Comment

Ecography 32: 379–381, 2009
doi: 10.1111/j.1600-0587.2009.05907.x

Response to Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review"

Carsten F. Dormann

C. F. Dormann (carsten.dormann@ufz.de), Dept of Computational Landscape Ecology, Helmholtz Centre for Environmental Research UFZ, Permoserstr. 15, DE-04318 Leipzig, Germany.

I was amused to read, in a criticism of another publication, that the criticised authors did not adhere to the "principles of ecological niche modelling" (McNyset and Blackburn 2006, p. 782). There aren't any such "principles". Ecological niche modelling really is statistics on species distribution data. All methods employed need to be statistically consistent, i.e. meet the assumptions made. (It would be a nice bonus if the analysis would also make ecological sense.) In the case discussed below, the crucial assumption is independence of data points, and it is usually violated in spatial distribution data, by spatial autocorrelation.

Autologistic regression (Augustin et al. 1996), and its generalisation the autocovariate regression, has up to now been the most popular way to deal with spatial autocorrelation in biogeographical analysis of species distribution (Dormann 2007b). In two recent papers analysing simulated data, this method has, however, performed worse than other methods (Dormann 2007a, Dormann et al. 2007), calling into question the validity of its use. In a comment, Betts et al. (2009, this issue) warn that these conclusions may be incorrect, and that autocovariate regression can be used when applied and interpreted properly. In the following paragraphs, I like to draw attention to two issues relevant in this debate: the use of the word "space", and the ecological interpretation of the values of the autocovariate (and other measures of aggregation or separation from spatial models in general), i.e. what the data tell us about the species.

I was seriously confused, when I realised that the autocovariate regression as applied in biogeography may be biased. In our review (Dormann et al. 2007), autocovariate regression was the only method that yielded a consistent bias in the parameter estimation, an observation confirmed by further simulations specifically looking at autologistic regression (Dormann 2007a). However, as Betts et al. pointed out correctly, all these analyses were carried out on the same underlying spatial structure, and if the environmental variable determining the virtual species' distribution (called "rain") was confounded with its spatial aggregation (what Betts et al. termed "collinearity of environment with space"), all these results may be fundamentally flawed. Since several previous studies have shown autologistic regression to be a reliable method (Wu and Huffer 1997, Huffer and Wu 1998, Hoeting et al. 2000, Hooten et al. 2003, Wintle and Bardos 2006), this explanation seems reasonable. So, why don't I simply concede my error and let everyone use autocovariate regression in peace?

There are several lines of argument that indicate that my results may not, after all, be wrong. A brief list: 1) environment and "space" are not correlated, unless we accept Betts et al.'s definition of "space". 2) Confirmation for the method comes only from studies that use an iterative implementation and missing data (see the studies cited above). 3) The idea of the autocovariate has an intrinsic circularity (since the new explanatory variable is constructed from the data to be explained: Dormann 2007a). 4) Two other methods using variables to represent spatial effects in a somewhat comparable way to the autocovariate are unbiased under the simulated conditions (spatial eigenvector mapping SEVM: Griffith and Peres-Neto 2006, Dormann et al. 2007, and a new wavelet method: Carl and Kühn 2007, Carl et al. 2008). Here, I shall only address the first topic, which I regard as most important to the differences between my view and that of Betts et al. (2009).

Spatial models attempt to correct for the non-independence of data points near each other, which may be connected through ecological and non-ecological processes (yielding spatial autocorrelation). To do so, they usually work through some quantification of spatial distances between sites (even if the model formula uses locations, i.e. longitude and latitude as inputs). Location, as one might have naively suspected, is thus not the same as "space" in the sense of spatial models. To avoid confusion, I prefer to use "neighbourhood" instead of "space" to refer to distance-related influences, irrespective of their location. The effect of a point's neighbourhood on this point can thus be referred to as the effect of "space" (in the words of Betts et al.) or of the neighbourhood. How can this be

measured? Betts et al. use the autocovariate to quantify this neighbourhood effect, and find it highly correlated with the environmental variable (''rain''). That is, in my view, problematic. When simulating the data, I had the following scenario in mind: the environment (''rain'') determines habitat suitability and hence occurrence probability. However, aggregation between individuals (e.g. by limited dispersal or colonial breeding) increases occurrences near presences, and decreases occurrences near absences (the effect of neighbourhood). Hence, the realised occurrence pattern is the result of both, habitat suitability and neighbourhood effects. The autocovariate, calculated from the occurrence pattern around the focal point, incorporates both of these effects, it is a compounded variable. Hence, I do not regard the autocovariate as a suitable measure of neighbourhood effects, as Betts et al. do: ''including aggregation in statistical models directly via autocovariates allows researchers to uncover, and further investigate such mechanisms'', (p.). A better measure of neighbourhood effects is the spatial error I used for simulating the data – which is not correlated with ''rain'' (Table 1). Furthermore, the autocovariate and the spatial error are uncorrelated (Table 1), also illustrating that the autocovariate encompasses more than only the neighbourhood effect.

So, where does that leave us? While I have to concede that I was negligent of the issue of collinearity of environment and neighbourhood when simulating the data, luckily they were not collinear, as claimed by Betts et al. What is indeed collinear is the autocovariate and environment, but I disagree that the autocovariate is a valid representation of neighbourhood effects. Because the autocovariate was constructed from the response, all processes affecting the response will also contribute to the autocovariate (e.g. environment, aggregation/separation, non-random noise such as observer bias, etc.).

I do not fully comprehend why MCMC implementations of the autocovariate method are apparently unbiased (Wu and Huffer 1997, Huffer and Wu 1998, Hoeting et al. 2000, Hooten et al. 2003). One crucial aspect is that in these cases the autocovariate is not estimated from occurrence data and then used as a new variable, but that the estimation of the autocovariate is part of the iterative model building process due to many missing data, for which the autocovariate has to be estimated. As such, it will be (initially) calculated from model residuals, which are (slightly) less correlated with the environment (Table 1). Thus, it may only be the inappropriate one-step-autocovariate that is in my view biased, which, however, is the standard way the autocovariate approach is applied in biogeographical research.

Table 1. Correlation coefficients (Pearson's r, n = 10) ± 1 standard error, and mean level of significance, for the environmental variable rain, spatial error (err), the autocovariate (in this case calculated with a size-2-neighbourhood and inverse weighting: ac) and an autocovariate calculated from the residuals of the regression on rain (rac). The residual autocovariate (rac) is reminiscent of the first step of an iterative procedure for calculating the autocovariate used, e.g. by Augustin et al. (1996). Although not very different from ac, a model with rain and rac is much less biased than a model containing rain and ac (coefficient for rain: true value: −0.002; ordinary GLM: −0.00215 ± 0.00038; ac: −0.00043 ± 0.00005; rac: −0.00294 ± 0.00047). Data are the binary occurrences in the ''snouter'' dataset of Dormann et al. (2007).

	err	ac	rac
rain	$0.137 \pm 0.038^{n.s.}$	$0.686 \pm 0.018^{***}$	$0.628 \pm 0.011^{***}$
err		$0.122 \pm 0.061^{n.s.}$	$0.115 \pm 0.065^{n.s.}$
ac			$0.922 \pm 0.018^{***}$

The second topic I would like to briefly touch upon is the ecological meaning of the autocovariate. Ecologically, spatial models contain information that non-spatial models cannot: an estimation of the range of aggregational or segregational processes. The ''range'' given in generalised least square models and, even more striking, the maps of spatial eigenvectors or spatial wavelets depict nicely at which spatial resolution clustering appears (Diniz-Filho and Bini 2005, Carl et al. 2008). Let us briefly reflect on what these ranges and maps actually mean (see also the discussion in Diniz-Filho and Bini 2005): any important (and in itself usually autocorrelated) explanatory environmental variable omitted from the model may cause spatial autocorrelation (due to model misspecification: Haining 2003). Also, incorrect description of the functional relationship can cause spatial autocorrelation (e.g. representing a non-linear effect only by a linear term in the model), and there may be various other mechanisms of model misspecification (Hastie et al. 2001, Dormann 2007c). Thus, the range of GLS and the maps of SEVM describe the combination of omitted variables and spatial ecological processes. That is the primary reason, why I have yet to see a study that convincingly extracts a meaningful ecological scale parameter from a species distribution analysis (which so far have been carried out on real data with necessarily unknown ''true'' parameters). While the studies of Betts et al. (2006, 2008) and Bourque and Desrochers (2006) are excellent examples of the importance of aggregational processes at the landscape scale, they do not show that quantitative information of the autocovariate is reliable. I do not share the optimism of Betts et al. that ''including aggregation in statistical models directly via autocovariates allows researchers to uncover, and further investigate such mechanisms'', at least not beyond a qualitative level. One study that claims to be able to do so uses four different specifications of the autocovariate, and a combination of abundance and occurrence data: only the rather intricate way in which these four analyses are interpreted in concert allows the identification of the underlying aggregation mechanism (van Teeffelen and Ovaskainen 2007); there is a reason why this study has a question mark at the end of its title: ''Can the cause of aggregation be inferred from species distributions?'' The jury is still out.

In conclusion, my questioning of the validity of the autocovariate regression has sparked criticism, which I can only partly quench. Betts et al. may well be right in their underlying criticism that these specific simulations were confounding the autocovariate approach with other problems, even though the claimed collinearity of environment and ''space'' is not existent. Also, while the simulated data may be flawed in various aspects, they may well be representative of real world data, where environmental variables are highly correlated with latitude and longitude. Given the other critical features of the autocovariate approach (listed in the fourth paragraph), I think the burden of proof now falls onto the proponents of this approach. They need to illustrate how we should use

autocovariate regression to yield as unbiased estimates as possible, and provide evidence that this method can give ecologically meaningful information on the spatial scale of ecological processes, ideally using simulated data. Since most analyses are carried out on atlas data, proposed methods need to be applicable to this type of data, too.

Acknowledgements – This work was supported by the Helmholtz Association (VH-NG 247). Many thanks to Brendan Wintle for checking this reply for ''glaring ambiguities'' and to Matthew Betts for constructive comments on an earlier version.

References

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

Betts, M. G. et al. 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. – Ecol. Model. 191: 197–224.

Betts, M. G. et al. 2008. Social information trumps vegetation structure in breeding-site selection by a migrant songbird. – Proc. R. Soc. B 275: 2257–2263.

Betts, M. G. et al. 2009. Comment on ''Methods to account for spatial autocorrelation in the analysis of species distributional data: a review''. – Ecography 32: 374–378.

Bourque, J. and Desrochers, A. 2006. Spatial aggregation of forest songbird territories and possible implications for area sensitivity. – Avian Conserv. Ecol. 1: 3, <www.ace-eco.org/vol1/iss2/art3/>.

Carl, G. and Kühn, I. 2007. Analyzing spatial ecological data using linear regression and wavelet analysis. – Stoch. Environ. Res. Risk Assess. 22: 315–324.

Carl, G. et al. 2008. A wavelet-based method to remove spatial autocorrelation in the analysis of species distributional data. – Web Ecol. 8: 22–29.

Diniz-Filho, J. A. and Bini, L. M. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. – Global Ecol. Biogeogr. 14: 177–185.

Dormann, C. F. 2007a. Assessing the validity of autologistic regression. – Ecol. Model. 207: 234–242.

Dormann, C. F. 2007b. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. – Global Ecol. Biogeogr. 16: 129–138.

Dormann, C. F. 2007c. Promising the future? Global change predictions of species distributions. – Basic Appl. Ecol. 8: 387–397.

Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.

Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses in exploiting relative location information. – Ecology 87: 2603–2613.

Haining, R. 2003. Spatial data analysis – theory and practice. – Cambridge Univ. Press.

Hastie, T. et al. 2001. The elements of statistical learning: data mining, inference, and prediction. – Springer.

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

Hooten, M. B. et al. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. – Landscape Ecol. 18: 487–502.

Huffer, F. W. and Wu, H. L. 1998. Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. – Biometrics 54: 509–524.

McNyset, K. M. and Blackburn, J. K. 2006. Does GARP really fail miserably? A response to Stockman et al. (2006). – Divers. Distrib. 12: 782–786.

van Teeffelen, A. J. A. and Ovaskainen, O. 2007. Can the cause of aggregation be inferred from species distributions? – Oikos 116: 4–16.

Wintle, B. A. and Bardos, D. C. 2006. Modeling species-habitat relationships with spatially autocorrelated observation data. – Ecol. Appl. 16: 1945–1958.

Wu, H. L. and Huffer, F. W. 1997. Modelling the distribution of plant species using the autologistic regression model. – Environ. Ecol. Stat. 4: 49–64.