RESEARCH REVIEW

Effects of incorporating spatial autocorrelation into the analysis of species distribution data

Carsten F. Dormann*

Department of Computational Landscape Ecology, UFZ Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany

ABSTRACT

Aim Spatial autocorrelation (SAC) in data, i.e. the higher similarity of closer samples, is a common phenomenon in ecology. SAC is starting to be considered in the analysis of species distribution data, and over the last 10 years several studies have incorporated SAC into statistical models (here termed 'spatial models'). Here, I address the question of whether incorporating SAC affects estimates of model coefficients and inference from statistical models.

Methods I review ecological studies that compare spatial and non-spatial models.

Results In all cases coefficient estimates for environmental correlates of species distributions were affected by SAC, leading to a mis-estimation of on average *c*. 25%. Model fit was also improved by incorporating SAC.

Main conclusions These biased estimates and incorrect model specifications have implications for predicting species occurrences under changing environmental conditions. Spatial models are therefore required to estimate correctly the effects of environmental drivers on species present distributions, for a statistically unbiased identification of the drivers of distribution, and hence for more accurate forecasts of future distributions.

*Correspondence: Carsten F. Dormann, Department of Computational Landscape Ecology, UFZ Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany. E-mail: carsten.dormann@ufz.de

Keywords

Autologistic regression, autoregressive model, spatial statistics, spatial autocorrelation, species distribution analysis, statistical biogeography.

INTRODUCTION

Ecologists have long been aware of spatial autocorrelation in their ecological data (Sokal & Oden, 1978a,b; Legendre, 1993; Koenig, 1999) and statistical methods for handling such complications have been available for almost as long (summarized in Cressie, 1993). However, these statistically correct methods have only recently been incorporated in distribution analyses. Still, most (i.e. > 80%) publications dealing with the analysis of spatial data in the ecological literature do not attempt to explicitly model spatial autocorrelation (SAC).

The effect of spatial autocorrelation on the interpretation of ordinary statistical methodology *in general* has been assessed several times (e.g. Liebhold & Sharov, 1998; Lennon, 2000; Dale & Fortin, 2002). It has been shown to influence both coefficients (for case studies see, e.g. Jetz & Rahbek, 2002; Lichstein *et al.*, 2002) and inference in statistical analyses (e.g. Chou & Soret, 1996; Fortin & Payette, 2002). For example, Tognelli and Kelt (2004) found a considerable change in the importance of explanatory variables after incorporating SAC: the second and third most important variables (vapour pressure and wind speed) in their non-spatial model dropped to become the least important in the spatial model.

The regression models employed for species distribution analyses model the expected value for a given set of environmental variables. The observed data points are the expected value plus an additional, unexplained noise — the variance. In truly independent data, the variance around the expected value is modelled as $\mathrm{Var}(\bar{x}) = \sigma^2/(n-1)$. In the spatially autocorrelated case, however, this variance has an additional component which specifies the covariance between values of *x* at locations *i* and *j*:

$$\mathrm{Var}(\bar{x}) = \sigma^2/(n-1) + [2/(n-1)^2]\sum_i \sum_{j(i<j)} \mathrm{Cov}(x_i, x_j)$$

(Haining, 2003; p. 275). This means that the larger the spatial autocorrelation the larger the covariance and the larger also the true variance around the expected value. Ignoring the second term in the above equation will lead to a downward-biased

estimate of σ^2 and accordingly incorrect tests of the significance of $\bar{x}$ in regression models. The statistical point thus is unambiguous (Griffith & Lagona, 1998): the statistical analysis of spatial data needs to incorporate spatial autocorrelation to avoid the pitfalls of spatial pseudoreplication (Hurlbert, 1984; Legendre, 1993), similar to a nested experiment requiring a mixed-model analysis to be structurally correct.

However, from the ecological point of view, spatial autocorrelation contains information one might not want to 'correct for' in the analysis. The most obvious ecological cause of spatial autocorrelation in species distribution data is dispersal (e.g. Austin, 2002; Epperson, 2005; Karst *et al.*, 2005; Lloyd *et al.*, 2005; Jones *et al.*, 2006). No species has globally dispersing offspring, and the density of propagules and progeny is usually decreasing with distance. Hence, the observed distribution pattern is the result of environmental factors as well as dispersal, competition and other ecological factors. Guisan and Thuiller (2005, p. 1002) have argued that predictive distribution models therefore need to explicitly model processes responsible for spatial autocorrelation (see also González-Megías *et al.*, 2005). Along a different line of argument, Diniz-Filho *et al.* (2003) see the use of spatial autocorrelation analysis as important at a different spatial scale, insofar as large-scale analyses (e.g. continental scale) may not need to incorporate SAC, since spatial variation occurs at a much larger scale than the ecological processes of dispersal and biotic interactions. However, from an ecological point of view spatial autocorrelation needs to be incorporated to account for dynamic ecological processes such as dispersal in static, statistical models (Austin, 2002).

Finally, several studies have analysed species distributions in a standard, non-spatial way, and found no evidence for spatial autocorrelation in model residuals (e.g. Higgins *et al.*, 1999; Hawkins & Porter, 2003; Bhattarai *et al.*, 2004; Flinn *et al.*, 2005; Warren *et al.*, 2005). The argument here is that 'A properly parametrized model (i.e. a model with the correct covariates) would reduce the need for the CAR [conditional autoregression] spatial structure' (B. D. Ripley, comment in Besag *et al.*, 1991). Legendre *et al.* (2002) called this 'spatial dependency' (i.e. SAC introduced into the response variable due to its dependence on an autocorrelated explanatory variable) as opposed to true spatial autocorrelation arising from ecological processes (e.g. dispersal). Therefore it has been argued that models with a high autocorrelation load in their residuals simply miss important ecological variables (Guisan & Thuiller, 2005). While this may be true, most of the time the 'correct' environmental variables will not be available for the analysis at the required spatial resolution (e.g. prey densities, pesticide application intensity, abundances of competitors, parasites and/or hosts) or at the necessary biological accuracy (e.g. temperature dependence of reproduction rate, habitat quality). So, for the time being, we have to manage with surrogate variables for such factors [climate, land use coverages, normalized difference vegetation index (NDVI), distances to specific habitats, etc.]. As a result, model residuals will probably display spatial autocorrelation.

In this review, I draw together evidence from the published literature on the effect of incorporating SAC into the analysis of spatial ecological data. I focus on species distribution data, together with spatial pattern in species richness or species performance. Two main questions shall be addressed: (1) does spatial autocorrelation influence the parameter estimates for species distribution data, or does intrinsic variability overshadow the problem; (2) does incorporating SAC lead to better-fitting statistical models?

METHODS

I searched for studies comparing a traditional distribution analysis with those incorporating a correction for spatial autocorrelation. As appropriate methods to deal with SAC I considered the following: autologistic regression (Augustin *et al.*, 1996; Gumpertz *et al.*, 1997; Wu & Huffer, 1997), generalized least square (GLS) regression (including simultaneous and conditional autoregressive models: Cliff & Ord, 1981; Anselin, 1988; Cressie, 1993; Anselin & Bera, 1998; Haining, 2003) and correction of significance levels (Clifford *et al.*, 1989; Dutilleul, 1993). Inclusion criteria for studies to this review were: (1) the distribution of organisms or ecological attributes (such as occurrence, abundance, biomass, species richness, etc.) was analysed; (2) both a traditional analysis [usually a generalized linear model (GLM) or generalized additive model (GAM)] and a spatial model were employed; and (3) results of this comparison were presented (although in some cases only in a qualitative way). Further restrictions (e.g. coefficients for the two model types need to be given; the analysis of residuals needs to be quantified) would have led to an even smaller set of studies.

Web of Science hits on the search phrase '"spatial autocorrelation" AND (ecology OR distribution)' were scanned for relevant studies. Additionally I searched for papers citing the benchmark study by Lichstein *et al.* (2002) and specifically for SAC-methods ('generalized least squares', 'autoregressive model' and 'autologistic regression'). Finally, I examined studies cited in each of the studies I used for this review. My search yielded 21 studies satisfying the inclusion criteria (Table 1). From each study I extracted the following information: (1) arrangement of samples (lattice/ points), (2) spatial extent/grain, (3) species/group, (4) response variable (presence/absence, richness, yield), (5) statistical methods, (6) size of neighbourhood, (7) type of autoregressive function, (8) quality of removal of SAC judged by semivariance, etc., (9) coefficients, SE and *P*-values of covariates, and (10) importance of spatial correction [R^2, deviance, Akaike information criterion (AIC)]. These data form the basis of the further analysis and review. In publications dealing with more than one species or group (e.g. Sanderson *et al.*, 2005) all modelled independent species were included as separate cases.

To quantify the effect of correcting for SAC on model coefficients, I used the following formula to transform spatial and non-spatial model coefficients (β_s and β_{ns}, respectively) into a 'relative SAC effect' [rSACe, comparable to the relative neighbour effect (RNE) in competition studies; Markham & Chanway, 1996]:

$$rSACe = \frac{|\beta_{ns} - \beta_s|}{\max(\beta_{ns}, \beta_s)}$$

This formula allows for a direct comparison of coefficients from different studies: the larger rSACe is, the greater is the difference

Table 1 Details of studies comparing spatial and non-spatial models in the analysis of distribution data

Study	Sampling*	Region	Area (km²)	Data points	Resolution (km)	Organism group†	Response variable‡	SAC method§	SAC range [cells]	Weighting function¶	Coefficient effect	r^2 (non-spatial)	r^2 (spatial)	Comments$
1	l	Europe	1.1×10^7	4419	50	p	PA	al	[1]	?	?	?	?	Means across 174 species models; neighbourhood set to 1, not tested or derived
2	p	Grampian	1300	1277	1.0	m	PA	al	7	?	?	?	?	SE of coefficients much higher in spatial model
3	p	California	2.7×10^5	34	?	p	r	SAR	?	?	?	?	?	Tables are wrong (S. Dark, pers. comm.)
4	l	Germany	4000	125	5.5	p	r	Dut	?	?	?	?	?	
5	l	Europe	1.2×10^7	250	220	b	r	GLS	?	exp	0.699	858	821	AIC
6	l	South America	1.8×10^7	374	220	b	r	GLS	7	sph	0.677	1974	1809	AIC
7	p	Reserve, NY	0.005	200	0.025	p	PA	Dut	?	?	?	?	?	Effective sample size decreased to 4.5–85% (of 200)
8	l	Iberian peninsula	6.0×10^5	240	50	p	r	GLS	?	sph	0.286	1655	1588	AIC
9	p	?	8.8×10^{-4}	64	0.0037	p	y	GLS	3	sph	0.083	?	?	
10	l	Africa	2.6×10^7	2605	100	b	r	CAR	2	?	0.250	0.76	0.66	
11	p	Norway	0.045	71	15	m	c	CAR	2.25	exp	0.266	269.4	268.1	AIC, difference n.s.
	p	Norway	0.045	71	15	m	c	SAR	2.25	exp	0.301	269.4	267.7	AIC, difference n.s.
	p	Quebec	2.6×10^{-5}	70	2.0×10^{-4}	a	c	GLS	0	exp	0.000	?	?	LR test for spatial dependence: $P = 0.71$
	p	Quebec	2.6×10^{-5}	70	2.0×10^{-4}	a	c	GLS	11	Gau	0.192	?	?	LR test for spatial dependence: $P = 0.02$
	p	Quebec	2.6×10^{-5}	70	2.0×10^{-4}	a	c	GLS	30	exp	0.785	?	?	LR test for spatial dependence: $P < 0.01$
	p	Quebec	2.6×10^{-5}	70	2.0×10^{-4}	a	c	GLS	4	Gau	0.691	?	?	LR test for spatial dependence: $P < 0.01$
12	p	Pennsylvania	2500	90	0.64	b	PA	al	3	inv	0.308	?	?	Commision and omission errors for spatial and non-spatial models given
13	p	Appalachians	600	1177	0.2	b	c	CAR	3	inv	0.083	0.53	0.55	
	p	Appalachians	600	1177	0.2	b	c	CAR	3	inv	0.200	0.22	0.25	
	p	Appalachians	600	1177	0.2	b	c	CAR	3	Inv	0.155	0.39	0.46	
14	p	Sweden	3750	313	2500	b	PA	al	?	1	0.316	144.8	141.2	AIC
	P	Sweden	3750	313	2500	b	PA	al	?	1	0.163	169.8	171.7	AIC
15	p	England	3	76	0.005	i	PA	al	2	inv	?	0.12	0.14	Deviance-based pseudo-R^2
	p	England	3	76	0.005	i	PA	al	2	inv	?	0.09	0.12	Deviance-based pseudo-R^2
	p	England	3	76	0.005	i	PA	al	2	inv	?	0.08	0.13	Deviance-based pseudo-R^2
	p	England	3	76	0.005	i	PA	al	32	inv	?	0.23	0.32	Deviance-based pseudo-R^2
	p	England	3	76	0.005	i	PA	al	4	inv	?	0.14	0.34	Deviance-based pseudo-R^2
	p	England	3	76	0.005	i	PA	al	5	inv	?	0.20	0.26	Deviance-based pseudo-R^2
	p	England	3	76	0.005	mo	PA	al	7	inv	?	0.21	0.34	Deviance-based pseudo-R^2
	p	England	3	76	0.005	mo	PA	al	20	inv	?	0.13	0.15	Deviance-based pseudo-R^2
	p	England	3	76	0.005	c	PA	al	10	inv	?	0.16	0.25	Deviance-based pseudo-R^2
	p	England	3	76	0.005	c	PA	al	1	inv	?	0.47	0.54	Deviance-based pseudo-R^2
16	l	Portugal	99,000	993	10	r	PA	al	2	inv	?	?	?	Data are prediction quality across 44 species
17	p	Tunisia	45,000	53	0.4	b	PA	al	?	inv	0.326	53.4	26.9	AIC
	p	Tunisia	45,000	53	0.4	b	c	al	?	inv	0.274	39.1	37.2	AIC
18	l	Colorado	0.001	256	0.002	p	c	CRH	7	?	?	?	?	Only change in P-value is given, which is significant in 4/9 cases
19	l	South America	1.8×10^7	1828	100	m	r	CAR	3	inv	0.096	0.79	0.84	
	l	South America	1.8×10^7	1828	100	m	r	SAR	3	inv	0.438	0.79	0.93	
20	p	Ontario	0.08	6811	0.001	p	s	GLS	4	exp	?	0.76	0.96	
21	l	Spain	10,000	152	10	b	PA	al	[1]	1	0.252	0.87	0.87	nb. size set to 1, not tested or derived

Table 1 *Continued*

Study	Sampling*	Region	Area (km²)	Data points	Resolution (km)	Organism group†	Response variable‡	SAC method§	SAC range [cells]	Weighting function¶	Coefficient effect	r² (non-spatial)	r² (spatial)	Comments$
	l	Spain	10,000	152	10	b	PA	al	[1]	1	0.348	0.80	0.83	nb. size set to 1, not tested or derived
	l	Spain	10,000	152	10	b	PA	al	[1]	1	0.663	0.88	0.90	nb. size set to 1, not tested or derived

?, No information given.

Studies: 1, Araújo and Williams (2000); 2, Augustin *et al.* (1996); 3, Dark (2004); 4, Deutschewitz *et al.* (2003); 5, Diniz-Filho *et al.* (2003); 6, Diniz-Filho and Bini (2005); 7, Fang (2005); 8, Ferrer-Castán and Vetaas (2005); 9, Gotway and Stroup (1997); 10, Jetz and Rahbek (2002); 11, Keitt *et al.* (2002); 12, Klute *et al.* (2002); 13, Lichstein *et al.* (2002); 14, Mörtberg and Karlström (2005); 15, Sanderson *et al.* (2005); 16, Segurado and Araújo (2004); 17, Selmi *et al.* (2003); 18, Thomson *et al.* (1996); 19, Tognelli and Kelt (2004); 20, Zhang *et al.* (2005); 21, Chou and Soret (1996).

*l, lattice; p, points.

†a, acari (mite); b, bird; c, crustacean; i, insect; m, mammal; mo, mollusc; p, plant; r, reptile/amphibian.

‡c, counts of abundance; PA, presence/absence; r, richness; s, size; y, biomass yield.

§al, autologistic regression; CAR, conditional autoregressive model; CRH, Clifford/Richarson/Hémon correction; Dut, Dutilleul's correction; GLS, generalized least squares; SAR, simultaneous autoregressive model.

¶1, non-weighted; exp, exponential; Gau, Gaussian; i, inverse of Euclidean distance; sph, spherical.

[] indicates that values were set *a priori*, rather than explored iteratively.

$AIC, Akaike Information Criterion; LR Test, likelihood-ratio test, nb site, neighborhood site.

between coefficient estimates from the spatial and non-spatial model. Calculation of effect sizes (Gurevitch & Hedges, 1999, 2001) as in a meta-analysis would have been preferable, but in only six studies were estimates of variance, standard deviation or standard error given. The rSACe should hence be interpreted cautiously, since the error on the coefficient estimates may be so large as to make differences insignificant.

Several parameters were log normally distributed (spatial resolution of study, area of study, rSACe). To summarize these values parametrically, I $\log_{10}$-transformed them before analysis and back-transformed them afterwards. Error bars were calculated on $\log_{10}$-transformed data as well, and back-transformed from log mean ± log SE.

The effect of correcting for SAC on overall model quality can be quantified by AIC, R^2 or deviance-based pseudo-R^2 values given in the publications. However, AIC and R^2 values are not directly comparable without the deviance of the intercept-only (null) model. Hence these values should be interpreted mainly qualitatively. All analyses were carried out using R version 2.1.1 (R Development Core Team, 2004).

RESULTS

Because 13 studies report on single species (e.g. their occurrence, abundance or some measure of their performance such as yield) while eight recorded species richness (see Table 1), I first analysed whether there was a systematic difference between these response types in terms of relative spatial autocorrelation effect. ANOVA results ($F_{1,22} = 0.91$, $P = 0.351$) indicate that there was no difference in rSACe. Additionally, in all analyses presented below,

response type was included both as a main and an interacting factor, and was not significant in all cases. I therefore present the studies together, irrespective of the type of response assessed.

Study regions ranged in size from 1000 m² (a field experiment: Gotway & Stroup, 1997) to 26 million km² (sub-Saharan Africa: Jetz & Rahbek, 2002). Accordingly resolution (or grain) ranged from approx. 25 m² (Sanderson *et al.*, 2005) to 220 km² (Diniz-Filho *et al.*, 2003; Diniz-Filho & Bini, 2005).

There was a highly significant correlation between the range of spatial autocorrelation (also called 'neighbourhood size', i.e. the distance over which spatial autocorrelation was taking effect), and spatial resolution of a study (Fig. 1). Phrased differently, when correcting for the differences in spatial resolution of the studies, neighbourhood size in terms of multiples of resolution was relatively constant across the entire range of study areas (back-transformed mean of $\log_{10}$-transformed data = 4.29, mean ± 1 SE = 5.00 and 3.68, respectively) and significantly different from 1 ($t_{1,29} = 9.48$, $P < 0.001$). This clearly shows that spatial autocorrelation was apparent in all cases. Among the vertebrates, mammals (three studies) and birds (seven studies) showed no difference in slope (ANCOVA of log-transformed data; interaction between resolution and group: $F_{1,6} = 0.004$, $P = 0.951$).

Effects of spatial autocorrelation on coefficient estimation

Across the 24 cases analysed the mean effect of spatial autocorrelation on regression coefficients (rSACe) was 0.25 (back-transformed mean of log-transformed data; mean + 1 SE = 0.31 and mean − 1 SE = 0.21). In *all* cases where comparisons between
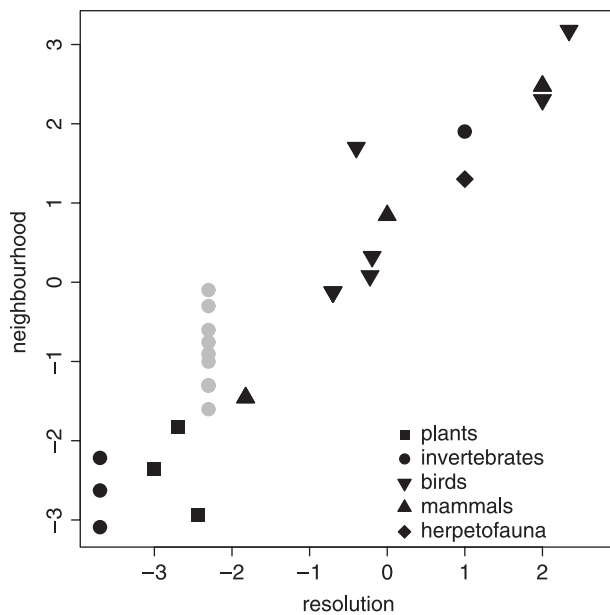
Figure 1 Neighbourhood size is related to the spatial resolution of a study: $y = x^{0.88\ (\pm\ 0.055)} - 0.70\ (\pm\ 0.119)$. $R^2 = 0.90$, $P < 0.001$, $n = 30$. Note that values on both axes were $\log_{10}$-transformed. Grey points represent freshwater organisms, all belonging to the data set of Sanderson *et al.* (2005). Note that the three rightmost points represent analyses of species richness, not species occurrence, abundance or performance.

the mean from 0.43 for non-spatial models to 0.49 for spatial models (Fig. 3). The average increase in adjusted model R^2 was 0.060 (min. = −0.10, max. = 0.20), with only one of the 20 spatial models performing worse than its non-spatial counterpart. Studies only giving AICs for spatial and non-spatial models report a significant decrease in AIC by incorporation of a SAC correction in six of nine cases.

In addition to those studies that have been quantitatively summarized above some studies provide strong evidence for the importance of SAC for modelling species distributions, but do not give the investigated parameters in the publications. Augustin *et al.* (1996) report R^2 values for four different spatial models, but not for the non-spatial. Araújo and Williams (2000) present a summary of 174 models for European tree species. Their analysis shows, without further quantification, a strong effect of incorporating a correction term for spatial autocorrelation. Kaboli *et al.* (2006) showed that the spatial autocovariate remained significant in all three of their models on richness, abundance and composition of bird communities in Iran. The

spatial and non-spatial models had been made, coefficients were biased.

There was no evidence for a change of rSACe with region size or resolution of the study (Kendall's correlation $\tau = 0.210$, $P = 0.162$ and $\tau = 0.252$, $P = 0.094$, respectively). Nor was there a significant difference in rSACe with respect to the separation into plants and animals or further into plant, vertebrates and invertebrates (in all cases Kruskal–Wallis and ANOVA $P > 0.4$; see Fig. 2).

Effects of spatial autocorrelation on model quality

Across the 20 cases which provide model R^2 values (or pseudo-R^2 in the case of non-normal errors) incorporating SAC increased


Figure 3 Difference in model fit between non-spatial and spatial models. This difference is significant in a paired Wilcoxon signed-rank test ($V = 15$, $P < 0.01$).

Figure 2 Relative effect of incorporating spatial autocorrelation into models on the model coefficients for different species groups. Bars ($\pm$ 1 SE) represent back-transformed means of log-transformed relative SAC effects, error bars are hence asymmetric. The difference between plants and vertebrates is not significant ($P = 0.223$, ANOVA on $\log(x + 0.01)$-transformed data).
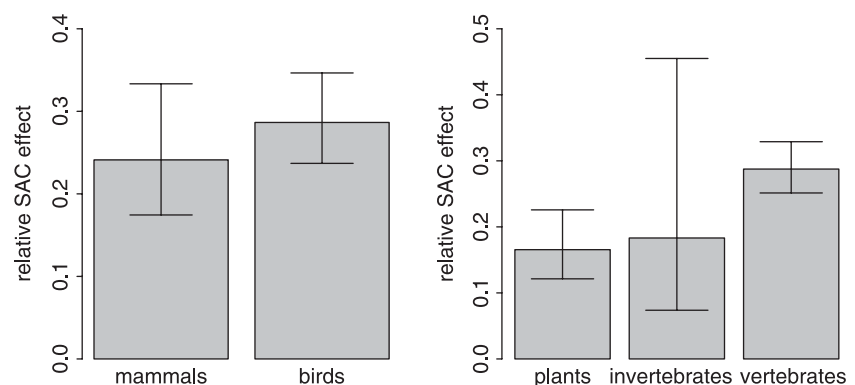
Dutilleul correction employed by Deutschewitz *et al.* (2003), Fang (2005) and Rodríguez *et al.* (2006) only corrects the significance value of a correlation (in both cases the effect was pronounced), but cannot be applied to model R^2. The same holds true for Thomson *et al.* (1996), who use a correction proposed by Clifford *et al.* (1989; see also Haining, 2003). These four studies together find a pronounced change (loss) in correlation significance in 18 of 35 correlations. Keitt *et al.* (2002) give likelihood-ratio tests for spatial dependence (significant in two of three cases), but not the deviances of the models. Klute *et al.* (2002) assess their model quality in terms of commission and omission errors (which were reduced by 50% and 30%, respectively, in spatial models), but not the fit as such. Segurado and Araújo (2004) give model quality across 44 different species being modelled. The spatial versions of their GLMs were significantly better (in terms of Kappa index). This did not hold true for GAMs, which were similar for both spatial and non-spatial types.

For the eight studies that presented both model coefficients and adjusted R^2 values no correlation between these two parameters was detectable (Kendall's $\tau = -0.028$, $P = 0.917$). This indicates no systematic relationship between model fit and bias in coefficients with respect to the comparison of spatial and non-spatial models.

DISCUSSION

Ecological causes of spatial autocorrelation

In this review I have considered both single-species distribution and performance data as well as species richness pattern. Obviously the causes of spatial autocorrelation are not necessarily the same for these two types of responses. Since composite measures of assemblages are also affected by the spatial autocorrelation in each of the contributing species, I will first discuss the mechanisms, both endogenous and exogenous (see review by Liebhold *et al.*, 2004), that introduce spatial autocorrelation into the distribution data for a single species.

Exogenous factors, such as climate, soil type, stochastic disturbances or even solar activity (Ranta *et al.*, 1997), may lead to a similar occurrence probability in neighbouring sites, simply because the external factors show a specific autocorrelation pattern. These exogenous factors can ideally be included into the statistical model as environmental covariates, reducing and even removing the residual spatial autocorrelation (e.g. Besag *et al.*, 1991; Higgins *et al.*, 1999; Fisher *et al.*, 2002; Hawkins & Porter, 2003; Bhattarai *et al.*, 2004; Flinn *et al.*, 2005; Warren *et al.*, 2005). If omitted, residuals are likely to display spatial autocorrelation.

Endogenous factors are due to the biology of the species under consideration: dispersal, colonial breeding, home-range size, competition, host availability, predation or parasitization risk, and so forth. Van Horne (2002) showed that mapping eagle occurrences at a spatial scale of 1 km caused a high level of spatial autocorrelation because these birds roam distances of tens of kilometres (see Scott *et al.*, 2002; for further examples). These causes of SAC are usually much more difficult to quantify, and

data are often scarce (e.g. Cain *et al.*, 2000). Most of these (most noticeably the interaction with other species) occur at small spatial scales (e.g. less than 1 km), and dispersal for plants and insects becomes extremely rare events at larger scales.

For species assemblage data (such as species richness, percentage of endemics, proportion of Red Data Book species, etc.) additional processes may introduce spatial autocorrelation. Most prominent among them is the omission of a variable relevant at the community scale (such as disturbance or management). Also possible are artefacts due to species-specific bias or different recorder density. Taxonomic specialists may, for example, subdivide plant species into more 'species' than a common botanist, or a recording team may sample one region more intensively than another, producing a bias unrelated to the environment. Because it is not always possible to correct for such artefacts, they may still show up as residual spatial autocorrelation. Spatial autocorrelation may also be introduced as a consequence of the sampling scheme (Fortin & Dale, 2005), when the regions of a known occurrence are sampled with higher intensity than regions of unclear occurrence (e.g. for Red List species). Finally, ecological interactions between species (competitive replacement) or founder effects in isolated habitat patches (fragmented landscapes, lakes) will add to SAC in assemblage data that is absent from the individual species distribution data.

Endogenous causes of SAC (dispersal, interspecific interactions, disturbance) can be expected to operate at smaller spatial scales (Guisan & Thuiller, 2005), and should hence be relevant for studies with a higher resolution. The present study shows that the range of spatial autocorrelation was very constant across seven orders of magnitude and various groups of organisms. This result suggests that various processes (biological and other) may contribute to spatial autocorrelation and that we cannot assume SAC to be a small-scale problem. As another consequence, resampling data at a coarser spatial scale, as a recommended treatment against SAC (Qi & Wu, 1996; Aubry & Debouzie, 2000; Guisan & Theurillat, 2000; Rossi & Nuutinen, 2004), will not necessarily solve the problem (see also Fortin & Dale, 2005; p. 248 *et seq*., for other reasons why sampling at coarser scales may not be adequate).

As can be concluded from the limited set of studies available so far, there is no evidence that models for plants and different animal groups differ in their susceptibility to SAC. The effect of SAC on regression coefficients was similar for all groups investigated. However, several studies found no spatial autocorrelation in the model residuals, and hence would not find an effect of including a SAC correction. These were mainly plant distribution analyses (e.g. Higgins *et al.*, 1999; but see Fisher *et al.*, 2002; Hawkins & Porter, 2003; Williams *et al.*, 2005), suggesting that plant distributions are less spatially autocorrelated than those of animals. The data analysed here were too sparse to assess differences between organism groups with respect to the range of SAC (Fig. 1). The large interval of study sizes and resolution is founded mainly on mammal and bird studies, while plants and invertebrates were only available for a much smaller spatial resolution interval. Only one study reported the effect of SAC for insects at the country scale (Luoto *et al.*, 2005).

Effects of spatial autocorrelation on model parameters

Ignoring SAC leads to two kinds of possible errors: biased parameter estimates and overly optimistic standard errors (for examples see Albert & McShane, 1995; Keitt *et al.*, 2002). The latter is important when model results are used for predicting species distributions for environmental change scenarios since it is another intrinsic problem adding to prediction uncertainty. The studies reviewed here are unambiguous with respect to the bias introduced by neglecting spatial autocorrelation: in all cases coefficients were affected. The large difference of close to 25% between the spatial and non-spatial models indicates that this effect is not only common, but also strong. To use spatial auto-correlation in the prediction is extremely difficult: Augustin *et al.* (1996) used a Gibbs sampler to iteratively recalculate occurrence probabilities using the autocovariate signal they detected in their data. Such an approach is very sensitive to starting conditions and may in many cases not converge (J. McPherson, pers. comm.).

However, not all previous studies which did not incorporate spatial autocorrelation are fundamentally flawed. Models may be wrongly specified because they contain the 'wrong' explanatory variables (e.g. they ignore environmental factors that are important), which may lead to far worse models than ignoring spatial autocorrelation (Haining, 2003, p. 273). Hence spatial auto-correlation is only one more potential problem of which ecologists should be aware. This review hopes to have shown that, based on currently available evidence, spatial autocorrelation is relevant across all groups of organisms and all spatial scales. The state of the art simply demands spatial models if residuals are spatially autocorrelated.

Effects of spatial autocorrelation on model building and fit

Why bother with better models? Usually the model fit of the non-spatial version was already good, and the improvement with the SAC correction term was only moderate (Fig. 3). In a so-far unique study, Chou and Soret (1996) demonstrated that models may still change qualitatively after accounting for SAC. They first simplified a non-spatial model, then added an autologistic term, and finally eliminated insignificant model terms again. These final models differed qualitatively from the final non-spatial model, and therefore hint that non-spatial models for autocorrelated data may be inadequately 'well fitting'. To date, no study has attempted to evaluate whether the more sophisticated 'true' spatial models are actually less biased than 'wrong' but simple non-spatial models (e.g. on artificial data with known causal relationships and controlled spatial autocorrelation). This is clearly a question deserving future research, since model predictions depend strongly on the bias introduced by the model structure (Reineking & Schröder, 2006).

At present, no standard approach to model construction and simplification under spatial autocorrelation has been established. Most commonly, a non-spatial model is constructed, simplified to contain only significant variables (usually using information-theoretical methods: Burnham & Anderson, 2002) and eventually the spatial version is employed. This is the case both for ordinary least square (OLS) models that in the final step are reformulated as GLS or conditional/simultaneous autoregressive model, and for GLMs employing an autocovariate (e.g. autologistic regression). Non-spatial OLS and GLM are much faster to run and hence to simplify than their spatial counterparts. Dale and Fortin (2002) argue that autologistic regression corrects the coefficient estimates for the presence of SAC, but does not correct the degrees of freedom employed. This result suggests that models may be wrongly specified, as model simplification will be based on incorrect likelihood ratios. The results of Chou and Soret (1996) illustrate how the autocovariate alters the model structure, but whether this is due to false degrees of freedom or due the 'true' effect of SAC remains obscure.

Outlook

Will species distribution models need to go spatial? Guisan and Thuiller (2005) claim that SAC models are barely transferable in space, because the spatial configuration of the landscape will usually be different in a new site, hence making the SAC correction (which is based on a neighbourhood size) inappropriate. This may be so. However, non-spatial models suffer the same deficiency, since their estimates are biased due to the same landscape configuration issue, only it is not captured by one spatial coefficient but rather influences all coefficient estimates. Landscape configuration is a variable worth considering as a covariate in species distribution models, especially if the landscape metric employed is based on ecological reasoning. This should alleviate some of the problems Guisan and Thuiller foresee. If, however, SAC is due to biological processes such as dispersal or species interactions, spatial models will be transferable (reusing their covariance structure or with Gibbs sampling) and more accurate because the spatial component represents some relevant biological process that cannot be accounted for by environmental covariates.

Some of the studies reviewed here were intended to investigate the effect of correcting for spatial autocorrelation as part of their study design. However, most of them needed to present their findings for both spatial and non-spatial models because the need to incorporate spatial autocorrelation is not yet generally perceived to be an important issue (e.g. it is not even mentioned in the excellent methodological review by Elith *et al.*, 2006). Only very recently studies have used corrections for spatial autocorrelation without having to justify their effect on model quality (e.g. Kühn *et al.*, 2003; Luoto *et al.*, 2005; Newbury & Simon, 2005; Orme *et al.*, 2005; Warren *et al.*, 2005; Worm *et al.*, 2005; Kaboli *et al.*, 2006; Stephenson *et al.*, 2006). The significance of the correction terms in their analyses is evidence that the residuals of their models were indeed spatially autocorrelated. These authors certainly represent a minority of spatial ecologists who have taken SAC matter-of-factly into consideration and included it in their models. For the time being, both spatial and non-spatial models should be presented to aid future comparisons on the effect of incorporating spatial autocorrelation on model estimates.

REFERENCES

Albert, P.S. & McShane, L.M. (1995) A generalized estimating equations approach for spatially correlated binary data: with an application to the analysis of neuroimaging data. *Biometrics*, **51**, 627–638.

Anselin, L. (1988) *Spatial econometrics: methods and models*. Kluwer, Dordrecht.

Anselin, L. & Bera, A. (1998) Spatial dependence in linear regression models with an introduction to spatial econometrics. *Handbook of applied economic statistics* (ed. by A. Ullah and D. Giles), pp. 237–289. Marcel Dekker, New York.

Araújo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.

Aubry, P. & Debouzie, D. (2000) Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling. *Ecology*, **81**, 543–553.

Augustin, N.H., Mugglestone, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Besag, J., York, J. & Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Bhattarai, K.R., Vetaas, O.R. & Grytnes, J.A. (2004) Fern species richness along a central Himalayan elevational gradient, Nepal. *Journal of Biogeography*, **31**, 389–400.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multi-model inference: a practical information-theoretical approach*. Springer, Berlin.

Cain, M.L., Milligan, B.G. & Strand, A.E. (2000) Long-distance seed dispersal in plant populations. *American Journal of Botany*, **87**, 1217–1227.

Chou, Y.H. & Soret, S. (1996) Neighborhood effects in bird distributions, Navarre, Spain. *Environmental Management*, **20**, 675–687.

Cliff, A.D. & Ord, J.K. (1981) *Spatial processes: models and applications*. Pion Limited, London.

Clifford, P., Richarson, S. & Hémon, D. (1989) Assessing the significance of the correlation between two spatial processes. *Biometrics*, **45**, 123–134.

Cressie, N.A.C. (1993) *Statistics for spatial data*. Wiley, New York.

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

Dark, S.J. (2004) The biogeography of invasive alien plants in California: an application of GIS and spatial regression analysis. *Diversity and Distributions*, **10**, 1–9.

Deutschewitz, K., Lausch, A., Kühn, I. & Klotz, S. (2003) Native and alien plant species richness in relation to spatial heterogeneity on a regional scale in Germany. *Global Ecology and Biogeography*, **12**, 299–311.

Diniz-Filho, J.A.F. & Bini, L.M. (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecology and Biogeography*, **14**, 177–186.

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.

Dutilleul, P. (1993) Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics*, **49**, 305–314.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.McC., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Epperson, B.K. (2005) Estimating dispersal from short distance spatial autocorrelation. *Heredity*, **95**, 7–15.

Fang, W. (2005) Spatial analysis of an invasion front of *Acer platanoides*: dynamic inferences from static data. *Ecography*, **28**, 283–294.

Ferrer-Castán, D. & Vetaas, O.R. (2005) Pteridophyte richness, climate and topography in the Iberian Peninsula: comparing spatial and nonspatial models of richness patterns. *Global Ecology and Biogeography*, **14**, 155–167.

Fisher, R.N., Suarez, A.V. & Case, T.J. (2002) Spatial patterns in the abundance of the coastal horned lizard. *Conservation Biology*, **16**, 205–215.

Flinn, K.M., Vellend, M. & Marks, P.L. (2005) Environmental causes and consequences of forest clearance and agricultural abandonment in central New York, USA. *Journal of Biogeography*, **32**, 439–452.

Fortin, M.J. & Dale, M.R.T. (2005) *Spatial analysis — a guide for ecologists*. Cambridge University Press, Cambridge.

Fortin, M.J. & Payette, S. (2002) How to test the significance of the relation between spatially autocorrelated data at the landscape scale: a case study using fire and forest maps. *Écoscience*, **9**, 213–218.

González-Megías, A., Gómez, J.M. & Sánchez-Pinero, F. (2005) Consequences of spatial autocorrelation for the analysis of metapopulation dynamics. *Ecology*, **86**, 3264–3271.

Gotway, C.A. & Stroup, W.W. (1997) A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biology, and Environmental Statistics*, **2**, 157–178.

Griffith, D.A. & Lagona, F. (1998) On the quality of likelihood-based estimators in spatial autoregressive models when the data dependence structure is misspecified. *Journal of Statistical Planning and Inference*, **69**, 153–174.

Guisan, A. & Theurillat, J.-P. (2000) Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia*, **30**, 353–384.

Guisan, A. & Thuiller, W. (2005) Predicting species distributions: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Gumpertz, M.L., Graham, J.M. & Ristaino, J.B. (1997) Autologistic model of spatial pattern of *Phytophthora* epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biology and Environmental Statistics*, **2**, 131–156.

Gurevitch, J. & Hedges, L.V. (1999) Statistical issues in ecological meta-analysis. *Ecology*, **80**, 1142–1149.

Gurevitch, J. & Hedges, L.V. (2001) Meta-analysis: combining the results of independent experiments. *Design and analysis of ecological experiments*, 2nd edn (ed. by S.M. Schreiner and J. Gurevitch), pp. 347–370. Chapman & Hall, London.

Haining, R. (2003) *Spatial data analysis — theory and practice.* Cambridge University Press, Cambridge.

Hawkins, B.A. & Porter, E.E. (2003) Does herbivore diversity depend on plant diversity? The case of California butterflies. *The American Naturalist*, **161**, 40–49.

Higgins, S.I., Richardson, D.M., Cowling, R.M. & Trinder-Smith, T.H. (1999) Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. *Conservation Biology*, **13**, 303–313.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological experiments. *Ecological Monographs*, **54**, 187–211.

Jetz, W. & Rahbek, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548–1551.

Jones, F.A., Hamrick, J.L., Peterson, C.J. & Squiers, E.R. (2006) Inferring colonization history from analyses of spatial genetic structure within populations of *Pinus strobus* and *Quercus rubra*. *Molecular Ecology*, **15**, 851–861.

Kaboli, M., Guillaumet, A. & Prodon, R. (2006) Avifaunal gradients in two arid zones of central Iran in relation to vegetation, climate, and topography. *Journal of Biogeography*, **33**, 133–144.

Karst, J., Gilbert, B. & Lechowicz, M.J. (2005) Fern community assembly: the roles of chance and the environment at local and intermediate scales. *Ecology*, **86**, 2473–2486.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism–environment interactions. *Ecography*, **25**, 616–625.

Klute, D.S., Lovallo, M.J. & Tzilkowski, M. (2002) Autologistic regression modeling of American woodcock habitat use with spatially dependent data. *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.B. Wall and F. Samson), pp. 335–343. Island Press, Washington, DC.

Koenig, W.D. (1999) Spatial autocorrelation of ecological phenomena. *Trends in Ecology & Evolution*, **14**, 22–26.

Kühn, I., May, R., Brandl, R. & Klotz, S. (2003) Plant distribution patterns in Germany — will aliens match natives? *Feddes Repertorium*, **114**, 559–573.

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

Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M.

& Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, **25**, 601–615.

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

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

Liebhold, A.M. & Sharov, A.A. (1998) Testing for correlation in the presence of spatial autocorrelation in insect count data. *Population and community ecology for insect management and conservation* (ed. by J. Baumgartner, P. Brandmayr & B.F.J. Manly), pp. 111–117. Balkema, Rotterdam.

Liebhold, A., Koenig, W.D. & Bjørnstad, O.N. (2004) Spatial synchrony in population dynamics. *Annual Review of Ecology Evolution and Systematics*, **35**, 467–490.

Lloyd, N.J., Mac Nally, R. & Lake, P.S. (2005) Spatial autocorrelation of assemblages of benthic invertebrates and its relationship to environmental factors in two upland rivers in southeastern Australia. *Diversity and Distributions*, **11**, 375–386.

Luoto, M., Poyry, J., Heikkinen, R.K. & Saarinen, K. (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**, 575–584.

Markham, J.H. & Chanway, C.P. (1996) Measuring plant neighbour effects. *Functional Ecology*, **10**, 548–549.

Mörtberg, U. & Karlström, A. (2005) Predicting forest grouse distribution taking account of spatial autocorrelation. *Journal for Nature Conservation*, **13**, 147–159.

Newbury, T.L. & Simon, N.P.P. (2005) The effects of clearcutting on snowshoe hare (*Lepus americanus*) relative abundance in central Labrador. *Forest Ecology and Management*, **210**, 131–142.

Orme, C.D.L., Davies, R.G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V.A., Webster, A.J., Ding, T.-S., Rasmussen, P.C., Ridgely, R.S., Stattersfield, A.J., Bennett, P.M., Blackburn, T.M., Gaston, K.J. & Owens, I.P.F. (2005) Global hotspots of species richness are not congruent with endemism or threat. *Nature*, **436**, 1016–1019.

Qi, Y. & Wu, J.G. (1996) Effects of changing spatial resolution on the results of landscape pattern analysis using spatial auto-correlation indices. *Landscape Ecology*, **11**, 39–49.

R Development Core Team (2004) *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org).

Ranta, E., Lindström, J., Kaitala, V., Kokko, H., Lindén, H. & Helle, E. (1997) Solar activity and hare dynamics: a cross-continental comparison. *The American Naturalist*, **149**, 765–775.

Reineking, B. & Schröder, B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.

Rodríguez, M.A., López-Sañudo, I.L. & Hawkins, B.A. (2006) The geographic distribution of mammal body size in Europe. *Global Ecology and Biogeography*, **15**, 173–181.

Rossi, J.P. & Nuutinen, V. (2004) The effect of sampling unit size on the perception of the spatial pattern of earthworm

(*Lumbricus terrestris* L.) middens. *Applied Soil Ecology*, **27**, 189–196.

Sanderson, R.A., Eyre, M.D. & Rushton, S.P. (2005) Distribution of selected macroinvertebrates in a mosaic of temporary and permanent freshwater ponds as explained by autologistic models. *Ecography*, **28**, 355–362.

Scott, J.M., Heglund, P.J., Morrison, M., Haufler, J.B. & Wall, W.A. (2002) *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, DC.

Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.

Selmi, S., Boulinier, T. & Faivre, B. (2003) Distribution and abundance patterns of a newly colonizing species in Tunisia oases: the Common Blackbird *Turdus merula*. *Ibis*, **145**, 681–688.

Sokal, R.R. & Oden, N.L. (1978a) Spatial autocorrelation in biology. I. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.

Sokal, R.R. & Oden, N.L. (1978b) Spatial autocorrelation in biology. II. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**, 229–249.

Stephenson, C.M., MacKenzie, M.L., Edwards, C. & Travis, J.M.J. (2006) Modelling establishment probabilities of an exotic plant, *Rhododendron ponticum*, invading a heterogeneous, woodland landscape using logistic regression with spatial autocorrelation. *Ecological Modelling*, **193**, 747–758.

Thomson, J.D., Weiblen, G., Thomson, B.A., Alfaro, S. & Legendre, P. (1996) Untangling multiple factors in spatial distributions: Lilies, gophers, and rocks. *Ecology*, **77**, 1698–1715.

Tognelli, M.F. & Kelt, D.A. (2004) Analysis of determinants of mammalian species richness in South America using spatial autoregressive models. *Ecography*, **27**, 427–436.

Van Horne, B. (2002) Approaches to habitat modeling: the tensions between pattern and process and between specificity and generality. *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.B. Wall and F. Samson), pp. 63–72. Island Press, Washington, DC.

Warren, T.L., Betts, M.G., Diamond, A.W. & Forbes, G.J. (2005) The influence of local habitat and landscape composition on cavity-nesting birds in a forested mosaic. *Forest Ecology and Management*, **214**, 331–343.

Williams, J.W., Seabloom, E.W., Slayback, D., Stoms, D.M. & Viers, J.H. (2005) Anthropogenic impacts upon plant species richness and net primary productivity in California. *Ecology Letters*, **8**, 127–137.

Worm, B., Sandow, M., Oschlies, A., Lotze, H.K. & Myers, R.A. (2005) Global patterns of predator diversity in the open oceans. *Science*, **309**, 1365–1136.

Wu, H.L. & Huffer, F.W. (1997) Modelling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics*, **4**, 49–64.

Zhang, L.J., Gove, J.H. & Heath, L.S. (2005) Spatial residual analysis of six modeling techniques. *Ecological Modelling*, **186**, 154–177.

BIOSKETCH

Carsten F. Dormann is an ecologist interested in the identification of drivers of species distribution at the landscape level. His focus is on methodological issues of statistical analyses and the investigation of causal mechanisms using field experiments and ecological modelling.

Editor: José Alexandre F. Diniz-Filho