Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2008) **17**, 59–71

RESEARCH PAPER

Spatial autocorrelation and the selection of simultaneous autoregressive models

W. Daniel Kissling[1,3]* and Gudrun Carl[2,3]

[1]*Community & Macroecology Group, Institute of Zoology, Department of Ecology, Johannes Gutenberg University of Mainz, D-55099 Mainz, Germany,* [2]*UFZ - Helmholtz Centre for Environmental Research, Department of Community Ecology, Theodor-Lieser-Str. 4, 06120 Halle, Germany,* [3]*Virtual Institute Macroecology, Theodor-Lieser-Str. 4, 06120 Halle, Germany*

*Correspondence: W. Daniel Kissling, Community & Macroecology Group, Institute of Zoology, Department of Ecology, Johannes Gutenberg University of Mainz, D-55099 Mainz, Germany.
E-mail: kissling@uni-mainz.de

ABSTRACT

Aim Spatial autocorrelation is a frequent phenomenon in ecological data and can affect estimates of model coefficients and inference from statistical models. Here, we test the performance of three different simultaneous autoregressive (SAR) model types (spatial error = SAR_{err}, lagged = SAR_{lag} and mixed = SAR_{mix}) and common ordinary least squares (OLS) regression when accounting for spatial autocorrelation in species distribution data using four artificial data sets with known (but different) spatial autocorrelation structures.

Methods We evaluate the performance of SAR models by examining spatial patterns in model residuals (with correlograms and residual maps), by comparing model parameter estimates with true values, and by assessing their type I error control with calibration curves. We calculate a total of 3240 SAR models and illustrate how the best models [in terms of minimum residual spatial autocorrelation (minRSA), maximum model fit (R^2), or Akaike information criterion (AIC)] can be identified using model selection procedures.

Results Our study shows that the performance of SAR models depends on model specification (i.e. model type, neighbourhood distance, coding styles of spatial weights matrices) and on the kind of spatial autocorrelation present. SAR model parameter estimates might not be more precise than those from OLS regressions in all cases. SAR_{err} models were the most reliable SAR models and performed well in all cases (independent of the kind of spatial autocorrelation induced and whether models were selected by minRSA, R^2 or AIC), whereas OLS, SAR_{lag} and SAR_{mix} models showed weak type I error control and/or unpredictable biases in parameter estimates.

Main conclusions SAR_{err} models are recommended for use when dealing with spatially autocorrelated species distribution data. SAR_{lag} and SAR_{mix} might not always give better estimates of model coefficients than OLS, and can thus generate bias. Other spatial modelling techniques should be assessed comprehensively to test their predictive performance and accuracy for biogeographical and macroecological research.

Keywords
Autoregressive process, biogeography, macroecology, model selection, neighbourhood structure, spatial model, spatial statistics, spatial weights, species richness.

INTRODUCTION

Spatial autocorrelation is a frequent phenomenon in ecological data because observations from nearby locations are often more similar than would be expected on a random basis (Legendre, 1993; Legendre & Legendre, 1998). This is especially true for species distribution data because they are inherently spatially structured (e.g. Jetz & Rahbek, 2002; Keitt *et al.*, 2002; Dark, 2004; Guisan *et al.*, 2006; Kissling *et al.*, 2007). Two types of spatial autocorrelation might be distinguished depending on whether endogenous or exogenous processes generate the spatial structure of species distributions (Legendre, 1993; Legendre & Legendre, 1998; Fortin & Dale, 2005). In the case of endogenous processes, the spatial pattern is generated by factors that are an inherent property of the variable itself ('inherent spatial autocorrelation'; Fortin & Dale, 2005), for instance distance-related

DOI: 10.1111/j.1466-8238.2007.00334.x
www.blackwellpublishing.com/geb

biotic processes such as reproduction, dispersal, speciation, extinction or geographical range extension (Legendre, 1993; Diniz-Filho *et al.*, 2003). On the other hand, spatial autocorrelation can be induced by exogenous processes that are independent of the variable of interest ('induced spatial dependence'; Fortin & Dale, 2005). These are most likely spatially structured environmental factors such as geomorphological processes, wind, energy input or climatic constraints, which can cause species distributions to be spatially structured (Legendre, 1993; Diniz-Filho *et al.*, 2003).

Irrespective of which processes cause the spatial structure of species distributions, the presence of spatial autocorrelation is problematic for classical statistical tests (ANOVA, correlation and regression) because these methods assume independently distributed errors (Legendre, 1993; Legendre & Legendre, 1998). The first problem relates to the inflation of type I errors, which means that confidence intervals are wrongly estimated when observations are not independent, and hence classical tests of significance of correlation or regression coefficients might be biased (Legendre, 1993; Lennon, 2000; Legendre *et al.*, 2002). The second problem applies to shifts in model coefficients between non-spatial and spatial regression models, which affects our ability to evaluate the importance of explanatory variables (Lennon, 2000; Lichstein *et al.*, 2002). This can be a serious shortcoming for hypothesis testing and inference from statistical models (Dormann, 2007) and might even invert the interpretation of environmental effects on species distributions (Kühn, 2007). One therefore needs to test for the presence of spatial autocorrelation in the residuals of regression models when modelling species distributions to evaluate whether type I errors and shifts in parameter estimates are likely to occur.

A number of methods exist to deal with spatial autocorrelation in ecological data (Cressie, 1993; Haining, 2003; Diniz-Filho & Bini, 2005; Fortin & Dale, 2005; Rangel *et al.*, 2006). One of these is spatial regression models, such as simultaneous autoregressive (SAR) models (Cressie, 1993; Haining, 2003), which augment the standard linear regression model with an additional term that incorporates the spatial autocorrelation structure of a given data set. This additional term is implemented with a 'spatial weights matrix' where the neighbourhood of each location (e.g. defined by distance) and the weight of each neighbour (e.g. closer neighbours might receive higher weights) need to be defined (e.g. Anselin & Bera, 1998; Fortin & Dale, 2005). The spatial dependence of a location on neighbouring sites is then modelled with a variance–covariance matrix based on the defined spatial weights matrix (for details see Cressie, 1993; Anselin, 1988, 2002; Anselin & Bera, 1998; Fortin & Dale, 2005). The spatial weights matrix in SAR models thus accounts for patterns in the response variable that are not predicted by explanatory variables, but are instead related to values in neighbouring locations.

Although SAR and other autoregressive models have been known for decades in the statistical literature (Besag, 1974; Cliff & Ord, 1981), their application in ecology and species distribution research has been limited up to now (e.g. Jetz & Rahbek, 2002; Keitt *et al.*, 2002; Lichstein *et al.*, 2002; Dark, 2004; Tognelli & Kelt, 2004; Kissling *et al.*, 2007). One reason might be that the implementation of autoregressive models is mathematically complex (Cressie, 1993) and computationally intensive (Rangel *et al.*, 2006), and freely available software packages have just recently become available (R Development Core Team, 2005; Rangel *et al.*, 2006). As a consequence, most applications of autoregressive models in ecology have so far restricted the range of available options to incorporate spatial interaction. For instance, most studies have not tested a variety of possible model specifications (e.g. different neighbourhood distances, model types or coding styles for the spatial weights matrix), nor have they systematically investigated their potential to account for spatial autocorrelation, including the precision of their parameter estimates. Moreover, model selection procedures, which allow the identification of a single best model or a set of models (Burnham & Anderson, 1998; Johnson & Omland, 2004), are largely absent for spatially autocorrelated data (see Hoeting *et al.*, 2006).

In this paper, we tested the potential of three different SAR model types (spatial error model, lagged model and mixed model) with 27 spatial weights matrices (based on nine neighbourhood distances and three different neighbourhood weights) to account for spatial autocorrelation in four artificial species distribution data sets with known spatial properties. All four data sets had the same relationship between the response variable and the two explanatory variables and only differed in the way that spatial autocorrelation was induced. This allowed us to systematically investigate the potential of SAR models to account for certain types of spatial autocorrelation structures, including the evaluation of the precision of parameter estimates and type I error controls. Moreover, we illustrate how the best SAR models can be selected from a range of model specifications using model selection procedures based on minimum residual spatial autocorrelation (minRSA), maximum model fit (R^2) and the Akaike information criterion (AIC). The construction and evaluation of SAR models was implemented with the free software R (R Development Core Team, 2005) to enable ecologists to freely use the methods presented here.

MATERIALS AND METHODS

Simultaneous autoregressive models

Simultaneous autoregressive models assume that the response at each location i is a function not only of the explanatory variable at i, but of the values of the response at neighbouring locations j as well (Cressie, 1993; Lichstein *et al.*, 2002; Haining, 2003). In SAR, the neighbourhood relationship is formally expressed in a $n \times n$ matrix of spatial weights (**W**), with elements (w_{ij}) representing a measure of the connection between locations i and j. The specification of the spatial weights matrix starts by identifying the neighbourhood structure of each cell. This neighbourhood can be identified by, for example, the adjacency of cells on a grid map, or by Euclidean or great circle distance (e.g. the distance along Earth's surface) to define cells within or outside a respective neighbourhood. The neighbours can further be weighted to give

closer neighbours higher weights and more distant neighbours lower weights. A number of methods are available for coding the spatial weights matrix (see Bivand, 2006), for example: (1) binary coding (locations are either listed as neighbours or not); (2) row-standardization (which scales the covariances based on the number of neighbours of each region in each row of the spatial weights matrix); or (3) variance stabilization (stabilizes variances by summing over all links, for details see Tiefelsdorf *et al.*, 1999). The final matrix of spatial weights **W** consists of zeros on the diagonal, and weights for the neighbouring locations (w_{ij}) in the off-diagonal positions.

Three different SAR models will be distinguished here depending on where the spatial autoregressive process is believed to occur (for details see Cliff & Ord, 1981; Anselin, 1988; Haining, 2003). The spatial error model (SAR_{err}) assumes that the autoregressive process is found only in the error term. This is most likely the case if spatial autocorrelation is not fully explained by the included explanatory variables ('induced spatial dependence'), e.g. if an important spatially structured explanatory variable has not been taken into account (Diniz-Filho *et al.*, 2003) or if spatial autocorrelation is an inherent property of the response variable itself ('inherent spatial autocorrelation'). For the SAR_{err}, the usual OLS regression model ($\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{e}$) is complemented by a term ($\lambda\mathbf{W}\boldsymbol{u}$) which represents the spatial structure ($\lambda\mathbf{W}$) in the spatially dependent error term ($\boldsymbol{u}$). The SAR_{err} thus takes the form

$$\mathbf{Y} = \mathbf{X\beta} + \lambda\mathbf{W}\boldsymbol{u} + \boldsymbol{e}$$

where λ is the spatial autoregression coefficient, **W** is the spatial weights matrix, $\boldsymbol{\beta}$ is a vector representing the slopes associated with the explanatory variables in the original predictor matrix **X**, and $\boldsymbol{e}$ represents the (spatially) independent errors.

Second, the SAR lagged model (SAR_{lag}) assumes that the autoregressive process occurs only in the response variable ('inherent spatial autocorrelation'), and thus includes a term ($\rho\mathbf{W}$) for the spatial autocorrelation in the response variable **Y**, but also the standard term for the explanatory variables and errors ($\mathbf{X\beta} + \boldsymbol{e}$) as used in an ordinary least squares (OLS) regression. The SAR_{lag} takes the form

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X\beta} + \boldsymbol{e}$$

where ρ is the autoregression coefficient, and the remaining terms are as above.

Finally, spatial autocorrelation can affect both response and explanatory variables (having both 'inherent spatial autocorrelation' and 'induced spatial dependence'). In this case, another term ($\mathbf{WX\gamma}$) must additionally appear in the model, which describes the autoregression coefficient ($\boldsymbol{\gamma}$) of the spatially lagged explanatory variables ($\mathbf{WX}$). The SAR mixed model (SAR_{mix}) takes the form

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X\beta} + \mathbf{WX\gamma} + \boldsymbol{e}.$$

For more details on SAR models and the estimation of the covariance matrices see Cliff & Ord (1981), Anselin (1988, 2002), Cressie (1993), Haining (2003) and Fortin & Dale (2005).

How to construct SAR models in R

The implementation of SAR models is illustrated here with the free software R (R Development Core Team 2005). The three SAR model types (SAR_{err}, SAR_{lag}, SAR_{mix}) are implemented in R in the library 'spdep' (Bivand, 2006). To use the SAR functions in R, one needs to specify the neighbourhood distances first, and the spatial weights matrix is then calculated by weighting the neighbours with a certain coding scheme (e.g. binary, row standardized or variance-stabilizing coding scheme, see above). A more detailed, annotated code to construct SAR models in R is given in Appendix S1 (see Supplementary Material).

Data

Four artificial data sets were created containing information on species distribution of a virtual organism and two environmental variables ('*rain*' and '*jungle*'). The four data sets were modified versions of the freely available R volcano data set (R Development Core Team, 2005), which gives topographic information on the Maunga Whau volcano near Auckland, New Zealand, on a 10 × 10 m grid. The extent of the grid (5307 grid cells) was first reduced to 1108 cells by simply increasing sea levels. Two explanatory variables, '*rain*' and '*jungle*', were then created. The variable *rain* (assumed to describe annual precipitation) was a significant determinant of the virtual organism distribution in all data sets, whereas the variable *jungle* (assumed to describe the percentage of jungle cover) did not have any explanatory power (noise). In all four data sets, the data on virtual organism distribution were normally distributed and had the same relationship to the explanatory variables *rain* and *jungle*, based on the following underlying model:

$$\text{expected value (virtual organism)} = 80 - (0.015 \times \textit{rain}) + (0 \times \textit{jungle})$$

The data sets differed, however, in the way in which spatial autocorrelation was induced. We first simulated three artificial data sets, which correspond to the mathematical formulation of the three model types for spatial externalities (Anselin, 2003). These data sets might therefore comprise ecologically borderline cases and were simulated with the aim of illustrating the performance of SAR models under extreme conditions. In all three cases we included spatial autocorrelation by multiplying a vector or matrix by the transpose of the Cholesky decomposition $\mathbf{C}^{T}$ of a pre-specified variance–covariance matrix. However, the data sets differed in where $\mathbf{C}^{T}$ was incorporated. In the 'error data', normally distributed errors containing spatial autocorrelation were added to the linear predictor ($\mathbf{Y} = \mathbf{X\beta} + \mathbf{C}^{T}\boldsymbol{e}$), and spatial autocorrelation was thus only present in the errors but not in the response variable or in the explanatory variables (see Fig. 1a). In real-case field data, this spatial autocorrelation pattern could, for instance, be caused by not taking into account the 'induced spatial dependence' of an important spatially structured explanatory variable (see Diniz-Filho *et al.*, 2003), or if spatial autocorrelation is an inherent property of the response variable itself ('inherent

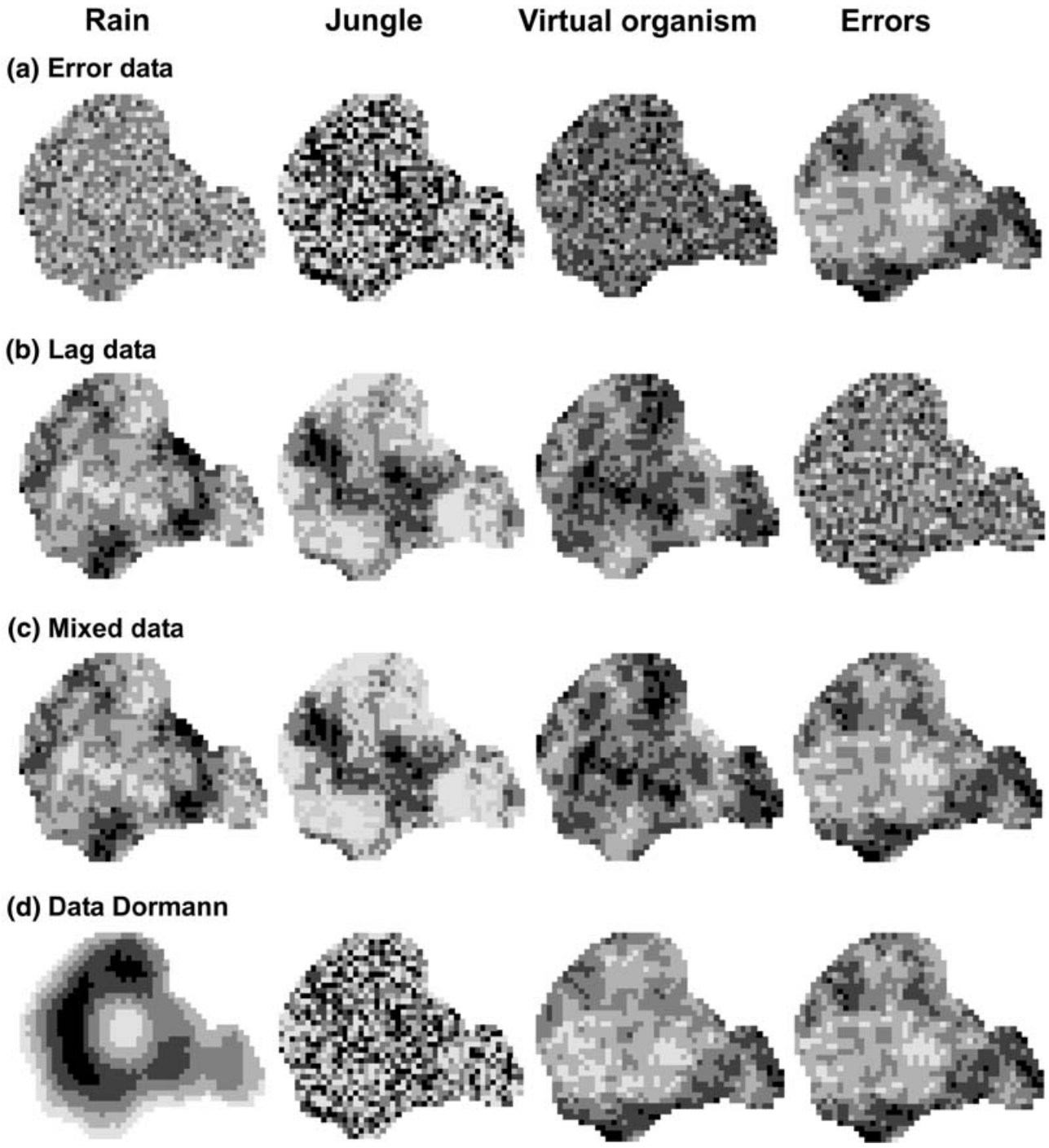

Figure 1 Spatial distribution of variables from four artificial data sets with different spatial autocorrelation structures. (a) Error data with spatial autocorrelation in errors only. (b) Lag data with spatial autocorrelation in both explanatory variables (*rain*, *jungle*) and in the distribution of the virtual organism, but not in the errors. (c) Mixed data with spatial autocorrelation in all variables. (d) Dormann data with spatial autocorrelation in virtual organism distribution and errors, and additional correlation (independent of errors) in *rain*. In all data sets, the relationship between response and explanatory variables is the same [E(virtual organism) = 80 − (0.015 × *rain*) + (0 × *jungle*)]. Equal-interval classification is shown, with light grey indicating minimum and black indicating maximum values. See text for more details.

spatial autocorrelation') and thus is not explained by the included explanatory variables. In the 'lag data', the spatial autocorrelation was incorporated in the explanatory variables only but not in the errors ($\mathbf{Y} = \mathbf{C}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}$), causing a spatial lag in the distribution of the virtual organism (Fig. 1b). This reflects the situation where all spatial autocorrelation in the response variable comes from exogenous processes ('induced spatial dependence'), here from one spatially structured environmental

variable. In the 'mixed data' (Fig. 1c), spatial autocorrelation was included in both the errors and the explanatory variables ($\mathbf{Y} = \mathbf{C}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}^{\mathrm{T}}\boldsymbol{e}$). Both explanatory variables, response and errors thus showed a spatially structured distribution (Fig. 1c). This pattern can arise if both endogenous and exogenous processes play a role (i.e. 'inherent spatial autocorrelation' and 'induced spatial dependence'). Note that the taxonomy used for describing the spatial autocorrelation in our data is similar to the formulation of the SAR model types, although the underlying regression models are not completely in line with them.

The fourth data set ('data Dormann', provided by C.F. Dormann) aimed to mimic ecological data and is the one currently used in a comprehensive evaluation of several statistical procedures to deal with spatial autocorrelation in statistical models (C.F. Dormann *et al.*, unpublished). In this fourth artificial data set, both explanatory variables (*rain*, *jungle*) were simulated with the same mean and the same variance as in the data above. The response variable (i.e. distribution of the virtual organism) was also calculated with the same formula as above, and normally distributed errors containing spatial autocorrelation (by multiplying them by the transpose of the same Cholesky decomposition $\mathbf{C}^{\mathrm{T}}$) were then incorporated in the distribution data for the virtual organism (Fig. 1d). In contrast to the data above, the spatial distribution of the variable *rain* was simulated to have a spatially structured pattern around the volcano in the centre of the map, with highest values in the western part of the study area. This spatial structure was created by adding a geographical pattern to *rain*. Hence, there is no collinearity between the *rain* pattern and the spatial distribution of the virtual organism and the errors (Fig. 1d). The variable *jungle* was purely randomly distributed in space (Fig. 1d).

SAR model performance

We first calculated all SAR models (SAR_{err}, SAR_{lag}, SAR_{mix}) with the same spatial weights matrix, i.e. an arbitrarily (but commonly) chosen neighbourhood distance of 1.5 and a coding style 'W' = row standardized (see Appendix S1). This was done for all four artificial data sets to illustrate the relative performance of SAR models without applying any model selection criteria. We compared the spatial autocorrelation pattern in model residuals using correlograms (Legendre & Fortin, 1989; Legendre, 1993), which plot Moran's *I* values (a measure for autocorrelation; Moran, 1950) on the *y*-axis against distance classes of sampling stations on the *x*-axis, and thus allow the assessment of the spatial autocorrelation pattern with increasing distance. Correlograms and Moran's *I* values were calculated with the function correlog() from the R package 'ncf' (Bjørnstad, 2005). We also plotted maps of model residuals to visualize their spatial pattern. Furthermore, we compared model parameter estimates for intercept, *rain* and *jungle* with the true (i.e. known) values (intercept, 80; *rain*, −0.015; *jungle*, 0). For comparison, we also did all calculations with simple OLS regressions for all data sets.

To assess the relative performance of parameter estimates in terms of type I errors (i.e. the probability α of falsely rejecting the null hypothesis H_0: $\beta = 0$) we calculated so-called calibration curves (see Fadili & Bullmore, 2002) where the observed number of type I errors (i.e. positive tests per 100 data realizations) is plotted against the expected number of type I errors (per 100 data realizations) across the full range of α. For this purpose, we generated 100 data realizations for each of the four data sets. The 100 data realizations of each artificial data set had exactly the same relationships as explained above. The only difference between the 100 realizations was that the normally distributed errors were randomly generated separately each time. We then calculated all models (SAR_{err}, SAR_{lag}, SAR_{mix} and OLS; SAR models with a neighbourhood distance of 1.5 and a coding style 'W') and recorded how often the *P* value of the (non-significant) variable *jungle* was falsely estimated to be $< \alpha$. The models work well when the observed number of type I errors equals the predicted one, i.e. when the calibration curve coincides with the line of identity given as a straight line in all plots (Fadili & Bullmore, 2002).

Model selection

Model selection can be helpful to identify a single best model or to make inferences from a set of multiple competing hypotheses (Johnson & Omland, 2004). Up to now, however, only a few model selection procedures have been tested for spatially autocorrelated data (e.g. Hoeting *et al.*, 2006). We therefore developed model selection procedures and selected the best SAR models from a range of models (see below) testing three model selection criteria: (1) minimum residual autocorrelation (minRSA); (2) maximum model fit (R^2); and (3) the AIC. R^2 values are not directly provided for SAR models, and maximum model fit was thus assessed with a pseudo-R^2 value (in the following simply referred to as R^2) calculated as the squared Pearson correlation between predicted and observed (i.e. true) values. We measured minRSA with correlograms (see above) by summing up the absolute Moran's *I* values in the first 20 distance classes of the correlogram (similar to C.F. Dormann *et al.*, unpublished). AIC values are directly provided for SAR models and allow the selection of models based on both model fit and model complexity (Burnham & Anderson, 1998). They are now being used widely in ecological and evolutionary studies (Johnson & Omland, 2004).

For the model selection procedures, we simulated 10 data realizations (similar to those above) for each of the four artificial data sets. We chose 10 data realizations here because a larger number of realizations would have been beyond our ability to test a great variety and number of SAR models (see below). For all 10 realizations of each of the four data sets we calculated the three SAR model types (SAR_{lag}, SAR_{mix}, SAR_{err}) with 27 different spatial weights matrices (i.e. 81 model specifications with 10 data realizations and four data sets = 3240 tested SAR models). The spatial weights matrices were constructed with nine neighbourhood distances (from 1 to 5, in steps of 0.5). Additionally, three coding styles ('B' = binary coding, 'W' = row standardized, and 'S' = variance stabilizing; see above) were tested. All SAR models were run with both explanatory variables (*rain*, *jungle*). For each of the 10 data realizations within each artificial data set we

selected the best model (based on minRSA, R^2 or AIC) from the 27 model specifications. We then assessed, for each of the four data sets, which SAR models and spatial weights matrices (i.e. neighbourhood distances and coding styles) were selected, and which performed poorly (in terms of parameter estimates) given a certain spatial autocorrelation structure in the data. To assess parameter estimates of intercept, *rain* and *jungle*, we calculated mean values (±SD) across the 10 data realizations from the best models (i.e. selected based on minRSA, R^2 or AIC, respectively), and compared them with the true values. For comparison, simple OLS regressions were also included in these analyses.

RESULTS

SAR model performance

The different spatial autocorrelation structures of the four artificial data sets (Fig. 1) were not equally detected by the different models tested. OLS models generally showed a spatial autocorrelation pattern in the residuals for all data sets except the lag data (Figs 2 & 3), which contained spatial autocorrelation in the explanatory variables and the response but not in the errors (Fig. 1). For an arbitrarily (but commonly) chosen spatial weights matrix with a neighbourhood distance of 1.5 and a coding style 'W' (= row standardized), the SAR_{lag} model was only able to remove the spatial autocorrelation in the lag data and the Dormann data but not in the two other data sets where spatial autocorrelation was induced in the errors (Figs 2 & 3). The SAR_{err} and SAR_{mix} models with this particular spatial weights matrix instead performed well, and were able to account for the spatial autocorrelation structures in all four data sets (Figs 2 & 3). However, parameter estimates from OLS and SAR models (with this single spatial weights matrix) were not always very precise, depending on the data set analysed (Fig. 4). Those from SAR_{err} models performed best whereas OLS, SAR_{lag} and SAR_{mix} model parameter estimates sometimes showed strong deviations from the true values (Fig. 4). Type I error control by OLS and SAR_{lag} was poor in all cases except for the lag data (Fig. 5, columns 1 and 3), whereas type I error control by SAR_{err} and SAR_{mix} was very good over the full range of probability thresholds (Fig. 5, columns 2 and 4).

Model selection for SAR

In contrast to above (and Fig. 4), the model selection procedures were designed to test a great variety of SAR models with different (but not arbitrarily chosen) spatial weights matrices. Most of the 360 SAR models that were selected by our model selection criteria (i.e. minRSA, R^2 or AIC) had a spatial weights matrix with a neighbourhood distance of 1 or 1.5 (83%) and a row standardized coding style 'W' (67%). However, SAR models with higher neighbourhood distances up to 5 distance units (e.g. SAR_{lag} models used with error and lag data) and coding styles 'B' (19%) and 'S' (14%) were also selected (see Appendix S2 for summary statistics from model selection). Parameter estimates of intercept, *rain* and *jungle* from these selected SAR models were usually very close to the true values (Fig. 6). However, some notable exceptions became apparent. SAR_{lag} and SAR_{mix} models sometimes showed strong (and unpredictable) deviations from the true parameter values, in particular for estimates of intercept and partly for *rain*. These biases in parameter estimates were evident in all data sets with spatial autocorrelation in the errors (error data, mixed data, data Dormann) but not in the lag data where autocorrelation was absent from the errors. Across all data

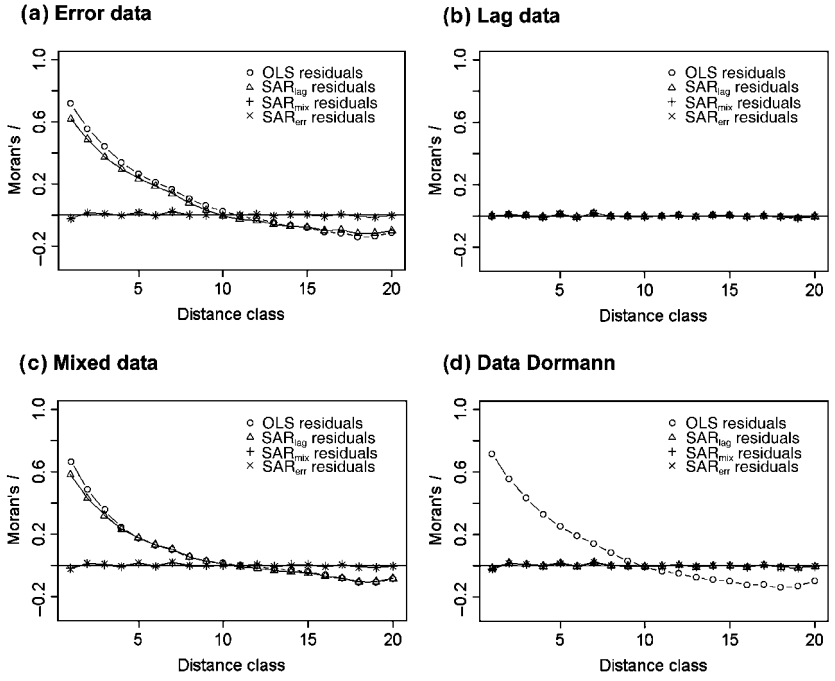

Figure 2 Correlograms for residuals from ordinary least squares (OLS) and three simultaneous autoregressive models (SAR_{err}, SAR_{lag}, SAR_{mix}) for four artificial data sets with different spatial autocorrelation structures (see Fig. 1). All models have the same relationship between response and explanatory variables [E(virtual organism) = 80 − (0.015 × *rain*) + (0 × *jungle*)]. The spatial weights matrix of all SAR models was calculated with a neighbourhood distance of 1.5 and a row standardized coding scheme ('W'). See text for more details.

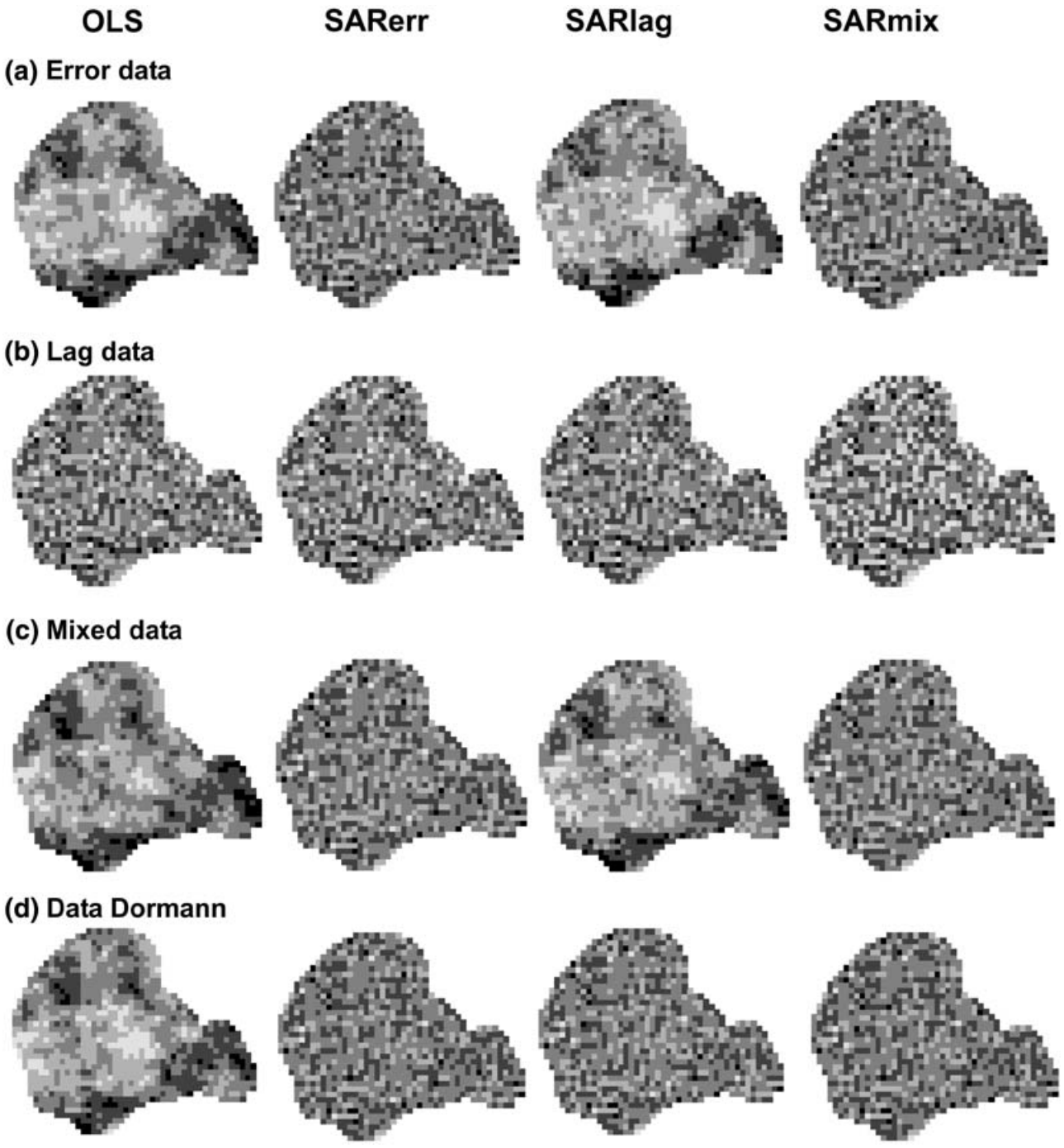

Figure 3 Residual maps illustrating the spatial distribution of residuals from non-spatial (ordinary least squares, OLS) and spatial simultaneous autoregressive (SAR_{lag}, SAR_{mix}, SAR_{err}) regression models. Models and data are the same as in Figs 1 & 2. Equal-interval classification is shown, with light grey indicating minimum and black indicating maximum residual values. See text for details on models.

sets, SAR_{err} models performed well and gave the most precise parameter estimates (Fig. 6), independent of the model selection criteria used (minRSA, R^2 or AIC). Parameter estimates from OLS regressions were unbiased (Fig. 6) although spatial autocorrelation was present in the OLS residuals (Figs 2 & 3; see also minRSA in Appendix S2). However, for the Dormann data, for example, the variance of the parameter estimate of *jungle* was very large (Fig. 6) and thus type I error control was poor (Fig. 5). For all data sets except the lag data, selected SAR models had higher R^2-values, lower AIC values and less spatial autocorrelation in the residuals (minRSA) than OLS regressions (see Appendix S2 for summary statistics from model selection). The lag data were correctly identified by SAR_{lag}, yielding the lowest AIC values. AIC values of SAR_{mix} were often almost as good as those of

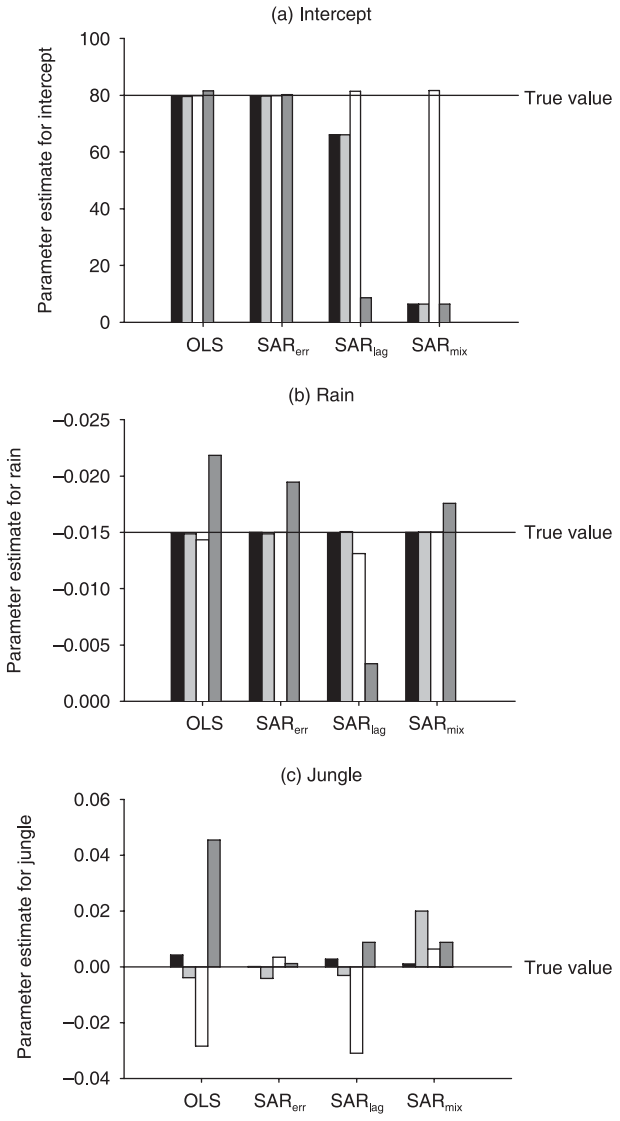

Figure 4 Parameter estimates from ordinary least squares (OLS) and simultaneous autoregressive models (SAR_{err}, SAR_{lag}, SAR_{mix}) for four artificial data sets with different spatial autocorrelation structures (black, error data; light grey, lag data; white, mixed data; dark grey, Dormann data). SAR models were calculated with a spatial weights matrix based on a neighbourhood distance of 1.5 and a row standardized coding scheme 'W', and correspond to Figs 2 & 3. The data sets are illustrated in Fig. 1.

SAR_{err}. SAR models with precise parameter estimates were always indicated by a combination of low AIC values and low minRSA values (see Appendix S2).

DISCUSSION

Simultaneous autoregressive models have the potential to reduce or remove the spatial pattern of model residuals and thus help to meet the assumption of independently distributed errors in regression models. However, our study shows that the performance of SAR models depends on model specification (i.e. model type, neighbourhood distance, coding styles of spatial weights matrices), and SAR model parameter estimates are not always more precise than those from OLS regressions. Our results indicate that SAR_{err} models are the most reliable SAR models in terms of precision of parameter estimates, reduction of spatial autocorrelation in model residuals and type I error control, independent of which kind of spatial autocorrelation is present in the data set. Other SAR models (SAR_{lag}, SAR_{mix}) and OLS regressions showed weak type I error control and/or unpredictable biases in parameter estimates when spatial autocorrelation was present in the errors. We do not therefore recommend them for real species distribution data where spatial autocorrelation is most likely to occur in model residuals, e.g. when important environmental variables have not been taken into account (Diniz-Filho *et al.*, 2003).

In our artificial data sets, the induced spatial autocorrelation structure was often removed when using SAR models with small neighbourhood distances (i.e. 1 or 1.5 distance units). This is consistent with some real ecological data sets where the spatial autocorrelation signature can be removed by using autoregressive models that incorporate information from neighbours immediately surrounding the focal cell (so-called first-order neighbourhoods, e.g. Jetz & Rahbek, 2002; Overmars *et al.*, 2003; Kissling *et al.*, 2007). However, other species distribution analyses show that higher-order neighbourhoods (i.e. larger distances) are necessary if the removal of spatial autocorrelation is attempted (e.g. Lichstein *et al.*, 2002; Tognelli & Kelt, 2004; Kühn, 2007). It is obvious that the degree of spatial autocorrelation depends on the data set analysed, and, consequently, it is difficult to decide *a priori* which neighbourhood structure (i.e. distance and coding style) is the most efficient one. We therefore suggest that ecologists should test a wide variety of SAR model specifications for each species distribution data set, and identify a single best model or a set of models (Burnham & Anderson, 1998; Johnson & Omland, 2004) based on one or more model selection criteria (see below).

Because statistical models aim to describe data, the preferred model selection criterion should be based on R^2 values because they describe model fit, or even better on AIC values, which are based on model fit and model complexity (Burnham & Anderson, 1998; Johnson & Omland, 2004). AIC values have also been suggested recently for spatially autocorrelated data when using geostatistical models (Hoeting *et al.*, 2006), but studies on AIC model selection with spatially autocorrelated data are otherwise largely lacking. To our knowledge, there is almost no information in the literature about whether minRSA (i.e. the reduction of spatial autocorrelation in model residuals) can also be a valid model selection criterion that identifies models with precise parameter estimates. In our model selection procedures, we could not find any difference in the precision of parameter estimates when SAR_{err} models were selected by minRSA, R^2 or AIC values. We therefore expect all three model selection criteria to be reliable when used with SAR_{err} models. In contrast, SAR_{lag} and SAR_{mix} models sometimes showed differences in the precision of parameter estimates depending on which model selection criterion was used (Fig. 6). However, there was no clear (i.e. systematic) trend in whether one of them is more reliable than

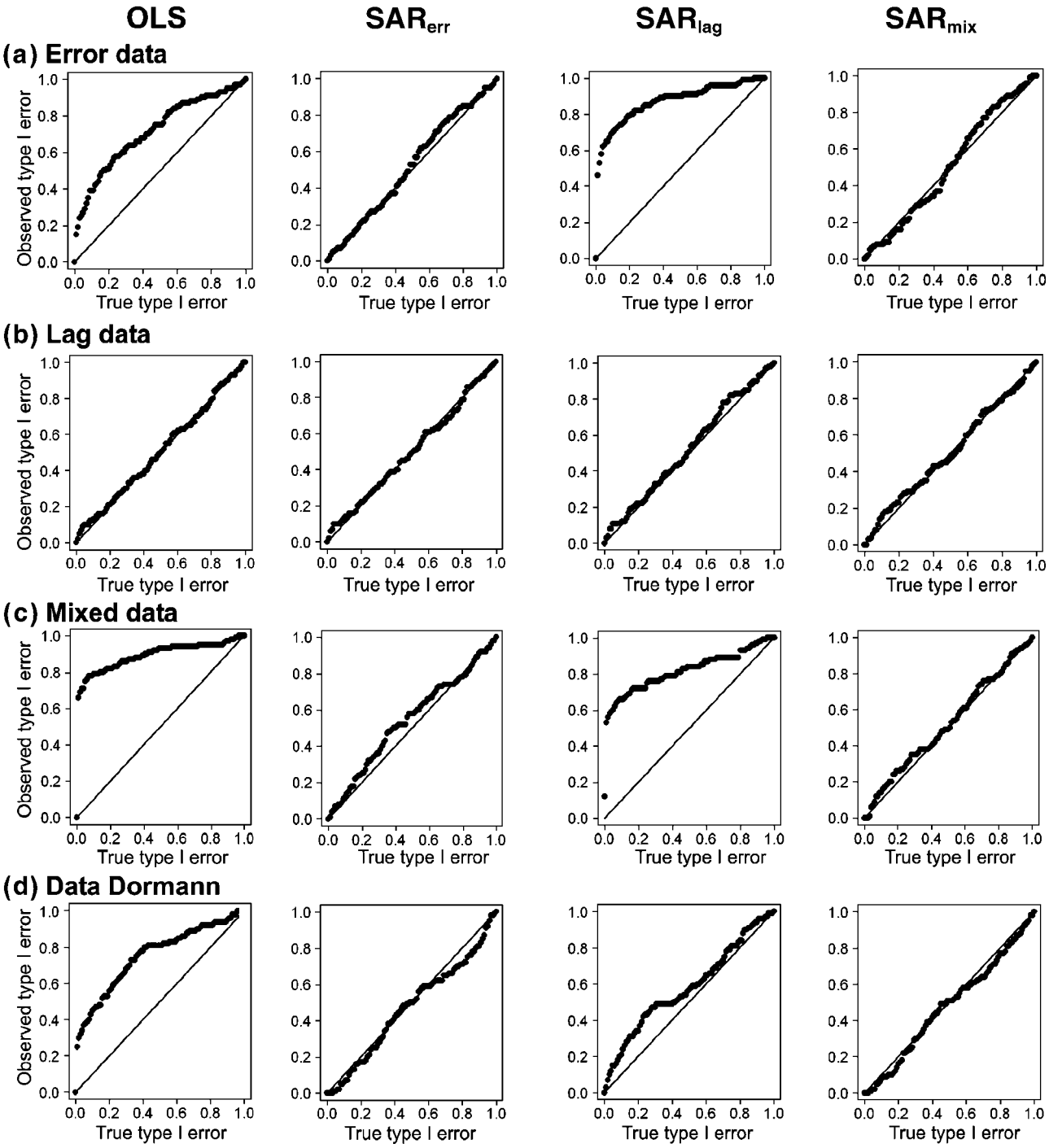

Figure 5 Type I error calibration curves for ordinary least squares (OLS) and simultaneous autoregressive models (SAR_{err}, SAR_{lag}, SAR_{mix}) from 100 data realizations. Illustrated are the observed versus predicted type I error probabilities for falsely estimating the (non-significant) variable '*jungle*' to be significant (i.e. $P < \alpha$, in the full range of α [0, 1]). Models perform well for a given data set (a, error data; b, lag data; c, mixed data; d, data Dormann) when the calibration curve coincides with the line of identity given as a straight line in all plots.

another. Overall, based on the performance of the SAR_{err} models, we recommend that AIC and minRSA should be used jointly to identify the most appropriate model.

Our study supports previous findings (e.g. Legendre *et al.*, 2002) that type I errors from traditional, non-spatial analyses are strongly inflated when spatial autocorrelation is present (see OLS in Fig. 5). In contrast, SAR_{err} and SAR_{mix} models were not prone to type I errors for all tested data sets (Fig. 5). However, SAR_{lag} models showed similar levels of type I error inflation than OLS (Fig. 5), indicating that both methods are unable to reject

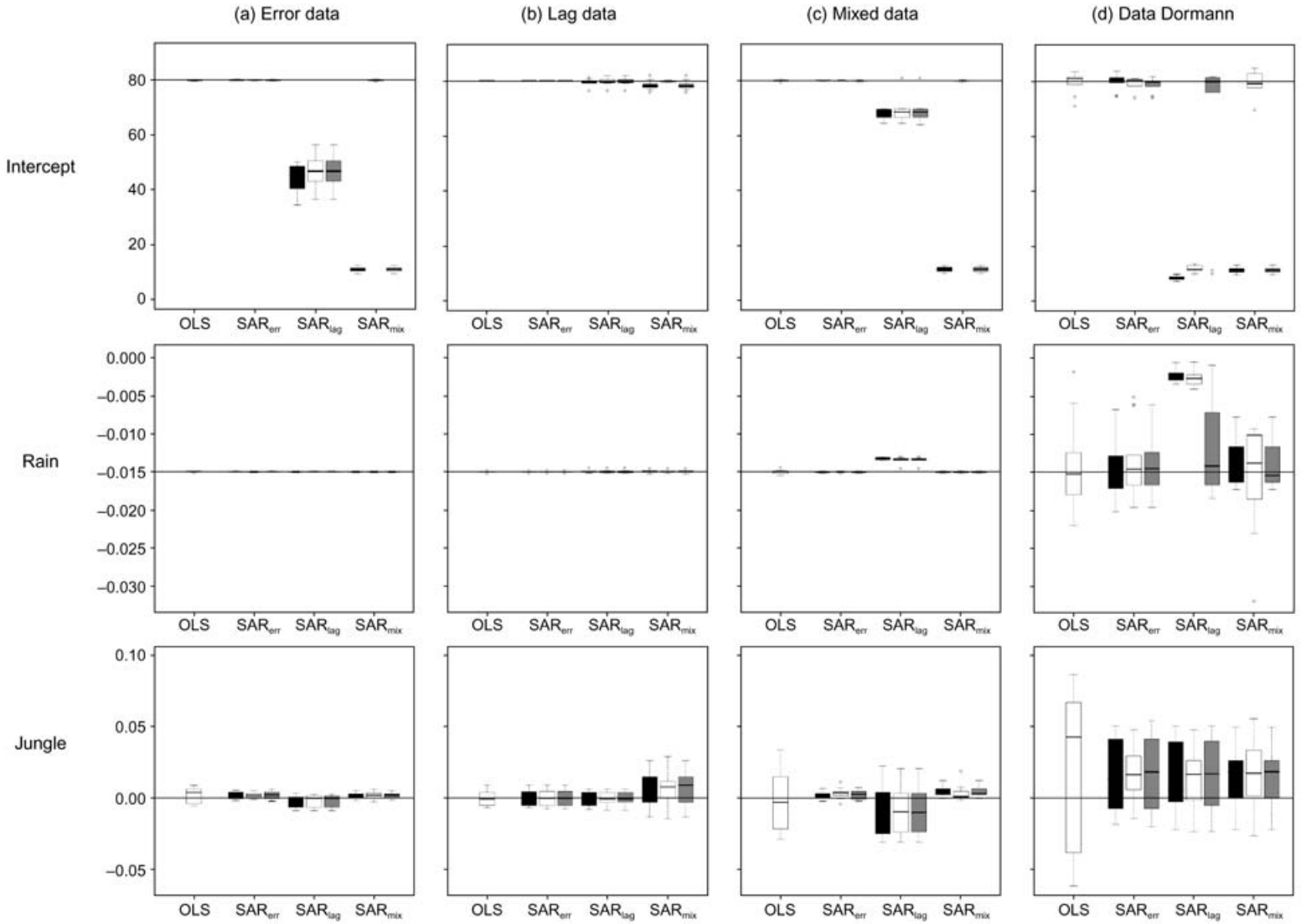

Figure 6 Box-and-whisker plots of parameter estimates for intercept (upper row), *rain* (middle row) and *jungle* (lower row) across 10 data realizations of four data sets (a, error data; b, lag data; c, mixed data; d, data Dormann) with different spatial autocorrelation structures. Ordinary least squares (OLS) values were derived from ordinary least squares regressions for the 10 data realizations within each data set. For simultaneous autoregressive models (SAR_{err}, SAR_{lag}, SAR_{mix}), values were obtained from model selection procedures based on minimum residual spatial autocorrelation (minRSA, black), maximum model fit (R^2, white) and Akaike information criterion (AIC, grey), respectively (see text for details). True parameter values of the four data sets are indicated as a straight line in each graph.

non-significant explanatory variables (here: *jungle*) if spatial autocorrelation is present in the residuals (a likely feature of real ecological data sets). The lag data, where all spatial autocorrelation in the response variable ('spatial lag') was caused by one spatially structured explanatory variable ('induced spatial dependence'), did not constitute a problem with regard to type I error control for any of the tested methods (Fig. 5). This indicates that type I errors are not inflated if the spatial structure of species distributions is caused only by those explanatory variables that are included in the model. However, in many situations we might not be able to include all important environmental variables, for instance if they are not available at a required spatial resolution or at the necessary biological accuracy (Diniz-Filho *et al.*, 2003; Dormann, 2007). This will cause spatial autocorrelation to be present in model residuals and thus can cause type I error inflation in OLS and SAR_{lag} models but not in SAR_{err} and SAR_{mix} (Fig. 5).

Apart from type I errors, the estimation of model coefficients is an additional difficulty in modelling species distributions with spatially autocorrelated data (Lennon, 2000; Diniz-Filho *et al.*, 2003; Dormann, 2007). Our study clearly showed that the selection of the SAR model type (SAR_{err}, SAR_{lag}, SAR_{mix}) can strongly influence parameter estimates, which might be even worse (e.g. for SAR_{lag} and SAR_{mix}) than parameter estimates from common OLS regressions (Figs 4 & 6). This is surprising, because many studies suggest (or simply assume) that parameter estimates (and hypotheses derived) from spatial models are generally better than those from OLS regressions (e.g. Lennon, 2000; Lichstein *et al.*, 2002; Dark, 2004; Tognelli & Kelt, 2004; Dormann, 2007; Kühn, 2007). Our results should thus cause us to be cautious about assuming that spatial regression techniques always provide better parameter estimates than OLS so long as it has not been demonstrated under which circumstances this is true. It is important to note, however, that our artificial data sets are simplifications of the real world since we have only one explanatory variable significantly correlated with the response, and hence there is no multicollinearity in our data. More comprehensive tests of SAR models and other spatial modelling techniques should be conducted to disentangle the influence of multiple, spatially autocorrelated explanatory variables on parameter estimation.

The examination of differences in parameter estimates between spatial and non-spatial methods might be helpful for improving our understanding of the ecological mechanisms behind the patterns we observe (Diniz-Filho *et al.*, 2003). Lennon (2000) suggested that parameter shifts between spatial and non-spatial multiple regression analyses are particularly strong if explanatory variables are spatially autocorrelated, and that environmental factors with less spatial autocorrelation are much more likely to be rejected by traditional, non-spatial analyses (so-called 'red shifts'). This could be a serious problem, because it would lead to a systematic bias in the choice of explanatory variables towards those that have the greater spatial autocorrelation. Diniz-Filho *et al.* (2003) supported this view by showing that spatial models de-emphasize explanatory variables with strong spatial autocorrelation and thus give more importance to variables acting at smaller spatial scales. Moreover, they interpreted this as a hierarchical effect, so that differences between spatial and non-spatial methods could reflect mechanisms at different spatial scales. Although our analyses were not designed to test these issues, our results support this last view because systematic shifts in parameter estimates between SAR and OLS were not observed when dealing with one spatially autocorrelated explanatory variable (Fig. 6).

The interpretation of parameter estimates and model coefficients from spatial models is now among the most important issues in geographical ecology (Lennon, 2000; Diniz-Filho *et al.*, 2003; Tognelli & Kelt, 2004; Dormann, 2007; Kühn, 2007). This is not simply a statistical discussion but has profound implications for biogeography, macroecology and global change research because biased estimates and incorrect model specifications will influence the testing of hypotheses and the prediction of species distributions (e.g. Diniz-Filho *et al.*, 2003; Dark, 2004; Guisan *et al.*, 2006; Dormann, 2007). Our study complements previous studies on species distribution and spatial autocorrelation (e.g. Keitt *et al.*, 2002; Legendre *et al.*, 2002; Lichstein *et al.*, 2002; Diniz-Filho *et al.*, 2003; Tognelli & Kelt, 2004; Dormann, 2007; Kühn, 2007) and is thus a further step towards a better understanding of the behaviour and potential of spatial methods. We propose to extend the ongoing comprehensive tests of non-spatial methods for modelling species distributions (e.g. Segurado & Araújo, 2004; Elith *et al.*, 2006) with a comprehensive assessment and full comparison of the various spatial modelling techniques (Cressie, 1993; Haining, 2003; Rangel *et al.*, 2006; C.F. Dormann *et al.*, unpublished). This should include an evaluation of the predictive performance and accuracy of spatial models under changing environmental conditions such as climate change (for good examples with non-spatial methods see Araújo *et al.*, 2005a, b; Hijmans & Graham, 2006). These methodological comparisons will help to identify the potential and pitfalls of the various spatial modelling techniques and might help to reduce uncertainty in model predictions.

ACKNOWLEDGEMENTS

We thank Carsten F. Dormann for inviting us to the Kohren-Sahlis spatial statistics workshop and for providing the 'Dormann data', Jana McPherson for providing helpful R-code, and Ingolf Kühn, J. Alexandre F. Diniz-Filho and one anonymous referee for stimulating and constructive comments on a draft manuscript. This work has also benefited from discussions with Roger Bivand, Ingolf Kühn, Carsten F. Dormann, Thiago F.L.V.B. Rangel, and colleagues from the Paper Discussion Club at the Department of Ecology, University of Mainz, Germany. The Virtual Institute for Macroecology funded by the Helmholtz Association organized a course on spatial and phylogenetic statistics where the principal methods used in this paper were introduced to us. W.D.K. is grateful to his PhD supervisor Katrin Böhning-Gaese for financial and institutional support, and for the freedom to work on the ideas outlined in this paper.

REFERENCES

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

Anselin, L. (2002) Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, **27**, 247–267.

Anselin, L. (2003) Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review*, **26**, 153–166.

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

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005a) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Araújo, M.B., Whittaker, R.J., Ladle, R.J. & Erhard, M. (2005b) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529–538.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Bivand, R. (2006) *Spdep: spatial dependence: weighting schemes, statistics and models. R package version 0.3-31*. Available online at http://cran.r-project.org/src/contrib/Descriptions/spdep.html

Bjørnstad, O.N. (2005) *ncf: spatial nonparametric covariance functions. R package version 1.0-8*. Available online at http://onb.ent.psu.edu/onb1/

Burnham, K.P. & Anderson, D.R. (1998) *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

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

Cressie, N.A.C. (1993) *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

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.

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**, 177–185.

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.

Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

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.S., 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.

Fadili, M.J. & Bullmore, E.T. (2002) Wavelet-generalized least squares: a new BLU estimator of linear regression models with 1/f errors. *NeuroImage*, **15**, 217–232.

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

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.Mc.C., Aspinall, R. & Hastie, T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, **43**, 386–392.

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

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12**, 2272–2281.

Hoeting, J.A., Davis, R.A., Merton, A.A. & Thompson, S.E. (2006) Model selection for geostatistical models. *Ecological Applications*, **16**, 87–98.

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

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution*, **19**, 101–108.

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

Kissling, W.D., Rahbek, C. & Böhning-Gaese, K. (2007) Food plant diversity as broad-scale determinant of avian frugivore richness. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 799–808.

Kühn, I. (2007) Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions*, **13**, 66–69.

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

Legendre, P. & Fortin, M.-J. (1989) Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–138.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*. Elsevier, Amsterdam.

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.

Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Overmars, K.P., de Koning, G.H.J. & Veldkamp, A. (2003) Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, **164**, 257–270.

Rangel, T.F.L.V.B., Diniz-Filho, J.A.F. & Bini, L.M. (2006) Towards an integrated computational tool for spatial analysis in macroecology and biogeography. *Global Ecology and Biogeography*, **15**, 321–327.

R Development Core Team (2005) *R: a language and environment for statistical computing*. R foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org

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

Tiefelsdorf, M., Griffith, D.A. & Boots, B. (1999) A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A*, **31**, 165–180.

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.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix S1 How to construct SAR models in R.

Appendix S2 Summary characteristics from model selection.

Appendix S3 Data table for analyses in Appendix S1.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1466-8238.2007.00334.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

BIOSKETCHES

W. Daniel Kissling is an ecologist and PhD student at the University of Mainz, Germany. He is interested in geographical ecology and biodiversity conservation, with a current focus on the macroecology of frugivore diversity. Apart from ecological complexity, nature and wild birds, he likes travelling, gardening and Latin American culture.

Gudrun Carl received her PhD in (theoretical) physics and gained experience in the field of mathematics. Her recent field of research is the development of methods for spatial and temporal analysis of environmental data. This paper was finalized when she was working for the UFZ-Helmholtz Centre for Environmental Research, Department of Community Ecology (http://www.ufz.de/index.php?en=10028).

Editor: José Alexandre F. Diniz-Filho