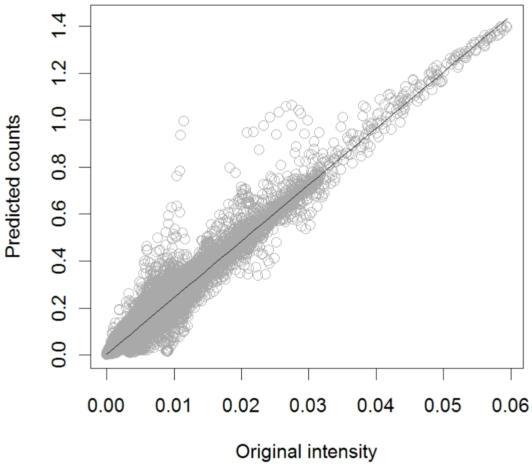

Fig. 5. Evaluation of the mapping accuracy for the map shown in Fig. 2e versus the original mapped density using 100% of samples (Fig. 1a).

although a significant correlation will typically be anticipated. Likewise, we use kernel smoother to estimate the intensities, but we then fit a variogram, which is obviously controlled by the amount of smoothing, i.e. value of the bandwidth, hence the variogram will often show artificially smooth shape, as shown in Fig. 4. The only way to avoid this problem is to estimate the bandwidth using some objective technique (which we failed to achieve in this example), or to scale the variogram fitted for the indicator variable (Fig. 4; right) to the density values scale.

3 Methods and materials

The computational framework used in this article follows the example described in the previous section (2.3), except it implies a larger number of predictors and several additional processing steps. A general workflow, as implemented in the R environment for statistical computing, is presented in Fig. 6. In order to fully understand all processing steps in detail, the interested readers can look at the R script provided via the contact authors' website.

The framework comprises six major steps. First, the occurrence locations are used to derive the density of a species for a given area based on the kernel smoother. Kernel density can be estimated in R using several methods; here we use the `density.ppp` method, as implemented in the `spatstat` package (Baddeley and Turner, 2005). In R, the smoothing parameter (bandwidth) can be estimated objectively; when it does not converges to a local minimum we use an *ad hoc* bandwidth selected as two times the average length of the block occupied by an individual ($2 \cdot \sqrt{area(B)/N}$). The output kernel density image can be coerced to the widely accepted spatial R format (`SpatialGridDataFrame`) of the `maptools/sp` package (Bivand et al., 2008); coercion to this format is important for further geostatistical analysis and export to GIS.

The second step is ENFA, which we run using the occurrence-only records. For ENFA, we use the `adehabitat` package, which is a collection of tools for the analysis of habitat selection by animals (Hirzel and Guisan, 2002; Calenge, 2006). Third, the resulting Habitat Suitability Index map (HSI, see further Fig. 8b and Fig. 11b) are used to generate the pseudo-absence locations. To achieve this, we use the `rpoint` method of the `spatstat` package. This method generates a random point pattern with the density of sampling proportional to the values of the weights map derived using Eq.(11).

In the fourth step, where possible, the simulated absence locations are reprojected to the Latitude/Longitude WGS84 system, exported to Google Earth (`writeOGR` method in `rgdal` package) and validated by an expert e.g. by doing photo-interpretation of high resolution satellite imagery.

Once we produce an equal number of occurrence and simulated absence locations, they can be packed together and used to build regression models using the ecological predictors. The residuals of the regression model are then analyzed for auto-correlation by fitting a variogram (`fit.variogram` method in `gstat`).

In the last, sixth step, after both the regression model and the variogram parameters have been determined, final predictions are generated using the generic `predict.gstat` method (Eq.13) as implemented in the `gstat` package (Pebesma, 2004; Bivand et al., 2008). More details on how to run regression-kriging and interpret its outputs can be found in Hengl (2007).

For a comparison, we also map the distribution of a species based on the occurrences+absences by fitting a binomial GLM. This is possible using the `glm` method in R, by setting a binomial link function (`binomial(link=logit)`). By using library `mgcv`, one can also fit Generalized Additive Models (GAM), using the same type of link function (`family=binomial`);

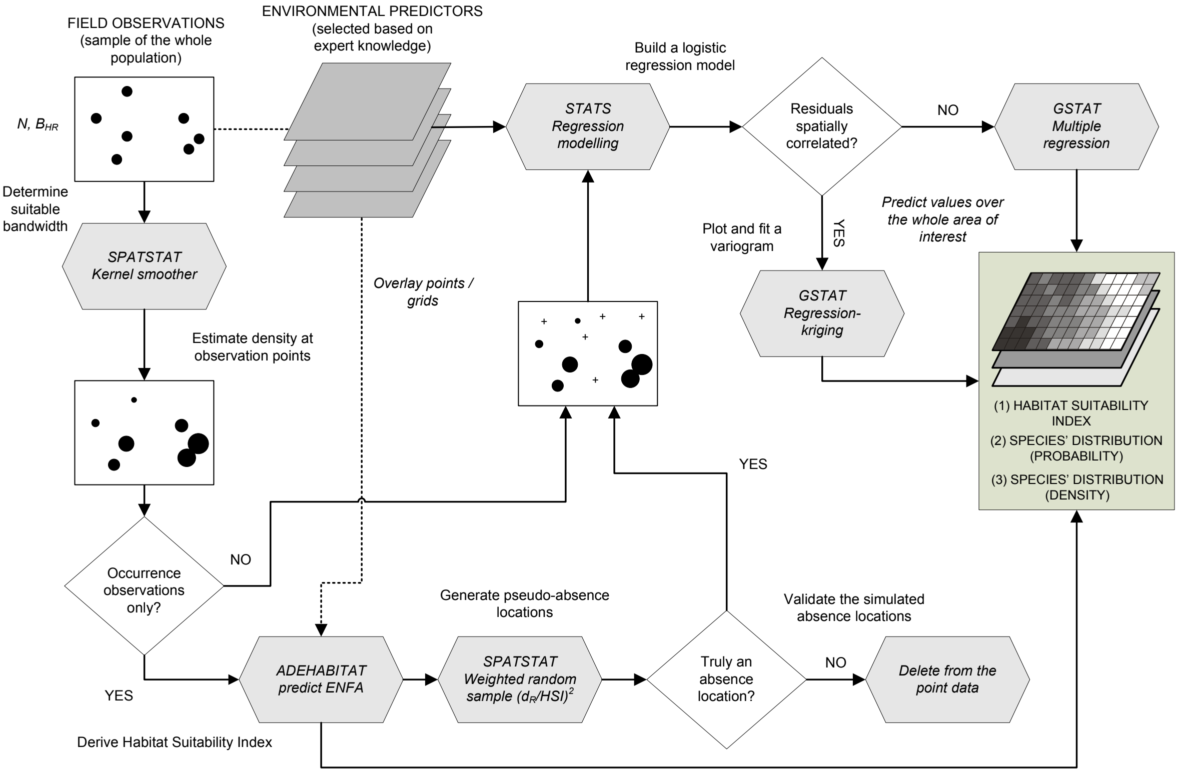
FIELD OBSERVATIONS
(sample of the whole population)
ENVIRONMENTAL PREDICTORS
(selected based on expert knowledge)
N, B_{HR}
Determine suitable bandwidth
SPATSTAT
Kernel smoother
Estimate density at observation points
Overlay points / grids
Build a logistic regression model
STATS
Regression modelling
Residuals spatially correlated?
NO
GSTAT
Multiple regression
YES
Plot and fit a variogram
Predict values over the whole area of interest
GSTAT
Regression-kriging
(1) HABITAT SUITABILITY INDEX
(2) SPECIES' DISTRIBUTION (PROBABILITY)
(3) SPECIES' DISTRIBUTION (DENSITY)
YES
Occurrence observations only?
NO
YES
Generate pseudo-absence locations
Validate the simulated absence locations
ADEHABITAT
predict ENFA
SPATSTAT
Weighted random sample $(d_R/HSI)^2$
Truly an absence location?
NO
Delete from the point data
Derive Habitat Suitability Index

Fig. 6. Data processing steps and related R packages used in this paper.

in this paper we focus on fitting linear models only. The output of running binomial GLM are probabilities, ranging from 0 to 1 (see further Fig. 8c and Fig. 11c).

The final results of running regression-kriging can be evaluated using the leave-one-out cross validation method, as implemented in the `krige.cv` method of `gstat` package (Pebesma, 2004). The algorithm works as follows: it visits a data point, predicts the value at that location by leaving out the observed value, and proceeds with the next data point. This way each individual point is assessed versus the whole data set. The results of cross-validation are used to pinpoint the most problematic locations, e.g. exceeding the three standard deviations of the normalized prediction error, and to derive the summary estimate of the map accuracy (Bivand et al., 2008, p.222–226).

We have tested this framework using occurrence-only records for two different species: distribution of root vole (*Microtes oeconomus*) in the Netherlands, and distribution of nests of white-tailed eagle (*Haliaeetus albicilla*) in Croatia. In both cases, we have jointly run analysis and then made the interpretation of the results and discussed strength and limitations of this framework.

4 Case studies

4.1 Root vole (Microtus oeconomus) in the Netherlands

The root vole (*Microtus oeconomus*) is a widespread, holarctic mouse species that inhabits the northern regions of Europe, Asia and Alaska. In Europe six subspecies are described (Mitchell-Jones et al., 2002). One of these subspecies, *Microtus oeconomus arenicola* is endemic to the Netherlands and listed as a species of conservation concern in the Habitats Directive of the European Union (van Apeldoorn, 2002). Its presence in the Netherlands is seen as a relict from the Ice Age and the Dutch population has no contact anymore with other European populations of the root vole. It is a good swimmer and well adapted to wetlands with varying water tables and has a high reproductive power. Therefore, root voles can swiftly recolonize wetlands after flooding.

It is thought that the Dutch root vole suffers heavily from competition with two other *Microtus*-species: the common cole (*Microtus arvalis*) and the field vole (*Microtus agrestis*) (van Apeldoorn et al., 1992; van Apeldoorn, 2002). On the isle of Texel, for example, the root vole was until recently the only occurring mouse species, which enable it to occupy a wider variety of habitats. Root vole populations are known to co-exist with populations of the other two *Microtus*-species on various locations in the country. Since these competitive species are not good swimmers, islands and large wetlands are the core areas of root voles, while smaller habitat patches in the vicinity of wetland throughout the country are places where the three species co-occur.

Following this knowledge about the biology of root vole, we selected two groups of environmental predictors to explain the distribution of root vole in the Netherlands: (1) habitat variables (wetland areas): `marsh` — marshland areas (0/1), `island` — island areas (0/1), `flooded` — flooded regions (0/1), `freat1` — duration of primary drainage in days (obtained from the `http://rijkswaterstaat.nl`), and `fgr` — map of the Physical Geographic Regions (denoting the same characteristics in physiography); and (2) biological factors: `nofvole` — indicator variable showing the areas in the north-west of the country where field voles are absent, `nofvole25` — 25 km wide band where root and field voles co-occur (all variables at 1 km resolution). Since the species are not mutually exclusive in most of the country on a landscape and/or local scale, other variables were sought fore that relate to the great ability of the root vole to recolonize adjacent areas from core areas. Hence, in addition to the maps showing locations of marshlands (`marsh`) and islands (`island`), we also used their density for 1 and 2 km search radiuses: (`island1km`, `island2km`, `marsh1km`, `marsh2km`), and `flooded2km`.

The occurrence records (562) of root vole were obtained from the Dutch organization for mammals (VZZ) (`http://www.vzz.nl/soorten/noordsewoelmuis/`). The records and environmental maps refer to the 2004–2007 period.

The occurrence records and derived kernel density is shown in Fig. 8a. The habitat suitability analysis shows that the potential spreading of the species is much larger than the actual locations show. The HSI map shown in Fig. 8b mainly follows the pattern of the primary drainage duration (`freat1`) and flooding intensity (`fgr`). The target variable (kernel density) is heavily skewed toward small values, so we used a log-transform for further modeling. The biplot graph of the principal component analysis output (Fig. 7), calculated using the sampling locations, shows four clusters of variables (a) `flooded`, `nofvole` and `fgr` (b) `marsh`, (c) `islands` and (d) `freat1`. Further Principal Component transformation of the original grid maps shows that PC1 explains 30% of total variance, PC2 20%, PC3 18%, PC4 10% and PC5 still 8% of the variation. The stepwise regression of PC-transformed variables reduces the number of variables as compared with the original variables from 8 to 9. The most significant predictors are now PC1 (`islands`) and PC3 (`flooded` and `marsh`). The PCA based-model is not statistically different from the model fitted using the original variables. The `gstat` fitted an exponential variogram model with a zero nugget, sill parameter of 0.00625 and a range parameter of 3.7 km to remaining residuals.

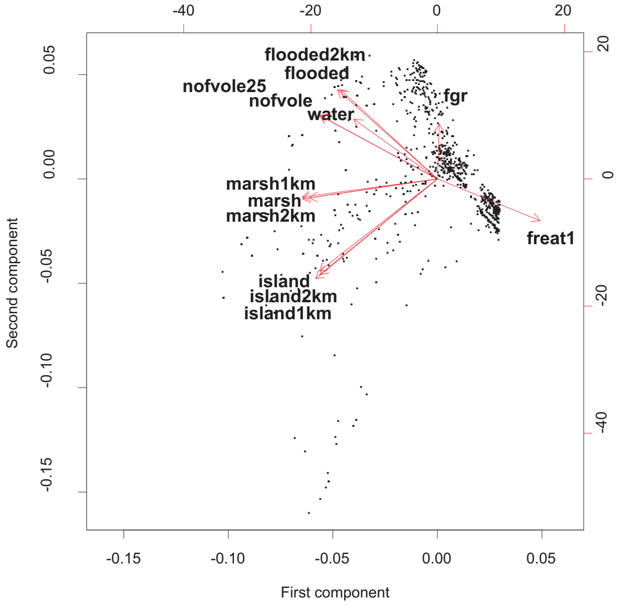

Fig. 7. Biplot showing the multicolinearity of the environmental predictors used to map distribution of root vole in the Netherlands: `marsh` — marshland areas (0/1), `island` — island areas (0/1), `flooded` — flooded regions (0/1), `freat1` — duration of primary drainage in days, `island1km`, `island2km`, `marsh1km`, `marsh2km`, and `flooded2km` — density of marshlands and flooded areas for 1 and 2 km search radiuses.

Regression analysis showed that, if occurrence-only data is used, the tailored predictors explain 71.0% of the variation. After including the simulated absence-observations the explained variation increases to 80.2%. The most significant predictors of root vole density are `marsh2km`, `flooded2km`, `freat1`, `islands2km`, `nofvolebuf25` and `water`.

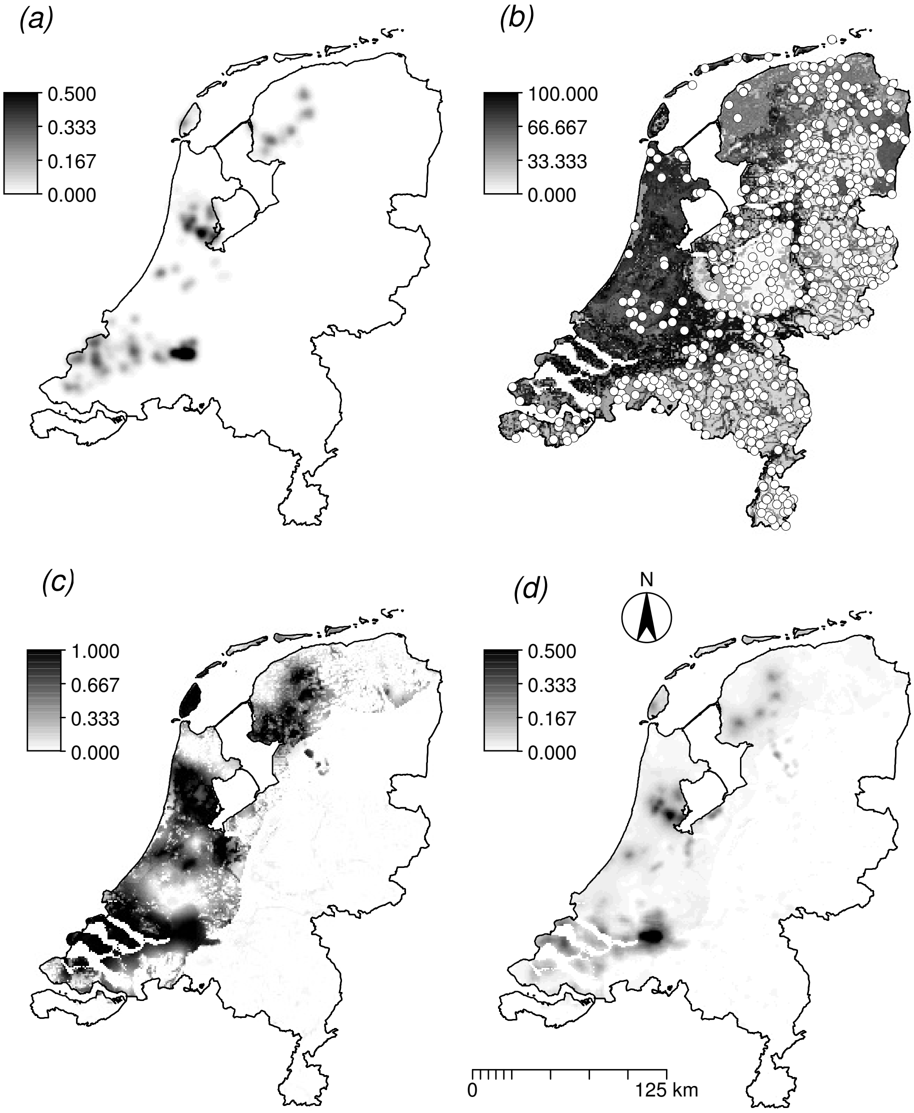

Fig. 8. Spatial prediction of root vole in the Netherlands: (a) the kernel density map (stretched to min–max range); (b) the Habitat Suitability Index and simulated pseudo-occurrence locations; (c) probabilities predicted using the Binomial GLM-based regression-kriging; and (d) the final predictions of densities produced using regression-kriging.

The final result of regression-kriging of 0/1 values and observation densities for root vole is shown in Figs. 8c and d. The root mean square prediction error at the leave-one-out validation points for model in Fig. 8d is 23% of the original variance; the regression-kriging model explains 98% of the original variance, which is quite high.

4.2 Nests locations of white-tailed eagle (Haliaeetus albicilla) in Croatia

In the second case study we focus on modeling the distribution of white-tailed eagle (*Haliaeetus albicilla*) in Croatia. At the beginning of the 1990s, about 80 pairs were recorded in Croatia (Tucker et al., 1994); a decade after, Croatia had 80–90 pairs. Some most recent records by Radović and Mikuska (2009) indicate a continuity of increase in population number in the period 2003–2006. This makes Croatia a country with the second largest population of *Haliaeetus albicilla* among the neighboring central European countries (Schneider-Jacoby et al., 2003; BirdLife International, 2004).

Haliaeetus albicilla breeds in various habitats but commonly needs sea coasts, lake shores, broad rivers, island and wetlands with high productivity. It breeds in different climates ranging from continental to oceanic. In Norway and Iceland nests are rarely placed above 300 m above sea level (Cramp, 2000). Same territories and eyries being occupied over many decades. Normally, only one or two alternate nests are built in a breeding territory (Helander and Stjernberg, 2002), which makes the nests most interesting for population distribution assessments. Breeding areas of *Haliaeetus albicilla* in Croatia are primarily alluvial wetlands along rivers Sava, Kupa, Drava and Mura (Pannonian plain), in Central and Eastern part of the country (Radović and Mikuska, 2009).

Following the habitat characteristics of *Haliaeetus albicilla*, we have prepared a total of 13 environmental predictors (all at 200 m resolution): `dem` — a Digital Elevation Map showing height of land surface; `canh` — derived as the difference between the topo-map DEM and the SRTM DEM, so that it reflects the height of canopy; `drailroad` — distance to rail roads; `droads` — distance to roads; `durban` — distance to urban areas; `dwater` — distance to water bodies; `pcevi1-4` — PCs from 12 MODIS Enchanced Vegetion Index (EVI) images obtained for the year 2005; `slope` — slope map derived using the DEM; `solar` — incoming solar insolation derived using the DEM; and `wetlands` — boolean map showing location of the wetlands. The proximity maps (`drailroad`, `droads`, `durban` and `dwater`) were derived from the vector features from the 1:100k topo-maps. `dem` and derivatives (`canh`, `slope` and `solar`) and EVI components are standard exhaustive predictors used for geostatistical mapping of environmental variables. The wetland habitats distribution map was obtained from the Croatian State Institute for Nature Protection (`http://www.cro-nen.hr/map/`). This is a boolean map (1/0) showing locations of the wetland areas, covered by both forests and swamps.

The nest positions used in this paper were recorded in the period 2003–2006. Altogether, 155 nest locations were recorded, out of which 125 locations showed clear signs of breeding (Radović and Mikuska, 2009). An additional 10 presumably active territories were detected but without knowing the exact position of the nests. Because of some problems during the fieldwork (minefields, flooded areas and extreme sensitivity of birds to our presence) the exact coordinates were taken for a total of 135 nests. We assume that this number represents about 80% of the total nests (N=169, B_{HR}=330 km^2), but this is hard to validate. Grlica (2007) most recently discovered some new breeding territories along Drava river coasts, but without recording the exact position of the nests.

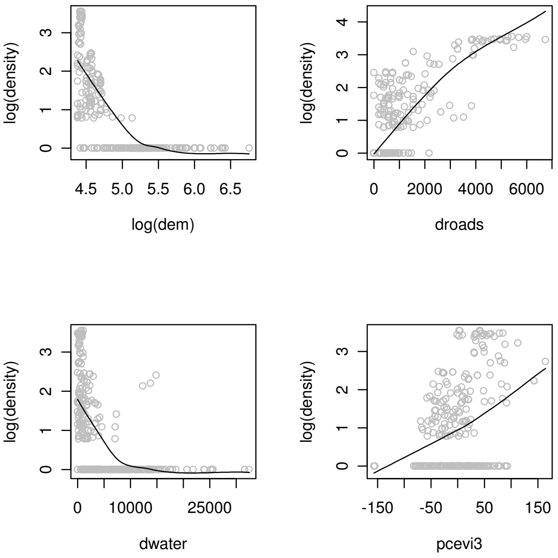

Fig. 9. Correlation plots between the log of nest densities and ecological predictors: `dem` — digital elevation model in meters; `droads` — distance to roads in meters; `dwater` — distance to water in meters; `pcevi` — the third component of the MODIS Enhanced Vegetation Index for year 2005.

The nest density estimated using a Gaussian kernel smoother with bandwidth set at 75% of the distance to the nearest neighbors (3.4 km) is shown in Fig. 11a. The areas with nest density close to zero are masked and 135 absence points generated using random sampling are shown in Fig. 11b. From these, 11 were found to fall in areas where potentially the species might occur, and were masked out from further analysis. We start by correlating the nest density estimated at observation points with the ecological predictors. If occurrence-only data are used, the ecological predictors explain 69% of variation of the target variable. Merging of the occurrence and absence observations gives 259 points in total, and the regression model explains 83% of variation. The most significant ecological predictors are `droads`, `wetlands`, `dem`, `pcevi3` and `dwater` (Fig. 9). Adding simulated absence locations was relatively inexpensive as it took only one day to validate simulated 135 locations.

The ecological predictors are highly inter-correlated and with skewed distributions. The biplot graph (Fig. 10) calculated at sampling locations shows that there are four clusters of predictors: (a) `dem` is correlated with `dwater` and `slope`; (b) `droads`, `durban`, `pcevi3`, `canh` and with `wetlands`; (c) `solar` and `pcevi4`; and (d)

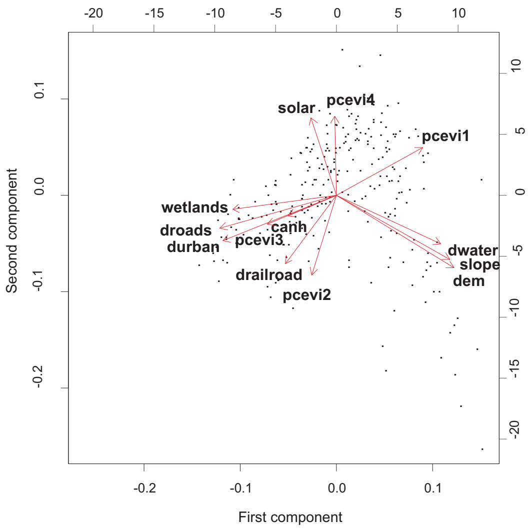

Fig. 10. Biplot showing the multicolinearity of the environmental predictors used to map distribution of white-tailed eagle: `dem` — digital elevation model; `canh` — height of canopy; `drailroads` — distance to rail roads; `droads` — distance to roads; `durban` — distance to urban areas; `dwater` — distance to water bodies; `pcevi1--4` — four PCs from 12 EVI images for year 2005; `slope` — slope map; `solar` — incoming solar insolation; and `wetlands` — boolean map showing location of wetlands.

`pcevi`.

The Principal Component transformation of the original predictors produces somewhat different picture. In this case, PC1 explains 80.1% of total variance and reflects mainly `pcevi01`, PC2 explains 7.9% of variance and reflects the position of `wetlands` and `dem`, PC3 explains 4.5% of variance, PC4 2.0% and PC5% 1.4% etc.

The step-wise regression shows that the most significant predictors of the nest density are PC2 (reflecting position of the wetlands and elevation) and PC1 (reflecting distance to roads and urban areas). Step-wise regression has much less problems in selecting the significant predictors if they are spatially independent. The number of significant predictors after the principal component transformation was reduced from 9 to 6; the adjusted R-square stays unchanged.

Further analysis of the residuals shows that they are spatially auto-correlated. We fitted an exponential variogram with 0 nugget, 0.263 sill parameter and range parameter of 5.2 km. The variogram for binomial GLM residuals is noisier than the variogram derived for densities. As expected, continuous variables (densities) are easier to model using geostatistics than the binary variables. This is true for both success of fitting a regression model and a for a success of fitting a variogram of residuals.

The accuracy of the map shown in Fig. 11a evaluated using the leave-one-out cross validation method shows that the map is fairly accurate: the root mean square prediction error at the validation points is only 16% of the original variance, or in other words, the regression-kriging model explains 94% of the original variance.

5 Discussion and conclusions

The results of the case studies described in this paper demonstrate that more informative and more accurate maps of the actual species' distribution can be generated by combining kernel smoothing, ENFA and regression-kriging. In order to improve estimation of regression model and final interpolation results, we advocate simulation of pseudo-absence data using inverted HSI and distance maps (Eq.11). This has shown to improve the regression models — the adjusted R-square increased from 0.69 to 0.83 for white-tailed eagle and from 0.71 to 0.80 for root vole — while improving the spreading of the points in feature space (see Fig. 12). This confirms the results of Chefaoui and Lobo (2008).

We believe that the method proposed in this article, as described in section 2.3, has several advantages over the known species' distribution modeling methods:

- The pseudo-absence locations are generated using a model-based design that spreads the points based on the geographical distance from the occurrence locations and the potential habitat. Compare with the purely heuristic approaches to generate the pseudo-absence by Chefaoui and Lobo (2008) or Jiménez-Valverde et al. (2008a).
- Both spatial auto-correlation structure and the trend component of the spatial variation are used to make spatial prediction of species' distribution. This leads to the Best Linear Unbiased Prediction of the presence/density values. Compare, for example, with the heuristic approach by Bahn and McGill (2007).
- Final output map shows distribution of a real physical parameter (number of individuals per grid cell) and can be directly validated using measures such as RMSE and similar. Compare with the often abstract

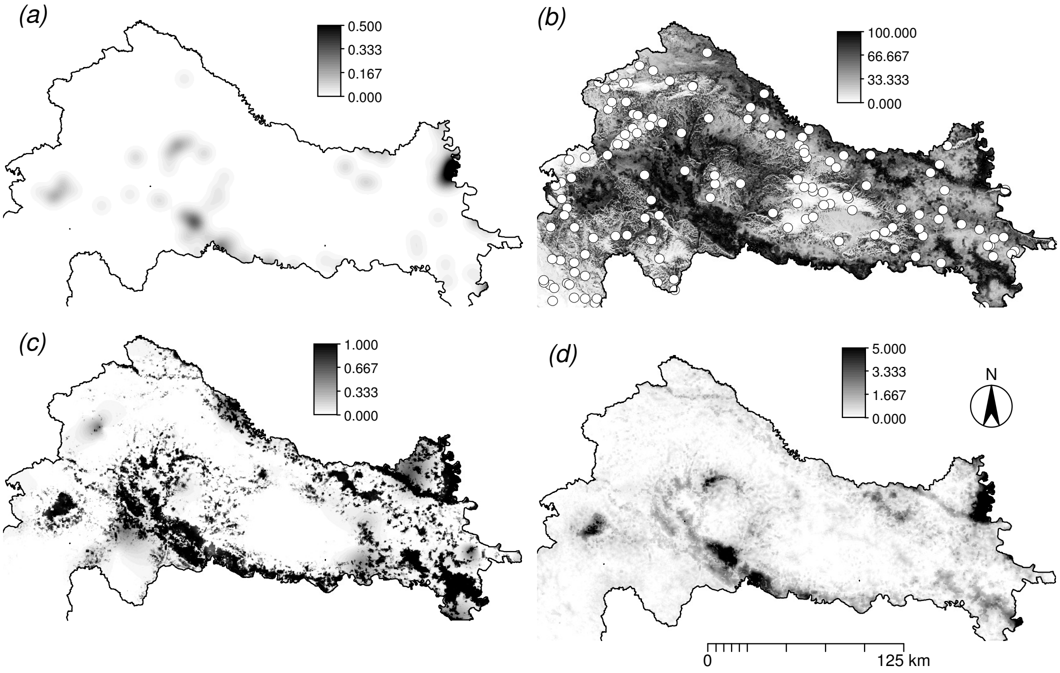

Fig. 11. Spatial prediction of white-tailed eagle in Croatia: (a) the kernel density map (stretched to min–max range); (b) the Habitat Suitability Index and simulated pseudo-occurrence locations; (c) probabilities predicted using the Binomial GLM-based regression-kriging; and (d) densities predicted using regression-kriging.

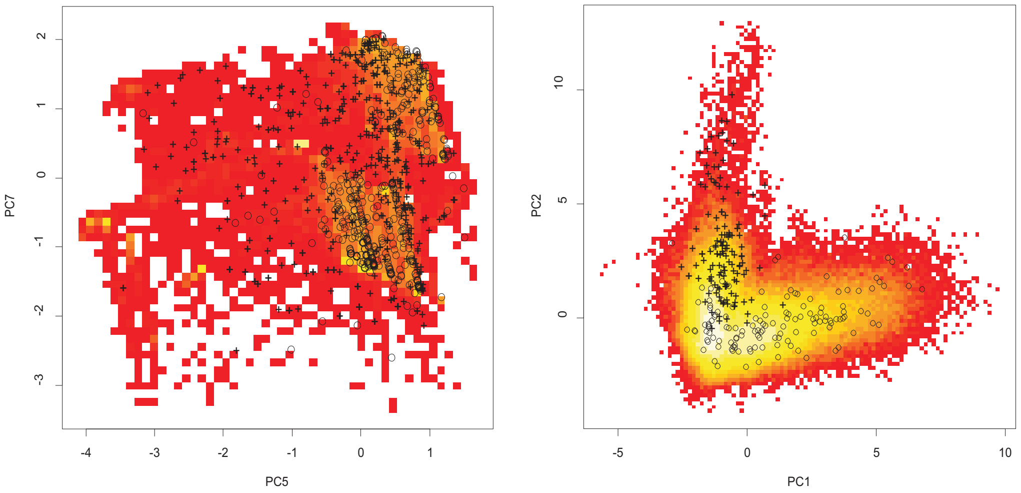

Fig. 12. Position of the occurrence (+) and the pseudo-absence (∘) locations when displayed in feature space (as defined using the most significant predictors): for root vole (left) and white-tailed eagle (right). Compare with Figs. 8a and 11b. The plot was produced using the hist2D function of the R package gplots.

evaluation measures (e.g Kappa, MaxKappa, AUC, adjusted D^2, AVI, CVI, Boyce index etc.) used in predictive habitat mapping (Hirzel et al., 2006).

- The whole mapping process can be automated in R, which is attractive for projects where the maps need to be constantly up-dated. The only interventions expected from a user is to provide an estimate of the total population of the species (N), the size of the area occupied (the home range area $area(B_{HR})$), and a list of environmental predictors.

Although we primarily advocate regression-kriging of relative densities, we are convinced that a species' distribution analyst should aim at producing all three types of maps: (1) the ENFA-based HSI map showing the potential habitat (Fig. 2b); (2) the species' distribution (probability) map (Fig. 2f); and (3) the species' distribution (density) map (Fig. 2e). ENFA can help understand the relationship between species and environmental conditions and generate pseudo-absence locations. The probability-based species' distribution map can be used to delineate home range areas (probability> 0.5), and the actual species' distribution map (density) quantifies the spreading of the species and can be used to estimate the number of individuals per area. Certainly, both binomial GLM using indicators and logistic regression-kriging using intensities are valid geostatistical techniques to handle this type of data.

In addition, visual validation of the simulated absence locations using Google Earth$^{\text{TM}}$ is fast, convenient and leads to more useful geostatistical models. The simulated absence points that are hard to validate visually (in the case of mapping the white-tailed eagle, any area close to wetlands and within natural forests), can be either omitted from the analysis or visited on the field. For example, in the case of mapping the white-tailed eagle in Croatia, only 11 simulated absence points (out of 135) were evaluated as unreliable and hence omitted from further analysis.

The proposed technique to generate pseudo-absences could be much improved. First, one could also build models that slowly increase the size of pseudo-absences until the prediction accuracy stabilizes. In this approach, we simply use a single number (number of pseudo-absences = number of presences), which is somewhat naïve approach. More absences can be generated for species that have narrow distributions/niche. Second, we ignore the fact that our pseudo-absences might be bias, so that our fitted model becomes over-optimistic. In the case of narrowly distributed species in a wide region, the selection of absences by our approach will generate absences far from the environmental conditions of presences, and possibly artificially increase the coefficient of variation. Both Chefaoui and Lobo (2008) and VanDerWal et al. (2009) clearly demonstrate that the way the pseudo-absence are generated has a significant impact on the resulting maps. More research is certainly needed to analyze impacts of techniques used to derive pseudo-absence, and the impacts they make on the success of prediction models.

Although the cross-validation statistics shows that we have produced a fairly accurate maps, in the case of mapping the distribution of root vole, it appears that the output map mainly reflects geometry of the points (note that even the buffer-based predictors we selected, also reflect geometry rather than environmental features). To prove this, we have excluded occurrence records from the most densely populated area (*Biesbosch*), only to see if the model would be able to predict the same pattern (extrapolation). The result of this exercise showed that our model is not successful in predicting the area that has been masked out, which finally means that the predictions by regression-kriging will be highly sensitive to how representative is the sample data set considering the whole population of this species.

Why does regression modeling performs poorer if only presence data is used? Obviously, the sampling designs are typically extremely biased considering the spreading of points in the feature space (Sutherland, 2006), which makes it very hard to estimate the true relationship between the distribution of a species and the ecological factors. It would be as if we would like to fit a model to estimate people's weight using their height, and then sample only extremely tall people. We illustrate this problem in Fig. 3 and 12, where you can compare spreading of the sampling points with occurrence only and with occurrence and simulated absence data. This shows that the occurrence only samples for specialized species are heavily clustered in the feature space (this is more distinct for the white-tailed eagle than for root vole). After addition of the absence locations, the feature space is much better represented, so that the output prediction maps become more reliable.

The geostatistical technique used in this paper could be expanded to accommodate even more complex data: spatio-temporal observations, multiscale predictors, clustered observations, trajectory-type of data, observations of multiple species and similar. In this article, we rely on the state-of-the-art geostatistical mapping techniques as implemented in the R package gstat. To run a GLM and then explore the residuals e.g. via variograms, is a routine practice, but it does not always tell the whole story. In the case of multiple regression, covariance matrix is used to account for spreading (clustering) of the points in the space. In our example (Fig. 11c), we fit a GLM that completely ignores location of the points, which is obviously not statistically optimal. In comparison, fitting a Generalized Linear Geostatistical Model (GLGM) can be more conclusive since we can model the spatial/regression terms more objectively (Diggle and Ribeiro Jr, 2007). This was, for example, the original motivation for the geoRglm and spBayes packages (Ribeiro Jr et al., 2003). However, GLGMs are not yet operational for geostatistical mapping purposes and R code will need to be adapted.

Automated retrieval and generation of distribution maps from biodiversity databases is possible but tricky. The biggest problem for such applications will be the quality of the occurrence records — especially their spatial reference that is extremely variable (from few meters to tens of kilometers), but also the sampling bias, and thematic quality of the records (incorrect taxonomic classification, incompleteness). Although Jimenez-Valverde and Lobo (2006) in general do not see the sampling bias as a big problem for the success of spatial prediction, in the case of regression-kriging the output maps will be heavily controlled by the sampling bias. Hence if you are considering implementing this framework, have in mind that your input data (point sample) should be a good spatial representation of the whole population (it is not so much about the size, but about how well are all presence locations represented geographically). Another issue is the computational burden of the framework we propose in Fig. 6 that can easily grow beyond the capacities of standard PCs. In fact, we could imagine that multiple species (all species in the GBIF database?) could be handled at the same time through a co-kriging framework, which would result in large quantity of models and combinations of models that would need to be fitted. The benefits of running the models jointly versus isolated modeling are obvious — this is rather a technical than conceptual problem. At this moment, we simply can not foresee when would such type of analysis become a reality.

References

Allouche, O., Steinitz, O., Rotem, D., Rosenfeld, A., Kadmon, R., 2008. Incorporating distance constraints into species distribution models. Journal of Applied Ecology 45 (2), 599–609.

Baddeley, A., 2008. Analysing spatial point patterns in R. CSIRO, Canberra, Australia.

Baddeley, A., Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software 12 (6), 1–42.

Bahn, V., McGill, B. J., 2007. Can niche-based distribution models outperform spatial interpolation? Global Ecology and Biogeography 16 (6), 733–742.

Berman, M., Diggle, P. J., 1989. Estimating weighted integrals of the second-order intensity of a spatial point process. Journal of the Royal Statistical Society B 51, 81–92.

BirdLife International, 2004. Birds in Europe: population estimates, trends and conservation status. BirdLife Conservation Series No. 12. BirdLife International, Cambridge, UK.

Bivand, R., Pebesma, E., Rubio, V., 2008. Applied Spatial Data Analysis with R. Use R Series. Springer, Heidelberg.

Calenge, C., 2006. The package “adehabitat” for the R software: A tool for the analysis of space and habitat use by animals. Ecological Modelling 197 (3-4), 516–519.

Chefaoui, R. M., Lobo, J. M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling 210, 478–486.

Cramp, S. (Ed.), 2000. The Complete Birds of the Western Palearctic. Concise Edition and CD-Rom Set. Oxford University Press, Oxford.

Diggle, P. J., 2003. Statistical Analysis of Spatial Point Patterns, 2nd Edition. Arnold Publishers.

Diggle, P. J., Ribeiro Jr, P. J., 2007. Model-based Geostatistics. Springer Series in Statistics. Springer.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, T. A., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberon, J., Williams, S., Wisz, M. S., Zimmermann, N. E., 2006. Novel methods improve prediction of species distributions from occurrence data. Ecography 29 (2), 129–151.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41 (2), 263–274.

Gotway, C. A., Stroup, W. W., 1997. A Generalized Linear Model approach to spatial data analysis and prediction. Journal of Agricultural, Biological, and Environmental Statistics 2 (2), 157–198.

Grlica, I. D., 2007. Monitoring of the *Riparia riparia* and *Haliaeetus albicilla* on the Drava river – project report (in Croatian). Croatian Academy of Sciences and Arts, Zagreb.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M. C. C., Aspinall, R., Hastie, T., 2006. Making better biogeographical predictions of species' distributions. Journal of Applied Ecology 43 (3), 386–392.

Helander, B., Stjernberg, M. (Eds.), 2002. Action plan for the conservation of white-tailed sea eagle. European Council / BirdLife International Sweden, Strasbourgh.

Hengl, T., 2007. A Practical Guide to Geostatistical Mapping of Environmental Variables. EUR 22904 EN. Office for Official Publications of the European Communities, Luxembourg.

Hirzel, A. H., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling 157 (2-3), 331–341.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling 199 (2), 142–152.

Jiménez-Valverde, A., Gómez, J., Lobo, J., Baselga, A., Hortal, J., 2008a. Challenging species distribution models: the case of Maculinea nausithous in the Iberian Peninsula. Annales Zoologici Fennici 45, 200–210.

Jimenez-Valverde, A., Lobo, J. M., 2006. The Ghost of Unbalanced Species Distribution Data in Geographical Model Predictions. Diversity and Distributions 12 (5), 521–524.

Jiménez-Valverde, A., Lobo, J. M., Hortal, J., 2008b. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14 (6), 885–890.

Kleinschmidt, I. Sharp, B. L., Clarke, G. P. Y., Curtis, B., Fraser, C., 2005. Use of Generalized Linear Mixed Models in the Spatial Analysis of Small-Area Malaria Incidence Rates in KwaZulu Natal, South Africa. American Journal of Epidemiology 153 (12), 1213–1221.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W. (Eds.), 2004. Applied Linear Statistical Models, 5th Edition. McGraw-Hill.

Legendre, P., Fortin, M. J., 1989. Spatial pattern and ecological analysis. Plant Ecology 80 (2), 107–138.

Miller, J., 2005. Incorporating Spatial Dependence in Predictive Vegetation Models: Residual Interpolation Methods. The Professional Geographer 57 (2), 169–184.

Miller, J., Franklin, J., Aspinall, R., 2007. Incorporating spatial dependence in predictive vegetation models. Ecological Modelling 202, 225–242.

Mitchell-Jones, A. J. G., Amori, W., Bogdanowicz, B., Krystufek, P. J. H. F., Reijnders, F., Spitzenberger, M., Stubbe, J., Thissen, J., Vohralík, V., Zima, J., 2002. The atlas of European mammals. Poyser Natural History, London.

Montgomery, D. C., 2005. Design and Analysis of Experiments, 6th Edition. Wiley, New York.

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30 (7), 683–691.

Pebesma, E. J., Duin, R. N. M., Burrough, P. A., 2005. Mapping sea bird densities over the North Sea: spatially aggregated estimates and temporal changes. Environmetrics 16 (6), 573–587.

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3-4), 231–259.

Radović, A., Mikuska, T., 2009. Population size, distribution and habitat selection of the white-tailed eagle *Haliaeetus albicilla* in the alluvial wetlands of Croatia. Biologia 64 (1), 1–9.

Rangel, T. F. L. V. B., Diniz-Filho, J. A. F., Bini, L. M., 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. Global Ecology & Biogeography 15 (7), 321–327.

Ribeiro Jr, P. J., Christensen, O. F., Diggle, P. J., 2003. geoR and geoRglm: Software for Model-Based Geostatistics. In: Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). Technical University Vienna, Vienna, pp. 517–524.

Schneider-Jacoby, M., Mohl, A., Schwarz, U., 2003. The white-tailed eagle in the Danube River Basin. In: Helander, B., Marquiss, M., Bowermann, W. (Eds.), Sea Eagle 2000. Swedish Society For Nature Conservation / SSF and Atta, Stockholm, pp. 133–140.

Stockwell, D., Peters, D., 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13 (2), 143–158.

Sutherland, W. J. (Ed.), 2006. Ecological Census Techniques: A Handbook, 2nd Edition. Cambridge University Press, Cambridge.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., Kadmon, R., 2007. A comparative evaluation of presence-only methods for modelling species distribution. Diversity & Distributions 13 (9), 397–405.

Tucker, G. M., Heath, M. F., Tomialojc, L., Grimmett, R. F. A., 1994. Birds in Europe: population estimates, trends and conservation status. BirdLife Conservation Series No. 3. BirdLife International, Cambridge, UK.

van Apeldoorn, R.-C., 2002. The root vole (*Microtus oeconomus arenicola*) in the Netherlands: Threatened and (un)adapted? Lutra 45 (2), 155–166.

van Apeldoorn, R. C., Hollander, H., Nieuwenhuizen, W., Van Der Vliet, F., 1992. The Root vole in the Delta area: is there a relation between habitat fragmentation and competetion at landscape scale? (in Dutch). Landschap : tijdschrift voor landschapsecologie en milieukunde 9 (3), 189–195.

VanDerWal, J., Shooa, L. P., Grahamb, C., Williams, S. E., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220, 589–594.

Worton, B. J., 1995. Using Monte Carlo simulation to evaluate kernel-based home range estimators. Journal of Wildlife Management (4), 794–800.

Zaniewski, A. E., Lehmann, A., Overton, J. M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological modelling 157 (2-3), 261–280.