geographical analysis

Geographical Analysis ISSN 0016-7363

A Wavelet-Based Extension of Generalized Linear Models to Remove the Effect of Spatial Autocorrelation

Gudrun Carl[1,2], Ingolf Kühn[1,2]

[1]UFZ—Helmholtz Centre for Environmental Research, Department Community Ecology (BZF), Theodor-Lieser-Strasse 4, 06120 Halle, Germany, [2]Virtual Institute Macroecology, Theodor-Lieser-Strasse 4, 06120 Halle, Germany

Biogeographical studies are often based on a statistical analysis of data sampled in a spatial context. However, in many cases standard analyses such as regression models violate the assumption of independently and identically distributed errors. In this article, we show that the theory of wavelets provides a method to remove autocorrelation in generalized linear models (GLMs). Autocorrelation can be described by smooth wavelet coefficients at small scales. Therefore, data can be decomposed into uncorrelated and correlated parts. Using an appropriate linear transformation, we are able to extend GLMs to autocorrelated data. We illustrate our new method, called the wavelet-revised model (WRM), by applying it to multiple regression with response variables conforming to various distributions. Results are presented for simulated data and real biogeographical data (species counts of the plant genus Utricularia [bladderworts] in grid cells throughout Germany). The results of our WRM are compared with those of GLMs and models based on generalized estimating equations. We recommend WRMs, especially as a method that allows for spatial nonstationarity. The technique developed for lattice data is applicable without any prior knowledge of the real autocorrelation structure.

Introduction

Biogeographical studies are often based on a statistical analysis of data sampled in a spatial context. The analysis of such spatial data is complicated by spatial autocorrelation (Cressie 1993; Legendre 1993; Lichstein et al. 2002). Positive spatial autocorrelation occurs when adjacent data points (i.e., data sampled at adjacent locations) are more likely to be similar than distant ones. Unfortunately, standard methods such as generalized linear models (GLMs) may yield wrong results for hypothesis testing, because the presence of autocorrelation violates the basic assumption of independently and identically distributed (i.i.d.) errors, and hence

Correspondence: Gudrun Carl, UFZ—Helmholtz Centre for Environmental Research, Department Community Ecology (BZF), Theodor-Lieser-Strasse 4, 06120 Halle, Germany
e-mail: gudrun.carl@ufz.de

inflates type I errors. This outcome has been demonstrated by simulation studies (Legendre et al. 2002). A variety of methods have been developed to correct the effects of spatial autocorrelation in normal linear models, but only a few are applicable for nonnormally distributed response variables. Dormann et al. (2007) present an overview of such modeling approaches. They state that the most flexible methods applicable to non-normal distributions are spatial generalized linear mixed models (GLMMs), generalized estimating equations (GEEs; see also Carl and Kühn 2007), and spatial eigenvector mapping (SEVM; see also Griffith and Peres-Neto 2006), whereas the autocovariate regressions tend to perform poorly with regard to parameter estimates. Both SEVM and autocovariate regressions seek to capture latent spatial dependency in additional covariates. In autocovariate regression models, however, the application to non-Gaussian responses is problematic due to an intractable normalizing constant. To circumvent this complication, Markov chain Monte Carlo (MCMC) maximum likelihood estimation procedures need to be used (Kaiser and Cressie 1997; Huffer and Wu 1998). SEVM models are based on a spatial filtering methodology; a geographic structure matrix is diagonalized, and a subset of the resulting eigenvectors is added as synthetic explanatory variables in a multiple regression (Getis and Griffith 2002; Griffith 2003). The issue, however, is that a computationally intensive model selection procedure may be needed to select the best subset of eigenvectors.

In this article, we propose an alternative strategy of spatial filtering to circumvent these problems. Here, the pre-filtering process is carried out by means of wavelets instead of eigenvectors and without any additional covariates. Accordingly, our aim is to present a wavelet analysis that removes the effects of spatial autocorrelation in multiple regression. This wavelet analysis is also applicable to a regression in which a response variable has a non-normal distribution. Therefore, it is an extension of the GLM. Fadili and Bullmore (2001) provide a wavelet-based method for linear regressions in the context of autocorrelated errors, but only for normal linear models. They apply their method to neurophysiological time series. The main statistical use of wavelets, however, has been in non-parametric regression, noise removal, time-frequency analysis, and digital image compression, which are quite different issues (e.g., Nason and Silverman 1995; Bruce and Gao 1996). With respect to biogeographical or ecological applications, wavelet analysis seems to be a relatively unemployed tool (Bradshaw and Spies 1992; Dale 1999; Dale et al. 2002; Keitt and Urban 2005; Xiangcheng et al. 2005; Keitt 2007).

Our aim is to introduce wavelets for autocorrelation removal in GLMs. This type of regression is known as *parametric* regression, because it is based on a model that requires the estimation of a finite number of parameters. This concept is quite different from *nonparametric* regression, which allows for denoising and adapts to unknown smoothness. Wavelet techniques provide an effective tool for nonparametric regression. Here denoising is done by wavelet shrinkage; that is, by thresholding techniques in the wavelet domain. Flandrin (1992) shows the decorrelation property of wavelets. He derives formulas for the correlation structure of wavelet

coefficients; that is, in the wavelet domain. Furthermore, Johnstone and Silverman (1997) provide an extension of thresholding methods to deal with correlated data. They recommend level dependent thresholding for denoising. Our case of *parametric* regression, however, is different in various aspects. First, denoising by means of wavelets is not what we want to do here. Instead, the amount of error or noise is estimated by the parametric regression itself. Second, thresholding for wavelet coefficients is not what we are allowed to do. Thresholding would alter estimates for parameters. Instead, projection into an orthogonal subspace specified by a linear wavelet transform ensures that estimates are unaltered. Third, decorrelation for the wavelet coefficients is not what we need here. Instead, errors should be uncorrelated in the data domain. We demonstrate how correlated parts of data are removed by total exclusion of specific orthogonal components.

To the best of our knowledge, wavelets have not been used to remove spatial autocorrelation in GLMs. While we have shown the applicability of wavelets previously with specific error distributions (Carl, Dormann, and Kühn 2008; Carl and Kühn 2008), we principally extend the method here to the GLM. Our previous papers evaluate the performance of our models through comparisons of regression parameters (mean, variance), correlograms, and error calibration curves. Here, we systematically investigate the potential of wavelet-revised models (WRMs) regarding efficient parameter estimation and valid testing. We analyze the efficiency of parameter estimators and the validity of testing procedures as functions of both the strength of autocorrelation and the sample size. Additionally, we study how the choice of appropriate wavelets and levels makes an impact on the results.

Our new technique, called the WRM, is developed for (regularly gridded) lattice data. It is applicable without any prior knowledge of the real autocorrelation structure. Our method is demonstrated by its application to simulated datasets. We compare WRMs to GLMs and GEEs for normal, binary, and Poisson data. Moreover, the WRM is illustrated by its application to a real large-scale spatial dataset comprising the geographic distribution of a plant genus in Germany.

Methods

The quasi-score equation

In cases of correlated observations, the score equation of a GLM can be extended to the GEE, $\mathbf{U}_{GEE} = (\mathbf{MX})'\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$, where $\mathbf{y}$ is an $(n \times 1)$ vector of responses, $\mathbf{X}$ is an $(n \times p)$ matrix of predictors, and $\mathbf{M} = diag\{\partial\mu_i/\partial\eta_i\}$ is a diagonal matrix (Dobson 2002, p. 202; Myers, Montgomery, and Vining 2002, p. 202). Here the (canonical) link function is $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \eta_i, \quad i = 1, 2, \ldots, n$, with the expected value of the response being $E(y_i) = \mu_i$, n is sample size, and $\boldsymbol{\beta}$ is a $(p \times 1)$vector of unknown parameters.

The variance-covariance matrix can be written as $\mathbf{V} = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$, where $\mathbf{A} = diag\{v_{ii}\} = diag\{var(y_i)\}$ is a diagonal matrix and $\mathbf{R} = E[\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'\mathbf{A}^{-1/2}]$ is the correlation matrix. Thus, an alternative way to write the score vector is

$\mathbf{U}_{GEE} = (\mathbf{W}^{1/2}\mathbf{X})'\mathbf{R}^{-1}\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$, where $\mathbf{W} = diag\{v_{ii}^{-1}(\partial\mu_i/\partial\eta_i)^2\}$ denotes a diagonal weight matrix.

Generalizing this score according to a linear transformation $\mathbf{P}$ from generalized least-squares theory yields the transformations $\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) \rightarrow \mathbf{P}\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ and $\mathbf{W}^{1/2}\mathbf{X} \rightarrow \mathbf{P}\mathbf{W}^{1/2}\mathbf{X}$, where $\mathbf{P}$ is an $(n \times n)$ matrix. Here the correlation matrix is transformed as follows: $\mathbf{R} \rightarrow E[\mathbf{P}\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'\mathbf{A}^{-1/2}\mathbf{P}'] = \mathbf{P}\mathbf{R}\mathbf{P}'$. Assume that $\mathbf{P}$ is chosen to decorrelate the model; that is, that the transformed correlation matrix is approximately the identity matrix $\mathbf{I}$. Using $\mathbf{P}\mathbf{R}\mathbf{P}' = \mathbf{I}$, the score equation can be simplified to the following transformed equation:

$$\mathbf{U} = (\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})' \quad \mathbf{P}\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})) = \mathbf{0} \tag{1}$$

A vector $\mathbf{U}$ is identified as a ''quasi-score function'' provided that the following three properties hold (McCullagh and Nelder 1989): (i) the vector $\mathbf{U}$ has a multivariate normal distribution, at least asymptotically; (ii) its expected value is $E(\mathbf{U}) = \mathbf{0}$; and (iii) its variance-covariance matrix is equivalent to the negative expected value of the partial derivatives of $\mathbf{U}$ with respect to $\boldsymbol{\beta}$; that is, $E(\mathbf{U}\mathbf{U}') = -E[\partial\mathbf{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}]$.Condition (ii) is fulfilled because $E(\mathbf{y}) = \boldsymbol{\mu}$. If we assume that $\mathbf{P}\mathbf{R}\mathbf{P}' = \mathbf{I}$, then the information matrix is

$$\mathbf{i} = E(\mathbf{U}\mathbf{U}') = (\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})'(\mathbf{P}\mathbf{W}^{1/2}\mathbf{X}) \tag{2}$$

and condition (iii) is fulfilled. To verify condition (i), we may assume that there is only one parameter β. Then $U_{GLM}/\sqrt{i_{GLM}}$ is normally distributed, and $U/\sqrt{i}$ is a linear combination of $U_{GLM}/\sqrt{i_{GLM}}$. Thus, $U/\sqrt{i}$ is normally distributed as well. Moreover, if $\mathbf{P}\mathbf{R}\mathbf{P}' = \mathbf{I}$, then $U/\sqrt{i} \sim N(0, 1)$, at least asymptotically.

Consequently, $\mathbf{U}$ (1) is a quasi-score function. Thus $\mathbf{U}$ and $\mathbf{i}$ can be used for the method of scoring. Therefore, we obtain the iterative solution

$$\mathbf{b}^{(m)} = ((\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})'\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})^{-1}(\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})'\mathbf{P}\mathbf{W}^{1/2}\mathbf{z} \tag{3}$$

where

$$\mathbf{P}\mathbf{W}^{1/2}\mathbf{z} = \mathbf{P}\mathbf{W}^{1/2}\mathbf{X}\mathbf{b}^{(m-1)} + \mathbf{P}\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}) \tag{4}$$

From equation (2), the asymptotic variance-covariance matrix of $\mathbf{b}$ is given by $\text{var}(\mathbf{b}) = \mathbf{i}^{-1} = ((\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})'\mathbf{P}\mathbf{W}^{1/2}\mathbf{X})^{-1}$. If $\mathbf{X}$ and $\mathbf{P}\mathbf{W}^{1/2}\mathbf{X}$ have full rank, then equation (3) has a unique solution.

To summarize the above: we derive an extension of the GLM, including tools for parameter estimation and statistical inference, from a quasi-score concept using a linear transformation. Recall that in standard GLM theory, the responses are assumed to be independent, and thus the variance-covariance matrix of $\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ is equivalent to the identity matrix $\mathbf{I}$. If $\mathbf{P} = \mathbf{I}$, then our quasi-score concept is equivalent to a standard GLM result. In general, however, responses as well as predictors and errors are correlated rather than independent random variables. Therefore, it is desirable to revise data and to remove autocorrelations. Our aim is to specify the matrix $\mathbf{P}$ to ensure that the variance-covariance matrix of the variables

$\mathbf{PA}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$ is close to the identity matrix $\mathbf{I}$. We demonstrate that one can achieve this goal by using wavelet decomposition. Our new method, the WRM, introduced in a more generalized framework in the subsection ''The WRM,'' provides this revision of data.

A two-dimensional (2-D) wavelet approach

One objective of our study is to analyze data points that are spatially distributed. For this purpose we use the 2-D wavelet approximation of a function $F(x,y)$, as follows (Bruce and Gao 1996, p. 44; Shumway and Stoffer 2000):

$$F(x,y)=\sum_{m=1}^{3}\sum_{j=1}^{J}\sum_{k_x,k_y} d^m_{j,k_x,k_y}\Psi^m_{j,k_x,k_y}(x,y) + \sum_{k_x,k_y} s_{J,k_x,k_y}\Phi_{J,k_x,k_y}(x,y) \tag{5}$$

where the functions Ψ^m_{j,k_x,k_y} and Φ_{J,k_x,k_y} are given orthogonal wavelet functions generated from mother and father wavelets, respectively. Scaling and translation produce wavelets to different levels j and shifts k—that is, dilations and locations—whereas m corresponds to different spatial directions. These three directions correspond to wavelets that operate horizontally, vertically, or diagonally. The wavelets Ψ are used to describe the detail and high-frequency parts of given data, whereas the wavelets Φ are used to describe the smooth and low-frequency parts. Accordingly, the wavelet coefficients d represent the detail part, while the coefficients s represent the smooth part of F (Müller et al. 2003). Multiresolution analysis is a decomposition of equation (5) into orthogonal image components:

$$F^i(x,y)=D^m_j=\sum_{k_x,k_y} d^m_{j.k_x,k_y}\Psi^m_{j,k_x,k_y}(x,y),\quad i=1,\ldots,3J \tag{6}$$

$$F^{3J+1}(x,y)=S_J=\sum_{k_x,k_y} s_{J,k_x,k_y}\Phi_{J,k_x,k_y}(x,y) \tag{7}$$

Thus the function F can be reconstructed by a sum of all $3J+1$ image components $F(x,y)=\sum_i^{3J+1} F^i(x,y)$where J is the resolution level (or number of scales).

The *discrete* wavelet transform calculates the coefficients for a finite set of discrete data points. It is equivalent to a matrix multiplication. Thus, the 2-D discrete wavelet transform **T** enables us to transform discrete image data **F**—that is, a matrix **F**—into a matrix of wavelet coefficients **TF**. Matrix **TF** consists of all wavelet coefficients in a specific hierarchical order. Without loss of generality, we can assume that both are $(2^J\times 2^J)$matrices.

This decomposition allows us to analyze 2-D data such as a matrix or a geographical pattern of an ecological or an environmental variable. A 2-D approach can be applied to both response variables and individual predictors in multiple regression models if the components of these variables occur in a spatial context; for example, if these components are sampled in a plane. Thus, we must convert

these vectors into matrices that reflect the special spatial form; then the 2-D transform can be applied to each matrix built in this way. Finally, we revert to vectors that allow us to continue as usual in linear regression analysis.

The WRM

For biomedical time series, Meyer (2003) shows that the smooth trend belongs to a subspace spanned by large-scale wavelets. Similarly, we expect that we can detect and extract the feature of autocorrelation by means of wavelet decomposition. In particular, autocorrelation can be described by smooth wavelet coefficients at a small scale; that is, at resolution level 1 or 2. Consequently, these coefficients should be set to zero to capture only the noncorrelated part. Therefore, three steps are executed in rotation: wavelet transform, wavelet coefficient selection, and back transform. For a good approximation, transform

$$\mathbf{F} \rightarrow \mathbf{PF} = \sum_{i}^{3J} \mathbf{F}^{i} \tag{8}$$

provides a tool to remove autocorrelation by keeping only the so-called detail matrices $\mathbf{D}_j^m$ defined by equation (6) and by removing the smooth matrix $\mathbf{S}_J$ given by equation (7). If this data preparation is carried out by means of Haar wavelets, for example, then it has a very intuitive meaning. This procedure basically averages data values within squared subareas and subtracts this average from the data values to remove autocorrelation (Carl, Dormann, and Kühn 2008). If transform **P** is performed at a small scale—that is, for $J = 1$ or $J = 2$—then the subareas are small and autocorrelation is reduced to nearly zero, as measured by Moran's I value (Carl and Kühn 2008). Therefore, equations (3) and (4) in conjunction with the meaning of **P** given by equation (8) form our new method, WRM (see Appendix). If observations are originally autocorrelated, the use of these equations leads to improved parameter estimates compared with standard GLM results.

Three issues deserve particular mention. First, in the case where the matrix **P** is equal to **T**, we already get an approximately diagonalized variance-covariance matrix in the wavelet domain. This is called the whitening or decorrelating property of the discrete wavelet transform. However, as a consequence of this operation, different variances for detail and smooth coefficients would arise (Fadili and Bullmore 2001). This causes the new problem to get estimates for the variances of the coefficients at each level. Robust estimates are hard to find, in particular, if non-Gaussian distributions are considered. Second, our approach, which contains wavelet coefficient selection and back transform, circumvents this problem by means of equation (8). However, to gain this advantage, we have to use an $(n \times n)$ matrix **P** that is rank-deficient, which implies that condition $\mathbf{PRP}' = \mathbf{I}$ cannot exactly be fulfilled. It still holds approximately though. Third, if **X** has full rank and the number of observations n is much greater than the number of predictors p, then **PX** and $\mathbf{PW}^{1/2}\mathbf{X}$ have full rank as well. Therefore, our quasi-score concept leading to equations (3) and (4) can be based on **P** given by equation (8).

Application

Implementing the WRM

Our computations are based on software packages housed in the R language and environment for statistical computing (R Development Core Team 2006). The tools for calculating wavelet transforms are available in the package *waveslim* (Whitcher 2005). We used either the functions *dwt.2d* and *idwt.2d* for discrete wavelet transform and inverse discrete wavelet transform, respectively, or the function *mra.2d* for multiresolution analysis.

Because of the truncation to finite sets in discrete wavelet transforms, boundary treatment rules must be provided. In the 2-D discrete wavelet transform, type *periodic* is implemented for boundary conditions, causing a restriction on the sample size. The number of rows and columns must be divisible by 2^j in order to perform dilation and location of wavelets. In general, however, one wishes to analyze samples of arbitrary size. For this reason, we decide to pad data with zeros until a quadratic matrix of required size is reached.

To show the effectiveness of our model, we compare several WRMs with GLMs and GEEs for normal, binary, and Poisson's data. The tools for calculating a GEE are available in R package *gee* (Carey et al. 2002) with function *gee* (Liang and Zeger 1986; Zeger and Liang 1986), and in R package *geepack* with function *geese* (Yan 2002; Yan and Fine 2004). *Fixed* and *use-defined* correlation structures work best in the cases considered here (Carl and Kühn 2007; Dormann et al. 2007).

Simulation

Simulations were performed to check the models for autocorrelation effects (Ross 1997; see also Haining, Griffith, and Bennet 1983; Heagerty and Lele 1998), and regular grids were generated for this purpose. The cells were assumed to be square. Values for two normally distributed predictors were randomly generated. In addition, normally distributed errors were randomly generated. Both the vector of errors and the vectors of predictors were multiplied by the transpose of the Cholesky decomposition of a correlation matrix. This procedure creates correlated normal random errors and predictors, enabling the calculation of correlated responses. On the one hand, normal responses are given as the sum of a linear component and correlated errors. On the other hand, the following steps transform these correlated normal variables into correlated binomial or Poisson outcomes: (i) scale to get the standard normal distribution; (ii) transform by means of their cumulative distribution function to get a uniform distribution; and (iii) use the inverse transform method to get binomial or Poisson outcomes.

The aforementioned correlation matrix includes specified spatial autocorrelation depending on the distances between the observations. 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 an exponential function: $AC = \rho^{d_{jh}}$. Here d_{jh} is the Euclidean distance between center points of grid cells, and ρ is the correlation parameter for nearest neighbors.

An application to simulated data

The real parameters of our simulations are $\beta_1 = 1$ and $\beta_2 = 0$, that is, the linear component consists of only the first predictor. In the models, however, the second predictor is included as a dummy variable. To assess model performance, we analyze the efficiency of parameter estimates b_2 and the validity of their hypothesis tests. The simulated datasets were generated for $n = 30$-by-30 grid cells and $n = 50$-by-50 grid cells, each with autocorrelation parameters $\rho = 0$, 0.25, 0.5, and 0.75. Our evaluations are based on 1,000 simulation runs in each case.

We consider the following models: the GLM, the GEE, and the WRM. To assess the WRM, we utilize three different families of orthogonal wavelets: haar, d4, and la16 (Percival and Walden 2000). This choice reflects a mixed assortment of samples for the following wavelet characteristics in the spatial domain: haar wavelets are symmetric, block-shaped (i.e., nonsmooth), and very compact; d4 wavelets are asymmetric, nonsmooth, and compact; la16 wavelets are nearly symmetric (i.e., least asymmetric), smoother, and wider. On the one hand, the application of haar wavelets is justified because of their capability of detecting edges in the spatial domain (Carl and Kühn 2008). This is an important aspect for minimizing border effects. On the other hand, smoother wavelets such as d4 and, in particular, la16 have a higher number of vanishing moments and better frequency localization properties. Therefore, they work better as high-pass filters, which is essential for transform **P**. Each of the three wavelet families is combined with smooth coefficients removed at the lowest resolution levels 1 and 2. Therefore, we label the WRM models as follows: haar-1, haar-2, d4-1, d4-2, la16-1, la16-2 (Table 1).

Alpargu and Dutilleul (2006) provide tools for an efficiency analysis of slope estimators. Accordingly, the mean squared error of β_2 is $MSE = \left(\sum_1^{1000} {b_2}^2\right)/1000$. Here we want to analyze efficiencies in relation to the GLM. The efficiency of any model M is defined as $Eff(M) = MSE(b_2, M)/MSE(b_2, GLM)$. Therefore, if $Eff(M)$ is greater (smaller) than 1, then model M is less (more) efficient than the GLM. In Table 1, we present these parameter estimate efficiencies for the models applied to our simulated normal, binomial, and Poisson random variables. Our comparisons emphasize the influence of sample size and strength of autocorrelation on efficiency. WRM models of level 2 are more efficient than those of level 1 for weak autocorrelation, and vice versa. WRM models of level 1 are more efficient than those of level 2 for strong autocorrelation because the strength of autocorrelation influences the range of autocorrelation. Furthermore, the WRM is more efficient in estimating the parameters than the GLM for cases of strong autocorrelation. The expectation is that the WRM is somewhat less efficient than the GEE because prior knowledge about the form of the error variance-covariance matrix is completely incorporated into the GEE models here. However, each of these GEE models is computable only for small samples. Note that there is no need for such prior knowledge in the WRM and that the sample size can be rather large.

To assess the relative performance of parameter estimates in terms of type I errors (i.e., the probability α of falsely rejecting the null hypothesis H_0: $\beta = 0$), we

Table 1 Results for Efficiencies of Parameter Estimates for All Models Estimated with Normal, Binomial, and Poisson Simulated Data

Model	$n = 30$-by-30				$n = 50$-by-50			
	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
Gaussian								
GEE	1.00	0.75	0.30	0.06	—	—	—	—
WRM haar-1	1.28	1.04	0.53	0.22	1.28	1.04	0.46	0.17
WRM haar-2	1.05	0.95	0.56	0.25	1.05	0.92	0.57	0.22
WRM d4-1	1.34	0.96	0.47	0.15	1.26	1.01	0.44	0.12
WRM d4-2	1.06	0.89	0.56	0.20	1.06	0.93	0.53	0.18
WRM la16-1	1.35	0.96	0.45	0.13	1.27	1.02	0.42	0.11
WRM la16-2	1.03	0.93	0.49	0.18	1.03	0.94	0.53	0.15
Binomial								
GEE	1.00	0.92	0.59	0.28	—	—	—	—
WRM haar-1	1.27	1.21	0.77	0.42	1.39	1.27	0.78	0.36
WRM haar-2	1.04	1.01	0.76	0.46	1.08	1.04	0.72	0.37
WRM d4-1	1.28	1.22	0.78	0.37	1.37	1.25	0.79	0.31
WRM d4-2	1.06	0.99	0.75	0.41	1.09	1.04	0.70	0.33
WRM la16-1	1.26	1.24	0.78	0.38	1.35	1.25	0.79	0.32
WRM la16-2	1.04	1.00	0.68	0.39	1.07	1.00	0.72	0.32
Poisson								
GEE	1.00	0.82	0.39	0.13	—	—	—	—
WRM haar-1	1.26	1.21	0.59	0.28	1.28	1.17	0.55	0.24
WRM haar-2	1.05	0.99	0.62	0.34	1.08	1.00	0.60	0.30
WRM d4-1	1.31	1.25	0.58	0.22	1.33	1.21	0.51	0.19
WRM d4-2	1.05	1.01	0.64	0.29	1.10	1.00	0.58	0.24
WRM la16-1	1.32	1.25	0.55	0.22	1.33	1.18	0.51	0.18
WRM la16-2	1.07	1.01	0.57	0.26	1.08	0.99	0.57	0.24

Efficiency is given as a function of sample size and strength of autocorrelation.

calculated the empirical probability of significance P. For this purpose, we recorded how often the probability value of the dummy variable was falsely estimated to be less than a theoretical significance level of 0.05. Then the empirical significance level P for 1,000 simulation runs is given by $P = n_r/1000$, where n_r is the number of rejections. A validity analysis takes into account that the standard deviation of P is $\sigma_P = \sqrt{P(1-P)/1000}$ (Alpargu and Dutilleul 2006). Hence we used the confidence band $P \pm 2\sigma_P$ (i.e., $0.035 < P < 0.065$) as a validity condition. The results are given in Fig. 1. All plots show the empirical significance level P calculated for the above-mentioned models and for an autocorrelation parameter ρ ranging from 0 to 0.75. The methods work well when their curves coincide with the confidence band, even for strong autocorrelation. The GLM considerably overestimates the true type I error, whereas both the GEE and WRM la16-2 yield very

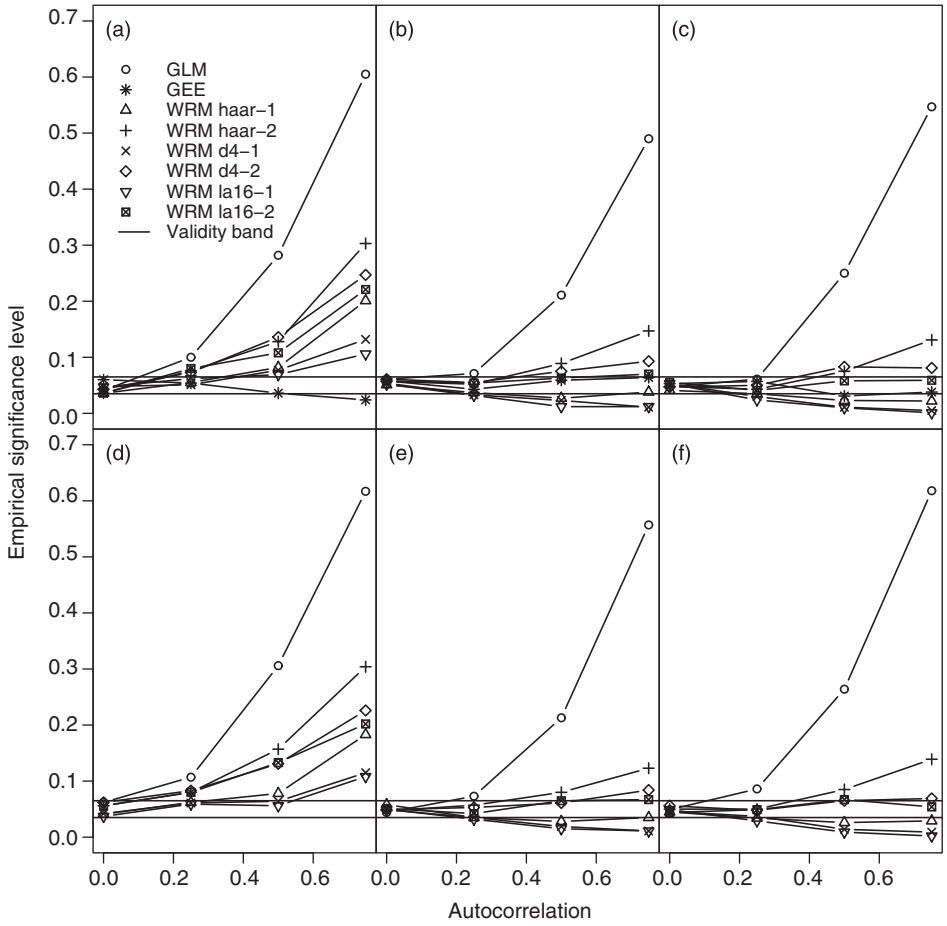

Figure 1. Performance of parameter estimates in terms of type I errors as the empirical probability of significance *P* of different methods for a theoretical significance level of 0.05. The validity band is given as $0.035 < P < 0.065$. The probabilities are calculated based on 1,000 randomly generated datasets for: (a) a Gaussian distribution, 30-by-30 grid cells; (b) a binomial distribution, 30-by-30 grid cells; (c) a Poisson distribution, 30-by-30 grid cells; (d) a Gaussian distribution, 50-by-50 grid cells; (e) a binomial distribution, 50-by-50 grid cells; and (f) a Poisson distribution, 50-by-50 grid cells.

good curves for binary and Poisson distributions. Although just missing the mark, WRM d4-2 and WRM haar-1 also provide good results.

An application to the flora of Germany

In this section, we apply the wavelet-revised methods to a real macroecological dataset. We relate environmental variables to plant species distribution in Germany. Information on species distribution is available from FLORKART (see http://www.floraweb.de), which contains species locations in a grid of 2,995 grid cells.

The cells of this lattice are 10′ longitude-by-6′ latitude—that is, about 11-by-11 km^2 in central Germany—and therefore almost square cells. Gridded datasets like these are typical for a plethora of species distribution atlases throughout Europe.

A dataset for regression has been chosen with Poisson distributed responses; that is, species count data of plant genus *Utricularia* (bladderworts), whose ecological behavior is well known. *Utricularia* is a genus of carnivorous plant species typical of mostly oligotrophic (i.e., nutrient-poor) ponds, pools, small lakes and reservoirs, ditches, and other small water bodies in swampy, boggy, or otherwise wet regions.

We chose three environmental variables: (i) the average *altitude* (in 1000 m units) per grid cell was calculated after the ARCDeutschland500 dataset, scale 1:500,000, provided by ESRI; (ii) the average annual *precipitation* from 1961 to 1990 (in 1,000 mm units) using data from the German weather service (Deutscher Wetterdienst [DWD]; homogenization and interpolation by the Potsdam Institute for Climate Impact Research [PIK]) averaged and referenced to our grid system within the project Modelling the Impact of Climate Change on Plant Distribution in Germany, funded by the Federal Agency for Nature Conservation (BfN) (Badeck et al. 2008); and (iii) the area of *moor* or *swamp* per grid cell according to the Corine Land Cover classes (in 10 km^2 units) (Statistisches Bundesamt 1997).

Wavelet-revised regression was carried out as previously described using haar, d4, or la16 wavelets and smooth coefficients removal at resolution level 1 or 2. The original grid cells data were padded with zeros to attain a 128-by-128 matrix. An impact study on zero padding and border effects yields the following results: only WRM d4-1 and WRM la16-1 provide a problematical spread of residuals toward the periphery. The d4 or la16 wavelets need to be applied at the appropriate resolution level. Haar wavelets, however, are less sensitive; that is, better in edge detection.

Poisson regression estimation results for *Utricularia* are given in Table 2. Here the performance of the WRM can be compared with that of the GEE and of the

Table 2 Results for Estimated Regression Parameters b_j and their *P*-values, Comparing Different Methods for the Plant genus *Utricularia* in Germany Treated as a Poisson Random Variable

Model	Intercept		Altitude		Precipitation		Moor/swamp	
	b_0	P	b_A	P	b_P	P	b_M	P
GLM	−0.63	<0.001	−0.93	<0.001	0.44	<0.001	0.60	<0.001
GEE	−0.37	0.0035	−0.90	<0.001	0.10	0.4790	0.46	<0.001
WRM haar-1	−0.72	0.0069	−1.31	0.0148	0.17	0.4928	0.44	<0.001
WRM haar-2	−0.60	<0.001	−0.86	0.0010	0.28	0.0772	0.46	<0.001
WRM d4-1	−0.57	0.0821	−1.55	0.0129	−0.37	0.1694	0.40	0.0015
WRM d4-2	−0.75	<0.001	−1.52	<0.001	0.34	0.0862	0.47	<0.001
WRM la16-1	−0.49	0.1533	−3.23	<0.001	−0.31	0.2492	0.40	0.0082
WRM la16-2	−0.54	0.0103	−1.22	0.0015	0.11	0.5864	0.45	<0.001

GLM. For Poisson data of this sample size, a proper GEE approach is difficult to establish. The number of correlation parameters that can be estimated in an iterative procedure is limited. Here, we present the results of an approximation that partly neglects correlations in GEEs. Moreover, we are able to discuss the results on the basis of our geographical and biological knowledge. Although *Utricularia* is restricted to wet areas, its count data should not be positively correlated with the predictor *precipitation* since the habitats of *Utricularia* depend on geology, geomorphology, and human land use, rather than on precipitation. In fact, the GLM results are corrected by the GEE and by all WRM models in this way. They thus provide more sensible results ecologically. The regression parameter for *precipitation* is reduced, and it is no longer significant at the $\alpha = 0.05$ level, as can be seen by its *P*-value.

Conclusion

The main contribution of this work is the development of an extension of GLMs for correcting data with respect to autocorrelation in a more general framework than previously done and a demonstration of its application to relevant datasets. The extension of GLMs is based on discrete wavelet transforms and is carried out through a two-dimensional analysis. Thus, multiresolution analysis gives the background for autocorrelation extraction from 2-D (image) data. After describing the model, we present an algorithm to estimate regression parameters, and we calculate the information matrix as a tool for hypothesis testing.

Using wavelets, one is able to reduce spatial autocorrelation of regression residuals, as can be measured by its Moran's *I* value (Carl and Kühn 2008), in a stepwise fashion. Our WRM, based on this characteristic, is shown to be more efficient in estimating the parameters than the GLM for datasets affected by substantial autocorrelation. The main difficulty arising from the GLM applied to such data is that autocorrelation affects its performance in terms of type I errors (i.e., the probability of falsely rejecting the null hypothesis). Therefore, we analyzed whether the WRM test statistic is valid and found that its empirical significance levels are very close to its theoretical counterparts. For binomial and Poisson distributed responses, for instance, we considered it best to utilize la16 wavelets at resolution level 2. In that case, the WRM test statistic is valid even for strong and long-range autocorrelation. Moreover, we compared the results with those of GEEs. These latter models are useful to correct autocorrelation effects successfully when the correlation structure is known, as in our simulated datasets. Wavelets, however, provide a powerful method for removing autocorrelation without any prior knowledge of the underlying correlation structure in lattice data.

Furthermore, by using a real biogeographical dataset and by basing the findings not only on statistics but also on prior knowledge in ecology and geography, we argue that the WRM results can be more plausible than the GLM results.

We therefore suggest the use of WRMs, especially when analyzing data observed on a regular two-dimensional lattice and characterized by large sample size. The WRM effectively removes spatial autocorrelation with or without any prior knowledge of the underlying correlation structure and is a computationally very fast and efficient procedure. Moreover, it is a method that allows for spatial nonstationarity, whereas stationarity is a basic and strong assumption for most of the standard methods.

Acknowledgements

We acknowledge support by the Virtual Institute for Macroecology, the Integrated Project ALARM: Assessing LArge scale environmental Risks with tested Methods (GOCE-CT-2003-506675) from European Commission within Framework Programme 6 as well as Sven Pompe (UFZ) and Franz Badeck (PIK) of the project Modelling the Impact of Climate Change and Plant Distribution in Germany, funded by the German Federal Agency for Nature Conservation (BfN) (FKZ: 805 81 00), for providing precipitation data.

Appendix

We give the following algorithm for calculating regression parameters via the WRM:

Step 1. Find an initial estimate $\mathbf{b}_0$.
Step 2. Use $\mathbf{b}_0$ to obtain $\boldsymbol{\mu}_0$.
Step 3. Calculate $\mathbf{A}_0$, $\mathbf{W}_0$, $\mathbf{X}_{\text{new}} = \mathbf{W}_0^{1/2}\mathbf{X}$, $\mathbf{z}_{\text{new}} = \mathbf{W}_0^{1/2}\mathbf{z}_0$.
Step 4. Create matrices for $\mathbf{z}_{\text{new}}$ and all columns of $\mathbf{X}_{\text{new}}$ according to the spatial structure.
Step 5. Perform multiresolution analysis on these matrices.
Step 6. Sum all orthogonal image components except the last one to pick up all detail components (and no smooth ones). This yields $\mathbf{Pz}_{\text{new}}$ and $\mathbf{PX}_{\text{new}}$.
Step 7. Transform matrices into vectors.
Step 8. Use equation (3) to estimate the regression model and to obtain the new estimator $\mathbf{b}_1$.
Step 9. Substitute $\mathbf{b}_1$ for $\mathbf{b}_0$ and return to step 2.
Step 10. Continue iterating until convergence is achieved.

References

Alpargu, G., and P. Dutilleul. (2006). ''Stepwise Regression in Mixed Quantitative Linear Models with Autocorrelated Errors.'' *Communications in Statistics—Simulation and Computation* 35, 79–104.

Badeck, F., S. Pompe, I. Kühn, and A. Glauer. (2008). ''Wetterextreme und Artenvielfalt, Zeitlich hochauflösende Klimainformationen auf dem Messtischblattraster und für Schutzgebiete in Deutschland.'' *Naturschutz und Landschaftsplanung* 40, 343–45.

Bradshaw, G. A., and T. A. Spies. (1992). ''Characterizing Canopy Gap Structure in Forests Using Wavelet Analysis.'' *Journal of Ecology* 80, 205–15.

Bruce, A., and H. Y. Gao. (1996). *Applied Wavelet Analysis with S-Plus*. New York: Springer.

Carey, V. J. ported to the R software by T. Lumley (versions 3.13 and 4.4) and B. Ripley (version 4.13). (2002). ''gee: Generalized Estimation Equation solver.'' R package version 4.13-10.

Carl, G., C. F. Dormann, and I. Kühn. (2008). ''A Wavelet-based Method to Remove Spatial Autocorrelation in the Analysis of Species Distributional Data.'' *Web Ecology* 8, 22–29.

Carl, G., and I. Kühn. (2007). ''Analyzing Spatial Autocorrelation in Species Distributions Using Gaussian and Logit Models.'' *Ecological Modelling* 207, 159–70.

Carl, G., and I. Kühn. (2008). ''Analyzing Spatial Ecological Data Using Linear Regression and Wavelet Analysis.'' *Stochastic Environmental Research and Risk Assessment* 22, 315–24.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Cambridge, UK: Wiley.

Dale, M. R. T. (1999). *Spatial Pattern Analysis in Plant Ecology*. Cambridge, UK: Cambridge University Press.

Dale, M. R. T., P. Dixon, M. J. Fortin, P. Legendre, D. E. Myers, and M. S. Rosenberg. (2002). ''Conceptual and Mathematical Relationships among Methods for Spatial Analysis.'' *Ecography* 25, 558–77.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl et al. (2007). ''Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review.'' *Ecography* 30, 609–28.

Fadili, M. J., and E. T. Bullmore. (2001). ''Wavelet-Generalized Least Squares: A New BLU Estimator of Linear Regression Models with 1/f Errors.'' *NeuroImage* 15, 217–32.

Flandrin, P. (1992). ''Wavelet Analysis and Synthesis of Fractional Brownian Motion.'' *IEEE Transactions on Information Theory* 38, 910–17.

Getis, A., and D. A. Griffith. (2002). ''Comparative Spatial Filtering in Regression Analysis.'' *Geographical Analysis* 34, 130–40.

Griffith, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering*. Berlin: Springer.

Griffith, D. A., and P. R. Peres-Neto. (2006). ''Spatial Modeling in Ecology: The Flexibility of Eigenfunction Spatial Analyses in Exploiting Relative Location Information.'' *Ecology* 87, 2603–13.

Haining, R., D. A. Griffith, and R. Bennet. (1983). ''Simulation Two-Dimensional Autocorrelated Surfaces.'' *Geographical Analysis* 15, 247–55.

Heagerty, P. J., and S. R. Lele. (1998). ''A Composite Likelihood Approach to Binary Spatial Data.'' *Journal of the American Statistical Association* 93, 1099–111.

Huffer, F., and H. Wu. (1998). ''Markov chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species.'' *Biometrics* 54, 509–24.

Johnstone, I. M., and B. W. Silverman. (1997). ''Wavelet Threshold Estimators for Data with Correlated Noise.'' *Journal of the Royal Statistical Society* 59, 319–51.

Kaiser, M., and N. Cressie. (1997). ''Modeling Poisson Variables with Positive Spatial Dependence.'' *Statistics and Probability Letters* 35, 423–32.

Keitt, T. H. (2007). ''On the Quantification of Local Variation in Biodiversity Scaling Using Wavelets.'' In *Scaling Biodiversity*, 168–80, edited by D. Storch, P. A. Marquet, and J. H. Brown. Cambridge, UK: Cambridge University Press.

Keitt, T. H., and D. L. Urban. (2005). ''Scale-specific Inference Using Wavelets.'' *Ecology* 86, 2497–504.

Legendre, P. (1993). ''Spatial Autocorrelation—Trouble or New Paradigm.'' *Ecology* 74, 1659–73.

Legendre, P., M. R. T. Dale, M. J. Fortin, J. Gurevitch, M. Hohn, and D. Myers. (2002). ''The Consequences of Spatial Structure for the Design and Analysis of Ecological Field Surveys.'' *Ecography* 25, 601–15.

Liang, K. Y., and S. L. Zeger. (1986). ''Longitudinal Data Analysis Using Generalized Linear Models.'' *Biometrika* 73, 13–22.

Lichstein, J. W., T. R. Simons, S. A. Shriner, and K. E. Franzreb. (2002). ''Spatial Autocorrelation and Autoregressive Models in Ecology.'' *Ecological Monographs* 72, 445–63.

McCullagh, P., and J. A. Nelder. (1989). *Generalized Linear Models*, 3rd ed. London: Chapman & Hall.

Meyer, F. G. (2003). ''Wavelet Based Estimation of a Semi Parametric Generalized Linear Model of fMRI Time-Series.'' *IEEE Transactions on Medical Imaging* 22, 315–22.

Müller, K., G. Lohmann, S. Zysset, and D. Y. von Cramon. (2003). ''Wavelet Statistics of Functional MRI Data and the General Linear Model.'' *Journal of Magnetic Resonance Imaging* 17, 20–30.

Myers, R. H., D. C. Montgomery, and G. G. Vining. (2002). *Generalized Linear Models*. New York: Wiley.

Nason, G. P., and B. W. Silverman. (1995). ''The Stationary Wavelet Transform and Some Statistical Applications.'' In *Wavelets and Statistics*, 281–99, edited by A. Antoniadis and G. Oppenheim. New York: Springer.

Percival, D. B., and A. T. Walden. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge, UK: Cambridge University Press.

R Development Core Team. (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ross, S. M. (1997). *Simulation*, 2nd ed. San Diego, CA: Academic Press.

Shumway, R. H., and D. S. Stoffer. (2000). *Time Series Analysis and Its Applications*. New York: Springer, Springer Texts in Statistics.

Statistisches Bundesamt. (1997). *Daten zur Bodenbedeckung für die Bundesrepublik Deutschland 1:100.000*. Wiesbaden, Germany: Statistisches Bundesamt.

Whitcher, B. (2005). ''waveslim: Basic Wavelet Routines for One-, Two- and Three-Dimensional Signal Processing.'' R package version 1.5.

Xiangcheng, M., R. Haibao, O. Zisheng, W. Wie, and M. Keping. (2005). ''The Use of the Mexican Hat and the Morlet Wavelets for Detection of Ecological Patterns.'' *Plant Ecology* 179, 1–19.

Yan, J. (2002). ''geepack: Yet Another Package for Generalized Estimating Equations.'' *R News* 2(3), 12–14.

Yan, J., and J. Fine. (2004). ''Estimating Equations for Association Structures.'' *Statistics in Medicine* 23, 859–74.

Zeger, S. L., and K. Y. Liang. (1986). ''Longitudinal Data Analysis for Discrete and Continuous Outcomes.'' *Biometrics* 42, 121–30.