ECOGRAPHY 29: 835–844, 2006

Importance of spatial autocorrelation in modeling bird distributions at a continental scale

Volker Bahn, Raymond J. O'Connor[†] and William B. Krohn

Bahn, V., O'Connor, R. J. and Krohn, W. B. 2006. Importance of spatial autocorrelation in modeling bird distributions at a continental scale. – Ecography 29: 835–844.

Spatial autocorrelation in species' distributions has been recognized as inflating the probability of a type I error in hypotheses tests, causing biases in variable selection, and violating the assumption of independence of error terms in models such as correlation or regression. However, it remains unclear whether these problems occur at all spatial resolutions and extents, and under which conditions spatially explicit modeling techniques are superior. Our goal was to determine whether spatial models were superior at large extents and across many different species. In addition, we investigated the importance of purely spatial effects in distribution patterns relative to the variation that could be explained through environmental conditions. We studied distribution patterns of 108 bird species in the conterminous United States using ten years of data from the Breeding Bird Survey. We compared the performance of spatially explicit regression models with non-spatial regression models using Akaike's information criterion. In addition, we partitioned the variance in species distributions into an environmental, a pure spatial and a shared component. The spatially-explicit conditional autoregressive regression models strongly outperformed the ordinary least squares regression models. In addition, partialling out the spatial component underlying the species' distributions showed that an average of 17% of the explained variation could be attributed to purely spatial effects independent of the spatial autocorrelation induced by the underlying environmental variables. We concluded that location in the range and neighborhood play an important role in the distribution of species. Spatially explicit models are expected to yield better predictions especially for mobile species such as birds, even in coarse-grained models with a large extent.

V. Bahn (volker.bahn@gmx.net) and R. J. O'Connor, Dept of Wildlife Ecology, Univ. of Maine, Orono, ME 04469-5755, USA (present address of V.B.: Dept of Biology, McGill Univ., Stewart Biol. Bldg., 1205 avenue Docteur Penfield, Montreal, QC H3A 1B1, Canada). – W. B. Krohn, U.S. Geological Survey, Maine Cooperative Fish and Wildlife Research Unit, Orono, ME 04469-5755, USA.

Documenting and understanding the distributions of organisms in space and time are central to the fields of biogeography, ecology, and conservation biology. Ecology has been defined as the study of the distribution and abundance of organisms (Andrewartha and Birch 1954, 1984, Krebs 1972). In conservation biology, knowledge of the actual or potential distribution of a species is indispensable for threatened and endangered species management and protected area planning (Scott and Csuti 1997). However, at most times the actual locations of individual organisms are unknown. The discipline of distribution modeling strives to fill this void by making probabilistic statements about the geographic distribution of species (Scott et al. 2002).

Accepted 19 September 2006
[†]deceased

ISSN 0906-7590

Distribution models that do not include spatial location explicitly assume that species' locations are independent in space and time. Such an assumption could be violated if a) the conditions defining the niche were autocorrelated; or b) species' locations were connected through dispersal or other behaviors that lead to spatial patterning such as aggregation or regular spacing. Lichstein et al. (2002) termed the former cause of spatial dependence exogenous and the latter endogenous.

Concerning endogenous sources of spatial patterning, species generally exhibit some dispersal, be it as seeds, juveniles or adults. Such dispersal events connect populations in space and time and have the potential to create dependence at varying spatial and temporal scales (Keitt et al. 2002, Lichstein et al. 2002). Standard habitat models do not account for population dynamics based on dispersal, such as source-sink populations and metapopulations (Pulliam 1988, Dias 1996, Hanski 1998). Consequently, a standard model will assign the same probability of occupancy to two habitat patches A and B, with similar physical characteristics, even if patch A is surrounded by excellent habitat and patch B is completely isolated from other suitable habitat. An expected consequence of dispersal among habitat patches is that the average similarity among population densities in patches decays with distance – one manifestation of spatial autocorrelation (Selmi and Boulinier 2001, Trenham et al. 2001, Keitt et al. 2002, Schiegg 2003).

In addition, environmental conditions underlying a species' niche are dependent in space and time, which exogenously causes spatio-temporal dependence or autocorrelation in species' distributions (Legendre 1993). Autocorrelation in abiotic and biotic resources has been observed for a long time, resulting in Tobler formulating the first law of geography as: "... everything is related to everything else, but near things are more related than distant things." (Tobler 1970: 236). In following these resources, species' distributions are also spatially autocorrelated (Legendre 1993, Lichstein et al. 2002).

If only exogenous autocorrelation was present in a species' distribution, the inclusion of all environmental determinants would suffice to create a valid model because they would implicitly carry all necessary spatial information (Diniz-Filho et al. 2003). In other words, if all autocorrelation in the distribution of a species is caused by autocorrelation in the distribution of the important resources and conditions (i.e. exogenous), inclusion of these conditions and resources as variables will lead to a complete model and not miss spatial information and relationships. If, however, endogenous autocorrelation is present – for example, due to dispersal, conspecific attraction or other behaviors leading to spatial patterning – the inclusion of all relevant environmental and resource determinants will not eliminate autocorrelation from residuals of the model and will lead to biases in variance and coefficient estimates, as well as model selection (Lennon 2000, Keitt et al. 2002).

The question remains, however, whether endogenous and exogenous autocorrelation in species distributions are of practical consequence to distribution modeling. This question is dependent on the temporal and spatial scale of the investigation and different authors have come to different conclusions. Typically, researchers working at small to medium extents and fine resolution found the explicit inclusion of spatial information beneficial or even crucial to their distribution models (Augustin et al. 1996, Selmi and Boulinier 2001, Keitt et al. 2002, Lichstein et al. 2002). At larger extents (in the order of hundreds of kilometers) Diniz-Filho et al. (2003) found the inclusion of environmental variables to be sufficient to eliminate autocorrelation in the residuals of a model for species richness of birds (however, see Hawkins et al. 2003). Similarly, Koenig (1998) found little evidence for spatio-temporal autocorrelation in the distribution of Californian landbirds.

The study presented here differs from the above studies because it a) covers a large extent (the conterminous United States); but b) concerns the distributions of single species (in contrast to Diniz-Filho et al. 2003 who modeled species richness); and c) deals with spatial autocorrelation only (in contrast to Koenig 1998 who investigated spatio-temporal autocorrelation).

At small extents, the connection among populations creating spatial autocorrelation in population sizes seems to be well established (Augustin et al. 1996, Thomson et al. 1996, Lichstein et al. 2002, Peakall et al. 2003). In contrast, at large extents, most of the ecological mechanisms suggested for pattern formation at small extents are not applicable (e.g. conspecific attraction, colonialism, short distance dispersal), leaving only long distance movements in the widest sense as a plausible mechanism.

In this paper, we investigate whether spatial effects are relevant to bird distribution modeling at a coarse, national scale. The goal was to determine whether individual species distributions show spatial patterns above and beyond what can be explained by spatial patterns in environmental conditions and if so, how strong these spatial patterns are in comparison to the explanatory power held by environmental conditions.

Methods

We compared standard distribution models, based on environmental and climate variables only, to spatially explicit models that also included spatial position and neighborhood relationships.

We used data from the Breeding Bird Survey (BBS) for the conterminous USA from 1981 to 1990. See Robbins et al. (1986), Sauer et al. (1994), and O'Connor et al. (1996) for detailed methods and discussion of the BBS. Bird data for individual species were summarized as presence/absence over complete routes each year and then expressed as incidences (i.e. proportion of years in which the species was present) over the ten years for each route. Using incidence instead of abundance has the advantage of being less sensitive to detection probabilities while being more closely related to abundance than are presence/absence data (Wright 1991, Hanski 1992). We transformed incidence values with an arcsine transformation (Freeman and Tukey 1950, Zar 1996: 283) to move toward normal distributions.

Only the 1189 BBS routes with the highest quality standard and at least seven years of data were included in the analysis. The starting points of routes were mapped to Environmental Monitoring and Assessment Program (EMAP) hexagons (White et al. 1992), which are 620 km^2 in size and ca 27 km apart from center to center.

We selected 108 species of breeding birds for the analysis (complete list in Appendix). Criteria for the selection were good coverage over the conterminous USA (>150 occupied routes) and sensitivity to coarse-scale predictors covered in our dataset ($R^2 > 0.5$ in initial regression tree models). Reasons for exclusion were extreme range shapes, such as long and narrow ranges along the border of the study area, or the extremely patchy distributions. Such distributions prevent meaningful spatial modeling.

We used bird ranges from NatureServe (Ridgely et al. 2003) to determine the study area and thus the routes to be included in the models for each of the selected species. This step was necessary because including the whole study area and all 1189 routes for all species would have meant that the many routes with zero incidences in each species would have dominated the models. Such models would have mostly modeled presence absence over the study area and not patterns of abundances within the study area. In addition, heavily skewed, zero-inflated distributions of incidence values would have led to violations in the assumptions of regression analyses and problems with the trend surfaces in the spatial models. However, because the NatureServe ranges were conservative and often excluded occupied BBS routes, we buffered all ranges by 150 km, a distance that proved to include almost all occupied BBS routes.

Our independent variables stemmed from research by O'Connor et al. (1996). They comprised 159 land cover variables from Loveland et al. (1991) derived from remotely sensed Advanced Very High Resolution Radiometer (AVHRR) and ancillary data including elevation, climate, ecoregions, and land resource areas. O'Connor et al. (1996) added a land cover type "urban." Additional variables were various measures of spatial configuration of habitat patches in hexagons (e.g. fractal dimension), climatic variables including seasonal temperatures and rainfall, and elevations from a digital elevation model. For more details on the land cover variables see O'Connor et al. (1996, 1999).

In total, there were 207 independent variables, 160 variables summarizing land cover information, twelve climate variables (January and July temperatures, precipitation, and derived variables such as seasonality), four variables from digital elevation models, and thirty-one other variables characterizing the land cover in terms of spatial configuration and fragmentation indices. Many of the land covers had a localized distribution (i.e. they did not occur at most locations) and the average number of effectively available variables at a single location was thus much smaller than 207.

The first step in the modeling process was the generation of regression tree (RT) models, which we used as a robust method for variable selection (Breiman 1984, Walker and Cocks 1991, De'ath and Fabricius 2000, Austin 2002). These models were also used to eliminate species whose environmental determinants were not well captured at a coarse scale. RT models were built with the library RPART (Therneau and Atkinson 1997), and were pruned to final size using the one standard error rule after twenty five-fold cross-validation (Breiman 1984).

Next the selected variables were included as third degree polynomials in ordinary least squares (OLS) regression models to allow for curvilinearity, which is modeled implicitly by the RT models. To eliminate redundant variables and/or their polynomials, we used backwards step-wise model selection by Akaike's information criterion (AIC). Polynomial terms of lower order were kept in the model with retained higher terms of the same variable even if they had not been selected by AIC.

We built three types of OLS regression models. The first type contained only environmental variables (ENV), the second type contained only the geographic coordinates of the hexagons (up to third degree polynomial form) as a trend surface (TREND), and the third combined the first two sets of variables in one model (ENV.TREND). The last type was the model we used for the secondary variable selection by AIC as described above, while the former were hierarchical subsets of the latter.

Finally, we used conditional autoregressive regressions (CAR) for spatially explicit modeling (Cressie 1993, Lichstein et al. 2002). CAR models include information on the residuals of neighboring locations (eq. 1) and are solved iteratively. Thus they capture fine-scale spatial autocorrelation, which is missed by the trend surface models. We determined the neighborhood by calculating eight models with neighborhood sizes between 50 and 400 km in 50 km steps, and selecting the neighborhood size leading to the model with the highest maximum

Table 1. Statistical models used in the study. Trend surface variables were the geographical coordinates of a site included as an up to 3rd degree polynomial. The neighborhood consisted of all neighbors within a certain distance (depending on the model), weighted by distance. The three regression types were regression trees (Tree), ordinary least squares regression (OLS), and conditional autoregressive regression (Spatial).

Model	Regression type	Variables included		
		Environment	Trend surface	Neighborhood
RT	Tree	Y	N	N
ENV	OLS	Y	N	N
TREND	OLS	N	Y	N
ENV.TREND	OLS	Y	Y	N
CAR.FULL	Spatial	Y	Y	Y
CAR.TREND	Spatial	N	Y	Y

likelihood estimate for each species. The influence of neighbors was inversely distance weighted with a spherical model (Kaluzny et al. 1996, Legendre and Legendre 1998). We used a spherical model for the distance weighting because this model consistently provided the best fit in preliminary variograms of bird data. We created CAR models with variable selections identical to the TREND model (CAR.TREND–only coordinates included) and ENV.TREND model (CAR.FULL–coordinates and environmental variables included). Table 1 shows an overview of all models used in this study. CAR models follow the equation:

$$Y = X\beta + \rho C(Y - X\beta) + \varepsilon \quad (1)$$

Where Y is the vector of dependent variables; X is a matrix of independent variables; C is a symmetric neighborhood matrix; β and ρ are coefficients; and ε is a vector of errors with the covariance matrix $\sigma^2(I - \rho C)^{-1}$, where I is the identity matrix.

The log-likelihood of all regression models was calculated in S-PLUS through a likelihood ratio test, in which model components can be set to zero. The non-spatial models had the coefficient ρ (rho) in front of the neighborhood element set to zero. Akaike's information criterion (Akaike 1981) was calculated from these log-likelihoods and from the number of parameters included in the models (including the intercept and the spatial coefficient ρ where appropriate). The proportion of variance explained by the model (R^2) was calculated from log-likelihoods according to the formula given by Nagelkerke (1991). Residuals from all (regular and spatial) regression models were visually inspected for deviations from a Gaussian distribution.

We partialled out the variation that could be ascribed to the environment, space and the "spatial component of the environmental influence" according to the method described in Borcard et al. (1992) and Legendre and Legendre (1998). Our method deviated from theirs in so far as our spatial component was not restricted to the coarse-scale effects captured by a trend surface, but also included the fine-scale effects captured by the neighborhood matrix in the CAR models. This method relies on differences in R^2s of selected models to allow inferences about partitions of the explained variation that cannot be directly modeled. It is not a fitted model itself, but rather relies on the results from the models described above and partitions the R^2s into new categories. The CAR.FULL model contains all elements used for modeling here (environmental variables, trend surface, and neighborhood matrix) and thus has the highest proportion of explained variance. The "error" partition is calculated as $1-R^2$ (CAR.FULL). Subtracting the R^2 of the ENV model from the R^2 of the CAR.FULL model gives the pure spatial partition of the explained variance ("Space") because the ENV model does not contain any explicit spatial information. Subtracting the R^2 of the CAR.TREND model from the R^2 of the CAR.FULL model gives the pure environmental partition of the explained variance ("Environment") because the CAR.TREND model does not contain any explicit environmental information. Subtracting the thus gained "Space" partition and "Environment" partition from the R^2 of the CAR.FULL model gives the "Shared" partition. This partition is neither clearly assignable to spatial effects, nor to environmental control. Instead it is thought to be some form of interaction between the two.

All statistical analyses were done in S-PLUS 6.2 (Anon. 2003) with the additional module SPATIAL (Kaluzny et al. 1996) and the add-on libraries RPART and MASS (Venables and Ripley 2002) (use of this product does not imply endorsement). The number given with $\pm$after statistics is the standard error (SE), unless noted otherwise.

Results

Table 2 shows a comparison among regression trees (RTs), ordinary least squares (OLS) regression models, and spatial regression models. See Table 1 for a description of the models. The sample size of included routes differed among species depending on how many routes fell within a species' range. The average sample size was 717 ± 28.7 routes per species (range: 161–1189).

Table 2. Comparison among regression trees (RT), ordinary least squares regression models (OLS) and spatial regression models. The values shown are averages of 108 individual models ± standard errors.

Model	R^2	Variables	Parameters*	AIC**
Regression tree	0.56 ± 0.013	4.9 ± 0.30	12.3 ± 0.74	n/a
Regular regression model				
ENV	0.50 ± 0.014	4.6 ± 0.29	14.3 ± 0.68	158.8 ± 9.91
Regular regression models with trend surface				
TREND	0.40 ± 0.017	1.9 ± 0.03	7.1 ± 0.11	261.8 ± 14.09
ENV.TREND	0.56 ± 0.013	6.5 ± 0.29	19.4 ± 0.70	69 ± 5.09
Spatial regression models				
CAR.FULL	0.60 ± 0.012	6.5 ± 0.29	20.4 ± 0.70	0 ± 0
CAR.TREND	0.49 ± 0.015	1.9 ± 0.03	8.1 ± 0.11	139.4 ± 8.27

* The number of parameters is the number of splits for RTs and the number of coefficients including the intercept for OLS regression models.
** Akaike's information criterion (AIC) scaled to the lowest value, which always was the CAR.FULL model. AIC cannot be calculated for RTs because they are not a likelihood based method.

The environmental variables passed on from the RT models were mostly retained during the AIC stepwise selection in the regression models. While 86 models retained all variables selected by the RT models, only 17 dropped one, four dropped two, and one dropped three variables. In addition, the median number of splits in the RT models (11, IQR: 7–15) was only slightly different from the median number of parameters (13, IQR: 8–18) in the environmental models (ENV), which was surprising given the very different structure of the models. However, the average R^2 of the RT models was 0.065 ± 0.007 higher than the average R^2 of the ENV models (0.498 ± 0.014). The average R^2 of the RT models (0.563 ± 0.013) was almost identical to the R^2 of the ENV.TREND models (0.562 ± 0.013), which contained a trend surface based on geographic coordinates in addition to the environmental variables. Note, however, that the variable selection was optimized by the RTs, which had the full set of independent variables to their disposal, while the OLS regression models only had the pre-selected variables to choose from.

The full spatial regression models (CAR.FULL) were a considerable improvement over the standard regression models, including those with environmental predictors only (ENV). The CAR.FULL models had an on average 0.102 ± 0.004 higher R^2, which is a 25.5 ± 0.02% improvement over the ENV models. However, the CAR.FULL models contained more parameters than the ENV models (median: 6, IQR: 6–7). The more meaningful statistic for comparing the goodness of fit between the two kinds of models is Akaike's Information Criterion (AIC), which penalizes for the number of parameters fitted in the model. The AIC values of the spatially explicit CAR.FULL models were, on average, 158.8 ± 9.9 points lower than those of the ENV models. According to the rule of thumb suggested by Burnham and Anderson (2002: 70), when comparing models, a difference in AIC of 2 or less lends substantial support to the competing model, a difference of 4–7 considerably less support and a difference >10 essentially lends no support to the inferior model.

The fully spatial CAR.FULL models also improved upon the regression models that contained a trend surface but not a neighborhood matrix (ENV.TREND). The R^2 of CAR.FULL models was on average 0.038 ± 0.003 higher than that of the ENV.TREND models, and the CAR.FULL models' AIC was on average 69.0 ± 5.1 points lower than that of the ENV.TREND models.

At the large extent of the study, using a niche based approach with environmental variables only (ENV) did not have more explanatory power than using pure spatial interpolation (CAR.TREND). Ecologically this means that the spatial position in the range and the incidences at neighboring locations are as important to the incidence value of a population as the local environmental conditions. The difference between ENV models and CAR.TREND models in R^2 was only 0.006 ± 0.009. The CAR.TREND models did not contain any environmental predictors and thus were pure spatial interpolations with fewer variables (the geographical coordinates) and parameters than the ENV models. The lower number of parameters led to a considerably lower average AIC value for the CAR.TREND models (AIC ENV–AIC CAR.TREND: 19.4 ± 11.1).

The average maximum neighborhood distance selected as giving the best model out of the 8 tested distances was 195.8 ± 7.2 km for the CAR.FULL models and 244.4 ± 8.9 km for the CAR.TREND models, which did not contain environmental variables. The increase in distance from CAR.FULL to CAR.TREND models could be explained by the spatial information carried implicitly in the environmental variables, which accounted for a part of the spatial autocorrelation in the CAR.FULL models but not the CAR.TREND models.

Table 3 shows the results of applying Borcard et al.'s (1992) partitioning sources of variation to the data. This technique yielded estimates of the proportions of variance associated with a non-spatial effect of the

Table 3. Partitioning of sources of variation in bird distributions according to Borcard et al.'s (1992) method. The four parts describe respectively the variation attributed to a purely local environmental effect (Environment), the spatial patterning in the dependent variable (Space), the spatial component of the environmental influence (Shared), and the unexplained variation or error in the model (Error). In addition to Borcard et al.'s (1992) original method, also shown is a partitioning based on fully spatial models (CAR.FULL), which captured fine-scale neighborhood effects in addition to the coarse-scale spatial effects captured in the original model's trend surfaces.

Approach	Environment	Space	Shared	Error
Trend surface	0.160 ± 0.009	0.064 ± 0.004	0.337 ± 0.017	0.438 ± 0.013
Full spatial	0.109 ± 0.006	0.102 ± 0.004	0.389 ± 0.016	0.400 ± 0.012
Difference	0.052 ± 0.003	-0.038 ± 0.003	-0.052 ± 0.003	0.038 ± 0.003
P-value*	<0.0001	<0.0001	<0.0001	<0.0001

* Paired Student's t-test with 107 degrees of freedom.

environmental variables (Environment), with a purely spatial patterning in the dependent variable (Space), and with the spatial component of the environmental influence (Shared) due to spatial patterning in the environmental variables. In contrast to Borcard et al.'s (1992) original technique, which used trend surfaces only, our results were based on a fully spatial CAR model, which captured both coarse- and fine-scale spatial patterns.

The purely environmental partition and the purely spatial partition were of similar size (Table 3: Environment and Space), 18 and 17% of the total explained variance, respectively. Approximately 65% of the explained variation in species' distributions, the largest part, however, was attributable to the spatial component of the environmental influence (Table 3: Shared).

Using a CAR model to determine spatial effects resulted in clear shifts in the three partitions from Borcard et al.'s original method, which uses trend surfaces only (Table 3). The partitions containing spatial elements increased (Space and Shared) at the cost of the size of the environmental part (Environment) and the unexplained part (Error). The increase in R^2 from the ENV.TREND model to the CAR.FULL model, which is identical to the increase of the Space partition from Borcard's original method to our derivation (0.038 ± 0.003) could be interpreted as the part of the pure spatial effect that is due to neighborhood effects, rather than location in the range. Such an interpretation would leave the remaining 0.064 ± 0.004 in the Space partition to variance explained by the location in the range.

Discussion

Spatial autocorrelation has positive and negative consequences for the modeling of species' distributions. Most authors focus on the negative consequences, such as the dependence among samples that decays with distance under positive spatial autocorrelation (Student 1914, Legendre and Fortin 1989, Legendre 1993, Dale and Fortin 2002, Keitt et al. 2002). Standard statistical models employed in distribution modeling, such as correlations and regressions, work under the assumption of independence in the residuals. Autocorrelated data violate this assumption and lead to inflated estimates in degrees of freedom, which lead to underestimates of variance and overestimation of the significance of effects (Student 1914, Legendre and Fortin 1989, Legendre 1993, Dale and Fortin 2002, Keitt et al. 2002).

However, spatially explicit models exist that can incorporate spatial autocorrelation with a low cost in terms of increased complexity. These models include truncated neighborhood matrices (Borcard and Legendre 2002), kriging (Legendre and Fortin 1989, van Horssen et al. 2002), autoregressive models (Augustin et al. 1996, Keitt et al. 2002, Lichstein et al. 2002), modified correlograms (Koenig and Knops 1998), and CART models with spatial dependence (Miller and Franklin 2002). The benefit of including autocorrelation in a model is not only that the statistical assumptions are better met, but also that the predictive power of a model is improved by incorporating additional information or predictors, such as the values at neighboring locations (Costanza and Ruth 2001). In many geostatistical applications, such as kriging, neighborhood information is the only predictor in the model, which equates to an elaborate form of spatial interpolation. In our study, models based exclusively on spatial trend and neighborhood information even performed marginally better than the standard models that included only environmental variables.

Out of the many techniques for spatial modeling listed above, we selected conditional autoregressive regressions (CAR) as the most appropriate technique for the situation. Kriging models are generally designed for spatial interpolation only and allow the inclusion of independent variables only in a very limited way (co-kriging). Using truncated neighborhood matrices to generate a large number of independent spatial variables for inclusion in a regular regression may have lead to very similar results. However, the method seemed less transparent and established to us compared to CAR models. Including a "contagion" in a regular regression (Augustin et al. 1996, Araujo and Williams 2000) or CART model (Miller and Franklin 2002) is similar to the inclusion of a neighborhood in a CAR model: a

(possible distance weighted) average of values (or residuals) at neighboring locations is included as independent variable. The difference is that CAR models solve the problem of mutual influence of neighbors iteratively, until a stable solution is found, while simply including a contagion in a regular regression only considers the influence of neighbors before they were themselves allowed to be influenced by neighbors. We (and Cressie (1993: 433)) found the iterative solution approach to be more appropriate.

Spatial models may also improve variable selection (Ellner and Seifu 2002, Keitt et al. 2002). Non-spatial models cannot account for autocorrelation and thus may incorrectly select variables purely because they have a similar autocorrelation as the dependent variable and not because they are good predictors (Lennon 2000, Ellner and Seifu 2002, Keitt et al. 2002).

While autocorrelation at fine scales has been well documented (Legendre 1993, Thomson et al. 1996, Lichstein et al. 2002), research at a coarse scale (hundreds to thousands of km) is still rare. We were able to demonstrate that spatial autocorrelation in bird distributions at such a coarse scale is important, and that spatial models are much better at handling spatially dependent data than are standard habitat-based regression models.

In contrast to the research on bird species richness by Diniz-Filho et al. (2003), we found strong spatial effects in individual bird species' distributions that did not disappear with the inclusion of environmental variables. The difference in results may be explained by the difference in dependent variables. As a compound measurement, species richness may smooth over spatial autocorrelation in individual species' distributions caused by endogenous mechanisms such as dispersal, leaving only environmental autocorrelation. Other reasons for the success of Diniz-Filho et al. (2003) in removing autocorrelation from residuals by including environmental variables could be attributed to the specific characteristics of the Palaearctic region, because Hawkins et al. (2003) found autocorrelation in residuals in all other five ecozones with similar methods.

Similarly, our finding of spatial autocorrelation effects differs from Koenig's (1998) findings from research on California birds. He only found spatio-temporal autocorrelation, or synchrony, in one out of eighty eight investigated species. However, finding synchrony in species distributions is much more demanding than finding spatial autocorrelation only. Tests of synchrony use individual observations of populations in space and time, and must accommodate variance in time as well as in space. In contrast, our focus on spatial autocorrelation in data that were averaged over ten years to minimize the effects of temporal variability gave more limited but clearer results.

The power of advanced spatial modeling techniques was further demonstrated using the invaluable method of partialling out sources of variance, pioneered by Borcard et al. (1992). They used trend surface models to capture spatial patterns, which included third degree or less polynomial geographic coordinates as variables in a regression model. However, with a sensible highest polynomial inclusion of the coordinates in the third degree, trend surfaces can capture only long-wave spatial patterns and cannot account for short-range autocorrelation (Meot et al. 1998). Incorporating relatively fine-scale neighborhood effects (here at the scale of tens to hundreds of km) on top of the coarse trend surface approaches resulted in a shift in the distribution of variance across the three partitions. While the purely environmental, niche-based factors experienced a relative loss in explanatory power, the purely spatial and shared spatial/environmental partitions gained in importance. This underscores the importance of neighborhood effects in bird distributions. Borcard and Legendre (2002) found similar shifts in partitions with an improved spatial approach to their own method. However, their method is more complicated than ours and was demonstrated only for one dimension (along a transect) in their paper.

On average, purely environmental effects and purely spatial effects each accounted for ca 18% of the explained variation. In contrast, on average 65% of the explained variation in the distributions was explained by the spatial component of the environmental influence. The pure environmental effect has to be understood as the immediate influence of the environmental conditions on survival and reproduction of the organism, ignoring any types of immigration and emigration, temporal movements, and influences of proximate habitats. The purely spatial component of species' distributions would then have to be interpreted as resulting from behaviors leading to dispersal (as defined below) or other mechanisms causing spatial patterns independent of environmental conditions, such as, for example, dispersal barriers or "shadows" of past distributions. We use dispersal here in the sense of Lidicker (1975) including every movement that constitutes leaving the home area for breeding, but not short-term exploratory and "round-trip" migratory movements. This inclusive definition of dispersal includes a wide range of behaviors such as breeding aggregations, natal dispersal, adult dispersal, common-wealth breeding systems, and predator avoidance.

Interpreting the meaning of the "spatial component of the environmental influence" or "shared" partition is difficult. This partition does not directly depend on one model or at least two hierarchically dependent models, as the other partitions do (Meot et al. 1998). Instead, it depends on the unrelated ENV and CAR.TREND models. The most reasonable ecological interpretation

of this shared partition would be the spatial configuration of required habitat elements and matrix. At the coarse scales examined in our study, the shared partition could also include an isolation effect: an otherwise perfectly suitable patch may be unoccupied because of extreme isolation from other habitat. In any case, the large size of this partition drives home the point that a non-spatial view of the niche is not sufficient for understanding a species' distribution.

The selection of variables through RT models was a viable alternative to step-wise selection methods in regression models (Austin 2002). Without RTs, variable selection including interactions and non-linear effects would have likely led to spurious results because of the high number of independent variables available compared to the number of data points (James and McCulloch 1990). Even if the large number of variables had caused the RTs to select a few spurious variables, the likelihood that they would have been retained in the regression models would have been small because the functional relationship between dependent and independent variables is very different between the two techniques. A pre-selection among independent variables based on ecological knowledge would have been highly desirable and should be best practices for individual species (Austin 2002). Here, however, the goal was to build numerous models with comparable and reproducible methods for statistical comparison, and manual selection was less important. This automated modeling methodology also explains the relatively low average R^2 of 60% among the models.

The variable selection had an additional caveat. Keitt et al. (2002) and Lennon (2000) found that spatial models selected different independent variables than non-spatial models, because non-spatial models tend to recover the missing spatial information by including environmental variables that happen to have a similar spatial structure. While we found that a visual comparison of spatially plotted residuals showed less spatial clustering in RT models than in regular regression models, they are not spatially explicit models. Therefore, they might be subject to the variable selection bias documented by Keitt et al. (2002) and thus might have biased variable selection towards variables carrying a useful spatial structure (at the right scale) rather than variables explaining the distributions of the species causally. However, if this had been the case, it would have strengthened the subsequent non-spatial regression models, selecting variables better suited to non-spatial models and weakened the spatially explicit models. Therefore, if this "red-shift" played a role, we would have rather underestimated the spatial effect in our partitioning than overestimated it.

Another caveat of our methodology lies in the possibility that important environmental variables were missed in the models. If that had been the case, the spatial structure of these missed variables, reflected in the spatial structure of the bird distributions, would likely have been picked up by the spatial model, inflating the spatial partition at the loss of the environmental partition. However, given the number and quality of environmental variables we included (160 land cover, twelve climate, and four variables from digital elevation models) and the coarse scale of our investigation, we consider this scenario to be very unlikely. In addition, it is also unlikely that our spatial model recovered 100% of the pure spatial patterns in the birds' distributions. One major assumption of the spatial model is that the spatial effect is stationary. That means the effect is assumed to be the same everywhere in the range of a species, which is almost certainly not true. Thus, the spatial model certainly missed some of the spatial effect in the species' distributions and thus the spatial partition is likely an underestimate.

The demonstrated superiority of spatial models has implications for conservation biology and ecology studies. Standard distribution modeling techniques underestimate the spatial coherence of populations and thus may lead to more fragmented protected area designs that overvalue core habitats and undervalue mediocre neighboring habitats or matrix. Spatial models paint a more realistic picture of the importance of neighboring habitats and populations.

Future work is needed to identify the causal mechanisms behind autocorrelation in species' distributions over large extents. Autocorrelation over large distances is most likely caused by some form of movement or dispersal of the organisms, be it as seeds, juveniles, or adults. The hypothesis that coarse-scale autocorrelation is caused by long distance dispersal links autocorrelation to other ecological theories based on dispersal such as source-sink populations and metapopulations (Pulliam 1988, Dias 1996, Hanski 1998), occupancy-abundance relationships (Gaston et al. 2000, Holt et al. 2002), range structure theory (Kirkpatrick and Barton 1997), the unified neutral theory of biodiversity and biogeography (Hubbell 2001), and synchronicity among populations (Koenig 1998). The high utility of spatial models for the investigation of the link between dispersal and autocorrelation patterns in species' distributions is, in our opinion, their most interesting contribution to ecological theory.

Acknowledgements – We are indebted to the following people for advice, interesting discussions, reviews and editorial support: Brian McGill, Brian Ripley, Deanna Newsom, William Halteman, Steven Campbell, and Stephen Matthews. Bird ranges were provided by NatureServe in collaboration with Robert Ridgely, James Zook, The Nature Conservancy – Migratory Bird Program, Conservation International – CABS, World Wildlife Fund – US, and Environment Canada – WILDSPACE. Thanks to the many thousands of volunteer observers and organizers who contributed to the BBS data under the auspice of the U.S. Geological Survey's (USGS) Patuxent Wildlife Research Center and the Canadian Wildlife Service's National Wildlife Research Center. This project was funded by the USGS's Gap Analysis Program, and is a contribution of the Maine Cooperative Fish

and Wildlife Research Unit (USGS, Univ. of Maine, Maine Dept of Inland Fisheries and Wildlife, and Wildlife Management Inst., cooperating), and is publication no. 2878 of the Maine Agriculture and Forest Experiment Station.

References

Akaike, H. 1981. Modern development of statistical methods. – In: Eykhoff, P. (ed.), Trends and progress in systems identification. Pergamon Press, pp. 169–184.

Andrewartha, H. G. and Birch, L. C. 1954. The distribution and abundance of animals. – Univ. of Chicago Press.

Andrewartha, H. G. and Birch, L. C. 1984. The ecological web: more on the distribution and abundance of animals. – Univ. of Chicago Press.

Anon. 2003. S-PLUS ver. 6.2. – Insightful.

Araujo, M. B. and Williams, P. H. 2000. Selecting areas for species persistence using occurrence data. – Biol. Conserv. 96: 331–345.

Augustin, N. H. et al. 1996. An autologistic model for the spatial distribution of wildlife. – J. Appl. Ecol. 33: 339–347.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Modell. 157: 101–118.

Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. – Ecol. Modell. 153: 51–68.

Borcard, D. et al. 1992. Partialling out the spatial component of ecological variation. – Ecology 73: 1045–1055.

Breiman, L. 1984. Classification and regression trees. – Wadsworth International Group.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach. – Springer.

Costanza, R. and Ruth, M. 2001. Dynamic systems modeling. – In: Costanza, R. et al. (eds), Institutions, ecosystems, and sustainability. Lewis Publ., pp. 21–29.

Cressie, N. A. C. 1993. Statistics for spatial data. – Wiley.

Dale, M. R. T. and Fortin, M.-J. 2002. Spatial autocorrelation and statistical tests in ecology. – Ecoscience 9: 162–167.

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

Dias, P. C. 1996. Sources and sinks in population biology. – Trends Ecol. Evol. 11: 326–330.

Diniz-Filho, J. A. F. et al. 2003. Spatial autocorrelation and red herrings in geographical ecology. – Global Ecol. Biogeogr. 12: 53–64.

Ellner, S. P. and Seifu, Y. 2002. Using spatial statistics to select model complexity. – J. Comput. Graph. Stat. 11: 348–369.

Freeman, M. F. and Tukey, J. W. 1950. Transformations related to the angular and the square root. – Ann. Math. Stat. 21: 607–611.

Gaston, K. J. et al. 2000. Abundance-occupancy relationships. – J. Appl. Ecol. 37: 39–59.

Hanski, I. 1992. Inferences from ecological incidence functions. – Am. Nat. 139: 657–662.

Hanski, I. 1998. Metapopulation dynamics. – Nature 396: 41–49.

Hawkins, B. A. et al. 2003. Productivity and history as predictors of the latitudinal diversity gradient of terrestrial birds. – Ecology 84: 1608–1623.

Holt, A. R. et al. 2002. Occupancy-abundance relationships and spatial distribution: a review. – Basic Appl. Ecol. 3: 1–13.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.

James, F. C. and McCulloch, C. E. 1990. Multivariate analysis in ecology and systematics panacea or Pandora's box? – In: Johnston, R. F. (ed.), Annual review of ecology and systematics. Annual Reviews, pp. 129–166.

Kaluzny, S. P. et al. 1996. S+SPATIALSTATS user's manual, ver. 1.0. – MathSoft.

Keitt, T. H. et al. 2002. Accounting for spatial pattern when modeling organism-environment interactions. – Ecography 25: 616–625.

Kirkpatrick, M. and Barton, N. H. 1997. Evolution of a species' range. – Am. Nat. 150: 1–23.

Koenig, W. D. 1998. Spatial autocorrelation in California land birds. – Conserv. Biol. 12: 612–620.

Koenig, W. D. and Knops, J. M. H. 1998. Testing for spatial autocorrelation in ecological studies. – Ecography 21: 423–429.

Krebs, C. J. 1972. Ecology: the experimental analysis of distribution and abundance. – Harper and Row.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – Ecology 74: 1659–1673.

Legendre, P. and Fortin, M.-J. 1989. Spatial pattern and ecological analysis. – Vegetatio 80: 107–138.

Legendre, P. and Legendre, L. 1998. Numerical ecology. – Elsevier.

Lennon, J. J. 2000. Red-shifts and red herrings in geographical ecology. – Ecography 23: 101–113.

Lichstein, J. W. et al. 2002. Spatial autocorrelation and autoregressive models in ecology. – Ecol. Monogr. 72: 445–463.

Lidicker, W. Z., Jr 1975. The role of dispersal in the demography of small mammals. – In: Golley, F. B. et al. (eds), Small mammals: their productivity and population dynamics. Cambridge Univ. Press, pp. 103–128.

Loveland, T. R. et al. 1991. Development of a land-cover characteristics database for the conterminous U.S. – Photogramm. Eng. Rem. Sens. 57: 1453–1463.

Meot, A. et al. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. – Environ. Ecol. Stat. 5: 1–27.

Miller, J. and Franklin, J. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. – Ecol. Modell. 157: 227–247.

Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. – Biometrika 78: 691–692.

O'Connor, R. J. et al. 1999. Linking continental climate, land use, and land patterns with grassland bird distribution across the conterminous United States. – Stud. Avian Biol. 19: 45–59.

O'Connor, R. J. et al. 1996. Spatial partitioning of environmental correlates of avian biodiversity in the conterminous United States. – Biodiv. Lett. 3: 97–110.

Peakall, R. et al. 2003. Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, *Rattus fuscipes*. – Evolution 57: 1182–1195.

Pulliam, H. R. 1988. Sources sinks and population regulation. – Am. Nat. 132: 652–661.

Ridgely, R. S. et al. 2003. Digital distribution maps of the birds of the western hemisphere, ver. 1.0. – NatureServe.

Robbins, C. S. et al. 1986. The breeding bird survey: Its first fifteen years. – Fish Wildlife Service. Resource Publi. 177, p. 196.

Sauer, J. R. et al. 1994. Observer differences in the North American breeding bird survey. – Auk 111: 50–62.

Schiegg, K. 2003. Environmental autocorrelation: curse or blessing? – Trends Ecol. Evol. 18: 212–214.

Scott, J. M. and Csuti, B. 1997. Gap analysis for biodiversity survey and maintenance. – In: Reaka-Kudla, M. L. et al. (eds), Biodiversity, II. Understanding and protecting our biological resources. Joseph Henry Press, pp. 321–340.

Scott, M. J. et al. (eds) 2002. Predicting species occurrences: issues of accuracy and scale. – Island Press.

Selmi, S. and Boulinier, T. 2001. Ecological biogeography of Southern Ocean islands: the importance of considering spatial issues. – Am. Nat. 158: 426–437.

Student (Gosset, W. S.) 1914. The elimination of spurious correlation due to position in time or space. – Biometrika 10: 179–181.
Therneau, T. M. and Atkinson, E. J. 1997. An introduction to recursive partitioning using the rpart routines. – Dept of Health Science Research, Mayo Clinic, Rochester, MN, USA.
Thomson, J. D. et al. 1996. Untangling multiple factors in spatial distributions: lilies, gophers, and rocks. – Ecology 77: 1698–1715
Tobler, W. R. 1970. A computer movie simulating urban growth in the detroit region. – Econ. Geogr. 46: 234–240
Trenham, P. C. et al. 2001. Spatially autocorrelated demography and interpond dispersal in the salamander *Ambystoma californiense*. – Ecology 82: 3519–3530.
van Horssen, P. W. et al. 2002. Uncertainties in spatially aggregated predictions from a logistic regression model. – Ecol. Modell. 154: 93–101.
Venables, W. N. and Ripley, B. D. 2002. Modern applied statistics with S. – Springer.
Walker, P. A. and Cocks, K. D. 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. – Global Ecol. Biogeogr. Lett. 1: 108–118.
White, D. et al. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. – Cartogr. Geogr. Inform. Syst. 19: 5–21.
Wright, D. H. 1991. Correlations between incidence and abundance are expected by chance. – J. Biogeogr. 18: 463–466.
Zar, J. H. 1996. Biostatistical analysis. – Prentice Hall.

Download the appendix as file E4621 from <www.oikos.ekol.lu.se/appendix>.

Subject Editor: Andrew Liebhold.