Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014

doi: 10.1111/2041-210X.12184

Improving species distribution models: the value of data on abundance

Christine Howard[1]*, Philip A. Stephens[1], James W. Pearce-Higgins[2], Richard D. Gregory[3] and Stephen G. Willis[1]

[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; and* [3]*The Royal Society for the Protection of Birds & European Bird Census Council, The Lodge, Sandy, Bedfordshire SG19 2DL, UK*

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 Species distribution models 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 (Araujo & 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

*Correspondence author: E-mail: christine.howard@durham.ac.uk

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 whether there are limitations to any benefit associated with abundance modelling.

Materials and 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 × 50 km Universal Transverse Mercator (UTM) grid, mostly representing the period from 1985 to 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, 1000–9999, 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 evapotransipration (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) 1-km resolution data base (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 RF implementation based on a conditional inference framework (Hothorn, Hornik & Zeileis 2006a,b; Strobl, Hothorn & Zeileis 2009; R Development Core Team 2012). As with other classification methods, RFs 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 optimize the binary split into each homogenous area (Hothorn, Hornik & Zeileis 2006a,b; 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 recognized 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 nonlinear, context-dependent relationships among 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 autocovariate (SAC). This term accounts for the greater degree of similarity between more proximate samples, which 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:

$$L = \log_{10}\left[\frac{1}{n}\Sigma_i^n \frac{1}{2} 10^{A_i}\right] \qquad \text{eqn 1}$$

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 10-fold cross-validation to reduce SAC between training and test data and to minimize overfitting. 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 Fig. 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 is 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 observe 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 data sets. 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 criticized 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 & Schröder 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 >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 >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 & Schröder 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 generalised additive model (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 (Fig. 1a,b; paired t-tests, kappa $t_{344} = 13{\cdot}23$, $P < 0{\cdot}01$, AUC $t_{344} = 3{\cdot}72$, $P < 0{\cdot}01$). 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 ($t_{344} = 3{\cdot}88$, $P < 0{\cdot}01$), 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 ($t_{344} = 3{\cdot}33$, $P < 0{\cdot}01$) 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

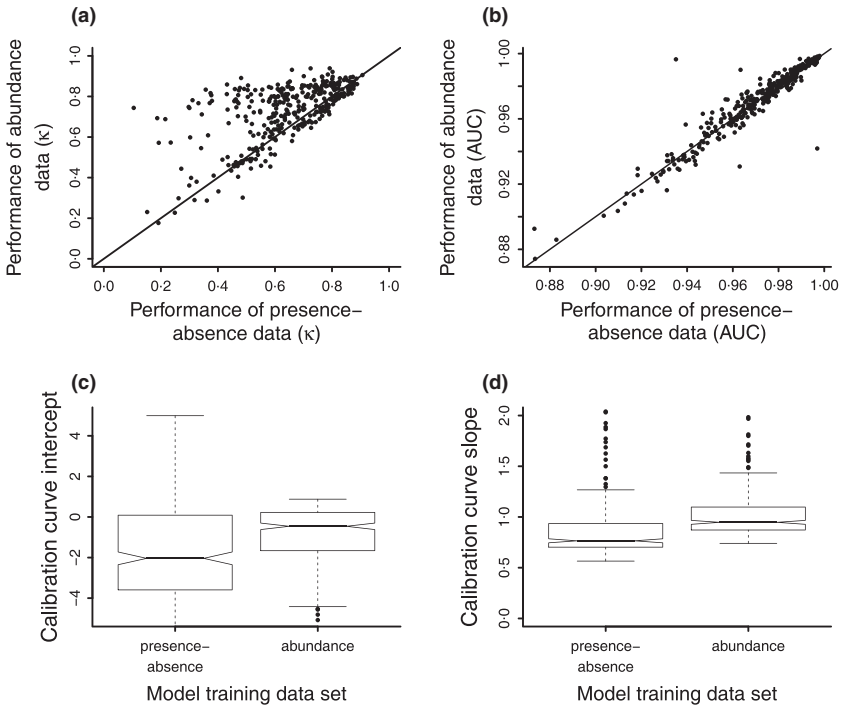

Fig. 1. Measures of model performance for each form of training data. (a) Cohen's kappa, (b) AUC, (c) Intercept of the model calibration curve and (d) slope of the model calibration curve (n = 345). Notches indicate the 95% confidence intervals of the median, with a lack of overlap indicating a significant difference at the 5% level.

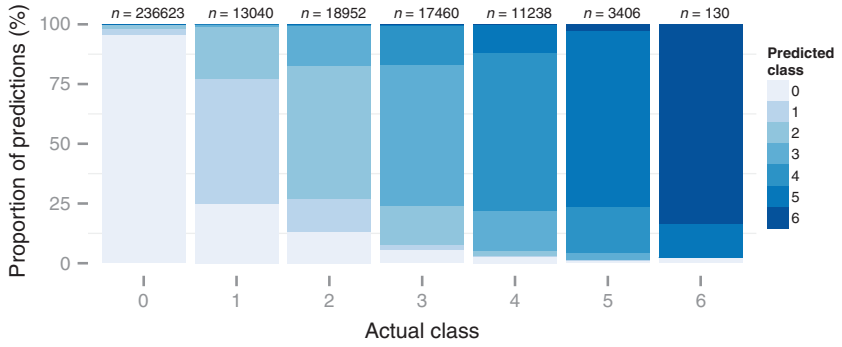

Fig. 2. Abundance predictions from abundance data trained models. Bars represent the mean proportion of predictions for each abundance class averaged across all species. *N* values indicate number of observed cells within each abundance class.

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; Fig. 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 (Fig. 3). For presence–absence data trained models with a poor to slight rating kappa score (i.e. <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 >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

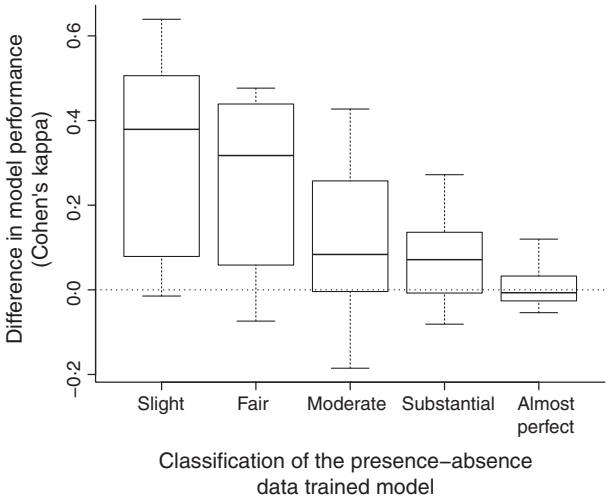

Fig. 3. Mean difference in Cohen's kappa scores between abundance data trained and presence–absence data trained models, Bins are based on the classification of the presence–absence data trained model according to Landis & Koch (1977). Positive values for differences in kappa score indicate an improvement in model fit, whilst negative values indicate a reduction in model fit.

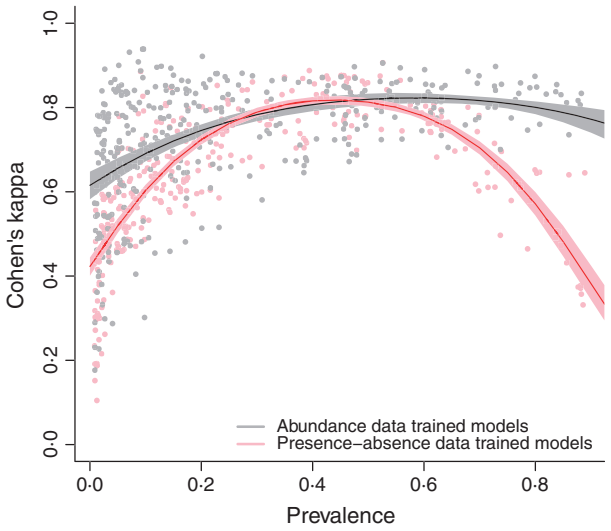

Fig. 4. Relationship between model accuracies, measured using Cohen's kappa statistic, for models trained on abundance data and those trained on presence–absence data. Shaded areas represent 95% bootstrapped confidence intervals of the mean.

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 *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 *et al.* 2005). Indeed, VanDerWal *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 colonization 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 1993), 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.

Acknowledgements

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. We would like to thank R.B. O'Hara, J. McPherson, B. Schröder, M. Spencer and 2 reviewers for their helpful insights and comments on the manuscript.

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 data set 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

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Araujo, M.B. & Peterson, A.T. (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**, 1527–1539.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Austin, M.P. & Van Niel, K.P. (2011) Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, **38**, 1–8.

Bagchi, R., Crosby, M., Huntley, B., Hole, D.G., Butchart, S.H.M., Collingham, Y. *et al*. (2013) Evaluating the effectiveness of conservation site networks under climate change: accounting for uncertainty. *Global Change Biology*, **19**, 1236–1248.

Barnard, R., Batten, S., Beaugrand, G., Buckland, C., Conway, D.V.P., Edwards, M. *et al*. (2004) Continuous plankton records: plankton atlas of the North Atlantic Ocean (1958–1999). II. Biogeographical charts. *Marine Ecology Progress Series*, **Supplement**, 11–75.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Brotons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Brown, J.H. (1984) On the relationship between abundance and distribution of species. *American Naturalist*, **124**, 255–279.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cumming, G. (2007) Global biodiversity scenarios and landscape ecology. *Landscape Ecology*, **22**, 671–685.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler, J.J. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Davey, C.M., Chamberlain, D.E., Newson, S.E., Noble, D.G. & Johnston, A. (2012) Rise of the generalists: evidence for climate driven homogenization in avian communities. *Global Ecology and Biogeography*, **21**, 568–578.

De'ath, G. & Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.

Dormann, C.F. (2007a) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Dormann, C.F. (2007b) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, **8**, 387–397.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G. *et al*. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Doswald, N., Willis, S.G., Collingham, Y.C., Pain, D.J., Green, R.E. & Huntley, B. (2009) Potential impacts of climatic change on the breeding and non-breeding ranges and migration distance of European Sylvia warblers. *Journal of Biogeography*, **36**, 1194–1208.

Elith, J., Graham, C.H, Anderson, R.P, Dudík, M., Ferrier, S., Guisan, A. *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Fortin, M.-J. & Dale., M.R.T. (Eds) (2005) *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, Cambridge, UK.

Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

Gibbons, D.W., Donald, P.F., Bauer, H.-G., Fornasari, L. & Dawson, I.K. (2007) Mapping avian distributions: the evolution of bird atlases: capsule an increasing proportion of atlases now map patterns of abundance but they are still a minority even though they require no more input of time or fieldworkers. *Bird Study*, **54**, 324–334.

Green, R.E., Collingham, Y.C., Willis, S.G., Gregory, R.D., Smith, K.W. & Huntley, B. (2008) Performance of climate envelope models in retrodicting

recent changes in bird population size from observed climatic change. *Biology Letters*, **4**, 599–602.
Gregory, R.D., Noble, D.G. & Custance, J. (2004) The state of play of farmland birds: population trends and conservation status of lowland farmland birds in the United Kingdom. *Ibis*, **146**, 1–13.
Gregory, R.D., van Strien, A., Vorisek, P., Meyling, A.W.G., Noble, D.G., Foppen, R.P.B. & Gibbons, D.W. (2005) Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 269–288.
Gregory, R.D., Willis, S.G., Jiguet, F., Vorisek, P., Klvanova, A., van Strien, A. *et al.* (2009) An indicator of the impact of climatic change on European bird populations. *PLoS ONE*, **4**, e4678.
Guisan, A. & Harrell, F.E. (2000) Ordinal response regression models in ecology. *Journal of Vegetation Science*, **11**, 617–626.
Guisan, A. & Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, **38**, 1433–1444.
Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.
Hagemeijer, E.J.M. & Blair, M.J. (1997) *The EBCC Atlas of European Breeding Birds: Their Distribution and Abundance*. T. & A.D Poyser, London.
Häring, T., Reger, B., Ewald, J., Hothorn, T. & Schröder, B. (2013) Regionalizing indicator values for soil reaction in the Bavarian Alps–from averages to multivariate spectra. *Folia Geobotanica*, 1–21.
Harrison, J.A. & Cherry, M. (1997) *The Atlas of southern African Birds*. BirdLife South Africa, Johannesburg.
Higgins, S.I., O'Hara, R.B. & Römermann, C. (2012) A niche for biology in species distribution models. *Journal of Biogeography*, **39**, 2091–2095.
Hothorn, T., Hornik, K. & Zeileis, A. (2006a) party: A Laboratory for Recursive Part(y)tioning.
Hothorn, T., Hornik, K. & Zeileis, A. (2006b) Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**, 651–674.
Huntley, B., Green, R.E., Collingham, Y. & Willis, S.G. (2007) *A Climatic Atlas of European Breeding Birds*. Durham University, The RSPB and Lynx Edicions, Barcelona.
Johnston, A., Ausden, M., Dodd, A.M., Bradbury, R.B., Chamberlain, D.E., Jiguet, F. *et al.* (2013) Observed and predicted effects of climate change on species abundance in protected areas. *Nature Climate Change*, **3**, 1055–1061.
Keith, D.A., Akcakaya, H.R., Thuiller, W., Midgley, G.F., Pearson, R.G., Phillips, S.J., Regan, H.M., Araujo, M.B. & Rebelo, T.G. (2008) Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models. *Biology Letters*, **4**, 560–563.
Kissling, W.D. & Carl, G. (2008) Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, **17**, 59–71.
Landis, J.R. & Koch, G.G. (1977) Measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.
Lawler, J.J., White, D., Neilson, R.P. & Blaustein, A.R. (2006) Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology*, **12**, 1568–1584.
Liaw, A. & Wiener, M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
Magness, D.R., Huettmann, F. & Morton, J.M. (2008) Using Random Forests to provide predicted species distribution maps as a metric for ecological inventory & monitoring programs. *Applications of Computational Intelligence in Biology: Current Trends and Open Problems*, **122**, 209–229.
Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.
Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, **15**, 59–69.
McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.
Mieszkowska, N., Milligan, G., Burrows, M.T., Freckleton, R. & Spencer, M. (2013) Dynamic species distribution models from categorical survey data. *Journal of Animal Ecology*, **82**, 1215–1266.
Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.
Mucher, C.A., Steinnocher, K.T., Kressler, F.P. & Heunks, C. (2000) Land cover characterization and change detection for environmental monitoring of pan-Europe. *International Journal of Remote Sensing*, **21**, 1159–1181.
New, M., Hulme, M. & Jones, P. (1999) Representing twentieth-century space-time climate variability. Part I: development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate*, **12**, 829–856.
Nielsen, S.E., Johnson, C.J., Heard, D.C. & Boyce, M.S. (2005) Can Models of presence-absence be used to scale abundance? Two case studies considering extremes in life history. *Ecography*, **28**, 197–208.
O'Grady, J.J., Reed, D.H., Brook, B.W. & Frankham, R. (2004) What are the best correlates of predicted extinction risk? *Biological Conservation*, **118**, 513–520.
Olden, J.D., Lawler, J.J. & Poff, N.L. (2008) Machine learning methods without tears: a primer for ecologists. *Quarterly Review of Biology*, **83**, 171–193.
Pagel, J. & Schurr, F.M. (2012) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, **21**, 293–304.
Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.
Pearce, J. & Ferrier, S. (2001) The practical value of modelling relative abundance of species for regional conservation planning: a case study. *Biological Conservation*, **98**, 33–43.
Pearson, R.G., Dawson, T.P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.
Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C. *et al.* (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704–1711.
Pollard, E. & Yates, T.J. (1993) *Monitoring Butterflies for Ecology and Conservation: The British Butterfly Monitoring Scheme*. Chapman & Hall, London, UK.
Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.
Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A. & Solomon, A.M. (1992) A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*, **19**, 117–134.
Preston, C.D., Pearman, D.A. & Dines, T.D. (2002) *New Atlas of the British and Irish Flora. An Atlas of the Vascular Plants of Britain, Ireland, the Isle of Man and the Channel Islands*. Oxford University Press, Oxford, UK.
R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Reineking, B. & Schröder B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.
Renwick, A.R., Massimino, D., Newson, S.E., Chamberlain, D.E., Pearce-Higgins, J.W. & Johnston, A. (2011) Modelling changes in species' abundance in response to projected climate change. *Diversity and Distributions*, **18**, 121–132.
Robertson, M.P., Cumming, G.S. & Erasmus, B.F.N. (2010) Getting the most out of atlas data. *Diversity and Distributions*, **16**, 363–375.
Robnik-Sikonja, M. (2004) Improving random forests. *Machine Learning: Ecml 2004, Proceedings*, **3201**, 359–370.
Santika, T. (2011) Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, **20**, 181–192.
Sauer, J.R., Hines, J.E., Fallon, J.E., Pardieck, K.L., Ziolkowski, D.J. Jr & Link, W.A. (2012) The North American Breeding Bird Survey, Results and Analysis 1966–2011. Version 07.03.2013 USGS Patuxent Wildlife Research Center, Laurel, MD [Online]. Available: http://www.mbr-pwrc.usgs.gov/bbs/.
Schmid, M., Hothorn, T., Maloney, K.O., Weller, D.E. & Potapov, S. (2011) Geoadditive regression modeling of stream biological condition. *Environmental and Ecological Statistics*, **18**, 709–733.
Schroder, B. & Seppelt, R. (2006) Analysis of pattern-process interactions based on landscape models – Overview, general concepts, and methodological issues. *Ecological Modelling*, **199**, 505–516.
Schurr, F.M., Pagel, J., Cabral, J.S., Groeneveld, J., Bykova, O., O'Hara, R.B. *et al.* (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography*, **39**, 2146–2162.
Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.
Segurado, P., Araujo, M.B. & Kunin, W.E. (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43**, 433–444.
Strobl, C., Hothorn, T. & Zeileis, A. (2009) Party on!. *R Journal*, **1**, 14–17.
Strobl, C., Malley, J. & Tutz, G. (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, **14**, 323–348.

Thuiller, W. (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9**, 1353–1362.
Thuiller, W., Araujo, M.B. & Lavorel, S. (2004) Do we need land-cover data to model species distributions in Europe? *Journal of Biogeography*, **31**, 353–361.
Van Horne, B. (1983) Density as a misleading indicator of habitat quality. *The Journal of Wildlife Management*, **47**, 893–901.
VanDerWal, J., Shoo, L.P., Johnson, C.N. & Williams, S.E. (2009) Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. *American Naturalist*, **174**, 282–291.
Vorpahl, P., Elsenbeer, H., Märker, M. & Schröder, B. (2012) How can statistical models help to determine driving factors of landslides? *Ecological Modelling*, **239**, 27–39.
Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 3–36.
Zurell, D., Jeltsch, F., Dormann, C.F. & Schröder, B. (2009) Static species distribution models in dynamically changing systems: how good can predictions really be? *Ecography*, **32**, 733–744.
Zurell, D., Grimm, V., Rossmanith, E., Zbinden, N., Zimmermann, N.E. & Schröder, B. (2012) Uncertainty in predictions of range dynamics: black grouse climbing the Swiss Alps. *Ecography*, **35**, 590–603.

Received 6 September 2013; accepted 4 March 2014
Handling Editor: Jana McPherson

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Figure S1. Correlogram indicating the mean (shaded areas show standard deviation) of the correlograms across all 345 species for raw data (black line) and for the residuals after model fitting (red line).

Data S1. Example R code to implement species abundance and distribution modelling using the party package, along with example code to calculate the spatial autocovariate term.