Research

Ecography 37: 001–012, 2014
doi: 10.1111/ecog.00749

Subject Editor: Miguel Araújo. Accepted 11 April 2014

Optimising long-term monitoring projects for species distribution modelling: how atlas data may help

Olatz Aizpurua, Jean-Yves Paquet, Lluís Brotons and Nicolas Titeux

O. Aizpurua and N. Titeux (nicolas.titeux@ctfc.es), Centre de Recherche Public – Gabriel Lippmann, Dépt Environnement et Agro-biotechnologies, 41 rue du Brill, LU-4422 Belvaux, Luxembourg. – L. Brotons, OA and NT, Forest Sciences Centre of Catalonia (CEMFOR-CTFC), Ctra. Sant Llorenç de Morunys, km 2, ES-25280 Solsona, Spain. LB also at: Centre de Recerca Ecològica i Aplicacions Forestals (CREAF), Univ. Autònoma de Barcelona, ES-08193 Bellaterra, Spain, and Inst. Català d'Ornitologia (ICO), Museu de Zoologia, Passeig Picasso s/n, ES-08003 Barcelona, Spain. NT also at: Univ. de Liège – Gembloux Agro-Bio Tech, Unité Biodiversité et Paysage, 2 Passage des Déportés, BE-5030, Gembloux, Belgium. – J.-Y. Paquet, Aves-Natagora, Dépt Études, 98 rue Nanon, BE-5000 Namur, Belgium.

Long-term biodiversity monitoring data are mainly used to estimate changes in species occupancy or abundance over time, but they may also be incorporated into predictive models to document species distributions in space. Although changes in occupancy or abundance may be estimated from a relatively limited number of sampling units, small sample size may lead to inaccurate spatial models and maps of predicted species distributions. We provide a methodological approach to estimate the minimum sample size needed in monitoring projects to produce accurate species distribution models and maps. The method assumes that monitoring data are not yet available when sampling strategies are to be designed and is based on external distribution data from atlas projects. Atlas data are typically collected in a large number of sampling units during a restricted timeframe and are often similar in nature to the information gathered from long-term monitoring projects. The large number of sampling units in atlas projects makes it possible to simulate a broad gradient of sample sizes in monitoring data and to examine how the number of sampling units influences the accuracy of the models. We apply the method to several bird species using data from a regional breeding bird atlas. We explore the effect of prevalence, range size and habitat specialization of the species on the sample size needed to generate accurate models. Model accuracy is sensitive to particularly small sample sizes and levels off beyond a sufficiently large number of sampling units that varies among species depending mainly on their prevalence. The integration of spatial modelling techniques into monitoring projects is a cost-effective approach as it offers the possibility to estimate the dynamics of species distributions in space and over time. We believe our innovative method will help in the sampling design of future monitoring projects aiming to achieve such integration.

Long-term wildlife monitoring is generally considered as an essential tool for biodiversity management and for research studies on biodiversity conservation (Gitzen et al. 2012). Monitoring projects primarily aim at delivering information on the changing status of key features of biodiversity (Lindenmayer et al. 2012). State variables are used to characterise the status of these features at different points in time with a view to assessing system state and inferring changes in state over time (Gitzen et al. 2012). State variables include, among others, species occupancy (MacKenzie et al. 2005, Kéry et al. 2009) or species abundance (Royle and Nichols 2003). In such projects, field data are often repeatedly collected over time in a network of sampling units according to standardised procedures (Gitzen et al. 2012). Previous studies have reported that monitoring projects have also the potential to provide an appropriate source of data to document the distribution of species in space (Brotons et al. 2007, Braunisch and Suchant 2010, Rodhouse et al. 2012). Mapping species distributions in space and documenting how they change over time may provide key information to guide effective landscape and conservation planning. Dynamic species distribution mapping may, therefore, be considered as an essential component of a biodiversity monitoring project (Brotons et al. 2007, Kéry et al. 2013). In any monitoring project, sampling units are, however, sparsely distributed over the region of interest, which is inconvenient for a straightforward mapping of species distributions.

Species distribution modelling is an increasingly used technique (Rodríguez et al. 2007) that can produce distribution maps based on monitoring data (Brotons et al. 2006). With these models, environmental variables describing the habitat conditions in the sampling units are related to records of species presence. These models are used to predict the species distribution beyond the sampling units in areas where species occurrence is unknown (Araújo and Guisan 2006, Elith et al. 2010). The use of models to predict species distributions is of key significance for biodiversity conservation (Guisan et al. 2013). Among several applications, models

may be used to identify the most important environmental conditions that influence species distributions or to guide the prioritization of management options amongst areas that vary in their suitability for the species (Titeux et al. 2007). Species distribution models are also often built to explore the impacts of environmental changes on future species distributions (Elith et al. 2010). Previous studies examined the use of monitoring data to generate species distribution models (Brotons et al. 2006, 2007, Braunisch and Suchant 2010) and showed that the integration of monitoring data into modelling approaches may contribute to understanding how species distributions change over time (De Cáceres and Brotons 2012, Rodhouse et al. 2012, Kéry et al. 2013).

Sampling design in a monitoring project typically results from a balance between the number of sampling units and the number of repeated surveys in these units to document the state variables with an acceptable level of precision (MacKenzie et al. 2005). A limited number of sampling units and a sufficient number of repeated surveys may be suited, and in some cases recommended, to derive unbiased estimates of the state variables (MacKenzie and Royle 2005, Kéry et al. 2009, MacKenzie 2012). This appropriate sampling design for monitoring purposes may, however, fail to produce enough spatial data to build relevant species distribution models (Brotons et al. 2007), because a small number of sampling units is known to induce inaccurate spatial models (Hernandez et al. 2006, Wisz et al. 2008, Jiménez-Valverde et al. 2009, Bean et al. 2012). This drawback can be avoided if dynamic species distribution mapping is explicitly considered when setting the objectives of the monitoring project and when making decisions about sampling design. At this stage of a project, existing monitoring data in the region of interest are, however, not yet available and other sources of information based on upfront sampling efforts are needed to help putting the monitoring project into place (Hooten et al. 2012).

Atlas projects are an interesting source of spatial information that may assist in making such pilot analysis. Two-stage sampling design (Thompson 2012) is increasingly implemented in 'last-generation' atlases (Estrada et al. 2004, Jacob et al. 2010, Maes et al. 2013): species presence or abundance is recorded in 1) primary sampling units to provide a picture of the species distribution across the whole region of interest but at coarse spatial resolution and in 2) a set of secondary sampling units nested within the primary ones to explore species distribution at finer resolution. Last-generation atlases are generally completed over considerable time periods and repeated at long time intervals (Dunn and Weston 2008), which prevents them from being suited to detect changes in species distributions with time scales matching decision-making needs. Interestingly, field sampling procedures for atlas data collection in secondary sampling units (e.g. bird or butterfly counts along transects) are often similar in nature to the procedures implemented in long-term monitoring projects (Van Swaay et al. 2008, Vorísek et al. 2008). Such kind of atlas data are generally collected only once during the atlas period, but in a large number of secondary sampling units to cover an important part of the region of interest at a fine spatial resolution (Carden et al. 2010, Maes et al. 2012). Hence, atlas data in secondary sampling units may be manipulated to imitate a broad gradient of sample sizes in a monitoring project and to build species distribution models with varying numbers of sampling units. Such an approach may, in turn, contribute to identifying how large the number of sampling units should be at the start of a monitoring project if dynamic species distribution mapping is set as an objective.

Here, we provide an innovative analytical framework using data from last-generation atlases to aid in the initial design of monitoring projects able to generate appropriate data for the production of accurate species distribution models and maps. We draw attention to important issues that are to be addressed if we are to generate and update species distribution maps as a direct output of long-term monitoring projects. This study illustrates how datasets derived from last-generation atlas projects can contribute to the integration of spatial modelling techniques into long-term monitoring studies in order to cost-efficiently estimate biodiversity dynamics in space and over time (Rodríguez et al. 2007).

Methods

An increasing number of atlas projects with two-stage sampling designs become available for different taxa worldwide (Estrada et al. 2004, Carden et al. 2010, Maes et al. 2012) and may support the integration of spatial modelling techniques into monitoring studies. The following analytical framework is of general interest as it can be applied to any dataset derived from such last-generation atlas projects. In the present study, we apply this innovative method to the 'Breeding Bird Atlas of Wallonia' (BBAW) data (Jacob et al. 2010).

Study area

Belgium is a heavily industrialized north-western European country with a high human population density. The southern part of Belgium (Wallonia, ca 16 850 km^2, Fig. 1a) is characterised by a strong gradient in landscape composition, from a densely populated and agriculture dominated landscape in the northwest to a hilly landscape with an important cover of forest and grassland in the southeast (Jacob et al. 2010).

Atlas data

During 2001–2007, 650 volunteer fieldworkers participated in the BBAW data collection. Data were collected across a range of spatial resolutions according to a two-stage sampling design and an additional territory-mapping procedure.

Grid-based procedure: primary sampling units

Based on regular field visits during day and night from February to August, fieldworkers were asked to report the presence, estimate the abundance and record the breeding evidence for all bird species in 40 km^2 (5 × 8 km) primary sampling units (n = 514, Fig. 1b). Fieldworkers paid particular attention to survey the different habitat types present in the primary sampling units. Abundance was estimated by fieldworkers in the form of 9 abundance classes derived from

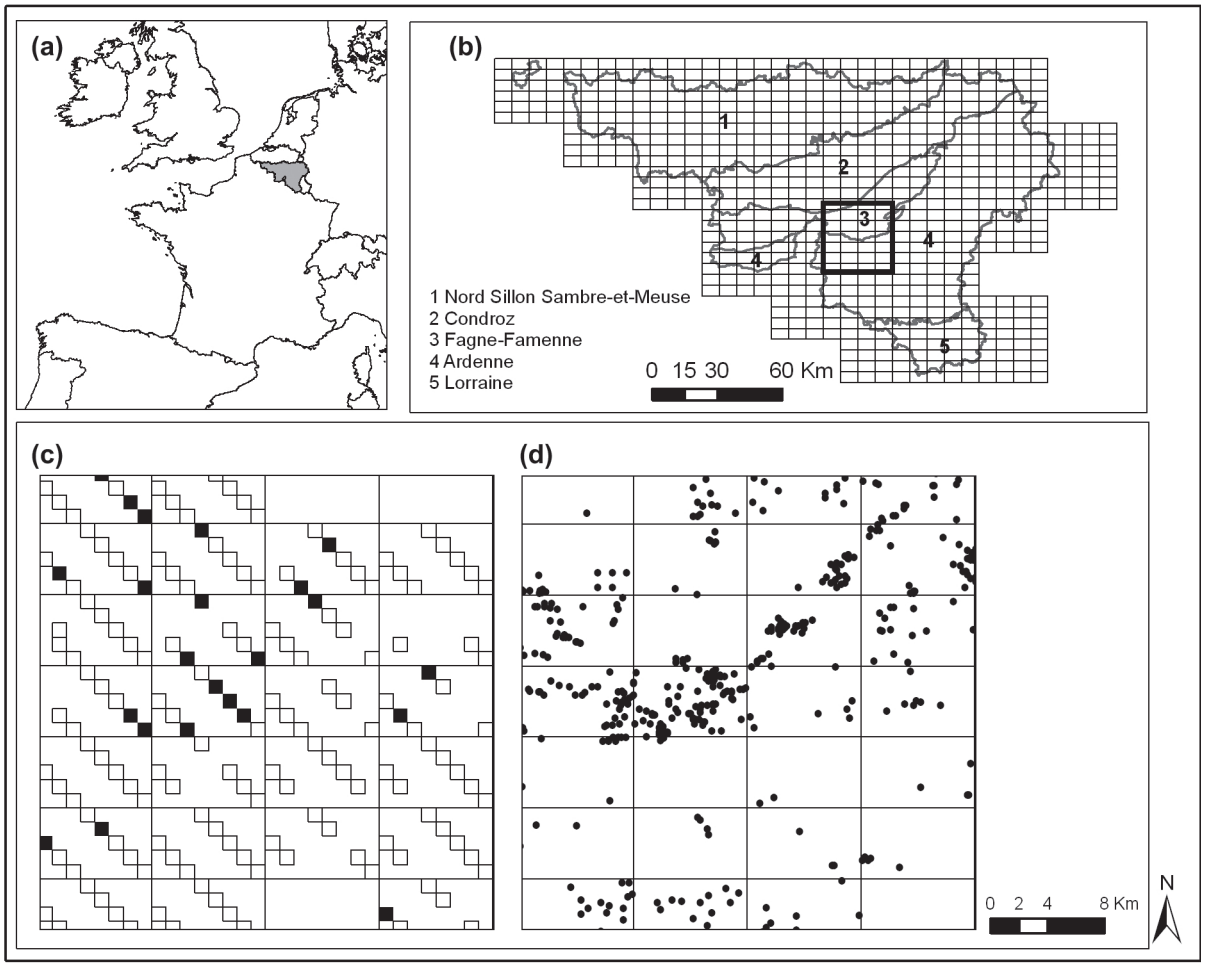

Figure 1. (a) Location of Wallonia in NW Europe. (b) Main ecological regions in Wallonia and grid system of the Breeding Bird Atlas of Wallonia with the 40-km^2 (5 × 8 km) primary sampling units. (c) Subset of the study area with the 1-km^2 secondary sampling units (black squares show an example with red-backed shrike *Lanius collurio* presence records collected during the transect-based procedure). (d) Same subset of the study area as in (c) with *L. collurio* territories (black dots) mapped during the simplified territory-mapping procedure.

a geometric progression with a common ratio set to 2 (see details in Jacob et al. 2010) and the central value of each class was used in subsequent analyses. The highest possible breeding evidence for each species was provided according to the EOAC classification, i.e. non-breeding, possible breeding, probable breeding and confirmed breeding (Timothy and Sharrock 1974).

Transect-based procedure: secondary sampling units

Secondary sampling units of 1-km^2 squares were selected according to a regular and systematic sampling design (see details in Jacob et al. 2010) so that all primary sampling units were geographically covered in the same way by the secondary sampling units (Fig. 1c). Within these secondary sampling units, transects were delineated by volunteer fieldworkers to cover the whole diversity of habitats in the squares. Fieldworkers walked during 1 h along these sampling routes in the first five hours after sunrise and twice a year during breeding season to record early and late breeders. Each breeding or non-breeding bird (detected either by sight or by sound) was recorded individually. In each secondary sampling unit, the transect-based procedure was conducted in only one year during the timeframe of the BBAW project. The number of secondary sampling units surveyed during the BBAW project (n = 2800) covered almost 17% of the study area.

Territory-mapping procedure

At the start of the BBAW project, bird species were classified in low-, moderate- and high-abundance species according to prior knowledge of their regional abundance. Based on territorial indications collected during the regular field visits conducted in the diversity of habitat types within the primary sampling units, fieldworkers were asked to map the locations of all detected territories or colonies of low- and moderate-abundance species. These locations were considered as the centres of the territories and were associated with an accuracy ranging from 100 to 500 m as estimated by the fieldworkers (Fig. 1d). This simplified territory-mapping procedure is a detailed and time-consuming technique and is unachievable over large areas on a regular basis.

Overview of the modelling approach

In our analytical framework (Fig. 2), we considered the data collected during the transect-based procedure in the secondary sampling units as equivalent to long-term monitoring data (Vorísek et al. 2008, Maes et al. 2012). We used these data as a basis to produce large-scale, fine-resolution species distribution models (hereafter 'transect-based models') and we manipulated the number of secondary sampling units in order to examine the effect of sample size on the performance of the models. The territory-mapping data covered the whole study area and provided the best available information on the distribution and habitat requirements of low- to moderate-abundance species. Therefore, we used territory-mapping data as a reference to evaluate the performance of the transect-based models. Then, we calculated the minimum sample size (i.e. minimum number of secondary sampling units) needed to reach an acceptable level of modelling performance based on three different evaluation measures. Finally, we evaluated for the whole set of species the effect of prevalence, range size and habitat specialization on the minimum sample size (redundancy analysis).

Transect-based model training

We randomly selected subsets of the available secondary sampling units to simulate a range of sample sizes in a long-term monitoring project (Jiménez-Valverde et al. 2009): 0.5% of the study area (sample size: n = 83 secondary sampling units), 1% (n = 166), 2% (n = 332), 4% (n = 664), 6% (n = 996), 8% (n = 1328) and 12% (n = 1992). In order

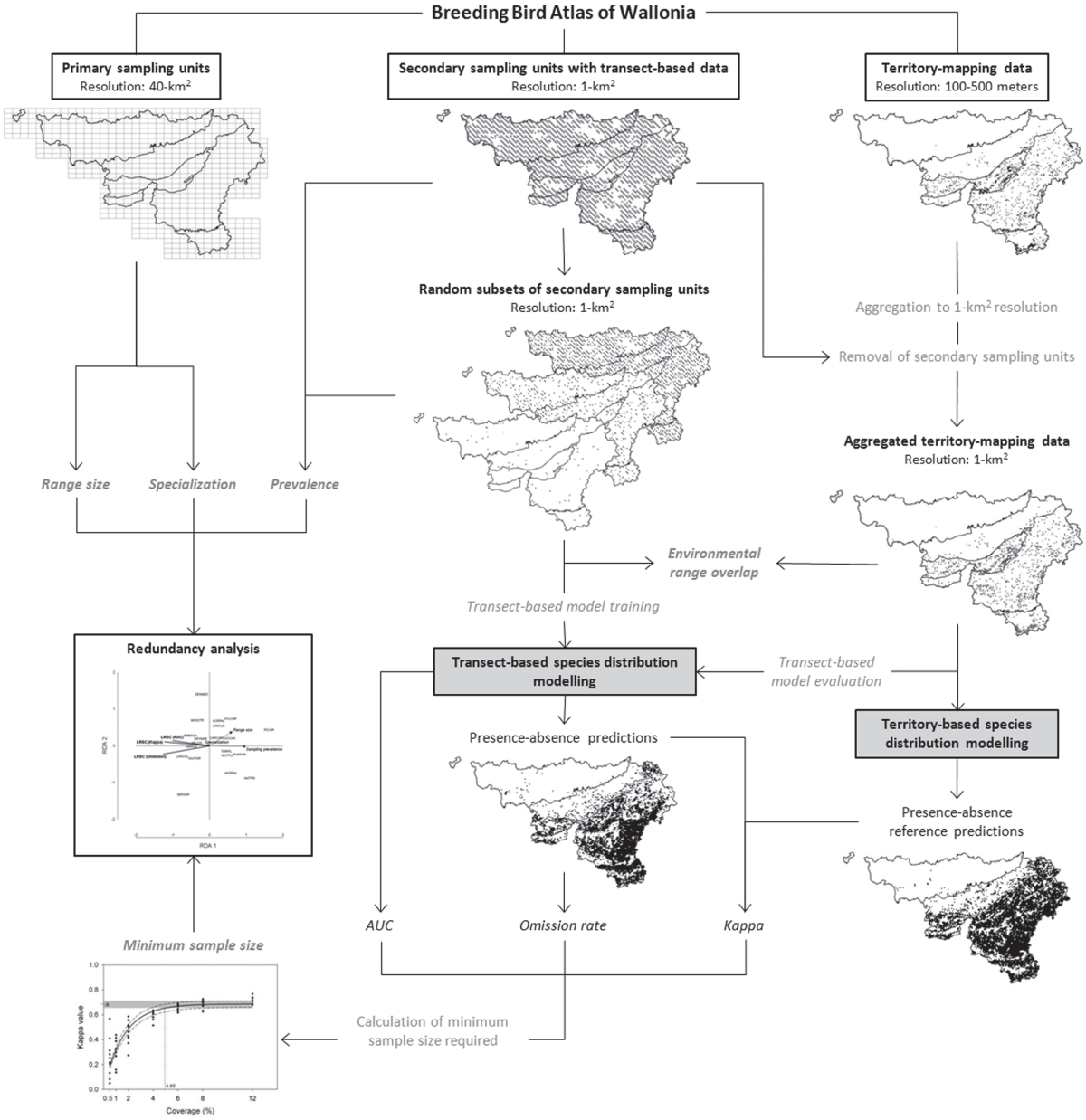

Figure 2. Overview of the modelling and analytical framework. Red-backed shrike *Lanius collurio* is used as an example.

for the subsets of secondary sampling units to be spread out over the whole environmental gradient in the study area, they were generated using a stratified random sampling procedure (Thompson 2012) with the main ecological regions in Wallonia as environmental strata (Jacob et al. 2010, Fig. 1b). We iterated this stratified random sampling with ten bootstrap replicates for each sample size.

We used 23 environmental variables (Supplementary material Appendix 1, Table A1) that characterize the most important habitat conditions for birds (e.g. elevation, climate, land cover and soil type) in southern Belgium (Jacob et al. 2010) as predictors in the models. These variables were sourced from available GIS data layers and sampled in the 1-km^2 squares that are completely within the boundaries of Wallonia (n = 16 600). We considered a secondary sampling unit as occupied by the species when at least one individual was recorded with breeding evidence during the transect-based procedure. The species that were included in the modelling exercise (hereafter 'focal species', Table 1) fulfilled four criteria: 1) they were recorded as present in all randomly generated subsets of secondary sampling units for model training, 2) territory-mapping data for model evaluation were available, 3) they are diurnal songbird species, and 4) their territory size or home range is on average lower than or close to the spatial resolution of the secondary sampling units.

Reliable absence data were unavailable and this issue may produce inaccurate presence–absence models (Brotons et al. 2004, Lobo et al. 2010). Hence, we applied the presence-only maximum entropy framework Maxent 3.3.1 (Phillips et al. 2006). Maxent is only moderately sensitive to sample size and outperforms other methods when sample size is small (Hernandez et al. 2006, Wisz et al. 2008, Bean et al. 2012).

For each focal species and sample size, model training was performed with the ten randomly generated subsets of secondary sampling units. The quadratic terms of the continuous environmental variables were included in addition to the linear functions. The continuous modelling outputs were converted into binary predictions by setting a threshold probability value above which the species was predicted as present. To set this value, we assumed that some presence records were located in unsuitable areas (Hirzel and Le Lay 2008) and we defined a threshold such that an omission rate of 10% was specified in the subsets of secondary sampling units used for model training (Martin et al. 2013). This method allows fixing a threshold that is independent of the false positive fraction, which is suitable in the case of presence-only data (Pearson et al. 2007).

Territory-based model training: reference distribution maps

Using the same 1-km^2 squares as for the environmental variables, we considered a square as occupied by a species when it enclosed the centre of at least one territory of the species recorded during the territory-mapping procedure. In order to avoid redundancy between the data used for model training and model evaluation (see below), we removed from the territory-mapping data the 1-km^2 squares that coincide with the set of secondary sampling units. The remaining territory-mapping data were used to build reference territory-based distribution models with the same environmental variables as for the transect-based models. Using a bootstrap approach, we fitted and averaged ten models for each focal species based on random selections of 70% of the territory-mapping data for model training. In order to create a reference distribution map for each focal species, the modelling outputs were converted into presence–absence predictions with the same threshold decision rule as for the transect-based models (10% of omission rate in the training data).

Table 1. Minimum sampling coverage (MSC: percentage of the study area) and sample size (MSS: number of secondary sampling units) needed to achieve an acceptable level of modelling performance according to omission rate, area under the curve of a ROC plot (AUC) and kappa value for each focal species (n = 20) used in this study. The species are listed by decreasing order of prevalence in secondary sampling units.

Species	Code	Prevalence	Range size	Specialization	Omission rate MSC/MSS	AUC MSC/MSS	Kappa MSC/MSS
Picus viridis	PICVIR	0.37 (high)	0.88 (wide)	0.86 (high)	1.78/295	0.88/145	2.48/411
Anthus trivialis	ANTTRI	0.30 (high)	0.71 (wide)	0.68 (high)	1.35/225	0.68/112	1.64/271
Pyrrhula pyrrhula	PYRPYR	0.26 (high)	0.78 (wide)	0.58 (low)	1.13/187	0.01/1	1.27/211
Anthus pratensis	ANTPRA	0.23 (high)	0.73 (wide)	0.46 (low)	1.96/325	0.73/122	1.91/317
Sylvia curruca	SYLCUR	0.22 (high)	0.85 (wide)	0.59 (low)	1.99/330	0.84/140	3.60/597
Cuculus canorus	CUCCAN	0.22 (high)	0.84 (wide)	0.44 (low)	1.92/319	0.83/138	2.49/413
Motacilla flava	MOTFLA	0.21 (high)	0.50 (restricted)	1.20 (high)	1.40/232	0.47/78	1.17/195
Carduelis carduelis	CARCAR	0.20 (high)	0.83 (wide)	0.41 (low)	2.22/369	0.75/124	1.19/198
Turdus pilaris	TURPIL	0.19 (high)	0.56 (restricted)	1.01 (high)	2.97/493	0.76/126	1.97/327
Streptopelia turtur	STRTUR	0.18 (high)	0.84 (wide)	0.43 (low)	2.54/422	0.97/160	4.49/746
Acrocephalus palustris	ACRPAL	0.17 (low)	0.85 (wide)	0.45 (low)	2.83/470	0.90/150	4.57/759
Dryocopus martius	DRYMAR	0.10 (low)	0.67 (restricted)	0.56 (low)	2.80/465	0.98/162	2.55/423
Muscicapa striata	MUSSTR	0.09 (low)	0.75 (wide)	0.46 (low)	6.16/1022	0.76/126	7.91/1313
Saxicola torquatus	SAXTOR	0.08 (low)	0.60 (restricted)	0.50 (low)	4.83/802	1.20/199	5.83/967
Dendrocopos medius	DENMED	0.08 (low)	0.62 (restricted)	1.03 (high)	2.73/453	1.08/179	4.69/778
Lanius collurio	LANCOL	0.07 (low)	0.52 (restricted)	0.71 (high)	4.08/678	0.99/165	4.95/821
Hippolais polyglotta	HIPPOL	0.06 (low)	0.53 (restricted)	0.74 (high)	5.80/964	1.11/185	8.46/1404
Miliaria calandra	MILCAL	0.05 (low)	0.27 (restricted)	1.47 (high)	3.62/601	0.94/156	4.18/694
Emberiza schoeniclus	EMBSCH	0.04 (low)	0.47 (restricted)	0.95 (high)	8.13/1350	1.78/296	8.50/1410
Serinus serinus	SERSER	0.03 (low)	0.38 (restricted)	0.63 (high)	7.76/1288	1.29/214	8.09/1342

Transect-based model evaluation

To evaluate the performance of the transect-based models, we first calculated an omission rate to measure the percentage of presence records in the evaluation territory-mapping data that were mistakenly classified as absences. Second, the area under the curve (AUC) of a receiver operating characteristic (ROC) plot was used as a threshold-independent measure of modelling performance (Fielding and Bell 1997). ROC plots were computed using presence and background data in the evaluation dataset. AUC values reflected the ability of the transect-based models to discriminate between presence data and a randomly selected secondary sampling unit (see details in Phillips et al. 2006, Jiménez-Valverde 2012). Third, we computed misclassification matrices to calculate the agreement between the binary predictions of the transect-based and the territory-based models based on the Cohen's kappa (Fielding and Bell 1997). The kappa value documented the extent to which the output of the transect-based models converged on those of the territory-based reference models (Hernandez et al. 2006).

Statistical analysis

Before analysing the modelling performance, we evaluated the extent to which the different subsets of secondary sampling units captured the range of environmental conditions used by the species. To do this, the range of all continuous variables was first normalized between 0 and 1 using a linear scaling transformation. Second, we calculated for each focal species and in each random subset of secondary sampling units, the difference between the maximum and the minimum values of the environmental variables associated with a presence record. Third, we calculated the arithmetic mean of these differences among all environmental variables to represent the width of the environmental range covered by the species in the random subsets of secondary sampling units. Fourth, we applied the same procedure for each focal species to the full set of evaluation territory-mapping data. Fifth, we computed the environmental range overlap for each focal species as the ratio between the environmental range covered by the species in the random subsets of secondary sampling units and in the territory-mapping data (Wisz et al. 2008, Feeley and Silman 2011).

Modelling performance was expected to increase with sample size and to level off beyond a sufficient number of secondary sampling units (Hernandez et al. 2006, Wisz et al. 2008). We plotted modelling performance measures against sample size and we fitted exponential functions to the data.

An exponential rise to maximum function was used for the AUC and kappa values:

$$y = a \times (1 - e^{-bx}) \quad (1)$$

Where y is the modelling performance measure, x is the sample size, a is the maximum asymptote y value, and b is the rise constant.

An exponential decay function was used for the omission rate:

$$y = y_0 + a \times e^{-bx} \quad (2)$$

Where y is the modelling performance measure, x is the sample size, y_0 is the minimum asymptote y value, $y_0 + a$ is the initial modelling performance measure when the sample size is equal to zero (forced to 1 in our case), and b is the decay constant.

For each modelling performance measure and each focal species separately, we calculated the minimum sample size (MSS, number of secondary sampling units) and sampling coverage (MSC, percentage of the study area) required to achieve an acceptable level of modelling performance, defined as the lowest x value for which the mean predicted y value was within the 95% confidence limits around the asymptote value (Fig. 3).

We calculated prevalence, range size and degree of habitat specialization for each focal species (Table 1) to evaluate how these features influence the MSS. The species prevalence was calculated from the whole set of secondary sampling units as the proportion of units in which the species was present. Species range size was calculated as the number of primary sampling units in which the species was recorded with probable or confirmed breeding evidence (McPherson et al. 2004). We used a k-means clustering analysis (Legendre and Legendre 2012) based on the continuous environmental variables (Supplementary material Appendix 1, Table A1) to allocate the primary sampling units to different habitat classes ($n = 10$ based on an analysis of the decrease in the total error sum of squares with increasing number of classes) and we used the species abundance data in the primary sampling units to calculate the degree of habitat specialization for each focal species as the coefficient of variation (= standard deviation/average) of the average species densities among the habitat classes (see details in Julliard et al. 2006). We used a redundancy analysis (RDA) to examine how much of the among-species variation in the MSS was explained by variation in prevalence, range size and habitat specialization (Legendre and Legendre 2012). In order to present the results in a simplified manner, the set of focal species was divided in equal-size categories according to prevalence (high- and low-prevalence species), range size (wide- and restricted-range species) and degree of habitat specialization (high- and low-specialization species).

Results

Range size was positively correlated with prevalence ($r = 0.70$, $p = 0.0006$) and negatively with habitat specialization ($r = -0.69$, $p = 0.0007$), but prevalence was not related to habitat specialization ($r = -0.19$, $p = 0.4131$). The training sample prevalence in the random subsets of secondary sampling units was independent of sample size (Fig. 4) and reflected the prevalence of the focal species in the whole set of sampling units (Table 1). In contrast, the proportion of the species environmental range represented in the subsets of secondary sampling units increased with sample size according to an exponential rise to maximum function (Fig. 5). This indicates that, even with the implementation of a stratified random sampling procedure, the complete range of conditions used by the species is only partly captured with very small sample sizes.

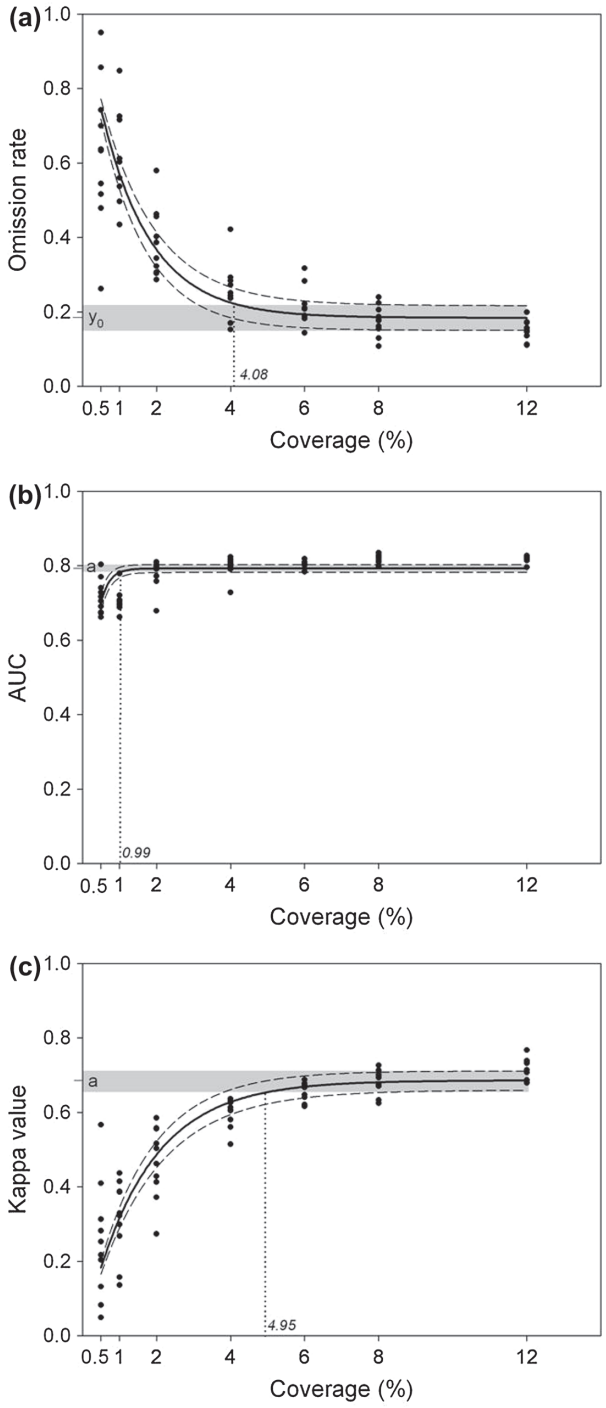

Figure 3. Identification of the minimum sampling coverage required to achieve an acceptable level of modelling performance according to (a) omission rate, (b) AUC and (c) kappa value for *Lanius collurio*. Black dots represent the modelling performance measures for the transect-based models fitted with the different subsets of secondary sampling units. Continuous and dashed black lines are the predicted average ± 95% confidence intervals after minimum square fit to the exponential function (Eq. 1 and 2). Grey areas represent the 95% confidence interval around the estimated (a) minimum or (b, c) maximum asymptote value. The dotted vertical lines indicate the minimum sampling coverage above which the modelling performance is considered to become stable.

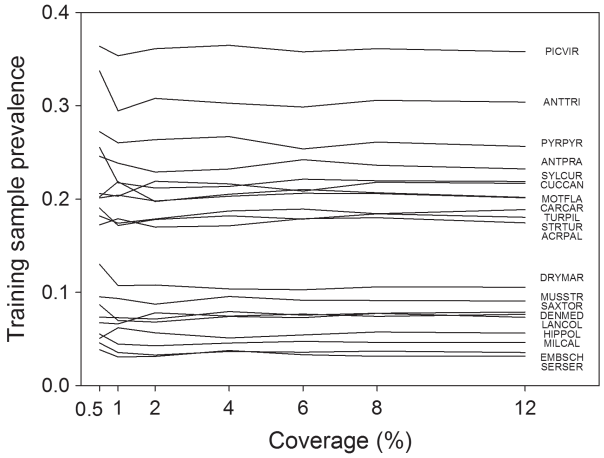

Figure 4. Average training sample prevalence of the different species along the gradient of sampling coverage. Sample prevalence was calculated as the proportion of secondary sampling units with species presence records in each individual subset of the units used to fit the models.

Figures 3 and 6 show the modelling performance (omission rate, AUC and kappa value) obtained with a number of sampling units covering 0.5 to 12% of the study area. Table 1 summarizes the minimum sample size (MSS) and coverage (MSC) calculated for each focal species according to the different modelling performance measures.

The constrained axes of the redundancy analysis (RDA, Fig. 7) explained together 57% of the total variance in the data ($R^2_{adjusted} = 0.47$). Only the first RDA axis was found to be statistically significant (permutation tests, $p < 0.001$), accounting for more than 97% of the explained variance. The first RDA axis was for the most part related to the prevalence of the species (canonical coefficient 2.39) and, to a much lower extent, to habitat specialization (0.12) and range size (0.04). Hence, the RDA results indicated that the prevalence of the focal species in the whole set of secondary sampling units had the most prominent influence on the MSS and that the effect of range size and habitat specialization can be considered negligible.

On average, the omission rate was lower for high-prevalence species than for low-prevalence species over the entire gradient of sample size, but the difference was decreasingly pronounced with an increasing sample size (Fig. 6). The exponential functions were estimated to reach their minimal value with a smaller sample size (MSS = 320 ± 29 SE) or sampling coverage (MSC = 1.93% ± 0.18%) in high-prevalence species than in low-prevalence species (MSS = 809 ± 106, MSC = 4.87% ± 0.64%). The AUC was only weakly sensitive to sample size and levelled off at smaller sample size in high-prevalence species (MSS = 115 ± 14, MSC = 0.69% ± 0.09%) than in low-prevalence species (MSS = 183 ± 15, MSC = 1.10% ± 0.09%). The kappa value increased consistently with sample size, thereby indicating that the predictions of the transect-based models gradually converged on those of the reference territory-based models. Kappa values were particularly affected by small sample size in low-prevalence species: they levelled off at smaller sample size in high-prevalence species (MSS = 369 ± 57, MSC = 2.22% ± 0.35%) than in

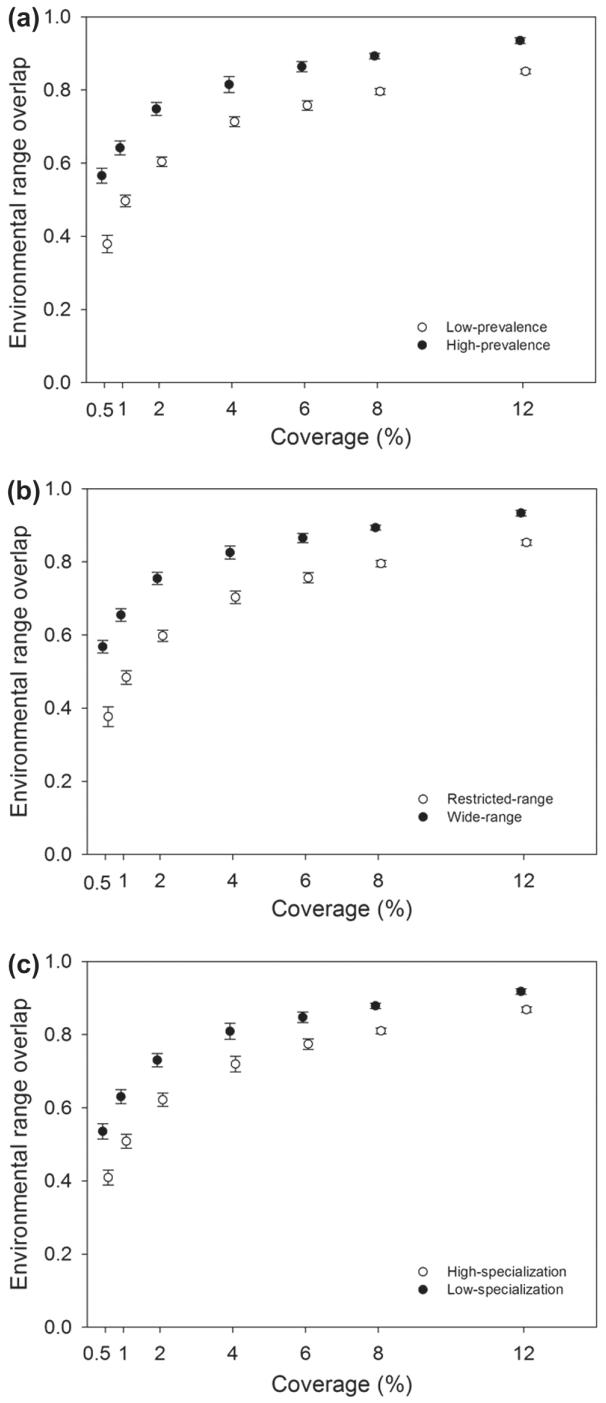

Figure 5. Average (± 95% confidence intervals) proportion of the species environmental range covered by the subsets of secondary sampling units along the gradient of sampling coverage for (a) high- and low-prevalence species, (b) wide- and restricted-range species, and (c) low- and high-specialization species.

low-prevalence species (MSS = 991 ± 111, MSC = 5.97% ± 0.67%).

Discussion

When designing a monitoring project to estimate biodiversity dynamics, a trade-off is typically made between spatial and temporal replication in the data collection strategy to minimize uncertainties associated with the estimation of changes in state variables over time (Rhodes and Jonzén 2011, Guillera-Arroita and Lahoz-Monfort 2012). In line with previous studies (De Cáceres and Brotons 2012, Rodhouse et al. 2012, Kéry et al. 2013), we argue that monitoring data may also be cost-effectively collected and used in species distribution models to document the spatial distribution of the species.

Although the influence of sample size on the performance of species distribution models is reported in many studies (Wisz et al. 2008), only few have addressed this issue when models are built with data from monitoring projects (Brotons et al. 2007). This is mostly due to the fact that dynamic distribution mapping is seldom explicitly addressed when setting the objectives of a project. If such an objective is integrated after the start of the project, the available data have been typically collected in a limited number of sampling units. This sampling design prevents from evaluating modelling performance over a broad gradient of sample sizes and from identifying how large the sample size should be to obtain an acceptable performance. On the other hand, monitoring data are unavailable when species distribution mapping is considered as an objective before the start of data collection. Other sources of information are therefore needed to help optimising the initial sampling design.

Here, we provide an analytical framework that makes use of data from large-scale last-generation atlases with two-stage sampling design to examine the influence of sample size on modelling performance and to identify how large the number of sampling units should be in a monitoring project to derive accurate species distribution maps. The method does not rely on existing data from already running monitoring projects and, hence, it may be applied before the start of field data collection when decisions about sampling design are made. The innovative idea was to consider part of the data collected during last-generation atlas projects as analogous to those derived from long-term monitoring projects (Van Swaay et al. 2008, Vorísek et al. 2008). In contrast with previous studies focusing on the link between monitoring projects and distribution modelling approaches (Brotons et al. 2007), the manipulation of atlas data allowed us to simulate a broad gradient of sample size in order to identify an optimal number of sampling units to achieve an acceptable modelling performance. The analytical framework may be easily implemented wherever such atlas data are available and where the sampling strategies of monitoring projects need to be optimised to map species distributions. Although we used bird data to illustrate our method, it is important to note that it may also be applied to other species groups for which atlas data are collected, at least partly, in the same way as in a monitoring project, such as in butterflies (Maes et al. 2013) or bats (Carden et al. 2010).

We showed that modelling performance was sensitive to particularly small sample sizes and reached an asymptote level beyond a sufficiently large number of sampling units. This result is especially interesting because it is generally assumed or reported that modelling performance increases with sample size (McPherson et al. 2004, Feeley and Silman 2011), without examining how large sample size should be to obtain sufficiently well-performing models. Wintle and

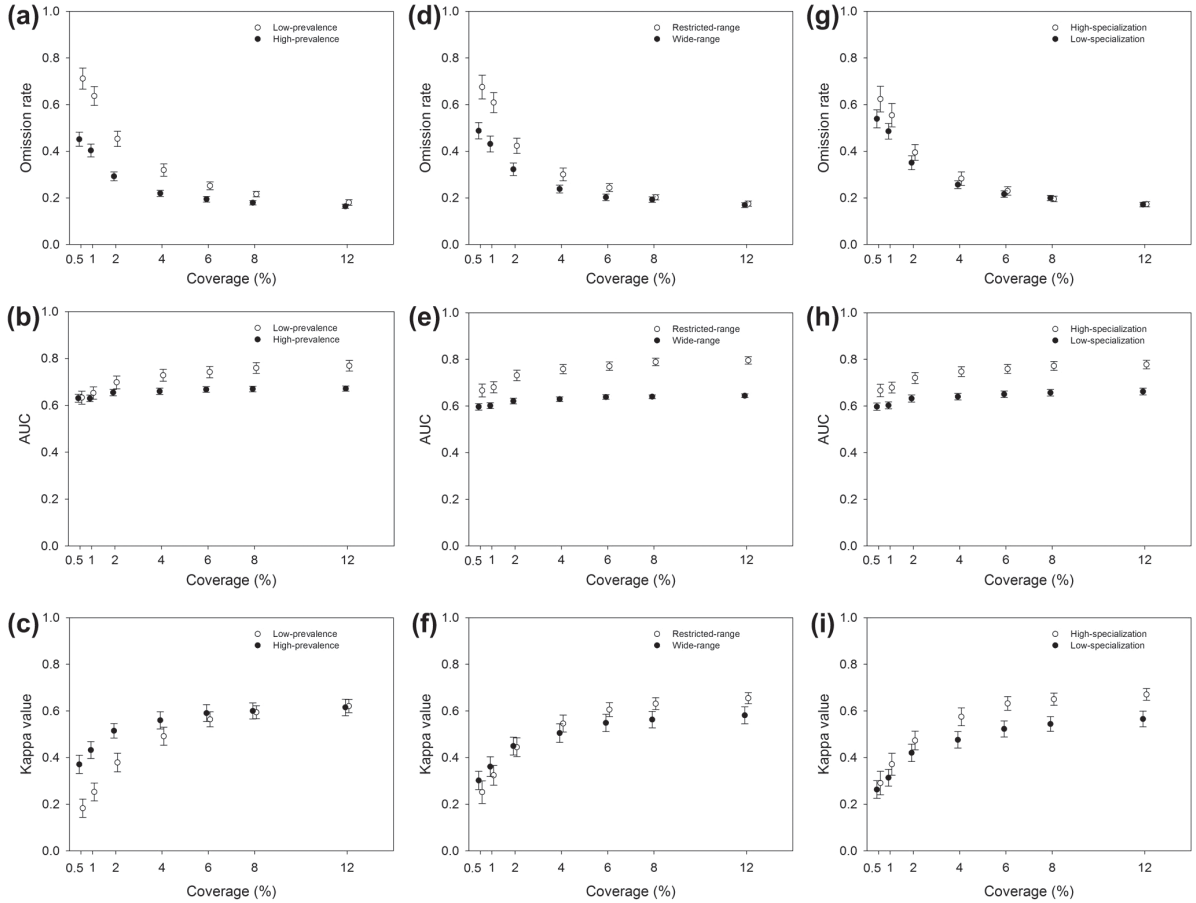

Figure 6. Average (± 95% confidence intervals) (a, d, g) omission rate, (b, e, h) AUC and (c, f, i) kappa values along the gradient of sampling coverage for high- and low-prevalence species, wide- and restricted-range species, and low- and high-specialization species.

Bardos (2006) and Jiménez-Valverde et al. (2009) have previously studied the influence of sample size on modelling performance and also showed that the effect of sample size becomes apparent only below a certain threshold, but their studies were conducted with virtual species and may only partly reflect monitoring data.

The prevalence of the species in the random subsets of secondary sampling units remained stable along the gradient of sample sizes and reflected the prevalence of the species in the whole set of sampling units. So, the link between modelling performance and sample size was independent of the proportion of sampling units with species presence records. Hence, we avoided the confusion between the effects of sample size and training sample prevalence (McPherson et al. 2004). In contrast, the extent to which the subsets of sampling units covered the range of environmental conditions used by the species was found to decrease with sample size and this contributes to explaining why the ability of the models to capture the environmental response of the species decreased markedly below a certain sample size. This issue underlines the importance of using a well-designed sampling procedure: the stratified random sampling that we implemented (see also Jiménez-Valverde et al. 2009) maximizes the chances to sample species distribution along the whole environmental gradient of the study area even when sample size decreases (Hortal and Lobo 2005, Thompson 2012). Below a certain sample size, the number of species presence records is, however, insufficient to cover the full range of environmental conditions used by the species and the modelling performance becomes less stable and much lower (see also Wintle and Bardos 2006, Wisz et al. 2008).

The minimum sample size required to ensure an acceptable level of modelling performance was strongly related to the prevalence of the species in the sampling units. On average, the minimum sample size was larger in low-prevalence species than in high-prevalence species. In contrast, the decrease in modelling performance with increasingly smaller sample size was found to be comparable in restricted- and wide-range species and in low- and high-specialization species. A large part of the among-species variance in the minimum sample size remained, however, unexplained and may be related to additional methodological issues or ecological processes. As imperfect detection of the species may confound the link between species distribution and environmental conditions, it is for instance warranted to analyse how detection rates may influence modelling performance. Rota et al. (2011) showed that using occupancy models to account for imperfect detection may contribute to improving modelling performance and relevance, especially in situations where detection probability varies along with environmental conditions (see also Kéry et al. 2010).

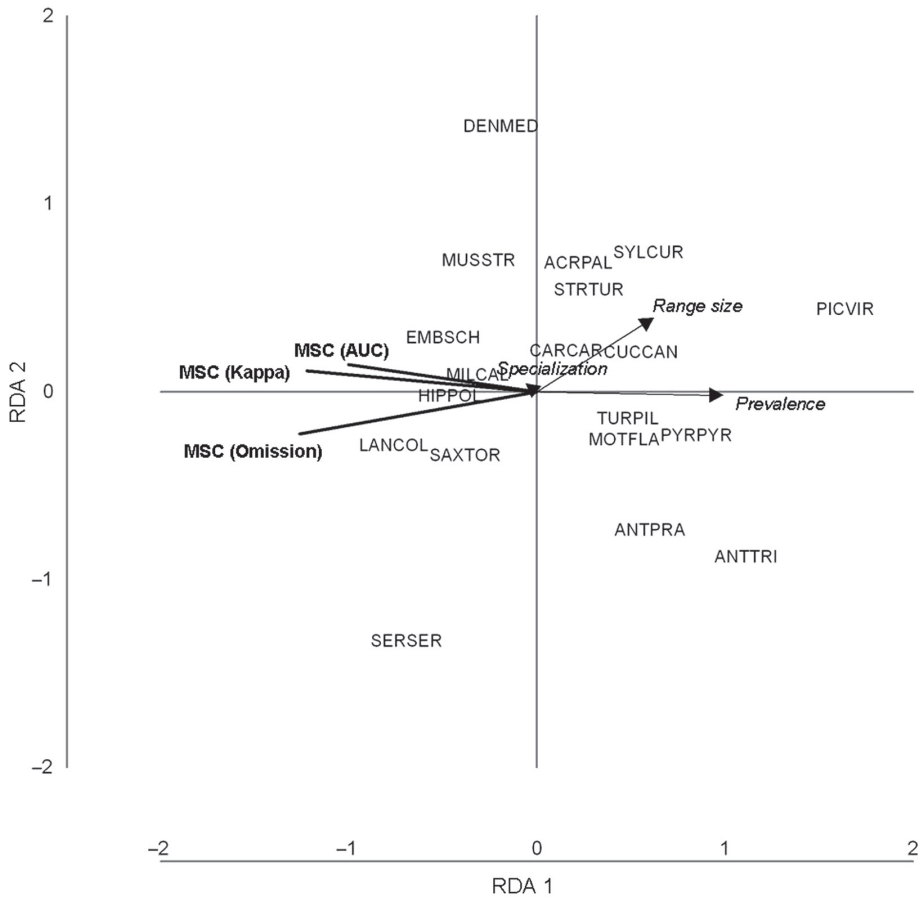

Figure 7. First two dimensions (RDA1 and RDA2) of the ordination space from the redundancy analysis (RDA, type-2 scaling). The explanatory variables (prevalence, range size and habitat specialization) are represented with arrows and the response variables (minimum sampling coverage according to omission rate, AUC and Kappa) are represented with bold black lines. Species are plotted using their code names (Table 1).

Such approaches are based on observation data collected during repeated surveys in the sampling units and are, therefore, only poorly suited to the context of the present study, as replicated observations are generally unavailable when setting the objectives of a monitoring project (but see Van Strien et al. 2013). Other ecological processes may have a direct or indirect influence on the minimum sample size. For instance, biotic interactions such as competition (with conspecific individuals or with other species) and predation may alter the location of the individuals in the landscape and shape the realized distribution of the species (Cadena and Loiselle 2007, Lima 2009). Although modelling tools become increasingly available to deal with this issue (Boulangeat et al. 2012, Wisz et al. 2013), such factors are probably beyond the scope of the analyses that could be done with the available atlas data.

We also have to stress the point that further work should include additional species because the set of species used in this study had to satisfy a number of criteria for the modelling exercise, which resulted in the use of a limited number of species that may only partly reflect the entire bird species assemblage. One of the most restrictive criteria was the availability of a sufficient amount of territory-mapping data to evaluate modelling performance. Such information was collected only for low- to moderate-abundance species in the atlas project. In order to increase the number of species in the analysis, a promising approach would be to use the increasingly available information on species distribution derived from web-based encoding systems for casual observation data (Sullivan et al. 2009).

When applying our innovative approach to the low- to moderate-abundance bird species in southern Belgium, a minimum sampling coverage of 4–5% (n = 664–830) was found to be needed in order to achieve an acceptable level of modelling performance for the majority of the studied species. Interestingly, Hoeting et al. (2000) and Wintle and Bardos (2006) obtained similar results with their simulated data reflecting plant and mammal distributions. However, the estimated minimum sampling coverage should probably not be considered as a rule of thumb. First, our results revealed considerable among-species variation in this minimum sample size. Second, the heterogeneity of the study area and the variables that are used to quantify the environmental conditions undoubtedly influence the number of sampling units needed to capture the link between species distribution and environmental conditions.

This application in southern Belgium illustrates that a substantial sampling coverage may be needed to derive

accurate species distribution models from long-term monitoring data. A sampling coverage of 4–5% of the study area is actually much higher than the coverage implemented in most of existing monitoring programmes worldwide. It may then become logistically difficult to find a trade-off between the number of sampling units and the number of repeated surveys in order for the same monitoring project to integrate in its objectives both the estimation of changes in occupancy or abundance and the mapping of species distribution. Interestingly, Hooten et al. (2009, 2012) used an optimal hybrid sampling design to combine different objectives in a single long-term monitoring project. In line with such an approach, a fixed subset of sampling units may be repeatedly surveyed within and between seasons (static design) to estimate species occupancy and detection probability, while a roving subset of sampling units may be surveyed less frequently over time (dynamic design) to increase spatial knowledge for distribution mapping. Both static and dynamic designs have advantages and disadvantages (MacKenzie and Royle 2005, Wikle and Royle 2005), but an appropriate allocation of sampling effort between fixed and roving units may contribute to combining several monitoring and mapping objectives (Hooten et al. 2009). In this context, our methodological approach constitutes a pilot analysis able to provide an initial estimate of the total number of sampling units needed when monitoring data are not yet available and to help putting the monitoring effort into place in order to reach one of the objectives. It is now important to consider additional optimisation criteria and to further integrate such approaches into a more general analytical framework to evaluate whether this initial sampling design will be suited to document species distribution dynamics and to estimate changes in the selected state variables or, alternatively, how the design has to be modified to better reach the multiple (and sometimes conflicting) objectives. In this respect, adaptive sampling design (Hooten et al. 2012) may prove a useful approach as it focuses on adjusting the sampling strategy on a regular basis as new information is gained in order to improve the cost-efficiency of the monitoring project.

Acknowledgements – We thank participants in the Breeding Bird Atlas of Wallonia project for fieldwork and Christophe Dehem, Marc De Sloover, Marc Fasol, Jean-Paul Jacob and Thierry Kinet for data or project management. Land use (COSW and IGN), land management (SIGEC) and soil (CNSW) maps were provided by the Direction Générale de l'Agriculture, des Ressources Naturelles et de l'Environnement (DGARNE) of the Service Public de Wallonie (SPW). The BBAW project was funded by the Service Public de Wallonie (SPW-DGO3). OA was funded by the National Research Fund, Luxembourg (FNR-AFR-PHD-08-63). NT and LB were funded through the EU BON project (contract no. 308454; FP7-ENV-2012, European Commission). This work has also been supported by the European infrastructure for biodiversity and ecosystem research (LifeWatch).

References

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Bean, W. T. et al. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. – Ecography 35: 250–258.

Boulangeat, I. et al. 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. – Ecol. Lett. 15: 584–593.

Braunisch, V. and Suchant, R. 2010. Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. – Ecography 33: 826–840.

Brotons, L. et al. 2004. Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. – Ecography 27: 437–448.

Brotons, L. et al. 2006. Spatial modeling of large-scale bird monitoring data: towards pan-European quantitative distribution maps. – J. Ornithol. 147: 29.

Brotons, L. et al. 2007. Updating bird species distribution at large spatial scales: applications of habitat modelling to data from long-term monitoring programs. – Divers. Distrib. 13: 276–288.

Cadena, C. D. and Loiselle, B. A. 2007. Limits to elevational distributions in two species of emberizine finches: disentangling the role of interspecific competition, autoecology, and geographic variation in the environment. – Ecography 30: 491–504.

Carden, R. et al. 2010. Irish bat monitoring schmes: ATLAS Republic of Ireland. Report for 2008–2009. – Bat Conservation Ireland.

De Cáceres, M. and Brotons, L. 2012. Calibration of hybrid species distribution models: the value of general-purpose vs. targeted monitoring data. – Divers. Distrib. 18: 977–982.

Dunn, A. M. and Weston, M. A. 2008. A review of terrestrial bird atlases of the world and their application. – Emu 108: 42–67.

Elith, J. et al. 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330–342.

Estrada, J. et al. 2004. Atles dels ocells nidificants de Catalunya (1999–2002). – Lynx Editions.

Feeley, K. J. and Silman, M. R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. – Divers. Distrib. 17: 1132–1140.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Gitzen, R. A. et al. 2012. Design and analysis of long-term ecological monitoring studies. – Cambridge Univ. Press.

Guillera-Arroita, G. and Lahoz-Monfort, J. J. 2012. Designing studies to detect differences in species occupancy: power analysis under imperfect detection. – Methods Ecol. Evol. 3: 860–869.

Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – Ecol. Lett. 16: 1424–1435.

Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Hirzel, A. H. and Le Lay, G. 2008. Habitat suitability modelling and niche theory. – J. Appl. Ecol. 45: 1372–1381.

Hoeting, J. A. et al. 2000. An improved model for spatially correlated binary responses. – J. Agric. Biol. Environ. Stat. 5: 102–114.

Hooten, M. B. et al. 2009. Optimal spatio-temporal hybrid sampling designs for ecological monitoring. – J. Veg. Sci. 20: 639–649.

Hooten, M. B. et al. 2012. Optimal spatio-temporal monitoring designs for characterizing population trends. – In: Gitzen, R. A. et al. (eds), Design and analysis of long-term ecological monitoring studies. Cambridge Univ. Press, pp. 443–459.

Hortal, J. and Lobo, J. M. 2005. An ED-based protocol for optimal sampling of biodiversity. – Biodivers. Conserv. 14: 2913–2947.

Jacob, J. P. et al. 2010. Atlas des oiseaux nicheurs de Wallonie. – Aves & Dépt de l'Etude du Milieu Naturel et Agricole (Service

Public de Wallonie – Direction générale de l'Agriculture, des Ressources naturelles et de l'Environnement).
Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. – Global Ecol. Biogeogr. 21: 498–507.
Jiménez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Community Ecol. 10: 196–205.
Julliard, R. et al. 2006. Spatial segregation of specialists and generalists in bird communities. – Ecol. Lett. 9: 1237–1244.
Kéry, M. et al. 2009. Trend estimation in populations with imperfect detection. – J. Appl. Ecol. 46: 1163–1172.
Kéry, M. et al. 2010. Predicting species distributions from checklist data using site-occupancy models. – J. Biogeogr. 37: 1851–1862.
Kéry, M. et al. 2013. Analysing and mapping species range dynamics using occupancy models. – J. Biogeogr. 40: 1463–1474.
Legendre, P. and Legendre, L. 2012. Numerical ecology. – Elsevier.
Lima, S. L. 2009. Predators and the breeding bird: behavioral and reproductive flexibility under the risk of predation. – Biol. Rev. 84: 485–513.
Lindenmayer, D. B. et al. 2012. Improving biodiversity monitoring. – Austral Ecol. 37: 285–294.
Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.
MacKenzie, D. I. 2012. Study design and analysis options for demographic and species occurrence dynamics. – In: Gitzen, R. A. et al. (eds), Design and analysis of long-term ecological monitoring studies. Cambridge Univ. Press, pp. 397–425.
MacKenzie, D. I. and Royle, J. A. 2005. Designing occupancy studies: general advice and allocating survey effort. – J. Appl. Ecol. 42: 1105–1114.
MacKenzie, D. I. et al. 2005. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. – Elsevier.
Maes, D. et al. 2012. Applying IUCN Red List criteria at a small regional level: a test case with butterflies in Flanders (north Belgium). – Biol. Conserv. 145: 258–266.
Maes, D. et al. 2013. Dagvlinders in Vlaanderen: nieuwe kennis voor betere actie. – Uitgeverij Lannoo nv.
Martin, Y. et al. 2013. Testing instead of assuming the importance of land use change scenarios to model species distributions under climate change. – Global Ecol. Biogeogr. 22: 1204–1216.
McPherson, J. M. et al. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? – J. Appl. Ecol. 41: 811–823.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Rhodes, J. R. and Jonzén, N. 2011. Monitoring temporal trends in spatially structured populations: how should sampling effort be allocated between space and time. – Ecography 34: 1040–1048.
Rodhouse, T. J. et al. 2012. Assessing the status and trend of bat populations across broad geographic regions with dynamic distribution models. – Ecol. Appl. 22: 1098–1113.
Rodríguez, J. P. et al. 2007. The application of predictive modelling of species distribution to biodiversity conservation. – Divers. Distrib. 13: 243–251.
Rota, C. T. et al. 2011. Does accounting for imperfect detection improve species distribution models? – Ecography 34: 659–670.
Royle, J. A. and Nichols, J. D. 2003. Estimating abundance from repeated presence–absence data or point counts. – Ecology 84: 777–790.
Sullivan, B. L. et al. 2009. eBird: a citizen-based bird observation network in the biological sciences. – Biol. Conserv. 142: 2282–2292.
Thompson, S. K. 2012. Sampling. – Wiley.
Timothy, J. and Sharrock, R. 1974. Minutes of the second meeting of the European Ornithological Committee. – Acta Ornithol. 14: 404–411.
Titeux, N. et al. 2007. Fitness-related parameters improve presence-only distribution modelling for conservation practice: the case of the red-backed shrike. – Biol. Conserv. 138: 207–223.
Van Strien, A. J. et al. 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. – J. Appl. Ecol. 50: 1450–1458.
Van Swaay, C. A. M. et al. 2008. Butterfly monitoring in Europe: methods, applications and perspectives. – Biodivers. Conserv. 17: 3455–3469.
Vorísek, P. et al. 2008. A best practice guide for wild bird monitoring schemes. – CSO/RSPB.
Wikle, C. K. and Royle, J. A. 2005. Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. – Environmetrics 16: 507–522.
Wintle, B. A. and Bardos, D. C. 2006. Modeling species–habitat relationships with spatially autocorrelated observation data. – Ecol. Appl. 16: 1945–1958.
Wisz, M. S. et al. 2008. Effects of sample size on the performance of species distribution models. – Divers. Distrib. 14: 763–773.
Wisz, M. S. et al. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. – Biol. Rev. 88: 15–30.

Supplementary material (Appendix ECOG-00749 at <www.ecography.org/readers/appendix>). Appendix 1.