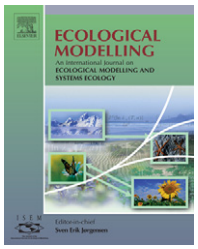
ECOLOGICAL MODELLING 207 (2007) 234–242

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Assessing the validity of autologistic regression

*Carsten F. Dormann**

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

ARTICLE INFO

Article history:
Received 11 July 2006
Received in revised form
30 April 2007
Accepted 7 May 2007
Published on line 20 June 2007

Keywords:
Presence–absence data
Spatial autocorrelation
Species distribution analysis

ABSTRACT

In autologistic regression models employed in the analysis of species' spatial distributions, an additional explanatory variable, the autocovariate, is used to correct the effect of spatial autocorrelation. The values of the autocovariate depend on the values of the response variable in the neighbourhood. While this approach has been widely used over the last ten years in biogeographical analyses, it has not been assessed for its validity and performance against artificial simulation data with known properties. I here present such an assessment, varying the range and strength of spatial autocorrelation in the data as well as the prevalence of the focal species. Autologistic regression models consistently underestimate the effect of the environmental variable in the model and give biased estimates compared to a non-spatial logistic regression. A comparison with other methods available for the correction of spatial autocorrelation shows that autologistic regression is more biased and less reliable and hence should be used only in concert with other reference methods.

1. Introduction

In 1996, Augustin et al. introduced a new approach to addressing the problem of spatial autocorrelation in species distribution data, which they termed autologistic regression (see also Gumpertz et al., 1997). This approach, detailed below, has quickly gained popularity among ecologists, as it provides an easy way to address a fundamental problem intrinsic to spatial data. Examples for its application include modelling the distribution of plant species (Wu and Huffer, 1997; Huffer and Wu, 1998; He et al., 2003; Boll et al., 2005), insects (Gumpertz et al., 2000; Dennis et al., 2002; Brownstein et al., 2003), amphibians (Knapp et al., 2003), birds (Osborne et al., 2001; Silva et al., 2002; Selmi et al., 2003; Mörtberg and Karlström, 2005; Betts et al., 2006; Piorecky and Prescott, 2006) and mammals (Mattson and Merrill, 2002; Edenius et al., 2003; Teterukovskiy and Edenius, 2003). Outside ecology, autologistic regression has been used in dental medicine (Kirkham et al., 2005), image analysis and remote sensing (Arbia et al., 1999; Koutsias, 2003) and manufacturing of integrated circuits (Ramirez and Taam, 2000). In a recent review, 8 out of 21 studies that compared spatial and non-spatial models used autologistic regression (Dormann, 2007).

Taking a closer look at the origin of autologistic regression shows that this method has been derived from statistical background (Besag, 1974), but recently divided into two different branches: that of auto-models more generally (which also included autoregressive models, Haining, 2003) and autologistic regression itself, mainly with applications in biogeography. It is this latter version of autologistic models that I address in this study, and which I assess for their validity. Validity, in this case, refers to the quality of parameter estimates, i.e. the usefulness of autologistic regression for statistical inference.

Two observations led me to investigate the general applicability of autologistic regression: Since autologistic regression is based on an *explanatory* variable being constructed from thenon-independent values of a *response* variable one might

* Tel.: +49 341 2353946; fax: +49 341 2353939.
E-mail address: carsten.dormann@ufz.de.
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.05.002

suspect circularity in the method. Moreover, autologistic regression as a method has not been tested on independent artificial data sets with known properties (but see Hoeting et al., 2000 for an assessment of model fit with artificial data). Addressing the latter deficit in this study, I want to explore the robustness and performance of the autologistic regression based on simulated data.

2. What is 'autologistic' regression?

The idea behind autologistic regression is to calculate an extra explanatory variable that captures the effect of other response values in the spatial neighbourhood. Neighbouring values are expected to be similar to the focal value for ecological reasons such as dispersal, which will lead to higher abundance of offspring near to the parent organism, or aggregative behaviour, leading, e.g., to colonial breeding (Danchin et al., 2004).

Auto-models were initially developed for the analysis of plant competition experiments (Mead, 1971; Ord, 1975) and then extended to spatial data in general (Besag, 1974). These auto-models fit a statistical model that explains the focal response (e.g. plant yield or plant presence) by the response values of its neighbours. Initially, only first-order neighbours in lattice systems were the focus, but this constraint was later relaxed. The resulting, general formulation of an auto-model for a binary response variable would then be (Besag, 1974):

$$\ln \frac{\pi_i}{1-\pi_i} = \alpha + \sum_j \beta_{ji} y_j + \varepsilon_i, \quad (1)$$

where π_i is the probability of presence at location i, α is the model intercept and $\beta_{i,j}$ is the coefficient that relates the observed occurrence y at location j to the predicted probability in i, and ε_i is the binomially distributed error.

This is a *basic autologistic* model as used in many graphic imaging and remote sensing routines to identify and correct missing or faulty data points (e.g. Arbia et al., 1999). In this general formulation, several different βs are fitted, one for each neighbour (i.e. four in the case of a first-order neighbourhood). All β_js become identical if isotropy is assumed. When used to correct missing or faulty data, the β_js are intially estimated based on available data, then the fitted values are compared to the observed, and outliers or missing points are replaced by the fitted values. The procedure is then repeated with the new values, until convergence. Many statistical publications deal with efficient ways to estimate the parameters for the final model, using pseudo-likelihood and various Markov Chain–Monte Carlo methods (see, e.g. Geyer and Thompson, 1992; Besag and Green, 1993; Wu and Huffer, 1997; Huffer and Wu, 1998; Pettitt et al., 2003; Friel and Pettitt, 2004).

More important to ecologists is the next step, namely to incorporate explanatory variables, such as climate variables in the case of species distributions. This step expands the model from a *basic autologistic model* to an *autologistic model with covariates*. Then, the above formula (assuming isotropy) becomes:

$$\ln \frac{\pi_i}{1-\pi_i} = \alpha + \beta s(y_i) + \sum_k \gamma_k x_{ki} + \varepsilon_i, \quad (2)$$

where $s(y_i)$ is a function that summarises the y-values in the neighbourhood of i (e.g. distance-weighted, see below) and x_{ki} are the values for k different environmental variables used to explain the occurrence of y in i. Although computationally difficult, this model can be fitted by maximum likelihood (Sherman et al., 2006).

Computation of the spatial term $s(y_i)$, known as the autocovariate, is straightforward in lattice data without missing values. This spatial term, the autocovariate, is calculated only once from the observed data and not changed during the model estimation process. Augustin et al. (1996) proposed to compute this term as a distance-weighted average of all cells j in the neighbourhood:

$$ac_i = s(y_i) = \frac{\sum_{j \neq i} w_j y_j}{\sum w_j}, \quad \text{with } w_j = d_{ij}^{-l}. \quad (3)$$

The data point i itself is excluded from the calculation. It is important to note here that there is no iterative recalculation of the spatial term $s(y_i)$ involved, in contrast to the above-mentioned applications in image analysis or data sets with incomplete observations (see, e.g., Huffer and Wu, 1998; Hoeting et al., 2000).

Distance-weighting can be of different forms, i.e. different values for exponent l in Eq. (3). Most commonly, l is set to zero, (giving equal weight to all neighbouring cells), unity (inverse distance-weighted) or two (inverse squared). As an alternative to distance-weighting, angle-weighting has been used to model bark-beetle attacks (Preisler, 1993). The size of the neighbourhood is found by examining several and choosing the size that reduces residual spatial autocorrelation most (e.g. Sanderson et al., 2005).

The autocovariate thus derived is then added to the model as additional explanatory variable (Eq. (2)). If one is also interested in the model estimates for the intercept, then the autocovariate should be standardised before incorporation into the model (i.e. subtraction of the mean and division by standard deviation, yielding an autocovariate with mean 0 and standard deviation 1). In principle, there are no constraints on the type of model to use the autocovariate with, but to date mainly generalised linear models (GLMs: e.g. Augustin et al., 1996, 1998; Betts et al., 2006; Boll et al., 2005), generalised additive models (GAMs: Knapp et al., 2003; Segurado and Araújo, 2004) and Bayesian frameworks (Hoeting et al., 2000; Osborne et al., 2001; Riiali et al., 2001; He et al., 2003; Friel and Pettitt, 2004; Reese et al., 2005) have been used with autocovariates. Many spatial models in a Bayesian framework are representations of a first-order autologistic regression (e.g. Högmander and Möller, 1995; Hoeting et al., 2000; Haining, 2003; He et al., 2003; Sherman et al., 2006). The main reasons for a Bayesian framework are additional complications such as unknown detection probabilities, unobserved sites, recorder bias, etc. (Latimer et al., 2006). The above described autologistic regression approach can be seen as the first iteration of an iterative fitting processes correcting for missing data and will yield estimates that are robust for regular lattice data (Augustin et al., 1996).

In the following I will only discuss the autologistic model in its simplest, and most commonly applied, form, i.e. according to Eq. (2), with an optimised estimate of $s(y_i)$ and a single parameter estimation for the autologistic covariate model.

3. Methods

I used two steps to assess autologistic regression models. In a first step, I analysed if autocovariates may produce spurious results, i.e. compensate for spatial autocorrelation although it is absent (leading to a type II error). The second step is based on realisations of data sets with known properties and spatial autocorrelation in the error. Thereby I test how well autologistic regression estimates the true parameters used to generate the distribution data.

3.1. *Autologistic regression on random data without spatial autocorrelation*

I generated a spatial realisation of 2500 cells of value 0 in a square lattice and randomly assigned a value of 1 to 25% of these cells. Next, I constructed an autocovariate for this data set, with neighbourhood sizes ranging from 1 to 30 cells. I then ran 30 *basic autologistic* models on these data (one for each neighbourhood size) and selected the one which reduced residual spatial autocorrelation most. This whole procedure was replicated 1000 times. Estimated coefficients at the link-scale were tested against 0 using a *t*-test.

3.2. *Autologistic regression on spatially autocorrelated data*

3.2.1. *Generating artificial distribution data*

This simulation analysis aimed to investigate the consistency of autologistic regression. To this end, I manipulated four factors that may affect regression analysis and then compared the performance of autologistic regression with ordinary logistic regression. The four factors are: (1) the strength of spatial autocorrelation, (2) the range of spatial autocorrelation, (3) prevalence and (4) the strength of the relationship between the response and the predictors, by adding unstructured noise.

The artificial data shall represent some virtual species inhabiting an island, and whose distribution is entirely dependent on precipitation. The size of the island (i.e. the number of cells) is 1108. Fig. 1 illustrates one example for the distribution data analysed. I will refer to the explanatory variable as 'rain' and an additional, spurious explanatory variable as 'djungle'. The data were constructed using a pre-defined functional relationship between the response variable (*y*) and a predictor variable: $E(y) = g^{-1}(3 - 0.003\text{rain})$, where $g(\cdot)$ is the logistic link function.

Spatial autocorrelation was introduced to the data as spatially autocorrelated errors. The process (detailed in below) entails three steps: first, the distance between all grid cells was calculated ("distance matrix"). This forms the basis for a second matrix ($\vec{\Omega}$), in which the distance is non-linearly transformed (to reflect the decay of importance with distance). Finally, the contribution of each point to the error at each other point is calculated by means of a weights matrix derived from

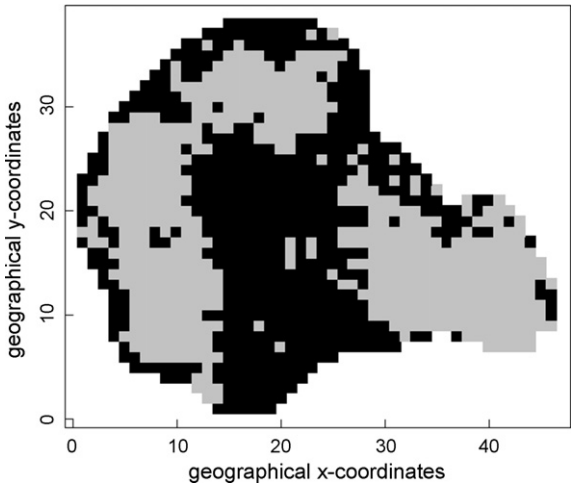


Fig. 1 – Example for a distribution map used in the simulations. Black represents presence, grey absence. Parameters for this simulation are: strength = 0.5, range = 0.625, prevalence = 0.5 and noise = off. The correlogram for this species is depicted in Fig. 2, solid line.

$\vec{\Omega}$. These errors are then simply added (which, in mathematical terms, is a matrix multiplication of a non-correlated error vector with the weights matrix).

More specifically, a weights matrix $\vec{W}$ was used to weight random noise according to the distance between data points. Let $\vec{D} = (d_{ij})$ be the (Euclidean) distance matrix of distances between the cells i and j. Then $\vec{\Omega} = (\omega_{ij})$ is a matrix defined as $\omega = e^{-\text{range } d_{ij}}$, with 'range' as a parameter controlling the range of spatial autocorrelation (see below). I can now calculate the weights matrix $\vec{W}$ (by Choleski decomposition) from $\vec{\Omega} = \vec{W}^{T}\vec{W}$. Hence the error in the above relationship between $\vec{y}$ and 'rain' is $\vec{\varepsilon} = \vec{W}\vec{\xi}$, with $\vec{\xi} \sim N(0, \sigma)$. The thereby produced error term $\vec{\varepsilon}$ was used to generate an occurrence probability p_i for each location i: $p_i = q_i + q_i(1 - q_i)\varepsilon_i$. This was then translated into binary responses y_i so that $(1/N)\sum_i^N y_i = \text{prevalence}$.

The strength of spatial autocorrelation was manipulated by the standard deviation of the normal distribution (σ) from which the error $\vec{\xi}$ was drawn. This error vector, $\vec{\varepsilon}$, was then re-scaled to standard deviation of 1, eliminating different absolute values of the error term but maintaining the degree of spatial similarity. Large values for the standard deviation meant that little of the error was contributed by spatially autocorrelated error, while low standard deviation led to a high level of spatial autocorrelation in the error. Strength, in the form of standard deviation of the normal distribution that the errors were sampled from, was varied over six steps (low to high strength of spatial autocorrelation: 8, 4, 2, 1, 0.5, 0.25).

The range of spatial autocorrelation *r*, i.e. the distance over which spatial autocorrelation is detectable, was also varied in six steps, from short- to far-distance autocorrelation. Spatial autocorrelation (SAC) decreased exponentially with distance d according to the function $\text{SAC} \sim e^{-\text{range } d}$. Range took the values 0.0625, 0.125, 0.25, 0.5, 1 and 2. The larger the value, the shorter the range of spatial autocorrelation.

To simulate models for differently common species, I altered the prevalence in five steps (0.05, 0.1, 0.25, 0.5, 0.75).

I thereby evaluated if autologistic regression is equally valid for rare and common species.

Finally, to see how the strength of the environment–species-relationship affects the autologistic regression, I also varied the degree of determination by adding unstructured noise. For model runs with unstructured noise, I added random values from a normal distribution with mean = 0 and standard deviation = 0.5.

All four factors (strength, range, prevalence and noise) were combined in a factorial experiment, yielding $6 \times 6 \times 5 \times 2 = 360$ different species distribution realisations. Each realisation yielded a value for each cell. Thresholds of which values to code as 0 or 1 were selected according to the prevalence value for this realisation.

3.2.2. *Analysing simulated distributions*

Each realisation was analysed with a non-spatial model (i.e. a generalised linear model with binomial error containing 'rain' and 'djungle' as explanatory variables) and a set of spatial models. Autocovariates for 30 different ranges (from 1 to 30 cells distance) were calculated. An autologistic regression, i.e. a GLM plus the autocovariate, was performed for all 30 spatial models separately. Then, the residuals of all these spatial models were compared and the one with the lowest spatial autocorrelation was selected.

Spatial autocorrelation was calculated as Global Moran's I (Fortin and Dale, 2005) up to a distance of 20. For computational reasons, the selection of the best performing neighbourhood was based on the area under the correlogram of residuals (until distance class 20), rather than Global Moran's I.

Based on the residuals of the non-spatial model, I calculated the realised strength and range of spatial autocorrelation, which differed from the intended due to the random noise in the realisations and the way the different factors interacted with one another (e.g. range and strength). Range was calculated as that distance class where the correlogram intercepted with the *x*-axis. Strength was quantified by the Global Moran's I index up to a distance of 20. These realised strengths and ranges were then used in the analysis of coefficient estimates, rather than the parameters used to set up the artificial data.

All simulations and analyses were carried out using the software R, Version 2.3.0 (Team, 2004), with the additional packages spdep, ncf, splancs and MASS.

4. Results

4.1. *Random pattern and autologistic regression*

Estimates for the autocovariate were close to, but significantly different from, the expected value of 0 (approximately normally distributed with mean ±S.D. = -0.0296 ± 0.062; *t*-test: $P < 0.001$; intercept = -1.100 ± 0.0017, $P < 0.001$). This led to a consistent underestimation of the intercept (i.e. the true mean) of the occupancy. Moreover, the estimate for the intercept was strongly negatively skewed (skewness = −2.98, kurtosis = 12.4), indicating that in several occasions it was very biased.

In all 1000 runs the estimate for the intercept (i.e. the mean) from the autologistic regression was lower than that of the non-spatial, true model (ranging from 0.247 to the true 0.250). Since the bias was usually small (albeit consistent), the deviance attributable to the autocovariate was always below 1%.

4.2. *Autologistic regression on artificial distribution data*

4.2.1. *Removal of residual spatial autocorrelation*

Autocovariate regression effectively removed spatial autocorrelation in the residuals (Fig. 2), independent of the parameter sets used to create the correlative signature. In contrast, residuals from the non-spatial model show essentially the same pattern as the raw data (data not shown).

4.2.2. *Parameter estimation*

Most important in the assessment of the validity of autocovariate regression is the ability of this method to correctly estimate the true parameter for the explanatory variable 'rain' in the artificial data.

Range, prevalence and unstructured noise impacted the difference between spatial and non-spatial models (interactions significant: see Table 1).

Model estimates for 'rain' differed greatly between the spatial and the non-spatial model (Fig. 3). Short-range spatial autocorrelation led to equally poor estimation of the rain-coefficient. As range increases, the non-spatial model approaches the true value of −0.003, while the 'rain'-effect of the spatial model moved closer to 0. Short-range spatial autocorrelation affected the non-spatial model particularly strongly, while this effect petered out as the range increases. However, in all cases was the coefficient estimate of the spatial model worse than that of the non-spatial.

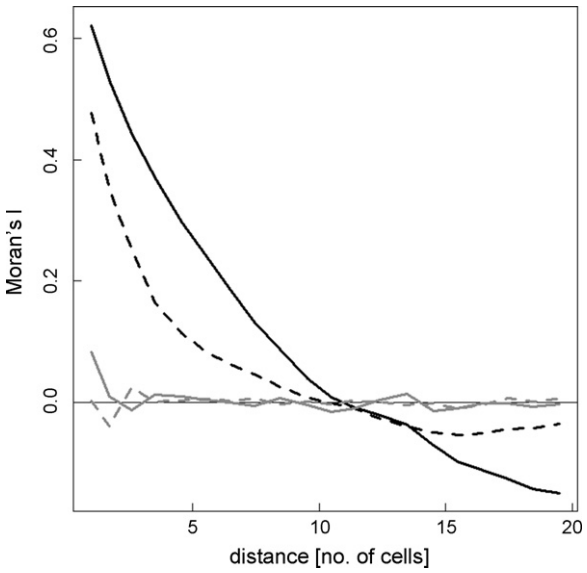


Fig. 2 – Effect of autocovariate on spatial autocorrelation in the residuals. Black lines are correlograms from non-spatial residuals, grey lines from autocovariate regression. Solid and hatched line differ in strength but not in range of spatial autocorrelation.

Table 1 – Analysis of variance of the influences on estimating the coefficient of 'rain'. 'Spatial' codes for the difference between spatial and non-spatial models; 'Moran' refers to the strength of spatial autocorrelation. Interactions with 'spatial' point at effects specifically influencing autologistic regression results

Term	d.f.	SS	F	P
Spatial	1	35.4E−05	512.00	<0.001
Range	1	0.76E−05	11.10	<0.001
Moran	1	0.86E−05	12.50	<0.001
Noise	1	1.57E−05	22.70	<0.001
Prevalence	1	4.76E−05	68.90	<0.001
Spatial:range	1	2.32E−05	33.60	<0.001
Spatial:noise	1	0.85E−05	2.20	<0.001
Spatial:prevalence	1	2.28E−05	32.90	<0.001
Moran:noise	1	0.38E−05	5.44	<0.05
Moran:prevalence	1	1.28E−05	18.50	<0.001
Residuals	709	49.0E−05		

A similar pattern emerged for the effect of prevalence (Fig. 4): spatial models consistently underestimated the true effect of 'rain'. There was little effect of prevalence on these estimations. Non-spatial models were severely impacted by the prevalence of species, only estimating the effect of 'rain' correctly for very high prevalences.

Changing the strength of the environment–response-relationship by adding unstructured noise altered parameter estimates for 'rain' only marginally in the autocovariate regression models (Fig. 5). Non-spatial estimates were further off the truth when unstructured noise was added. These results show that the expectation of worse estimates in noisier data only holds for the non-spatial model. The autocovariate regression was able to capture some of the additional noise.

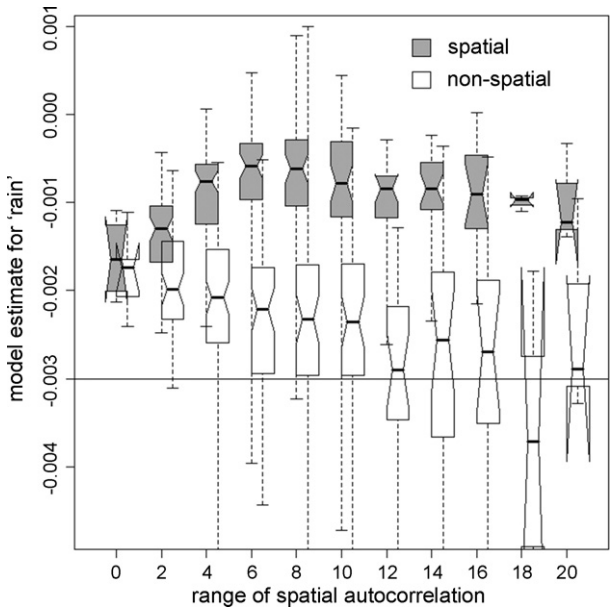


Fig. 3 – Coefficient estimates for 'rain' for spatial and non-spatial models, depending on the range of spatial autocorrelation across the 360 simulations. Line at y = −0.003 marks the true value of the coefficient. Boxes give first and third quartile, whiskers extend to the entire range of values. Non-overlapping notches are indicative of significant differences of the median (solid line in the boxes). Non-spatial models were set off slightly for better graphical representation.

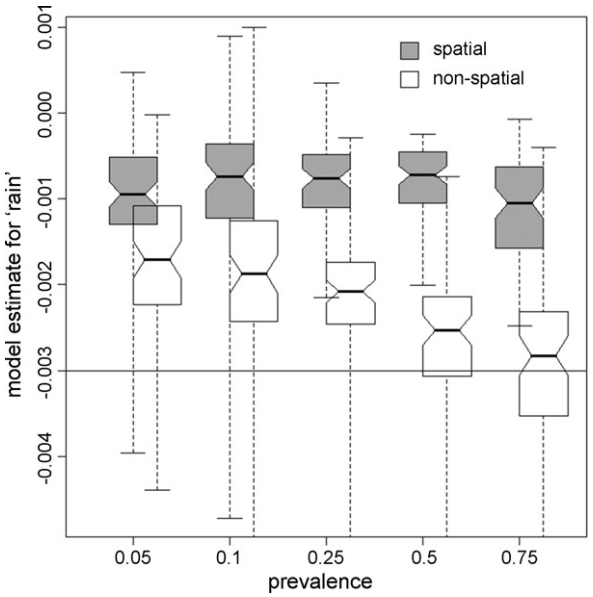


Fig. 4 – Coefficient estimates for 'rain' for spatial and non-spatial models, depending on the prevalence of a species. Line at y = −0.003 marks the 'true' coefficient. See Fig. 3 for further details.

In addition to the above-mentioned interactions with the spatial/non-spatial model approach, several other effects were significantly affecting the model for 'rain'-estimates (Table 1). Since they impacted the spatial as well as the non-spatial model, I will not explore them at length. It may suffice to mention that the strength of spatial autocorrelation (effect 'Moran') significantly effected parameter estimation, with

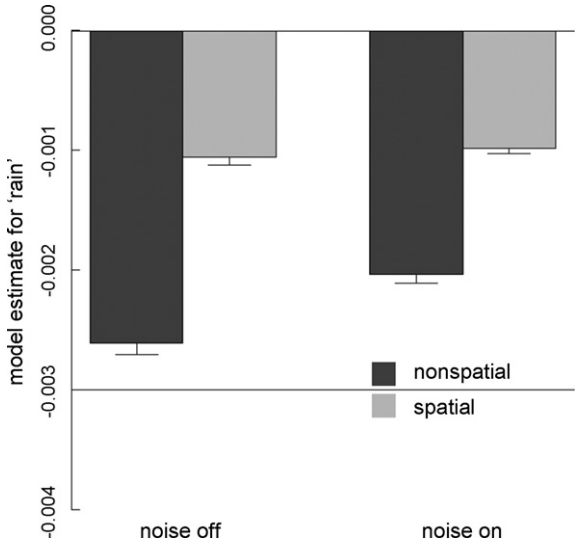


Fig. 5 – Coefficient estimate for 'rain' for spatial and non-spatial models, depending on the additional presence of spatially uncorrelated noise. Line at y = −0.003 marks the 'true' coefficient. Error bars represent ±1 standard error.

stronger autocorrelation yielding better (i.e. closer to the truth) estimates for 'rain'. This effect is slightly reduced when additional noise is added (Moran:noise interaction).

Also important is the error attached to the coefficient estimate. Since spatial autocorrelation leads to non-independent observations and hence to inflated degrees of freedom, a method correcting for SAC should have larger standard errors for the coefficient estimates than a non-spatial model. Across the 360 simulations, non-spatial standard errors for 'rain' were always lower than those of the autocovariate model (mean = 3.31×10^{-4} versus 4.26×10^{-4}). The same holds true for the errors on 'djungle'-estimates (1.56×10^{-2} versus 1.85×10^{-2}). As a consequence, the spurious effect of 'djungle' was more often significant ($P < 0.05$) in the non-spatial model than in the spatial one (24 versus 16 cases, which is 6.7% versus 4.4%). Thus, the non-spatial models wrongly detected an effect of 'djungle' slightly more often than would be expected by chance alone (i.e. 5%). The flip-side of the better performance of the autologistic regression with respect to spurious effects is that it is also more conservative with respect to true effects. In the 360 simulations, the non-spatial model detected a 'rain'-effect in 337 cases, compared to only 217 for the autologistic regression (i.e. false negatives in 10% versus 40% of all simulations).

The analysis of the estimates for 'djungle' showed that these estimates were not influenced by the type of model. Rather, species prevalence and the interaction of range and noise were significantly affecting the estimates for 'djungle'. Since also the overall fit of the ANOVA-model was rather poor (adjusted $R^2 = 0.025$), this analysis was not revealing with respect to the focus of the study.

5. Discussion

5.1. *Parameter estimates based on autologistic regression*

The results of this study show that autologistic regression leads to a consistent underestimation of the true parameters. This finding is corroborated by the regression coefficients reported by Klute et al. (2002), who compared autologistic regression to a non-spatial GLM when analysing woodcock occurrences. Their six regression parameters were all lower (on average by 30%) in the autologistic version of the model. Obviously, it is impossible to guess how much of this is a true effect due to reduced spatial autocorrelation in the residuals or due to the bias that the autocovariate introduced. Without giving details on the coefficients, also Segurado et al. (2006) report on a slight 'overcompensation' effect of their autocovariate. However, in two different studies the opposite effect was observed: for two bird species (*Capercaillie* and *Hazel grouse*) the coefficients to the four environmental variables in the respective autologistic models were higher (Mörtberg and Karlström, 2005), while for blackbird occurrences in African oasis the first principle component axis received higher weight in the autologistic model (Selmi et al., 2003). Again, since the true values are unknown it is impossible to judge whether the spatial model is actually better. Finally, in the multi-species study by Chou and Soret (1996) the final models differed considerably as a consequence of incorporating an autocovariate. Thus while the impact on parameter estimates was consistent with the present study (slight reduction in absolute parameter values), their study additionally shows how the autocovariate affects model selection.

The way spatial autocorrelation influences the accuracy of coefficient estimates (i.e. standard errors) has implications for statistical inference and model selection. I found that the non-spatial models identified the true effect of rain in over 90% of cases, while the autologistic model only about 60%. Not only would we thus infer that rain was unimportant in a large number of data sets, but also would this effect be deleted when employing model simplification (Crawley, 2002). A predictive model based on this analysis would hence not contain the truly driving variable. On the other hand, based on the non-spatial model one would regard 'djungle' as relevant in about 7% of cases, slightly more than the 5% expected from the probability threshold of 0.05. I hence conclude that the wrongful omission of 'rain' in the spatial model is a more serious issue than the wrongful inclusion of the spurious 'djungle'-effect in the non-spatial model.

Spatial autocorrelation causes the problem of inflated degrees of freedom: because data are not independent of each other, but are treated that way in non-spatial models, the are less effective degrees of freedom (Legendre, 1993). After an analysis of spatial data, residuals should always be checked for spatial autocorrelation; if spatial autocorrelation in the residuals is detected, spatial models should be used to avoid spatial pseudo-replication (Fortin and Dale, 2005; Dormann, 2007). A recent study showed that coefficients may be inverted by spatial autocorrelation (Kühn, 2007), demonstrating the need for spatial models.

Although care was taken to create artificial data in a controlled way, the data analysed in this study will not be universally representative of species distribution data (Hoeting et al., 2000). This potential shortcoming does not invalidate the assessment in the present analysis, since it does show in any case that the basic autologistic model employed so frequently in the ecological literature *may* be biased. As we have no known true parameters under real-world conditions, this study at least casts doubt on its universal validity. More extensive, and perhaps more realistic, distribution patterns could address this issue. In particular, real data analyses often suffer from the problem that not all 'true' determinants of the response variable are quantifiable. It is unknown, how spatial autocorrelation resulting from spatial dependence on an unknown explanatory variable may affect autologistic regression. The difference to the present study is that the spatial interdependence is not only in the error term, but also in some latent variables. However, the 360 different data sets employed here represent a fair range of distribution patterns not explicable as random number artefacts.

5.2. *Ecological information in autologistic regression?*

It has been argued (e.g. Sanderson et al., 2005) that the optimum neighbourhood size of the autocovariate may be a useful indicator of the range of the ecological process causing spatial autocorrelation (such as dispersal, homerange size, etc.). I investigated this idea with the simulations: There was only a

weak and moreover negative correlation between the realised range of spatial autocorrelation in the artificial data and the optimum neighbourhood size for the autocovariate (across all 360 simulations: Pearson's $r = -0.33$, $P < 0.001$; correlation with the intended range was even weaker). This indicates that the mechanism producing the spatial autocorrelation was not captured by the autocovariate and inference from the optimum neighbourhood size of the autocovariate to the spatial dimension of the underlying mechanism may be poor also in real data.

Van Teeffelen and Ovaskainen (2007), in a recent paper, used autologistic regression to analyse the causes of aggregation pattern in simulated data. They calculated autocovariates in different ways and found that they were affected differently for different causes of aggregation. Their study shows at least that the range of best-fitting autologistic models was in many cases not matching the spatial scales of aggregation processes. Only by comparing different model approaches were they able to recover the simulation process from the data.

5.3. *Alternative methods to correct for spatial autocorrelation in binary lattice data*

In a recent overview, Dormann et al. (2007) review several alternative methods to correct for spatial autocorrelation in ecological data. For binary data such as here, four other methods were mentioned in addition to autologistic regression (sorted in ascending order of computer runtime): general estimating equations (GEE: Carl and Kühn, 2007), mapping of eigenvectors (ME: Dray et al., 2006; Griffith and Peres-Neto, 2006; Tiefelsdorf and Griffith, 2007), spatial generalised linear mixed models (GLMM: Breslow and Clayton, 1993; Gotway and Stroup, 1997) and spatial Bayesian models (e.g. Latimer et al., 2006). For comparison with the autologistic results I employed the first three (Bayesian models using Gaussian random fields did not converge on the artificial data: S. Bierman, personal communication).

As Fig. 6 shows, the estimates for 'rain' and 'djungle' differ substantially between the different models. It shows an obvious shift towards 0 for the autologistic regression and an opposite trend for ME. ME also produced the by far most variable parameter estimates. GLMM, GEE and the non-spatial model differed only marginally, with the spatial GLMM and GEE slightly less prone to underestimating the effect of 'rain'. Estimates for 'djungle' were virtually identical. Overall, these results indicate that none of the four spatial models (autologistic, ME, GEE and GLMM) provide a consistently improved estimate over the non-spatial model for these specific data.

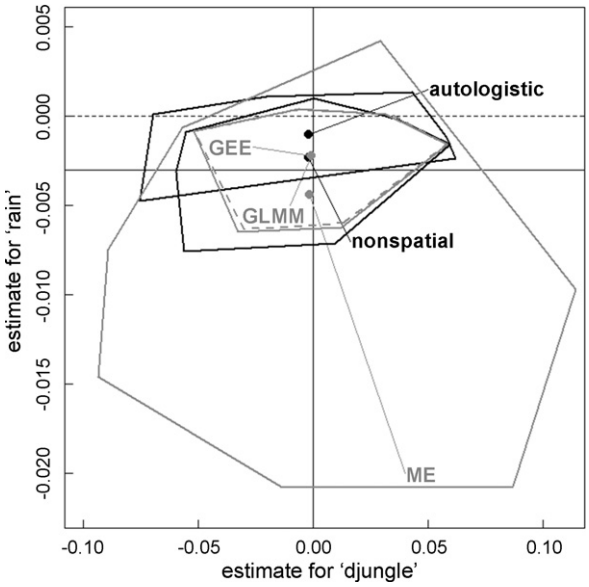


Fig. 6 – Estimates for the coefficients of 'rain' and 'djungle' across the 360 different simulations. The polygon depicts the convex hull (all estimates are within this polygon) and the dot represent the grand mean for each method (point of gravity of the polygon). GEE and GLMM share the same point of gravity. Dashed line refers to GEE, solid to GLMM. ME refers to mapping of spatial eigenvectors as implemented in R.

6. Conclusion

This study shows that basic autologistic regression, as employed in the ecological literature on species distribution analyses, suffers some severe drawbacks, most importantly potentially strongly biased estimation of model parameters. Since alternative methods to correct for spatial autocorrelation are available, autologistic regression should only be used when backed-up by another method, to ensure that parameter estimates are not influenced unduly.

Acknowledgments

Many thanks to Jana McPherson for raising and discussing some of the points addressed in this paper, to Gudrun Carl for helping me to generate spatially autocorrelated data, and to Stijn Bierman, Gudrun Carl, Jana McPherson, Boris Schröder and an anonymous reviewer for comments on an earlier version. This research has been partly funded by the Helmholtz Association (VH-NG-247).

REFERENCES

Arbia, G., Benedetti, R., Espa, G., 1999. Contextual classification in image analysis: an assessment of accuracy of ICM. Comput. Stat. Data Anal. 30, 443–455.

Augustin, N., Mugglestone, M., Buckland, S., 1996. An autologistic model for the spatial distribution of wildlife. J. Appl. Ecol. 33, 339–347.

Augustin, N., Mugglestone, M., Buckland, S., 1998. The role of simulation in modelling spatially correlated data. Environmetrics 9, 175–196.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. B 36, 192–236.

Besag, J., Green, P.J., 1993. Spatial statistics and Bayesian computation. J. R. Stat. Soc. B 55, 25–38.

Betts, M.G., Diamond, A.W., Forbes, G.J., Villard, M.-A., Gunn, J., 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. Ecol. Model. 191, 197–224.

Boll, T., Svenning, J.C., Vormisto, J., Normand, S., Grandez, C., Balslev, H., 2005. Spatial distribution and environmental

preferences of the Piassaba palm *Aphandra natalia* (Arecaceae) along the Pastaza and Urituyacu rivers in Peru. Forest Ecol. Manage. 213, 175–183.
Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88, 9–25.
Brownstein, J.S., Holford, T.R., Fish, D., 2003. A climate-based model predicts the spatial distribution of the Lyme disease vector *Ixodes scapularis* in the United States. Environ. Health Perspect. 111, 1152–1157.
Carl, G., Kühn, I., 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. Ecol. Model., 207, 159–170.
Chou, Y.H., Soret, S., 1996. Neighborhood effects in bird distributions, Navarre, Spain. Environ. Manage. 20, 675–687.
Crawley, M., 2002. Statistical Computing: An Introduction to Data Analysis using S-Plus. Wiley, New York.
Danchin, T., Giraldeau, L.-A., Valone, T.J., Wagner, R.H., 2004. Public information: from nosy neighbors to cultural evolution. Science 305, 487–491.
Dennis, R., Shreeve, T., Sparks, T., Lhonore, J., 2002. A comparison of geographical and neighbourhood models for improving atlas databases. The case of the French butterfly atlas. Biol. Conserv. 108, 143–159.
Dormann, C.F., 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecol. Biogeogr. 16, 129–138.
Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davis, R., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, submitted.
Dray, S., Legendre, P., Peres-Neto, P., 2006. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecol. Model. 196, 483–493.
Edenius, L., Vencatasawmy, C.P., Sandstrom, P., Dahlberg, U., 2003. Combining satellite imagery and ancillary data to map snowbed vegetation important to reindeer *Rangifer tarandus*. Arct. Antarct. Alpine Res. 35, 150–157.
Fortin, M.J., Dale, M.R.T., 2005. Spatial Analysis—A Guide for Ecologists. Cambridge University Press, Cambridge.
Friel, N., Pettitt, A.N., 2004. Likelihood estimation and inference for the autologistic model. J. Comput. Graph. Stat. 13, 232–246.
Geyer, C.J., Thompson, E.A., 1992. Constrained Monte Carlo maximum-likelihood for dependent data. J. R. Stat. Soc. B: Meth. 54, 657–699.
Gotway, C.A., Stroup, W.W., 1997. A generalized linear model approach to spatial data analysis and prediction. J. Agric. Biol. Environ. Stat. 2, 157–178.
Griffith, D.A., Peres-Neto, P.R., 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses in exploiting relative location information. Ecology 87, 2603–2613.
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. J. Agric. Biol. Environ. Stat. 2, 131–156.
Gumpertz, M.L., Wu, C.-T., Pye, J.M., 2000. Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. Forest Sci. 46, 95–107.
Haining, R., 2003. Spatial Data Analysis—Theory and Practice. Cambridge University Press, Cambridge.
He, F.L., Zhou, J., Zhu, H.T., 2003. Autologistic regression model for the distribution of vegetation. J. Agric. Biol. Environ. Stat. 8, 205–222.
Hoeting, J.A., Leecaster, M., Bowden, D., 2000. An improved model for spatially correlated binary responses. J. Agric. Biol. Environ. Stat. 5, 102–114.
Högmander, H., Möller, J., 1995. Estimating distribution maps from atlas data using methods of statistical image analysis. Biometrics 51, 393–404.
Huffer, F.W., Wu, H.L., 1998. Markov Chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Biometrics 54, 509–524.
Kirkham, J., Kaur, R., Stillman, E.C., Blackwell, P.G., Elcock, C., Brook, A.H., 2005. The patterning of hypodontia in a group of young adults in Sheffield, UK. Arch. Oral Biol. 50, 287–291.
Klute, D.S., Lovallo, M.J., Tzilkowski, M., 2002. Autologistic regression modeling of American woodcock habitat use with spatially dependent data. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.B., Samson, F. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, pp. 335–343.
Knapp, R.A., Matthews, K.R., Preisler, H.K., Jellison, R., 2003. Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. Ecol. Appl. 13, 1069–1082.
Koutsias, N., 2003. An autologistic regression model for increasing the accuracy of burned surface mapping using landsat thematic mapper data. Int. J. Remote Sens. 24, 2199–2204.
Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Divers. Distrib. 13, 66–69.
Latimer, A.M., Wu, S., Gelfand, A.E., Silander, J.A., 2006. Building statistical models to analyze species distributions. Ecol. Appl. 16, 33–50.
Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659–1673.
Mattson, D.J., Merrill, T., 2002. Extirpations of grizzly bears in the contiguous United States, 1850–2000. Conserv. Biol. 16, 1123–1136.
Mead, R., 1971. Models for interplant competition in irregularly distributed populations. Stat. Ecol. 2, 13–32.
Mörtberg, U., Karlström, A., 2005. Predicting forest grouse distribution taking account of spatial autocorrelation. J. Nat. Conserv. 13, 147–159.
Ord, J.K., 1975. Estimation methods for models of spatial interaction. J. Am. Stat. Assoc. 70, 120–126.
Osborne, P., Alonso, J., Bryant, R., 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. J. Appl. Ecol. 38, 458–471.
Pettitt, A.N., Friel, N., Reeves, R., 2003. Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. J. R. Stat. Soc. B: Stat. Meth. 65, 235–246.
Piorecky, M.D., 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.
Preisler, H.K., 1993. Modeling spatial patterns of trees attacked by bark-beetles. Appl. Stat.: J. R. Stat. Soc. C 42, 501–514.
Ramirez, J., Taam, W., 2000. An autologistic model for integrated circuit manufacturing. J. Qual. Technol. 32, 254–262.
Reese, G.C., Wilson, K.R., Hoeting, J.A., Flather, C.H., 2005. Factors affecting species distribution predictions: a simulation modeling experiment. Ecol. Appl. 15, 554–564.
Riiali, A., Penttinen, A., Kuusinen, M., 2001. Bayesian mapping of lichens growing on trees. Biomet. J. 43, 717–736.
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.
Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modeling species distributions. J. Biogeogr. 31, 1555–1568.
Segurado, P., Araújo, M.B., Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. J. Appl. Ecol. 43, 433–444.

Selmi, S., Boulinier, T., Faivre, B., 2003. Distribution and abundance patterns of a newly colonizing species in Tunisia oases: the common blackbird *Turdus merula*. J. Appl. Ecol. 145, 681–688.

Sherman, M., Apanasovich, T.V., Carroll, R.J., 2006. On estimation in binary autologistic spatial models. J. Stat. Comput. Simul. 76, 167–179.

Silva, T., Reino, L.M., Borralho, R., 2002. A model for range expansion of an introduced species: the common waxbill *Estrilda astrild* in Portugal. Divers. Distrib. 8, 319–326.

Team, R.D.C., 2004. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org).

Teterukovskiy, A., Edenius, L., 2003. Effective field sampling for predicting the spatial distribution of reindeer (*Rangifer tarandus*) with help of the Gibbs sampler. Ambio 32, 568–572.

Tiefelsdorf, M., Griffith, D.A., 2007. Semi-parametric filtering of spatial autocorrelation: the eigenvector approach. Environ. Plan. A 39, 1193–1221.

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

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