Review & synthesis

Ecography 30: 609–628, 2007
doi: 10.1111/j.2007.0906-7590.05171.x

Subject Editor: Carsten Rahbek. Accepted 3 August 2007

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

Carsten F. Dormann, Jana M. McPherson, Miguel B. Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Kühn, Ralf Ohlemüller, Pedro R. Peres-Neto, Björn Reineking, Boris Schröder, Frank M. Schurr and Robert Wilson

C. F. Dormann (carsten.dormann@ufz.de), Dept of Computational Landscape Ecology, UFZ Helmholtz Centre for Environmental Research, Permoserstr. 15, DE-04318 Leipzig, Germany. – J. M. McPherson, Dept of Biology, Dalhousie Univ., 1355 Oxford Street Halifax NS, B3H 4J1 Canada. – M. B. Araújo, Dept de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, CSIC, C/ Gutiérrez Abascal, 2, ES-28006 Madrid, Spain, and Centre for Macroecology, Inst. of Biology, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark. – R. Bivand, Economic Geography Section, Dept of Economics, Norwegian School of Economics and Business Administration, Helleveien 30, NO-5045 Bergen, Norway. – J. Bolliger, Swiss Federal Research Inst. WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland. – G. Carl and I. Kühn, Dept of Community Ecology (BZF), UFZ Helmholtz Centre for Environmental Research, Theodor-Lieser-Strasse 4, DE-06120 Halle, Germany, and Virtual Inst. Macroecology, Theodor-Lieser-Strasse 4, DE-06120 Halle, Germany. – R. G. Davies, Biodiversity and Macroecology Group, Dept of Animal and Plant Sciences, Univ. of Sheffield, Sheffield S10 2TN, U.K. – A. Hirzel, Ecology and Evolution Dept, Univ. de Lausanne, Biophore Building, CH-1015 Lausanne, Switzerland. – W. Jetz, Ecology Behavior and Evolution Section, Div. of Biological Sciences, Univ. of California, San Diego, 9500 Gilman Drive, MC 0116, La Jolla, CA 92093-0116, USA. – W. D. Kissling, Community and Macroecology Group, Inst. of Zoology, Dept of Ecology, Johannes Gutenberg Univ. of Mainz, DE-55099 Mainz, Germany, and Virtual Inst. Macroecology, Theodor-Lieser-Strasse 4, DE-06120 Halle, Germany. – R. Ohlemüller, Dept of Biology, Univ. of York, PO Box 373, York YO10 5YW, U.K. – P. R. Peres-Neto, Dept of Biology, Univ. of Regina, SK, S4S 0A2 Canada, present address: Dept of Biological Sciences, Univ. of Quebec at Montreal, CP 8888, Succ. Centre Ville, Montreal, QC, H3C 3P8, Canada. – B. Reineking, Forest Ecology, ETH Zurich CHN G 75.3, Universitätstr. 16, CH-8092 Zürich, Switzerland. – B. Schröder, Inst. for Geoecology, Univ. of Potsdam, Karl-Liebknecht-Strasse 24-25, DE-14476 Potsdam, Germany. – F. M. Schurr, Plant Ecology and Nature Conservation, Inst. of Biochemistry and Biology, Univ. of Potsdam, Maulbeerallee 2, DE-14469 Potsdam, Germany. – R. Wilson, Área de Biodiversidad y Conservación, Escuela Superior de Ciencias Experimentales y Tecnología, Univ. Rey Juan Carlos, Tulipán s/n, Móstoles, ES-28933 Madrid, Spain.

Species distributional or trait data based on range map (extent-of-occurrence) or atlas survey data often display spatial autocorrelation, i.e. locations close to each other exhibit more similar values than those further apart. If this pattern remains present in the residuals of a statistical model based on such data, one of the key assumptions of standard statistical analyses, that residuals are independent and identically distributed (i.i.d), is violated. The violation of the assumption of i.i.d. residuals may bias parameter estimates and can increase type I error rates (falsely rejecting the null hypothesis of no effect). While this is increasingly recognised by researchers analysing species distribution data, there is, to our knowledge, no comprehensive overview of the many available spatial statistical methods to take spatial autocorrelation into account in tests of statistical significance. Here, we describe six different statistical approaches to infer correlates of species' distributions, for both presence/absence (binary response) and species abundance data (poisson or normally distributed response), while accounting for spatial autocorrelation in model residuals: autocovariate regression; spatial eigenvector mapping; generalised least squares; (conditional and simultaneous) autoregressive models and generalised estimating equations. A comprehensive comparison of the relative merits of these methods is beyond the scope of this paper. To demonstrate each method's implementation, however, we undertook preliminary tests based on simulated data. These preliminary tests verified that most of the spatial modeling techniques we examined showed good type I error control and precise parameter estimates, at least when confronted with simplistic simulated data containing

spatial autocorrelation in the errors. However, we found that for presence/absence data the results and conclusions were very variable between the different methods. This is likely due to the low information content of binary maps. Also, in contrast with previous studies, we found that autocovariate methods consistently underestimated the effects of environmental controls of species distributions. Given their widespread use, in particular for the modelling of species presence/absence data (e.g. climate envelope models), we argue that this warrants further study and caution in their use. To aid other ecologists in making use of the methods described, code to implement them in freely available software is provided in an electronic appendix.

Species distributional data such as species range maps (extent-of-occurrence), breeding bird surveys and biodiversity atlases are a common source for analyses of species-environment relationships. These, in turn, form the basis for conservation and management plans for endangered species, for calculating distributions under future climate and land-use scenarios and other forms of environmental risk assessment.

The analysis of spatial data is complicated by a phenomenon known as spatial autocorrelation. Spatial autocorrelation (SAC) occurs when the values of variables sampled at nearby locations are not independent from each other (Tobler 1970). The causes of spatial autocorrelation are manifold, but three factors are particularly common (Legendre and Fortin 1989, Legendre 1993, Legendre and Legendre 1998): 1) biological processes such as speciation, extinction, dispersal or species interactions are distance-related; 2) non-linear relationships between environment and species are modelled erroneously as linear; 3) the statistical model fails to account for an important environmental determinant that in itself is spatially structured and thus causes spatial structuring in the response (Besag 1974). The second and third points are not always referred to as spatial autocorrelation, but rather spatial dependency (Legendre et al. 2002). Since they also lead to autocorrelated residuals, these are equally problematic. A fourth source of spatial autocorrelation relates to spatial resolution, because coarser grains lead to a spatial smoothing of data. In all of these cases, SAC may confound the analysis of species distribution data.

Spatial autocorrelation may be seen as both an opportunity and a challenge for spatial analysis. It is an opportunity when it provides useful information for inference of process from pattern (Palma et al. 1999) by, for example, increasing our understanding of contagious biotic processes such as population growth, geographic dispersal, differential mortality, social organization or competition dynamics (Griffith and Peres-Neto 2006). In most cases, however, the presence of spatial autocorrelation is seen as posing a serious shortcoming for hypothesis testing and prediction (Lennon 2000, Dormann 2007b), because it violates the assumption of independently and identically distributed (i.i.d.) errors of most standard statistical procedures (Anselin 2002) and hence inflates type I errors, occasionally even inverting the slope of relationships from non-spatial analysis (Kühn 2007).

A variety of methods have consequently been developed to correct for the effects of spatial autocorrelation (partially reviewed in Keitt et al. 2002, Miller et al. 2007, see below), but only a few have made it into the ecological literature. The aims of this paper are to 1) present and explain methods that account for spatial autocorrelation in analyses of spatial data; the approaches considered are: autocovariate regression, spatial eigenvector mapping (SEVM), generalised least squares (GLS), conditional autoregressive models (CAR), simultaneous autoregressive models (SAR), generalised linear mixed models (GLMM) and generalised estimation equations (GEE); 2) describe which of these methods can be used for which error distribution, and discuss potential problems with implementation; 3) illustrate how to implement these methods using simulated data sets and by providing computing code (Anon. 2005).

Methods for dealing with spatial autocorrelation

Detecting and quantifying spatial autocorrelation

Before considering the use of modelling methods that account for spatial autocorrelation, it is a sensible first step to check whether spatial autocorrelation is in fact likely to impact the planned analyses, i.e. if model residuals indeed display spatial autocorrelation. Checking for spatial autocorrelation (SAC) has become a commonplace exercise in geography and ecology (Sokal and Oden 1978a, b, Fortin and Dale 2005). Established procedures include (Isaaks and Shrivastava 1989, Perry et al. 2002): Moran's I plots (also termed Moran's I correlogram by Legendre and Legendre 1998), Geary's c correlograms and semi-variograms. In all three cases a measure of similarity (Moran's I, Geary's c) or variance (variogram) of data points (i and j) is plotted as a function of the distance between them (d_{ij}). Distances are usually grouped into bins. Moran's I-based correlograms typically show a decrease from some level of SAC to a value of 0 (or below; expected value in the absence of SAC: $E(I) = -1/(n-1)$, where n = sample size), indicating no SAC at some distance between locations. Variograms depict the opposite, with the variance

between pairs of points increasing up to a certain distance, where variance levels off. Variograms are more commonly employed in descriptive geostatistics, while correlograms are the prevalent graphical presentation in ecology (Fortin and Dale 2005).

Values of Moran's I are assessed by a test statistic (the Moran's I standard deviate) which indicates the statistical significance of SAC in e.g. model residuals. Additionally, model residuals may be plotted as a map that more explicitly reveals particular patterns of spatial autocorrelation (e.g. anisotropy or non-stationarity of spatial autocorrelation). For further details and formulae see e.g. Isaaks and Shrivastava (1989) or Fortin and Dale (2005).

Assumptions common to all modelling approaches considered

All methods assume spatial stationarity, i.e. spatial autocorrelation and effects of environmental correlates to be constant across the region, and there are very few methods to deal with non-stationarity (Osborne et al. 2007). Stationarity may or may not be a reasonable assumption, depending, among other things, on the spatial extent of the study. If the main cause of spatial autocorrelation is dispersal (for example in research on animal distributions), stationarity is likely to be violated, for example when moving from a floodplain to the mountains, where movement may be more restricted. One method able to accommodate spatial variation in autocorrelation is geographically weighted regression (Fotheringham et al. 2002), a method not considered here because of its limited use for hypothesis testing (coefficient estimates depend on spatial position) and because it was not designed to remove spatial autocorrelation (see e.g. Kupfer and Farris 2007, for a GWR correlogram).

Another assumption is that of isotropic spatial autocorrelation. This means that the process causing the spatial autocorrelation acts in the same way in all directions. Environmental factors that may cause anisotropic spatial autocorrelation are wind (giving a wind-dispersed organism a preferential direction), water currents (e.g. carrying plankton), or directionality in soil transport (carrying seeds) from mountains to plains. He et al. (2003) as well as Worm et al. (2005) provide examples of analyses accounting for anisotropy in ecological data, and several of the methods described below can be adapted for such circumstances.

Description of spatial statistical modelling methods

The methods we describe in the following fall broadly into three groups. 1) Autocovariate regression and spatial eigenvector mapping seek to capture the spatial configuration in additional covariates, which are then added into a generalised linear model (GLM). 2) Generalised least squares (GLS) methods fit a variance-covariance matrix based on the non-independence of spatial observations. Simultaneous autoregressive models (SAR) and conditional autoregressive models (CAR) do the same but in different ways to GLS, and the generalised linear mixed models (GLMM) we employ for non-normal data are a generalisation of GLS. 3) Generalised estimating equations (GEE) split the data into smaller clusters before also modelling the variance-covariance relationship. For comparison, the following non-spatial models were also employed: simple GLM and trend-surface generalised additive models (GAM: Hastie and Tibshirani 1990, Wood 2006), in which geographical location was fitted using splines as a trend-surface (as a two-dimensional spline on geographical coordinates). Trend surface GAM does not address the problem of spatial autocorrelation, but merely accounts for trends in the data across larger geographical distances (Cressie 1993). A promising tool which became available only recently is the use of wavelets to remove spatial autocorrelation (Carl and Kühn 2007b). However, the method was published too recently to be included here and hence awaits further testing.

We also did not include Bayesian spatial models in this review. Several recent publications have employed this method and provide a good coverage of its implementation (Osborne et al. 2001, Hooten et al. 2003, Thogmartin et al. 2004, Gelfand et al. 2005, Kühn et al. 2006, Latimer et al. 2006). The Bayesian approach to spatial models used in these studies is based either on a CAR or an autologistic implementation similar to the one we used as a frequentist method. The Bayesian framework allows for a more flexible incorporation of other complications (observer bias, missing data, different error distributions) but is much more computer-intensive then any of the methods presented here.

Beyond the methods mentioned above, there are also those which correct test statistics for spatial autocorrelation. These include Dutilleul's modified t-test (Dutilleul 1993) or the CRH-correction for correlations (Clifford et al. 1989), randomisation tests such as partial Mantel tests (Legendre and Legendre 1998), or strategies employed by Lennon (2000), Liebhold and Gurevitch (2002) and Segurado et al. (2006) which are all useful as a robust assessment of correlation between environmental and response variables. As these methods do not allow a correction of the parameter estimates, however, they are not considered further in this study. In the following sections we present a detailed description of all methods employed here.

1. Autocovariate models

Autocovariate models address spatial autocorrelation by estimating how much the response variable at any one site reflects response values at surrounding sites. This is achieved through a simple extension of generalised linear models by adding a distance-weighted function of neighbouring response values to the model's explanatory variables. This extra parameter is known as the autocovariate. The autocovariate is intended to capture spatial autocorrelation originating from endogenous processes such as conspecific attraction, limited dispersal, contagious population growth, and movement of censused individuals between sampling sites (Smith 1994, Keitt et al. 2002, Yamaguchi et al. 2003).

Adding the autocovariate transforms the linear predictor of a generalised linear model from its usual form, $\mathbf{y}=\mathbf{X\beta}+\boldsymbol{\varepsilon}$, to $\mathbf{y}=\mathbf{X\beta}+\rho\mathbf{A}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a vector of coefficients for intercept and explanatory variables $\mathbf{X}$; and ρ is the coefficient of the autocovariate $\mathbf{A}$.

$\mathbf{A}$ at any site i may be calculated as:

$$A_i = \sum_{j \in k_i} w_{ij} y_j \quad \text{(the weighted sum) or}$$

$$A_i = \frac{\sum_{j \in k_i} w_{ij} y_j}{\sum_{j \in k_i} w_{ij}} \quad \text{(the weighted average),}$$

where y_j is the response value of $\mathbf{y}$ at site j among site i's set of k_i neighbours; and w_{ij} is the weight given to site j's influence over site i (Augustin et al. 1996, Gumpertz et al. 1997). Usually, weight functions are related to geographical distance between data points (Augustin et al. 1996, Araújo and Williams 2000, Osborne et al. 2001, Brownstein et al. 2003) or environmental distance (Augustin et al. 1998, Ferrier et al. 2002). The weighting scheme and neighbourhood size (k) are often chosen arbitrarily, but may be optimised (by trial and error) to best capture spatial autocorrelation (Augustin et al. 1996). Alternatively, if the cause of spatial autocorrelation is known (or at least suspected), the choice of neighbourhood configuration may be informed by biological parameters, such as the species' dispersal capacity (Knapp et al. 2003).

Autocovariate models can be applied to binomial data (''autologistic regression'', Smith 1994, Augustin et al. 1996, Klute et al. 2002, Knapp et al. 2003), as well as normally and Poisson-distributed data (Luoto et al. 2001, Kaboli et al. 2006).

Where spatial autocorrelation is thought to be anisotropic (e.g. because seed dispersal follows prevailing winds or downstream run-off), multiple autocovariates can be used to capture spatial autocorrelation in different geographic directions (He et al. 2003).

2. Spatial eigenvector mapping (SEVM)

Spatial eigenvector mapping is based on the idea that the spatial arrangement of data points can be translated into explanatory variables, which capture spatial effects at different spatial resolutions. During the analysis, those eigenvectors that reduce spatial autocorrelation in the residuals best are chosen explicitly as spatial predictors. Since each eigenvector represents a particular spatial patterning, SAC is effectively allowed to vary in space, relaxing the assumption of both spatial isotropy and stationarity. Plotting these eigenvectors reveals the spatial patterning of the spatial autocorrelation (see Diniz-Filho and Bini 2005, for an example). This method could thus be very useful for data with SAC stemming from larger scale observation bias.

The method is based on the eigenfunction decomposition of spatial connectivity matrices, a relatively new and still unfamiliar method for describing spatial patterns in complex data (Griffith 2000b, Borcard and Legendre 2002, Griffith and Peres-Neto 2006, Dray et al. 2006). A very similar approach, called eigenvector filtering, was presented by Diniz-Filho and Bini (2005) based on their method to account for phylogenetic non-independence in biological data (Diniz-Filho et al. 1998). Eigenvectors from these matrices represent the decompositions of Moran's I statistic into all mutually orthogonal maps that can be generated from a given connectivity matrix (Griffith and Peres-Neto 2006). Either binary or distance-based connectivity matrices can be decomposed, offering a great deal of flexibility regarding topology and transformations. Given the non-Euclidean nature of the spatial connectivity matrices (i.e. not all sampling units are connected), both positive and negative eigenvalues are produced. The non-Euclidean part is introduced by the fact that only certain connections among sampling units, and not all, are considered. Eigenvectors with positive eigenvalues represent positive autocorrelation, whereas eigenvectors with negative eigenvalues represent negative autocorrelation. For the sake of presenting a general method that will work for either binary or distance matrices, we used a distance-based eigenvector procedure (after Dray et al. 2006) which can be summarized as follows: 1) compute a pairwise Euclidean (geographic) distance matrix among sampling units: $\mathbf{D}=[d_{ij}]$; 2) choose a threshold value t and construct a connectivity matrix using the following rule:

$$\mathbf{W} = [w_{ij}] = \begin{cases} 0 & \text{if } i = j \\ 0 & \text{if } d_{ij} > t \\ [1-(d_{ij}/4t)^2] & \text{if } d_{ij} \le t \end{cases}$$

where t is chosen as the maximum distance that maintains connections among all sampling units being connected using a minimum spanning tree algorithm

(Legendre and Legendre 1998). Because the example data we use represent a regular grid (see below), $t = 1$ and thus w_{ij} is either 0 or $1–1/4^2 = 0.9375$ in our analysis. Note that we can change 0.9375 to 1 without affecting eigenvector extraction. This would make the matrix fully compatible with a binary matrix which is the case for a regular grid. 3) Compute the eigenvectors of the centred similarity matrix: $(\mathbf{I}–\mathbf{1}\mathbf{1}^T/n)\mathbf{W}(\mathbf{I}–\mathbf{1}\mathbf{1}^T/n)$, where $\mathbf{I}$ is the identity matrix. Due to numerical precision regarding the eigenvector extraction of large matrices (Bai et al. 1996) the method is limited to ca 7000 observations depending on platform and software (but see Griffith 2000a, for solutions based on large binary connectivity matrices). 4) Select eigenvectors to be included as spatial predictors in a linear or generalised linear model. Here, a model selection procedure that minimizes the amount of spatial autocorrelation in residuals was used (see Griffith and Peres-Neto 2006 and Appendix for computational details). In this approach, eigenvectors are added to a model until the spatial autocorrelation in the residuals, measured by Moran's I, is non-significant. Our selection algorithm considered global Moran's I (i.e. autocorrelation across all residuals), but could be easily amended to target spatial autocorrelation within certain distance classes. The significance of Moran's I was tested using a permutation test as implemented in Lichstein et al. (2002). This potentially renders the selection procedure computationally intensive for large data sets (200 or more observations), because a permutation test must be performed for each new eigenvector entered into the model. Once the location-dependent, but data-independent eigenvectors are selected, they are incorporated into the ordinary regression model (i.e. linear or generalized linear model) as covariates. Since their relevance has been assessed during the filtering process model simplification is not indicated (although some eigenvectors will not be significant).

3. Spatial models based on generalised least squares regression

In linear models of normally distributed data, spatial autocorrelation can be addressed by the related approaches of generalised least squares (GLS) and autoregressive models (conditional autoregressive models (CAR) and simultaneous autoregressive models (SAR)). GLS directly models the spatial covariance structure in the variance-covariance matrix Σ, using parametric functions. CAR and SAR, on the other hand, model the error generating process and operate with weight matrices that specify the strength of interaction between neighbouring sites.

Although models based on generalised least squares have been known in the statistical literature for decades (Besag 1974, Cliff and Ord 1981), their application in ecology has been very limited so far (Jetz and Rahbek 2002, Keitt et al. 2002, Lichstein et al. 2002, Dark 2004, Tognelli and Kelt 2004). This is most likely due to the limited availability of appropriate software that easily facilitates the application of these kinds of models (Lichstein et al. 2002). With the recent development of programs that fit a variety of GLS (Littell et al. 1996, Pinheiro and Bates 2000, Venables and Ripley 2002) and autoregressive models (Kaluzny et al. 1998, Bivand 2005, Rangel et al. 2006), however, the range of available tools for ecologists to analyse spatially autocorrelated normal data has been greatly expanded.

Generalised least squares (GLS)

As before, the underlying model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with the error vector $\boldsymbol{\varepsilon} = N(0,\boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is called the variance-covariance matrix. Instead of fitting individual values for the variance-covariance matrix $\boldsymbol{\Sigma}$, a parametric correlation function is assumed. Correlation functions are isotropic, i.e. they depend only on the distance s_{ij} between locations i and j, but not on the direction. Three frequently used examples of correlation functions C(s) also used in this study are exponential ($C(s) = \sigma^2 \exp(-r/s)$), Gaussian ($C(s) = \sigma^2 \exp(-r/s))^2$) and spherical ($C(s) = \sigma^2(1 - 2/\pi(r/s\sqrt{1 - r^2/s^2} + \sin^{-1} r/s))$, where r is a scaling factor that is estimated from the data).

Some restrictions are placed upon the resulting variance-covariance matrix Σ: a) it must be symmetric, and b) it must be positive definite. This guarantees that the matrix is invertible, which is necessary for the fitting process (see below). The choice of correlation function is commonly based on a visual investigation of the semi-variogram or correlogram of the residuals.

Parameter estimation is a two-step process. First, the parameters of the correlation function (i.e. scaling factor r in the examples used here) are found by optimizing the so called profiled log-likelihood, which is the log-likelihood where the unknown values for $\boldsymbol{\beta}$ and σ^2 are replaced by their algebraic maximum likelihood estimators. Secondly, given the parameterization of the variance-covariance matrix, the values for $\boldsymbol{\beta}$ and σ^2 are found by solving a weighted ordinary least square problem:

$$\left(\sum^{-1/2}\right)^T \mathbf{y} = \left(\sum^{-1/2}\right)^T \mathbf{X}\boldsymbol{\beta} + \left(\sum^{-1/2}\right)^T \boldsymbol{\varepsilon}$$

where the error term $(\boldsymbol{\Sigma}^{-1/2})^T\boldsymbol{\varepsilon}$ is now normally distributed with mean 0 and variance $\sigma^2\mathbf{I}$.

Autoregressive models

Both CAR and SAR incorporate spatial autocorrelation using neighbourhood matrices which specify the

relationship between the response values (in the case of CAR) or residuals (in the case of SAR) at each location (i) and those at neighbouring locations (j) (Cressie 1993, Lichstein et al. 2002, Haining 2003). The neighbourhood relationship is formally expressed in a $n \times n$ matrix of spatial weights ($\mathbf{W}$) 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. Usually, a binary neighbourhood matrix $\mathbf{N}$ is formed where $n_{ij} = 1$ when observation j is a neighbour to observation i. This neighbourhood can be identified by the adjacency of cells on a grid map, or by Euclidean or great circle distance (e.g. the distance along earth's surface), or predefined according to a specific number of neighbours (e.g. a neighbourhood distance of 1.5 in our case includes the 8 adjacent neighbours). The elements of $\mathbf{N}$ can further be weighted to give closer neighbours higher weights and more distant neighbours lower weights. The matrix of spatial weights $\mathbf{W}$ consists of zeros on the diagonal, and weights for the neighbouring locations (w_{ij}) in the off-diagonal positions. A good introduction to the CAR and SAR methodology is given by Wall (2004).

Conditional autoregressive models (CAR)

The CAR model can be written as (Keitt et al. 2002):

$$\mathbf{Y} = \mathbf{X\beta} + \rho\mathbf{W}\,(\mathbf{Y} - \mathbf{X\beta}) + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} = N(0, \mathbf{V}_c)$. If $\sigma_i^2 = \sigma^2$ for all locations i, the covariance matrix is $\mathbf{V}_C = \sigma^2\,(\mathbf{I} - \rho\mathbf{W})^{-1}$, where $\mathbf{W}$ has to be symmetric. Consequently, CAR is unsuitable when directional processes such as stream flow effects or prevalent wind directions are coded as non-Euclidean distances, resulting in an asymmetric covariance matrix. In such situations, the closely related simultaneous autoregressive models (SAR) are a better option, as their $\mathbf{W}$ need not be symmetric (see below). For our analysis, we used a row-standardised binary weights matrix for a neighbour-distance of 2 (Appendix).

Simultaneous autogressive models (SAR)

SAR models can take three different forms (we use the notation presented in Anselin 1988), depending on where the spatial autoregressive process is believed to occur (see Cliff and Ord 1981, Anselin 1988, Haining 2003, for details). The first SAR model assumes that the autoregressive process occurs only in the response variable (''lagged-response model''), and thus includes a term ($\rho\mathbf{W}$) for the spatial autocorrelation in the response variable $\mathbf{Y}$, but also the standard term for the predictors and errors ($\mathbf{X\beta} + \boldsymbol{\varepsilon}$) as used in an ordinary least squares (OLS) regression. Spatial autocorrelation in the response may occur, for example, where propagules disperse passively with river flow, leading to a directional spatial effect. The SAR lagged-response model (SAR lag) takes the form

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

(which is equivalent to $\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}$), where ρ is the autoregression parameter, $\mathbf{W}$ the spatial weights matrix, and $\boldsymbol{\beta}$ a vector representing the slopes associated with the predictors in the original predictor matrix $\mathbf{X}$.

Second, spatial autocorrelation can affect both response and predictor variables (''lagged-mixed model'', SAR mix). Ecologically, this adds a local aggregation component to the spatial effect in the lag-model above. In this case, another term ($\mathbf{WX}\boldsymbol{\gamma}$) must also appear in the model, which describes the regression coefficients ($\boldsymbol{\gamma}$) of the spatially lagged predictors ($\mathbf{WX}$). The SAR lagged-mixed model takes the form

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

Finally, the ''spatial error model'' (SAR err) assumes that the autoregressive process occurs only in the error term and neither in response nor in predictor variables. The model is most similar to the CAR, with no directionality in the error. In this case, the usual OLS regression model ($\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$) is complemented by a term ($\lambda\mathbf{W}\boldsymbol{\mu}$) which represents the spatial structure ($\lambda\mathbf{W}$) in the spatially dependent error term ($\boldsymbol{\mu}$). The SAR spatial error model thus takes the form

$$\mathbf{Y} = \mathbf{X\beta} + \lambda\mathbf{W}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

where λ is the spatial autoregression coefficient, and the rest as above. SAR and CAR are related to each other, but the terms $\rho\mathbf{W}$ used in both CAR and SAR are not identical. As noted above, in CAR, $\mathbf{W}$ must be symmetrical, whereas in SAR it need not be. Let $\rho\mathbf{W}$ of the CAR be called $\mathbf{K}$ and $\rho\mathbf{W}$ of the SAR be called $\mathbf{S}$. Then any SAR is a CAR with $\mathbf{K} = \mathbf{S} + \mathbf{S}^T - \mathbf{S}^T\mathbf{S}$ (Haining 2003). Assuming constant variance σ^2, the formal relationship between the error variance-covariance matrices in GLS, SAR, and CAR is as follows: $\mathbf{V}_{GLS} = \sigma^2\mathbf{C}(s)$; $\mathbf{V}_{CAR} = \sigma^2\,(\mathbf{I}–\mathbf{K})^{-1}$ and $\mathbf{V}_{SAR} = \sigma^2\,(\mathbf{I}–\mathbf{S})^{-1}(\mathbf{I}–\mathbf{S}^T)^{-1}$, with $\mathbf{K}$ and $\mathbf{S}$ as defined above. Thus CAR and SAR models are equivalent if $\mathbf{V}_{CAR} = \mathbf{V}_{SAR}$. The relationship between specific values in correlation matrix $\mathbf{C}$ and weight matrix $\mathbf{W}$ is not straightforward, however. In particular, spatial dependence parameters that decrease monotonically with distance do not necessarily correspond to spatial covariances that decrease monotonically with distance (Waller and Gotway 2004). An extensive comparison of the impact of different model formulations on parameter estimation and type I error control is given by Kissling and Carl (2007) using simulated datasets with different spatial autocorrelation structures.

Spatial generalised linear mixed models (GLMM)

Spatial generalised linear mixed models are generalised linear models (GLMs) in which the linear predictor may contain random effects and within-group errors may be spatially autocorrelated (Breslow and Clayton 1993, Venables and Ripley 2002). Formally, if Y_{ij} is the j-th observation of the response variable in group i,

$$E[Y_{ij}|\zeta_i] = g^{-1}(\eta_{ij}) \qquad \text{and} \qquad \eta_{ij} = x_{ij}\boldsymbol{\beta} + z_{ij}\zeta_i,$$

where g is the link function, η is the linear predictor, $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ are coefficients for fixed and random effects, respectively, and x and z are the explanatory variables associated with these effects. Conditional on the random effects $\boldsymbol{\zeta}$, the standard GLM applies and the within-group distribution of $\mathbf{Y}$ can be described using the same error distributions as in GLM.

Since the GLMM is often implemented based on so-called penalized quasi-likelihood (PQL) methods (Breslow and Clayton 1993, Venables and Ripley 2002) around the GLS-algorithm (McCullough and Nelder 1989), we can use it in a similar way, i.e. fitting the structure of the variance-covariance-matrix to the data (see GLS above), albeit with a different error distribution. In cases where spatial data are available from several disjunct regions, GLMMs can thus be used to fit overall fixed effects while spatial correlation structures are nested within regions, allowing the accommodation of regional differences in e.g. autocorrelation distances, and assuming autocorrelation only between observations within the same region (Orme et al. 2005, Davies et al. 2006, Stephenson et al. 2006).

4. Spatial generalised estimating equations (GEE)

Liang and Zeger (1986) developed the generalised estimating equation (GEE) approach which is an extension of generalised linear models (GLMs). When responses are measured repeatedly through time or space, the GEE method takes correlations within clusters of sampling units into account by means of a parameterised correlation matrix, while correlations between clusters are assumed to be zero. In a spatial context such clusters can be interpreted as geographical regions, if distances between different regions are large enough (Albert and McShane 1995). We modified the approach of Liang and Zeger to use these GEE models for spatial, two-dimensional datasets sampled in rectangular grids (see Carl and Kühn 2007a, for more details). Fortunately, estimates of regression parameters are fairly robust against misspecification of the correlation matrix (Dobson 2002). The GEE approach is especially suited for parameter estimation rather than prediction (Augustin et al. 2005).

Firstly, consider the generalised linear model $E(\mathbf{y}) = \boldsymbol{\mu}$, $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ where $\mathbf{y}$ is a vector of response variables, $\boldsymbol{\mu}$ the expected value, g^{-1} the inverse of the link function, $\mathbf{X}$ the matrix of predictors, and $\boldsymbol{\beta}$ the vector of regression parameters. Minimization of a quadratic form leads to the GLM score equation (Diggle et al. 1995, Dobson 2002, Myers et al. 2002)

$$\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = 0,$$

where $\mathbf{D}^T$ is the transposed matrix of $\mathbf{D}$ of partial derivatives $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$. Secondly, note that the variance of the response can be replaced by a variance-covariance matrix $\mathbf{V}$ which takes into account that observations are not independent. In GEEs, the sample is split up into m clusters and the complete dataset is ordered in a way that in all clusters data are arranged in the same sequence: $E(\mathbf{y}_j) = \boldsymbol{\mu}_j$, $\boldsymbol{\mu}_j = g^{-1}(\mathbf{X}_j\boldsymbol{\beta})$. Then the variance-covariance matrix has block diagonal form, since responses of different clusters are assumed to be uncorrelated. One can consequently transform the score equation into the following form

$$\sum_{j=1}^{m} \mathbf{D}_j^T\mathbf{V}_j^{-1}\ (\mathbf{y}_j - \boldsymbol{\mu}_j) = 0,$$

which sums over all clusters j. This equation is called the generalised estimating equation or the quasi-score equation.

For spatial dependence the following correlation structures for $\mathbf{V}$ are important: 1) Fixed. The correlation structure is completely specified by the user and will not change during an iterative procedure. Referred to here as GEE. 2) User defined. Correlation parameters are to be estimated, but one can specify that certain parameters must be equal, e.g. that the strength of correlation is always the same at a certain distance. Referred to here as geese.

First, we consider the GEE model with fixed correlation structure. In order to predetermine the correlation structure we have good reasons to assume that the correlation decreases exponentially with increasing spatial distance in ecological applications. Therefore, we use the function

$$\alpha = \alpha_1^{d_{ij}}$$

for computation of correlation parameters α. Here d_{ij} is the distance between centre points of grid cells i and j and α_1 is the correlation parameter for nearest neighbours. The parameter α_1 is estimated by Moran's I of GLM residuals. In this way we obtain a full $n \times n$ correlation matrix with known parameters. Thus clustering is not necessary.

In the user defined case we build a specific variance-covariance matrix in block diagonal form with 5 unknown correlation parameters (corresponding to the five different distance classes in a 3×3 grid) which

have to be calculated iteratively. The dispersion parameter as a correction of overdispersion can be calculated as well.

Example analysis using simulated data

To illustrate and compare the various approaches that are available to incorporate SAC into the analysis of species distribution data, we constructed artificial datasets with known properties. The datasets represent virtual species distribution data (for example species atlases) and environmental (such as climatic) covariates, available on a lattice of 1108 square cells imposed on the surface of a virtual island (Fig. 3).

Generation of artificial distribution data

The basis for the virtual island is a subset of the volcano data set in R, which consists of a digital elevation model for Auckland's Maunga Whau Volcano in New Zealand (Anon. 2005). Two uncorrelated (Pearson's $r = 0.013$, $p = 0.668$) environmental variables were created based on the altitude-component of this data set: ''rain'' and ''djungle''. These data are available as electronic appendix and are depicted in Fig. 3. While ''rain'' is a rather deterministic function of altitude (including a rain-shadow in the east), ''djungle'' is dominated by a high noise component. Data are given in the Appendix.

On this lattice the species distribution data, y_i (with i an indicator for cell ($i = 1, 2 \ldots, 1108$)), were simulated as a function of one of the two artificial environmental predictors, $rain_i$. Onto this functional relationship, we added a spatially correlated noise component we refer to as error ε_i. The covariate $rain_i$ can for example be thought of as estimates of the total annual amount of rainfall in cell i. We simulated the three most commonly available types of species distribution data; continuous, binary and count data, using the normal distribution and approximations of the Poisson and binomial distributions respectively. The following models were used to simulate the artificial data: 1) normally distributed data: $y_i = 80 - 0.015 \times rain_i + 10 \times \varepsilon_i$. 2) Binary data: $y_i = 0$ if $p_i < 0.5$, and $y_i = 1$ if, $p_i \geq 0.5$, where $p_i = q_i + \sqrt{q_i(1-q_i)}\varepsilon_i$, and $q_i = \frac{e^{3-0.003*rain_i}}{1 + e^{3-0.003*rain_i}}$. 3) Poisson data: $y_i = \text{round}(k_i + \sqrt{k_i}\varepsilon_i)$, where $k_i = e^{3-0.001*rain_i}$, and round is an operator used to round values to the nearest integer. This led to simulated data with no over- or underdispersion.

A weight matrix $\mathbf{W}$ was used to simulate the spatially correlated errors ε_i using weights according to the distance between data points. Let $\mathbf{D} = (d_{ij})$ be the (Euclidean) distance matrix for the distances between cells i and j ($d_{ij} = 0$ if $i = j$). On our lattice, the distance between the mid-points of neighbouring cells is $d_{ij} = 1$. Then, $\mathbf{\Omega} = (\omega_{ij})$ is a matrix defined as $\omega_{ij} = \exp(-\rho \times d_{ij})$, ρ ($\rho \geq 0$) is a parameter that determines the decline of inter-cell correlation in errors with inter-cell distance. The strength of spatial autocorrelation increases with increasing values of ρ (there is no spatial autocorrelation if $\rho = 0$). Here, we used a value of $\rho = 0.3$, which resulted in strongly correlated errors in neighbouring cells ($\omega_{ij} = 0.74$, if $d_{ij} = 1$), but a steep decline of autocorrelation with increasing distance. A weights matrix $\mathbf{W}$ was calculated (by Choleski decomposition) using $\mathbf{\Omega} = \mathbf{W}^T\mathbf{W}$. Finally, the spatially correlated errors are given by $\boldsymbol{\varepsilon} = \mathbf{W}^T\boldsymbol{\xi}$, with $\boldsymbol{\xi}$ drawn from the standard normal distribution.

Analysis of simulated data

For each error distribution, ten data sets were created, each using a random realisation of the spatially autocorrelated errors, using random draws of ξ_i. These data sets were then submitted to statistical analyses in which the response variables were modelled using a number of different linear models for the normally distributed data, and generalized linear models with the binomial distribution and logit-link for the binary data, and Poisson distribution and log-link for the count data: $E(y_i) = g^{-1}(\alpha + \beta \times rain_i + \gamma \times djungle_i)$, where g are the corresponding link functions (identity for the normal distribution). The variable ''djungle'' was entered into all of the statistical models as an additional predictor of the response. This was done to be able to assess the models' ability to distinguish random noise from meaningful variables.

Simulations and analyses were primarily carried out (see Appendix for implementation details and R-code) using the statistical programming software R (Anon. 2005), with packages gee (Carey 2002), geepack (Yan 2002, 2004), spdep (Bivand 2005), ncf (Bjørnstad and Falck 2000) and MASS (Venables and Ripley 2002). Calculations for the spatial eigenvector mapping were originally performed in Matlab using routines later ported to R (spdep) by Roger Bivand and Pedro Peres-Neto. Additional functions (Appendix) to work generalised estimating equations on a 2-D lattice were written by Gudrun Carl (Carl and Kühn 2007a). See also Table 1 for alternative software.

As most of the statistical methods tested allow for some flexibility in the precise structure of their spatial component, several models per method were calculated for each simulated dataset. This allowed us to identify the model configuration that most successfully accounted for spatial autocorrelation in the data at hand, by, for example, varying the distance over which spatial autocorrelation was assumed to occur, or its

Table 1. Methods correcting for spatial autocorrelation and their software implementations. This list is not exhaustive but represents the major software developments in use.

Method	R-package[1]	Computational intensity[2]	Other[3]
Autocovariate regression	spdep	low	
Autoregressive models[4] (CAR, SAR)	spdep	medium	GeoDa, Matlab*, SAM, SpaceStat, S-plus†
Bayesian analysis		very high	WinBUGS/GeoBUGS
Generalised linear mixed model	MASS	very high	SAS (glimmix)
Generalised estimating equations	gee, geepack	low	SAS
Generalised least squares[4]	MASS, nlme	high	SAS, SAM
Spatial eigenvector mapping	spdep	very high	Matlab, SAM

[1] for most R-packages (<http://www.r-project.org>) an equivalent for S-plus is available.
[2] low, medium, high and very high refer roughly to a few seconds, several minutes, a few hours and several hours of CPU-time per model (1108 data points on a Pentium 4 dual core, 3.8GHz, 2GB RAM).
[3] GeoDa: freeware: <http://www.geoda.uiuc.edu>.
[4] for normally distributed error only.
Matlab: <http://www.mathworks.com>, with EigMapSel – a matlab compiled software to perform the eigenvector selection procedure for generalised linear models (normal, logistic and poisson) – available in ESA's Electronic Data Archive (Griffith and Peres-Neto 2006).
SAM: spatial analysis for macroecology; freeware under: <http://www.ecoevol.ufg.br/sam/>.
SAS: statistical analysis system; commercial software: <http://www.sas.com>.
SpaceStat: commercial software: <http://www.terraseer.com/products/spacestat.html>.
S-plus: commercial software: <http://www.insightful.com>.
WinBUGS/GeoBUGS: freeware: <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>.
*requires the free ''Spatial Econometric toolbox'': <http://www.spatial-econometrics.com>.
†requires additional module ''spatial''.

functional form. Inferior models were discarded, so that the results section below reports only on the best configuration for each approach. We used residuals based on fitted values and which were as such calculated from both the spatial and the non-spatial model components. For each, we report the following details: 1) model coefficients (and their standard errors); since the true parameters are known, we can directly judge the quality of coefficient estimation; 2) removal of SAC (global Moran's I, i.e. Moran's I computed across neighbourhood up to a distance of 20, and correlograms, which plot Moran's I for different distance classes); 3) spatial distribution of residuals (map).

Results of simulations

It is worth pointing out that the main aim of this study is to illustrate the different methods by applying them to the same data sets. The ten realisations of one type of spatial autocorrelation do not allow us to provide a comprehensive evaluation of the relative merits of each of the methods considered. Such evaluation is beyond the scope of this review paper, and will depend on the data set and question under study. Nonetheless, some interesting results emerged from our simulations.

Spatial and non-spatial models differed considerably in terms of the spatial signature in their residuals (Table 2; Fig. 2 and 3). Residual maps for OLS/GLM and GAM exhibit clusters of large residuals of the same sign (Fig. 3), indicating that these models were not able to remove all spatial autocorrelation from the data. In our case we know that this is due neither to the omission of an important variable nor an incorrect functional relationship, but a simulated aggregation mechanism in the errors. In comparison, all spatial models managed to decrease spatial autocorrelation in the residuals (Fig. 2), although not all were able to completely eliminate it. Geese performed worst in this regard. Our simulations are not comprehensive enough, however, to allow us to deduce what the influence of this incomplete removal of SAC might be on parameter estimation or hypothesis testing.

Another – though inconsistent – difference between the spatial and non-spatial models especially with binary data was that standard errors of the coefficient estimates for ''rain'' and ''djungle'' were often larger for the spatial models (Fig. 1). For normal and Poisson data, differences in coefficient estimates between spatial and non-spatial models were relatively small, and statistical inference was not affected. Only autocovariate model and SAR lag provided consistently incorrect estimates of the spatially autocorrelated parameter ''rain''.

Most model approaches performed well with respect to type I and II error rates for the normal and Poisson data, correctly identifying ''rain'' as a significant effect (Table 2). An exception was autocovariate regression, which severely and consistently underestimated the effects of rain (Table 2, Fig. 1). Model performance was worse for data with a binomial error structure than for models with normal or Poisson error structure.

Table 2. Model quality: spatial autocorrelation in the model residuals (given as global Moran's I) and mean estimates for the coefficients "rain" and "djungle" (±1 SE across the 10 simulations). True coefficient values are given in the first row for each distribution in italics. ***, ** and ^{ns} refer to median significance levels of $p < 0.001$, < 0.01 and > 0.1, respectively, across the 10 realisations. See Fig. 1 for abbreviations.

		Moran's I	Coefficients	
			"rain"	"djungle"
Normal			*−0.015*	*0.0*
	GLM	0.016 ± 0.026	$-0.0143 \pm 0.0010^{***}$	0.0220 ± 0.0508^{ns}
	GAM	−0.002 ± 0.002	$-0.0125 \pm 0.0028^{***}$	0.0130 ± 0.0370^{ns}
	autocov	−0.001 ± 0.000	-0.0004 ± 0.0007^{ns}	0.0141 ± 0.0309^{ns}
	GLS exp	−0.001 ± 0.000	$-0.0140 \pm 0.0033^{***}$	0.0162 ± 0.0248^{ns}
	CAR	0.000 ± 0.000	$-0.0145 \pm 0.0022^{***}$	0.0156 ± 0.0324^{ns}
	SAR err	−0.001 ± 0.000	$-0.0144 \pm 0.0028^{***}$	0.0156 ± 0.0253^{ns}
	GEE	−0.001 ± 0.001	$-0.0141 \pm 0.0031^{***}$	0.0162 ± 0.0255^{ns}
	geese	0.002 ± 0.005	$-0.0141 \pm 0.0017^{***}$	0.0256 ± 0.0276^{ns}
	SEVM	−0.001 ± 0.001	$-0.0132 \pm 0.0009^{***}$	0.0163 ± 0.0257^{ns}
Binomial			*−0.003*	*0.0*
	GLM	0.006 ± 0.011	$-0.0022 \pm 0.0003^{***}$	0.0052 ± 0.0130^{ns}
	GAM	−0.002 ± 0.001	$-0.0006 \pm 0.0013^{**}$	0.0016 ± 0.0162^{ns}
	autocov	−0.001 ± 0.001	-0.0006 ± 0.0004^{ns}	0.0029 ± 0.0167^{ns}
	GLMM	0.001 ± 0.001	$-0.0042 \pm 0.0006^{***}$	0.0025 ± 0.0096^{ns}
	GEE	−0.001 ± 0.001	$-0.0021 \pm 0.0007^{***}$	0.0024 ± 0.0093^{ns}
	geese	0.000 ± 0.003	$-0.0021 \pm 0.0004^{***}$	0.0048 ± 0.0101^{ns}
	SEVM	−0.001 ± 0.000	$-0.0028 \pm 0.0006^{***}$	0.0084 ± 0.0190^{ns}
Poisson			*−0.001*	*0.0*
	GLM	0.018 ± 0.024	$-0.0010 \pm 0.0000^{***}$	0.0006 ± 0.0018^{ns}
	GAM	0.002 ± 0.002	$-0.0005 \pm 0.0001^{***}$	0.0005 ± 0.0018^{ns}
	autocov	0.010 ± 0.010	$-0.0001 \pm 0.0001^{*}$	0.0010 ± 0.0018^{ns}
	GLMM	0.001 ± 0.001	$-0.0010 \pm 0.0001^{***}$	0.0006 ± 0.0009^{ns}
	GEE	0.001 ± 0.001	$-0.0010 \pm 0.0001^{***}$	0.0006 ± 0.0009^{ns}
	geese	0.005 ± 0.005	$-0.0010 \pm 0.0001^{***}$	0.0008 ± 0.0010^{ns}
	SEVM	0.001 ± 0.001	$-0.0010 \pm 0.0001^{***}$	0.0008 ± 0.0019^{ns}

When applied to such data, autocovariate regression (9 false negatives) and GAM (3 false negatives) were rather prone to type II errors (results not shown). Moreover, the spurious effect of djungle would have been retained in the model in several cases (based on a significance level of $\alpha = 0.05$: 6 normal, 2 binomial and 1 Poisson model of those presented in Table 2), resulting in type I errors (rejecting a null hypothesis although it was true).

The ability of simultaneous autoregressive models (SAR) to correctly estimate parameters depended heavily on SAR model structure. For instance, using a lagged response model in our artificial dataset yielded much poorer coefficient estimates for "rain" than using an error model (Fig. 1). This was to be expected, since our artificial distribution data was created such that its spatial structure most closely resembled that of the SAR error model.

We used an exponential distance decay function to generate the spatial error (see above). Hence, we would also expect those methods to perform best in which a correlation function can be defined accordingly (i.e. GLS, GLMM and GEE). While indeed the exponential GLS yielded better coefficient estimates than the spherical model, the Gaussian model and the GEE using a different exponential function were equivalent, as were methods that did not specify the correlation structure in such a way (e.g. SAR, Fig. 2). However, parameterisation for GEE resulted from the Moran's I correlogram, mimicking the distance decay function, though not using the original correlation function.

Limitations of our simulations

Our example analysis above was meant to illustrate the application of the presented methods to species distribution data. As such, it remained a cartoon of the complexity and difficulties posed by real data. Among the potential factors that may influence the analysis of species distribution data with respect to spatial autocorrelation, we like to particularly mention the following.

Missing environmental variables. As mentioned in the introduction, SAC can be caused by omitting an important variable from the model or misspecifying its functional relationship with the response (Legendre

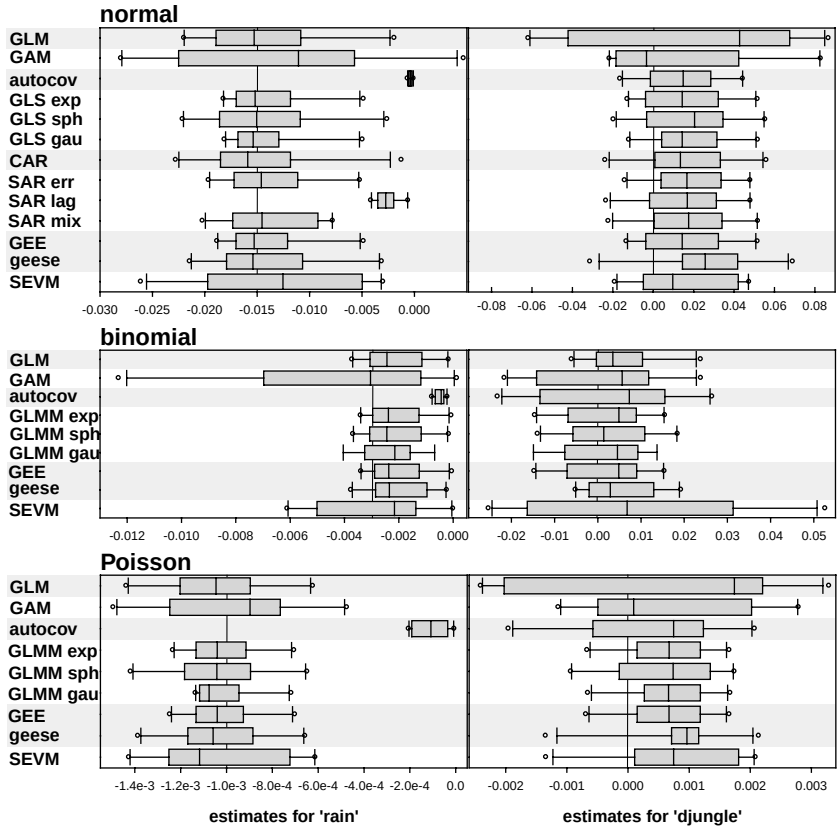

Fig. 1. Comparison of non-spatial (GLM) and spatial modelling approaches for data with normally, binomially and Poisson-distributed errors. Box, whiskers and dots refer to 25/75%, 10/90% and outliers of estimates across 10 realisations of the same parameter set. Vertical lines indicate true values of parameter. GEE and geese refer to generalised estimating equations with fixed and user-defined correlation structures, respectively. GAM represents a trend-surface regression. GLS and GLMM refer to generalised least squares-based models with exponential, spherical and Gaussian correlation structure, respectively. SAR (simultaneous autoregressive model) was analysed as error, lag and mixed models. SEVM stands for the spatial eigenvector mapping approach.

1993). This is certainly often a problem in real data, where the ecological determinants of a species' niche are not necessarily known and good spatial coverage may not be available for all the important factors. Also, moderate collinearity among environmental variables may lead models to exclude one or more variables which would be important in explaining the species' spatial patterning.

Biased spatial error. The autocorrelated error we added in our simulated data had no bias in geographical (stationarity of the error) or parameter space. Hence our non-spatial models performed similar to the spatial ones with regards to parameter estimation, as opposed to removal of SAC. This may or may not be very different in real data, where both non-stationarity (Ver Hoef et al. 1993, Brunsdon et al. 1996, Foody 2004, Osborne et al. 2007) and bias in parameter space (e.g. less complete data coverage in warmer regions) can be found (Lennon 2000, but see Hawkins et al. 2007 for an opposite view).

Mapping bias or mapping heterogeneity can cause spatial autocorrelation in real data. If real data resulted from several different regional mapping schemes with different protocols or from different people performing the mapping, data can differ systematically across a grid with being more similar within a region and more different across.

Spatial autocorrelation at different spatial scales. Several of the methods presented build a ''correction structure'' across all spatial scales (i.e. the variance-covariance matrices in GLS-based models as well as the spatial eigenvectors), but others do not (the autocovariate and the cluster in geese have one specific spatial scale). Even the former may be dominated by patterns at one spatial scale, underestimating effects of another.

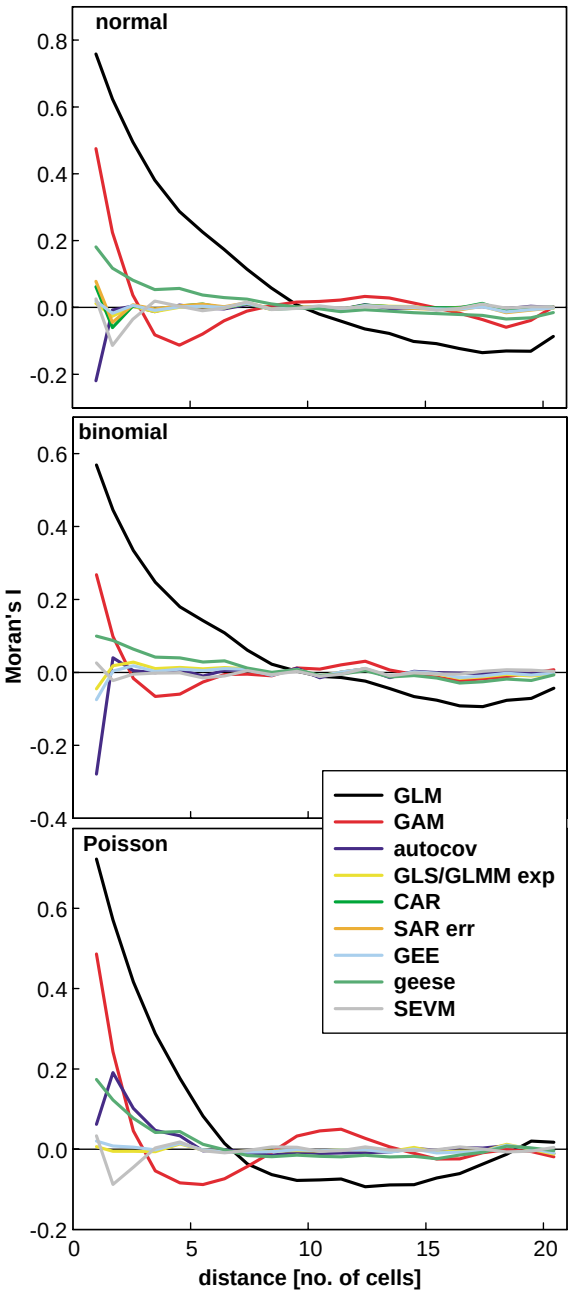

Fig. 2. Correlograms of one realisation for each of the three different distributions (normal, binomial, Poisson) and the methods compared. See Fig. 1 for abbreviations.

Finally, small sample sizes make the estimation of model parameters unstable. Adding the additional parameters for spatial models will further destabilise model parameterisation. Also, patterns of SAC in small data sets will hinge on very few data points which may distort the spatial correction.

Discussion

The analysis of species distribution data has reached high statistical sophistication in recent years (Elith et al. 2006). However, even the most advanced and computer-intensive statistical procedures are no guarantee for improving our understanding of the determinants of species distributions, nor of our ability to predict species distributions under altered environmental conditions (Araújo and Rahbek 2006, Dormann 2007c). One critical step in statistical modelling is the identification of the correct model structure. As pointed out for experimental ecology in 1984 by Hurlbert, designs analysed without consideration of the nested nature of subsampling are fundamentally flawed. Spatial autocorrelation is a subtle, less obvious form of subsampling (Fortin and Dale 2005): samples from within the range of spatial autocorrelation around a data point will add little independent information (depending on the strength of autocorrelation), but unduly inflate sample size, and thus degrees of freedom of model residuals, thereby influencing statistical inference.

We have presented an overview of different modelling approaches for the analysis of species distribution data in which environmental correlates of the distribution are inferred. All these methods can be implemented in freely available software packages (Table 1). In choosing between the methods, the type of error distribution in the response variable will be an important criterion. For normal data, GLS-based methods (GLS, SAR, CAR) can be used efficiently. The most flexible methods, addressing SAC for different error distributions, are spatial GLMMs, GEEs and SEVM. The autocovariate method, too, is flexible, but performed very poorly with regards to coefficient estimation in our analyses. We encourage users to try a number of methods, since there is often not enough mechanistic information to choose one specific method a priori. One can use AIC or alike to compare models (Link and Barker 2006). Note that a "proper" (perfectly correctly specified) model would not require the kind of correction the above methods undertake (Ripley in comments to Besag 1974). In the absence of a perfect model, however, doing something is better than doing nothing (Keitt et al. 2002).

With the exception of autocovariate regression, differences in parameter estimates and inference between spatial and non-spatial models were small for our simulated data. This was possibly a result of the type of spatial autocorrelation in, and the simplistic nature of, these data (see section "Limitations of our simulations"). However, spatial autocorrelation can also reflect failure to include an important environmental driver in the analysis or inadequate capture of its non-linear effect, so that its spatial autocorrelation cannot be accounted for by non-spatial models (Besag et al. 1991,

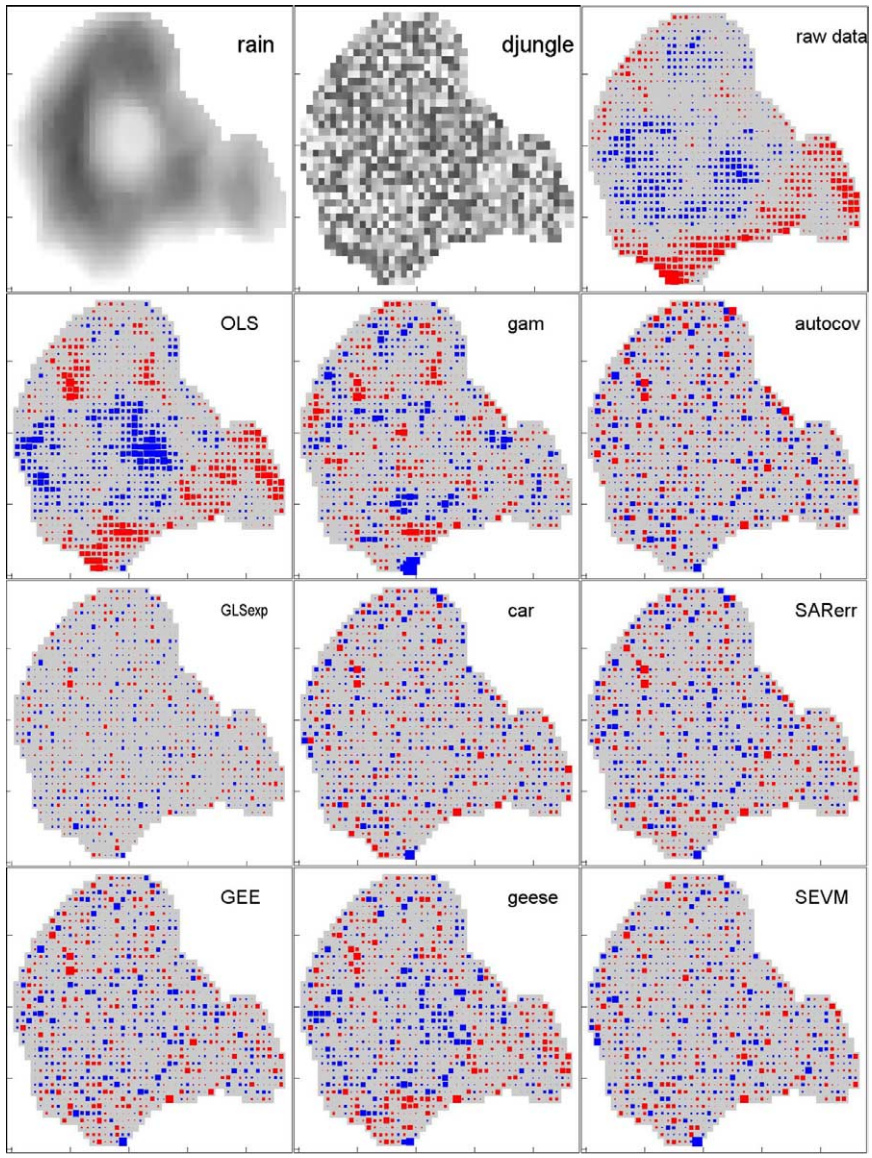

Fig. 3. The two environmental covariates, raw distributional data and residual maps for the different methods for one realisation of the data with normally distributed errors. Blue indicates positive, red negative values. Quadrat size is proportional to the value of the residual (or raw data value), but scaling is different for every plot (since a comparison of quadrat sizes would simply be a comparison of model fit). See Fig. 1 for abbreviations.

Legendre et al. 2002). In either case, spatial autocorrelation can make a large difference for statistical inference based on spatial data (for review see Dormann, 2007a, b, c; for drastic cases of this effect see Tognelli and Kelt 2004 and Kühn 2007). How to interpret these differences, especially the shifts in parameter estimates between spatial and non-spatial models commonly observed in real data, remains controversial. While Lennon (2000) and others (Tognelli and Kelt 2004, Jetz et al. 2005, Dormann 2007b, Kühn 2007) argue that spatial autocorrelation in species distribution models may well bias coefficient estimation, Diniz-Filho et al. (2003) and Hawkins et al. (2007) found non-spatial model to be robust and unbiased for several data sets. So far, no extensive simulation study has been carried out to investigate how spatial versus non-spatial methods perform under different forms and causes of SAC. Implementing a lagged autocorrelation structure to simplistic data did not reveal a bias in parameter estimation in OLS (Kissling and Carl 2007), consistent with the results of Hawkins et al. (2007).

One of the two most striking findings of our analyses is the high error rate of the autocovariate method. Most methods for normally distributed data yielded coefficient estimates for "rain" that were acceptable, including the non-spatial ordinary least square regression (Fig. 1). However, two models performed poorly: both the autocovariate regression and the lag version of the simultaneous autoregressive model showed a very consistent and strong bias, leading to severe underestimation (in absolute terms) of model coefficients. A similar pattern was also found for the non-normally distributed errors, identifying autocovariate regression as a consistently worse performer than the other approaches. The poor performance of the autocovariate regression approach in our study with regards to parameter estimation contrasts with earlier evaluations of this method (Augustin et al. 1996, Huffer and Wu 1998, Hoeting et al. 2000, He et al. 2003), but is in line with more recent ones (Dormann 2007a, Carl and Kühn 2007a). These earlier studies used more sophisticated parameter estimation techniques, suggesting that the inferiority of autocovariate models in our simulation may partly result from our simplistic (but not unusual) implementation of the method. Moreover, two of the earlier studies were undertaken in the context of many missing values: Augustin et al. (1996) used only 20% of sites in their study area for model training; Hoeting et al. (2000) used between 3.8 and 5.8%. This may have diminished the influence of any autocovariate and perhaps explains why in these studies the autocovariate did not overwhelm other model coefficients (as it did in ours). A final reason for the discrepancy in findings may be that our artificial data simulated spatial autocorrelation in the error structure, whereas other simulations created spatial structure directly in the response values, which more closely reflects the assumptions underlying autocovariate models.

The second interesting finding is the overall higher variability of results for binary data. While for normal- and Poisson-distributed residuals all model approaches (apart from autocovariate regression) yielded similar results and little variance across the ten realisations (Fig. 1), a different pattern emerged for binary (binomial) data. We attribute this to the relatively low information content of binary data (Breslow and Clayton 1993, Venables and Ripley 2002), making parameterisation of the model very dependent on those data points that determine the point of inflexion of the logistic curve (McCullough and Nelder 1989). This phenomenon has been noted before (McCullough and Nelder 1989), and remains relevant for species distribution models, where the majority of studies are based on the analysis of presence-absence data (Guisan and Zimmermann 2000, Guisan and Thuiller 2005).

Tricks and tips

Each of the above methods has its quirks and some require fine-tuning by the analyst. Without attempting to cover these comprehensively, we here hint at some areas for each method type which require attention.

In autocovariate regression, neighbourhood size and type of weighting function are potentially sensitive parameters, which can be optimised through trial and error. It seems, however, that small neighbourhood sizes (such as the next one to two cells) often turn out best, and that the type of weighting function has relatively little effect. This was the case in our analysis as well as in published studies investigating different neighbourhood sizes (for review see Dormann 2007b). Another important aspect of autocovariate models is the approach chosen to dealing with missing data, which may lead to cells without neighbours ("islands"). Since the issue arises for all modelling methods, we shall briefly discuss it here. Missing data can be overcome by a) omission (Klute et al. 2002, Moore and Swihart 2005); b) strategic choice of neighbourhood structure (Smith 1994); c) estimating missing response values by initially ignoring spatial autocorrelation and regressing known response values against explanatory variables other than the autocovariate (Augustin et al. 1996, Teterukovskiy and Edenius 2003, Segurado and Araújo 2004); and d) as in c), but then refining it through an iterative procedure known as the Gibbs sampler (Casella and George 1992). This procedure is computationally intensive, but has been found to yield the best results (Augustin et al. 1996, Wu and Huffer 1997, Osborne et al. 2001, Teterukovskiy and Edenius 2003, Brownstein et al. 2003, He et al. 2003). Simulation studies further suggest that a) parameter estimation is poor when the autocovariate effect is strong relative to the effect of other explanatory variables (Wu and Huffer 1997, Huffer and Wu 1998); b) the precision of parameter estimates varies with species prevalence, i.e. the number of presence records relative to the total sample size (Hoeting et al. 2000); and c) autocovariate models adequately distinguish between meaningful explanatory variables and random covariates (Hoeting et al. 2000) (but not in our study). Both simulation and empirical studies also indicate that autocovariate models achieve better fit than equivalent models lacking the autocovariate term (Augustin et al. 1996, Hoeting et al. 2000, Osborne et al. 2001, He et al. 2003, McPherson and Jetz 2007).

For spatial eigenvector mapping, computational speed becomes an issue for large datasets. Although the calculation of eigenvectors itself is rapid, optimising the model by permutation-based testing combinations of spatial eigenvectors is computer-intensive. Diniz-Filho and Bini (2005) argue that the identity of the selected eigenvectors is indicative of the spatial scales at

which spatial autocorrelation takes effect, making this method potentially very interesting for ecologists. The implementation used in our analysis requires little arbitration and hence should be explored more widely. Note that SEVM, in the way that was applied here, is based on a different modelling philosophy. Its declared aim is to remove residual spatial autocorrelation, unlike all other methods described above, which simply provide a mathematical way to incorporate SAC into the analysis.

For the GLS-based methods (GLS and the spatial GLMM), estimation of the correlation structure functions (i.e. the parameter r) can be rather unstable. As a consequence some models yield $r = 0$ (i.e. no spatial autocorrelation incorporated) or $r \approx \infty$, with the GLS model returning what is in fact a non-spatial GLM or nonsensical results, respectively. This problem can be overcome by inclusion of a ''nugget'' term that reduces the correlation at infinitesimally small distances to a value below 1, or, even better, a specification of r based on a semi-variogram of the residuals (Littell et al. 1996, Kaluzny et al. 1998). The common justification for a nugget term are measurement errors (on top of the spatially correlated error); including a nugget effect can stabilize the estimation of the correlation function (Venables and Ripley 2002).

Autoregressive models (SAR and CAR) require a decision on the weighting scheme for the weights matrix, for which there is not always an a priori reason. The main options are row standardised coding (sums over all rows add up to N), globally standardised coding (sums over all links add up to N), dividing globally standardised neighbours by their number (sums over all links add up to unity), or the variance-stabilising coding scheme proposed by Tiefelsdorf et al. (1999, pp. 167–168), i.e. sums over all links to N. In our analysis, the row standardised coding was most often the superior choice, which is in line with other studies (Kissling and Carl 2007), but the binary and the variance-stabilising coding scheme also resulted in good models. SAR and CAR models did not differ much in our analysis. According to Cressie (1993), CAR models should be preferred in terms of estimation and interpretation, although SAR models are preferred in the econometric context (Anselin 1988). Either approach can be relatively slow for large data sets (sample size $> 10\,000$) due to the estimation of the determinant of $(\mathbf{I}-\rho\mathbf{W})$ for each step of the iteration. Note that Bayesian CAR models do not require the computation of such a determinant and can therefore be particularly suitable for data on large lattices (Gelfand and Vounatsou 2003). For SAR models, identification of the correct model structure is recommended and model selection procedures can help to reduce bias (Kissling and Carl 2007). The Lagrange-test (Appendix) can also help here. However, SAR error models generally perform better than SAR lag or even SAR mix models when tackling simulated data containing autocorrelation in lagged predictors (or response and predictors), as recently demonstrated in a more comprehensive assessment of SAR models using different spatially autocorrelated datasets (Kissling and Carl 2007).

Generalised estimating equations require high storage capacity for solving the GEE score equation without clustering as we used it in our fixed model. Application of the fixed model will therefore be limited for models on data with larger sample size, but the method is very suitable for missing data and non-lattice data. The need in storage capacity is considerably reduced by cluster models, such as our user-defined model. But clustering requires attention to three steps in the analysis: cluster size, within-cluster correlation structure and allocation of cells to clusters. To find the best cluster size for the analysis, we recommend investigating clusters of 2×2, 3×3 and 4×4. In real data, these cluster sizes have been sufficient to remove spatial autocorrelation (Carl and Kühn 2007a). Several different correlation structures should be computed initially, e.g. to allow for anisotropy. Finally, allocation of cells to clusters can start in different places. Depending on the starting point (e.g. top right or north west), cells will be placed in different clusters. Choosing different starting points will give the analyst an idea of the (in our experience limited) importance of this issue. Computing time is short.

Autocorrelation in a predictive setting

Spatial autocorrelation may arise for a number of ecological reasons, including external environmental and historical factors limiting the mobility of organisms, intrinsic organism-specific dispersal mechanisms and other behavioural factors causing the spatial aggregation of populations and species in the landscapes. In addition to these factors, spatial autocorrelation can also be caused by observer bias and differences in sampling schemes and sampling effort. Overall, spatial autocorrelation occurs at all spatial scales from the micrometre to hundreds of kilometres (Dormann 2007b), possibly for a whole suite of reasons. Since these reasons are mostly unknown, one cannot readily derive a spatial correlation structure for an entirely new, unobserved region. Augustin et al. (1996) and others (Hoeting et al. 2000, Teterukovskiy and Edenius 2003, Reich et al. 2004) have, however, successfully used the Gibbs sampler (Casella and George 1992) to derive predictions for unobserved areas within the study region (interpolation), and He et al. (2003) extrapolated autologistic predictions through time to examine possible effects of climate change.

Interpolation, i.e. the prediction of values within the parameter and spatial range, can be achieved by several of the presented methods. An advantage of GLS is that the spatially correlated error can be predicted for sites where no observations are available, based on the values of observed sites (e.g. kriging). The same holds true for the spatial GLMM. For autocovariate regression and spatial eigenvector mapping, in contrast, interpolation is more complicated, requiring use of the aforementioned Gibbs-sampler.

When models are projected into new geographic areas or time periods the handling of spatial autocorrelation becomes more problematic (if not impossible). Extrapolation in time, for example, is necessarily uncertain, particularly if biotic interactions – and with them spatial autocorrelation patterns – could change as each species responds differentially to climate change. However; most of the statistical methods used for prediction in time neglect important processes such as migration, dispersal, competition, predation (Pearson and Dawson 2003, Dormann 2007c), or at least assume many of them to remain constant. One might therefore argue that, while taking the autocorrelation structure as constant adds one more assumption, the use of spatial parameters at least helps to derive better models. Extrapolation in space, in contrast, is not recommended: the variance-covariance matrix parameterised in GLS approaches, for example, may look very different in other regions, even for the same organism. Hence, extrapolation can only be based on the coefficient estimates, not on the spatial component of the model. Extrapolation is further complicated by model complexity. The use of non-linear predictors and interactions between environmental variables will increase model fit, but compromises transferability of models in time and space (Beerling et al. 1995, Sykes 2001, Gavin and Hu 2006). Our study therefore did not compare methods' abilities to either make predictions to new geographic areas or extrapolate beyond the range of environmental parameters.

Bayesian approaches

Our review focused on frequentist methods. Bayesian methods, which allow prior beliefs about data to be incorporated in the calculation of expected values, offer an alternative. Experience and a good understanding of the influence of prior distributions and convergence assessment of Markov chains are crucial in Bayesian analyses. Thus, if therefore the question of interest can be addressed using more robust, less computationally intensive methods, there is no real need to apply the ''Bayesian machinery'' (Brooks 2003). The spatial analyses as presented in this paper can be done straightforwardly using non-Bayesian methods. However, Bayesian methods for the analyses of species distribution data are more flexible; they can be more easily extended to include more complex structures (Latimer et al. 2006). Models can for example be extended to a multivariate setting when several (correlated) counts of different species in each grid cell are to be modelled, or when both count and normally distributed data are to be modelled within the same framework (Thogmartin et al. 2004, Kühn et al. 2006). Bayesian methods are also a generally more suitable tool for inference in data sets with many missing values, or when accounting for detection probabilities (Gelfand et al. 2005, Kühn et al. 2006).

Wishlist

In this study, we introduced a wide range of statistical tools to deal with spatial autocorrelation in species distribution data. Unfortunately, none of these tools directly represents dynamic aspects of ecological reality (e.g. dispersal, species interaction): all the methods examined remain phenomenological rather than mechanistic. Therefore they are unable to disentangle stochastic and process-introduced spatial autocorrelation. Disentangling these sources of spatial autocorrelation in the data would be particularly important for the analysis of species that are not at equilibrium with their environmental drivers (e.g. newly introduced species expanding in range or species that have undergone population declines due to overexploitation). Moreover, it would be desirable to extend the statistical approaches used here to model multivariate response variables, such as species composition (see Kühn et al. 2006, for an example). Similarly, presence-only data, as commonly found for museum specimens, cannot be analysed with the above methods, nor are we aware of any method suitable for such data. While in principle it is possible to incorporate temporal and/or phylogenetic components into species distribution models (e.g. into GEEs, GLMMs and Bayes), this has not yet been attempted. It also would be desirable to have methods available that allow for the strengths of spatial autocorrelation to vary in space (non-stationarity), since stationarity is a basic and strong assumption of all the methods used here (except perhaps SEVM). Finally, the issue of variable selection under spatial autocorrelation has received virtually no coverage in the statistical literature, and hence the effect of spatial autocorrelation on the identification of the best-fitting model, or candidate set of most likely models, still remains unclear.

Authors' contributions and acknowledgements – Data were created by CFD, GC and FS. Analyses and manuscript sections describing each method were carried out as follows: autocovariate regression: JMM; SEVM: PRPN and RB;

GAM: JB, RO and CFD; GLS: BR and WJ; CAR: BS; SAR: WDK; GLMM: FMS and RGD; GEE: GC and IK. Further analyses, figure and table preparation and initial drafting were carried out by CFD. All authors contributed to writing the final manuscript.

We also would like to thank Pierre Legendre, Carsten Rahbek, Alexandre Diniz-Filho, Jack Lennon and Thiago Rangel for comments on an earlier version. This contribution is based on the international workshop ''Analysing Spatial Distribution Data: Principles, Applications and Software'' (GZ 4850/191/05) funded by the German Science Foundation (DFG), awarded to CFD. CFD acknowledges funding by the Helmholtz Association (VH-NG-247). JMM's work is supported by the Lenfest Ocean Program. MBA, GC, IK, RO & BR acknowledge funding by the European Union within the FP 6 Integrated Project ''ALARM'' (GOCE-CT-2003-506675). GC acknowledges a stipend from the federal state ''Sachsen-Anhalt'', Ministry of Education and Cultural Affairs. WDK & IK acknowledge support from the ''Virtual Inst. for Macroecology'', funded by the Helmholtz Association (VH-VI-153 Macroecology). RGD was supported by NERC (grant no. NER/O/S/2001/01257). PEPN research was supported by NSERC.

References

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

Anon. 2005. R: a language and environment for statistical computing. – R Foundation for Statistical Computing.

Anselin, L. 1988. Spatial econometrics: methods and models. – Kluwer.

Anselin, L. 2002. Under the hood: issues in the specification and interpretation of spatial regression models. – Agricult. Econ. 17: 247–267.

Araújo, M. B. and Williams, P. H. 2000. Selecting areas for species persistence using occurrence data. – Biol. Conserv. 96: 331–345.

Araújo, M. B. and Rahbek, C. 2006. How does climate change affect biodiversity? – Science 313: 1396–1397.

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

Augustin, N. H. et al. 1998. The role of simulation in modelling spatially correlated data. – Environmetrics 9: 175–196.

Augustin, N. H. et al. 2005. Analyzing the spread of beech canker. – For. Sci. 51: 438–448.

Bai, Z. et al. 1996. Some large-scale matrix computation problems. – J. Comput. Appl. Math. 74: 71–89.

Beerling, D. J. et al. 1995. Climate and the distribution of *Fallopia japonica*–use of an introduced species to test the predictive capacity of response surfaces. – J. Veg. Sci. 6: 269–282.

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

Besag, J. et al. 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). – Ann. Inst. Stat. Math. 43: 1–59.

Bivand, R. 2005. spdep: spatial dependence: weighting schemes, statistics and models. – R package version 0.3–17.

Bjørnstad, O. N. and Falck, W. 2000. Nonparametric spatial covariance functions: estimation and testing. – Environ. Ecol. Stat. 8: 53–70.

Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. – Ecol. Modell. 153: 51–68.

Breslow, N. E. and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. – J. Am. Stat. Assoc. 88: 9–25.

Brooks, S. P. 2003. Bayesian computation: a statistical revolution. – Phil. Trans. R. Soc. A 361: 2681–2697.

Brownstein, J. S. et al. 2003. A climate-based model predicts the spatial distribution of the Lyme disease vector *Ixodes scapularis* in the United States. – Environ. Health Persp. 111: 1152–1157.

Brunsdon, C. et al. 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. – Geogr. Analys. 28: 281–298.

Carey, V. J. 2002. gee: generalized estimation equation solver. Ported to R by Thomas Lumley (ver. 3.13, 4.4) and Brian Ripley. – <www.r-project.org>.

Carl, G. and Kühn, I. 2007a. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. – Ecol. Modell. 207: 159–170.

Carl, G. and Kühn, I. 2007b. Analyzing spatial ecological data using linear regression and wavelet analysis. – Stochast. Environ. Res. Risk Assess., in press.

Casella, G. and George, E. I. 1992. Explaining the Gibbs sampler. – Am. Stat. 46: 167–176.

Cliff, A. D. and Ord, J. K. 1981. Spatial processes: models and applications. – Pion.

Clifford, P. et al. 1989. Assessing the significance of the correlation between two spatial processes. – Biometrics 45: 123–134.

Cressie, N. A. C. 1993. Statistics for spatial data. – Wiley.

Dark, S. J. 2004. The biogeography of invasive alien plants in California: an application of GIS and spatial regression analysis. – Div. Distrib. 10: 1–9.

Davies, R. G. et al. 2006. Human impacts and the global distribution of extinction risk. – Proc. R. Soc. B 273: 2127–2133.

Diggle, P. J. et al. 1995. Analysis of longitudinal data. – Clarendon.

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

Diniz-Filho, J. A. F. et al. 1998. An eigenvector method for estimating phylogenetic inertia. – Evolution 52: 1247–1262.

Diniz-Filho, J. A. F. et al. 2003. Spatial autocorrelation and red herrings in geographical ecology. – Global Ecol. Biogeogr. 12: 53–64.

Dobson, A. J. 2002. An introduction to generalized linear models. – Chapman and Hall.

Dormann, C. F. 2007a. Assessing the validity of autologistic regression. – Ecol. Modell. 207: 234–242.
Dormann, C. F. 2007b. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. – Global Ecol. Biogeogr. 16: 129–138.
Dormann, C. F. 2007c. Promising the future? Global change predictions of species distributions. – Basic Appl. Ecol. 8: 387–397.
Dray, S. et al. 2006. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). – Ecol. Modell. 196: 483–493.
Dutilleul, P. 1993. Modifying the t test for assessing the correlation between two spatial processes. – Biometrics 49: 305–314.
Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.
Ferrier, S. et al. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. – Biodiv. Conserv. 11: 2275–2307.
Foody, G. M. 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. – Global Ecol. Biogeogr. 13: 315–320.
Fortin, M. J. and Dale, M. R. T. 2005. Spatial analysis – a guide for ecologists. – Cambridge Univ. Press.
Fotheringham, A. S. et al. 2002. Geographically weighted regression: the analysis of spatially varying relationships. – Wiley.
Gavin, D. G. and Hu, F. S. 2006. Spatial variation of climatic and non-climatic controls on species distribution: the range limit of *Tsuga heterophylla*. – J. Biogeogr. 33: 1384–1396.
Gelfand, A. E. and Vounatsou, P. 2003. Proper multivariate conditional autoregressive models for spatial data analysis. – Biostatistics 4: 11–25.
Gelfand, A. E. et al. 2005. Modelling species diversity through species level hierarchical modelling. – Appl. Stat. 54: 1–20.
Griffith, D. A. 2000a. Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. – Lin. Algebra Appl. 321: 95–112.
Griffith, D. A. 2000b. A linear regression solution to the spatial autocorrelation problem. – J. Geogr. Syst. 2: 141–156.
Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses in exploiting relative location information. – Ecology 87: 2603–2613.
Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Modell. 135: 147–186.
Guisan, A. and Thuiller, W. 2005. Predicting species distributions: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Gumpertz, M. L. et al. 1997. Autologistic model of spatial pattern of *Phytophthora* epidemic in bell pepper: effects of soil variables on disease presence. – J. Agricult. Biol. Environ. Stat. 2: 131–156.
Haining, R. 2003. Spatial data analysis – theory and practice. – Cambridge Univ. Press.
Hastie, T. J. and Tibshirani, R. J. 1990. Generalized additive models. – Chapman and Hall.
Hawkins, B. A. et al. 2007. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. – Ecography 30: 375–384.
He, F. L. et al. 2003. Autologistic regression model for the distribution of vegetation. – J. Agricult. Biol. Environ. Stat. 8: 205–222.
Hoeting, J. A. et al. 2000. An improved model for spatially correlated binary responses. – J. Agricult. Biol. Environ. Stat. 5: 102–114.
Hooten, M. B. et al. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. – Landscape Ecol. 18: 487–502.
Huffer, F. W. and Wu, H. L. 1998. Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. – Biometrics 54: 509–524.
Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological experiments. – Ecol. Monogr. 54: 187–211.
Isaaks, E. H. and Shrivastava, R. M. 1989. An introduction to applied geostatistics. – Oxford Univ. Press.
Jetz, W. and Rahbek, C. 2002. Geographic range size and determinants of avian species richness. – Science 297: 1548–1551.
Jetz, W. et al. 2005. Local and global approaches to spatial data analysis in ecology. – Global Ecol. Biogeogr. 17: 97–98.
Kaboli, M. et al. 2006. Avifaunal gradients in two arid zones of central Iran in relation to vegetation, climate, and topography. – J. Biogeogr. 33: 133–144.
Kaluzny, S. P. et al. 1998. S-plus spatial stats user's manual for Windows and Unix. – Springer.
Keitt, T. H. et al. 2002. Accounting for spatial pattern when modeling organism-environment interactions. – Ecography 25: 616–625.
Kissling, W. D. and Carl, G. 2007. Spatial autocorrelation and the selection of simultaneous autoregressive models. – Global Ecol. Biogeogr., in press.
Klute, D. S. et al. 2002. Autologistic regression modeling of American woodcock habitat use with spatially dependent data. – In: Scott, J. M. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, pp. 335–343.
Knapp, R. A. et al. 2003. Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. – Ecol. Appl. 13: 1069–1082.
Kupfer, J. A. and Farris, C. A. 2007. Incorporating spatial non-stationarity of regression coefficients into predictive vegetation models. – Landscape Ecol. 22: 837–852.
Kühn, I. 2007. Incorporating spatial autocorrelation may invert observed patterns. – Div. Distrib. 13: 66–69.
Kühn, I. et al. 2006. Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. – New Phytol. 72: 127–139.
Latimer, A. M. et al. 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.
Legendre, P. and Fortin, M.-J. 1989. Spatial pattern and ecological analysis. – Vegetatio 80: 107–138.
Legendre, P. and Legendre, L. 1998. Numerical ecology. – Elsevier.
Legendre, P. et al. 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.
Liang, K. Y. and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. – Biometrika 73: 13–22.
Lichstein, J. W. et al. 2002. Spatial autocorrelation and autoregressive models in ecology. – Ecol. Monogr. 72: 445–463.
Liebhold, A. M. and Gurevitch, J. 2002. Integrating the statistical analysis of spatial data in ecology. – Ecography 25: 553–557.
Link, W. A. and Barker, R. J. 2006. Model weights and the foundations of multimodel inference. – Ecology 87: 2626–2635.
Littell, R. C. et al. 1996. SAS system for mixed models. – SAS Publ.
Luoto, M. et al. 2001. Determinants of distribution and abundance in the clouded apollo butterfly: a landscape ecological approach. – Ecography 24: 601–617.
McCullough, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. – Ecography 30: 135–151.
Miller, J. et al. 2007. Incorporating spatial dependence in predictive vegetation models. – Ecol. Modell. 202: 225–242.
Moore, J. E. and Swihart, R. K. 2005. Modeling patch occupancy by forest rodents: incorporating detectability and spatial autocorrelation with hierarchically structured data. – J. Wildl. Manage. 69: 933–949.
Myers, R. H. et al. 2002. Generalized linear models. – Wiley.
Orme, C. D. L. et al. 2005. Global hotspots of species richness are not congruent with endemism or threat. – Nature 436: 1016–1019.
Osborne, P. E. et al. 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. – J. Appl. Ecol. 38: 458–471.
Osborne, P. E. et al. 2007. Non-stationarity and local approaches to modelling the distributions of wildlife. – Div. Distribut., in press.
Palma, L. et al. 1999. The use of sighting data to analyse Iberian lynx habitat and distribution. – J. Appl. Ecol. 36: 812–824.
Pearson, R. G. and Dawson, T. P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? – Global Ecol. Biogeogr. 12: 361–371.
Perry, J. N. et al. 2002. Illustrations and guidelines for selecting statistical methods for quantifying spatial paterrns in ecological data. – Ecography 25: 578–600.
Pinheiro, J. C. and Bates, D. M. 2000. Mixed-effect models in S and S-plus. – Springer.
Rangel, T. F. L. V. B. et al. 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. – Global Ecol. Biogeogr. 15: 321–327.
Reich, R. M. et al. 2004. Predicting the location of northern goshawk nests: modeling the spatial dependency between nest locations and forest structure. – Ecol. Modell. 176: 109–133.
Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.
Segurado, P. et al. 2006. Consequences of spatial autocorrelation for niche-based models. – J. Appl. Ecol. 43: 433–444.
Smith, P. A. 1994. Autocorrelation in logistic regression modelling of species' distributions. – Global. Ecol. Biogeogr. Lett. 4: 47–61.
Sokal, R. R. and Oden, N. L. 1978a. Spatial autocorrelation in biology. I. Methodology. – Biol. J. Linn. Soc. 10: 199–228.
Sokal, R. R. and Oden, N. L. 1978b. Spatial autocorrelation in biology. II. Some biological implications and four applications of evolutionary and ecological interest. – Biol. J. Linn. Soc. 10: 229–249.
Stephenson, C. M. et al. 2006. Modelling establishment probabilities of an exotic plant, *Rhododendron ponticum*, invading a heterogeneous, woodland landscape using logistic regression with spatial autocorrelation. – Ecol. Modell 193: 747–758.
Sykes, M. T. 2001. Modelling the potential distribution and community dynamics of lodgepole pine (*Pinus contorta* Dougl. ex. Loud.) in Scandinavia. – For. Ecol. Manage. 141: 69–84.
Teterukovskiy, A. and 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.
Thogmartin, W. E. et al. 2004. A hierarchical spatial model of avian abundance with application to Cerulean warblers. – Ecol. Appl. 14: 1766–1779.
Tiefelsdorf, M. et al. 1999. A variance-stabilizing coding scheme for spatial link matrices. – Environ. Plann. A 31: 165–180.
Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. – Econ. Geogr. 46: 234–240.
Tognelli, M. F. and Kelt, D. A. 2004. Analysis of determinants of mammalian species richness in South America using spatial autoregressive models. – Ecography 27: 427–436.
Venables, W. N. and Ripley, B. D. 2002. Modern applied statistics with S. – Springer.
Ver Hoef, J. M. et al. 1993. Spatial models for spatial statistics: some unification. – J. Veg. Sci. 4: 441–452.
Wall, M. M. 2004. A close look at the spatial structure implied by the CAR and SAR models. – J. Stat. Plann. Infer. 121: 311–324.
Waller, L. A. and Gotway, C. A. 2004. Applied spatial statistics for public health data. – Wiley.
Wood, S. N. 2006. Generalized additive models. – Chapman and Hall/CRC.

Worm, B. et al. 2005. Global patterns of predator diversity in the open oceans. – Science 309: 1365–1369.
Wu, H. L. and Huffer, F. W. 1997. Modelling the distribution of plant species using the autologistic regression model. – Environ. Ecol. Stat. 4: 49–64.
Yamaguchi, N. et al. 2003. Habitat preferences of feral American mink in the Upper Thames. – J. Mammal. 84: 1356–1373.

Download the appendix as file E5171 from <www.oikos.ekol.lu.se/appendix>.

Yan, J. 2002. geepack: yet another package for generalized estimating equations. – R News 2: 12–14.
Yan, J. 2004. geepack: generalized estimating equation package. – R package version 0.2–10.

Ecography

E5171

Dorman, F. C., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M. and Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 000–000.

Introduction

The publication and this supplement serve the purpose of making specific statistical methods available to ecologists. These users are usually not statisticians, and we attempt to relate sometimes rather sophisticated methodologies to the desperate analyst. However, analysing species distribution data is a tricky thing, with many potential pitfalls along the way. Neither do we attempt to address all open questions, nor will we be able to produce a cookbook recipe for all types of analyses. What we do attempt is a) a decision tree about which spatial autocorrelation modelling method to use when, and b) software implementation aids for these methods. We opted for using the software package R (www.r-project.org), which is extremely flexible, versatile - and free.

The following pages cannot be understood without some advanced statistical knowledge or without the paper this code accompanies. We assume a basic understanding of generalised linear models for non-normally distributed data. Most details on the methods are provided in the main paper, while these pages are primarily for implementing the methods. The person mainly responsible for the implementation of that method is given in the parentheses in the section title.

Decision tree

The first partition in our decision tree is dictated by the type of response variable to be analysed. More methods are available for data derived from a normal distribution (data whose residuals are normally distributed) than data of alternative distributional form (e.g. binomial, Poisson). Typical examples of ecological data with normally distributed errors include abundance, species richness, or functional diversity per unit area, crop yield and catch per unit effort. Examples of data following a binomial distribution are the presence or absence of a species and the success or failure of seed germination. In contrast, whale sightings, counts of herd size or the number of ectoparasites per sheep are data more likely to follow a Poisson (or negative binomial) distribution.

The second partition refers to computational efficiency. Some methods are so computer-intensive that they may not be suitable for use with very large data sets.

method	residuals	computational intensity
GAM	normal, Poisson, binomial	low
autogressive models (SAR/CAR)	normal	medium-high
GLS	normal	medium-high
GEE	normal, Poisson, binomial	low
autocovariate regression	normal, Poisson, binomial	low
spatial GLMM	normal, Poisson, binomial	very high
Spatial Eigenvector Mapping	normal, Poisson, binomial	very high

1.1. Preparing the data and non-spatial analysis (Carsten F. Dormann)

All following analyses are illustrated using data organized as an XYZ-table (or, in the R nomenclature), data.frame. X and Y are the spatial coordinates, while Z is the response variable. Additionally, we have explanatory variables, here "rain" and "djungle". The data we used for the analysis are available as supplementary material ("snouterdata.txt").

To get into the right mood, we start with *non-spatial* models, both for normally and non-normally distributed residuals:

Reading the data

```
#set your path here!
setwd <- "C:/Data/..."
#read in data:
snouter.df <- read.table("snouterdata.txt", header=T, sep="\t")
```

Run non-spatial models

GLMs

```
summary(ols1 <- lm(snouter1.1 ~ rain + djungle, data=snouter.df))
summary(binom1 <- glm(snouter2.1 ~rain+djungle, family=binomial,
data=snouter.df))
summary(pois1 <- glm(snouter3.1 ~rain+djungle, family=poisson,
data=snouter.df))
```

GAMs (Janine Bolliger & Ralf Ohlemüller)

There are two commonly used GAM-packages in R: **mgcv** (by Simon Woods) and **gam** (by Trevor Hastie). They are different, since there is no single definition of what a GAM is, but they are also similar (see the comparison in Faraway's book: Extending the Linear Model). Here, we use **mgcv**. We use the spline only for the geographical component, treating rain and djungle as fixed effects. This practically amounts to a trend-surface regression with rain and djungle as covariates.

```
library(mgcv)
summary(gam.norm <- gam(snouter1.1 ~ rain + djungle + s(X, Y),
data=snouter.df, family=gaussian))
summary(gam.bino <- gam(snouter2.1 ~ rain + djungle + s(X, Y),
data=snouter.df, family=binomial))
summary(gam.pois <- gam(snouter3.1 ~ rain + djungle + s(X, Y),
data=snouter.df, family=poisson))
```

The degree of spatial autocorrelation in the residuals can be assessed using correlograms. They depict Moran's across all distance classes.

Plotting/calculating spatial autocorrelation

```
# First, install the library ncf (http://onb.ent.psu.edu/onb1/);
# for windows:
install.packages("ncf",contriburl="http://asi23.ent.psu.edu/onb1/R/windows"
)
# for linux:
#install.packages("ncf",contriburl="http://asi23.ent.psu.edu/onb1/R/src")
require(ncf)
?correlog
model <- ols1 # or binom1 or pois1
correlog1.1 <- correlog(snouter.df$X, snouter.df$Y, residuals(model),
na.rm=T, increment=1, resamp=0)

# now plot only the first 20 distance classes:
par(mar=c(5,5,0.1, 0.1))
plot(correlog1.1$correlation[1:20], type="b", pch=16, cex=1.5, lwd=1.5,
xlab="distance", ylab="Moran's I", cex.lab=2, cex.axis=1.5); abline(h=0)

# make a map of the residuals:
plot(snouter.df$X, snouter.df$Y, col=c("blue",
"red")[sign(resid(model))/2+1.5], pch=19,
cex=abs(resid(model))/max(resid(model))*2, xlab="geographical x-
coordinates", ylab="geographical y-coordinates")

# calculate Moran's I values explicitly for a certain distance,
# and to test for its significance:
require(spdep)
snouter.nb <- dnearneigh(as.matrix(snouter.df[1:2]), 0, 20) #give lower and
upper distance class here!
snouter.listw <- nb2listw(snouter.nb) #turns neighbourhood object into a
weighted list
#this next step takes often several minutes to run:
GlobMT1.1<- moran.test(residuals(model), listw=snouter.listw)
```

Now we are set to start with the spatial analysis.

1.2. Methods for normally distributed residuals

The best established methods here are autoregressive models and Generalized Linear Models. Spatial filtering is a very new method. For details on GEE and autocovariate regression models see next section.

Autoregressive Models in R

AR are implemented in the library **spdep**. Several functions can be invoked for the regression itself, depending on which assumptions are made about the cause of spatial autocorrelation (errorsarlm, lagsarlm, spautolm). A comparison of these different autoregressive models is very advisable, either using model selection procedures (e.g. Kissling & Carl 2007) or the Lagrange multiplier test (see SAR below).

Simultaneous Autoregressive Models (SAR) (W. Daniel Kissling)

1. spatial SAR error model ("SARerr"): function `errorsarlm()`
2. spatial SAR lag model ("SARlag"): function `lagsarlm()` with `type="lag"`
3. spatial SAR mixed model ("SARmix"): function `lagsarlm()` with `type="mixed"`

All SAR models require normally distributed errors. To use the SAR functions, one first needs to specify a spatial weights matrix which is to be incorporated in the SAR functions as a "listw" object. This spatial weights matrix is constructed by first defining the neighbourhood with the dnearneigh() function and then weighting the neighbours with the nb2listw() function by choosing a certain coding scheme (e.g., "B" = binary coding, "W" = row standardised, or "S" = variance-stabilising coding scheme).
Here we only use a neighbourhood distance of 1.5 and a coding scheme "W", but other spatial weights matrices can be defined and implemented. The Lagrange Multiplier diagnostics function "lm.LMtests()") is used to compare the different SAR model types. Rather than entering the regression formula directly, we here call a previous non-spatial model (ols1) for this. A model selection procedure can also be used (Kissling & Carl 2007).

```
require(spdep)

# Data preparation
snouter.df <- read.table("snouterdata.txt", header=T, sep="\t")

# Define coordinates, neighbours, and spatial weights
coords<-as.matrix(cbind(snouter.df$X,snouter.df$Y))

#Define neighbours up to a certain distance
nb1.5<-dnearneigh(coords,0,1.5) #you can use e.g. distance 1 or 2 instead

#Spatial weights
nb1.5.w<-nb2listw(nb1.5, glist=NULL, style="W", zero.policy=FALSE)

#1. Spatial SAR error model
sem.nb1.5.w <- errorsarlm(ols1, listw=nb1.5.w)
summary(sem.nb1.5.w) #Gives a summary of the SAR error model

# 2. Spatial SAR lag model
slm.nb1.5.w <- lagsarlm(ols1, listw=nb1.5.w, type="lag")
summary(slm.nb1.5.w) #Gives a summary of the SAR lag model

# 3. Spatial SAR mixed model
smm.nb1.5.w <- lagsarlm(ols1, listw=nb1.5.w, type="mixed")
summary(smm.nb1.5.w) #Gives a summary of the SAR mixed model


# Lagrange multiplier diagnostics for model comparison
#To test which model is most appropriate:

lm.LMtests(lm1, nb1.5.w, test="all")
```

Conditional Autoregressive Models (CAR) (Boris Schröder)

```
require(spdep)

#Make a matrix of coordinates
coords<-as.matrix(cbind(data$X,data$Y))
```

```
#Define neighbourhood
nb<-dnearneigh(coords, 0, 2) # 2nd order neighbourhood

# define spatial weights matrix
w<-nb2listw(nb, style="W", zero.policy=FALSE)

# specify and run CAR model

cem<-spautolm(snouter1.1 ~ rain + djungle, data=snouter.df, listw=w,
family="CAR") # CAR error model
summary(cem)
```

Generalized Least Square Models in R (Björn Reineking)

GLS are fitted using the function gls {**nlme**} or gls.fit {**MASS**}. Internally, also SAR and CAR methods call one of them. gls {**nlme**} offers to specify the expected form of autocorrelation in the correlation argument. With the correlation option in the gls-call, the different spatial correlation structures can be specified.

```
require(nlme)
?gls #for syntax help
?corClasses #for help with correlation functions available

summary(gls.exp <- gls(snouter1.1 ~ rain + djungle, data=snouter.df,
correlation=corExp(form=~X+Y)))
summary(gls.gauss <- gls(snouter1.1 ~ rain + djungle, data=snouter.df,
correlation=corGaus(form=~X+Y)))
summary(gls.spher <- gls(snouter1.1 ~ rain + djungle, data=snouter.df,
correlation=corSpher(form=~X+Y)))
```

1.3. Methods also for non-normally distributed residuals (e.g. Poisson or binomial)

Autocovariate regression in R (Jana McPherson)

Autocovariate regression was conceived for binary data (as "autologistic regression") by Augustin et al. (1996). Here we offer the identical methodology to be extended to any type of data, since the logic behind the approach is independent of the error distribution. Note, however, that to date only autologistic regressions have found their way into literature, and we cannot warrant for the statistical correctness of this approach. In R, a function autocov_dist {**spdep**} provides the means to create a new explanatory variable which is entered in addition to the other explanatory variables in the model.

```
require(spdep)
?autocov_dist

# prepare neighbour lists for spatial autocorrelation analysis
nb.list <- dnearneigh(as.matrix(snouter.df[,c("X", "Y")]), 0, 5)
nb.weights <- nb2listw(nb.list)

#Make a matrix of coordinates
coords<-as.matrix(cbind(data$X,data$Y))

# compute the autocovariate based on the above distance and weight
ac <- autocov_dist(snouter1.1,  coords,  nbs = 1,  type = "inverse")
```

```
# now run a linear model with the autocovariate as additional explanatory
variable:
fm <- lm(snouter1.1 ~ rain + djungle + ac,  data=snouter.df)
summary(fm)
```

Generalized Estimation Equations in R (Gudrun Carl)

Two different GEE packages are available in R: gee {**gee**} and geese {**geepack**}. Data preparation for the analysis is a significant part of the exercise, while the models themselves work fast and efficient. The following code and helper functions shall aid with the data preparation.

Copy the following code (starting and ending with #%%%%%%%%%%) into a text-file and save as "geefunctions.R". It is also provided as supplementary material, which you can include into an R-session writing: `source("geefunctions.R")`.

```
#%%%%%%%%%%
# Helper functions for the spatial usage of gee and geepack
# Code written by Gudrun Carl, 2005

#################################################################

#########################################################################
dat.nn<-function(data,n){
#########################################################################
# A function to generate clusters and order variables and
# to produce a data frame with response, predictors, coordinates, and
# 3 new parameters:
# o for order
# id for identifying clusters and
# waves for identifying members of clusters
#
# Arguments
# data      a data frame with response and predictors and
#           in the last columns with ordered cartesian coordinates
# n         for maximal cluster size  n*n
#########################################################################

l<-dim(data)[2]
OST<-data[,l-1]
NORD<-data[,l]
ko<-OST-min(OST)
idx<-(ko-(ko%%(n)))/n+1
ks<-NORD-min(NORD)
idy<-(ks-(ks%%(n)))/n+1
ie<-(idy-1)*max(idx)+idx
idwx<-ko%%(n)+1
idwy<-ks%%(n)+1
wav<-(idwy-1)*n+idwx
data<-as.matrix(data)
o<-order(ie,wav)
id<-ie[o]
waves<-wav[o]
dat.new1<-data[o,]
dat.new2<-cbind(dat.new1,o,id,waves)
dat.new<-as.data.frame(dat.new2)
}
```

```
#########################################################################
a.gee<-function(mgee,n,type="glm",corstr="independence",quad=T) {
#########################################################################
# A function to order correlation parameters of Generalized Estimating
# Equation Models
# Arguments
# mgee       matrix or vector of correlation parameters according to model
# n          for maximal cluster size n*n
# type       type of model
#            "glm", "gee", "geese" are allowed
# corstr     correlation structure
#            "independence", "exchangeable", "userdefined" are allowed
# quad       by default quadratic correlation structure
#            for model "geese" and "userdefined" correlation only
#########################################################################

if(n==2)n3<-6
if(n==3)n3<-36
if(n==4)n3<-120
a<-rep(0,n3)
if(type=="glm") a<-a
if(type=="gee"){
  if(corstr=="exchangeable") a[c(1:n3)]<-mgee[1,2]
  if(corstr=="independence") a<-a
}
a<-as.vector(a)

if(type=="geese") {
 if(corstr=="userdefined"){
if(quad) {
if(n==2)  {
 a<-rep(0,6)
 a[c(1,2,5,6)]<-mgee[1]
 a[c(3,4)]<-mgee[2]
}
if(n==3)  {
 a<-rep(0,36)
 a[c(1,3,9,11,18,22,24,27,29,33,34,36)]<-mgee[1]
 a[c(2,6,14,21,23,35)]<-mgee[2]
 a[c(4,10,12,17,25,28,30,32)]<-mgee[3]
 a[c(5,7,13,15,16,20,26,31)]<-mgee[4]
 a[c(8,19)]<-mgee[5]
}
if(n==4)  {
 a<-rep(0,120)
 a[c(1,4,16,19,30,33,46,55,58,66,69,76,79,88,93,96,100,103,106,109,
114,115,118,120)]<-mgee[1]
 a[c(2,8,17,23,37,50,56,62,67,73,83,92,94,101,116,119)]<-mgee[2]
 a[c(3,12,27,41,54,57,95,117)]<-mgee[3]
 a[c(5,18,20,32,34,45,59,68,70,78,80,87,97,102,104,108,110,113)]<-mgee[4]
 a[c(6,9,21,22,24,31,36,38,44,49,60,63,71,72,74,77,82,84,86,91,98,
105,107,112)]<-mgee[5]
 a[c(7,13,26,28,40,42,43,53,61,85,99,111)]<-mgee[6]
 a[c(10,25,35,48,64,75,81,90)]<-mgee[7]
 a[c(11,14,29,39,47,52,65,89)]<-mgee[8]
 a[c(15,51)]<-mgee[9]
}}
if(!quad) a<-mgee
}
if(corstr=="exchangeable") a[c(1:n3)]<-mgee
if(corstr=="independence") a<-a
```

```
 }
a<-as.vector(a)
}

##############################################################################
clus.sz<-function(id){
##############################################################################
# A function to calculate sizes of clusters
# Argument
# id     vector which identifies the clusters
##############################################################################

clus<-rep(0,length(id))
k0<-0
k1<-1
for(i in 2:length(id)) { i1<-i-1
if(id[i]==id[i1]) {k1<-k1+1
if(i==length(id)) {k0<-k0+1
                   clus[k0]<-k1}}
if(id[i]!=id[i1]) {k0<-k0+1
                   clus[k0]<-k1
                   k1<-1
if(i==length(id)) {k0<-k0+1
                   clus[k0]<-k1 }}}
clusz<-clus[clus>0]
}

##############################################################################
zcor.quad<-function(zcor,n,quad=TRUE) {
##############################################################################
# A function to create a quadratic correlation structure
# zcor    an object of class "genZcor" (see: geepack)
# n       for maximal cluster size n*n
# quad    by default quadratic correlation structure
##############################################################################

if(quad) {
if(n==2)  {
 zcorn<-matrix(0,dim(zcor)[1],2)
 zcorn[,1]<-zcor[,1]+zcor[,2]+zcor[,5]+zcor[,6]
 zcorn[,2]<-zcor[,3]+zcor[,4]
}
if(n==3)  {
 zcorn<-matrix(0,dim(zcor)[1],5)
 zcorn[,1]<-zcor[,1]+zcor[,3]+zcor[,9]+zcor[,11]+zcor[,18]+zcor[,22]+
 zcor[,24]+zcor[,27]+zcor[,29]+zcor[,33]+zcor[,34]+zcor[,36]
 zcorn[,2]<-zcor[,2]+zcor[,6]+zcor[,14]+zcor[,21]+zcor[,23]+zcor[,35]
 zcorn[,3]<-zcor[,4]+zcor[,10]+zcor[,12]+zcor[,17]+zcor[,25]+zcor[,28]+
 zcor[,30]+zcor[,32]
 zcorn[,4]<-zcor[,5]+zcor[,7]+zcor[,13]+zcor[,15]+zcor[,16]+zcor[,20]+
 zcor[,26]+zcor[,31]
 zcorn[,5]<-zcor[,8]+zcor[,19]
}
if(n==4)  {
 zcorn<-matrix(0,dim(zcor)[1],9)
  zcorn[,1]<-zcor[,1]+zcor[,4]+zcor[,16]+zcor[,19]+zcor[,30]+zcor[,33]+
zcor[,46]+zcor[,55]+zcor[,58]+zcor[,66]+zcor[,69]+zcor[,76]+
zcor[,79]+zcor[,88]+zcor[,93]+zcor[,96]+zcor[,100]+zcor[,103]+
zcor[,106]+zcor[,109]+zcor[,114]+zcor[,115]+zcor[,118]+zcor[,120]
  zcorn[,2]<-zcor[,2]+zcor[,8]+zcor[,17]+zcor[,23]+zcor[,37]+zcor[,50]+
zcor[,56]+zcor[,62]+zcor[,67]+zcor[,73]+zcor[,83]+zcor[,92]+
```

```
zcor[,94]+zcor[,101]+zcor[,116]+zcor[,119]
  zcorn[,3]<-zcor[,3]+zcor[,12]+zcor[,27]+zcor[,41]+zcor[,54]+zcor[,57]+
zcor[,95]+zcor[,117]
  zcorn[,4]<-zcor[,5]+zcor[,18]+zcor[,20]+zcor[,32]+zcor[,34]+zcor[,45]+
zcor[,59]+zcor[,68]+zcor[,70]+zcor[,78]+zcor[,80]+zcor[,87]+
zcor[,97]+zcor[,102]+zcor[,104]+zcor[,108]+zcor[,110]+zcor[,113]
  zcorn[,5]<-zcor[,6]+zcor[,9]+zcor[,21]+zcor[,22]+zcor[,24]+zcor[,31]+
zcor[,36]+zcor[,38]+zcor[,44]+zcor[,49]+zcor[,60]+zcor[,63]+
zcor[,71]+zcor[,72]+zcor[,74]+zcor[,77]+zcor[,82]+zcor[,84]+
zcor[,86]+zcor[,91]+zcor[,98]+zcor[,105]+zcor[,107]+zcor[,112]
  zcorn[,6]<-zcor[,7]+zcor[,13]+zcor[,26]+zcor[,28]+zcor[,40]+zcor[,42]+
zcor[,43]+zcor[,53]+zcor[,61]+zcor[,85]+zcor[,99]+zcor[,111]
  zcorn[,7]<-zcor[,10]+zcor[,25]+zcor[,35]+zcor[,48]+zcor[,64]+zcor[,75]+
zcor[,81]+zcor[,90]
  zcorn[,8]<-zcor[,11]+zcor[,14]+zcor[,29]+zcor[,39]+zcor[,47]+zcor[,52]+
zcor[,65]+zcor[,89]
  zcorn[,9]<-zcor[,15]+zcor[,51]
}
}
if(!quad) zcorn<-zcor
zcorn<-as.matrix(zcorn)
}

#############################################################################
res.gee<-function(formula,data,n,clusz=NA,zcor=NA,a=NA,b,R=NA,
fam="b",type="resid")  {
#############################################################################
# A function to calculate fitted values and residuals
# for Generalized Estimating Equation Models
# for gaussian or binary data (with logit link) or Poisson data (log link)
# Arguments
# formula     a formula expression
# data        a data frame
# n           for maximal cluster size n*n
# clusz       an object of class "clus.sz"
# zcor        an object of class "genZcor"
# a           a vector of correlation parameters
#             for clusters only
#             as an object of class "a.gee"
# b           a vector of regression parameters beta
# R           a square matrix of correlation parameters
#             for full dimension (=number of observations)  only
# fam         family
#             "g", "b", "p" (gaussian, binary, Poisson) are allowed
# type        "fitted" calculates fitted values
#             "resid" calculates residuals
#
#############################################################################

l<-dim(data)[2]
ieo<-data[,l-1]
if(n!=dim(data)[1]) {
n2<-n*n
n3<-n2*(n2-1)/2
n4<-n2-1
n5<-n2-2
for(i in 1:dim(zcor)[1]){
for(k in 1:n3){
if(zcor[i,k]==1) zcor[i,k]<-a[k]  }}
lc<-length(clusz)
z2<-matrix(0,lc,n3)
```

```
for( j in 1:n3) {
k3<-0
k2<-0
for(i in 1:lc) {
if(clusz[i]!=1) {
k2<-k3+1
k3<-clusz[i]*(clusz[i]-1)/2+k3
for(k in k2:k3)
z2[i,j]<-zcor[k,j]+z2[i,j] }}}
if(n==2)
iod<-c(1,1,2)
if(n==3)
iod<-c(1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,4,4,4,4,5,5,5,6,6,7)
if(n==4)
iod<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,
3,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,
6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,8,8,8,8,8,8,8,9,9,9,9,9,9,10,10,10,10,10,
11,11,11,11,12,12,12,13,13,14)
cs<-0
v<-matrix(0,length(ieo),length(ieo))
vgl<-rep(0,n2)
for(i in 1:lc) {clu<-clusz[i]
if(clu!=1) {
v1<-matrix(0,n2,n2)
if(n==2)
{  v1[1,2:4]<-z2[i,1:3]
   v1[2,3:4]<-z2[i,4:5]
   v1[3,4]<-z2[i,6]  }
if(n==3)
{  v1[1,2:9]<-z2[i,1:8]
   v1[2,3:9]<-z2[i,9:15]
   v1[3,4:9]<-z2[i,16:21]
   v1[4,5:9]<-z2[i,22:26]
   v1[5,6:9]<-z2[i,27:30]
   v1[6,7:9]<-z2[i,31:33]
   v1[7,8:9]<-z2[i,34:35]
   v1[8,9]<-z2[i,36]  }
if(n==4)
{  v1[1,2:16]<-z2[i,1:15]
   v1[2,3:16]<-z2[i,16:29]
   v1[3,4:16]<-z2[i,30:42]
   v1[4,5:16]<-z2[i,43:54]
   v1[5,6:16]<-z2[i,55:65]
   v1[6,7:16]<-z2[i,66:75]
   v1[7,8:16]<-z2[i,76:84]
   v1[8,9:16]<-z2[i,85:92]
   v1[9,10:16]<-z2[i,93:99]
   v1[10,11:16]<-z2[i,100:105]
   v1[11,12:16]<-z2[i,106:110]
   v1[12,13:16]<-z2[i,111:114]
   v1[13,14:16]<-z2[i,115:117]
   v1[14,15:16]<-z2[i,118:119]
   v1[15,16]<-z2[i,120]   }
for(i1 in 1:length(iod)) {
i2<-iod[i1]
if(var(v1[i2,1:n2])==0) {for(k in i2:n5) {k1<-k+1
                                          v1[k,]<-v1[k1,]
                                          v1[k1,]<-vgl[]}}}
for(i1 in 1:length(iod)){
i3<-iod[i1]+1
if(var(v1[1:n2,i3])==0) {for(k in i3:n4) {k1<-k+1
```

```
                                        v1[,k]<-v1[,k1]
                                        v1[,k1]<-vgl[]}}}

clu1<-clu-1
for(k in 1:clu1) {csk<-cs+k
f1<-2
for(k1 in f1:clu) {k2<-cs+f1
v[csk,k2]<-v1[k,k1]
f1<-f1+1 }}
for(k in 1:clu) {csk<-cs+k
v[csk,csk]<- 0.5 } }
if(clu==1) {cs1<-cs+1
v[cs1,cs1]<-0.5 }
cs<- cumsum(clusz)[i]  }
v<-v+t(v)
}
if(n==dim(data)[1]) v<-R
ww<-solve(v)

s.geese<-svd(ww,LINPACK=T)
d.geese<-diag(sqrt(s.geese$d))
w<-s.geese$u%*%d.geese%*%t(s.geese$u)

x.matrix<-model.matrix(formula)
fitted<-x.matrix%*%b
fitted<-fitted[1:length(ieo)]
if(fam=="p") fitted<-exp(fitted)
if(fam=="b") fitted<-exp(fitted)/(1+exp(fitted))

if(type=="fitted")resgeeseo<-fitted
if(type=="resid"){
if(fam=="g") rgeese<-data[,1]-fitted
if(fam=="p") rgeese<-(data[,1]-fitted)/sqrt(fitted)
if(fam=="b") rgeese<-(data[,1]-fitted)/sqrt(fitted*(1-fitted))
rsgeese<-w%*%rgeese
resgeeseo<-rsgeese[1:length(ieo)]
}

resgeeseo<-as.vector(resgeeseo)
}


############################################################

#%%%%%%%%%%


# here the real analysis with GEE starts:
# define dependent variable "snouter" (response)
response <- "snouter1.1"

# load necessary libraries and codes
require(gee)
require(geepack)
source("geefunctions.R") # attached as text file


# GEE ------------------------------------------------
# gee model with fixed correlation structure

# first prepare a dataframe
```

```
data <-
data.frame(snouter.df[,c("snouter1.1",names(snouter.df[c(3:4,1:2)]))])
attach(data)
nn <- nrow(data)
coord <- cbind(X, Y)

# next compute glm model:
mglm <- glm(snouter1.1 ~ rain + djungle, family=gaussian, data=data)
resglm <- resid(mglm)
corglm <- correlog(X, Y, resglm, na.rm=T, increment=1,
resamp=1,legacy=FALSE)

# fit of autocorrelation for gee fixed model
alpha <- corglm$correlation[1]
idn <- rep(1,nn)
D <- as.matrix(dist(coord))
R <- alpha^D

# then gee model
mgee <- gee(snouter1.1 ~ rain + djungle, family=gaussian, data=data,
id=idn, R=R, corstr="fixed")
summary(mgee)[1:7]

# residuals and fitted values extracted by homemade functions
resgee <-res.gee(snouter1.1 ~ rain + djungle, data=data, nn, b=mgee$coeff,
R=R, fam="g", type="resid")
fitgee <-res.gee(snouter1.1 ~ rain + djungle, data=data, nn, b=mgee$coeff,
R=R, fam="g", type="fitted")
detach(data)

# GEESE ------------------------------------------------
# geese model with user defined correlation structure
# cluster size, nr=3 --> 3*3 cells in a cluster
nr <- 3

# first prepare a new dataframe
data.cluster <- dat.nn(snouter.df[,c("snouter1.1",
names(snouter.df[c(3:4,1:2)]))], n=nr)
attach(data.cluster)

# prepare cluster sizes, waves within clusters and quadratic correlation
structure
# by functions from geepack and homemade functions
clusz <- clus.sz(id)
zcor <- genZcor(clusz=clusz, waves=waves, corstrv="unstructured")
zcorq <- zcor.quad(zcor, n=nr, quad=T)

# then geese model
mgeese <- geese(snouter1.1 ~ rain + djungle, family=gaussian,
data=data.cluster, id=id, corstr="userdefined", zcor=zcorq)
summary(mgeese)

# residuals and fitted values extracted by homemade functions
ageese <- a.gee(mgeese$a,n=nr,type="geese",corstr="userdefined",quad=T)
resgeese.cluster <- res.gee(snouter1.1 ~ rain + djungle,
data=data.cluster,n=nr,clusz,zcor,ageese,mgeese$b,fam="g",type="resid")
fitgeese.cluster <- res.gee(snouter1.1 ~ rain + djungle,
data=data.cluster,n=nr,clusz,zcor,ageese,mgeese$b,fam="g",type="fitted")
resgeese <- resgeese.cluster[order(o)]
fitgeese <- fitgeese.cluster[order(o)]
detach(data.cluster)
```

Spatial Generalized Linear Mixed Model in R (Frank Schurr)

This is an inofficial abuse of a Generalized Linear Mixed Model function (glmmPQL {**MASS**}), which is a wrapper function for lme {**nlme**}, which in turn internally calls gls {**nlme**}. Hence the presented method is actually a generalization of gls to non-normal data. It requires a "cheat" to make it do what we want, however. This code produces the identical results as an official spatial GLMM in SAS (proc glimmix) and can hence be trusted.

```
require(MASS)
?glmmPQL
?corClasses

snouter.df <- read.table("snouterdata.txt", head=T)

#define a grouping factor that assigns all observations to the same group
group <- factor(rep("a",nrow(snouter.df)))
snouter.df <- cbind(snouter.df, group)


attach(snouter.df) #For some reason, the data have to be attached AND
specified in the formula!
# GLMM fits ----------------------
#exponential correlation structure

model.e <- glmmPQL(snouter2.1 ~ rain + djungle, random=~1|group,
data=snouter.df,
correlation=corExp(form=~X+Y), family=binomial))

#Gaussian correlation structure
model.g <- glmmPQL(snouter2.1 ~ rain + djungle, random=~1|group,
data=snouter.df, correlation=corGaus(form=~X+Y), family=binomial))

#spherical correlation structure
model.s <- glmmPQL(snouter2.1 ~ rain + djungle, random=~1|group,
data=snouter.df, correlation=corSpher(form=~X+Y), family=binomial))

detach(snouter.df)
```

Spatial Eigenvector Mapping in R (Pedro Peres-Neto)

This method has only recently seen the light of day. Two steps are involved in R: (1) setting up the correct neighbourhood, (2) running the acutal eigenvector generator and selector. You may also want to check ?SpatialFiltering which "only" corrects standard errors of an OLS but does not affect parameter estimates.

```
require(spdep)
?ME

snouter_sp <- SpatialPixelsDataFrame(as.matrix(snouter.df[,2:1]),
snouter.df)
nb1.0 <- dnearneigh(coordinates(snouter_sp), 0, 1.0)
nb1.0_dists <- nbdists(nb1.0, coordinates(snouter_sp))
nb1.0_sims <- lapply(nb1.0_dists, function(x) (1-((x/4)^2)) )
ME.listw <- nb2listw(nb1.0, glist=nb1.0_sims, style="B")
```

```
sevm1 <- ME(snouter1.1 ~ rain + djungle, data=snouter.df, family=gaussian,
listw=ME.listw)

# modify the arguments "family" according to your error distribution
```

Text file: <http://www.oikos.ekol.lu.se/appendixdown/snouterdata.txt>.
Text file: <http://www.oikos.ekol.lu.se/appendixdown/geefunctions.R>.