

# Updating bird species distribution at large spatial scales: applications of habitat modelling to data from long-term monitoring programs

Lluís Brotons<sup>1,2\*</sup>, Sergi Herrando<sup>2</sup> and Magda Pla<sup>1</sup>

<sup>1</sup>Centre Tecnològic Forestal de Catalunya, Àrea de Biodiversitat, Pujada del Seminari s/n, 25280 Solsona, Spain, <sup>2</sup>Institut Català d'Ornitologia, Museu de Zoologia, Passeig Picasso s/n, 08003 Barcelona, Catalonia, Spain

## ABSTRACT

Mapping of species distributions at large spatial scales has been often based on the representation of gathered observations in a general grid atlas framework. More recently, subsampling and subsequent interpolation or habitat spatial modelling techniques have been incorporated in these projects to allow more detailed species mapping. Here, we explore the usefulness of data from long-term monitoring (LTM) projects, primarily aimed at estimating trends in species abundance and collected at shorter time intervals (usually yearly) than atlas data, to develop predictive habitat models. We modelled habitat occupancy for 99 species using a bird LTM program and evaluated the predictive accuracy of these models using independent data from a contemporary and comprehensive breeding bird atlas project from the same region. Habitat models from LTM data using generalized linear modelling were significant for all the species and generally showed a high predictive power, albeit lower than that from atlas models. Sample size and species range size and niche breadth were the most important factors behind variability in model predictive accuracy, whereas the spatial distribution of sampling units at a given sample size had minor effects. Although predictive accuracy of habitat modelling was strongly species dependent, increases in sample size and, secondarily, a better spatial distribution of sampling units should lead to more powerful predictive distribution models. We suggest that data from LTM programs, now established in a large number of countries, has the potential for being a major source of good quality data suitable for the estimation and regularly update of distributions at large spatial scales for a number of species.

## Keywords

Atlas work, bird conservation, breeding bird survey, generalised linear models, sampling design.

\*Correspondence: Lluís Brotons, Centre Tecnològic Forestal de Catalunya, Àrea de Biodiversitat, Pujada del Seminari s/n, 25280 Solsona, Spain. E-mail: lluis.brotons@ctfc.es

## INTRODUCTION

Knowing where a species occurs and recording changes in its distribution has major implications in ecology, species management, and conservation planning (Araújo & Williams, 2000; Bani *et al.*, 2002). This has led to many large-scale mapping or atlas projects where the aim has been to document distributions in a systematic manner, usually on a grid system. The greatest development and application of atlas methodology has been evident in the assessment of bird distribution (Donald & Fuller, 1998; Underhill & Gibbons, 2002). Historically, mapping of species at large spatial scales has been based on the collection of observations and their comprehensive representation on a continuous spatial grid. However, detecting species reliably requires considerable effort (Kery & Schmid, 2005; McPherson *et al.*, 2006), and atlas works commonly aim at covering large areas. A

trade-off is likely to appear between atlas spatial coverage and the final effort to accurately cover it. In general, this compromise has led to bird atlases being completed at considerable time intervals (i.e. during 3–6 years) and repeated at long time intervals (i.e. 20-year periods). Furthermore, the sampling procedure conducted in atlas works usually gives complete coverage of the study region but only describing presence–absence at relatively coarse resolutions (Donald & Fuller, 1998). Traditional atlas approaches often provide poor spatial information at small scales (McPherson *et al.*, 2006). However, second generation atlases have recently complemented the wide-covering, coarse-resolution surveys with a more detailed subsampling within the main units in order to gather information on the species distribution at more local spatial scales (1 × 1 km or 2 × 2 km) to produce distribution/abundance maps at finer-scale resolutions (Gibbons *et al.*, 1993; Schmid *et al.*, 1998; Johnson & Sargeant, 2002).

In parallel to the constant improvement of atlas works, long-term monitoring (LTM) projects have been developed with their main objective being the estimation of temporal trends in abundance, mostly for common and widespread species (Link & Sauer, 1998). LTM programs are based on a network of sampling locations where the relative abundance of a number of species is recorded at given time steps (typically each year) and have been implemented for birds across a number of countries, usually using representative and well-designed sampling strategies (Gregory *et al.*, 2005; Kery & Schmid, 2005).

Bird LTM programs provide us with a great deal of spatial data that have the potential to be used to create maps showing changes in species distribution and abundance (Jiguet *et al.*, 2005), which may lead to a progressive convergence of objectives between LTM and atlas projects [i.e. Breeding Bird Survey data in North America (Price *et al.*, 1995)]. This convergence may provide substantial advantages in terms of the periodicity of produced maps because LTM is a data source that may be used to update species distributions at shorter time intervals.

Two main different alternatives are available, permitting the use of information gathered in a subsample of small spatial units to map the species distribution in a continuous manner: interpolation techniques and niche or habitat-based models (Guisan & Zimmermann, 2000). Atlas works have commonly used interpolation techniques which aim at estimating unknown data from neighbour values (Gibbons *et al.*, 1993). This approach appears as reasonable if data are spatially well distributed and the density of locations high. However, this is often not the case in LTM programs, for which coverage is frequently less spatially exhaustive than that used to obtain interpolated distributions in second-generation atlas works. Therefore, in these cases surrogate environmental data may probably prove useful to estimate species distribution beyond sampled locations. Recent developments in numerical methods and increasing availability of digital environmental data sources have boosted application of habitat-based models in ecology (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005). This approach is based on the hypothesis that if species environmental associations can be robustly established, one may use them to estimate species distributions through the identification of suitable habitat in areas from which faunal data have not been recorded but environmental information is available.

With the exception of few pilot methodological studies (Pearce & Ferrier, 2001; Thogmartin *et al.*, 2004, 2006), habitat modelling has not been widely applied to LTM data. In this study, we use data from a bird LTM program, the Catalan common bird survey, to produce habitat-based maps and assess its predictive accuracy using independent atlas data. It is important to stress that predictive habitat modelling was used in our study with a predictive rather than inductive goal. In such circumstances, accuracy of model predictions is more important than significance of particular ecological terms (Legendre & Legendre, 1998). We also investigated different factors that are known to be related to the predictive accuracy of habitat models: the role of sample size and the distribution of sample locations (Hirzel & Guisan, 2002), and ecological factors (Brotons *et al.*, 2004; Seoane *et al.*, 2005). We then derive recommendations on how to

use LTM data to map species distributions. To our knowledge this is the first time that mapping from LTM data is extensively used to generate habitat models from a large number of species and evaluated using independent atlas data from the same area (but see Carrascal *et al.*, 2006 for a similar approach at a regional scale).

## METHODS

### Study area

Catalonia is a region situated in north-east Spain, comprising an area of about c. 32,000 km<sup>2</sup>, roughly the size of Belgium (Fig. 1). The region is located in the Mediterranean Basin, and despite its area, it is remarkably heterogeneous including a range of landscapes from alpine habitats to coastal marshes and from evergreen forests to steppes.

### Catalan common bird survey (SOCC)

We used LTM data from the Catalan common bird survey (SOCC, from the Catalan name ‘*Seguiment d’Ocells Comuns a Catalunya*’), a project started in 2002. The SOCC is based on the line transect approach (Bibby *et al.*, 2000) in which observers record all individuals of all bird species sighted or heard on a 3-km walked itinerary divided into three 1-km sections. In this study we used 1-km sections as sampling units. SOCC transects are visited twice during the breeding season, the first between 15 April and 15 May, and the second between 15 May and 15 June, thus covering the periods of maximum activity of early and late breeders and minimizing the presence of migrants (see [www.ornitologia.org](http://www.ornitologia.org)).

There are two different types of SOCC transects according to the criteria used in their selection: (1) priority transects (located in a randomly chosen 10-km square within each of five biogeographically determined strata) or (2) non-priority transects (located in a 10-km square freely chosen by the observer). These two possibilities aim at attracting volunteers to the main areas of interest without losing any possible contributor because of distant or undesired census locations. At present 226 SOCC transects (112 priority and 114 non-priority transects) have been conducted at least in 1 year during the period 2002–2005 (Fig. 1). The mean number of available years (maximum of 4 years) per transect during this period was 3.01. We included in our study 99 species appearing in at least 10 different transects during the 2002–2005 period.

Since the SOCC program is essentially based on volunteer observers, the survey is constrained by the number of available sampling transects and is prone to poor spatial cover of remote areas, resulting in, regionally, a biased sampling distribution (Fig. 1). To investigate the potential effect of sample size and sampling distribution on the performance of habitat modelling using LTM data, we assessed the impact of varying sample set size on the predictive power of habitat models on three different scenarios of data partitioning. In the first scenario, we assessed changes in predictive power by randomly selecting subsets (25%,



**Figure 1** Maps showing the location of (a) Catalonia, (b) the location of the 226 transects of the Catalan Breeding Bird Survey (SOCC) conducted during the period 2002–05, and (c) the location of the 3077  $1 \times 1$  km UTM squares surveyed during the Catalan Breeding Bird Atlas 1999–2002. Map (b) also shows the different counties in which Catalonia is subdivided.

50%, 75%, or 100%) of the complete data set of available 1-km transect sections (RANDOM scenarios). In the second scenario, we reduced sample sizes by reducing the total number of sections from the most intensively sampled areas (SELECTIVE scenarios). This was done by randomly selecting subsets of the data using a probability function based on the distribution of priority transects and the density of transect locations per county (Fig. 1). Finally, we assessed the impact of varying field effort for the same sample size on the predictive power using the entire transects and not their sections as main sampling units for modelling (EFFORT scenario). The EFFORT scenario results in more intensive sampling effort per unit (mean values from a 3-km transect vs. a 1-km section), but increases their spatial independence (larger mean distance between units). In this scenario, the

entire data set was reduced to separate 226 units (the number of SOCC transects), and the overall available sample size is more restricted than in the other two scenarios.

### Catalan Breeding Bird Atlas (CBBA)

We used species occurrence from the Catalan Breeding Bird Atlas (CBBA, Estrada *et al.*, 2004) as independent data to evaluate predictive accuracy of habitat models based on SOCC data. The CBBA is a large-scale survey that covered between 1999 and 2002 the whole of the Catalonia using a grid-based 10-km Universal Transverse Mercator (UTM) squares. A subsampling of a total of 3077 1-km squares (approximately 9% of the total area) was selected to conduct standardized intensive surveys of species

presence in a stratified fashion to cover the main habitat types present within each of the 10-km squares (Hirzel & Guisan, 2002).

We also used the CBBA data to investigate the role of species range size and niche breadth in determining variability in model predictive accuracy based on LTM data (Hepinstall *et al.*, 2002). Species range size was estimated from the total number of UTM 10-km atlas squares in which the species was recorded (Estrada *et al.*, 2004). We obtained a measure of niche breadth (habitat amplitude) using the following procedure. First, we performed a matrix of Pearson distances from species habitat selection patterns presented in the CBBA (Estrada *et al.*, 2004). This was done in order to estimate the ecological distance ( $dist(hab_x, hab_y)$ ) between the habitats used in their study. Distances between habitats were then weighted by the habitat preferences and finally, niche breadth was calculated using all habitats in which each species was present as:

$$A = Var \left[ \frac{[weight_{hab_x} * weight_{hab_y} * dist(hab_x, hab_y)]}{weight_{hab_x} * weight_{hab_y}} \right]$$

Since species prevalence (proportion of occurrences in a sample) has been shown to affect model predictive accuracy, some authors have warned about the need to control for prevalence when assessing the role of variables such as range size and niche breadth on such measure (McPherson *et al.*, 2004). To allow the independent assessment of prevalence from range size and niche breadth, we conducted a Principal Component Analysis using as original variables species niche breadth, range size (CBBA 10-km squares), and the two highly correlated measures ( $r = 0.96$ ) of prevalence in our data sets (species prevalence in CBBA 1-km squares, and in 1-km SOCC sections). After a varimax transformation of the principal components maximizing their correspondence to the original variables, we succeed to obtain three independent components that were then used in further analyses: (1) a niche breadth component positively associated to the original niche breadth variable ( $r = 0.99$ ), (2) a range size component positively related to the number of 10-km squares occupied by the species in Catalonia ( $r = 0.69$ ), and (3) an independent prevalence component separating species with low and high prevalence values in the CBBA data ( $r = 0.95$ ) and SOCC data ( $r = 0.90$ ).

### Habitat-based models

We used 39 environmental variables for model building (Appendix 1) which were generated from available digital layers (see Brotons *et al.*, 2004; Estrada *et al.*, 2004). All environmental variables were generated for each 1-km UTM square in Catalonia and for 1 km<sup>2</sup> buffers around the central point of each SOCC section and if possible (i.e. land use maps), estimated from different data sources so that they better matched the sampling periods of each of the surveys.

We conducted occupancy models using presence/absence data over the 2002–2005 period from SOCC transect sections (SOCC models) using generalized linear modelling with binomial error distribution (GLM). GLM have been extensively tested elsewhere

and have proved robust in a number of independent situations (Manel *et al.*, 1999; Pearce & Ferrier, 2000b; Osborne *et al.*, 2001). We included as potential predictors in model building all linear and quadratic terms. To select the most parsimonious model using the Akaike Information Criterion (AIC), we applied an automatic stepwise model-selection procedure starting from a null model containing the intercept only (Chambers & Hastie, 1997). Quadratic terms were included only if they improved their linear counterpart. We also included interactions between environmental variables when their ecological interpretation was highly justified (i.e. interaction between shrubs and precipitation allowing to account for hypothetical wetness gradients in shrub vegetation). The number of years that a SOCC transect was surveyed was included as a covariable in habitat models to correct for differences in sampling effort among transects.

### Evaluation of habitat models

Results from probabilistic models may be judged as successful if the predicted probabilities above a certain cut-off correspond with observed occurrences, the values below the cut-off with absences and the prediction errors (false positives and false negatives) are low (Fielding & Bell, 1997; Pearce & Ferrier, 2000a). Since the choice of an adequate cut-off may be problematic, we use a more powerful approach based on the receiver operating characteristic (ROC) plot which is based on a series of misclassification matrices computed for a range of cut-offs from 0 to 1. The ROC curve then plots on the *y*-axis the true-positive fraction, against the false-positive fraction from the same misclassification matrix (Fielding & Bell, 1997; Pearce & Ferrier, 2000a). The area under the ROC curve (AUC) is a convenient measure of overall fit and commonly varies between 0.5 (for chance performance) and 1 (perfect fit).

For a given species and sampling scenario (RANDOM, SELECTIVE, EFFORT), we calibrated 10 models by randomly selecting an 80% of the available sections and then evaluated the predictive accuracy of the models on independent field data. Each of these independent tests measured predictive accuracy by means of AUC on a different random selection of 30% of CBBA data (1-km squares). Finally, the mean AUC of these 10 cases was used as estimator of predictive accuracy per species and scenario.

We considered that CBBA occurrence habitat models (CBBA models) can be currently used as the best models available for each species in the region (Estrada *et al.*, 2004). Hence, for each species, we also measured the disagreement between the predictive accuracy of SOCC habitat models and the models obtained in the context of the CBBA project. CBBA models used exactly the same environmental variables used in this study to calibrate SOCC models. However, the two kinds of models differed in that CBBA models included a second modelling step that involved adding spatial autocovariables to the best environmental model previously selected (Augustin *et al.*, 1996). This allowed identifying the best environmental model available given our predictors and complementing those with information about the spatial structure of the species data (i.e. spatial autocorrelation). CBBA habitat model predictions were evaluated using a cross-validation



procedure by randomly assigning 70% of occurrence values for each species to a calibration data set, and the 30% of remaining occurrences to an evaluation data set [the same number of cases used in the SOCC model evaluation (Guisan & Zimmermann, 2000)].

### Comparison of habitat models

We assessed how predictive accuracy of habitat models varied between methods by means of repeated measures ANCOVA designs using model type as within-subject factors in the design according to species, and the three components obtained from the principal component analyses as continuous predictors. When comparing models estimated from different sample sizes, sample size category was considered as an additional within-subject factor.

We also investigated discrepancies between SOCC and CBBA model accuracy by means of GLM designs using as predictors the ecological characteristics of the species and the amount of variability explained by the CBBA models at each of the two steps conducted (environmental and spatial, Estrada *et al.*, 2004). In this way, we could assess whether differences in predictive ability between the models were associated to species ecology per se or to the potential of the models to adequately describe species habitat with the variables employed.

## RESULTS

### Overall accuracy of SOCC models and comparison with reference CBBA models

Overall model accuracy estimated with the SOCC data performed better than random in all the species analysed (Appendix 2). Furthermore, the evaluation of their predictive accuracy on independent atlas (CBBA) field data provides acceptable to excellent results (AUC values significantly different from 0.5 and ranging between 0.63 and 0.97, Figs 2 & 3). However, AUC values from SOCC models were consistently lower than AUC values calculated from reference CBBA models ( $F_{1,88} = 376.57$ ,  $P < 0.0001$ , Fig. 3). We found that while consistently lower, AUC values for SOCC models could be easily predicted by AUC values from CBBA models due to the strong and positive relationship between the two data sets ( $r^2 = 0.81$ ,  $F_{1,97} = 417.00$ ,  $P < 0.0001$ ). Species distribution in ecological space had a major role in determining model accuracies. Once corrected for prevalence, both SOCC and CBBA occupancy models showed higher predictive performance when applied to specialist species ( $F_{1,88} = 33.42$ ,  $P < 0.0001$ , Fig. 3) and restricted range species ( $F_{1,88} = 45.29$ ,  $P < 0.0001$ , Fig. 3).

We were able to predict up to 46% of the variability in the differences between SOCC and reference CBBA models with differences being lower for wide-ranging ( $F_{1,85} = 7.99$ ,  $P < 0.01$ ) and generalist species ( $F_{1,85} = 7.65$ ,  $P < 0.01$ ). However, the most important factors behind differences between the two models were related to the type and amount of variability explained by reference CBBA models. In particular, species with a higher amount of variability explained by spatial variables in CBBA models showed larger discrepancies between the two occupancy

models ( $F_{1,85} = 11.21$ ,  $P < 0.01$ ), whereas these differences decreased with the amount of variability explained by environmental variables ( $F_{1,85} = 38.10$ ,  $P < 0.0001$ ) (Fig. 4). That is, compared to reference CBBA models, SOCC models performed relatively better for generalist, wide-ranging species and for species showing loose spatial structures and being well modelled with environmental data.

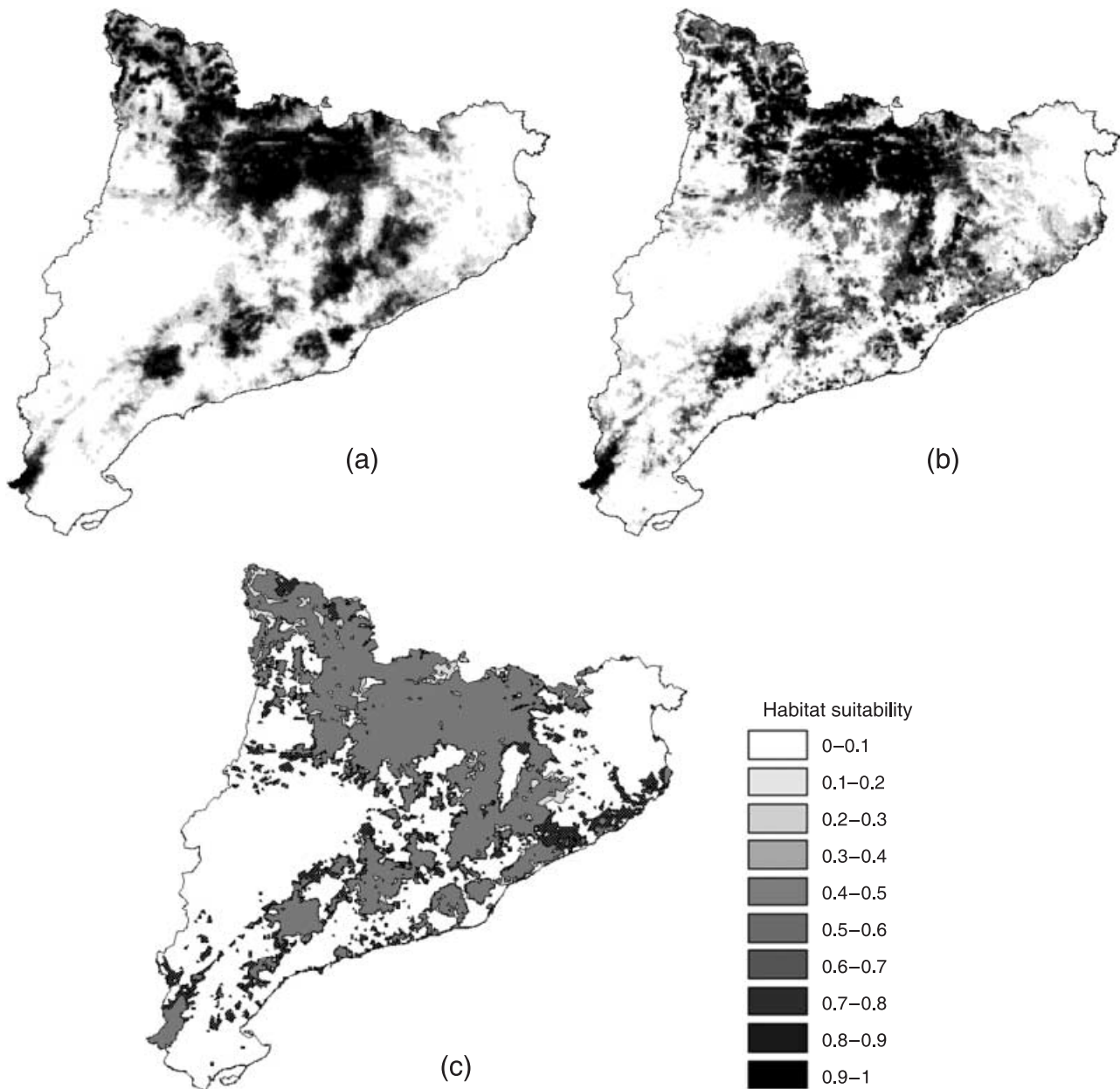
### Effects of data availability on SOCC model predictive performance

Once the significant positive effect of prevalence was taken into account ( $F_{1,86} = 5.95$ ,  $P < 0.01$ ), sample size had a strong impact on the predictive performance of SOCC models resulting in a sharp decrease in AUC values at smaller sample sizes (RANDOM scenario,  $F_{3,258} = 74.206$ ,  $P < 0.0001$ , Fig. 5). Changes in predictive performance of SOCC models in relation to sample size were similar for species with different niche breadth ( $F_{3,258} = 0.36$ , NS). However, these changes differed according to species range, since a relative poorer increase in the performance of models at larger sample sizes was found for wider ranging species ( $F_{3,258} = 4.30$ ,  $P < 0.01$ , Fig. 5).

The sampling scenario applied to data reduction did not have any significant global effect on the predictive performance of SOCC models (scenario RANDOM vs. SELECTIVE,  $F_{1,87} = 3.08$ , NS). The increase in predictive performance with sample size was apparent for the two subsampling procedures ( $F_{2,174} = 222.09$ ,  $P < 0.0001$ ), and we only detected a minor but significant lower predictive accuracy of RANDOM scenario models compared to SELECTIVE scenario models at the subsample of 25% of the data ( $F_{2,174} = 6.08$ ,  $P < 0.01$ ). Therefore, a random subsampling of the original available data was generally as efficient as a subsampling procedure that aimed at a more spatially homogeneous distribution of the monitored sections. We finally compared the predictive performance of models conducted on 25% of the data (scenarios RANDOM and SELECTIVE) with a model that used the 3-km transects and not 1-km sections as analytical units (scenario EFFORT). Such comparison resulted in a small, but significant increase in predictive performance of models conducted on complete 3-km transect data than in disaggregated 1-km section data ( $F_{2,172} = 3.79$ ,  $P < 0.05$ , Fig. 6).

## DISCUSSION

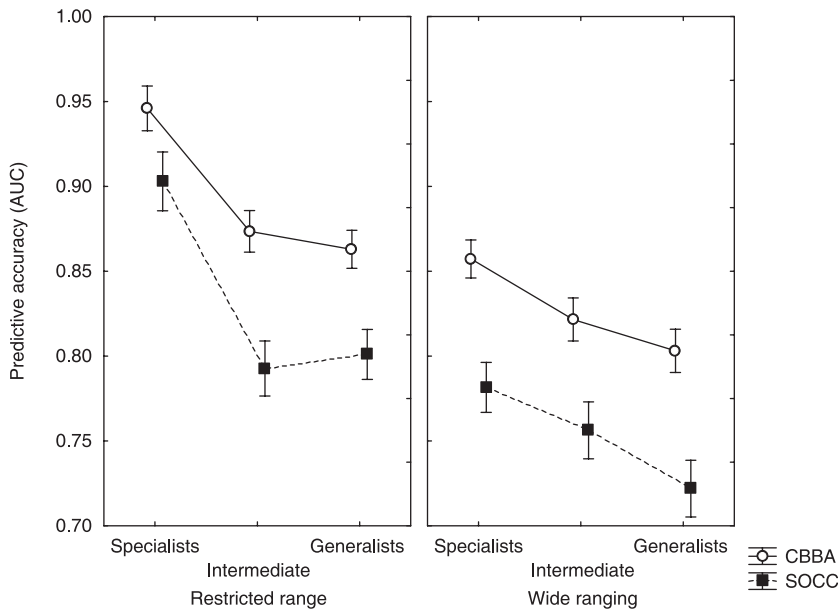
Spatial data collected in the framework of LTM programs have a substantial potential for producing species distribution maps based on habitat suitability modelling. Predictive accuracy of the models was closely linked to species range size and niche breadth but also to the available sample size to the modelling exercise (McPherson *et al.*, 2004). That is, modelling accuracy of species distribution was determined by how difficult it was to capture species responses to environment and by how much data were available to capture those responses. Hirzel & Guisan (2002), using a similar approach on a single species, concluded that sample size was the major single factor behind predictive accuracy of predictive habitat models and that sampling strategy was in



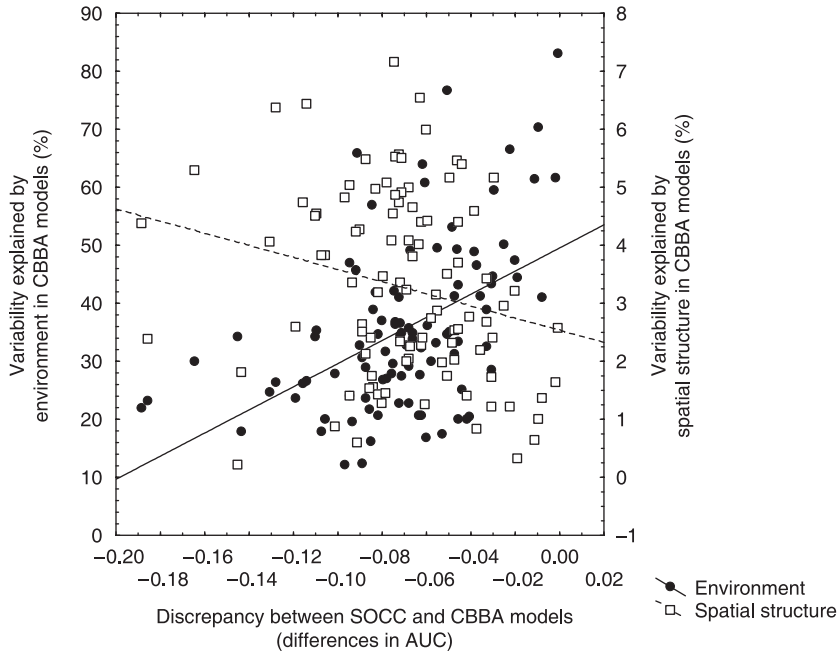
**Figure 2** Maps showing the predicted distributions of (a) the atlas GLM occupancy model [Catalan Breeding Bird Atlas (CBBA), AUC = 0.91], (b) the long-term monitoring data GLM occupancy model [Catalan Breeding Bird Survey (SOCC), AUC = 0.85], and (c) the discrepancies between the two methods for one of the species analysed, the coal tit *Parus ater*. The discrepancy map was created by crossing predicted binary presence/absence maps after choosing for each modelling method a threshold maximizing specificity and sensitivity. In (c) black, areas where the SOCC model predicted species presence and CBBA model absence. Light grey, areas where CBBA model predicted presence and SOCC model absence. Dark grey indicates coincidence in model predictions.

fact of secondary importance in their case. This is supported by our results since at a given sample size, predictive accuracy of random and spatially selective subsets was similar, indicating that enhancing the spatial representation of the data did not significantly contribute to a better model except for smaller sample sizes. Our results also suggest that disaggregation of field data sampling units may have a significant positive effect on model predictive accuracy due, predominantly, to an increase in sample size.

Predictive accuracy of models conducted on SOCC data ranked always lower than those conducted in the atlas framework. In the first place, CBBA models were conducted using a larger sample size than SOCC models (derived from about 6200 h of sampling effort for atlas data vs. about 1580 h for SOCC data). This more extensive sampling effort was probably important in allowing CBBA models to better capture niche structure for the different species, especially so for less abundant ones (which is in line with our finding that predictive accuracy differences in models



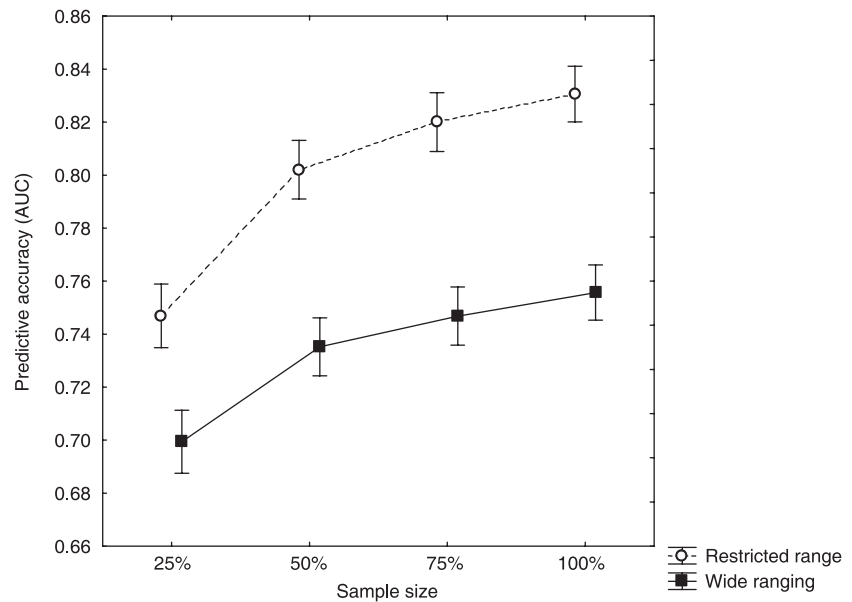
**Figure 3** Mean model accuracy for Catalan Breeding Bird Survey (SOCC) models and Catalan Breeding Bird Atlas (CBBA) reference models in relation to species range size and niche breadth as estimated from the two components of the principal component analyses. The two variables are categorized to facilitate the interpretation of their relationship with predictive model accuracy. Whiskers represent the standard error of the estimates.



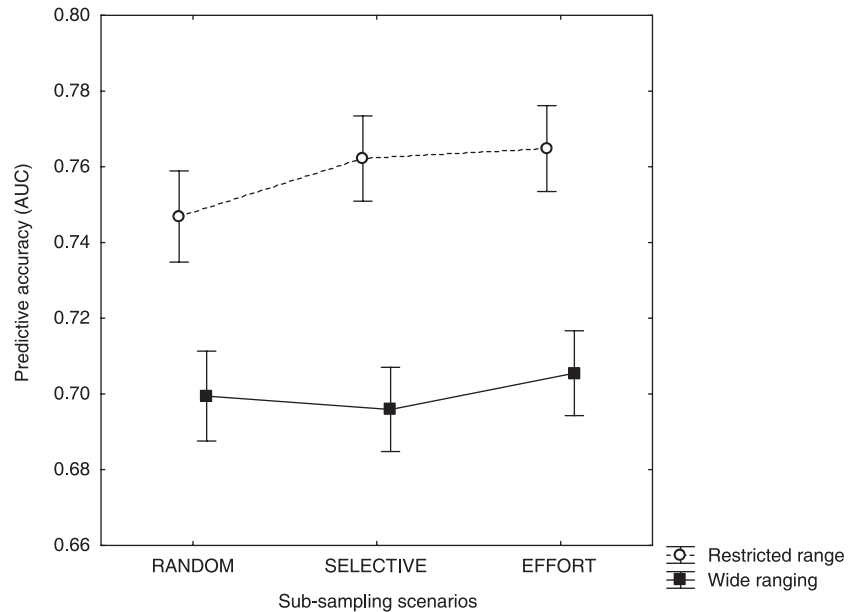
**Figure 4** Representation of the main factors behind the differences in predictive accuracy of Catalan Breeding Bird Survey (SOCC) models and reference Catalan Breeding Bird Atlas (CBBA) models. The two factors represented are the two variables describing the proportion of the total variability accounted by either environmental or spatial variables (see Methods).

conducted on scarce and wide-ranging species increase with sample size, Fig. 5). Second, it should be also stressed that, in contrast with CBBA models, SOCC models did not include information about the occurrence of the species in the surroundings (i.e. spatial structure). Since the amount of variability explained in CBBA models accounted for some of the discrepancy between these and SOCC models, it is probable that the incorporation of reliable spatial information enhances model predictive accuracy to some degree. This may be done by modelling spatial autocovariables including species spatial information (Estrada *et al.*, 2004), by spatial kriging of residuals from environmental models (Pebesma *et al.*, 2005) or by explicitly incorporating spatial information in the model formulation (Thogmartin *et al.*, 2004).

We found that ecology plays a critical role in determining predictive accuracy in models developed with both SOCC and CBBA models. In particular, wide-ranging species and generalists were modelled less accurately than more specialist and selective species. This result agrees with those of Stockwell & Peterson (2002), Brotons *et al.* (2004), and Seoane *et al.* (2005), who also reported poorer habitat models for either more generalist or abundant species. Stockwell & Peterson (2002) offered as a biological explanation for this observation that widespread species often show local or regional subpopulations that differ in ecological characteristics. Therefore, modelling all these subpopulations together would effectively overestimate the species ecological breath and hence reduce model accuracy. Another potentially simpler explanation is that species described to have



**Figure 5** Mean model accuracy for Catalan Breeding Bird Survey models in relation to available sample size RANDOM subsampling scenario and species range size. Whiskers represent the standard error of the estimates.



**Figure 6** Mean model accuracy for Catalan Breeding Bird Survey (SOCC) reference models in relation to species range size and to the spatial distribution of subsamples in three different scenarios using 25% of original SOCC section data (RANDOM, SELECTIVE) and the combined original 3-km transect data (EFFORT). Whiskers represent the standard error of the estimates.

wider distributions or use a wide range of habitats in one area might not be limited by any of the measured predictive factors at the scale at which models are fitted. Overall, these results suggest that limitations caused by species-specific traits associated to species ecology (i.e. range size and niche breadth) will be difficult to circumvent by either statistical approaches or increasing sampling effort and should always be kept in mind when conducting predictive habitat models for different species (Seoane *et al.*, 2005).

### Long-term monitoring programs and spatial modelling: perspectives and applications

Given the number of LTM programs currently running in many countries, application of spatial modelling techniques to these

data may prove a major contributor to conservation and land use planning in many areas. Spatial mapping of LTM data may substantially enhance the general efficiency of large-scale biodiversity assessments by adding a potentially useful spatially explicit component allowing accurate representation of species distributions. Furthermore, spatial mapping from LTM may be integrated in current projects specifically aimed at mapping species distributions at large spatial scales. For instance during atlas work periods, spatial mapping derived from LTM data may become an integral part of atlas methodology covering more common species. However, some limitations of habitat modelling such as the difficulty to account for fine-scale habitat structure should be kept in mind to enhance proper use of distributions maps derived from LTM data.



First, many authors consider maps generated by habitat or niche modelling as equivalent to potential distribution maps (Guisan & Zimmermann, 2000). Since they rely on the species responses to environmental gradients, occurrence of the appropriate combination of environmental variables is likely to induce the presence of the species under consideration, but this may not be always the case. Due to unrecorded environmental variables, historical reasons, habitat fragmentation, or others, niche modelling may predict species presences where a given species is certainly known to be absent. Whereas our models predicted the occurrence of most species with high accuracy, some additional steps may be added to assure that final relative abundance maps corresponded as accurately as possible as real rather than potential distribution maps (Pearce *et al.*, 2001). A possibility is to filter out hypothetical occurrence areas for each species from the known distribution of the species gathered from expert knowledge (Pearce *et al.*, 2001) or coarse resolution field atlas data (Estrada *et al.*, 2004). In fact, LTM data may complement more traditional mapping approaches which may in turn allow model refinements leading to a progressive integration between the two types of surveys (Carrascal *et al.*, 2006).

How to improve monitoring programs to obtain more reliable distribution maps? From our results we suggest that if used for mapping purposes based on habitat modelling, LTM should benefit from the effort aimed at increasing sample size. Such an increase in sampling effort is also likely to benefit also trend estimation, which is the main aim for which most of them have been launched. There is, however, a trade-off between the number of locations possibly sampled and the distance at which they are located. From our results, we suggest that LTM data based on long transects or possibly on other methods (i.e. point counts) may be disaggregated in smaller sampling units, transect sections in our case, leading to significant increases in the predictive accuracy of habitat models. The optimal degree of disaggregation to develop accurate habitat models from LTM data should be further investigated and is likely to depend on factors such as minimum unit size, species ecology, and spatial distribution of the sampling locations. The spatial coverage of the sampling scheme is likely to be an important factor in many cases and, therefore, improving this feature should be also favoured. At some point, good spatial coverage may facilitate the implementation of spatial interpolation techniques that may increase the predictive accuracy of habitat models especially for species with loose association with environmental variables. Here also, further investigation of the role that the spatial component of species distributions plays in the development of habitat models is needed (Segurado *et al.*, 2006).

Finally, we have showed that LTM data are well suited for occurrence data. Occurrence data have been often found to be a good surrogate of abundance (Pearce & Ferrier, 2001). Since, LTM programs often collect count or density data, they have the potential to be used for more informative modelling of abundance data. It is expected that combining presence/absence modelling and abundance models will better fit the data when factors determining occurrence are different from those determining abundance. Although, contradictory evidence exists about the predictive accuracy of abundance models in ecology, many studies

show that at least for a number of species it is possible to model abundance successfully (Welsh *et al.*, 2000; Pearce & Ferrier, 2001). At least when factors determining occurrence differed to some degree from those determining abundance, the combining of presence/absence modelling and abundance models may result in a better option.

## ACKNOWLEDGEMENTS

We thank all volunteers that participated in the Catalan Breeding Bird Atlas (CBBA) and the Catalan Common Bird Survey (SOCC). The two projects are financed by the Department of Environment and Housing of the Catalan Government, and the 'Fundació Territori i Paisatge de Caixa Catalunya'. We thank Miquel Caceres for statistical advice in building the niche breadth index, and Frederic Jiguet and other members of the EBCC (European Bird Census Council) spatial modelling working group for enriching comments and discussions. This work is a contribution to the project CGL2005-00031/BOS founded by the Spanish Ministry of Education and Science. L.B. benefited from a Ramon y Cajal contract from the Spanish government.

## REFERENCES

- Araújo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.
- Augustin, N.H., Muggleston, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.
- Bani, L., Baietto, M., Bottoni, L. & Massa, R. (2002) The use of focal species in designing a habitat network for a lowland area in Lombardy, Italy. *Conservation Biology*, **16**, 826–831.
- Bibby, C., Burguess, N.D. & Hill, D.A. (2000) *Bird census techniques*. Cambridge University Press, Cambridge.
- 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.
- Carrascal, L.M., Seoane, J., Palomino, D., Alonso, C.L. & Lobo, J.M. (2006) Species-specific features affect the ability of census derived models to map avian distribution. *Ecological Research*, **21**, 681–691.
- Chambers, J.M. & Hastie, T. (1997) *Statistical models in S*. Chapman & Hall, London.
- Donald, P.F. & Fuller, R.J. (1998) Ornithological atlas data: a review of uses and limitations. *Bird Study*, **45**, 129–145.
- Estrada, J., Pedrocchi, V., Brotons, L. & Herrando, S. (2004) *Catalan breeding bird atlas (1999–2002)*. Lynx, ed., Barcelona.
- Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.
- Gibbons, D.W., Reid, J.B. & Chapman, R.A. (1993) *The new atlas of breeding birds in Britain and Ireland: 1988–1991*. T & A D Poyser, London.
- 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 of London Series B-Biological Sciences*, **360**, 269–288.
- Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.
- Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.
- Hepinstall, J.A., Krohn, W.B. & Sader, S.A. (2002) *Effects of niche width on the performance and agreement of avian habitat models. Predicting species occurrences: issues of accuracy and scale* (ed. by M.S. Scott, P. Heglund, M.L. Morrison, M.G. Raphael, A. Wall and F.B. Smason), Island Press, Washington.
- Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331–341.
- Jiguet, F., Julliard, R., Couvet, D. & Petiau, A. (2005) Modeling spatial trends in estimated species richness using breeding bird survey data: a valuable tool in biodiversity assessment. *Biodiversity and Conservation*, **14**, 3305–3324.
- Johnson, D.H. & Sargeant, G.A. (2002) Predicting species occurrences: issues of accuracy and scale. *Towards better atlases: improving presence-absence information* (ed. by M.S. Scott, P. Heglund, M.L. Morrison, M.G. Raphael, A. Wall and F.B. Smason). Island Press, Washington.
- Kery, M. & Schmid, H. (2005) Estimating species richness: calibrating a large avian monitoring programme. *Journal of Applied Ecology*.
- Legendre, P. & Legendre, L. (1998) *Numerical ecology*. Elsevier, Amsterdam.
- Link, W.A. & Sauer, J.R. (1998) Estimating population change from count data: Application to the North American Breeding Bird Survey. *Ecological Applications*, **8**, 258–268.
- Manel, S., Dias, J.M. & Ormerod, S.J. (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337–347.
- 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.
- McPherson, J.M., Jetz, W. & Rogers, D.J. (2006) Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions-possibilities and limitations. *Ecological Modelling*, **192**, 499–522.
- Osborne, P.E., Alonso, J.C. & Bryant, R.G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458–471.
- Pearce, J.L., Cherry, K., Drielsma, M., Ferrier, S. & Whish, G. (2001) Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412–424.
- Pearce, J. & Ferrier, S. (2000a) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.
- Pearce, J. & Ferrier, S. (2000b) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127–147.
- 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.
- Pebesma, E.J., Duin, R.N.M. & Burrough, P.A. (2005) Mapping sea bird densities over the North Sea: spatially aggregated estimates and temporal changes. *Environmetrics*, **16**, 573–587.
- Price, J.T., Droege, S. & Price, A. (1995) *Summer atlas of North American birds*. Academic Press, London.
- Schmid, H., Luder, R., Naef-Daenzer, R., Graf, R. & Zbinden, N. (1998) *Atlas de oiseaux nicheurs de Suisse et du Liechtenstein en 1993–96*. Station Ornithologique Suisse, Sempach, Switzerland.
- Segurado, P., Araújo, M.B. & Kunin, W.E. (2006) Consequences of spatial autocorrelation on niche-based models. *Journal of Applied Ecology*, **43**, 433–444.
- Seoane, J., Carrascal, L.M., Alonso, C.L. & Palomino, D. (2005) Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling*, **185**, 299–308.
- Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.
- Thogmartin, W.E., Knutson, M.G. & Sauer, J.R. (2006) Predicting regional abundance of rare grassland birds with a hierarchical spatial count model. *Condor*, **108**, 25–46.
- Thogmartin, W.E., Sauer, J.R. & Knutson, M.G. (2004) A hierarchical spatial model of avian abundance with application to Cerulean warblers. *Ecological Applications*, **14**, 1766–1779.
- Underhill, L.G. & Gibbons, D.W. (2002) Conserving bird biodiversity: general principles and applications. *Mapping and monitoring bird populations: their conservation uses* (ed. by K. Norris and D.J. Pain). University Press, Cambridge.
- Welsh, A.H., Cunningham, R.B. & Chambers, R.L. (2000) Methodology for estimating the abundance of rare animals: Seabird nesting on North East Herald Cay. *Biometrics*, **56**, 22–30.

**Appendix 1** Environmental variables (ENV) used to generate habitat suitability models used in the comparison of the Catalan Breeding Bird Atlas and the Catalan Breeding Bird Survey data. Unless otherwise mentioned, variables refer to  $1 \times 1$  km squares correspond to means obtained from averaging individual values from pixels contained each  $1 \times 1$  km square. Cartographic sources are indicated when necessary.

Descriptor type	Variable description [units]	Range
Forest	Coniferous forest*	0–400
	Esclerophyllous forest*	0–397
	Deciduous forest*	0–400
	<i>Pinus halepensis</i> forest†	0–393
	<i>Pinus sylvestris</i> forest†	0–400
	<i>Abies alba</i> – <i>Pinus uncinata</i> forest pixels in $1 \times 1$ km squares†	0–400
	<i>Pinus nigra</i> forest†	0–400
	Other <i>Pinus</i> forest†	0–394
	<i>Quercus suber</i> forest†	0–400
	<i>Quercus ilex</i> forest†	0–400
	<i>Quercus humilis</i> forest†	0–393
	Other deciduous forest†	0–400
	Distance to nearest forest patch [log m]†	0–10
Agriculture	Dry herbaceous cropland (cereals)*	0–400
	Irrigated herbaceous cropland pixels (alfalfa, corn)*	0–400
	Dry arboreal cropland (olive tree, almond)*	0–400
	Irrigated arboreal cropland (fruit trees)*	0–371
	Vineyard*	0–387
Low vegetation cover	Scrub*	0–400
	Bare ground (rocks)*	0–389
	Wetland vegetation*	0–400
Landscape	Number of land uses in $1 \times 1$ km squares (based on land use cover 1997, urban and industrial categories clumped)*	1–11
Human impact	Low density urbanization*	0–379
	Distance to cities > 10.000 h [log m]†	0–11
	Infrastructure (transport network and urban areas)*	0–400
	Distance to main roads of the primary road network [log m]*	0–10
	Distance to roads of the secondary road network [log m]*	0–10
Climate	Mean solar radiation‡ [10 kJm <sup>2</sup> /day]	19–896
	Mean accumulated summer precipitation (June–September) [l m <sup>-2</sup> ]‡	50–500
	Mean accumulated of mean winter temperatures (December–March) [°C]‡	–50–105
	Mean accumulated of mean annual temperatures (January–December) [°C]‡	0–150
Productivity	First PCA factor from NDVI monthly data (April–July)§	–4.1–2.6
	Second PCA factor from NDVI monthly data (April–July)§	–4.2–3.4
Topography	Mean altitude [m]*	0–2780
	Mean slope [degrees]*	0–39
	Standard deviation of slope [degrees]*	0–15.2
	Distant to rivers [log m]*	3.3–8.8
Geography	Mean latitude [degrees]*	2.70–3.80
	Mean longitude [degrees]*	45.70–46.50

\*‘Institut Cartogràfic de Catalunya’ (ICC).

†‘Departament de Medi Ambient de la Generalitat de Catalunya’ (DAM).

‡‘Centre de Recerca Ecològica i Aplicacions Forestals’ (CREAF).

§NOAA satellite data.

**Appendix 2** Species characteristics used to evaluate the potential of long-term monitoring (LTM) data for spatial modelling. Model accuracy as estimated using the area under the ROC curve (AUC). AUC values are shown for reference Catalan Breeding Bird Atlas (CBBA) models (Estrada *et al.*, 2004) and for models conducted on LTM data [Catalan Breeding Bird Survey (SOCC) models]. AUC results for SOCC models correspond to the best models obtained using 1-km section data for all available transects and evaluated on independent data from the CBBA data set. Range size was estimated as the total number of 10 × 10 km UTM squares occupied by the species in Catalonia. Prevalence indicates the number of occurrences for each species within the CBBA data set ( $N = 3077$  1 × 1 km UTM squares) and the SOCC data set ( $N = 684$  1 km long transect sections). Niche breadth was estimated from species habitat selection patterns in main habitat types derived from 1 × 1 km square data (see Estrada *et al.*, 2004 for further details).

Name	Range size	Prevalence CBBA data	Prevalence SOCC data	Niche breadth	AUC CBBA models	AUC SOCC models
<i>Acrocephalus arundinaceus</i>	94	132	31	— *	0.89	0.83
<i>Acrocephalus scirpaceus</i>	72	102	60	—	0.93	0.86
<i>Aegithalos caudatus</i>	247	1274	351	0.50	0.82	0.79
<i>Alauda arvensis</i>	144	272	77	0.59	0.88	0.80
<i>Alcedo atthis</i>	74	102	30	—	0.89	0.71
<i>Alectoris rufa</i>	264	710	179	0.42	0.79	0.75
<i>Anthus campestris</i>	82	101	38	0.62	0.82	0.76
<i>Anthus trivialis</i>	64	112	59	0.44	0.91	0.82
<i>Apus apus</i>	251	2045	250	0.48	0.84	0.71
<i>Apus melba</i>	94	352	56	0.58	0.78	0.63
<i>Carduelis cannabina</i>	243	996	207	0.67	0.84	0.77
<i>Carduelis carduelis</i>	296	1812	428	0.44	0.87	0.78
<i>Carduelis chloris</i>	292	1637	407	0.44	0.81	0.72
<i>Certhia brachydactyla</i>	281	1452	420	0.50	0.83	0.74
<i>Cettia cettia</i>	239	751	201	0.44	0.88	0.81
<i>Charadrius dubius</i>	84	101	42	—	0.90	0.76
<i>Cisticola juncidis</i>	173	575	141	0.33	0.92	0.85
<i>Clamator glandarius</i>	56	73	28	0.47	0.86	0.67
<i>Columba livia</i>	223	668	144	0.31	0.84	0.77
<i>Columba oenas</i>	108	299	36	0.50	0.86	0.73
<i>Columba palumbus</i>	300	2464	586	0.54	0.82	0.75
<i>Corvus corax</i>	181	645	119	0.56	0.76	0.65
<i>Corvus corone</i>	155	801	167	0.49	0.89	0.83
<i>Corvus monedula</i>	61	120	32	0.37	0.86	0.79
<i>Coturnix coturnix</i>	203	400	125	0.59	0.86	0.81
<i>Cuculus canorus</i>	299	1462	414	0.44	0.82	0.71
<i>Delichon urbicum</i>	290	1019	174	0.38	0.76	0.68
<i>Dendrocopos major</i>	221	763	249	0.41	0.84	0.79
<i>Emberiza calandra</i>	260	1035	215	0.51	0.90	0.87
<i>Emberiza cia</i>	195	719	160	0.47	0.89	0.82
<i>Emberiza cirrus</i>	294	1674	324	0.53	0.83	0.75
<i>Emberiza citrinella</i>	49	79	35	0.43	0.97	0.91
<i>Emberiza hortulana</i>	91	119	37	0.71	0.83	0.76
<i>Erithacus rubecula</i>	266	1734	452	0.56	0.92	0.90
<i>Falco tinnunculus</i>	236	812	175	0.68	0.76	0.67
<i>Fringilla coelebs</i>	251	1540	368	0.52	0.94	0.91
<i>Galerida cristata</i>	225	983	188	0.44	0.90	0.87
<i>Galerida theklae</i>	98	223	24	0.30	0.88	0.84
<i>Gallinula chloropus</i>	159	355	101	—	0.89	0.81
<i>Garrulus glandarius</i>	217	1699	448	0.52	0.86	0.79
<i>Hippolais polyglotta</i>	229	842	249	0.50	0.79	0.72
<i>Hirundo rustica</i>	293	2040	384	0.45	0.86	0.79
<i>Jynx torquilla</i>	203	270	95	0.60	0.74	0.65
<i>Lanius collurio</i>	93	200	59	0.36	0.92	0.88
<i>Lanius meridionalis</i>	92	176	43	0.47	0.82	0.73
<i>Lanius senator</i>	188	532	139	0.27	0.83	0.73
<i>Loxia curvirostra</i>	86	212	54	0.48	0.89	0.85
<i>Lullula arborea</i>	258	914	210	0.67	0.85	0.78

## Appendix 2 continued.

Name	Range size	Prevalence CBBA data	Prevalence SOCC data	Niche breadth	AUC CBBA models	AUC SOCC models
<i>Luscinia megarhynchos</i>	297	2039	481	0.54	0.88	0.81
<i>Melanocorypha calandra</i>	26	43	30	0.29	0.98	0.93
<i>Merops apiaster</i>	213	1114	233	0.45	0.85	0.79
<i>Monticola solitarius</i>	135	251	42	0.43	0.85	0.83
<i>Motacilla alba</i>	244	990	243	0.58	0.77	0.67
<i>Motacilla cinerea</i>	169	369	85	0.61	0.83	0.71
<i>Muscicapa striata</i>	166	354	109	0.48	0.72	0.67
<i>Myiopsitta monachus</i>	28	38	45	— *	0.99	0.93
<i>Oenanthe hispanica</i>	107	312	43	0.20	0.93	0.88
<i>Oenanthe oenanthe</i>	63	121	70	0.26	0.97	0.88
<i>Oriolus oriolus</i>	273	983	317	0.50	0.77	0.69
<i>Parus ater</i>	200	853	210	0.55	0.91	0.89
<i>Parus caeruleus</i>	277	1591	413	0.50	0.86	0.79
<i>Parus cristatus</i>	248	1213	335	0.63	0.84	0.80
<i>Parus major</i>	305	2436	563	0.58	0.83	0.83
<i>Passer domesticus</i>	300	1994	419	0.43	0.88	0.84
<i>Passer montanus</i>	186	724	217	0.34	0.89	0.82
<i>Petronia petronia</i>	149	207	40	0.57	0.74	0.70
<i>Phasianus colchicus</i>	55	50	34	—	0.84	0.79
<i>Phylloscopus bonelli</i>	272	1288	293	0.51	0.87	0.76
<i>Phylloscopus collybita</i>	207	731	289	0.46	0.88	0.79
<i>Pica pica</i>	237	1294	316	0.40	0.91	0.84
<i>Picus viridis</i>	290	1379	388	0.62	0.72	0.66
<i>Prunella modularis</i>	77	254	88	0.41	0.96	0.96
<i>Ptyonoprogne rupestris</i>	177	426	47	0.57	0.86	0.78
<i>Pyrrhonorax pyrrhonorax</i>	22	232	45	0.40	0.95	0.86
<i>Pyrrhula pyrrhula</i>	79	269	77	0.44	0.95	0.90
<i>Regulus ignicapilla</i>	238	1066	328	0.52	0.88	0.83
<i>Regulus regulus</i>	47	124	41	0.37	0.98	0.96
<i>Remiz pendulinus</i>	68	135	38	0.31	0.96	0.88
<i>Saxicola torquatum</i>	283	1159	296	0.52	0.74	0.70
<i>Serinus citrinella</i>	53	153	47	0.38	0.96	0.94
<i>Serinus serinus</i>	307	2334	547	0.47	0.90	0.79
<i>Sitta europaea</i>	126	374	111	0.51	0.90	0.87
<i>Streptopelia decaocto</i>	192	409	180	0.30	0.88	0.81
<i>Streptopelia turtur</i>	241	1161	236	0.51	0.81	0.73
<i>Sturnus unicolor</i>	180	519	93	0.38	0.87	0.70
<i>Sturnus vulgaris</i>	251	1303	294	0.42	0.87	0.81
<i>Sylvia atricapilla</i>	282	1702	491	0.50	0.82	0.77
<i>Sylvia borin</i>	141	215	97	0.50	0.78	0.66
<i>Sylvia cantillans</i>	221	894	235	0.53	0.87	0.76
<i>Sylvia hortensis</i>	112	149	37	0.68	0.82	0.74
<i>Sylvia melanocephala</i>	241	1512	360	0.46	0.93	0.89
<i>Sylvia undata</i>	193	435	107	0.63	0.86	0.77
<i>Tetrax tetrax</i>	22	57	33	0.26	0.97	0.97
<i>Troglodytes troglodytes</i>	271	1566	402	0.54	0.87	0.81
<i>Turdus merula</i>	308	2614	599	0.56	0.89	0.81
<i>Turdus philomelos</i>	223	841	275	0.50	0.84	0.78
<i>Turdus torquatus</i>	42	95	26	0.36	0.98	0.97
<i>Turdus viscivorus</i>	245	1038	259	0.66	0.79	0.73
<i>Upupa epops</i>	267	1297	291	0.47	0.82	0.74

\*Niche breadth for water birds and invasive species was not available due to biases derived from the habitat categories used.