Improving species distribution models: the value of data on abundance

Running title: Improving species distribution models

Christine Howard¹
Philip A. Stephens¹
James W. Pearce-Higgins²
Richard D. Gregory³
and Stephen G. Willis¹

¹ School of Biological and Biomedical Sciences, Durham University, Mountjoy Science Site, Durham, DH1 3LE, UK
² British Trust for Ornithology, The Nunnery, Thetford, Norfolk, IP24 2PU, UK
³ The Royal Society for the Protection of Birds & European Bird Census Council, The Lodge, Sandy, Bedfordshire, SG19 2DL, UK

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12184
This article is protected by copyright. All rights reserved.
Summary:

1. Species distribution models (SDMs) are important tools for forecasting the potential impacts of future environmental changes but debate remains over the most robust modelling approaches for making projections.

2. Suggested improvements in SDMs vary from algorithmic development through to more mechanistic modelling approaches. Here, we focus on the improvements that can be gained by conditioning SDMs on more detailed data. Specifically, we use breeding bird data from across Europe to compare the relative performances of SDMs trained on presence-absence data and those trained on abundance data.

3. SDMs trained on presence-absence data, with a poor to slight fit according to Cohen’s kappa, show an average improvement in model performance of 0.32 (se ±0.12) when trained on abundance data. Even those species for which models trained on presence-absence data are classified as good to excellent show a mean improvement in Cohen’s kappa score of 0.05 (se ±0.01) when corresponding SDMs are trained on abundance data. This improved explanatory power is most pronounced for species of high prevalence.

4. Our results illustrate that even using coarse scale abundance data, large improvements in our ability to predict species distributions can be achieved. Furthermore, predictions from abundance models provide a greater depth of
information with regard to population dynamics than their presence-absence model counterparts. Currently, despite the existence of a wide variety of abundance data sets, species distribution modellers continue to rely almost exclusively on presence-absence data to train and test SDMs. Given our findings, we advocate that, where available, abundance data rather than presence-absence data can be used to more accurately predict the ecological consequences of environmental change. Additionally, our findings highlight the importance of informative baseline data sets. We therefore recommend the move towards increased collection of abundance data, even if only coarse numerical scales of recording are possible.

**Key words:**
Species distribution modelling, ordinal abundance data, presence-absence data, random forests, model performance

**Introduction**

To determine the impacts of future climate and habitat changes on species, ecologists increasingly use species distribution models (SDMs) to quantify species-environment relationships (Guisan & Thuiller 2005). SDMs are now widely used and frequently refined (Guisan & Rahbek 2011; Higgins, O'Hara & Römermann 2012). Nevertheless, confidence in the predictive power of these models continues to be undermined by conceptual, biotic and algorithmic flaws, which include uncertainty regarding variable selection (Austin & Van Niel 2011), unrealistic model assumptions (Schroder & Seppelt 2006; Dormann 2007b), and lack of agreement over the
classification of basic concepts (Segurado & Araújo 2004; Araújo & Guisan 2006; Austin 2007). As a result, ongoing debate concerns the strengths and limitations of SDMs and potential areas for their improvement (Araújo & Peterson 2012). Suggested areas of development range from the incorporation of land cover variables and biotic interactions, to accounting for spatial autocorrelation (Guisan & Thuiller 2005; Araújo & Guisan 2006; Dormann 2007a; Bagchi et al. 2013) and incorporating biological traits (Higgins, O'Hara & Römermann 2012). Methodological improvements may well increase the predictive performance of SDMs (Araújo & Guisan 2006; Austin 2007). Additionally, we might consider what could be achieved by improving the information available for training data sets. Although the relative value of presence-only and presence-absence data has been widely discussed (Brotons et al. 2004; Elith et al. 2006; Pearson et al. 2006), a third, more detailed form of data is available for many taxa in some regions: abundance data. This may either be an index of abundance, for example based on frequency of reporting rates (Harrison & Cherry 1997), or an estimate of true population size, such as derived from surveys accounting for detectability (Renwick et al. 2011). In addition to providing additional information that may be better related to conservation status (Gregory, Noble & Custance 2004; Johnston et al. 2013), extinction risk (O'Grady et al. 2004) and community structure and function (Davey et al. 2012), the greater information content of abundance data could also result in models with a greater ability to discriminate species' range boundaries, and to produce more accurate models of presence-absence. At present, however, there is no indication of the magnitude of improvements in SDMs that could be gained through using abundance rather than presence-absence data.
Based on the assumption that local abundance is an indicator of habitat quality, SDMs derived from abundance data may reflect the importance of key demographic and environmental factors such as carrying capacity (Pearce & Ferrier 2001). Van Horne (1983) cautioned against the assumption that abundance can be used as an indicator of habitat quality, as some environmental factors and species characteristics, such as detectability, can reduce the probability of a positive correlation between abundance and habitat quality. Nevertheless, by using abundance data and increasing the information available to train SDMs, we may be able to improve our ability to predict occurrence. It is therefore important to understand the extent to which structuring presence-absence data through the use of abundance data improves model performance in cases where land cover and spatial autocorrelation have already been incorporated.

A curvilinear relationship between predictive performance of SDMs and prevalence has been widely reported in the literature (Manel, Williams & Ormerod 2001; McPherson, Jetz & Rogers 2004; Allouche, Tsoar & Kadmon 2006), especially when fit is assessed using the kappa statistic (Santika 2011). A positive relationship between range size and mean abundance has also been reported within many taxonomic groups (Brown 1984). With this in mind, we would expect the mean abundance of low prevalence species to be uniformly low across their range, and therefore abundance values to be little more informative than presence-absence data. We might therefore expect the predictive capabilities of models trained on abundance data and models trained using presence-absence data to converge at low levels of prevalence.
Here, we use a machine learning technique, random forests, to model the distribution of European breeding-bird atlas data across the scale of the continent. We analyse the relative performance of models trained on abundance data and those trained on presence-absence data. Additionally, we investigate the role of prevalence on the performance of these models to determine if there are limitations to any benefit associated with abundance modelling.

Methods

Data

Spatial abundance data were available for 345 species of European breeding birds from the EBCC (European Bird Census Council) Atlas of breeding birds (Hagemeijer & Blair 1997). These data record a logarithmically scaled, categorical estimate of the abundance of each species across a 50 x 50 km Universal Transverse Mercator (UTM) grid, mostly representing the period from 1985-1988 (data for a few areas were drawn from slightly earlier/later censuses). Population size estimates are based on a 7-point scale, including 6 logarithmically scaled categories (1-9, 10-99, 100-999, 1000-9,999, 10,000-99,999, ≥100,000 breeding pairs) and 0. These categorical abundance data were simplified to presence/absence data to enable a comparison of the performance of SDMs trained on the two types of data.

Environmental variables

Bioclimatic variables were derived from a global compilation (New, Hulme & Jones 1999) for the 30 year period 1961-1990. This consisted of four bioclimatic variables (mean temperature of the warmest month (MTWM), mean temperature of the coldest month (MTCO), growing degree days above 5° (GDD5), and the annual ratio of
actual to potential evapotranspiration (APET). These variables were calculated at the same resolution as the species data, using the formulation in Prentice et al. (1992).

The specific bioclimatic variables were chosen because all have been shown to describe both the range extents (Thuiller, Araujo & Lavorel 2004; Huntley et al. 2007; Doswald et al. 2009) and abundance patterns (Green et al. 2008; Gregory et al. 2009) of European birds.

Land cover variables were derived from the Pan-European Land Cover (PELCOM) 1km resolution database (Mucher et al. 2000). These data were aggregated to provide percentage coverage at the same resolution as the species data. In total, eight land cover classifications were used: forest, grassland, urban, arable, wetland, coastal, shrub land, marine and barren.

**Statistical modelling**

Random forest (RF) models were used to model species’ distributions from both the abundance and the presence-absence data. This machine learning technique is a bootstrap based classification and regression trees (CART) method (Cutler et al. 2007). Here, to account for a high degree of correlation between climatic covariates (with Pearson’s r ranging between 0.61 and 0.9) and the potential for biased variable selection, we use the party package in R, which uses a random forest implementation based on a conditional inference framework (Hothorn, Hornik & Zeileis 2006a; Hothorn, Hornik & Zeileis 2006b; Strobl, Hothorn & Zeileis 2009; R Development Core Team 2012). As with other classification methods, random forests draw bootstrap samples and a subset of predictors to construct multiple classification trees (Prasad, Iverson & Liaw 2006). The classification trees find optimal binary splits in the selected covariates to partition the sample recursively into
increasingly homogenous areas with respect to the class variable (Cutler et al. 2007). Under the conditional inference framework, unbiased variable selection is achieved by using a linear statistic to test the relationship between covariate and response, selecting the covariate with the minimum $P$-value. This linear statistic is also used to optimise the binary split into each homogenous area (Hothorn, Hornik & Zeileis 2006a; Hothorn, Hornik & Zeileis 2006b; Strobl, Hothorn & Zeileis 2009). In the case of ordinal response variables, a score vector reflecting the ‘distances’ between class levels is combined linearly with the linear statistic altering both the selection and binary splitting of variables according to the scale of the ordinal response data (Hothorn, Hornik & Zeileis 2006b).

Random forests make few assumptions about the distribution of variables, are robust to over-fitting, and are widely recognised to produce good predictive models (Breiman 2001; Liaw & Wiener 2002; Prasad, Iverson & Liaw 2006). These models typically outperform traditional regression based approaches to species distribution modelling and are ideal for modelling categorical and ordinal data (Lawler et al. 2006; Magness, Huettmann & Morton 2008; Marmion et al. 2009). More established approaches to ordinal data modelling include proportional odds and continuation ratio ordinal regression models (Guisan & Harrell 2000). However these models have limiting assumptions, such as parallelism between classes, and lack the flexibility to identify non-linear, context dependent relationships amongst predictor variables (De’ath & Fabricius 2000; Olden, Lawler & Poff 2008; Strobl, Malley & Tutz 2009).

To account for spatial autocorrelation we included a measure of the surrounding abundance of conspecifics in the first order neighbouring UTM grid cells (Segurado, Araujo & Kunin 2006) as a spatial auto-covariate (SAC). This term
accounts for the greater degree of similarity between more proximate samples, that arises through distance-related biological process and spatially structured environmental processes (Dormann et al. 2007). We account for potential spatial autocorrelation in our abundance-based models by calculating an indicator of surrounding abundance for each UTM grid cell, using the following equation:

\[ L = \log_{10} \left( \frac{1}{n} \sum_{i} \frac{1}{2} A_i \right) \]  

(1)

where: \( L \) = surrounding local abundance, \( n \) = number of adjacent cells, \( A \) = categorical abundance, \( i \) = abundance category index. The log scaled abundance categories in the adjacent cells are back transformed to the mid-points of the relevant categories; these are averaged and re-transformed to the log scale. For models based on presence-absence data, the spatial autocovariate used the same equation, except that the abundance categories \( (A_i) \) were converted to binary (presence-absence) data. Models were fitted using ten-fold cross validation to reduce SAC between training and test data and to minimise over-fitting. We used correlograms to compare autocorrelation in the model residuals with autocorrelation present in the raw data. Correlograms plot a measure of spatial autocorrelation, Moran’s I (Moran 1950), between grid cells as a function of the distance between them (Fortin & Dale. 2005; Dormann et al. 2007; Kissling & Carl 2008). A value of zero of Moran’s I for within model residuals indicates an absence of spatial autocorrelation. Therefore, a significant deviation from zero suggests that the model is not adequately accounting for spatial autocorrelation (Dormann et al. 2007). Here, we note that all of our models showed substantial reductions in residual spatial autocorrelation when compared to that present in the raw data (see supporting information figure s1). R code to implement species abundance and distribution modelling using the party package,
along with code to calculate the spatial autocovariate term is available in the supporting information.

Predictions of the probability of a species occurring at each abundance class were based on the number of votes for each class from the 1000 classifiers that comprised each forest (Robnik-Sikonja 2004). Predicted probability across the abundance classes are summed to give a predicted probability of occurrence, whilst predicted ordinal abundance are based on the class with the majority vote. Ordinal predictions from the distribution model based on abundance data were converted to presence-absence data to enable a direct comparison to recorded presence-absence data.

Model fits of simulated presence-absences derived from the abundance (after conversion to presence-absence data) and presence-absence models to observed presence-absence data were assessed using three methods, which included measures of both model calibration and discrimination. We used two measures of discrimination, which indicate the ability of a model to discriminate between species presence and absence. First, the kappa statistic measures model accuracy whilst correcting for accuracy expected to occur by chance (Cohen 1960); we used this on the simulated occurrences from the cross validated datasets. Kappa is the most widely used measure of discrimination and performance for presence-absence models (Manel, Williams & Ormerod 2001; Pearson, Dawson & Liu 2004; Segurado & Araújo 2004; Allouche, Tsoar & Kadmon 2006) but is criticised for being inherently dependent on prevalence and the often arbitrary choice of threshold value (Allouche, Tsoar & Kadmon 2006; Freeman & Moisen 2008). Our second measure of discrimination, therefore, was a threshold independent measure of model
performance, the area under the receiver operating characteristic (ROC) curve (AUC) (Manel, Williams & Ormerod 2001; Thuiller 2003; Brotons et al. 2004).

As a measure of model calibration, we used calibration curves to assess agreement between the logits of the predicted probabilities and the observed proportions of occurrence in the test data (Zurell et al. 2009). The slope and intercept of this regression can provide a measure of model bias and spread (Pearce & Ferrier 2000). Model bias is the systematic over- or under-estimation of the probability of occurrence across the range of a species and results in an upwards or downwards shift of the regression line, causing the intercept to deviate from zero (Reineking & der 2006). The slope of the regression line, fitted to the predicted and observed values on x and y logit axes respectively, indicates the spread of the data. If predicted values lower than 0.5 overestimate the probability of occurrence whilst predicted values greater than 0.5 underestimate the probability of occurrence the slope of the regression line will be greater than one. Conversely a gradient of less than one indicates that predicted values lower than 0.5 are underestimating the probability of occurrence, whilst predicted values greater than 0.5 overestimate the probability of occurrence (Pearce & Ferrier 2000). A perfectly calibrated model will have an intercept of zero and a slope of one (Reineking & der 2006; Zurell et al. 2009; Vorpahl et al. 2012).

We used a paired t-test on logit-transformed data to assess differences between the predictive performances, according to kappa, of models trained on each data set. The effect of prevalence (the proportion of presences out of 2813 cells) on predictive accuracy was assessed using a GAM, after controlling for species (to account for the paired nature of the data set). The model was fitted with a binomial
error structure with a logit link and included species as a random effect, using the mgcv package in R (Wood 2011; R Development Core Team 2012).

Results

Models trained on abundance data, and later converted to presence-absence predictions, were significantly more discriminating than models trained on presence-absence data (Figure 1 a and b; paired t-tests, kappa $t_{344}=13.23$, $p<0.01$, AUC $t_{344}=3.72$, $p<0.01$). Measures of model calibration also showed improved performance in the models trained on abundance data, when compared with models trained on presence-absence data. The measures of the intercept of the calibration curve were significantly different between the two models ($t_{344}= 3.88$, $p<0.01$), with 74% of abundance models having an intercept closer to zero than their presence-absence trained counterpart. This significant difference is also true for the slope of the model calibration curves ($t_{344}= 3.33$, $p<0.01$) with the slopes of the calibration curves from 76% of models showing a greater tendency towards 1 when trained with abundance data rather than presence-absence data. Furthermore, models trained on abundance data generally fitted the observed abundance data well with a mean weighted Cohen’s kappa score (Landis & Koch 1977) of 0.73 (se ± 0.01) (figure 2). The magnitude of the improvement in model performance associated with abundance-trained models varied with the performance of the presence-absence data trained model (Figure 3). For presence-absence data trained models with a poor to slight rating kappa score (i.e. less than 0.2) (Landis & Koch 1977), mean kappa improved by 0.32 (se ± 0.12). Unsurprisingly, the magnitude of benefit declined with the fit of the original model, with minimal improvements among presence-absence data trained models that rated as almost perfect (i.e. with a kappa score greater than 0.8).
Improvements in model accuracy resulting from the use of abundance data depended on the metric of model accuracy used. When that metric was kappa, improvements were most marked for models that had performed poorly when presence-absence data were used (Fig. 3). Poorer performing presence-absence models tended to be those associated with high or low prevalence species (Fig. 4). Indeed, when kappa was used as the metric of model accuracy, a GAM showed that prevalence had a significant quadratic effect on model accuracy ($z=2.55, p=0.01$, $z=1.38, p=0.17$), and that the modelling method was also a significant categorical explanator ($z=2.317, p=0.02$). There was a marginally significant but weak interaction between prevalence and modelling method ($z=0.18, p=0.85$, $z=2.02$, $p=0.04$; Fig. 4). By contrast, when AUC was used as the metric of model accuracy, improvements owing to the use of abundance data were unrelated to both prevalence and the fit of the equivalent presence-absence model.

Discussion

Here we demonstrate the significant improvements in the accuracy of SDMs that can be achieved from using abundance data to train species distribution models. By including measures of abundance we derive a more accurate assessment of the relative suitability of habitats, thereby improving predictive performance. A lack of differentiation between low and high quality habitats may lead to model bias in the presence-absence trained models. For example, occurrences in low quality, wide-ranging habitats will outweigh records from high quality, scarce habitats. Due to the large number of observations, the relative importance of these low quality habitats will be over-weighted in models trained on presence-absence data (Brotons et al. 2004).
We also show a hump-shaped relationship between species prevalence and model predictive accuracy. A variety of hypotheses on the causal factor behind this association already exist in the literature (Segurado & Araújo 2004; Allouche, Tsoar & Kadmon 2006; Santika 2011). Here, however, the interacting effects of method and prevalence on model performance are of greater interest. The marginal interaction shows that models built using abundance data generally outperform those built with presence-absence data, particularly for species with low prevalence. This contrasts with expectations based on the positive relationship between range size and local abundance (Brown 1984), which suggest that model performance would converge at low prevalence, owing to the relative lack of differentiation between presence-absence and abundance data (Brotons \textit{et al.} 2004).

Our results suggest that models trained on abundance data are better able to identify the relative suitability of habitats, than those trained on presence-absence data. The question naturally arises: what biological explanations could underlie this finding? The relationship between environmental suitability and abundance has been widely discussed (Pearce & Ferrier 2001; Nielsen \textit{et al.} 2005). Indeed, VanDerWal \textit{et al.} (2009) demonstrated that spatial patterns of abundance could be predicted using habitat suitability inferred from models based on presence-absence data alone. Using models based on abundance data (rather than presence-absence data), the relative suitability of habitats can be modelled with even greater refinement. This is because information about the suitability of habitats is lost when treating all presences as equal, regardless of the abundance of individuals that the habitat supports. By considering abundance, presences - which are uninformative in presence-absence modelling - gain structure, improving the models’ ability to
discriminate between fine scale differences in habitat quality. This could be particularly pronounced in situations in which the presence of a species is determined by habitat features that occur at a finer scale than that at which the model is fitted (Brotons et al. 2004). For instance, microclimates within a cell may render small patches of that cell suitable for low numbers of individuals, even where the mean climate of the cell is unsuitable; presence-absence data alone would suggest that the mean climate of that cell is as suitable as that of a cell with suitable climate throughout. Additionally this increased level of model refinement and ability to discriminate between finer scale differences in habitat quality may prove beneficial when using the model to project across alternative regions or time periods.

Our results suggest that even coarse scale abundance data can deliver large improvements in predicting spatial patterns of occurrence. With this in mind, why are spatial distribution modellers not driving the collection of abundance data? Gibbons et al. (2007), suggested that collecting abundance data for bird atlases is no more costly or resource demanding than collecting presence-absence data. Abundance data also provide valuable baselines against which to assess future changes (Cumming 2007). Changes in abundance will be much more rapidly apparent, and hence more rapidly detected, than changes in presence-absence patterns across ranges (which are dependent upon colonisation and extinction events) (Gregory et al. 2005). Furthermore, categorical abundance data allow for the use of new and more informative modelling techniques such as density structured models and dynamic range modelling (Keith et al. 2008; Zurell et al. 2012; Mieszkowska et al. 2013). By integrating demographic data with range dynamics, these models aim to reduce bias in future range projections (Pagel & Schurr 2012; Schurr et al. 2012).
Additionally, existing methods for modelling ordinal data, such as proportional odds models, are being improved by integration with boosting approaches. These algorithms improve prediction accuracy and avoid the overfitting problems associated with a maximum-likelihood approach (Schmid et al. 2011; Häring et al. 2013). By including population dynamics, dynamic SDMs allow for the temporal aspects of a species’ distribution to be investigated, including future abundance trends and species persistence. This in turn allows for a detailed assessment of the long term value of a site for species conservation. It is clear that not only can abundance data trained models predict the distribution of a species with a greater degree of accuracy, but that the information provided by these models is much richer than those predictions provided by distribution modelling.

Currently, many global data sets already contain measures of the local abundance of species (Robertson, Cumming & Erasmus 2010). Aside from periodic atlases, many of these provide annually repeated census data across a broad range of taxa including butterflies (Pollard & Yates 1994), birds (Sauer et al. 2012), vascular plants (Preston, Pearman & Dines 2002), and plankton (Barnard et al. 2004). Despite this array of data, species distribution modellers continue to use presence-absence data to train and test SDMs, choosing to focus on methodological development to enhance model performance (Guisan & Thuiller 2005; Araújo & Guisan 2006; Elith et al. 2006; Pearson et al. 2006; Higgins, O'Hara & Römermann 2012). To our knowledge, only two papers have attempted to use these abundance data to model species’ abundance at a large scale (Renwick et al. 2011; Johnston et al. 2013), yet here we show that relatively slight increases in the information content of a training data set (the change from binary presence-absence data to a log-scaled...
set of seven abundance categories) result in significant improvements in model performance. Given this improvement in model accuracy, combined with the creation of better baseline data sets, where existing abundance data are available, we advocate the use of abundance models as tools to predict the ecological consequences of environmental change. Where such data do not exist, we recommend that abundance data be collected alongside presence/absence data because, even if only relatively coarse numerical scales of recording are possible, the benefits are considerable.

Acknowledgments

CH is funded by a National Environment Research Council (NERC) training grant with a British Trust for Ornithology (BTO) CASE partnership. CH is supervised by SGW, PAS and JPH.

Data Accessibility

EBCC atlas data: Use of the data is administered via the EBCC Executive Committee and the data extraction and handling is currently done by staff at SOVON in the Netherlands or the BTO in the UK, according to agreed rules. Those interested in using this dataset should contact the EBCC Chair, Ruud Foppen (ruud.foppen@sovon.nl) about the conditions for obtaining the data.

R scripts: uploaded as online supporting information

References


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.

