Dreesman, Tutz:

Nonstationary conditional models for spatial data based on varying coefficients

Sonderforschungsbereich 386, Paper 150 (1999)

Online unter: http://epub.ub.uni-muenchen.de/

Nonstationary conditional models for spatial data based on varying coefficients

Johannes Dreesman[1] and Gerhard Tutz[2]

[1]NLGA, Department of Epidemiology, Roesebeckstr. 4-6, 30449 Hannover.
[2]LMU München, Institute of Statistics, Akademiestr. 1 /I, 80799 München.

Abstract: The analysis of spatial data by means of Markov random fields usually is based on strict stationarity assumptions. Although these assumptions rarely hold, they are necessary in order to obtain parameter estimates. For Gaussian data the necessary assumptions are mean- and covariance stationarity. While simple techniques are available to deal with violations of mean stationarity, the same is not true for covariance stationarity. In order to handle mean nonstationarity as well as covariance nonstationarity, we propose the modelling by spatially varying coefficients. This aproach not only yields more appropriate models for nonstationary data but also can be used to detect violations of the stationarity assumptions. The method is illustrated by use of the well known wheat yield data.

Keywords: Markov Random Fields, Local Likelihood, Pseudolikelihood, Wheat Yield Data

1 Introduction

In several statistical application areas like image analysis and the analysis of agricultural trials, data are typically collected on a regular lattice of measurement points. The spatial dependence structure of such data can be modelled in terms of a Markov random field. Though being stochastically more complex, Markov random fields may be seen as a multidimensional extension of common Markov chains. The Markov random field model for Gaussian data is reviewed in section 2 and the estimation of the model's parameters is treated in section 3.
One of the best analysed lattice data sets are the wheat-yields of Mercer and Hall (1911). Figure 1 gives an impression of these data, which were obtained in an uniformity trial, i.e. the same treatment was applied to each of the 20×25 plots. In this figure the wheat-yields are visualized by grey values varying from white, denoting low yields, to black, denoting high yields.
In many seminal contributions to the theory of Gaussian Markov random fields these data have been used as the exemplary application, for example by Whittle (1954) and Besag (1974). However, both authors had to realize that their models did not fit very well if the criterion was the comparison between theoretical spatial covariances and sample covariances. Since Mercer & Hall (1911) had stated the adequacy of the Gaussian distribution, Besag (1974) suspected that the dissatisfying fit is caused by nonstationarity, which has been detected by Patankar (1954). Künsch (1985) and Cressie (1993) took

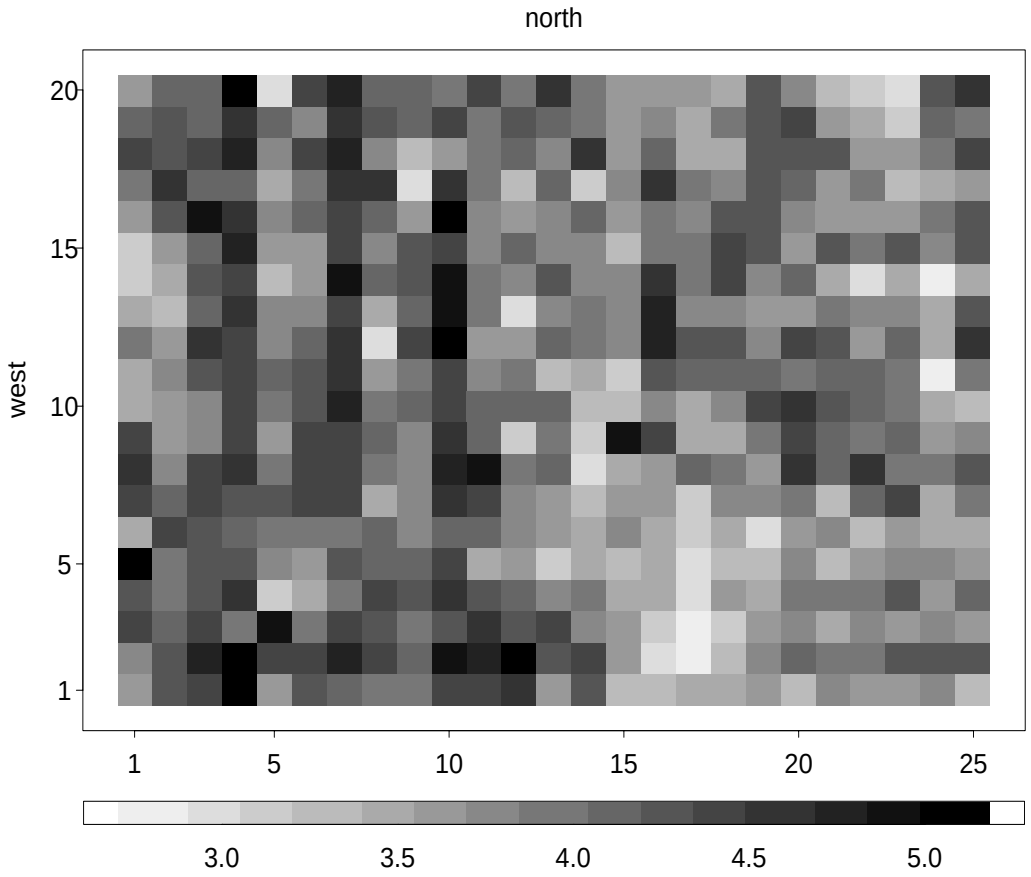

Figure 1: Wheat-yields in lbs on 20×25 lattice.

up the hypothesis of nonstationarity and proposed several techniques for detrending the data before fitting Markov random fields. A brief description will be given in section 4.
The objective of this paper is to propose a model which allows for nonstationarity and shows if and where violations are to be suspected. This approach, which is based on varying coefficients as introduced by Hastie and Tibshirani (1993), is developed in section 5 and 6. In section 7 the varying coefficient model is applied to the wheat-yield data. Finally, in section 8 the fit is compared to the results obtained from the stationary model.

2 Conditional Gaussian models

Let $L = \{(i,j)\}_{i=1,\ldots,I,j=1,\ldots,J}$ denote a regular lattice and $\mathbf{X} = \{X_{ij}\}_{(i,j)\in L}$ be a corresponding set of random variables. Let the dependence structure of the elements of $\mathbf{X}$ be defined by means of their full conditional distributions. A simple model, which specifies the conditional distribution of each X_{ij}, given

all observations x_{rs} except at site (i,j), is given by

$$X_{ij}|\{x_{rs}\}_{(r,s)\neq(i,j)} \sim N\left(\eta_{ij}, \tau^2\right), \tag{1}$$

$$\begin{aligned}\eta_{ij} &= \mu + (x_{i-1,j} + x_{i+1,j} - 2\mu)\ \beta_1 + (x_{i,j-1} + x_{i,j+1} - 2\mu)\ \beta_2 \\ &= \beta_0 + (x_{i-1,j} + x_{i+1,j})\ \beta_1 + (x_{i,j-1} + x_{i,j+1})\ \beta_2. \end{aligned} \tag{2}$$

Furthermore the restriction $|\beta_1| + |\beta_2| < 1/2$ is imposed on the coefficients in order to obtain a proper joint model. Equation (1) and (2) represent a conditional Gaussian Markov random field with respect to the first order neighbourhood system, where each internal site (i,j) has the neighbours $\{(i-1,j),(i+1,j),(i,j-1),(i,j+1)\}$. In the same way higher order models may be specified, where the neighbourhood includes more neighbours and consequently a higher number of coefficients is required. Models of this conditional Gaussian type, which will be referred to as CG-models, yield a joint Gaussian distribution. Switching to vector notation, the joint distribution can be given in the form

$$\mathbf{x} := (X_{11}, \ldots, X_{IJ})' \sim N\left(\boldsymbol{\mu}, (\mathbf{I} - \mathbf{B})^{-1}\mathbf{T}\right), \tag{3}$$

where $\mathbf{x}$ and the mean vector $\boldsymbol{\mu} := \mu\mathbf{1}$ are of length $I \cdot J$ and the matrices have dimension $I \cdot J \times I \cdot J$. $\mathbf{T}$ is the diagonal matrix $diag(\tau^2)$ and $\mathbf{I}$ denotes the identity matrix. The matrix $\mathbf{B}$ is determined by the parameters $\beta_1, \ldots, \beta_q$ in a simple way. Let $\mathbf{B}_{\cdot,ij}$ be the column of $\mathbf{B}$ which corresponds to x_{ij}, then in the first order setting $\mathbf{B}_{\cdot,ij}$ is such that $\mathbf{x}'\mathbf{B}_{\cdot,ij}$ is equal to $(x_{i-1,j} + x_{i+1,j})\beta_1 + (x_{i,j-1} + x_{i,j+1})\beta_2$ (see Besag, 1974, or Cressie, 1993, Section 6.4).
It should be noted that in (3) the residuals are not independent. This may easily be seen by considering the residual vector $\boldsymbol{\epsilon} := (\mathbf{I} - \mathbf{B})(\mathbf{x} - \boldsymbol{\mu})$. One obtains the covariance matrix $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \mathbf{T}(\mathbf{I} - \mathbf{B}')$, which obviously is nondiagonal since $\mathbf{B}$ is nondiagonal while $\mathbf{T}$ is diagonal.

3 Parameter estimation

Maximum-likelihood estimation of the coefficients of CG-models is not straightforward because calculation of the likelihood leads to computational problems. The problems arise at different stages depending on the formulation that is used. If the likelihood is computed directly from the joint density, the problems are due to the normalizing term, which contains the determinant of the covariance matrix. As seen in section 2, this matrix is nondiagonal and its dimension is determined by the number of measurement points $I \cdot J$, which is 500 for the wheat-yield data and can amount to several thousands in image analysis problems. Since the computational effort is of the order $(I \cdot J)^3$, problems of this dimension soon become intractable. If on the other hand the likelihood is derived from the full conditionals, every variable X_{ij} occurs several times, once as response variable and four times as part of the condition. The corresponding likelihood contributions are not independent and hence the calculation of the joint likelihood is rather troublesome.

Therefore Besag (1974) proposed a method called coding. The basic idea is to divide up the lattice into disjoint sublattices such that there are no neighbours within the same sublattice. The next step is to select one sublattice and take only likelihood contributions corresponding to response variables from this sublattice. Then likelihood contributions are independent and from the factorization one may derive an estimator with well known theoretical properties. The number of coding sets and the fraction of data contained in each one depend on the order of the neighbourhood considered. With a first order neighbourhood one obtains at least two coding sets. Unfortunately, there are as many coding estimators as there are coding sets, but so far no method is known how to reasonably combine these estimators. Therefore one has to decide for one of them in order to obtain the desired theoretical properties.
More efficient but less appealing is the pseudolikelihood method (Besag 1975), where all likelihood contributions are factorized as if they were independent. For the coefficient vector $\boldsymbol{\beta}' := (\beta_0, \beta_1, \ldots, \beta_q)$ of the CG-model the pseudolikelihood leads to the simple least squares estimator

$$\hat{\boldsymbol{\beta}}_{PL} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}. \tag{4}$$

Here $\mathbf{Z} := (\mathbf{1}, \mathbf{z}_1, \ldots, \mathbf{z}_q)$ is a design matrix with $\mathbf{z}_s$ corresponding to β_s for $s = 1, \ldots, q$. Let $\mathbf{z}_s$ be indexed in the same way as $\mathbf{x}$, then for example $\mathbf{z}_1 = (\ldots, z_{1,ij}, \ldots)' = (\ldots, x_{i-1,j} + x_{i+1,j}, \ldots)'$ since the two neighbours $x_{i-1,j}$ and $x_{i+1,j}$ are connected to x_{ij} through β_1. The coding estimator for the coding set $C \subset L$ can be given in the same simple way by just replacing $\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_q$ by the subvectors which contain only those components $x_{ij}, z_{1,ij}, \ldots, z_{q,ij}$ where $(i,j) \in C$ holds.
For the wheat-yield data the fit of a first order model yields the pseudolikelihood estimates

$$\beta_0 = 0.05, \qquad \beta_1 = 0.142, \qquad \beta_2 = 0.343, \tag{5}$$

indicating that the dependence within rows (west-east) is weaker than within columns (north-south). The coding estimates are quite similar.
Besides coding and pseudolikelihood, alternative methods have been proposed for the analysis of CG-models. In use are also Markov chain Monte Carlo methods (Younes, 1988, and Geyer, 1991) or numerical approximations to the likelihood (see for example Besag and Moran, 1975). Since all these methods require a high amount of computation, they are not appropriate for repeated fitting and will not be considered here further.

4 Stationarity as a problem in real data sets

Parameter estimation for CG-models requires the field to be stationary. Usually second order stationarity is assumed, i.e. the mean and the covariance function are not allowed to depend on the position on the lattice. One postulates

$$E(X_{ij}) = \mu \qquad \text{for all } (i,j) \in L, \tag{6}$$

$$cov(X_{ij}, X_{i+u,j+v}) = \gamma(u,v) \qquad \text{for all } (i,j), (i+u, j+v) \in L. \tag{7}$$

Thus, the covariance of two variables is assumed to depend only on their distance vector. For the models considered in Section 2, these assumptions are already fulfilled since the βs and τ^2 are the same for the full conditionals at any site.
In real data sets stationarity most often is a doubtful assumption. Although for the Mercer & Hall wheat-yield data the first order dependence structure seems not unreasonable, Cressie (1993) concludes that the constant term β_0 in (2) actually should be a varying coefficient $\beta_0(i, j)$, allowing for a non-stationary field. This is illustrated by Figure 2. The top panel shows the mean within each row and the bottom panel shows the mean within each column. Rows as well as columns are identified by an index. The index for rows ranging from 1 to 20 indicates steps from south to north and the index for columns ranging from 1 to 25 indicates steps from west to east (see also Figure 1). In particular the columns show rather strongly varying means with a decreasing tendency towards the eastern plots. In order to correct the mean instationarity across columns, Künsch (1985) subtracted the means of the columns and fitted a first order CG-model to the corrected data. Cressie (1993, Section 4.5) proposed a similar technique, called median polishing, where both, the rows and the columns, are corrected for their medians. This is achieved by alternately applying a correction procedure to the rows and the columns until a stop criterion is fulfilled. As Cressie remarks the large scale variation (spatial trend) must be taken into account before the parameters of the small scale variation (spatial dependence) can be interpreted.

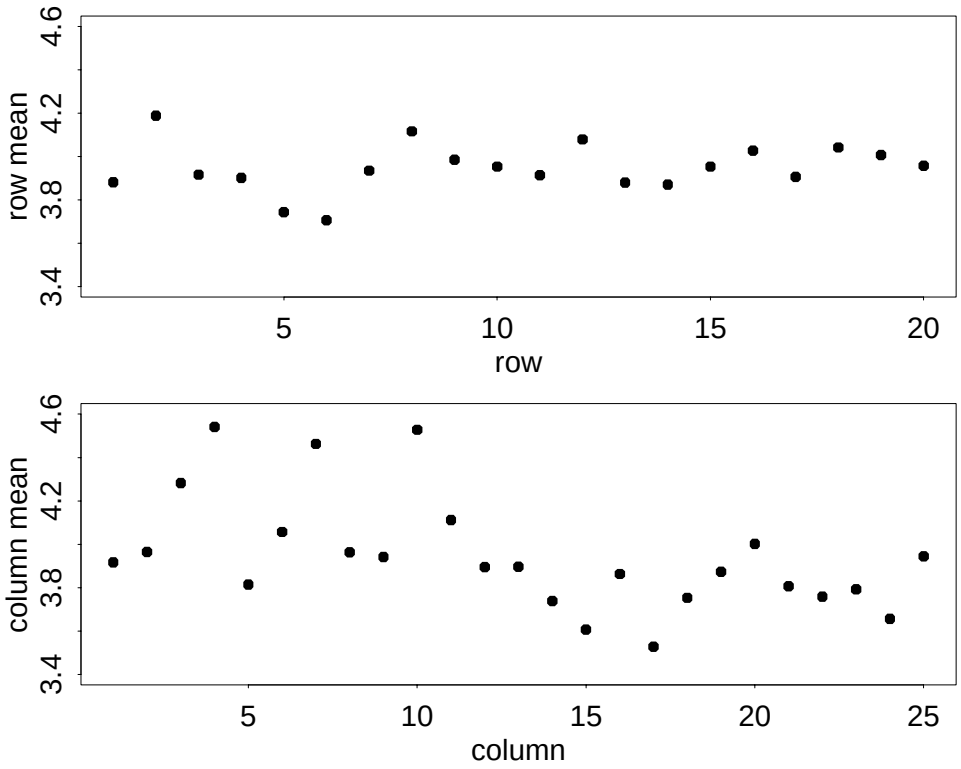

Figure 2: Means within rows (above) and within columns (below).

However, for the wheat-yield data additional problems occur since the covariance stationarity assumption also seems to be violated. This may be seen from Figure 3, where the serial correlations within rows and columns are given. The smoothed curves show that there is variation in both dimensions which cannot be neglected. Künsch's (1985) correction yields slightly lower serial correlation for rows, but due to his correction method the correlation structure within columns remains unchanged. Moreover, Künsch's investigations have shown that the covariance structure, which he expressed in terms of spectra, differs substantially between the left (west) and the right (east) half of the field. These differences become already obvious by visual inspection of the data in Figure 1. From this figure one gets the impression that neighbours in the columns are stronger connected in the west half of the field than in the east half, while neighbours in the rows seem to be strongest connected in the east half, especially in the south.

5 Varying coefficient models

Since methods like subtracting row means only allow to remove mean nonstationarity, different methods have to be developed to account for covariance nonstationarity. Hastie & Tibshirani (1993) considered a rather general model, which includes several smoothing approaches like generalized additive models and semiparametric models as special cases. In a simple regression context the model for Gaussian data is given by

$$Y = \beta_0(U) + X_1\beta_1(U) + \ldots + X_q\beta_q(U) + \epsilon, \tag{8}$$

where $\epsilon \sim N(0, \sigma^2)$. The essential point is that the parameters and therefore the strength of the effects of the covariates is modified by the external variable U, the so called effect modifier. The varying coefficients $\beta_s(U)$ in (8) are assumed to be smooth functions which have to be estimated nonparametrically. It is immediately seen that $\beta_s(U) = \beta_s, s = 0, 1, \ldots, q$, yields a parametric model, whereas $q = 0$ yields a simple smooth model.
An obvious way to handle nonstationarity in CG-models is by considering a varying coefficient approach of the form

$$X_{ij}|\{x_{rs}\}_{(r,s)\neq(i,j)} \sim N\left(\eta_{ij}, \tau^2(i,j)\right), \tag{9}$$

$$\eta_{ij} = \beta_0(i,j) + (x_{i-1,j} + x_{i+1,j})\ \beta_1(i,j) + (x_{i,j-1} + x_{i,j+1})\ \beta_2(i,j), \tag{10}$$

where parameters may vary smoothly across the field. In this context smoothness means that for sites that are close to each other the parameter values have to be more similar than they have for sites that are far apart. Since the parameters are allowed to depend on the location, stationarity is no longer assumed.
The crucial point is how to estimate the parameters in (10). Hastie & Tibshirani (1993) suggest a penalized least squares approach. This approach rises problems of invariance. If the covariates are transformed, the penalizing term changes its meaning (see also the discussion of the paper of Hastie &

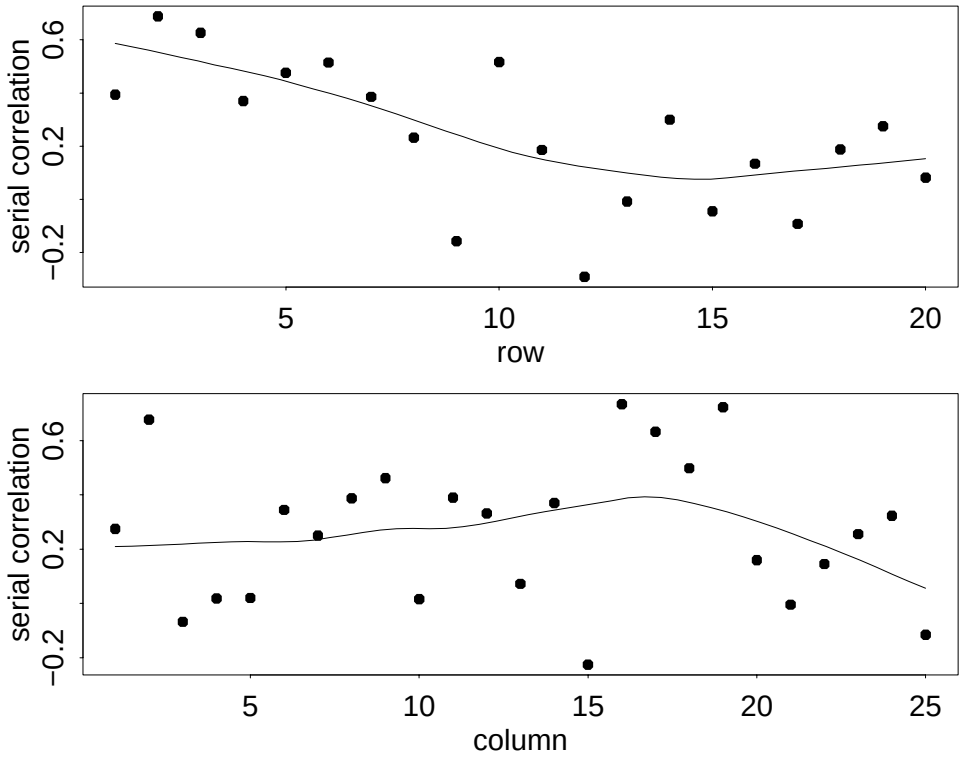

Figure 3: Serial correlations within rows (above) and within columns (below).

Tibshirani). As an alternative estimation concept in the following a version of local likelihood estimation is considered.

6 Local estimation techniques

Local likelihood approaches have been introduced in the context of regression smoothing by Tibshirani & Hastie (1987) and further developed by several authors (e.g. Fan & Gijbels, 1996). The extension to varying coefficients has been considered e.g. by Tutz & Kauermann (1997) as follows. A set of values of the effect modifier is chosen as target points and at each target point an estimate for the varying coefficient is calculated by maximization of the local likelihood. The local likelihood is constructed by attributing an individual weight to the likelihood contribution of each observation. This weight depends on the distance between the target point and the value of the effect modifier at the observation point. With increasing distance to the target point observations receive lower weights. The estimates at all target points yield more or less smooth curves or surfaces, respectively, over the space of the effect modifier.
In the present context the calculation of the localized likelihood bears the same computational problems as the calculation of the total likelihood. Therefore we will consider localization of the pseudolikelihood estimator (PL). The underlying principle is the same as for local likelihood estimation approaches.

The observations around the target point (i, j) are weighted down by use of the weight function

$$w_\lambda((i,j),(r,s)) = cK\left(\frac{d((i,j),(r,s))}{\lambda}\right), \tag{11}$$

where d is a distance function, K is an unimodal, symmetric kernel function and λ is a smoothing parameter, denoted as bandwidth, which determines the amount of smoothing. The normalizing constant c is chosen by $c = K(0)^{-1}$, yielding the weight "1" at the target point and lower weights in the neighbourhood. The kernel function chosen here is the tricube-function of Cleveland (1979), where $d(.,.)$ denotes the Euclidean distance of the two arguments. Thus the weights have the form

$$w_\lambda((i,j),(r,s)) = \begin{cases} \left(1 - \left(\frac{d((i,j),(r,s))}{\lambda}\right)^3\right)^3, & \text{if } 0 \le d((i,j),(r,s)) \le \lambda; \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Based on the weights (12) the local pseudolikelihood estimator is given by

$$\hat{\beta}(i,j) = (\mathbf{Z}'\mathbf{W}_{ij}^{\lambda}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{W}_{ij}^{\lambda}\mathbf{x}), \tag{13}$$

where $\mathbf{W}_{ij}^{\lambda} = diag(w_\lambda((i,j),(1,1)), w_\lambda((i,j),(1,2)), \ldots, w_\lambda((i,j),(I,J)))$. Since the local coding estimator is constructed in a similar way, we give formulae only for the local pseudolikelihood estimator.

6.1 Choice of the smoothing parameter

The choice of the bandwidth is crucial because it determines the smoothness of the curve and therefore the trade-off between the bias and the variance of the estimates. Increasing λ decreases the variance, but tends to increase the bias because the estimate at target point (i, j) might involve more terms with β-values different from $\beta(i, j)$. If λ tends to ∞, all observations are receiving equal weights and in the limit the fixed parameter model is assumed to hold. On the other hand decreasing λ leads to an increasing variance and a decreasing bias. If λ tends to zero, the number of observations involved in an estimate at any point becomes small and the estimated curves are quite jagged.

For purely exploratory purposes one may simply try several smoothing parameters and pick one by eye-sight. In the same spirit one may consider a family of smoothed curves by allowing bandwidth from a grid (see Marron & Chung, 1997). However, often it is preferable to have at least a reasonable proposal from a data driven procedure, which automatically chooses an adequate smoothing parameter. An often used criterion is cross validation. With the focus on quadratic loss one chooses the parameter λ which minimizes

$$CV(\lambda) = \frac{1}{n}\sum_{s=1}^{n}(y_s - \hat{\mu}_s^{-s,\lambda})^2, \tag{14}$$

where $\hat{\mu}_s^{-s,\lambda}$ is the smoothed estimate of the expected response $E(y_s)$. The superscript $-s$ denotes that the estimate is computed without observation y_s. In the case of correlated residuals the simple cross validation criterion $CV(\lambda)$ is no longer appropriate. It is known that in spline smoothing of independent data the estimate of the vector $(\hat{\mu}_1^{-s,\lambda}, \ldots, \hat{\mu}_n^{-s,\lambda})$ may be computed in the same way as the estimate of the "full data" vector $(\hat{\mu}_1^{\lambda}, \ldots, \hat{\mu}_n^{\lambda})$. One simply has to substitute the uninformative observation $\hat{\mu}_s^{-s,\lambda}$ for the original observation y_s. For data with correlated residuals, however, $\hat{\mu}_s^{-s,\lambda}$ is no longer uninformative. Van der Linde (1984) derives the uninformative substitute to be given by $\hat{\mu}_s^{-s,\lambda} + \hat{\epsilon}_s^{-s,\lambda}$, where $\hat{\epsilon}_s^{-s,\lambda}$ is the estimated residual. In the corresponding cross-validation criterion

$$CV_{dep}(\lambda) = \frac{1}{n} \sum_{s=1}^{n} (y_s - (\hat{\mu}_s^{-s,\lambda} + \hat{\epsilon}_s^{-s,\lambda}))^2 \tag{15}$$

only the difference between the observation and the estimate $\hat{\mu}_s^{-s,\lambda} + \hat{\epsilon}_s^{-s,\lambda}$ is evaluated, where the estimate $\hat{\epsilon}_s^{-s,\lambda}$ contains the information which is due to the residuals' correlation structure. In the present case the latter criterion has the form

$$CV_{dep}(\lambda) = \frac{1}{I \cdot J} \sum_{i,j} (X_{ij} - (\hat{\eta}_{ij}^{-(i,j),\lambda} + \hat{\epsilon}_{ij}^{-(i,j),\lambda}))^2, \tag{16}$$

where $\hat{\eta}_{ij}^{-(i,j),\lambda}$ is the predictor resulting from pseudo likelihood estimation without observation (i,j). The estimated residual $\hat{\epsilon}_{ij}^{-(i,j),\lambda}$ is given by

$$\hat{\epsilon}_{ij}^{-(i,j),\lambda} = \Sigma_{ij,-(i,j)} (\Sigma_{-(i,j),-(i,j)})^{-1} (\mathbf{x}_{-(i,j)} - \hat{\boldsymbol{\eta}}_{-(i,j)}^{-(i,j),\lambda}). \tag{17}$$

Here $\mathbf{x}_{-(i,j)} = (x_{11}, x_{12}, \ldots)$ is the vector of observations without x_{ij}, and $\hat{\boldsymbol{\eta}}_{-(i,j)}^{-(i,j),\lambda}$ is the corresponding vector of predictors without $\hat{\eta}_{(i,j)}^{-(i,j),\lambda}$, as indicated by the subscript. Again the superscript of $\hat{\boldsymbol{\eta}}_{-(i,j)}^{-(i,j),\lambda}$ gives details of the estimation procedure, in particular, that the observation at site (i,j) was omitted when estimating this vector locally with bandwidth λ. $\Sigma_{ij,-(i,j)}$ is the vector $(\sigma_{ij,rs})_{(r,s)\neq(i,j)}$, extracted from the covariance matrix and $\Sigma_{-(i,j),-(i,j)}$ denotes the covariance matrices submatrix $(\sigma_{rs,tu})_{(r,s),(t,u)\neq(i,j)}$.
When using cross validation for bandwidth selection in the case of Markov random fields one has to account for the correlation in the residuals. Therefore we choose the λ-value, which minimizes (16). The specific algorithm consists of two nested loops.

- The outer loop is over a set of values λ^* from the space of the smoothing parameter λ.

  - In the inner loop each point (i,j) of the lattice L is chosen as target point.

1. At the current target point (i,j) we estimate the coefficients locally using the current bandwidth λ^*. The contribution to the pseudolikelihood-estimator, where x_{ij} is the response, is omitted. The estimate is denoted as $\widehat{\boldsymbol{\beta}}(i,j)^{-(i,j),\lambda^*}$.
2. With this estimate, we calculate $\widehat{\eta}_{ij}^{-(i,j),\lambda^*}$, which is the predictor for x_{ij}.
3. With the same estimate, we calculate the vector of predictors $\widehat{\boldsymbol{\eta}}_{-(i,j)}^{-(i,j),\lambda^*}$ and the vector of residuals $\widehat{\boldsymbol{\epsilon}}_{-(i,j)}^{-(i,j),\lambda^*} := \mathbf{x}_{-(i,j)} - \widehat{\boldsymbol{\eta}}_{-(i,j)}^{-(i,j),\lambda^*}$. These residuals together with the coefficients $\widehat{\boldsymbol{\beta}}(i,j)^{-(i,j),\lambda^*}$ allow us to predict the residual at the target point $\widehat{\epsilon}_{ij}^{-(i,j),\lambda^*}$ using (17).
4. Finally we compare the sum of the predicted value $\widehat{\eta}_{ij}^{-(i,j),\lambda^*}$ and the predicted residual $\widehat{\epsilon}_{ij}^{-(i,j),\lambda^*}$ with the observed value x_{ij} and calculate the remaining error.

– After the inner loop has finished, we compute $CV_{dep}(\lambda^*)$, the mean of the squared remaining errors.

- After the outer loop has finished, CV_{dep} is interpolated between the λ^*-values and presented as function of λ.

For the wheat-yield data the results from cross validation for the local pseudolikelihood estimator are given in Figure 4. The curve shows an unique minimum around $\lambda = 20$.
Alternative concepts for bandwidth selection include plug-in estimates, which have been investigated e.g. by Härdle, Hall & Marron (1988) and Ruppert, Steather & Wand (1995). A modification for correlated observations has been given by Opsomer (1995). There has been some discussion on the optimality of criteria of bandwidth selection mostly in terms of asymptotic optimality, in the course of which e.g. Loader (1995) gives several arguments for using cross validation.

6.2 Bias correction

It has already been mentioned how local estimators lead to biased estimates. Under weak regularity assumptions Kauermann & Tutz (1995) developed an approximation for the bias of local likelihood estimates and proposed to use the approximation for an additive bias correction. The efficiency of this bias correction has been shown in Kauermann, Müller & Carroll (1997).
Let $\boldsymbol{\beta}(u^*)$ be the coefficient vector at the target point u^* and $\{\boldsymbol{\beta}\}$ the set of coefficient vectors at the target points $u_1, \ldots, u_N$, then an approximation

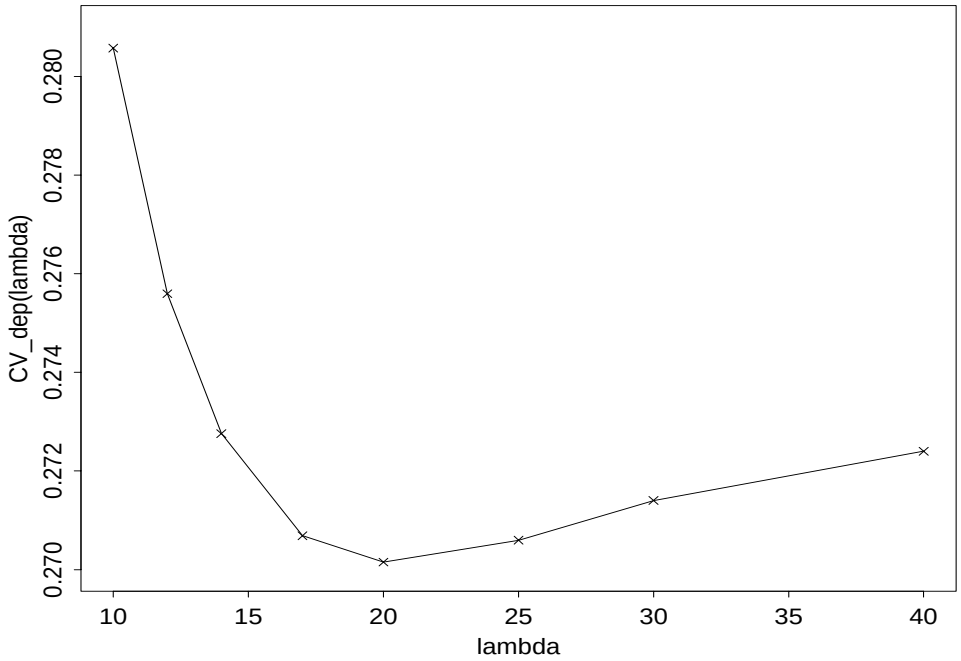

Figure 4: Mean squared error cross validation for dependent observations.

for the bias of the local likelihood estimator in model (8) is given by:

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}^{\lambda}(u^*) - \boldsymbol{\beta}(u^*)) &\approx b_\lambda\left(\boldsymbol{\beta}(u^*), \{\boldsymbol{\beta}\}\right) \\ &= \left(\sum_{n=1}^{N} w_\lambda(u^*, u_n)\frac{1}{\sigma^2}(1, \mathbf{x}_n')'(1, \mathbf{x}_n')\right)^{-1} \\ &\quad \sum_{n=1}^{N} w_\lambda(u^*, u_n)(1, \mathbf{x}_n')'\frac{1}{\sigma^2}\left((1, \mathbf{x}_n')\boldsymbol{\beta}(u_n) - (1, \mathbf{x}_n')\boldsymbol{\beta}(u^*)\right). \end{aligned}$$

The bias $b_\lambda\left(\boldsymbol{\beta}(u), \{\boldsymbol{\beta}\}\right)$ can be estimated by plugging in the estimates instead of the true coefficients. Then $b_\lambda\left(\widehat{\boldsymbol{\beta}}(u), \{\widehat{\boldsymbol{\beta}}\}\right)$ may be subtracted from the estimated curves of the single coefficients directly in order to obtain a bias corrected estimate.

7 Wheat-yield data revisited

In Section 4 it has already been remarked that for the Mercer & Hall wheat-yield data stationarity seems not to be fulfilled. Therefore a model with varying coefficients has been fitted by means of local pseudolikelihood. The bandwidth $\lambda = 20$ was chosen according to the cross validation results (see section 6.1) and in addition bias correction has been applied.
From Figure 5 it is seen that the estimate $\widehat{\beta}_1(i, j)$, which describes the dependence in row-direction, is strongly increasing when proceeding from the north-west corner (top left) to the south-east corner (bottom right). The

variation from north to south may already be seen from Figure 3, where the top panel shows that the serial correlation is higher within rows with low index, i.e. rows in the south. However, the methodology underlying Figure 3 restricts consideration to the north-south or east-west direction. The diagonal effect seen in Figure 5 may not be seen from investigating correlations within columns or rows.
From Figure 6 it becomes obvious that the estimate $\widehat{\beta}_2(i,j)$, which represents the dependence in column-direction, is decreasing essentially from west to east with particularly high values in the south-west corner. This effect could not be seen from the serial correlation plots in Figure 3. It is primarily caused by the strong shifts between neighbouring column means in the west, which is shown in the bottom panel of Figure 2. The main common feature of Figure 5 and 6 is the variation from west to east. This corresponds to the findings of Künsch (1985) that the covariance structure in the east and the west half differ substantially (cf. section 4).
The estimate $\widehat{\beta}_0(i,j)$ is presented in Figure 7. It is decreasing from north to south. The variation is hard to interpret, because the intercept term is influenced by the neighbourhood coefficients $\widehat{\beta}_1(i,j)$ and $\widehat{\beta}_2(i,j)$ and therefore can not be considered as a separate feature.

8 Comparison of nonstationary model versus stationary model

The varying coefficient approach has already led to further insight into the spatial structure of the wheat-yield data. But in order to confirm the extracted features as well as for inferential purposes like prediction, one may want to decide between models and select the most appropriate. Here, one has to decide between a parametric model and a semiparametric model. This is currently a field of intensive research and we focus on one approach which is based on confidence intervals around the nonparametric fit. The parametric model is considered inappropriate if the estimates of coefficients are not covered by these intervals (see e.g. Bowman and Young, 1996). In order to estimate these intervals, bootstrap techniques may be applied. Here, it is advisable to generate the bootstrap sample by Markov-chain-Monte-Carlo-methodology. Let the parametric model (1) and (2) and the pseudolikelihood estimate (5) be the starting constituents. After generating the first realization, the field is updated componentwise, using the current values in the neighbourhood. Actually, instead of updating component by component, the complete coding sets were updated simultaneously. After every 10th update the current field was used to generate a new bootstrap estimate. Estimation was done by local pseudolikelihood with bandwidth $\lambda = 20$ and an additional bias correction procedure, as applied to the wheat-yield data before. Since the bootstrap realizations are not independent, the simulation procedure could rather be seen as a MCMC-algorithm, which is standard for simulating Markov random fields.

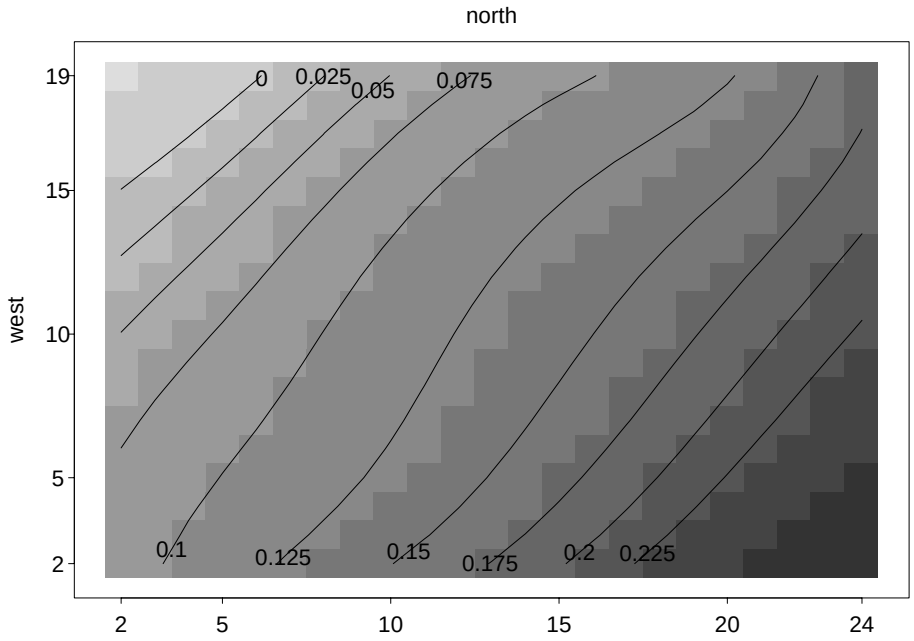

Figure 5: Spatial variation of $\widehat{\beta_1}$ estimated with the pseudolikelihood method and $\lambda = 20$.

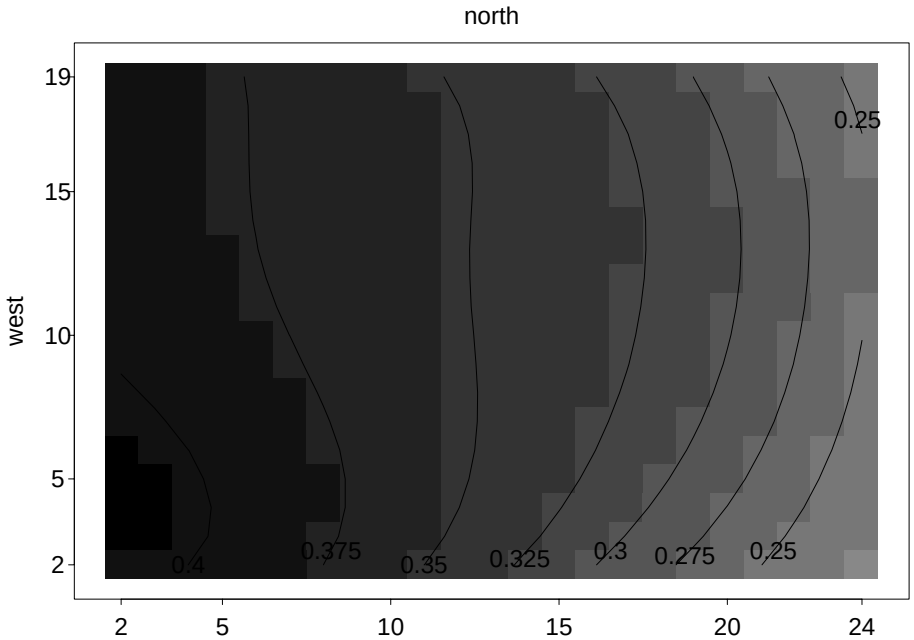

Figure 6: Spatial variation of $\widehat{\beta_2}$ estimated with the pseudolikelihood method and $\lambda = 20$.

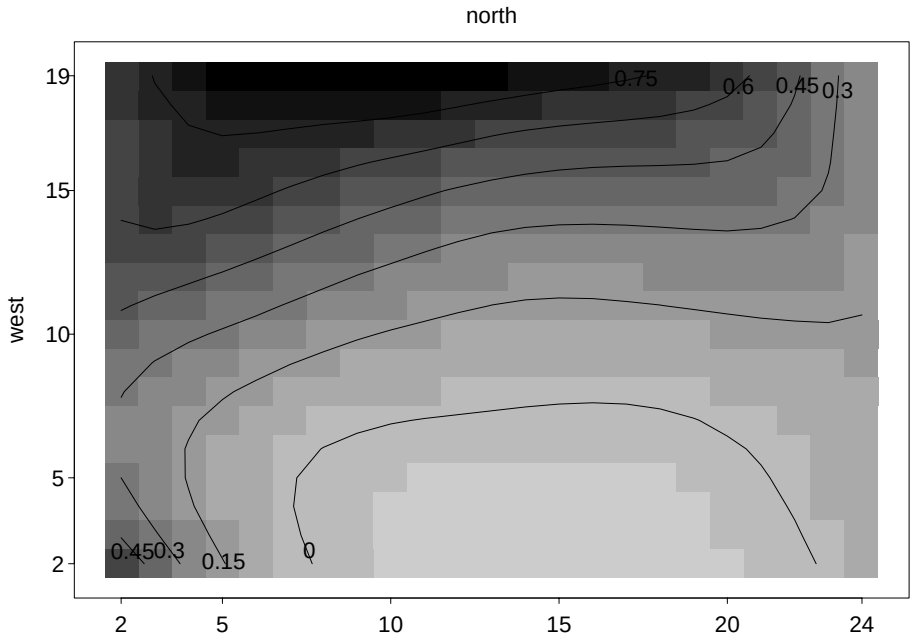

Figure 7: Spatial variation of $\widehat{\beta_0}$ estimated with the pseudolikelihood method and $\lambda = 20$.

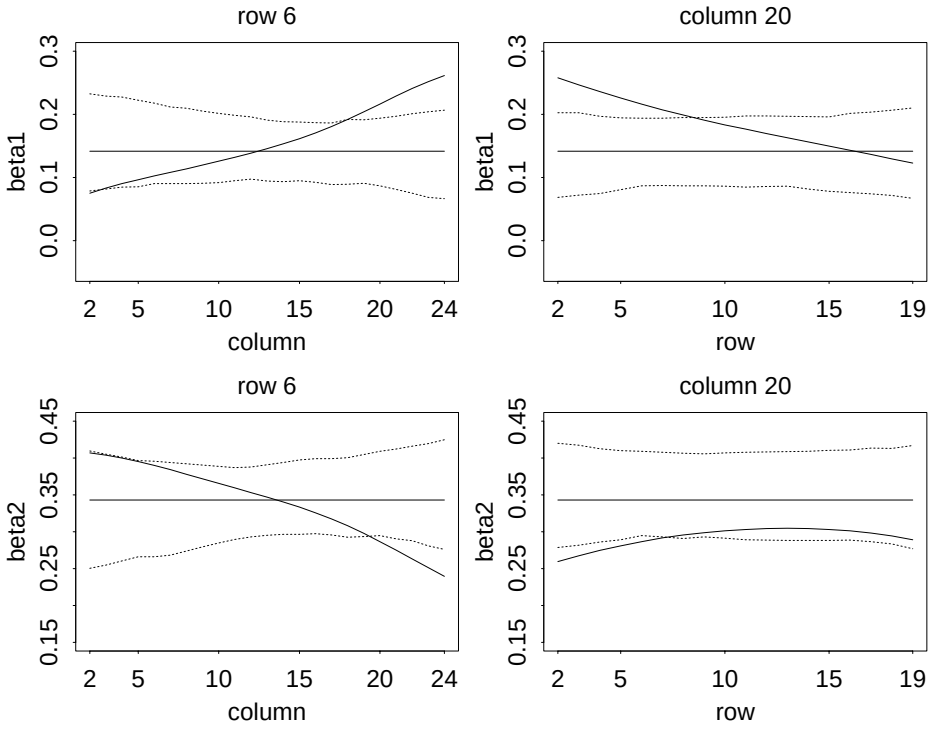

Figure 8: Bootstrap 90% confidence intervals around $\widehat{\beta_1}$ (top) and $\widehat{\beta_2}$ (bottom) and varying coefficients.

Since there is no way to present a set of two-dimensional surfaces in one figure, only part of the results are presented here. Figure 8 shows values of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ for several sections along rows and columns. The horizontal solid line shows the estimate of the stationary Gaussian model, while the solid curve gives the nonparametric estimates (compare Figure 6 and Figure 7). The dotted curves give pointwise 90% confidence intervals based on the bootstrap estimates.
It becomes obvious that the coverage of the nonparametric estimates by the confidence intervals is not satisfactory, indicating that the stationary model is unable to account for the underlying spatial variation. This is new strong support of the suspicion that the wheat-yield data are not stationary.

9 Concluding remarks

Varying coefficients are an useful tool to discover nonstationarities in spatial data. Compared with other approaches to nonstationary spatial data which require detrending prior to fitting a Markov random field, as suggested by Künsch (1985) and Cressie (1993), we see two main advantages. First, Markov random fields with spatially varying coefficients can take into account not only mean nonstationarity but also covariance nonstationarity. Second, the model is fitted without the arbitrariness which is inherent to detrending. The estimation procedure is easily implemented by incorporating weights into the fitting procedure. In the present paper it is mostly considered as an exploratory tool. Future research should include more formal tools to decide upon the appropriateness of the semiparametric model and the investigation of the extent to which parameters may vary and still yield a proper joint model. Moreover, more flexibility is desirable, especially concerning the specification of the weights. An adaptive bandwidth specification and a separate bandwidth for each direction should improve the fit further.
The area of spatial statistics has found considerable interest recently. When finishing the present paper, a special issue of The Statistician (1998, Part 3), which is devoted to spatial data and local statistics, came to our knowledge. In this issue Unwin & Unwin give a survey of recent developments and Brunsdon et al. considered geographically weighted regression. In contrast to Brunsdon et al. we consider a different class of models, namely the more structured case of conditional Gaussian models, and therefore the estimation procedures are quite different. Consequently, we use different ways of cross validation and differing bootstrap procedures. Nevertheless, the basic approach to develop instruments for the exploration of spatial structures by local modelling is the same as in Brunsdon et al. (1998).

References

Besag, J.E. (1974), "Spatial interaction and the statistical analysis of lattice systems," Journal of the Royal Statistical Society, B 36, 192–225.

Besag, J.E. (1974), "Statistical analysis of non-lattice data," The Statistician 24, 179–195.

Besag, J.E., and Moran, P.A.P. (1975), "On the estimation and testing of spatial interaction in Gaussian lattice processes," Biometrika 82, 733–746.

Bowman, A., and Young, S. (1996), "Graphical comparison of nonparametric curves," Applied Statistics 45, 83–98.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998), "Geographically weighted regression - modelling spatial non-stationarity," The Statistician 47, 431–443.

Cleveland, W.S. (1979), "Robust locally weighted regression and smoothing scatterplots," Journal of the American Statistical Association 74, 829–836.

Cressie, N. (1993), Statistics for Spatial Data (Revised Edition). New York: Wiley.

Fan, J., and Gijbels, I. (1996), Local polynomial modelling and its applications. London: Chapman & Hall.

Geyer, C.J. (1991), "Markov chain Monte Carlo maximum likelihood," in E.M. Keramidas (ed.), Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface, 156–163.

Härdle, W., Hall, P., and Marron, J.S. (1988), "How far are automatically chosen regression parameters from their optimum," Journal of the American Statistical Association 83, 86–99.

Hastie, T., and Tibshirani, R. (1993), "Varying-coefficient models," Journal of the Royal Statistical Society, B 55, 757–796.

Kauermann, G., Müller, M., and Carroll, R.J. (1997), "The efficiency of bias-corrected estimators for nonparametric kernel estimation based on local estimation functions," Statistics and Probability Letters 37, 41–47.

Kauermann, G. and Tutz, G. (1995), "Local likelihood estimation and bias reduction in varying-coefficient models," Forschungsbericht 95-5, Technische Universität Berlin. Forschungsberichte des Fachbereichs Informatik.

Loader, C.R. (1995), "Old faithfull erupts: Bandwidth selection reviewed," Technical Report, AT&T Bell Laboratories.

Marron, J.S. & Chung, S.S. (1997), "Presentation of smoothers: The family approach," Technical Report, University of North Carolina.

Mercer, W.B., and Hall, A.D. (1911), "The experimental error of field trials," Journal of Agricultural Science (Cambridge) 4, 107–132.

Opsomer, J.D. (1995), "Estimating a function by local linear regression when the errors are correlated," Preprint 95-42, Department of Statistics, Iowa State University.

Ruppert, D., Steather, S.J., and Wand, M.P. (1995), "An effective bandwidth selector for local least squares regression," Journal of the American Statistical Association 90, 1257–1270.

Tibshirani, R., and Hastie, T. (1987), "Local likelihood estimation," Journal of the American Statistical Association 82, 559–567.

Tutz, G., and Kauermann, G. (1995), "Varying coefficients in multivariate generalized linear models, a local likelihood approach," Forschungsbericht 95-5, Technische Universität Berlin. Forschungsberichte des Fachbereichs Informatik.

Unwin, A., and Unwin, D. (1998) "Exploratory spatial data analysis with local statistics," The Statistician 47, 415–421.

van der Linde, A. (1994), "On cross-validation for smoothing splines in the case of dependent observations," The Australian Journal of Statistics 36(1), 67–73.

Younes, L. (1988), "Estimation and annealing for Gibbsian fields," Annales de l'Institut Henri Poincaré 24, 269–294.