



# Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data

Stijn M. Bierman, Adam Butler, Glenn Marion and Ingolf Kühn

*S. M. Bierman (stijn.bierman@wur.nl), A. Butler and G. Marion, Bioss – Biomathematics and Statistics Scotland, James Clerk Maxwell Building, The King's Building, Edinburgh, EH9 3JZ, UK. (Present address of S. M. B.: IMARES, Inst. for Marine Resources and Ecosystem Studies, PO Box 68, NL-1970 AB IJmuiden, The Netherlands.) – I. Kühn, UFZ – Helmholtz Centre for Environmental Research Leipzig-Halle, Dept of Community Ecology (BZF), Theodor-Lieser-Str. 4, DE-06120 Halle, Germany.*

Biological atlases are, for many species, the only source of information on their distribution over large geographical areas, and are widely used to inform models of the environmental distribution of species. Such data are not collected using standardized survey techniques, however, and spatial variations in coverage (the relative extent or completeness of records) may lead to variations in the probability that the species will be recorded at locations where it is present (the “recording probability”). If spatial patterns in recording probabilities are correlated with key environmental variables, then biased estimates of the relationships between environmental variables and species distributions may be obtained. We outline a general statistical framework for modelling the environmental distribution of species using, known as Bayesian Image Restoration (BIR). BIR can be used in combination with any species distribution model, but in addition allows us to account for spatial heterogeneity in recording probabilities by utilizing expert knowledge on spatial patterns in coverage. We illustrate the methodology by applying it to maps of the recorded distribution of two plant species in Germany, taken from the German atlas of vascular plants. We find that estimated spatial patterns in recording probabilities for both species are correlated with key environmental variables. Consequently, different relationships between the probability of presence of a species and environmental variables were obtained when the species distribution models were parameterised within a BIR framework. Care must be taken in the application of BIR, since the resulting inferences can depend strongly upon the modelling assumptions that are adopted. Nevertheless, we conclude that BIR has the potential to make better use of uncertain information on species distributions than conventional methods, and can be used to formally investigate the robustness of inferences on the environmental distribution of species to assumptions concerning spatial patterns in recording probabilities.

Biological atlases are databases that consist of records of observed presences of species in cells of a rectangular grid that has been superimposed on the landscape. A species is recorded as present in a grid cell if the database holds at least one record of that species from a location within that grid cell. The data usually consist of the results of both coordinated regional or national surveys and collated historical records, resulting from reported sightings, which have often been collected by thousands of volunteers over many years. Biological atlas data are increasingly used in, for example, reserve design (Araújo et al. 2004), bioclimate envelope modeling (Thuiller et al. 2005), or the mapping of species-specific traits (Kühn et al. 2006).

Difficulties with the statistical interpretation of biological atlas data are widely recognized (Rich and Woodruff 1996, Telfer et al. 2002, Pearce and Boyce 2006). In particular, the recording probability (or alternatively: “detection probability”), the probability that a species is recorded given that it is present somewhere within a grid cell, can be expected to vary between grid cells. We note

that we will use the terms “detection probability” and “recording probability” interchangeably in this paper. Experts often realize that there are spatial patterns in the relative coverage (“coverage” hereafter) of locations (typically cells on a regular grid). We define coverage as the relative extent or completeness of records from these locations obtained by an atlas project (Rich and Woodruff 1996, Telfer et al. 2002). Differences in coverage between grid cells can be expected to be the major determinant of differences in the recording probabilities of a given species between these grid cells. Spatial patterns in coverage may result from a large number of processes such as spatial variation in recording activity (e.g. numbers of volunteers), accessibility of terrain, or the ability or willingness of (groups of locally organized) individuals to hand over their records to a central (national or supra-national) database. For example, systematic differences among German regions (i.e. political units) based not only on sampling intensity but also on differences in the taxonomic concepts and recording lists used were identified by Mahecha and

Schmidtlein (2008). This is a major problem in the use of atlas data for informing species distribution models, because biased estimates of species–environment relationships may be obtained if spatial patterns in coverage are correlated with key environmental variables (Hirzel et al. 2002, Gu and Swihart 2004). For clarity, we note that spatial patterns in recording activity (possibly coinciding with environmental variables) can also be referred to as “sampling bias”, whereas recording probabilities can be assumed to be a function of the detectability of species at specific locations. Here, we focus on describing a formal statistical framework for incorporating spatial patterns in recording (or detection) probabilities, arising from sampling bias (which leads to patterns in detectability), into species distribution models.

Because atlas data are not collected using standardised survey protocols, the main body of (robust) design-based methods in the ecological statistical literature for the estimation of species recording probabilities, such as those based on repeated visits to sites (Mackenzie et al. 2002, Tyre et al. 2003, MacKenzie 2005, Sargeant et al. 2005, Royle et al. 2007), are not applicable. Also, atlas data do not contain information on the abundance of species, and modeling approaches which rely on relating local abundance to local probability of recording the presence of a species (Royle et al. 2007) cannot be used. Instead, explicitly model-based approaches to estimate recording probabilities will have to be developed, which formally utilize expert knowledge on spatial patterns in coverage. This means that, out of necessity, inferences of these models will depend strongly upon the modeling assumptions made, and it will typically be difficult to assess these assumptions. Nevertheless, it seems worthwhile to develop such a modeling framework in order to make potentially better use of existing uncertain information on species distribution and expert knowledge, and to be able to formally investigate the robustness of inferences on the environmental distribution of species to assumptions concerning spatial patterns in recording probabilities.

Here, we propose that expert knowledge on coverage can be formally utilized through Bayesian Image Restoration (BIR) techniques, within the statistical framework as proposed by Heikkinen and Högmänder (1994). The structure of the data in gridded species distribution maps is essentially the same as the data in binary images which are degraded because of the process of non-detection. Heikkinen and Högmänder draw on the extensive statistical image analysis literature (Besag 1986) in which techniques are described to infer the underlying “true” image by modeling both the properties of the image itself (the species distribution) and the process of degrading (non-detection); hence the term “image restoration”. In general terms, BIR combines a model for the recording process and a species distribution model. Within the BIR framework, the parameters of a species distribution model can be estimated jointly with location-specific recording probabilities by relating these probabilities to a proxy variable for coverage. It is important to note that this will only be possible under the assumption that species are not misclassified. Thus, a recorded presence of a species at a location will have to be assumed to represent a true presence. For the taxa under study here, however, misclassification rates are thought to be much lower than non-recording rates (Kühn unpubl.).

Here, we introduce the BIR methodology into the species distribution modeling literature, by describing both the theory underlying BIR, and how to implement it using WinBugs (Spiegelhalter et al. 1999) a freely available software package commonly used to implement Bayesian methods. Furthermore, we extend the work of Heikkinen and Högmänder (1994) by incorporating environmental (climatic, geological and land-use) variables as predictors to model the true underlying spatial distribution of the species. We illustrate the methodology by modeling the distribution of plant species as recorded in a digitized distribution atlas of plant species covering the whole of Germany, using a variable that experts believe reflects the relative difference in the completeness of coverage.

## Methods

### Species distribution data

FLORKART (<[www.floraweb.de](http://www.floraweb.de)>) contains >14 million records on plant species in grid cells of 10' longitude by 6' latitude (ca 130 km<sup>2</sup>; Fig. 1). FLORKART is essentially a central database including all provincial and regional mapping schemes of vascular plants in Germany, maintained by the Federal Agency for Nature Conservation (Bundesamt für Naturschutz) on behalf of the German Network for Phytodiversity (NetPhyD). The data were collected by thousands of volunteers who contributed records of the flora in their respective grid cells.

We illustrate the BIR methodology by modeling the distribution of two plant species as recorded in FLORKART in relation to environmental variables, namely: *Papaver argemone* (prickly poppy), and *Galium pumilum* (slender bedstraw). The distribution of *Galium pumilum* (which is native to Germany) in FLORKART (Fig. 2a) includes several so-called micro-species of which *G. sterneri* is restricted to sand dunes along the North-Sea coast and two other are restricted to higher altitudes in south-eastern Bavaria (*G. valdepilosum*) and the Alps (*G. anisophyllum*). *Galium pumilum*, though wide spread across Germany, is locally rare. It is a species of acidic nutrient poor grasslands, and open shrubs, groves and oak forests (Oberdorfer 1994, Jäger and Werner 2002). In Germany, such conditions are mostly found at mid- to high altitudes. *Papaver argemone* became naturalized in Germany during the course of the Neolithic expansion of agricultural land use practices (Coward et al. 2008). It is widespread (Fig. 2b) and regularly found on sandy, nutrient rich (but avoiding lime) fields and disturbed areas such as road verges or railway tracks (Oberdorfer 1994, Jäger and Werner 2002).

We do not have access to covariate information on these specific habitat requirements. Furthermore, at the coarse spatial grain of the FLORKART grid, the associations of these species with their habitats may (partly) be obscured. Therefore, broad-scale environmental variables on the climate, land use and geology are often used as proxy variables for these habitat conditions when modelling biological atlas data. These proxy variables do not necessarily determine the distribution of a species, but can represent several relevant environmental factors that do. For example, as described above, suitable habitat conditions for

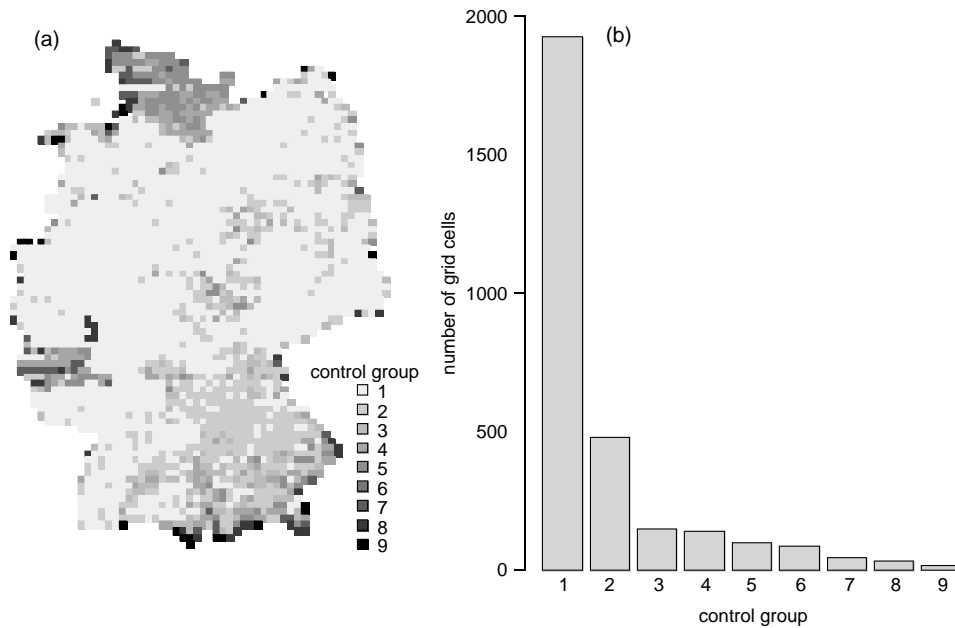


Figure 1. Expert knowledge on spatial variation in the coverage of the German atlas of vascular plants (FLORKART). (a) Map of the proxy variable for spatial variation in relative coverage. The proxy variable consists of 9 different categories (“control groups”), and each grid cell is allocated to one of these (see text). Recording probabilities will be estimated independently for each category (but assumed to be identical for each grid cell within the same category). (b) The number of grid cells in each of the 9 different control groups.

*Galium pumilum* tend to occur at mid- to high altitudes in Germany. Here, we use covariates on the geology, land use and climate to model the underlying distribution of the species. Geological data was aggregated from the Geological Survey Map of Germany (Bundesanstalt für Geowissenschaften und Rohstoffe 1993), and we used the areas of substrate classes lime, sand, and loess per grid cell. The following land use data, provided by the Corine Land Cover data sets (Statistisches Bundesamt 1997) were used as explanatory variables: area of agricultural fields, agricultural grasslands, (semi-)natural grasslands and deciduous forests. For climate data, we used the mean annual temperature (1960–1990),

mean annual precipitation (1950–1980) and mean wind speed (10 m above ground, 1960–1990) provided by the German Meteorological Service (Deutscher Wetterdienst, Dept Klima und Umwelt), and interpolated into the grid cells we used by the Federal Agency for Nature Conservation.

### Expert knowledge on relative coverage

Coverage is thought to be heterogeneous over the FLORKART grid, and is evaluated by experts (coordinators of the mapping project) using 50 “control” species (Kühn et al.

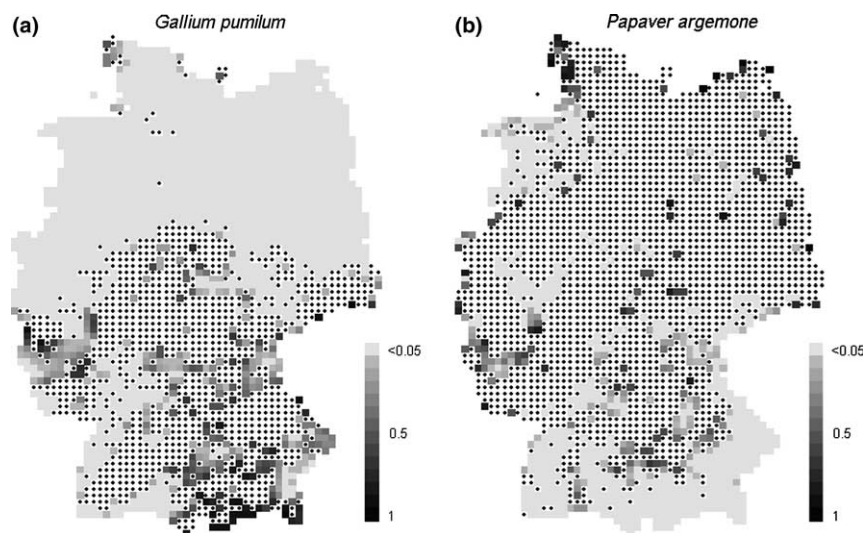


Figure 2. The distribution maps of *Galium pumilum* (a) and *Papaver argemone* (b), as estimated by applying the autologistic species distribution models within a Bayesian Image Restoration framework. Depicted are the recorded presences of the species in grid cells (white cells with filled circles), and the posterior probabilities that species were present in grid cells where the species were not recorded (gray-tones; see legend on maps).

2004). The set of control species consisted of the 45 most common species in Germany (Krause 1998) and five additional species that were relatively inconspicuous or regarded by volunteers as difficult to identify. All of these 50 species could reasonably be assumed to occur in every 6' longitude by 10' latitude grid cell in Germany. The 5 inconspicuous species were included to minimize potential observer bias towards species that are relatively common (abundant in the grid cells) and easy to identify. In each grid cell the number of recorded species out of these 50 control species was counted, yielding a variable between 0 and 50 that we assume is related to the probabilities that species were present in a grid cell, but not recorded. In this paper, we estimated grid-cell specific probabilities of non-detection in 9 different groups of grid cells ("control groups" hereafter), with the following number of control species out of a total of 50 present: 50 (group 1), 49 (group 2), 48 (group 3), 47 (group 4), 46 (group 5), 44–45 (group 6), 40–44 (group 7), 20–39 (group 8), and 0–19 (group 9). The numbers of grid cells that fall within these different control groups are given in Fig. 1. This 9-level categorical variable was believed to be a good proxy variable for relative differences in the coverage for relatively common and relatively conspicuous species. The chosen number of categories (9) is arbitrary, but was high enough to provide enough spatial resolution to reflect the spatial pattern in coverage, while still ensuring that there are enough grid cells in each group to provide the power to estimate these probabilities.

Recording probabilities will be estimated for each category, and assumed to be identical for all grid cells within the same control group. Furthermore, recording probabilities for the different categories of grid cells will be estimated independently. Thus, we will make no a-priori assumption about the potential relationship between control group number and recording probabilities, even though the expectation is that estimated recording probabilities will decrease with increasing control group number.

### Incorporating recording probabilities in species presence/absence models

To calculate the probability that the absence of a recorded presence of a species in a grid cell represents an unobserved "true" presence, we need a model to calculate the probability that the species was present in the grid cell at the time of the survey. This probability of presence is calculated using a presence/absence model, whose parameters we will denote by  $\theta$ . The presence/absence model is explained in detail in the next section. In order to define the detection model, let  $O_i$  denote the recorded (observed) state of grid cell  $i$  ( $O_i=0$  for a recorded absence, and  $O_i=1$  for a recorded presence), and  $M_i$  the unobserved "true" state of grid cell  $i$  ( $M_i=0$  if the species was absent, and  $M_i=1$  if the species was present). Here, we assume that all recorded presences in the data set represent true presences:  $P(M_i=1|O_i=1)=1$ . In contrast, all grid cells with no recorded presences will have a certain probability,  $P(M_i=1|O_i=0)$ , to represent true presences, depending on the probability of non-detection,  $q_i$ , and the probability that grid cell  $i$  was occupied given the parameterized presence/absence model

( $\theta$ : see next section). The probability that a species was in truth present in grid cell  $i$ , while no presence was recorded in that grid cell, can be calculated using Bayes theorem:

$$P(M_i=1|O_i=0, \theta) = \frac{P(O_i=0|M_i=1)P(M_i=1|\theta)}{P(O_i=0|M_i=1)P(M_i=1|\theta) + P(O_i=0|M_i=0)P(M_i=0|\theta)}, \quad (1a)$$

with  $P(O_i=0|M_i=1)$  the probability of non-detection,  $P(M_i=1|\theta)$  and  $P(M_i=0|\theta)$  the probabilities that the species was present and absent from grid cell  $i$  given the parameterised presence/absence model respectively, and  $P(O_i=0|M_i=0)$  the probability that a grid cell was recorded to be empty given that the species was in truth absent. Since we assume that presence will never be recorded when the focal species is in truth absent from a grid cell, or  $P(O_i=0|M_i=0)=1$ , we can rewrite eq. 1a as:

$$P(M_i=1|O_i=0, \theta) = \frac{q_k}{q_k + P(M_i=0|\theta)/P(M_i=1|\theta)}, \quad (1b)$$

with  $q_k=P(O_i=0|M_i=1)$  the probability of non-detection in grid cell  $i$ , which belongs to one of the 9 categories  $k=\{1, 2, \dots, 9\}$ .

Thus, we will estimate non-recording probabilities separately for the 9 different groups of grid cells, and assume that these probabilities are constant within each of these groups of grid cells. Recording probabilities are simply calculated as 1 minus the non-recording probabilities:  $1 - q_k$ .

### The presence/absence model

To model the probabilities that species were present or absent in grid cells (the species distribution model), we use the autologistic model (Augustin et al. 1996). However, we note that the BIR framework would lend itself to any other model that predicts species presence/absence. The autologistic model can be defined as follows:

$$P(M_i=1|\theta) = \frac{\exp(a_0 + a_1 N_i + a_2 X_i^1 + \dots + a_n X_i^n)}{1 + \exp(a_0 + a_1 N_i + a_2 X_i^1 + \dots + a_n X_i^n)}, \quad (2)$$

where  $\theta=(a_0, a_1, \dots, a_n)$  denotes the set of model parameters of the presence/absence model, and  $X_i^1 \dots X_i^n$  the value of environmental covariates in the grid cells (see

below for further explanation). The covariate  $N_i = \frac{\sum_{j=1}^{k_i} M_{ij}}{J_i}$

indicates the "neighborhood weight", with  $j$  an index for grid cells in the neighborhood of cell  $i$ , and  $J_i$  the number of grid cells in the neighborhood of grid cell  $i$ . Grid cells both directly adjacent and diagonal to grid cell  $i$  are included in the neighborhood (a second-order neighborhood), so  $J_i=8$  for all grid cells except those at the edges of the map. The neighborhood weight is included as a predictor, because it is often reasonable to assume spatial smoothing in the spatial distribution of species at very coarse spatial grains such as in the German plant atlas.

We fitted the autologistic models to the observed maps of recorded presences of the two species within a Bayesian framework (explained below) under two scenarios: 1) without estimating recording probabilities, and thus assuming that all pseudo-absences represented true absences, and: 2) jointly estimating the parameters of the autologistic model and the spatially varying recording probabilities: the BIR framework. Variable selection using the BIR model is unpractical because of the long computing times involved in estimating the parameters. Therefore, we modeled each of the species presence/absence data sets using the classical (not Bayesian) generalized linear model with binomially distributed errors and a logit link. This meant that models could be fitted quickly, at the expense of ignoring potential non-detections. For each species, a subset of the variables from a full model including all variables as regressors was chosen using backwards stepwise selection as implemented by the stepAIC function in R (Venables and Ripley 2002, R Development Core Team 2004). The stepAIC function drops covariates from the model if their contribution to the overall explained variance is not significant, as determined by the change in the Akaike information criterion between the model with and without this covariate. All covariates that were selected in the final models were chosen as covariates in the Bayesian autologistic models with and without BIR (scenarios 1 and 2 as described above).

### Estimating the model parameters

We take a Bayesian approach to the estimation of all unknowns,  $v$ , which in the BIR case consist of the distribution model parameters,  $\theta$ , the non-recording probabilities,  $q_k$ , and the underlying presence/absence status  $M_i$  of all grid cells with no recorded presence. Estimation of  $v$  is based on the posterior distribution formed by combining prior knowledge about these unknowns with information provided by the data ( $Y$ ) using Bayes theorem (see for example Brooks 2003):

$$f(v|Y) \propto f(Y|v) \times \pi(v),$$

where  $\pi(v)$  denotes the prior knowledge about  $\theta$ , and  $f(Y|v)$  the likelihood of observing the data given the parameters.

Markov chain Monte Carlo (MCMC) algorithms can be used to draw a set of values from the joint posterior distribution  $f(v|Y)$  (Brooks 2003). These sets of values can be used to obtain summary statistics (such as the mean and standard deviation) of the posterior distribution of each individual parameter. It is possible to implement the autologistic model within a BIR framework using WinBugs (Spiegelhalter et al. 1999), which greatly facilitates the implementation of this methodology. We provide annotated WinBugs code in Supplementary material Appendix 1. For the interested reader, and those who do not wish to use WinBugs to implement the models, we describe the MCMC algorithm that one can use to iteratively draw values from the conditional posterior distributions of the parameters in more detail in Supplementary material Appendix 2.

We adopted vague priors for all parameters, indicating that we assumed no knowledge on the parameters prior to fitting the model to the data, and that the inferences were dominated by the likelihood. Our priors for the

non-recording probabilities of the 9 different groups of grid cells were:  $q_k \sim \text{Beta}(1,1)$ , which is identical to the prior belief that all values for these probabilities that lie in the interval (0,1) were equally likely. Our priors for all parameters of the autologistic model were location invariant ( $\sim \infty, \infty$ ).

## Results

The following environmental variables (apart from the spatial weight) were selected as predictors in the autologistic model for the distribution of *Galium pumilum*, after the stepwise model selection using AIC: altitude, altitude-squared, and area of deciduous forest. This is consistent with our knowledge of the distribution of this species, whose suitable habitat in Germany tends to be found at mid- to high altitudes and in oak forests (see the motivating data section). Precipitation and temperature were selected as environmental predictors in the model for the distribution of *Papaver argemone*. This reflects the fact that this species is widespread, but largely absent from the south of Germany which tends to be relatively colder and receives more precipitation.

Non-recording probabilities for grid cells in control group 1 (grid cells with all 50 control species present) were estimated to be small for both species, with posterior means (95% CI) of  $q_1 = 0.012$  (0.001, 0.032) for *Galium pumilum*, and  $q_1 = 0.002$  (0, 0.007) for *Papaver argemone*. Thus, the majority of grid cells were estimated to have small non-recording probabilities (about two-thirds of all grid cells fall within control group 1: Fig. 1b). However, in line with our expectation, estimates of  $q_k$  tended to increase with increasing control group number  $k$  for both species (Fig. 3). For example, for *Papaver argemone*, grid cells in control groups 8 and 9 had associated non-recording probabilities close to 1, indicating that these grid cells can approximately be regarded as missing data. For *Galium pumilum*, the estimated non-recording probabilities increased more or less continuously with increasing control group number, although similar values were estimated for grid cells in control groups 4, 5 and 6 (Fig. 3). However, for *Papaver argemone*, grid cells in control groups 4, 5 and 6 had smaller associated non-recording probabilities than grid cells with control group number 3 (Fig. 3).

The restored species distribution maps are given in Fig. 2, which depicts the posterior mean probability of presence in each grid cell. The estimated number (mean and standard deviation) of grid cells with no recorded presence that were in truth occupied by *Galium pumilum* (false absences) was 123 (20.6) out of 918 grid cells with no recorded presence (13.4%). A slightly smaller number of false absences was estimated for *Papaver argemone*: 88 (12.4) out of 809 grid cells with no recorded presence (10.9%). The spatial distribution of the estimated recording probabilities was such that the recording probabilities were correlated with environmental variables that were selected in the autologistic models as predictors of species presence, in particular altitude (Fig. 4, 5). As a direct consequence of this, estimates of the parameters of the autologistic models of one of the species (*Galium pumilum*) changed when these models were applied within a BIR framework (Table 1 and

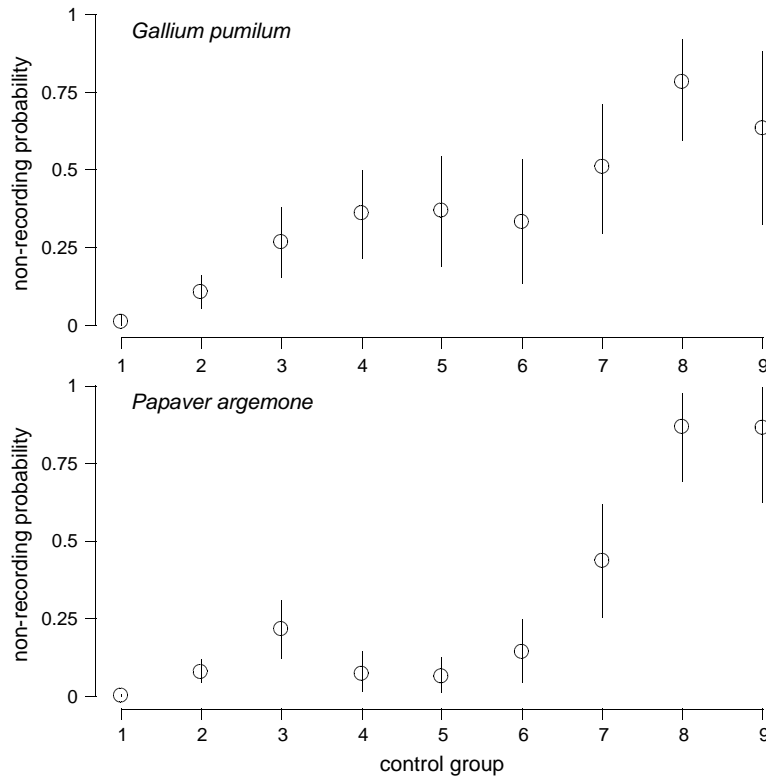


Figure 3. The estimated non-recording probabilities in the grid cells belonging to the 9 different control groups ( $q_k = P(O_i = 0 | M_i = 1)$ ,  $k = \{1, 2, \dots, 9\}$ ). Depicted are the means (circles) and 95% credible intervals (vertical segments) of the posterior distributions of  $q_k$  for *Galium pumilum* and *Papaver argemone*.

Fig. 4a, b). The estimated values for the intercepts were higher within the BIR framework, reflecting the higher number of true presences in the restored species distribution maps. The amount of spatial smoothing was also always estimated to be stronger within the BIR framework, because (for these species) non-detection events were distributed such that they caused small gaps within the distribution of the species (Fig. 2). For *Papaver argemone*, changes in estimated means and standard deviations of the posterior distributions of the slopes for precipitation and temperature were small, and our conclusion would be that these slopes were significant both with and without the BIR framework. In contrast, for *Galium pumilum*, changes in parameter estimates for the slopes of the environmental predictors were pronounced (Table 1). A significant quadratic effect of altitude on the probability of presence of *Galium pumilum* was estimated using the autologistic model without BIR (Table 1), and a sharp increase from low- to mid altitudes, followed by a sharp decrease from mid- to high altitudes was predicted (Fig. 4). However, this quadratic effect was estimated to be insignificant when the autologistic model was fitted using BIR (Table 1), resulting in a very different predicted relationship between altitude and probability of presence (Fig. 4). This change is due to the correlation in the spatial patterns in altitude and recording probabilities (Fig. 4). Consequently, many grid cells without recorded presences at mid-to-high altitudes (the north of Germany is mostly flat, while altitudes tend to increase towards the south of Germany) fell into control groups  $k=2$  or higher

and were thus estimated to be part of the underlying true distribution of the species.

## Discussion

We have illustrated how BIR can be used to jointly model the environmental distribution and spatially varying recording probabilities of species, by combining the information contained in the recorded presences in biological atlases and expert knowledge on the existence of spatial patterns in the relative coverage. The estimated recording probabilities for two species of vascular plants, *Galium pumilum* and *Papaver argemone*, exhibited substantial spatial heterogeneity, with the patterns of variation coinciding with key environmental variables which were used in the distribution model for these species. It is well known that biased estimates of species-environment relationships may be obtained if spatial patterns in recording probabilities are correlated with spatial patterns in key environmental variables (Hirzel et al. 2002, Gu and Swihart 2004, Hortal et al. 2007). However, existing methods to analyze presence-only data, such as ecological niche factor analysis (Hirzel et al. 2002), envelope or profiling methods (Walker and Cocks 1991), or re-sampling methods (Ferrier et al. 2002, Engler et al. 2004) are not parameterized within a framework which allows a formal assessment of the potential impact of spatial patterns in coverage (sampling bias) on the inferred species distribution models. Thus, these methods implicitly assume that key environmental drivers of the species distribution

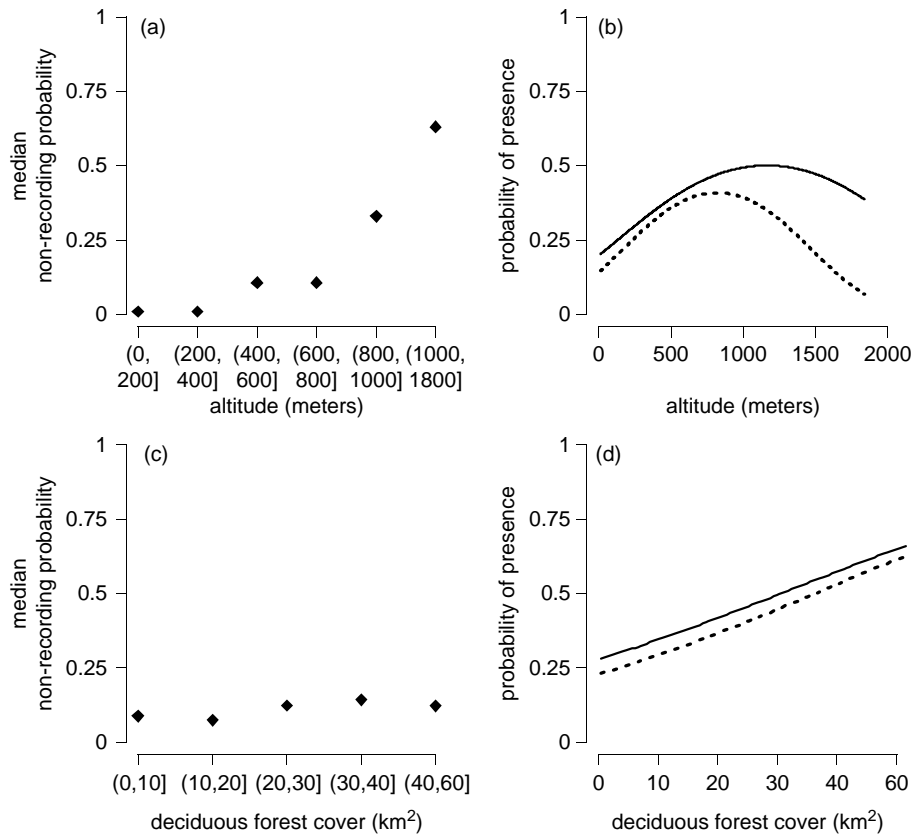


Figure 4. Correlations between environmental variables and estimated non-recording probabilities of grid cells, and the environmental gradients in the probability of species presence as inferred with or without BIR, for *Galium pumilum*. (a) The median value of the estimated non-recording probabilities (means of the posterior distributions for  $q_k$ ) of grid cells with similar altitudes increased sharply with increasing altitude. (b) The relationship between the probability of presence and altitude, as inferred with (solid line) and without BIR (dotted line). (c) The median values of estimated non-recording probabilities of grid cells were not strongly related to amounts of deciduous forest cover. (d) The relationship between probability of presence and deciduous forest cover, as inferred with (solid line) and without BIR (dotted line).

and detection rates are uncorrelated. Due to the estimated relationships between environmental variables and the probability of recording (Fig. 4) our perception of the environmental distribution of the species changed when we applied the logistic models within the BIR framework (Table 1, Fig. 4b). The estimates obtained using BIR can be regarded as better estimates provided that the estimated relationship between the proxy variable for coverage and detection probabilities, and the obtained distribution map is deemed to provide a more plausible description of reality than those obtained under the assumption of perfect detection. At the very least, the estimates obtained under BIT provide alternative estimates, as obtained under different assumptions regarding the observation process.

We note that we have used the term “expert knowledge” on relative coverage loosely here. Expert knowledge on relative coverage in this case does not mean that there is no data available to construct proxy variables. Nevertheless, we use this terminology because it is the judgement of experts that it is plausible that such a proxy variable is related to spatial variation in coverage for the species of interest. Subsequently, BIR can be used to estimate this relationship. Here we have used a number of species that can reasonably be expected to occur in all grid cells to construct such a proxy variable. However, other information could be used

to construct proxy variables, such as the slopes of species accumulation curves as a measure of survey completeness (Hortal et al. 2004, Hortal and Lobo 2005), or spatial measures of recording activity such as numbers of submitted records (Soberón et al. 2007) or published floras (Rich 2006).

The main aim of this paper was to introduce and illustrate BIR methodology for the analysis of species atlas data. We have presented here only the basic BIR framework, and this could be extended or altered. Other models than the autologistic may be used for the probability of presence of the species. In fact, any parametric model in which probabilities of presence  $P(M=1|\theta)$  in eq. 1a are predicted will be suitable. Also, we did not make any a-priori assumption on the functional form of the relationship of the proxy variable for coverage (control group number) and recording probabilities. However, it will be possible to, for example, constrain recording probabilities to increase with increasing values of the proxy variable (Heikkinen and Högmander 1994). The methodology presented in this paper extends the methodology of Heikkinen and Högmander (1994) by including a species distribution model in which the distribution of the species is modelled as a function of environmental variables. Also, BIR can be seen as a generalisation of the methodology as

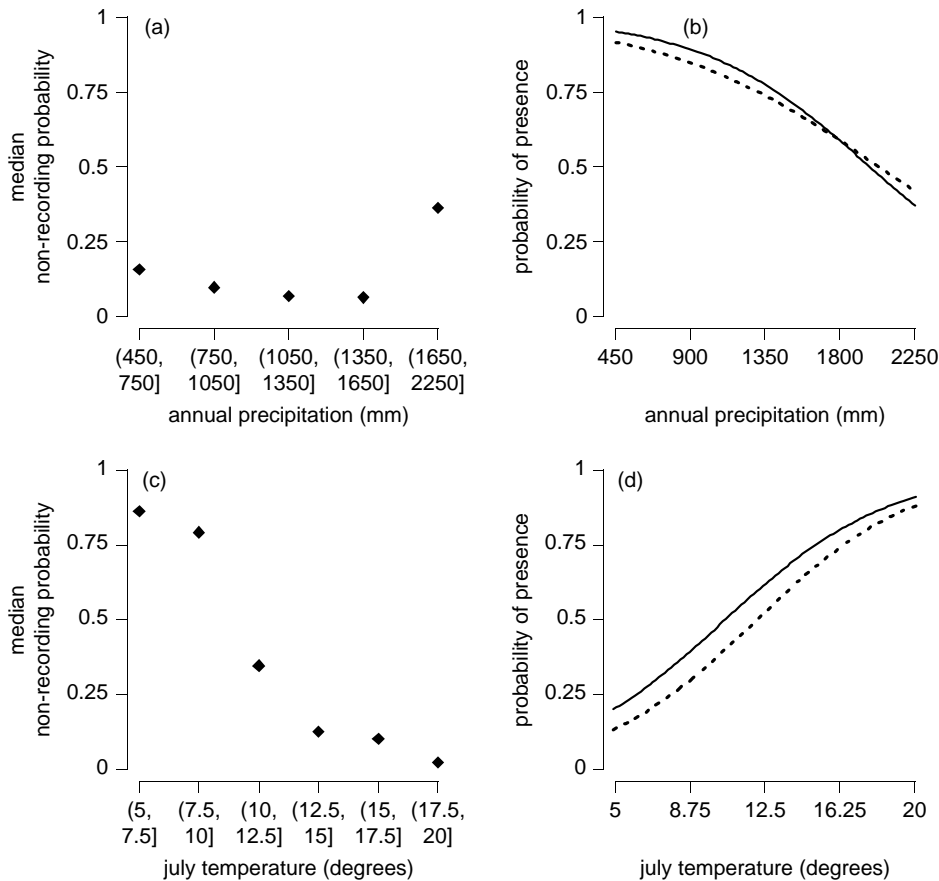


Figure 5. Correlations between environmental variables and estimated non-recording probabilities of grid cells, and the environmental gradients in the probability of species presence as inferred with or without BIR, for *Papaver argemone*. (a) The median value of the estimated non-recording probabilities (means of the posterior distributions for  $q_k$ ) of grid cells with similar amounts of annual precipitation. (b) The relationship between the probability of presence and annual precipitation, as inferred with (solid line) and without BIR (dotted line). (c) The median value of the estimated non-recording probabilities of grid cells with similar July temperature increased sharply with temperature. (d) The relationship between probability of presence and July temperature, as inferred with (solid line) and without BIR (dotted line).

introduced by Augustin et al. (1996) to fit the autologistic model to distribution data where a proportion of squares was completely unsurveyed, by assigning non-detection probabilities of 1 to these squares (thus treating them as missing data). Instead, BIR allows for varying non-detection probabilities over the grid, and can be used to estimate these

in conjunction with the parameters of the species distribution model.

Table 1. The estimated means and standard deviations (comparable to standard errors of classical statistics) of the posterior distributions of the parameters of the autologistic models (eq. 2), fitted either with or without Bayesian Image Restoration. The variable “weight” indicates the neighborhood weight ( $N_i$ ; eq. 2).

	Parameter	Without BIR Mean (SD)	With BIR Mean (SD)
<i>Galium pumilum</i>	Intercept	-0.976 (0.068)	-0.793 (0.086)
	Altitude	0.0023 (0.0004)	0.00183 (0.00048)
	Altitude <sup>2</sup>	-0.002 (0.0005)	-0.001 (0.001)
	Forest	1.958 (0.548)	1.820 (0.599)
	Weight	5.297 (0.232)	5.856 (0.290)
<i>Papaver Argemone</i>	Intercept	1.455 (0.063)	1.779 (0.096)
	Precipitation	-0.0012 (0.0004)	-0.0016 (0.0004)
	Temperature	0.205 (0.048)	0.197 (0.054)
	Weight	5.31 (0.24)	5.68 (0.24)

Because the inferences made using BIR are strongly model-based, it is a modeling framework best suited for making best use of data for taxa where the apparent distribution maps in biological atlases are thought to be degraded by heterogeneity in coverage. However, even for taxa where the distribution maps are thought to reflect the true underlying distribution well, BIR can be usefully applied to test the assumption that potential spatial patterns in recording probabilities are of negligible influence on the estimated parameters of species.

We note that we have assumed that recording probabilities are unrelated to the underlying distribution model (mathematically; our assumption is that  $P(O=0|M=1) = P(O=0|M=1, \theta)$ ). This may be unrealistic, if for example the species is more abundant in the most suitable parts of its range and if detection increases with abundance (which may be likely: see for example Royle et al. 2007). However, since our model for the underlying distribution in this case describes only the probability of presence this would require a link to be made between  $P(M=1|\theta)$  and  $P(O=0|M=1)$  which is beyond the scope of this paper. We note that Royle

et al. (2007) have made this link, but they had abundance data available. Their method is conceptually similar, in that an underlying distribution and abundance model is used which included both a spatial model describing variation in the abundance of territories of a bird and an detection model. However, they relied on the availability of repeated visits to territories which yields independent information on detection probabilities. We stress that such information is much better, and yields inferences that are less strongly model-based than the BIR methodology described here, which we propose should only be utilized for species atlas data in the absence of no independent information on detection probabilities or abundance data.

The BIR methodology presented here has allowed us to make use of all of the information available (all grid cells with none or at least one recorded presence and the control variable to capture coverage) and to relax the critical assumption that the recorded presences provide a good representation of the environmental distribution of the species. Also, the uncertainty in the imputed underlying (partly unobserved) distribution of the species is reflected in the estimated variance of the parameters of the species distribution model. However, we stress that Bayesian image restoration techniques for the modeling of species distributions should be applied with care. The validity of the results depend on the reliability of the model assumptions, and it is therefore important to make these as biologically plausible as possible. There are four main components to this. Firstly, the probability that a species is recorded, if it is in truth absent from a location, is assumed to be zero. This is not necessarily the case, because a species may be misclassified for another species which is absent from a location. The validity of this assumption will therefore have to be assessed by experts on a species by species basis. Secondly, the validity of the recording model depends on the availability of good proxy variables for coverage. Thirdly, the validity of the underlying species distribution model is dependent on the availability of suitable explanatory variables, and/or the existence of other characteristics of the species distribution such as spatial smoothness in probabilities of presence. In particular, it should be realized that it is possible that the recording model will partly correct for biases in the species distribution model by assigning high non-recording probabilities to areas where probabilities that the species occurs are predicted to be higher than is the case in reality. Part of the consideration in formulating a good distribution model is to distinguish between absences of environmental origin (i.e. the suitability of the environment) and absences of contingent origin (i.e. where the environment is suitable but the species absent for a different reason). We note that we have assumed that the two plant species here are present in all grid cells with suitable habitat (thus no absences of contingent origin). This assumption seems reasonable given that the coarseness of the grid cells will mask local dynamics (such as local extinctions and re-colonizations), and the fact that these are species native to Germany and thus have had enough time to establish themselves throughout the country. Fourthly, the less complete the distribution data (i.e. the lower detection probabilities are), the larger the reliance on the model will become, while at the same time there will be less information to inform this model. The combination of both a very uncertain model for the

distribution and little information to estimate detection probabilities will result in a large set of possible outcomes, ranging for example from an inferred distribution that is similar to that of the raw data through to a distribution that fills almost the whole map. If the available distribution data are nearly complete, and thus give a good description of the distribution of the species, there will be much information in the data to inform both the model for the distribution and the recording probabilities. This will result in inferences that are less dependent on the model assumptions. Here, we have used distribution data where most absences represent true absences, and this is the situation in we think which BIR is most successful in terms of inferring possible true distributions of species, whilst estimating spatial variation in detection rates without making strong prior assumptions about either the distribution model or detection rates. However, when detection rates are very low (if most true presences are unobserved), it is unlikely that BIR can be used successfully without assuming strong prior knowledge regarding the distribution of the species and detection rates.

We conclude by summarizing at what point in the modeling of species atlas data we believe BIR will provide a useful tool. First of all, methods other than BIR are available which will make better use of the available information if independent information on detection probabilities is available, for example from repeated visits (Mackenzie et al. 2002, Tyre et al. 2003, MacKenzie 2005, Sargeant et al. 2005, Royle et al. 2007). A suite of methods exists for the modelling of presence-only data, which explicitly account for sample selection bias. These methods have been found to produce reasonable estimates of species distribution models when compared to independent better quality “genuine” presence-absence data (see for example Elith and Leathwick 2007, Phillips et al. 2009). In particular, Phillips et al. (2009) propose to use presence records of other species as proxy variables for sampling bias, which is in spirit similar to what we have done in this paper. However, in the absence of independent “genuine” presence-absence data, the quality of the modeling approach will be impossible to assess. Furthermore, these methods do not formally model detection probabilities, and the uncertainty in the inferences arising from unknown (spatial patterns in) detection probabilities cannot be assessed. We believe that the stage in the modeling of atlas data at which BIR will be most usefully applied is when it has been established that there are no major taxonomic problems (thus when it can be assumed that there will be few misclassifications), and when a good overview exists of the extent of survey bias (see also Hortal et al. 2007). We propose to use BIR only when in addition to the atlas data knowledge exists on spatial patterns in coverage that can be captured in the form of proxy variables. BIR provides a tool to estimate the relationship between such possible proxy variables and detection probabilities, and to assess the reliance of inferences regarding the distribution of species on the assumptions that are made regarding the spatial patterns in detection probabilities.

*Acknowledgements* – We thank the many volunteers who have collected the data over many years and Rudolf May and Hans Fink of the Federal Agency for Nature Conservation who provided the data on behalf of the German Network for Phytodiversity. Thanks

also to Mark Brewer for his help in the development of the WinBugs code. Finally, we thank Joaquín Hortal, an anonymous reviewer, and the subject editor, whose critical comments have greatly improved this manuscript. This work has been partly funded by the European Union within the FP 6 Integrated Project “ALARM” (GOCE-CT-2003-506675) (Settele et al. 2005), and the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

## References

- Araújo, M. B. et al. 2004. Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. – *Global Change Biol.* 10: 1618–1626.
- Augustin, N. H. et al. 1996. An autologistic model for the spatial distribution of wildlife. – *J. Appl. Ecol.* 33: 339–347.
- Besag, J. 1986. On the statistical analysis of dirty pictures. – *J. R. Stat. Soc. B* 48: 259–302.
- Brooks, S. P. 2003. Bayesian computation: a statistical revolution. – *Phil. Trans. R. Soc. A* 361: 2681–2697.
- Bundesanstalt für Geowissenschaften und Rohstoffe 1993. Geologische Karte der Bundesrepublik Deutschland 1:1 000 000. – Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover.
- Coward, F. et al. 2008. The spread of Neolithic plant economies from the Near East to northwest Europe: a phylogenetic analysis. – *J. Archaeol. Sci.* 35: 42–56.
- Elith, J. and Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. – *Divers. Distrib.* 13: 265–275.
- Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – *J. Appl. Ecol.* 41: 263–274.
- Ferrier, S. et al. 2002. Extended statistical approaches to modeling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modeling. – *Biodivers. Conserv.* 11: 2275–2307.
- Gu, W. and Swihart, R. K. 2004. Absent or present? Effects of non-detection of species occurrence on wildlife-habitat models. – *Biol. Conserv.* 116: 195–203.
- Heikkinen, J. and Högmander, H. 1994. Fully Bayesian approach to image restoration with an application in biogeography. – *Appl. Stat.* 43: 569–582.
- Hirzel, A. H. et al. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? – *Ecology* 83: 2027–2036.
- Hortal, J. and Lobo, J. M. 2005. An ED-based protocol for the optimal sampling of biodiversity. – *Biodivers. Conserv.* 14: 2913–2947.
- Hortal, J. et al. 2004. Butterfly species richness in mainland Portugal: predictive models of geographic distribution patterns. – *Ecography* 27: 68–82.
- Hortal, J. et al. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). – *Conserv. Biol.* 21: 853–863.
- Jäger, E. J. and Werner, K. 2002. Rothmaler, Exkursionsflora von Deutschland. Gefäßpflanzen: Kritischer Band, 9 ed. – Spektrum, Berlin.
- Krause, A. 1998. Floras Alltagskleid oder Deutschlands 100 häufigste Pflanzenarten. – *Natur und Landschaft* 73: 486–491.
- Kühn, I. et al. 2004. The flora of German cities is naturally species rich. – *Evol. Ecol. Res.* 6: 749–764.
- Kühn, I. et al. 2006. Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. – *New Phytol.* 172: 127–139.
- MacKenzie, D. I. 2005. Was it there? Dealing with imperfect detection for species presence/absence data. – *Aust. N. Z. J. Stat.* 47: 65–74.
- Mackenzie, D. I. et al. 2002. Estimating site occupancy rates when recording probabilities are less than one. – *Ecology* 83: 2248–2255.
- Mahecha, M. D. and Schmidtlein, S. 2008. Revealing biogeographical patterns by nonlinear ordinations and derived anisotropic spatial filters. – *Global Ecol. Biogeogr.* 17: 284–296.
- Oberdorfer, E. 1994. Pflanzensoziologische Exkursionsflora, 7 ed. – Ulmer.
- Pearce, J. L. and Boyce, M. S. 2006. Modeling distribution and abundance with presence-only data. – *J. Appl. Ecol.* 43: 405–412.
- Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – *Ecol. Appl.* 19: 181–197.
- R Development Core Team 2004. R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria.
- Rich, T. C. G. 2006. Floristic changes in vascular plants in the British Isles: geographical and temporal variation in botanical activity 1836–1988. – *Bot. J. Linn. Soc.* 152: 303–330.
- Rich, T. C. G. and Woodruff, E. R. 1996. Changes in the vascular plant floras of England and Scotland between 1930–1960 and 1987–188: the BSBI monitoring scheme. – *Biol. Conserv.* 75: 217–229.
- Royle, J. A. et al. 2007. Hierarchical spatial models of abundance and occurrence from imperfect survey data. – *Ecol. Monogr.* 77: 465–481.
- Sargeant, G. A. et al. 2005. Markov chain Monte Carlo estimation of species distributions: a case study of the swift fox in western Kansas. – *J. Wildl. Manage.* 69: 483–497.
- Settele, J. et al. 2005. ALARM: assessing LARge scale environmental Risks for biodiversity with tested Methods. – *GAIA* 14: 69–72.
- Soberón, J. et al. 2007. Assessing completeness of biodiversity databases at different spatial scales. – *Ecography* 30: 152–160.
- Spiegelhalter, D. J. et al. 1999. WINBUGS version 1.2 user manual. – MRC Biostatistics Unit, Inst. of Public Health, Cambridge, UK.
- Statistisches Bundesamt 1997. Daten zur Bodenbedeckung für die Bundesrepublik Deutschland 1:100.000. – Statistisches Bundesamt, Wiesbaden.
- Telfer, M. G. et al. 2002. A general method for measuring relative change in range size from biological atlas data. – *Biol. Conserv.* 107: 99–109.
- Thuiller, W. et al. 2005. Climate change threats to plant diversity in Europe. – *Proc. Nat. Acad. Sci. USA* 102: 8245–8250.
- Tyre, A. J. et al. 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. – *Ecol. Appl.* 13: 1790–1801.
- Venables, W. N. and Ripley, B. D. 2002. Modern applied statistics with S, 4 ed. – Springer.
- Walker, P. A. and Cocks, K. D. 1991. HABITAT – a procedure for modeling a disjoint environmental envelope for a plant or animal species. – *Global Ecol. Biogeogr.* 1: 108–118.

Download the Supplementary material as file E5798 from <[www.oikos.ekol.lu.se/appendix](http://www.oikos.ekol.lu.se/appendix)>.