



## Comment on “Methods to account for spatial autocorrelation in the analysis of species distributional data: a review”

Matthew G. Betts, Lisa M. Ganio, Manuela M. P. Huso, Nicholas A. Som, Falk Huettmann, Jeff Bowman and Brendan A. Wintle

*M. G. Betts (Matthew.Betts@oregonstate.edu), L. M. Ganio, M. M. P. Huso and N. A. Som, Dept of Forest Ecosystems and Society, Oregon State Univ., Corvallis, OR 97331, USA. – F. Huettmann, Biology and Wildlife Dept, Inst. of Arctic Biology, Univ. of Alaska-Fairbanks, Fairbanks, AK 99775-7000, USA. – J. Bowman, Ontario Ministry of Natural Resources, Wildlife Research and Development Section, 2140 East Bank Drive, Peterborough, ON K9J 7B8, Canada. – B. A. Wintle, School of Botany, Environmental Science, The Univ. of Melbourne, Victoria, Australia.*

In a recent paper, Dormann et al. (2007) (hereafter Dormann et al.) conducted a review of approaches to account for spatial autocorrelation in species distribution models. As the review was the first of its kind in the ecological literature it has the potential to be an important and influential source of information guiding research. Although many spatial autocovariance approaches may seem redundant in the spatial processes they reflect, seemingly subtle differences in approach can have major implications for the resulting description of the data and conclusions drawn. Though Dormann et al.'s review of the available approaches was a step in the right direction, we think that their simulation study ignored important concepts which leads us to question some of their conclusions.

One of Dormann et al.'s primary conclusions was that parameter estimates for most spatial modeling techniques were not strongly biased except in the case of autocovariate models. In the autocovariate model, as implemented by Dormann et al., the parameter representing the effect of environmental variables on species distributions (the coefficient for rain) was consistently underestimated. For this reason Dormann et al. cautioned the use of autocovariate approaches. This caution reiterated findings from a similar simulation in which (Dormann 2007) argued that autocovariate logistic regression models used for binomially distributed data (autologistic models) would be biased and unreliable. These results appear to be in direct contrast to earlier evaluations of this method (Augustin et al. 1996, Hoeting et al. 2000, He et al. 2003) and need to be considered seriously, as autocovariate approaches are now widely used in ecology (Piorecky and Prescott 2006, Wintle and Bardos 2006, McPherson and Jetz 2007, van Teeffelen and Ovaskainen 2007, Miller et al. 2007); for instance the seminal paper on autologistic regression (Augustin et al. 1996) has now been cited 222 times (Web of Science accessed 8 September 2008). Simplified implementation

and interpretation of these models may result in misleading conclusions.

Our critique is on three grounds. First, we show that the change Dormann et al. observed in the parameter estimate between the autocovariate approach and the true value is due to multicollinearity between environment and space. Variation shared among parameters is a common occurrence in ecological models and can rarely be avoided (Graham 2003); however, it can be directly measured using hierarchical partitioning approaches (Chevan and Sutherland 1991). Second, there are situations in which autocovariate approaches offer the opportunity to incorporate effects of behavioural and population processes into ecological models. This may result in greater understanding of these processes even though interpretation of the estimated coefficients themselves may not be possible. Third, we highlight that statistical regression models are developed for different objectives than outlined by Dormann et al. In particular, the goal of predicting future or non-sampled observations invites a very different model-building strategy than the goal of interpretation of model coefficients (Hastie and Tibshirani 1990). Because of this, we argue that comparison of models with different objectives should not be limited to an evaluation of only bias. We show that the autocovariate approach can be a useful model if minimizing prediction error is the objective. For brevity, in this paper we focus on autologistic regression for Bernoulli distributed data however, we believe our arguments are applicable to autocovariate methods used with Poisson and normally distributed data.

### Multicollinearity of space and environment

Dormann et al. generated artificial distribution data in which a hypothetical species was positively influenced by rainfall. The authors also simulated spatially correlated

errors. The realization of this data generation process was a species distributed as a function of only rainfall and space; this simulation could be thought of as reflecting the realistic scenario that a species is influenced by both the environment and some sort of aggregative process (e.g. dispersal limitation, conspecific attraction; see below). Examination of a map of Dormann et al.'s simulated data clearly reveals a species that is clustered in space (Fig. 1). However, because rainfall itself is positively spatially autocorrelated (Fig. 2), there is overlap in the effects of environmental and aggregative processes on species clustering.

Autocovariate models include a covariate ( $autocov_i$ ) to model the influence of  $k_i$  neighbors at a distance ( $h_{ij}$ ) from a focal site  $i$ :

$$autocov_i = \frac{\sum_{j=1}^{k_i} w_{ij} y_j}{\sum_{j=1}^{k_i} w_{ij}}$$

The autocovariate,  $autocov_i$ , is a weighted average of  $k$  values in the neighbourhood of cell  $i$ . The weight given to any neighbouring point  $j$  is  $w_{ij} = 1/h_{ij}$  where  $h_{ij}$  is (usually) the Euclidean distance between points  $i$  and  $j$ . If the species is present at point  $j$  then  $y_j = 1$ , otherwise  $y_j = 0$  (Augustin et al. 1996). This covariate is added to a generalized linear model (glm) to account for the variation explained by space. In this case, the observed data are the presence or absence of the species,  $Y_i$  which is Bernoulli distributed with a mean  $\rho_i$ . Then the glm is:

$$\begin{aligned} \text{logit}(\rho_i) &= \ln(\rho_i/1 - \rho_i) \\ &= \beta_0 + \beta_1 \text{rainfall}_i + \beta_2 \text{autocov}_i \end{aligned}$$

Where  $\beta_0$  is the model intercept,  $\beta_1$  and  $\beta_2$  are parameter estimates for rainfall and the autocovariate respectively, and  $\text{rainfall}_i$  and  $\text{autocov}_i$  are the values of predictor variables at the  $i$ th site.



Figure 1. One of ten spatial distributions of the hypothetical species generated by Dormann et al. (2007). This distribution shows strong spatial aggregation in the species that is due to both the environment (in this case rainfall) and spatial processes. Black and gray shaded points are species presences and absences respectively.

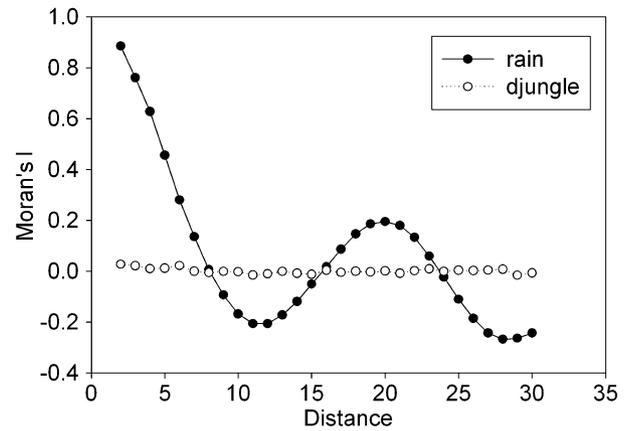


Figure 2. Degree of spatial autocorrelation, as measured by Moran's I, in two environmental variables simulated by Dormann et al. (2007). Distance is measured as number of cells.

In the case of the Dormann et al. data, we expected some of the variation in species presence to be shared by rainfall and the autocovariate. To test this hypothesis, we used the hierarchical partitioning method (Whittaker 1984, Chevan and Sutherland 1991, Lawler and Edwards 2006) to estimate the amount of deviance that is: a) explained independently by the environmental variable (rainfall;  $R_I$ ) or b) independently by space (the autocovariate;  $A_I$ ), c) jointly explained by both variables ( $R_I + A_I$ ), and d) explained by rain in a simple regression model ( $R_T$ ). As expected, over the 10 datasets simulated by Dormann et al., the proportion of the deviance explained by rain that was shared by the autocovariate was large ( $([R_I + A_I]/R_T) \times 100 = 56 \pm 43\%$  SD). In contrast, only  $6 \pm 1\%$  of the explained deviance in species distribution could be independently attributed to rainfall ( $(R_I/\text{Total explained}) \times 100$ ; Fig. 3).

If two predictor variables in the same model overlap in their contribution to the model, coefficients of both variables may change radically in comparison to the case

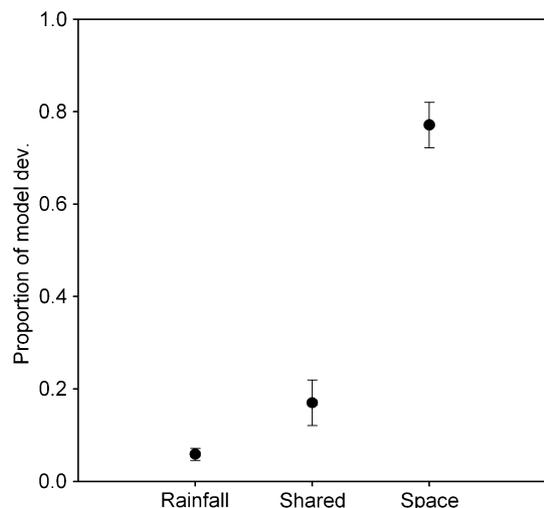


Figure 3. The proportion of variation explained independently by rain and space and explained jointly by both variables (Shared) across the ten datasets simulated by Dormann et al. (2007). Error bars show SE.

where each variable occurs on its own in a simple regression model (Wonnacott and Wonnacott 1981). The absolute magnitude of partial regression coefficients increases with increasing collinearity (Petraitis et al. 1996). The reason for the bias observed by Dormann et al. was that information about species occurrence was shared by the autocovariate and rainfall. The spatial aggregation effect simulated by Dormann et al. was large and correlated with rainfall so that the presence of the autocovariate in the model changed the estimated coefficient for the effect of rainfall on species occurrence.

To demonstrate this point further we present two tests. First, if our argument is true we expect there to be a negative correlation between the proportion of deviance shared by rainfall and space ( $R_J + A_J$ ) and the degree to which the coefficient for rainfall changes as a function of including or excluding the autocovariate. Using the 10 datasets simulated by Dormann et al., we found this pattern to be strongly supported ( $r = -0.95$ ,  $p < 0.0001$ ) (Fig. 4). If there was little overlap in the deviance explained by rainfall and space, coefficients for rain did not change with the addition of the autocovariate. Large multicollinearity corresponded to shrinkage of rain coefficients by a factor of  $\sim 3$ .

Second, we predicted that if there is, on average, no correlation between space and an environmental variable, there should be no apparent change to the environmental variable parameter estimate. Using the methods of Dormann et al., we simulated binary response data to construct 10 datasets. Rather than using rainfall as the “true” predictor, in this instance we used the variable “djungle” also presented in Dormann et al. Djungle itself is not spatially autocorrelated (Fig. 2). We simulated data so that occurrence of the hypothetical species was positively associated ( $\beta = 0.24$ ) with djungle. As in Dormann et al. we added normally distributed spatially-autocorrelated error to the logit of the response variable. In this case parameter estimates for the explanatory variable

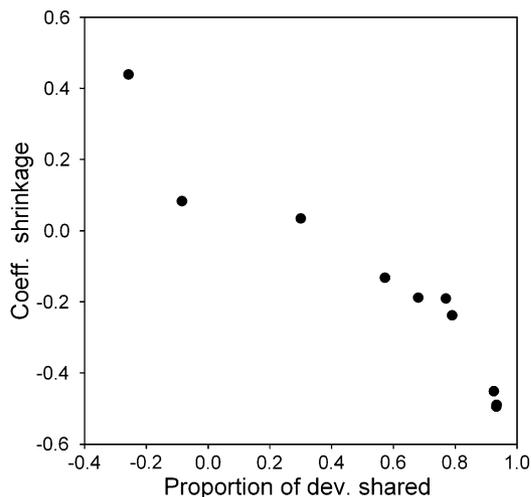


Figure 4. Relationship between the proportion of variation in species distribution explained by rain that is shared with space (see text), and the degree to which coefficients of rain shrink from the logistic to autologistic model (calculated as:  $\log(\hat{\beta}(\text{auto covariate})/\hat{\beta}(\text{glm}))$ ). Negative values thus represent coefficient shrinkage and positive values coefficient expansion.

from our simulation were higher than expected (mean  $\hat{\beta} = 0.292 \pm 0.06$  [SE]). This contrasts sharply with the results of Dormann et al. who found coefficients of the spatially autocorrelated predictor variable, rainfall, to shrink by a factor of five ( $-0.003/-0.0006$ ; p. 618). By changing only the spatial structure of the explanatory variable alone we completely reversed the results reported by Dormann et al. In this case, the increased value of the environmental coefficient was due to the fact that the jointly explained deviance for the autocovariate and djungle was negative ( $-10 \pm 1\%$  SE).

It is important to discuss why the eight other methods tested by Dormann et al. to account for spatial autocorrelation in binary data do not exhibit the same apparent bias in parameter estimates. As noted by Dormann et al., only autocovariate regression and spatial eigenvector mapping (SEVM) methods account for spatial autocorrelation via additional explanatory variables. None of the other investigated methods include spatial structure as fixed explanatory variables; thus it is not possible to confound environmental variables with spatial structure in the mean as these components exist in separate parts of the model (Kissling and Carl 2008). Not surprisingly, the SEVM approach also suffers from the potential for space-environment confounding (Griffith and Peres-Neto 2006). However, this multicollinearity can apparently be resolved via extracting the eigenfunctions of the matrix  $[I-H]C[I-H]$  where C and I are the connectivity and identity matrices as described by Dormann et al. and H is the common hat matrix (Myers 1990). It is not clear from the text if Dormann et al. utilized this procedure in the analysis of their simulated data. Nevertheless, SEVM appears to offer some promise for avoiding problems of multicollinearity while retaining fixed explanatory variables to account for spatial autocorrelation.

In sum, simulations presented by Dormann et al. revealed a change in the coefficient estimate caused by adding an autocovariate to a logistic regression model. This change is no more of a problem however, than it is in any other statistical model with multiple explanatory variables that are collinear to some degree (Wonnacott and Wonnacott 1981). In the case of Dormann et al., the coefficient change was particularly severe because data were simulated in such a way as to result in high collinearity between the autocovariate and the environmental covariate. Such collinearity is not uncommon in nature so the simulation was not unrealistic, but it makes it impossible to attribute “cause” to spatial vs environmental variables (Hawkins et al. 2007). Of course, without manipulative experiments and/or data simulations, it is impossible to attribute cause in any case. When variables are confounded (i.e. there is jointly shared information) researchers can only quantify this (using some form of hierarchical partitioning) and design a future experiment that explicitly disentangles the effects of space and environment.

### The importance of behavioural and population processes

A wide range of processes govern the distribution of species, many of which are not directly related to environmental

variables. Such processes include dispersal limitation, inter- and intra-specific competition, evolutionary history, territoriality, and conspecific attraction. Dormann et al. correctly point out that spatial autocorrelation can thus be seen as an opportunity as well as a challenge (Legendre 1993). However, the authors attributed little discussion to how the various models and methods reviewed can help to address questions in behavioural and population ecology.

Autocovariate methods explicitly include aggregation in predictive models by adding additional covariates (see second formula above). That is, autocovariates occur in models as fixed effects that change the mean value of the response. Other models (e.g. GLMM) deal with aggregation by including specific sources of variation (autocorrelation) in a random component of the model for the mean response, but do not generally change the estimated mean. By including an autocovariate, i.e. an “endogenous” source of spatial autocorrelation (Currie 2007), it is possible to identify the potential presence of aggregative processes that are operating to affect species distributions (Fig. 3). Unfortunately, because coefficients may not be directly interpretable in instances of multicollinearity, direct estimation of the degree to which aggregative processes occur is limited. However, as noted above, hierarchical partitioning approaches offer promise for uncovering instances where substantial independent variation is explained by space. For example, in recent years, two separate studies conducted in different regions of North America reported high degrees of fine-scale spatial autocorrelation in the abundance of a neotropical migrant warbler, the black-throated blue warbler (Bourque and Desrochers 2006, Betts et al. 2006). In both studies this autocorrelation was hypothesized to be due to conspecific attraction, perhaps as a result of use of social cues in habitat selection. These correlative results have now been confirmed experimentally (Hahn and Silverman 2007, Betts et al. 2008). Including aggregation in statistical models directly via autocovariates allows researchers to uncover, and further investigate such important mechanisms.

### Regression prediction versus coefficient estimation

Models developed for prediction may include covariates whose functional link to the response is not obvious but which are excellent predictor variables. Quality coefficient estimation and quality prediction do not necessarily coincide, and researchers focused on either aspect should be keenly aware of what metrics should be emphasized in their regression analysis (Myers 1990, p. 133). Coefficient bias is an appropriate metric for analyses whose goal is to accurately describe biological relationships. However, when the goal is accurate prediction, minimizing prediction error is most important (Myers 1990). Dormann et al. chose bias in a parameter estimate to compare models whose objectives may not be unbiased parameter estimates, but might rather be prediction success, as in the autocovariate model.

Including spatial autocorrelation in models directly as autocovariates is thus not only of interest to behavioural and population ecologists but is likely to be useful for species distribution modelers (Segurado et al. 2006). Prediction success of species distribution models may improve with the

inclusion of spatial variables because they explicitly measure endogenous sources of spatial autocorrelation. In a simulation study, Wintle and Bardos (2006) found that autologistic models had better fit and better predictive performance than logistic models under a range of sampling conditions. Indeed, using Dormann et al.’s simulated data, we calculated the area under the receiver operating characteristic curve (AUC), a measure of prediction success (Manel et al. 2001) for two models: a) rain (environmental variable only), and b) rain + autocovariate (environmental variable and space). The mean AUC for the autologistic models was significantly higher than for environment only models (autologistic:  $85 \pm 1$  SE, environment only:  $66 \pm 3$  SE,  $t=6.32$ ,  $p < 0.001$ ). This is consistent with the findings of Augustin et al. (1996) who argued that autocovariate models were best suited for prediction in ecology rather than necessarily being useful for inference.

Interpolation to new environments (Bahn and McGill 2007) can be achieved with the autologistic model if one utilizes a Gibbs sampling estimation and prediction method (Augustin et al. 1996, Wintle and Bardos 2006). Such improved prediction success has been found in subsequent studies using autocovariate approaches (Hoeting et al. 2000, Osborne et al. 2001, He et al. 2003, Knapp et al. 2003, Duff and Morrell 2007, McPherson and Jetz 2007). However, because such methods tend to be computer intensive, many studies that use autocovariates do not use the Gibbs sampler making it difficult to predict independent data (Betts et al. 2006). Species distribution modeling would benefit greatly from the development of a “user friendly” interface for calculating MCMC-generated autocovariates. Such algorithms are becoming more accessible to ecologists (Wintle and Bardos 2006, McPherson and Jetz 2007).

In summary we argue that the apparent bias caused by autocovariate approaches reported by Dormann et al. is just due to shared explained variation between the environmental and spatial variables. Dormann et al.’s parameter estimates were changed owing to multicollinear variables. Such joint contributions to explained variation are likely to occur frequently in nature as many environmental variables are correlated. However, this problem in autocovariate models was recognized early on by its developers who cautioned against its use for inference (Augustin et al. 1996). Thus, if the research objective is increasing prediction success autocovariate approaches are a viable option. If the primary objective is parameter estimation, other models (e.g. GLMM) that include space as a random effect may be more appropriate.

### References

- Augustin, N. H. et al. 1996. An autologistic model for the spatial distribution of wildlife. – *J. Appl. Ecol.* 33: 339–347.
- Bahn, V. and McGill, B. J. 2007. Can niche-based distribution models outperform spatial interpolation? – *Global Ecol. Biogeogr.* 16: 733–742.
- 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.
- Bourque, J. and Desrochers, A. 2006. Spatial aggregation of forest songbird territories and possible implications for area sensitivity. – *Avian Conserv. Ecol.* 1: 3.
- Chevan, A. and Sutherland, M. 1991. Hierarchical partitioning. – *Am. Stat.* 45: 90–96.
- Currie, D. J. 2007. Disentangling the roles of environment and space in ecology. – *J. Biogeogr.* 43: 2009–2011.
- Dormann, C. F. 2007. Assessing the validity of autologistic regression. – *Ecol. Model.* 207: 234–242.
- 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.
- Duff, A. A. and Morrell, T. E. 2007. Predictive occurrence models for bat species in California. – *J. Wildl. Manage.* 71: 693–700.
- Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. – *Ecology* 84: 2809–2815.
- 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.
- Hahn, B. A. and Silverman, E. D. 2007. Managing breeding forest songbirds with conspecific song playbacks. – *Anim. Conserv.* 10: 436–441.
- Hastie, T. J. and Tibshirani, R. J. 1990. *Generalized additive models*, 1st ed. – Chapman and Hall.
- Hawkins, B. A. et al. 2007. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. – *Ecography* 30: 375–384.
- He, F. L. et al. 2003. Autologistic regression model for the distribution of vegetation. – *J. Agric. Biol. Environ. Stat.* 8: 205–222.
- Hoeting, J. A. et al. 2000. An improved model for spatially correlated binary responses. – *J. Agric. Biol. Environ. Stat.* 5: 102–114.
- Kissling, W. D. and Carl, G. 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. – *Global Ecol. Biogeogr.* 17: 59–71.
- Knapp, R. A. et al. 2003. Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. – *Ecol. Appl.* 13: 1069–1082.
- Lawler, J. J. and Edwards, T. C. 2006. A variance-decomposition approach to investigating multiscale habitat associations. – *Condor* 108: 47–58.
- Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – *Ecology* 74: 1659–1673.
- Manel, S. et al. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. – *J. Appl. Ecol.* 38: 921–931.
- McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. – *Ecography* 30: 135–151.
- Miller, J. et al. 2007. Incorporating spatial dependence in predictive vegetation models. – *Ecol. Model.* 202: 225–242.
- Myers, R. H. 1990. *Classical and modern regression with applications*. – PWS-Kent Publ. Company.
- Osborne, P. E. et al. 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. – *J. Appl. Ecol.* 38: 458–471.
- Petraitis, P. S. et al. 1996. Inferring multiple causality: the limitations of path analysis. – *Funct. Ecol.* 10: 421–431.
- Piorecky, M. D. and Prescott, D. R. C. 2006. Multiple spatial scale logistic and autologistic habitat selection models for northern pygmy owls, along the eastern slopes of Alberta's Rocky Mountains. – *Biol. Conserv.* 129: 360–371.
- Segurado, P. et al. 2006. Consequences of spatial autocorrelation for niche-based models. – *J. Appl. Ecol.* 43: 433–444.
- van Teeffelen, A. J. A. and Ovaskainen, O. 2007. Can the cause of aggregation be inferred from species distributions? – *Oikos* 116: 4–16.
- Whittaker, J. 1984. Model interpretation from the additive elements of the likelihood function. – *Appl. Stat.* 33: 52–64.
- Wintle, B. A. and Bardos, D. C. 2006. Modeling species-habitat relationships with spatially autocorrelated observation data. – *Ecol. Appl.* 16: 1945–1958.
- Wonnacott, T. H. and Wonnacott, R. J. 1981. *Regression: a second course in statistics*, 1st ed. – Wiley.