available at www.sciencedirect.com

ScienceDirect

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

Analyzing spatial autocorrelation in species distributions using Gaussian and logit models

G. Carl[a,b,*], I. Kühn[a,b]

[a] UFZ - Centre for Environmental Research Leipzig-Halle, Department Community Ecology (BZF), Germany
[b] Virtual Institute Macroecology, Theodor-Lieser-Strasse 4, 06120 Halle, Germany

ARTICLE INFO

Article history:
Received 21 November 2006
Received in revised form
20 April 2007
Accepted 24 April 2007
Published on line 1 June 2007

Keywords:
Clustered binary data
Generalized estimating equations
Logistic regression
Macroecological method
Moran's *I*
Spatial autocorrelation

ABSTRACT

Analyses of spatial distributions in ecology are often influenced by spatial autocorrelation. While methods to deal with spatial autocorrelation in Normally distributed data are already frequently used, the analysis of non-Normal data in the presence of spatial autocorrelation are rarely known to ecologists. Several methods based on the generalized estimating equations (GEE) are compared in their performance to a better known autoregressive method, namely spatially simultaneous autoregressive error model (SSAEM). GEE are further used to analyze the influence of autocorrelation of observations on logistic regression models. Originally, these methods were developed for longitudinal data and repeated measures models. This paper proposes some techniques for application to two-dimensional macroecological and biogeographical data sets displaying spatial autocorrelation. Results are presented for both computationally simulated data and ecological data (distribution of plant species richness throughout Germany and distribution of the plant species *Hydrocotyle vulgaris*). While for Normally distributed data SSAEM perform better than GEE, GEE provide far better results than frequently used autologistic regressions and remove residual spatial autocorrelation substantially when having binary data.

1. Introduction

Recently, it became obvious that spatial autocorrelation is often present in macroecological or biogeographical datasets. This means that observations close to each other geographically are more likely to be similar than those far away from each other. One therefore might have a lack of independence in a dataset. Even if we are not interested in analyzing the spatial structure within a dataset (i.e., the relationship of the response to spatial variables such as longitude and latitude) we should care for spatial autocorrelation as it can result in severe problems (Legendre, 1993; Legendre et al., 2002). The non-independence of outcomes represents a form of pseudoreplication and overestimates the effectively available degrees of freedom (Dutilleul, 1993; Legendre et al., 2002). Situations with correlated observations are not unusual in the statistical analysis of spatial data so that ordinary regression models are not appropriate. In particular, in geological and ecological applications, responses of regression models can be locally correlated in all geographical directions. Therefore, spatial autocorrelation may result in an underestimation of coefficients variances and incorrect parameter estimates of statistical models (Anselin and Bera, 1998; Lennon, 2000; Haining, 2003).

For data with Normally distributed responses, spatially autoregressive linear models are available and already used by

[*] *Corresponding author at*: UFZ - Centre for Environmental Research Leipzig-Halle, Department Community Ecology (BZF), Germany. Tel.: +49 345 558 5311.

E-mail addresses: gudrun.carl@ufz.de (G. Carl), ingolf.kuehn@ufz.de (I. Kühn).

0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.04.024

ecologists (Lichstein et al., 2002; Dark, 2004; Kissling and Carl, in press). Augustin et al. (1996) applied an autologistic method to handle presence/absence data which was later widely used by several authors (Wu and Huffer, 1997; Osborne et al., 2001; Segurado and Araujo, 2004; Luoto et al., 2005). This method can also be extended to other distributions (Haining, 2003). Furthermore, binary spatial regression models for spatially aggregated data have been proposed using spatial Markov random fields (Pettitt et al., 2002). Bayesian estimation methods are available using Markov chain Monte Carlo to approximate the posterior distributions (Besag et al., 1991; Kühn et al., 2006, for an application in ecology). The approach we present here is based on generalized estimating equations which is an extension of generalized linear models (GLM) (Liang and Zeger, 1986). This method allows for a correlation structure for longitudinal (one-dimensional, e.g., time series) data analysis. This special data layout is necessary when responses are measured repeatedly on the same subject or unit across time. Therefore, the GEE method takes into account correlations within these units, while correlations between units can be assumed to be zero. Such units or clusters can also occur in a spatial context, i.e., as independent regions. A GEE approach for the analysis of such kind of data was developed by Albert and McShane (1995).

In macroecological or biogeographical datasets natural clusters comply usually large regions as they represent biogeographical regions or so-called eco-regions (Korsch, 1999; Metzger et al., 2005). Though such a natural structure might exist, adjacent clusters would be neither absolutely independent nor necessarily present at the desired scale. This is the case, in particular, for lattice data, i.e., where maps are divided into grid cells of arbitrary size. We apply the methods introduced by Zeger and Liang (1986) by either having the entire area as one cluster or making up artificial clusters combining only a few grid cells. In the second case intra-cluster correlations are incorporated, while inter-cluster correlations are neglected. Therefore, potentially existing correlations among adjacent grid cells which belong to different clusters will not be used. We think, however, it is better to involve only those autocorrelations within clusters and ignore those between clusters than to completely ignore autocorrelation et all. We can thus apply generalized estimating equations to gridded datasets and analyze whether they produce more accurate values.

We are only aware of a few studies in ecology that actually use GEE to correct for spatial autocorrelation (Gotway and Stroup, 1997; Gumpertz et al., 2000; Augustin et al., 2005). But the authors of the cited papers used the GEE method only for examples of rather small sample size where data clustering is neither naturally present nor artificially to impose. We are not aware of any study that tests the performance of the method for data of gridded maps and large sample size.

In this paper, we first give an overview of the theory. To this aim we describe the GEE method and the use of standardized Pearson residuals and spatial correlograms. We then apply the method to synthetic data with specified spatial correlation and describe some aspects of implementation using the free software R (R Development Core Team, 2006). For comparison we analyze several modelling techniques. The GEE approach is especially suited for parameter estimation rather than prediction (Augustin et al., 2005). Thus, we compared results of several models regarding autocorrelation removal and efficiency of parameter estimates. Real examples are provided for both Normal and binary macroecological datasets on plant distribution in Germany. Our aim is to discuss several GEE models in comparison to linear models (LM) and spatially simultaneous autoregressive error models (SSAEM) for Normal response variables and to standard generalized linear models and autologistic regression models (autologistic) for binary response.

2. Statistical background

2.1. GEE method

Generalized estimating equations are an extension of generalized linear models and allow for correlated responses (Diggle et al., 1995). Firstly, consider the score equation of generalized linear models (Dobson, 2002; Myers et al., 2002).

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

where $\mathbf{y}$ is a vector of response variables. The expected value $\boldsymbol{\mu}$ is given by $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ with g^{-1} is the inverse of the link function, $\mathbf{X}$ the matrix of predictors, and $\boldsymbol{\beta}$ is the vector of regression parameters. Furthermore, $\mathbf{D}'$ 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 spatially independent. $\mathbf{V}^{-1}$ is the inverse of matrix $\mathbf{V}$ given by

$$\mathbf{V} = \phi\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2},$$

where $\mathbf{A}$ contains variances as usual in GLM, ϕ the dispersion parameter, and $\mathbf{R}$ is an additional matrix to incorporate the correlation structure. For formal reasons the diagonal matrix $\mathbf{A}$ is splitted up into two square roots of $\mathbf{A}$. In case that $\mathbf{R}$ is a specified matrix the score Eq. (1) can be used to estimate the regression parameters $\boldsymbol{\beta}$ in a similar way as in GLM. If the parameters in $\mathbf{R}$ are not given, it will be necessary to estimate all of them in an iterative procedure until convergence is achieved. However, when the set of unknown parameters is too large, this procedure may lead to convergence problems and may not longer be applicable.

We are able to reduce the number of unknown parameters if a cluster-specific model can be used. In this context, clusters are groups with certain sets of observations. Cluster models are models that recognize correlations within clusters, and neglect correlations between them. The complexity of the model can strongly be reduced by assuming the same kind of spatial autocorrelation in all clusters. Eq. (1) can thus be transformed into an equation of more practical value. For this purpose, suppose the sample is split up into m clusters and the complete dataset is ordered in a way that data of specific clusters follow each other. So we obtain, e.g., for the response variable $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_m')$. Here the model can be written in the form $\boldsymbol{\mu}_j = g^{-1}(\mathbf{x}_j'\boldsymbol{\beta})$, $j = 1, 2, \ldots, m$ for each cluster separately. Furthermore, with regard to each cluster a variance–covariance matrix $\mathbf{V}_j$ exists containing off-diagonal elements because of the dependence on intra-cluster observations. The complete variance–covariance matrix, however, will be obtained in a block diagonal form $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m)$, since inter-cluster responses are assumed to be uncorrelated. As a consequence Eq. (1) yields the so-called quasi-score equation

$$\sum_{j=1}^{m} \mathbf{D}_j' \mathbf{V}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j) = 0, \tag{2}$$

which sums over all clusters. Here the matrix $\mathbf{V}_j$ has a corresponding decomposition in the form

$$\mathbf{V}_j = \phi \mathbf{A}_j^{1/2} \mathbf{R}_j \mathbf{A}_j^{1/2}.$$

If correlation parameters are not given or datasets are large, Eq. (2) is the preferable form in comparison to Eq. (1). But it has seriously to be checked if applications to gridded datasets from macroecology are allowed since there is no natural clustering (Hosmer and Lemeshow, 2000; Myers et al., 2002). Fortunately, estimates of regression parameters are fairly robust against misspecification of the correlation matrix (Dobson, 2002).

2.2. Residual autocorrelation

One can define a vector $\boldsymbol{r}$ of standardized residuals for GEE models

$$\boldsymbol{r} = \mathbf{R}^{-1/2} \mathbf{A}^{-1/2} (\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{b})), \tag{3}$$

where $\boldsymbol{b}$ is the estimate of $\boldsymbol{\beta}$ and $\mathbf{R}^{-1/2} = (\mathbf{R}^{-1})^{1/2}$ (Lumley, 1996). To this end we decompose the inverse symmetric correlation matrix $\mathbf{R}^{-1}$ into a product of two symmetric matrices $\mathbf{R}^{-1/2}$ using the singular-value decomposition of a matrix. These standardized residuals are Pearson residuals standardized by the working correlation. They should be spatially independent if the used correlation structure is true. We will analyze these residuals by means of spatial correlograms and compare the results to those of Pearson residuals of GLM.

For this purpose we use Moran's I (e.g., Lichstein et al., 2002) given by

$$I = \frac{(1/S)\sum_i \sum_h w_{ih}(r_i - r_{\mathrm{mean}})(r_h - r_{\mathrm{mean}})}{(1/n)\sum_i (r_i - r_{\mathrm{mean}})^2}. \tag{4}$$

Here one has to introduce "lag distance" intervals for the spatial structure under consideration. The factor w_{ih} is a weight that equals one if the distance of the variables r_i and r_h belongs to this interval and zero otherwise. S is the sum of weights for a given interval and n is the total number of residuals. If there is no spatial autocorrelation, the expected value of I is $-1/(n-1)$, which can be approximated by 0 if n is large.

2.3. Simulations

Simulations were performed to check the models for autocorrelation effects. For this purpose regular grids were generated. The number of grid cells is 32 × 32 and the cells were assumed to be square. Values for two Normally distributed predictors were randomly generated, and linearly combined using specified parameters (intercept and two slopes). In addition, Normally distributed errors $\boldsymbol{\varepsilon} \sim N(0, 1)$ were randomly generated. The vector of errors was multiplied by the Cholesky decomposition of a variance–covariance matrix. This procedure creates correlated Normal random errors. Finally, we are able to simulate correlated responses. On one hand Normal responses are given as the sum of linear component and correlated errors. On the other hand the following steps transform these correlated Normal variables into correlated binary outcomes: (1) scale to get a marginally standard Normal distribution, (2) transform by their cumulative distribution function to get a uniform distribution, and (3) use the inverse transform method to get binomial responses (Ross, 1997).

The correlation matrix includes specified spatial autocorrelation depending on the distances between the points of measurements (e.g., centre points of grid cells). In our case this correlation is assumed to be equal for each pair of equal distance. In this way we have introduced an isotropic spatial autocorrelation structure by the following exponential function

$$\alpha = \alpha_1^{d_{ih}}. \tag{5}$$

Here d_{ih} is the distance between centre points of grid cells and α_1 is the correlation parameter for nearest neighbours. Note that two scale parameters are necessary for binary response data. The first one ensures the correct fit in case of uncorrelated errors. To check the fit GLM regressions for 1000 simulated datasets are performed. The second scale parameter has to preserve the specified error variance when the correlation is incorporated.

2.4. Software details

Our computations are based on software packages in the computer language R (R Development Core Team, 2006). The tools for SSAEM and autologistic computations are available in package spdep (Bivand et al., 2006) with functions named *errorsarlm* and *autocov_dist*. Here the calculations were carried out for 8 and 4 neighbours, respectively. The tools for calculating GEEs are available in package gee (Carey et al., 2006) with function named *gee* (Liang and Zeger, 1986; Zeger and Liang, 1986) and in package geepack (Yan, 2004) with function named *geese* (Yan, 2002; Yan and Fine, 2004).

2.5. Fitting GEE models

In our case of spatial dependence the following correlation structures are used in GEE models:

1. *Fixed*. The correlation structure is completely specified by the user and will not change during an iterative procedure.
2. *Quadratic*. Correlation parameters are to be estimated. But one can require that certain parameters must be equal. We set correlations to be equal within a cluster if the corresponding grid cells are of equal distances. It follows that for instance a square cluster of 2 × 2 square grid cells is completely characterized by 2 correlation parameters according to the 2 possible distances between these cells.

Table 1 – Results for 1000 randomly simulated 32×32 datasets with Normally distributed responses

Method	a	Mean(b_0)	Mean(b_1)	Mean(b_2)	Var(b_0)	Var(b_1)	Var(b_3)
LM		−0.9970	2.9959	−2.0005	0.0660	0.0236	0.0242
GEE *exch.* 2 × 2	1	−1.0011	2.9967	−1.9972	0.0451	0.0138	0.0136
GEE *quadr.* 2 × 2	2	−1.0038	2.9985	−1.9963	0.0424	0.0128	0.0126
GEE *exch.* 3 × 3	1	−1.0014	2.9976	−1.9984	0.0477	0.0149	0.0157
GEE *quadr.* 3 × 3	5	−1.0007	2.9983	−1.9999	0.0408	0.0114	0.0113
GEE *exch.* 4 × 4	1	−1.0007	2.9961	−1.9970	0.0516	0.0167	0.0165
GEE *quadr.* 4 × 4	9	−1.0034	2.9960	−1.9945	0.0523	0.0170	0.0165
GEE *fixed*	0	−0.9994	2.9968	−2.0003	0.0356	0.0096	0.0094
SSAEM		−0.9980	2.9971	−2.0008	0.0388	0.0105	0.0105

Means and variances for estimated regression parameters (intercept = b_0, slope of first predictor = b_1, slope of second predictor = b_2) compared for different methods. The real parameters are $\beta_0 = -1$, $\beta_1 = 3$, and $\beta_2 = -2$. The number a gives the number of correlation parameters to be estimated in GEE models.

These 2 correlation parameters are equal for all clusters. This imposes an isotropic correlation structure on each cluster.

3. *Exchangeable*. All correlations within clusters are equal such that only one parameter is to be iteratively estimated. It is the same one for all clusters.

We use the *fixed* structure to solve Eq. (1) where no clustering is necessary. In ecological applications we have good reasons to assume that the correlation decreases with increasing spatial distance. Therefore, we use the function (5) for computation of correlation parameters α for *fixed* structure. The parameter α_1 is estimated by Moran's I (4) at distance interval lag = 1 from the GLM residuals. This estimate is substituted in the GEE procedure so that all correlation parameters are given.

Furthermore, we use the structures *exchangeable* and *quadratic* to solve the cluster model of Eq. (2). Note that clusters combine some grid cells into a so-called neighbourhood. This can easily be done by an additional lattice with a mesh size of several grid cells. All grid cells within such a mesh are joined into a cluster. Here these clusters are, in general, of square shape. As we have pointed out above, it is necessary to take into consideration the influence of clustering data. Thus, different values for square cluster units were chosen: 2 × 2, 3 × 3 and 4 × 4-clusters. Note that clusters need not to be complete. Usually, there are several fragments at the outside margin of the map.

All models under consideration are summarized in the first column of Table 1. The number of correlation parameters to be estimated in GEE models is given in the second column.

3. Application to simulated data

In Tables 1 and 2 we show the efficiency of parameter estimates. Table 1 presents results for 1000 randomly generated datasets of sample size 32 × 32. All simulated datasets were created with Normally distributed responses and equal regression parameters. The real parameters are $\beta_0 = -1, \beta_1 = 3$, and $\beta_2 = -2$. This table provides the averaged estimates of regression parameters for different methods. We note that all means for intercept and slopes fit very well. However, we recognize differences in the variances. The maximum values for variances and, therefore, the lowest efficiency for parameter estimates can clearly be found for LM. The GEE *fixed* model has the least variances. This is, however, not surprising due to the simulated datasets. Here *fixed* correlation means that we know the correlation function because the same exponential function (5) was used in simulation and evaluation. SSAEM models are almost as good as GEE *fixed* models. Furthermore, the variances of cluster models are always clearly smaller than the

Table 2 – Results for 1000 randomly simulated 32×32 datasets with binary distributed responses

Method	a	Mean(b_0)	Mean(b_1)	Mean(b_2)	Var(b_0)	Var(b_1)	Var(b_3)
GLM		−1.0162	2.9990	−1.9840	0.2365	0.2454	0.1603
GEE *exch.* 2 × 2	1	−1.0168	3.0006	−1.9847	0.1560	0.2090	0.1246
GEE *quadr.* 2 × 2	2	−1.0177	2.9984	−1.9815	0.1547	0.2097	0.1238
GEE *exch.* 3 × 3	1	−1.0094	2.9970	−1.9862	0.1554	0.2075	0.1254
GEE *quadr.* 3 × 3	5	−1.0084	2.9909	−1.9809	0.1446	0.2060	0.1183
GEE *exch.* 4 × 4	1	−1.0034	2.9960	−1.9919	0.1616	0.2069	0.1260
GEE *quadr.* 4 × 4	9	−1.0030	2.9918	−1.9873	0.1641	0.2093	0.1314
GEE *fixed*	0	−1.0132	3.0002	−1.9832	0.1440	0.2015	0.1170
Autologistic		−4.7609	5.5820	−3.7207	0.5217	0.2475	0.2724

Means and variances for estimated regression parameters (intercept = b_0, slope of first predictor = b_1, slope of second predictor = b_2) compared for different methods. The real parameters are $\beta_0 = -1$, $\beta_1 = 3$, and $\beta_2 = -2$. The number a gives the number of correlation parameters to be estimated in GEE models.

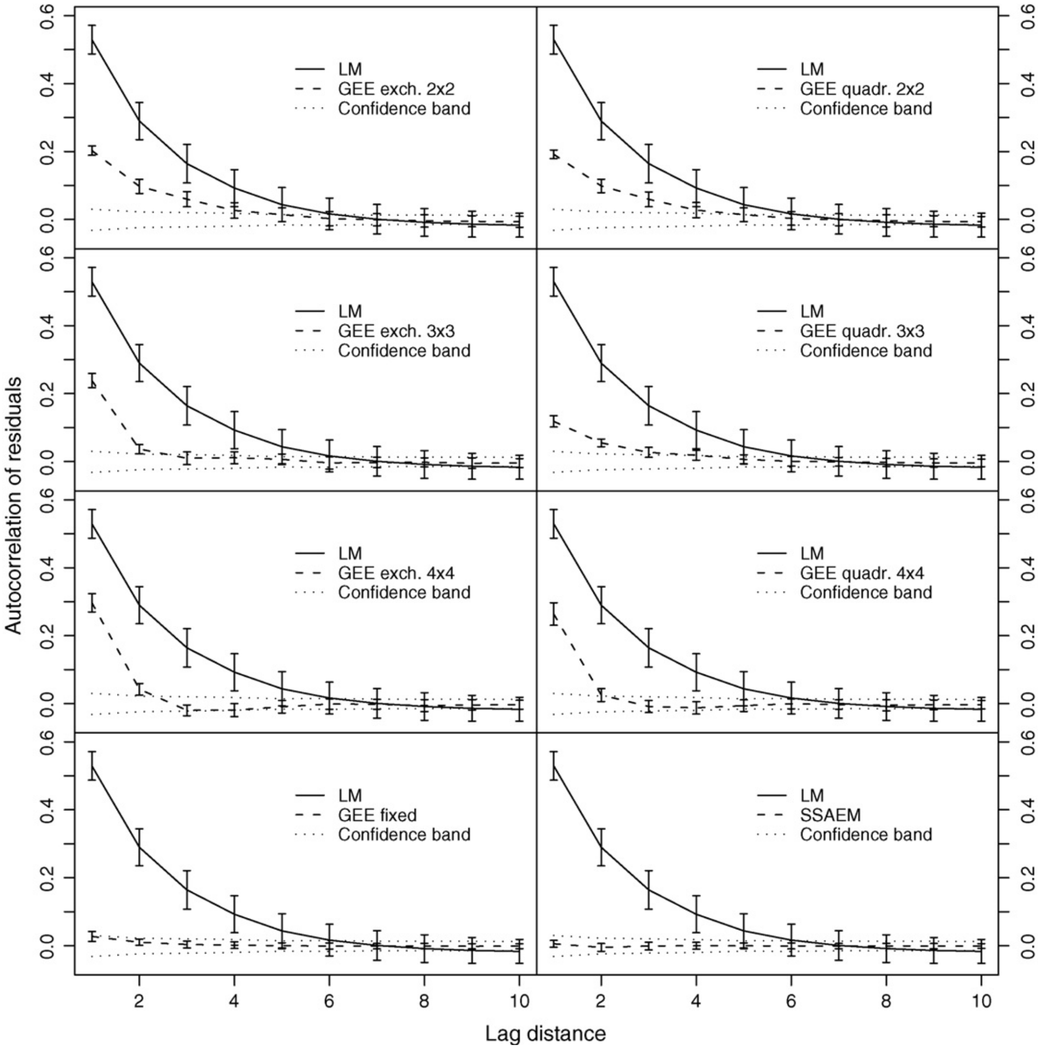

Fig. 1 – The average of residual autocorrelation for 1000 generated datasets of size 32 × 32. The responses are Normally distributed.

ones of LM. The GEE *quadr*. 3 × 3 model provides the largest efficiency gain among the cluster models under consideration.

In Table 2 corresponding results are given for datasets with binary responses. As for Normal data, we can draw the conclusion that variances decrease in GEE models compared to GLM. The GEE *quadr*. 3 × 3 model provides the largest efficiency gain among the cluster models as above for Normal data. Note that efficiency gains and reduced bias in standard error estimators result in more accurate inferences on β. However, the autologistic regression showed a very strong bias identifying such models as worse performer than the GEE approaches.

To know to what extent spatial autocorrelation is well described by GEE models we calculated Moran's I values for the standardized residuals by Eq. (4) and compared them to those of LM residuals. In Fig. 1 we present the average of residual autocorrelation for 1000 generated datasets of size 32 × 32. The responses are Normally distributed. These correlograms are calculated for the first 10 lag distance intervals. The first one, i.e., lag = 1 was chosen so that in all 8 cardinal directions nearest neighbours were included which means that there are 8 nearest neighbours at most (i.e., queen's neighbourhood). In Fig. 1 eight plots are given. In all plots LM results and confidence bands are the same, while the other models differ. As can be seen, the average correlation for nearest neighbours is about 0.53 in LM. SSAEM and also the GEE models essentially reduce the levels of autocorrelation. The GEE *fixed* model is nearly as good as SSAEM. Both are better than cluster models, in particular, for lag = 1. For GEE with *exchangeable* correlation structure, the Moran's I for lag = 1 increases with cluster size, while for GEE with *quadratic* correlation structure, the 3 × 3 clusters are best and 4 × 4 clusters worst.

The application to the average of residual autocorrelation for 1000 generated datasets of binary responses is shown in Fig. 2. Again, in all plots GLM results and confidence bands are the same, while GEE models differ. As for Normal data the *fixed* correlation structure is a good way to minimize autocorrelation. However, *fixed* structure means known structure in the case of simulated data as mentioned above. But also the GEE models with *exchangeable* and *quadratic* correlation structure are better than GLM. As we can see from our simulation, the

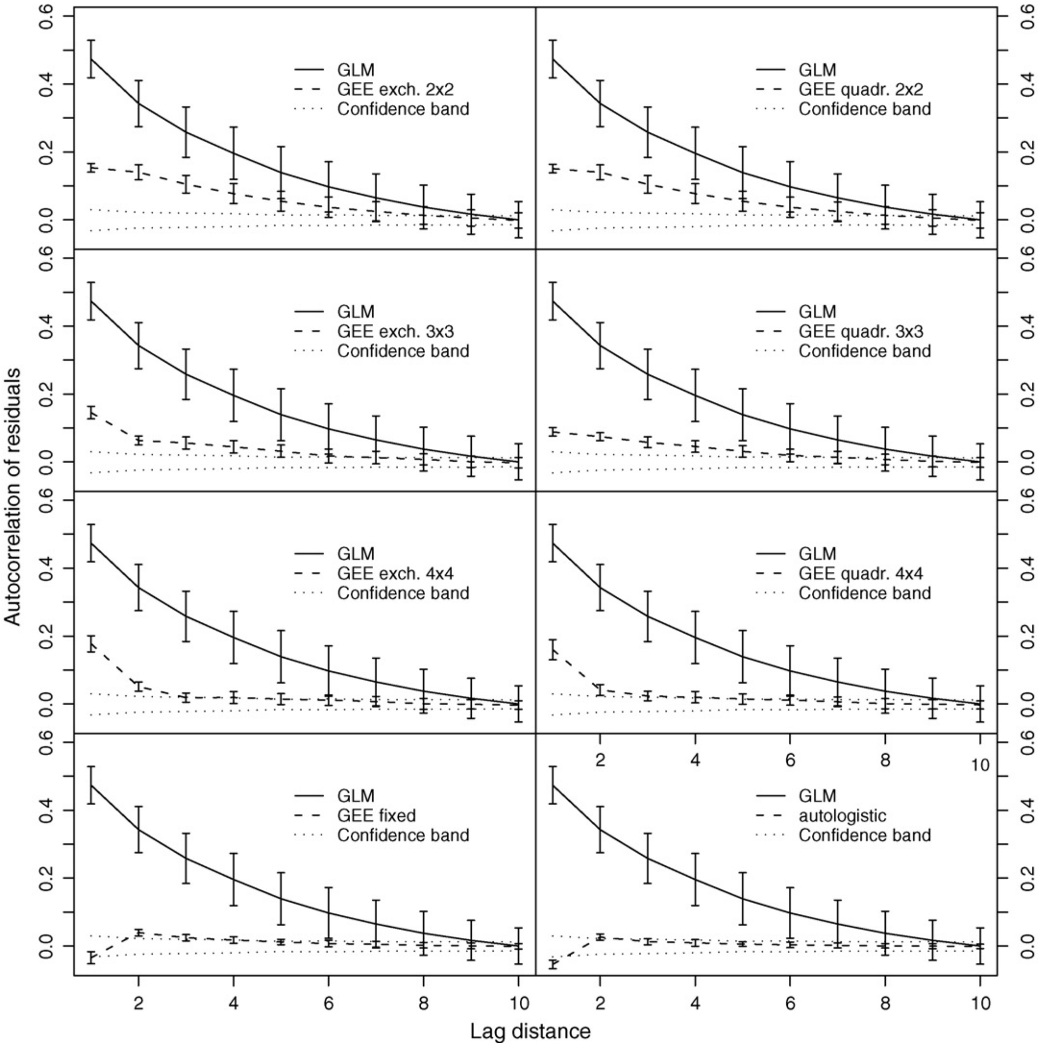

Fig. 2 – The average of residual autocorrelation for 1000 generated datasets of size 32 × 32. The responses are binary distributed.

level of autocorrelation of these GEE models is half or less of the autocorrelation of GLM if nearest neighbours are taken into account. The best cluster model seems to be the GEE *quadratic* 3 × 3 model as above in Tables 1 and 2. Here nearest neighbour correlation is reduced to about 0.09. The efficiency of the autologistic method in reducing spatial autocorrelation is comparable to the GEE *fixed* model.

For both distributions we can say that larger clusters have better fits for larger lags, but the nearest neighbour correlation will not be described in a better way. The GEE *quadratic* 3 × 3 models seem to be optimal. Thus, we reduce our further investigations to the study of 2 × 2 and 3 × 3 clusters.

Next we have to take into account that the pooling of grid cells in clusters is not naturally predetermined by the data structure. In our case neighbouring grid cells are combined to adjacent clusters of defined size. Aggregation depends on the initial point that is arbitrarily chosen by the user. This implies that the choice of neglected inter-cluster correlations is arbitrary as well. Nevertheless, the number of neglected correlations is almost completely specified by cluster size and sample size and the degree of arbitrariness is diminished as the number of clusters gets large. In Fig. 3 we present the residual autocorrelation averaged over all possible cluster settings (characterized by all possible translations leading to different settings compared to the original one). The calculations used an example of a generated 32 × 32 dataset per distribution. As can be seen, the errors are small even for 3 × 3 cluster models at small lag distance.

For Normal data we expect that the autocorrelation incorporated in the simulated data is completely absorbed in the response variable and is therefore detectable by Moran's *I* from the GLM residuals. In particular, for the GEE *fixed* model it is important to see how good α_1 is estimated by Moran's *I*. Moreover, for all methods which require the estimation of correlation parameters, these estimates can be compared to the true ones used in the simulation. We give the comparison of true and estimated correlation parameters for 4 different simulation scenarios (Table 3). Instead of Eq. (5) we used a modified exponential function

$$\alpha = \alpha_1^{(d_{ih}^{\nu})}, \tag{6}$$

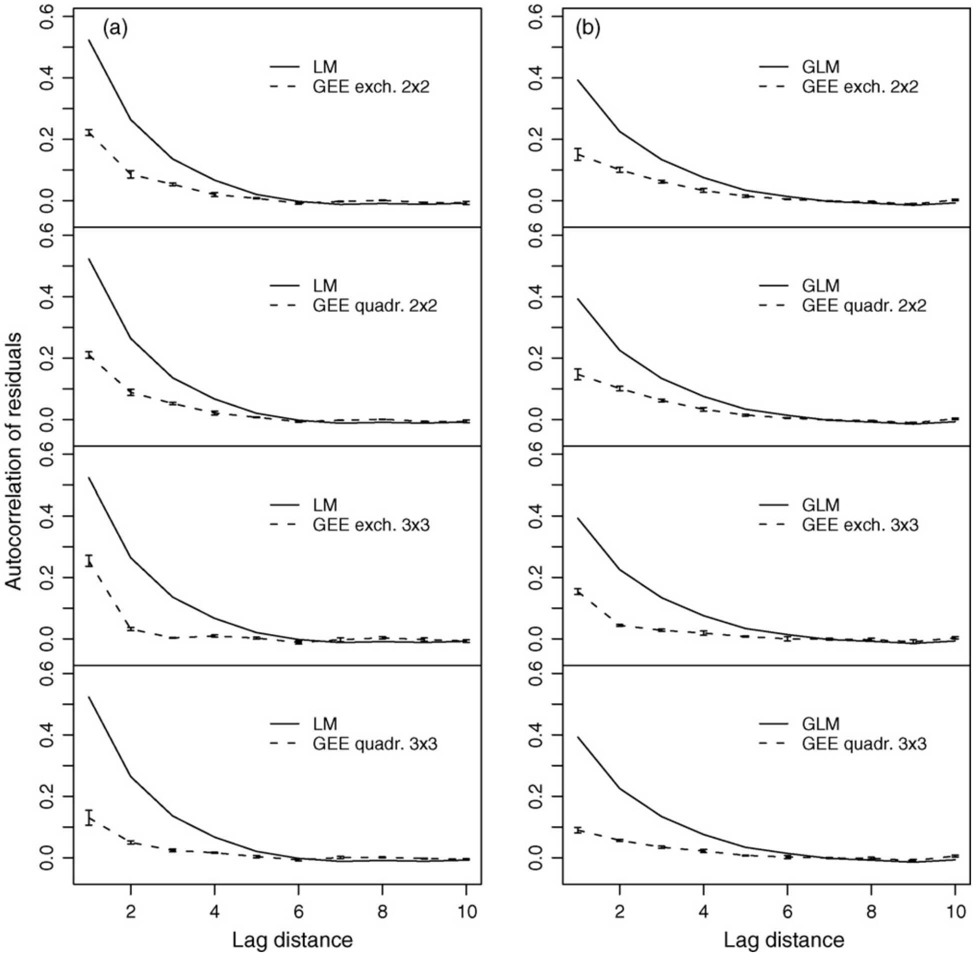

Fig. 3 – The residual autocorrelation averaged over different cluster settings. The responses of the arbitrarily chosen 32 × 32 dataset are (a) Normally and (b) binary distributed.

Table 3 – Comparison of true and estimated correlation parameters for 4 different simulation scenarios

Method	α_1	v	Mean($\alpha_{2\times2}$)	Mean($\alpha_{3\times3}$)	Mean($\alpha_{4\times4}$)
(1)					
True values	0.400	0.600	0.375	0.307	0.258
GEE *fixed*	0.382	0.699	–	–	–
GEE *exch.*	–	–	0.355	0.288	0.235
GEE *quadr.*	–	–	0.356	0.287	0.235
(2)					
True values	0.600	0.600	0.578	0.514	0.463
GEE *fixed*	0.542	0.714	–	–	–
GEE *exch.*	–	–	0.516	0.452	0.385
GEE *quadr.*	–	–	0.517	0.450	0.386
(3)					
True values	0.600	1.000	0.562	0.452	0.367
GEE *fixed*	0.590	1.048	–	–	–
GEE *exch.*	–	–	0.554	0.444	0.355
GEE *quadr.*	–	–	0.555	0.441	0.357
(4)					
True values	0.800	1.000	0.776	0.700	0.631
GEE *fixed*	0.772	1.082	–	–	–
GEE *exch.*	–	–	0.752	0.664	0.592
GEE *quadr.*	–	–	0.753	0.662	0.594

where an additional parameter υ is responsible for the range of autocorrelation. For the GEE *fixed* model, υ is estimated by the following formula: $\upsilon = \log(\log \alpha_5 / \log \alpha_1)/\log 5$. The parameters α_1 and α_5 are estimated by the Moran's I (4) of GLM residuals at distance intervals lag = 1 and lag = 5, respectively. In Table 3 we present the average of estimates for 100 generated datasets per scenario. Moreover, the correlation parameters are averaged per cluster. As can be seen, the deviations are small even for strong ($\alpha_1 = 0.8$) or long-range ($\upsilon = 0.6$) autocorrelation.

4. Application to the flora of Germany

In this section we apply the GEE methods to real macroecological datasets. We relate environmental variables to plant species distribution in Germany. Information on species distribution is available from FLORKART (see www.floraweb.de) which contains species location in a grid of 2995 grid cells. The cells of this lattice are 10′ longitude ×6′ latitude, i.e., about 11 km × 11 km, and therefore almost square cells. We selected species data for two regression models which differ in the response variable.

1. A dataset has been built with the Normally distributed number of all plant species found per grid cell (and ranging from 0 to nearly 1200).
2. A dataset for logistic regression has been chosen with binary distributed responses for presence/absence of the plant species *Hydrocotyle vulgaris* (see Fig. 4(a)).

We only choose two environmental variables (see Fig. 4(b and c)): (1) the average altitude (in 100 m units) per grid cell was calculated after the ARCDeutschland500 dataset, scale 1:500,000, provided by ESRI. (2) Mean annual temperature based on a 1 km^2 grid scale was provided by the "Deutscher Wetterdienst, Department Klima und Umwelt". Recording period for temperature data was 1951–1980.

We analyzed both datasets by the methods described above regarding the residual autocorrelation and the regression parameters $\boldsymbol{\beta}$. Here we used the cluster models with cluster sizes: 2 × 2 and 3 × 3, according to our conclusions from simulated data. In fact, 4 × 4 clusters lead to matrices $\mathbf{V}$ which are misspecified, i.e., not positive definite. In the GEE *fixed* model, we found better fits for long-range autocorrelation when correlations are computed by the modified exponential function (6).

Residual autocorrelations are presented in Fig. 5. All plots for Normal data of plant species richness in Germany (Fig. 5(a)) include the LM results and confidence bands for comparison. The calculations for confidence intervals were based on a Moran's I statistic of regression residuals. The corresponding software is provided in package spdep (Bivand et al., 2006). The results for binary data for presence/absence of *Hydrocotyle vulgaris* are presented in Fig. 5(b). Here calculations for confidence bands are no longer applicable. Moreover, GEE *fixed* did not yield results due to large matrices and thus computational problems. In that case the clustering of datasets is indispensable. For Normal data GEE *fixed* and SSAEM reduce the correlation to nearly zero. The cluster model which reduced residual autocorrelation best was GEE *quadratic* 3 × 3 in the case of Normal data. In the case of binary data the GLM autocorrelation is 0.2 at lag = 1 and thus rather small. Here all cluster models are of similar quality. Among the models, the GEE *quadratic* 3 × 3 model yields the minimal autocorrelation at lag = 1, while GEE *exchangeable* 3 × 3 yields the minima at lag = 2 and lag = 3.

Moreover, we compared the regression coefficients for the two predictors: altitude (Altit.) and temperature (Temp.). We deliberately chose two collinear predictors to test the ability of the models to discriminate between them and find the correct one. Species richness is usually a function of energy (i.e., temperature) (Currie, 1991; Hawkins et al., 2003) while *Hydrocotyle vulgaris* is distributed in the lowlands in Northern Germany due to specific environmental parameters different to temperature (Haeupler and Schönfelder, 1989; Benkert et al., 1996). Note that the assumption of isotropic spatial autocorrelation is reasonable for such regression models. There is no overall favoured direction in our data since neither wind,

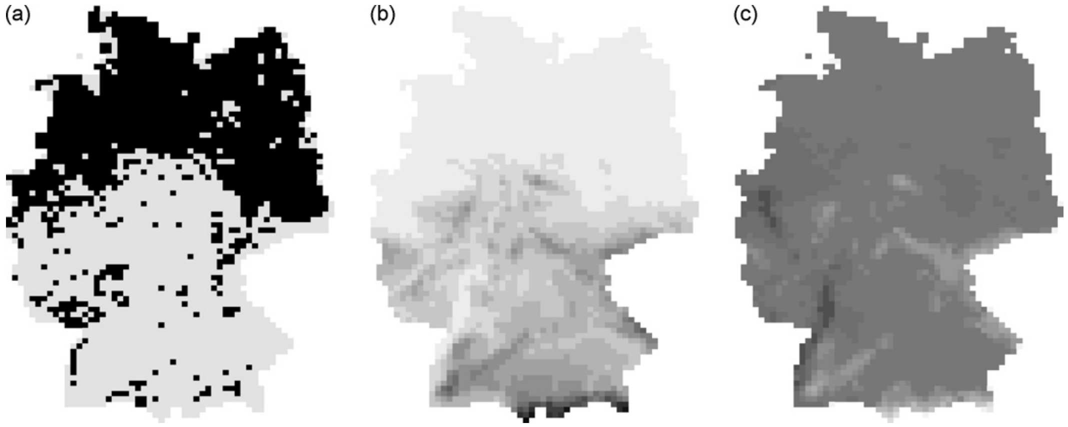

Fig. 4 – Distribution of data across Germany for (a) presence/absence (dark/light) of the plant species *Hydrocotyle vulgaris*, (b) altitude ranging from 0 (light) to nearly 18.23 (dark) (in 100 m units) and (c) temperature ranging from −0.5 (light) to 10.5 (dark) (in °C).

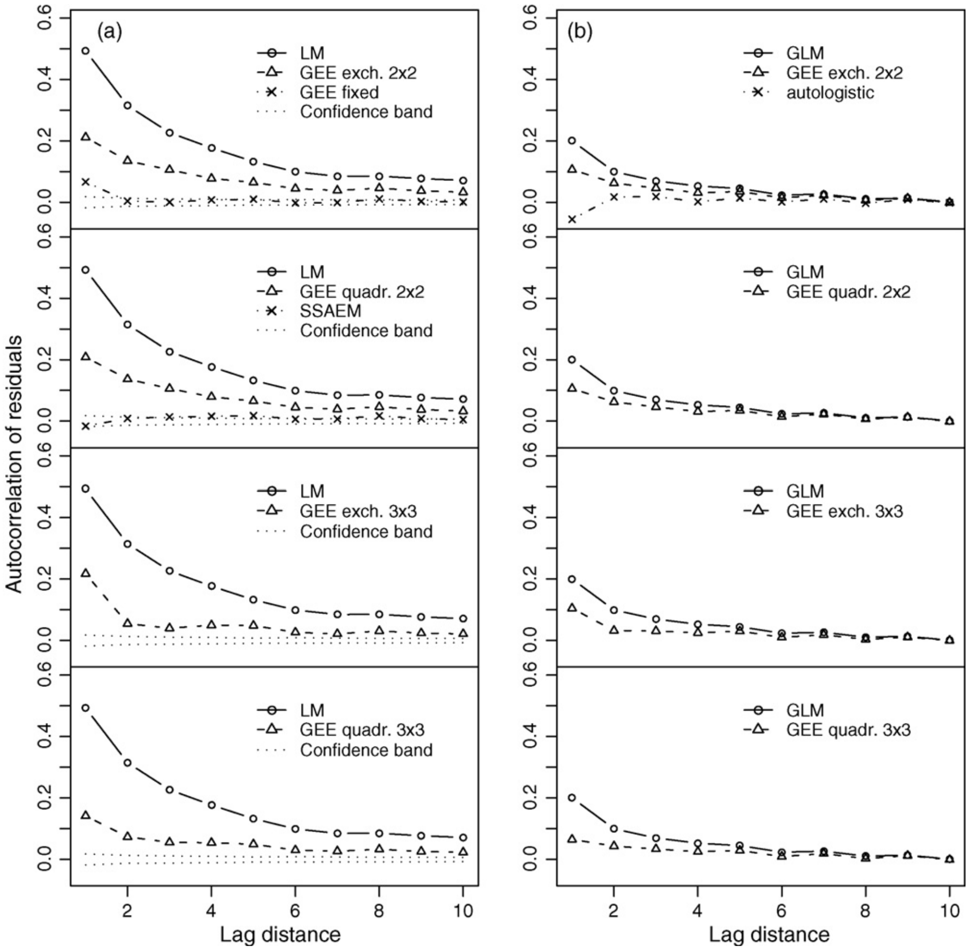

Fig. 5 – Autocorrelation of residuals (a) for Normal data of plant species richness in Germany and (b) for binary data for presence/absence of *Hydrocotyle vulgaris* in Germany.

water currents, soil transports nor any routes of migration constantly follow one certain direction across the area of analysis.

For Normal data the slopes for altitude and temperature decrease with increasing cluster size and the smallest slopes can be found for GEE *quadratic* 3 × 3, GEE *fixed* and SSAEM (Table 4). Note that the very small coefficients of altitude for GEE *fixed* and SSAEM are no longer significant. These models were found to be optimal for reducing autocorrelation (see Fig. 5(a)).

The results for binary data for presence/absence of *Hydrocotyle vulgaris* are given in Table 5. Here all GEE cluster models provide similar values for the slopes. GEE *fixed* did not yield results (see above). The GEE *exchangeable* 3 × 3 model provides the smallest absolute values. Note that in GEE models the regression parameters for temperature are not significant. The

Table 4 – Parameters for regression coefficients under various models for Normal data of plant species richness in Germany

Method	Intercept	Altit.	Temp.
GLM	−57.2	31.2***	78.7***
GEE *exch.* 2 × 2	27.7	25.8***	69.4***
GEE *quadr.* 2 × 2	36.0	25.6***	68.5***
GEE *exch.* 3 × 3	87.2	20.4***	63.4***
GEE *quadr.* 3 × 3	165.5**	18.1***	54.6***
GEE *fixed*	211.8***	3.3	47.7***
SSAEM	230.0***	0.8	51.6***

** Significant: $0.001 < p < 0.01$.
***Significant: $p < 0.001$.

Table 5 – Parameters for regression coefficients under various models for presence/absence of *Hydrocotyle vulgaris* in Germany

Method	Intercept	Altit.	Temp.
GLM	5.57***	-1.09***	−0.40***
GEE *exch.* 2 × 2	3.93**	−1.04***	−0.22
GEE *quadr.* 2 × 2	4.04***	−1.04***	−0.24
GEE *exch.* 3 × 3	3.75**	−0.98***	−0.22
GEE *quadr.* 3 × 3	3.80**	−1.00***	−0.22
Autologistic	−2.18**	−0.19***	−0.01

** Significant: $0.001 < p < 0.01$.
***Significant: $p < 0.001$.

results of the autologistic model are always very different to the ones of the GEE models.

5. Discussion

We presented a strategy for including autocorrelation in logistic regression models. We compared spatially simultaneous autoregressive error models (Cressie, 1993; Anselin and Bera, 1998) and autologistic regression models with generalized estimating equations. GEEs offer valuable methods for ecological applications. Paradis and Claude (2002) introduced a theoretical framework to correct phylogenetic autocorrelation using this method which was applied by Duncan and Blackburn (2004). In a spatial context, this method was used by Lennon et al. (2003) and Augustin et al. (2005). Nevertheless, our study is the first one that we are aware of which tests the performance for data of gridded maps and large sample size.

Autologistic regression is a more often used approach to analyze spatially autocorrelated binary data (Augustin et al., 1996; Wu and Huffer, 1997; Osborne et al., 2001; Segurado and Araujo, 2004; Luoto et al., 2005). However, estimating autologistic regression models is very time consuming, and do not correct for the available degrees of freedom. Furthermore, the incorporated covariable is the mean of the predicted values in the neighbourhood. This spatial covariate is at least potentially highly correlated with the response variable. Due to this, problems in model selection can arise (Dalgaard, 2002) as the autologistic variable might interfere with important ecological variables by being a better predictor and corrupt other important parameter estimates.

To examine the behaviour of GEE methods in detail we checked their performance for simulated datasets. We compared correlograms of the average Moran's *I* of standardized residuals. These provided a basis to decide which model is best at removing the effect of spatial autocorrelation, so that inferential results could be trusted. Although the real parameters of the German plant data analyses are unknown, due to general ecological theories (Currie, 1991; Hawkins et al., 2003) and knowledge on distribution and ecology (Haeupler and Schönfelder, 1989; Benkert et al., 1996), we can conclude, however, that those GEE models which reduced spatial autocorrelation best were also the ones which yielded the most sensible results. It is important to recognize that (1) all used GEE models are better than standard GLM and (2) the performance of the autologistic regression approach is very poor in our study with regards to parameter estimation. We can therefore not recommend the use of autologistic methods to account for spatial autocorrelation.

Due to a misspecification of the variance covariance structure in a model that does not account for spatial autocorrelation, significances could be overestimated (Anselin and Bera, 1998; Lennon, 2000). Diniz-Filho et al. (2003) discusses the problem whether one should take spatial autocorrelation into account or not quite carefully. They conclude that spatial autocorrelation can impose problems but that not necessarily all analyses ignoring spatial autocorrelation are flawed. Nevertheless, if the effect of spatial autocorrelation is not analysed, it remains unclear whether regression coefficients would be affected. As we have shown in our examples of real data, coefficients as well as significances may change (Kühn, 2007). So in cases where spatial autocorrelation is ignored, we have no idea about the correct relationship between variables in analyses that ignore spatial autocorrelation at all. It might be absolutely correct, or not.

6. Recommendations

The following recommendations result from our experience with using GEE and many other analyses to account for spatial autocorrelation and are only rules of thumb.

With Normally distributed, spatially autocorrelated data, we would preferably use ordinary spatially autoregressive methods such as SSAEM. SSAEM yields good results compared to LM and GEE and needs no specification for the correlation structure (Kissling and Carl, in press).

Having binary data, we recommend GEE. The use of GEE models reduced the autocorrelation of the residuals consid-

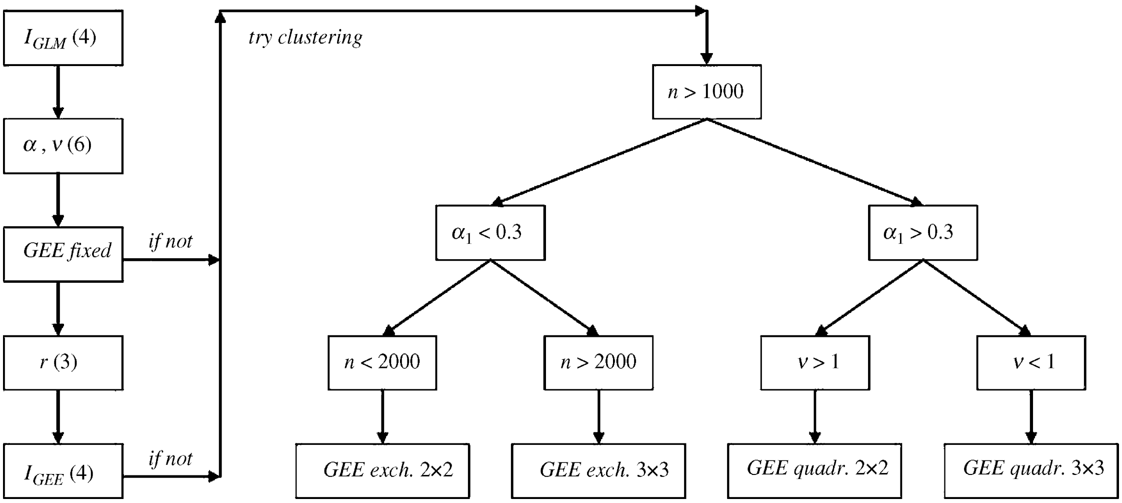

Fig. 6 – Block diagram of the GEE models. (1) Start with a GEE *fixed* model (left). Parameters are given by Eq. (4) for GLM residuals and Eq. (6). (2) If the dataset is too large or the autocorrelation removal is not acceptable, try clustering (right).

erably as measured by Moran's *I*. The GEE *fixed* models give the best results whenever the correlation structure can be specified by the user. Here the autocorrelation is reduced to nearly zero. We would therefore recommend the use of GEE *fixed* models (without clustering) wherever applicable (see Fig. 6, left). This depends on two problems: (1) The selection of a sensible correlation structure and (2) memory storage capacity for computations.

(1) In case that spatial stationarity and isotropy of autocorrelation can be assumed we can use the Moran's *I* (4) of GLM Pearson residuals to fit a correlation function. Here we recommend Eq. (6) as useful function (see (first) and (second) box in Fig. 6).
(2) Note the high need of memory storage capacity for datasets of large sample size without clustering (symbolized by the decision in the (third) box in Fig. 6).

If the dataset is too large (regarding to available memory) or one has neither prior knowledge nor sensible assumptions on the autocorrelation structure, cluster models can also be used, provided that the cluster size and the number of unknown parameters is small. We achieved this by setting all autocorrelation parameters in a cluster to an identical value (i.e., *exchangeable*) or defining a specific square (i.e., *quadratic*) structure where all correlation coefficients of the same distance are equal within clusters. Under these conditions also cluster models can decrease residual autocorrelation. Large clusters and a large number of unknown parameters can all lead to statistical and computational difficulties. However, the size of clusters and the number of parameters are related to the sample size. A decision tree (when to use which cluster model) is given in Fig. 6(right). We therefore think that cluster models will be useful especially for large datasets.

Acknowledgements

Gudrun Carl acknowledges a stipend from the federal state "Sachsen-Anhalt", Ministry of Education and Cultural Affairs. The authors acknowledge Integrated Project "ALARM: Assessing LArge scale environmental Risks with tested Methods" (GOCE-CT-2003-506675) (Settele et al., 2005) from European Commission within Framework Programme 6. We thank Boris Schröder for his valuable comments and suggestions on previous versions of this article, and thousand of volunteers mapping the flora of Germany which is incorporated in FLORKART, provided by the Federal Agency for Nature Conservation, namely Rudolf May, on behalf of the German Network for Phytodiversity (NetPhyD).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2007.04.024.

REFERENCES

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

Anselin, L., Bera, A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A., Giles, D. (Eds.), Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp. 237–289.

Augustin, N.H., Kublin, E., Metzler, B., Meierjohann, E., von Wuhlisch, G., 2005. Analyzing the spread of beech canker. Forest Sci. 51, 438–448.

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

Benkert, D., Fukarek, F., Korsch, H., 1996. Verbreitungsatlas der Farn- und Blütenpflanzen Ostdeutschlands. Fischer.

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

Bivand, R., et al., 2006. spdep: spatial dependence: weighting schemes, statistics and models. R package version 0.3–32.

Carey, V. J., 2006. Ported to R by Thomas Lumley (versions 3.13, 4.4, version 4.13)., B. R. gee: Generalized Estimation Equation solver. R package version 4.13-11.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, Cambridge.

Currie, D.J., 1991. Energy and large-scale patterns of animal-species and plant-species richness. Am. Nat. 137, 27–49.

Dalgaard, P., 2002. Introductory Statistics with R. Springer, New York.

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

Diggle, P.J., Liang, K.Y., Zeger, S.L., 1995. Analysis of Longitudinal Data. Clarendon, Oxford.

Diniz-Filho, J.A.F., Bini, L.M., Hawkins, B.A., 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, second ed. Chapman and Hall, London.

Duncan, R.P., Blackburn, T.M., 2004. Extinction and endemism in the new zealand avifauna. Global Ecol. Biogeogr. 13, 509–517.

Dutilleul, P., 1993. Modifying the t test for assessing the correlation between two spatial processes. Biometrics 49, 305–314.

Gotway, C.A., Stroup, W.W., 1997. A generalized linear model approach to spatial data analysis and prediction. J. Agric. Biol. Environ. Stat. 2 (2), 157–178.

Gumpertz, M.L., Wu, C., Pye, J.M., 2000. Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. Forest Sci. 46 (1), 95–107.

Haeupler, H., Schönfelder, P., 1989. Atlas der Farn- und Blütenpflanzen der Bundesrepublik Deutschland. Ulmer.

Haining, R.P., 2003. Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge.

Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guegan, J.F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M., Porter, E.E., Turner, J.R.G., 2003. Energy, water, and broad-scale geographic patterns of species richness. Ecology 84, 3105–3117.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, second ed. Wiley, New York.

Kissling, W., Carl, G. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecol. Biogeogr., in press, doi:10.1111/j.1466-8238.2007.00334.x.

Korsch, H., 1999. Chorologisch-ökologische Auswertungen der Daten der floristischen Kartierung Deutschlands. Schriftenreihe für Vegetationskunde 30, 1–200.

Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Diver. Distrib. 13, 66–69.

Kühn, I., Bierman, S.M., Durka, W., Klotz, S., 2006. Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. New Phytol. 172, 127–139.

Legendre, P., 1993. Spatial autocorrelation—trouble or new paradigm. Ecology 74, 1659–1673.

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25, 601–615.

Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. Ecography 23, 101–113.

Lennon, J.T., Smith, V.H., Dzialowski, A.R., 2003. Invasibility of plankton food webs along a trophic state gradient. Oikos 103, 191–203.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrica 73, 13–22.

Lichstein, J.W., Simons, T.R., Shriner, S.A., Franzreb, K.E., 2002. Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. 72 (3), 445–463.

Lumley, T., 1996. Xlispstat tools for building gee models. J. Stat. Software 1.

Luoto, M., Poyry, J., Heikkinen, R.K., Saarinen, K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. Global Ecol. Biogeogr. 14, 575–584.

Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Mucher, C.A., Watkins, J.W., 2005. A climatic stratification of the environment of Europe. Global Ecol. Biogeogr. 14, 549–563.

Myers, R.H., Montgomery, D.C., Vining, G.G., 2002. Generalized Linear Models. Wiley, New York.

Osborne, P.E., Alonso, J.C., Bryant, R.G., 2001. Modelling landscape-scale habitat use using gis and remote sensing: a case study with great bustards. J. Appl. Ecol. 38, 458–471.

Paradis, E., Claude, J., 2002. Analysis of comparative data using generalized estimating equations. J. Theor. Biol. 218, 175–185.

Pettitt, A.N., Weir, I.S., Hart, A.G., 2002. A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. Stat. Comput. 12, 353–367.

R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3–900051-07–0. http://www.R-project.org.

Ross, S.M., 1997. Simulation, second ed. Academic Press, San Diego.

Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.

Settele, J., Hammen, V., Hulme, P., Karlson, U., Klotz, S., Kotarac, M., Kunin, W., Marion, G., O'Connor, M., Petanidou, T., Peterseon, K., Potts, S., Pritchard, H., Pyšek, P., Rounsevell, M., Spangenberg, J., Steffan-Dewenter, I., Sykes, M., Vighi, M., Zobel, M., Kühn, I., 2005. Alarm: Assessing large scale environmental risks for biodiversity with tested methods. GAIA - Ecol. Perspect. Sci. Hum. Econ. 14, 69–72.

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

Yan, J., 2002. geepack:yet another package for generalized estimating equations. R News 2 (3), 12–14.

Yan, J., 2004. geepack: Generalized Estimating Equation Package. R package version 0.2–10.

Yan, J., Fine, J., 2004. Estimating equations for association structures. Stat. Med. 23, 859–874.

Zeger, S.L., Liang, K.Y., 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42, 121–130.