Kriging by Local Polynomials

GUDMUND HØST[1]

November 11, 1997

Abstract

A flexible framework for prediction of a random process with unknown trend and correlated residuals is presented. Our approach is motivated by a local parametric model, and we derive a locally optimal predictor of the process at unobserved locations. Comparisons with local regression estimation and Kriging are made, and we show that the proposed class of methods provides a bridge between these two approaches. A procedure for parameter estimation and model selection is suggested. The method is illustrated through a simulation study and through an application to European sulphate data.

KEY WORDS: Correlated data; Cross-validation; neighborhood Kriging; Local modeling; Local regression; Nonparametric regression; Trend estimation.

[1]Norwegian Computing Center, Box 114 Blindern, 0314 Oslo, Norway.

1 Introduction

Data from physical and environmental processes often exhibit mixture of low-frequency (trend) and high-frequency (residual) components. It is often difficult to justify stationarity of the underlying process or a parametric form of the trende function. It may be equally hard to justify an uncorrelated residual process. The purpose of this paper is to present a method for prediction of a process with unspecified trend function and autocorrelated residuals. For convenience, our notation is for a process on the line, but the framework extends easily to higher dimensions.

Consider a continuous random process $y(s)$, where s is a location on the real line. Let $y(s)$ have the decomposition

$$y(s) = f(s) + v(s), \tag{1}$$

where $f(s)$ is the trend (or mean) function and $v(s)$ is the residual process. The first and second order moments of the residual process are

$$\begin{aligned} \mathrm{E}\,\{v(s)\} &= 0, & (2)\\ \mathrm{Cov}\,\{v(s'), v(s'')\} &= \sigma^2\,\rho(|s' - s''|; a). & (3) \end{aligned}$$

Here, σ^2 is the variance and $\rho(\cdot; a)$ is a positive correlation function. We will refer to the parameter a as the correlation range. The correlation range is the distance beyond which the correlation function can be neglected. In general, σ^2, a and $\rho(\cdot; a)$ must be estimated from data.

Denote the data by $\boldsymbol{y} = (y(s_1), y(s_2), \ldots, y(s_n))'$, with data locations $\{s_1, \ldots, s_n\}$. Given the data $\boldsymbol{y}$, we consider the smoothing problem of estimating f and the prediction problem of inferring y at an arbitrary location x.

Our approach is motivated by a local, parametric approximation to the trend function f within a window of radius h. We derive an "optimal" trend estimator and predictor for this local model. The global properties will be governed by the bandwidth h and a kernel function $K_h(\cdot)$. We obtain a framework in which both polynomial regression estimation (Hastie and Loader 1993; Fan and Gijbels 1996) and Kriging prediction (Journel and Huijbregts 1978, pp. 303–443; Ripley 1981, pp. 44–50; and Cressie 1991, pp. 105–182) can be described as special cases.

The performance of nonparametric kernel estimators of f when the residuals are correlated has been studied by various authors, see for example Hart and Wehrly (1986), Altman (1990) and Hart (1991). Some related results for spline smoothing are given by Diggle and Hutchinson (1989). An important result of all these authors is that neglecting (positive) correlation of the residuals may lead to grossly undersmoothed trend estimates, and they suggest various procedures for adjusting the bandwidth. In contrast to such approaches, our method explicitly incorporates correlation structure in the trend estimator.

Taking the trend as a linear combination of low-order polynomials, the Kriging predictor is the best linear unbiased predictor (BLUP) of the process at an arbitrary location, in the sense that it minimizes mean squared prediction error over all linear predictors (Cressie 1991 pp.172–173, Goldberger 1962). On the other hand, prior knowledge of the functional form of f is often unavailable, and then the resulting Kriging predictor will be biased.

In contrast, our proposed predictor is specifically designed to account for an unknown trend. The trade-off between bias and variance is determined by a bandwidth h, and the optimal bandwidth for a data set can be selected through cross-validation to minimize an estimate of mean squared prediction error. We will see that Kriging is a special case of our proposed predictor, obtained by letting h tend to infinity. Consequently, Kriging prediction will tend to have larger bias and smaller variance than the predictor which uses optimal bandwidth.

Practitioners frequently apply Kriging only to data local to the prediction location, a procedure we will refer to as neighborhood Kriging. This ad hoc procedure will reduce the bias of the predictor, but the prediction error estimate usually reported does not incorporate bias effects. Furthermore, neighborhood Kriging may produce predictions that are discontinous when the support of data in the neighborhood changes, due to the crude windowing of the data. The present approach allows for inclusion of bias effects in the

prediction error estimate. Also, smoothness properties of the predictions can be controlled by choosing an appropriate kernel function $K_h(\cdot)$.

Our article is structured as follows: In Section 2, the local polynomial trend estimator and predictor are introduced, and some finite sample properties are given. Section 3 contains a demonstration of the method on simulated data and an application to European sulphate data. We give some final remarks in Section 4.

2 Mathematical Framework

Under model (1)–(3), the data can be expressed

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{v}, \tag{4}$$

$$\mathrm{E}\,\{\boldsymbol{v}\} = \boldsymbol{0}, \tag{5}$$

$$\mathrm{Var}\,\{\boldsymbol{v}\} = \sigma^2\,\boldsymbol{R}. \tag{6}$$

Here, the elements of the vectors $\boldsymbol{f}$ and $\boldsymbol{v}$ are the values of the trend f and residual v at the data locations $\{s_1, \ldots, s_n\}$. Furthermore, the elements of the $n \times n$ matrix $\boldsymbol{R}$ are the correlations between the residuals at the data locations. We will refer to the model (1)–(3) or its finite-dimensional representation (4)–(6) as our *global* model. It is our most general representation of the problem, but it is of limited use since the trend is completely unspecified. Our first goal is to estimate f at an arbitrary location x by a local polynomial.

2.1 Local Regression Revisited

To motivate our method, we start by giving an alternative interpretation of local regression. The usual local polynomial regression trend estimate at location x can be written

$$\widetilde{f}(x) = \boldsymbol{b}'(x)(\boldsymbol{B}'\boldsymbol{K}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{K}_h\boldsymbol{y}. \tag{7}$$

Here, $\boldsymbol{b}(x)$ is a $(q+1)$-vector of polynomials of increasing order, $\boldsymbol{b}'(x) = (1, x, \ldots, x^q)$, and the columns of $\boldsymbol{B}$ are these polynomials evaluated at the data locations, $(\boldsymbol{B})_{ij} = b_i(s_j) = s_j^{i-1}$; $i = 1, \ldots, q+1$; $j = 1, \ldots, n$. Furthermore, $\boldsymbol{K}_h$ is an $n \times n$ diagonal matrix of weights with i'th element

$$(\boldsymbol{K}_h)_i = h^{-1} K[(s_i - x)/h],$$

and $K(\cdot)$ is a symmetric density function supported on $(-1, 1)$. The dependence of $\boldsymbol{K}_h$ on x will be suppressed in our notation. Alternatively, we may re-express (7) as

$$\widetilde{f}(x) = \boldsymbol{b}'(x)\, \widetilde{\boldsymbol{\beta}},$$

where $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{B}'\boldsymbol{K}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{K}_h\boldsymbol{y}$.

For $h \to \infty$, we may recognize $\widetilde{\boldsymbol{\beta}}$ as the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$. We introduce a *local polynomial model* for $\boldsymbol{y}$ by asking the following question. Under which model is $\widetilde{\boldsymbol{\beta}}$ the OLS estimate of $\boldsymbol{\beta}$ for arbitrary h? The answer is

$$\boldsymbol{y} = \boldsymbol{B}\,\boldsymbol{\beta} + \boldsymbol{\omega}, \tag{8}$$

$$\mathrm{E}\,\{\boldsymbol{\omega}\} = \boldsymbol{0}, \tag{9}$$

$$\mathrm{Var}\,\{\boldsymbol{\omega}\} = \boldsymbol{K}_h^{-1/2}\,\boldsymbol{R}\,\boldsymbol{K}_h^{-1/2}. \tag{10}$$

The local polynomial model (8)–(10) is not meant to give a true representation of the data $\boldsymbol{y}$, it is merely a mathematical construction which will prove useful in defining and motivating our method. In particular, the residual

correlation may have different interpretations under the two models. Before giving an interpretation of (8)–(10), we propose the correponding local polynomial model for the process $y(s)$

$$\begin{aligned} y(s) &= \boldsymbol{b}'(x)\boldsymbol{\beta} + \omega(s), && (11) \\ \mathrm{E}\,\{\omega(s)\} &= 0, && (12) \\ \mathrm{Cov}\,\{\omega(s), \omega(t)\} &= K_h^{-1/2}(s-x)\, K_h^{-1/2}(t-x)\, \rho(|s-t|), && (13) \end{aligned}$$

where $s \in [x-h, x+h]$ and $K_h(s-x) = h^{-1}K[(s-x)/h]$. We may regard $K_h^{-1}(s-x)$ as a pseudo-variance function. In the local polynomial model, only data within the window $[x-h, x+h]$ are given finite pseudo-variance. Consequently, the estimators $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{f}(x)$ assign non-zero weights only to local data. An informal interpretation is that we take the local polynomial model as a prior model of the data, the prior model being specified by h and $K(\cdot)$. Although (8)–(13) is formally defined for all data locations, the local model will be used and interpreted *only* for data locations within the window.

Later use of the term *local* in this paper will always refer to the local polynomial model (8)–(10) or (11)–(13). For example, we call the estimate $\widetilde{\boldsymbol{\beta}}$ the local ordinary least squares (LOLS) estimate, since it is the OLS estimate under the local model (8)–(10). Similarly, $\widetilde{\boldsymbol{\beta}}$ is the local best linear unbiased estimate (LBLUE) of $\boldsymbol{\beta}$ and $\widetilde{f}(x)$ is the local best linear unbiased predictor (LBLUP) of $y(x)$ when the residuals are uncorrelated ($\boldsymbol{R} = \boldsymbol{I}$).

2.2 Accounting for Correlated Residuals

An extension of local polynomial regression in the present situation is to derive the LBLUE and LBLUP for correlated residuals. We may apply standard theory of multivariate regression (for example Mardia, Kent and Bibby 1979, pp. 171–173), to the local model (8)–(10) on $[x-h, x+h]$. Then the LBLUE may be expressed

$$\widehat{\boldsymbol{\beta}} \;=\; (\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{y}. \tag{14}$$

Here, $\boldsymbol{C}_h$ is a matrix of weights defined as follows. First, denote by $\boldsymbol{R}_h$ the local correlation matrix, i.e. the matrix where the elements are correlations between the residuals at the data locations within $[x-h, x+h]$. Also, denote by $\boldsymbol{R}_h^{-1}$ the inverse of $\boldsymbol{R}_h$. Each element of $\boldsymbol{R}_h^{-1}$ corresponds to a pair of data locations (x_i, x_j) within the window $[x-h, x+h]$. Now, transfer the elements of $\boldsymbol{R}_h^{-1}$ to the $n \times n$ matrix $\{\boldsymbol{R}_h^{-1}\}_0$ according to the ordering of the data vector $\boldsymbol{y}$, and set all other elements of $\{\boldsymbol{R}_h^{-1}\}_0$ equal to zero. Then the matrix $\{\boldsymbol{R}_h^{-1}\}_0$ is just the inverse local correlation matrix, ordered in correspondence to the data vector $\boldsymbol{y}$. Finally, $\boldsymbol{C}_h$ is defined by kernel-smoothing of $\{\boldsymbol{R}_h^{-1}\}_0$, $\boldsymbol{C}_h = \boldsymbol{K}_h^{1/2}\{\boldsymbol{R}_h^{-1}\}_0\boldsymbol{K}_h^{1/2}$. The dependence of $\boldsymbol{C}_h$ on location x will be suppressed in our notation.

The estimate $\widehat{\boldsymbol{\beta}}$ given in (14) could alternatively be called the local general-

ized least squares estimate of $\boldsymbol{\beta}$. The corresponding trend estimate is

$$\begin{aligned} \widehat{f}(x) &= \boldsymbol{b}'(x)\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{b}'(x)(\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{y}. \end{aligned} \tag{15}$$

This estimate is *not* LBLUP for $y(x)$ when the residuals are correlated. Following Goldberger (1962), and using the local model (8)–(10), the LBLUP is

$$\widehat{y}(x) = \widehat{f}(x) + \boldsymbol{r}_h'\boldsymbol{C}_h(\boldsymbol{y} - \boldsymbol{B}\widehat{\boldsymbol{\beta}}), \tag{16}$$

with $\boldsymbol{r}_h = K_h^{-1/2}(0)\{\boldsymbol{K}_h^{-1/2}\}_0\,\boldsymbol{r}$. Here, $\boldsymbol{r}$ is an n-vector with its i'th element $\rho(|s_i - x|; a)$. Furthermore, $\{\boldsymbol{K}_h^{-1/2}\}_0$ is the diagonal $n \times n$ matrix with i'th element

$$\left(\{\boldsymbol{K}_h^{-1/2}\}_0\right)_i = \begin{cases} h^{1/2}\,K^{-1/2}[(s_i - x)/h] & \text{if } s_i \in (x-h, x+h), \\ 0 & \text{otherwise.} \end{cases}$$

The predictor (16) is the "universal Kriging" predictor (Cressie 1991, pp. 151–157) of y under the local model (11)–(13). Thus, we propose to use $\widehat{f}$ for smoothing and $\widehat{y}$ for prediction of the process at arbitrary locations.

Some properties and special cases are discussed in the following remarks:

1. When predicting at locations further than a from the data locations, $\boldsymbol{r}$ will be close to zero and the predictor will be close to $\widehat{f}(x)$. On the other hand, assume we want to predict the process at a data location

s_i. Denote by $\boldsymbol{r}_i'$ the i'th row of $\{\boldsymbol{R}_h\}_0$ and by $\boldsymbol{r}^{(j)}$ the j'th column of $\{\boldsymbol{R}_h^{-1}\}_0$. Since $\boldsymbol{R}_h\boldsymbol{R}_h^{-1} = \boldsymbol{I}$, we have

$$\boldsymbol{r}_i'\ \boldsymbol{r}^{(j)} = \begin{cases} 1 & \text{if } i{=}j, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\boldsymbol{r}_h'\boldsymbol{C}_h = K_h^{-1/2}(0)\ \boldsymbol{r}_i'\ \{\boldsymbol{R}_h^{-1}\}_0\ \boldsymbol{K}_h^{1/2} = (0,\ldots,0,1,0,\ldots,0)$, i.e. a vector where the only non-zero element is a 1 at the i'th position. It follows from (16) that $\widehat{y}(s_i) = y(s_i)$, implying that the curve $\widehat{y}(x)$ will pass through the data values at the data locations. Thus, $\widehat{y}$ combines the "exact" interpolation property of Kriging with a local regression type of trend estimate.

2. Suppose the residuals are uncorrelated, and x is not a data location. Then $\boldsymbol{r}_h = \boldsymbol{0}$ and, by comparing with (7) and (15), we see that

$$\widehat{y}(x) = \widehat{f}(x) = \widetilde{f}(x).$$

This is the traditional local polynomial regression trend estimator (Hastie and Loader 1993).

3. Let the bandwidth h be much greater than the study interval, $h \gg 1$. Then Taylor expansion gives

$$K_h(s-x) = h^{-1}K(0) + O(h^{-3}).$$

To leading order in h^{-1}, we get the generalized least squares estimate

$$\widehat{\boldsymbol{\beta}} \approx (\boldsymbol{B}'\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{R}^{-1}\boldsymbol{y}. \tag{17}$$

Furthermore, we get $\boldsymbol{r}_h'\boldsymbol{C}_h \approx \boldsymbol{r}'\boldsymbol{R}^{-1}$, and the predictor reduces to

$$\widehat{y}(x) \approx \boldsymbol{b}'(x)\widehat{\boldsymbol{\beta}} + \boldsymbol{r}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{B}\widehat{\boldsymbol{\beta}}). \tag{18}$$

If we use the approximation (17) for $\widehat{\boldsymbol{\beta}}$ we may recognize the right hand side of (18) as the universal Kriging predictor (see Cressie 1991, pp. 151–157).

4. An advantage of the proposed local method as compared to global methods is that the inversion of the correlation matrix $\boldsymbol{R}$ is replaced by a sequence of inversions of smaller local correlation matrices $\boldsymbol{R}_h$. This may be a computational advantage for large data sets and may be done by parallel processing.

2.3 Finite Sample Properties

We showed in Section 2.2 that our proposed estimator and predictor are optimal under the local model (8)–(10). However, the main interest is in the performance of these methods under the global model (1)–(3). In this section, we state some finite sample properties of $\widehat{f}(x)$ and $\widehat{y}(x)$. Some asymptotics are given in Høst (1996). Using in-fill asymptotics (Cressie (1991), p.100), the trend estimator is related to the Nadaraya-Watson estimator and the traditional local linear estimator. It is argued that accounting for correlated

residuals gives smaller bias and larger variance asymptotically. However, except for boundary effects the differences found in Høst (1996) are minor. The reason is that the configuration of data locations is not important asymptotically.

2.3.1 Bias for Finite Samples

For $s \in [x-h, x+h]$, suppose f can be expanded in a Taylor series f_q around x:

$$f(s) = f_q(s) + Q_{q+1}.$$

Here, Q_{q+1} is the remainder and $f_q(s)$ is the polynomial

$$\begin{aligned} f_q(s) &= f(x) + h(\frac{s-x}{h})f'(x) + \ldots + \frac{h^q}{q!}(\frac{s-x}{h})^q f^{(q)}(x) \qquad (19) \\ &= b_0 + h(\frac{s-x}{h})b_1 + \ldots + h^q(\frac{s-x}{h})^q b_q. \end{aligned}$$

Denote by $\widehat{\boldsymbol{\beta}}_{GLS}$ the generalized least squares (GLS) estimate of the series coefficients of f_q. Then we have

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{B}'\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{R}^{-1}\boldsymbol{y}.$$

According to basic results in multivariate analysis (see for example Mardia, Kent and Bibby 1979, p. 184), any linear unbiased estimate of $\boldsymbol{\beta}$ is of the form $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{A}\boldsymbol{y}$, with $\boldsymbol{A}\boldsymbol{B} = 0$. The proposed estimator corresponds

to choosing

$$\boldsymbol{A} = (\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{C}_h - (\boldsymbol{B}'\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{R}^{-1}.$$

It is easily verified that $\boldsymbol{AB} = 0$, therefore $\widehat{\boldsymbol{\beta}}$ is a linear unbiased estimate of $\boldsymbol{\beta}$. This shows that $\widehat{f}(x) = \boldsymbol{b}'(x)\widehat{\boldsymbol{\beta}}$ is a linear unbiased estimate of $f_q(x)$, i.e. of any polynomial of order $\leq q$. Our particular choice of the matrix $\boldsymbol{A}$ corresponds to a *local* Taylor expansion, and ensures that the remainder Q_{q+1} is of order $O(h^{q+1})$. Whereas both $\widehat{f}(x)$ and $\widehat{f}_{GLS}(x) = \boldsymbol{b}'(x)\widehat{\boldsymbol{\beta}}_{GLS}$ are linear unbiased estimates of a q'th order polynomial, the bias in estimating an unknown smooth function will be of order $O(1)$ for $\widehat{f}_{GLS}(x)$ and of order $O(h^{q+1})$ for the proposed estimator.

The true bias of $\widehat{f}(x)$ is

$$\begin{aligned} \text{Bias}\{\widehat{f}(x)\} &= \text{E}\{\widehat{f}(x) - f(x)\} \\ &= \boldsymbol{b}'(x)(\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{f} - f(x). \end{aligned}$$

This may be approximated using the Taylor expansion (19). We get

$$\text{Bias}\{\widehat{f}(x)\} = \frac{1}{(q+1)!}\boldsymbol{b}'(x)(\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{s}^{q+1}f^{(q+1)}(x) + O(h^{q+2}). \quad (20)$$

Here, we have introduced the notation $\boldsymbol{s}^q = ((s_1 - x)^q, \ldots, (s_n - x)^q)'$. Thus, the bias may be estimated by plugging in an estimate of the $(q+1)$'th derivative $f^{(q+1)}(x)$. We estimate $f^{(q+1)}(x)$ by fitting a local polynomial trend of order $\geq q+1$ to the process and taking the $(q+1)$'th derivative of this local

polynomial (see Hastie and Loader 1993, p. 125). Typically, a larger h would be used in estimating $f^{(q+1)}(x)$ than in estimating $f(x)$, because estimates of higher order derivatives tend to have large variance. A detailed study of the properties of the bias estimate (20) is beyond the scope of the present work.

The predictor can be written

$$\widehat{y}(x) = \widehat{f}(x) + \boldsymbol{r}_h' \boldsymbol{C}_h (\boldsymbol{y} - \widehat{\boldsymbol{f}}),$$

where $\widehat{\boldsymbol{f}} = \boldsymbol{B}\widehat{\boldsymbol{\beta}}$ is the vector of trend estimates at the data locations. Taking the expectation, we get

$$\begin{aligned} \mathrm{E}\,\{\widehat{y}(x)\} &= \mathrm{E}\,\{\widehat{f}(x)\} + \boldsymbol{r}_h' \boldsymbol{C}_h (\boldsymbol{f} - \mathrm{E}\,\{\widehat{\boldsymbol{f}}\}) \\ &= \mathrm{E}\,\{\widehat{f}(x)\} - \boldsymbol{r}_h' \boldsymbol{C}_h \,\mathrm{Bias}\,\{\widehat{\boldsymbol{f}}\}. \end{aligned}$$

This last expression can be used to calculate the *prediction bias* of $\widehat{y}(x)$:

$$\begin{aligned} \mathrm{Bias}\,\{\widehat{y}(x)\} &= \mathrm{E}\,\{\widehat{y}(x) - y(x)\} \\ &= \mathrm{Bias}\,\{\widehat{f}(x)\} - \boldsymbol{r}_h' \boldsymbol{C}_h \,\mathrm{Bias}\,\{\widehat{\boldsymbol{f}}\}. \end{aligned} \tag{21}$$

By following the argument given at the end of Section 2.2 we can verify that the prediction bias is zero at the data locations. Far away from the data locations the prediction bias will be close to the bias of the trend estimate. The trend estimate inherent in Kriging is $\widehat{f}_{GLS}$. Therefore, the Kriging predictor will have greater bias than the proposed predictor in sparsely sampled regions when the trend is different from a q'th order polynomial.

2.3.2 Variance for Finite Samples

Using the notation of Section 2.3.1, the variance of $\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\text{Var } \{\widehat{\boldsymbol{\beta}}\} &= \text{Var } \{\widehat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{A}\boldsymbol{y}\} \\
&= \text{Var } \{\widehat{\boldsymbol{\beta}}_{GLS}\} + 2 \text{ Cov } \{\widehat{\boldsymbol{\beta}}_{GLS},\ \boldsymbol{A}\boldsymbol{y}\} + \boldsymbol{A}\text{Var } \{\boldsymbol{y}\}\boldsymbol{A}' \\
&= \text{Var } \{\widehat{\boldsymbol{\beta}}_{GLS}\} + 2\sigma^2(\boldsymbol{B}'\boldsymbol{R}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}' + \sigma^2\boldsymbol{A}\boldsymbol{R}\boldsymbol{A}' \\
&= \text{Var } \{\widehat{\boldsymbol{\beta}}_{GLS}\} + \sigma^2\boldsymbol{A}\boldsymbol{R}\boldsymbol{A}',
\end{aligned}$$

where the last equality follows from $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}$. Now, the variance of $\widehat{f}(x)$ is

$$\begin{aligned}
\text{Var } \{\widehat{f}(x)\} &= \boldsymbol{b}'(x) \text{ Var } \{\widehat{\boldsymbol{\beta}}\}\, \boldsymbol{b}(x) \\
&= \text{Var } \{\widehat{f}_{GLS}(x)\} + \sigma^2\boldsymbol{b}'(x)\boldsymbol{A}\boldsymbol{R}\boldsymbol{A}'\boldsymbol{b}(x) \geq \text{Var } \{\widehat{f}_{GLS}(x)\},
\end{aligned}$$

because $\boldsymbol{A}\boldsymbol{R}\boldsymbol{A}'$ is positive definite. Thus, by localizing the trend estimate, the bias is decreased and the variance is increased.

The prediction variance is

$$\text{Var } \{\widehat{y}(x) - y(x)\} = \sigma^2(1 - 2\boldsymbol{\alpha}'\boldsymbol{r} + \boldsymbol{\alpha}'\boldsymbol{R}\boldsymbol{\alpha}), \tag{22}$$

where

$$\boldsymbol{\alpha} = \boldsymbol{C}_h\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{b}(x) + \left[\boldsymbol{I} - \boldsymbol{C}_h\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{C}_h\boldsymbol{B})^{-1}\boldsymbol{B}'\right]\boldsymbol{C}_h\boldsymbol{r}_h. \tag{23}$$

At the data locations the prediction variance is zero, because the predictor reproduces the observations. At distances greater than a from the data

locations, the prediction variance will tend to Var $\{\widehat{f}(x)\}$. In comparison, the Kriging prediction variance will tend to Var$\{\widehat{f}_{GLS}(x)\}$. Therefore, the proposed predictor will have greater prediction variance than Kriging when predicting at distant locations.

We have seen that the proposed predictor has smaller prediction bias and greater prediction variance than the Kriging predictor. In practical situations, we would usually be more interested in the mean squared error of prediction:

$$\begin{aligned} \text{MSEP } \{\widehat{y}(x)\} &= \text{E } \{[\widehat{y}(x) - y(x)]^2\} \\ &= \text{Var } \{\widehat{y}(x) - y(x)\} + \text{Bias}^2 \{\widehat{y}(x)\}. \end{aligned} \tag{24}$$

Here, the variance term is given by (22)–(23) and the bias term is given by (21). The proposed predictor may have smaller or greater MSEP than Kriging, but in Kriging the prediction variance is used as an estimate of MSEP. Such an estimate neglects a bias term which is $O(1)$ when the trend is not a q'th order polynomial. This is improved upon in the neighborhood Kriging approach. Using only data within a radius h, the bias of neighborhood Kriging is $O(h^{q+1})$, but this bias is neglected in the MSEP estimate. The prediction error estimate we propose will therefore be more realistic than the prediction error estimate usually reported in Kriging procedures.

In practical applications the true model is unknown. Then the above con-

siderations may not apply, because $\rho(\cdot;\cdot)$, σ^2, a and h must be estimated from data. On the other hand, a comparison of prediction methods on real data may be assessed through cross-validation. This will be pursued in Section 3.2.2.

3 Examples

In this section, we illustrate the aspects of our method through a constructed example and an application to a data set with observations in a two-dimensional domain. The purpose is to check the performance of our method in some practical situations.

3.1 Prediction of Simulated Data

The purpose of this section is to present a procedure for estimation of covariance parameters and bandwidth from simulated data. We also make a comparison with Kriging. We use the trend function $f(s) = 10s^3 - 15s^4 + 6s^5$, which mimics the typical behavior of growth curves, see Hart & Wehrly (1986). Furthermore, we take $\sigma^2 = 0.1^2$ and use an exponential correlation function $\rho(\Delta; a) = \exp(-3\Delta/a)$ with $a = 0.054$. One realisation of this process is simulated at 80 locations drawn from a uniform density on $(0, 1)$. The simulated data and the underlying trend is shown in Figure 1.

Most methods for estimating spatial covariance parameters assume the trend is known, while in our case the unknown trend will have to be estimated using the (unknown) correlation range a. For given h, we suggest the following procedure for parameter estimation. First, subtract a trend estimate as-

suming zero correlation from the data. Then estimate $\widehat{\sigma^2}$ and a from the fitted residuals and re-estimate the trend f using the current estimate of a. This procedure is iterated. For estimating covariance parameters from fitted residuals, we used the weighted least squares method of Cressie (1991), pp. 95–97, applied to lags less than h. An alternative method is demonstrated in Section 3.2. An iterative procedure similar to the procedure used here was originally suggested by Cochrane and Orcutt (1949) in the context of linear regression with autocorrelated errors. Although relevant, a detailed evaluation of alternative methods for parameter and bandwidth estimation is beyond the scope of the present work.

We fit a local linear model to the simulated data using the Epanechnikov kernel $K(t) = 0.75(1-t^2)\, I_{[-1,1]}$ and compare with the linear trend (universal) Kriging predictor. For the local model, covariance parameters were estimated for a range of plausible bandwidths and the mean integrated squared error of prediction (MISEP) was obtained by integrating (24) over the study interval. MISEP was estimated numerically for each bandwidth by cross-validation, using the proposed predictor as given by (16) with $q = 1$. The resulting MISEP is shown in Figure 2, and we see that there is a minimum at $h_0 = 0.17$. This is in close agreement with the optimal bandwidth of 0.218, which is obtained for given σ^2 and a. The corresponding parameter estimates were $\widehat{\sigma}^2 = 0.0792^2$ and $\widehat{a} = 0.0348$.

The bias and variance of the trend estimator forms the basis for the prediction error estimates. The variance of the trend estimator was estimated by plugging in the estimated covariance parameters and the bias was estimated by formula (20). This requires estimation of the second derivative of the trend, and involves again the choice of a bandwidth h_1. For any fixed bandwidth, we would expect estimates of higher order derivatives of a function to have larger variance than estimates of a function value. Therefore, the optimal bandwidth for estimating $f''(x)$ is likely to be larger than the optimal bandwidth for estimating $f(x)$. We used $h_1 = 2h_0 = 0.34$, which gave a reasonably smooth estimate. Procedures for objective selection of bandwidth for derivative estimation should be investigated in the future. Figure 3 shows the estimated and true bias. We see that the bias estimate captures some of the true structure, with the largest disagreement towards the ends of the estimation interval. Figure 4 shows the estimated and true MSE of the trend estimator. Both estimated and true MSE are smallest in the interior of the study interval, and largest near the ends. The estimated MSE is smaller than the true MSE at all locations, mainly because σ^2 is underestimated in this example.

Parameters in the linear trend kriging model was estimated by setting $h = \infty$ in the proposed model. We obtained $\widehat{\sigma}_K^2 = 0.225^2$ and $\widehat{a}_K = 0.868$. The estimated parameter values are larger in this case, because parts of the trend function is now interpreted as residual fluctuations. Predicted values

for Kriging and the local method is shown in Figure 5. We see that the Kriging predictor is nearly interpolating the data linearly, while the proposed predictor is pulled towards the local trend estimate.

A better way of comparing the two predictors is by cross-validation. Our cross-validation exercise omits one data location at a time, and predicts that location using data from all other locations. The procedure is repeated for each data location, resulting in 80 cross-validated data values. The absolute difference between the cross-validated data value and the observed data value at the same location is denoted the *true prediction error*. The true prediction errors for Kriging and local polynomial Kriging is shown in Figure 6. We see that local polynomial Kriging generally has smaller true prediction errors than Kriging. Indeed, the RMS prediction error for local polynomial Kriging is 0.0526 and for Kriging 0.0578. A feature of the Kriging framework is the ability to provide a prediction error estimate. The estimated Kriging RMS prediction error is 0.0326, while the estimated local polynomial Kriging estimate is 0.0491. Hence, the local polynomial method gives somewhat better predictions and much more realistic prediction error estimates than Kriging in this example.

3.2 Prediction of European Sulphate Data

We apply the method to prediction of sulphate concentrations from Europe for January, 1989, and compare with ordinary Kriging through cross-validation. The proposed framework is easily extended to this two-dimensional example with the following modifications. The kernel function is replaced by a product of one kernel function for each spatial dimension. Furthermore, the length of the $\boldsymbol{b}(\boldsymbol{x})$-vector and the number of rows in the $\boldsymbol{B}$-matrix will now be $1/2(q+1)(q+2)$. Finally, the calculation of the bias term will involve estimation of all partial derivatives of $f(\boldsymbol{x})$ of order $q+1$.

The sulphate data were collected daily through the "Co-operative Program for Monitoring and Evaluation of the Long Range Transmission of Air Pollutants in Europe". The data set was provided by the Norwegian Institute for Air Research (NILU) and a description is given in Schaug, Pedersen, Skjelmoen and Kvalvågnes (1993). For January 1989, we used 64 averaged values of sulphate concentrations, measured in units of milligrams of sulfur per liter $[\mu g(S)/l]$. The study area is showed in Figure 7. As concentration data are always positive, it is convenient to operate on a logarithmic scale. In particular, the concentrations referred to in this article will be in units of log $[\mu g(S)/l]$.

3.2.1 Parameter Estimation

As trend model in this example we use a local constant. The procedure for parameter estimation used in this two-dimensional example deviates slightly from the description given in the previous sections. As kernel function, we use the product of two biweight kernels,

$$K(t_1, t_2) = \left[\frac{15}{16}(1 - t_1^2)(1 - t_2^2)\right]^2 I_{[-1,1]}(t_1)\, I_{[-1,1]}(t_2),$$

The bandwidth and spatial covariance parameters for the local constant model are estimated by the procedure described in Section 3.1, with the exception that maximum likelihood is used for estimation of covariance parameters, as suggested by Mardia and Marshall (1984). Some limitations are discussed in Warnes and Ripley (1897) and in Mardia and Watkins (1989).

In our present cross-validation estimate of MISEP, data locations in sparsely sampled regions are given larger weights than data locations in densely sampled regions. Again, we used the exponential correlation function, and after 4 iterations the parameter estimates stabilized at $\widehat{a} = 867\ km$, $\widehat{\sigma^2} = 0.73\ \{\log\ [\mu g(S)/l]\}^2$ and $\widehat{h} = 3000\ km$. Figure 8 shows the fitted variogram function $\gamma(\tau) = \sigma^2[1 - \rho(\tau)]$ for the proposed model. Also shown in this figure are squared differences of residuals between station pairs and the number of station pairs within each lag-average. The squared differences are averaged over $150\,km$ lags. For the purpose of this example, the variogram

function seems to give a reasonable fit to the empirical lag-averages when the separating distance is less than the selected bandwidth of $h = 3000\,km$.

For the Kriging model (assuming a constant trend) the maximum likelihood estimates of the covariance parameters were $\widehat{a}_K = 1780\ km$ and $\widehat{\sigma^2}_K = 1.28\ \{\log\ [\mu g(S)/l]\}^2$. The estimated variance and range parameters are larger in this case, because low-frequency fluctuations in the data are incorporated in the residual variability. The fitted parametric variogram functions and lag-averaged squared differences between station pairs for the constant trend model are shown in Figure 9.

3.2.2 Cross-validation

Comparisons between the proposed method and Kriging are made by analysis of cross-validation estimates for predicted values and prediction errors. The true prediction errors for the proposed method and the Kriging approach are shown in Figure 10. We see that the two methods both have small errors for locations in the central part of the domain, which is the most densely sampled region. An exception is site 3, where both methods give poor predictions. Near the boundary of the study area, the proposed method seems to have the smallest prediction errors (sites 21, 23, 29, 34, 39, 46). The RMS true prediction errors for the proposed method was 0.747 and for Kriging 0.793.

The RMS estimated prediction errors were 0.653 for the proposed method and 0.692 for Kriging.

We see the difference between the two methods is not large for this example when averaging over all data. However, there are indications that the proposed method gives better predictions in sparsely sampled regions. An important feature of the proposed predictor as compared to Kriging is the improved ability to translate information from densely sampled regions to sparsely sampled regions. This is because the predictor incorporates a more detailed trend structure than Kriging. In applied problems, inference for sparsely sampled regions may be quite important.

4 Conclusions

An advantage of the proposed framework is that it can account for both correlation structure and a general trend function. Our framework also includes some familiar methods of smoothing and prediction as special cases. In particular, if the residuals are uncorrelated we get usual local polynomial regression, and when the smoothing parameter h is infinite we get Kriging.

Nonparametric regression estimation is a tool for estimating the conditional expectation of a process given the data. This is also the purpose of Kriging prediction, but under different model assumptions. While Kriging can be regarded as a high-pass filter designed to utilize structure in the residuals, nonparametric regression estimation may be viewed as a low-pass filter, intended to reveal trend structure. In keeping with this terminology, the predictor we propose combines a low-pass filter and a high-pass filter to take advantage of both high-frequency and low-frequency structure. However, for the model to be identifiable for a given set of data, the process should have distinct high-frequency and low-frequency structure, and sufficiently large sample size for separating these scales. This potential problem may be of little concern if the purpose of the analysis is prediction rather than inference of model parameters.

The expense of using the proposed framework for prediction purposes is

having to choose the smoothing parameter h. However, practitioners will frequently apply the Kriging predictor only to some neighborhood of the prediction location. Therefore, one will need some smoothing parameter (for example radius of the neighborhood) also in this case. The proposed framework can be regarded as a theoretical basis for neighborhood Kriging, and we propose a coherent prediction error estimate for this situation.

A strength of the proposed trend estimator is that it explicitly accounts for boundary effects and unevenly spaced data locations through incorporation of residual correlation structure. It is also consistent with the parametric trend estimator when the residuals are correlated, because it is optimal (in the best linear unbiased sense) as the bandwidth tends to infinity. Furthermore, our unified framework for smoothing and prediction allows for cross-validation for bandwidth selection. This is more problematic in direct application of a conventional nonparametric smoother, because smoothing and prediction is not equivalent when the errors are correlated.

When the bandwidth h is fixed, the problems of estimating covariance parameters within the proposed framework are similar to the problems encountered in Kriging. Various procedures for parameter estimation of spatial processes are suggested in the literature (Ripley 1981; Cressie 1991; Mardia and Marshall 1984; Zimmerman 1989; Hjort and Omre 1994). In some situations, there is a natural choice of h to use with one of these parameter estimation

procedures. In other problems, our current recommendation is to combine a reasonable procedure with some cross-validation criterion to assess the bandwidth. As a final diagnostic check, we recommend an assessment of *individual* cross-validated prediction errors. Particular attention should be given to data locations in sparsely sampled regions, since this is where prediction methods are likely to differ the most. These points are illustrated in the sulphate data example of Section 3.2.2, but see Høst, Omre and Switzer (1995) for a related application.

The literature of local polynomial regression is abundant in extensions and modifications to the "basic" local polynomial regression approach, many of which may apply to the present situation. Some authors advocate fixing the number of local observations used in the estimation, giving a variable bandwidth (Cleveland 1979; Müller and Stadtmüller 1987; Fan and Gijbels 1992; Fan and Gijbels 1996). The properties of such variable bandwidth predictors deserve further investigations. A further extension would be to allow for the covariance of the process to vary with location. This would give prediction error estimates that are location-specific, not only dependent on sampling geometry. Hopefully, progress along these lines can be reported in the future.

References

Altman, N. S. (1990), 'Kernel smoothing of data with correlated errors', *Journal of the American Statistical Association* **85**(411), 749–759.

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

Cochrane, D. & Orcutt, G. H. (1949), 'Application of least squares regression to relationships containing autocorrelated terms', *Journal of the American Statistical Association* **44**, 32–61.

Cressie, N. (1991), *Statistics for Spatial Data*, Wiley, New York.

Diggle, P. J. & Hutchinson, M. F. (1989), 'On spline smoothing with correlated errors', *Australian Journal of Statistics* **31**, 166–182.

Fan, J. & Gijbels, I. (1992), 'Variable bandwidth and local linear regression smoothers', *AnlsStat* **20**, 2008–2036.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

Goldberger, A. S. (1962), 'Best linear unbiased prediction in the generalized linear regression model', *Journal of the American Statistical Association* **57**, 369–375.

Hart, J. D. (1991), 'Kernel regression estimation with time series errors', *Journal of Royal Statistical society, series B* **53**(1), 173–187.

Hart, J. D. & Wehrly, T. E. (1986), 'Kernel regression estimation using repeated measurement data', *Journal of the American Statistical Association* **81**(396), 1080–1088.

Hastie, T. & Loader, C. (1993), 'Local regression: Automatic kernel carpentry', *Statistical Science* **8**, 120–143.

Hjort, N. L. & Omre, H. (1994), 'Topics in spatial statistics', *Scandinavian Journal of Statistics* **21**(4), 289–357. With discussion.

Høst, G. (1996), Contributions to the analysis of spatial and spatial-temporal data, Dr.scient. thesis, Department of Mathematics, Statistics Division, University of Oslo.

Høst, G., Omre, H. & Switzer, P. (1995), 'Spatial interpolation errors for monitoring data', *Journal of the American Statistical Association* **90**, 853–861.

Journel, A. G. & Huijbregts, C. J. (1978), *Mining Geostatistics*, Academic Press, London.

Mardia, K. V. & Marshall, R. J. (1984), 'Maximum likelihood estimation of models for residual covariance in spatial regression', *Biometrika* **71**, 135–146.

Mardia, K. V. & Watkins, A. J. (1989), 'On the multimodality of the likelihood in the spatial linear model', *Biometrika* **76**, 289–295.

Müller, H. G. & Stadtmüller, U. (1987), 'Variable bandwidth kernel estimators of regression functions', *AnlsStat* **15**, 610–625.

Ripley, B. (1981), *Spatial Statistics*, Wiley, New York.

Schaug, J., Pedersen, U., Skjelmoen, J. E. & Kvalvågnes, I. (1993), Data report 1991. Part 1: Annual summaries, EMEP/CCC–Report 4/93, NILU, Lillestrøm, Norway.

Warnes, J. J. & Ripley, B. D. (1987), 'Problems with likelihood estimation of covariance functions of spatial Gaussian processes', *Biometrika* **74**, 640–642.

Zimmerman, D. (1989), 'Computationally efficient restricted maximum likelihoood estimation of generalized covariance functions', *Mathematical Geology* **21**, 655–672.

List of Figures

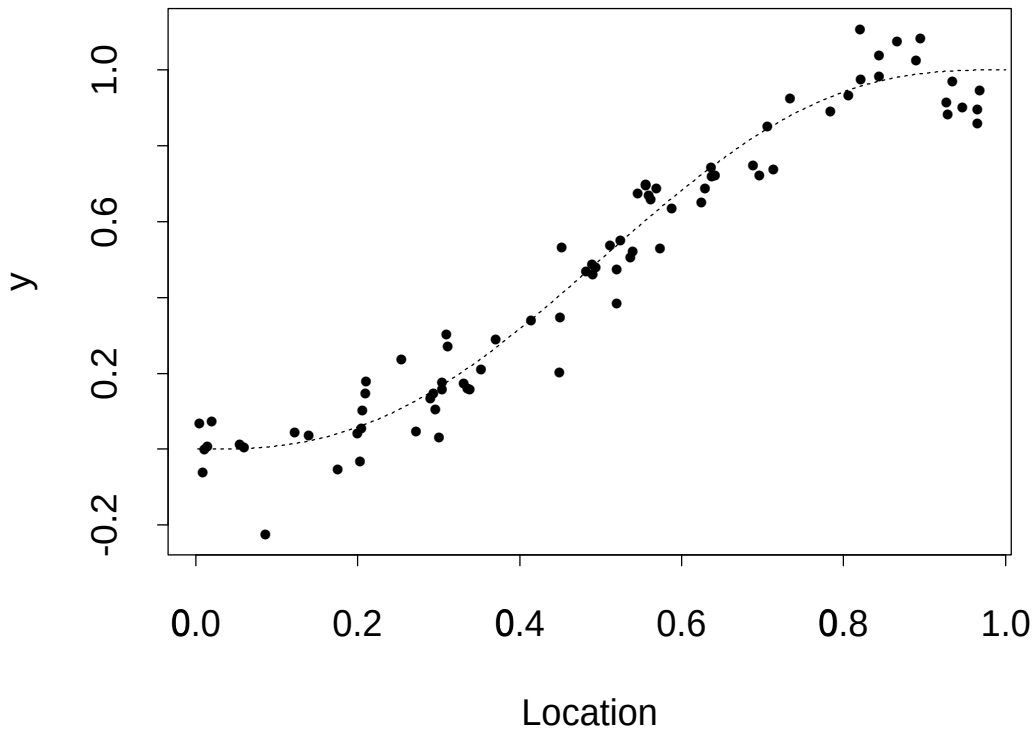

Figure 1: *Simulated data (dots) and underlying trend function (broken line)*

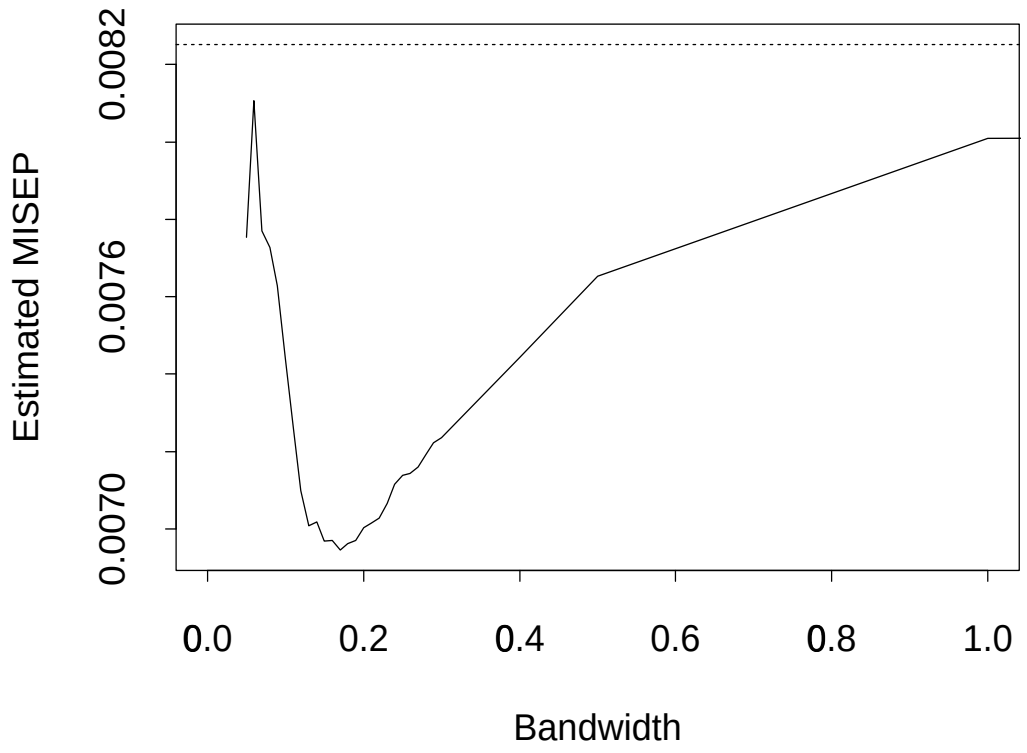

Figure 2: *Mean Integrated Squared Error of Prediction (MISEP) as estimated by cross-validation of simulated data. Proposed method (full line) and Kriging (broken line)*

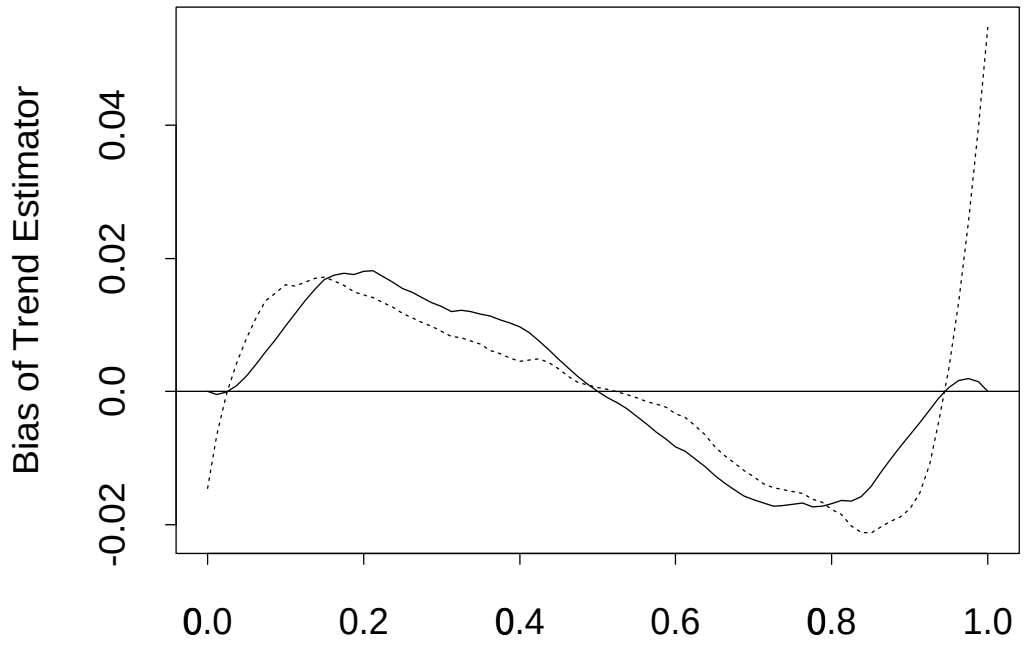

Figure 3: *True bias of trend estimator (full line) and estimated bias (broken line)*

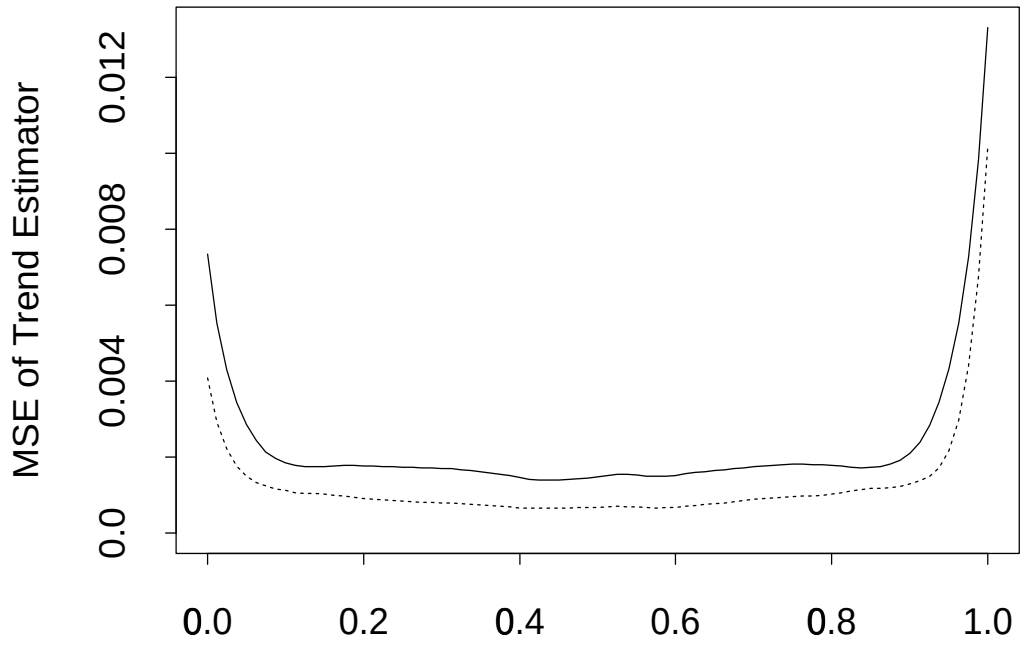

Figure 4: *True mean squared error (MSE) of trend estimator (full line) and estimated MSE (broken line)*

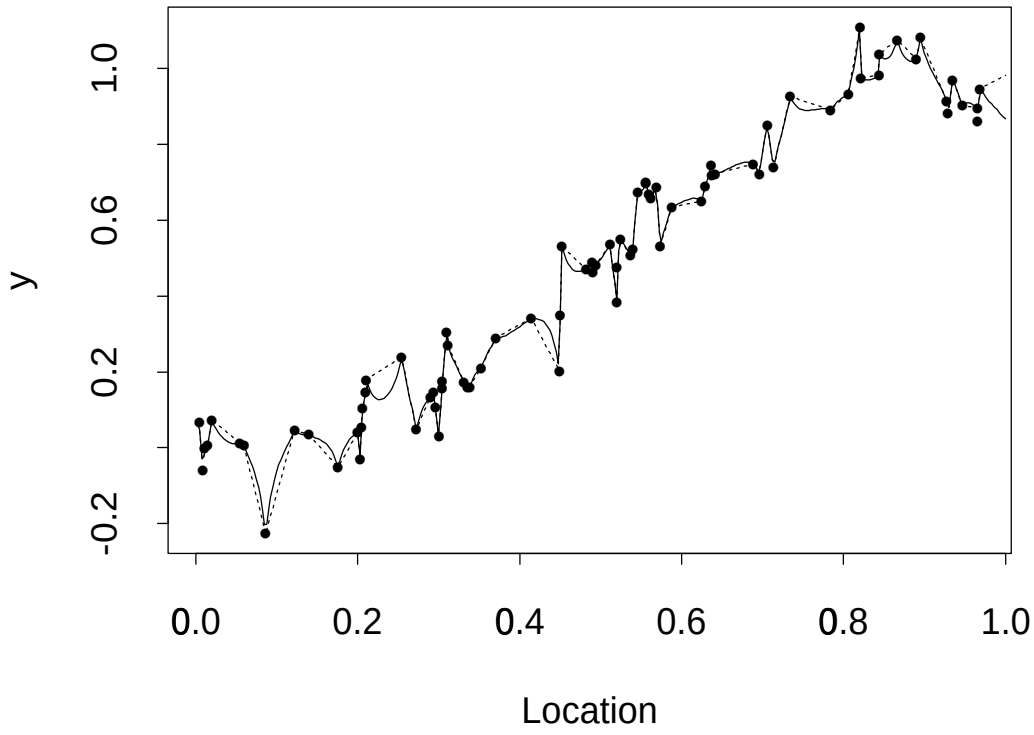

Figure 5: *Proposed predictor (full line) and Kriging predictor (broken line)*

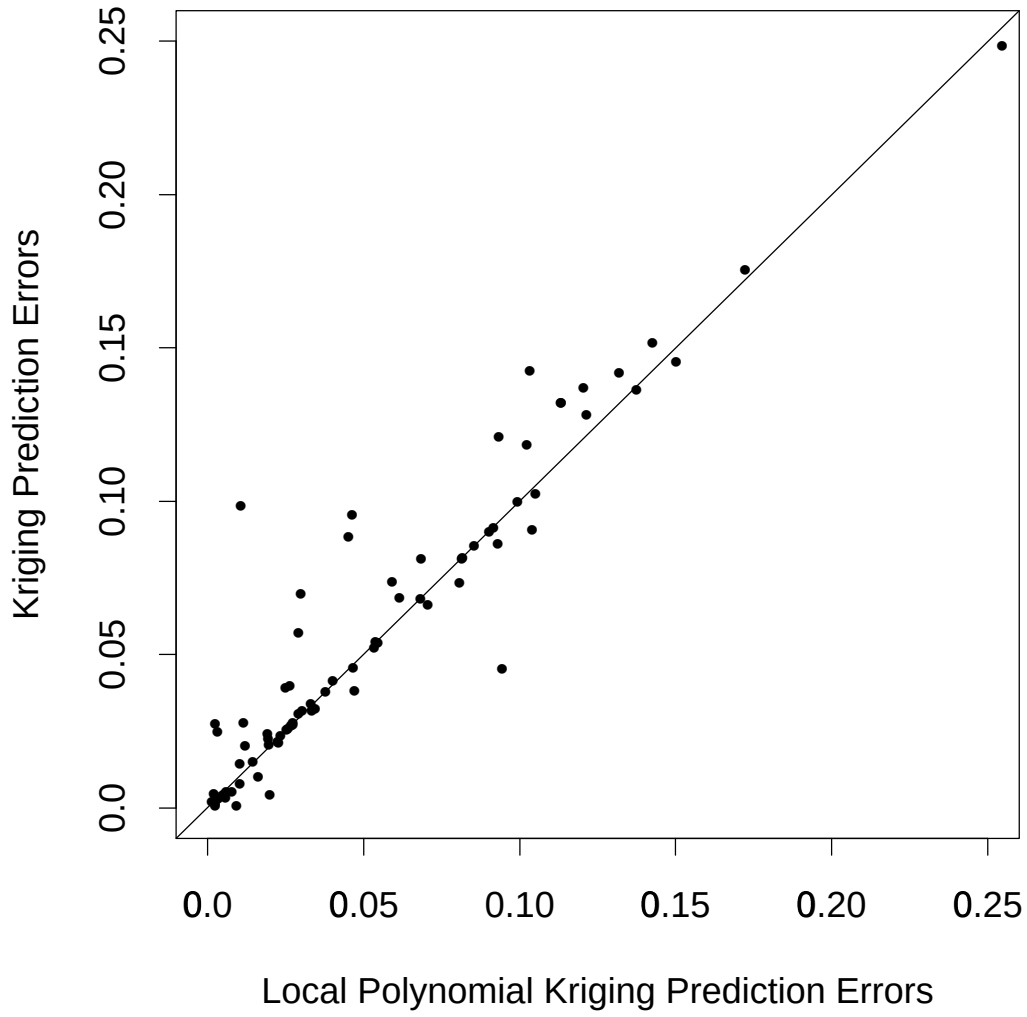

Figure 6: *True prediction errors for the proposed method and for Kriging approach*

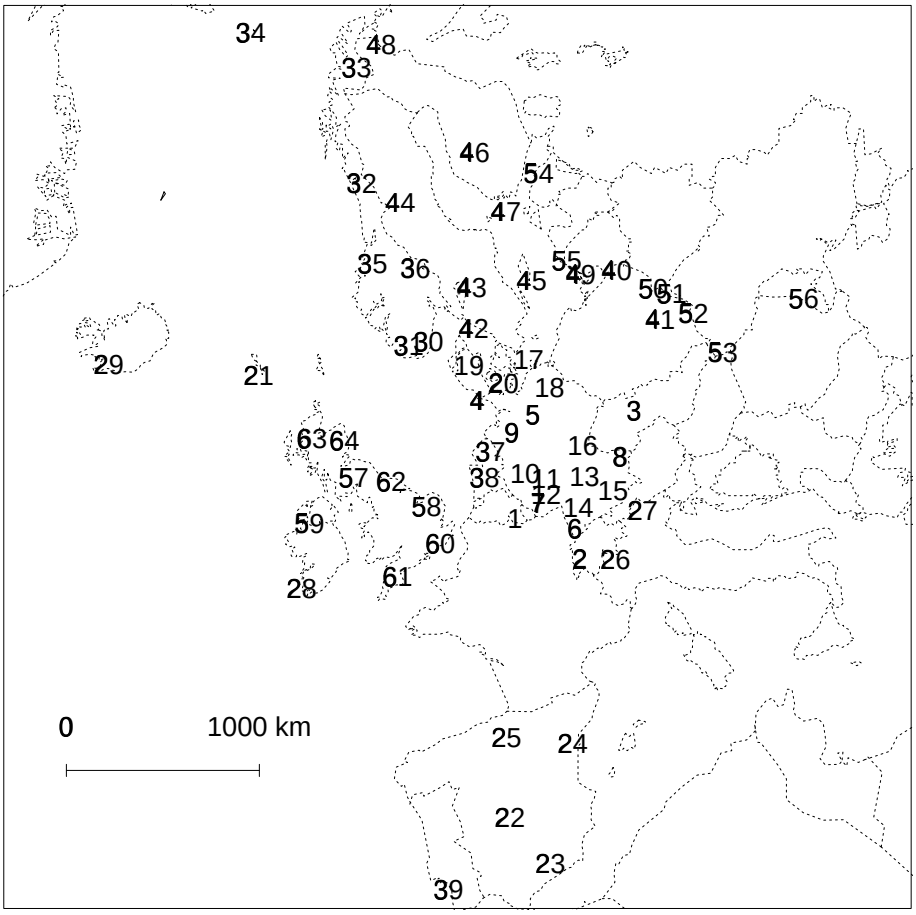

Figure 7: *Study area with data locations*

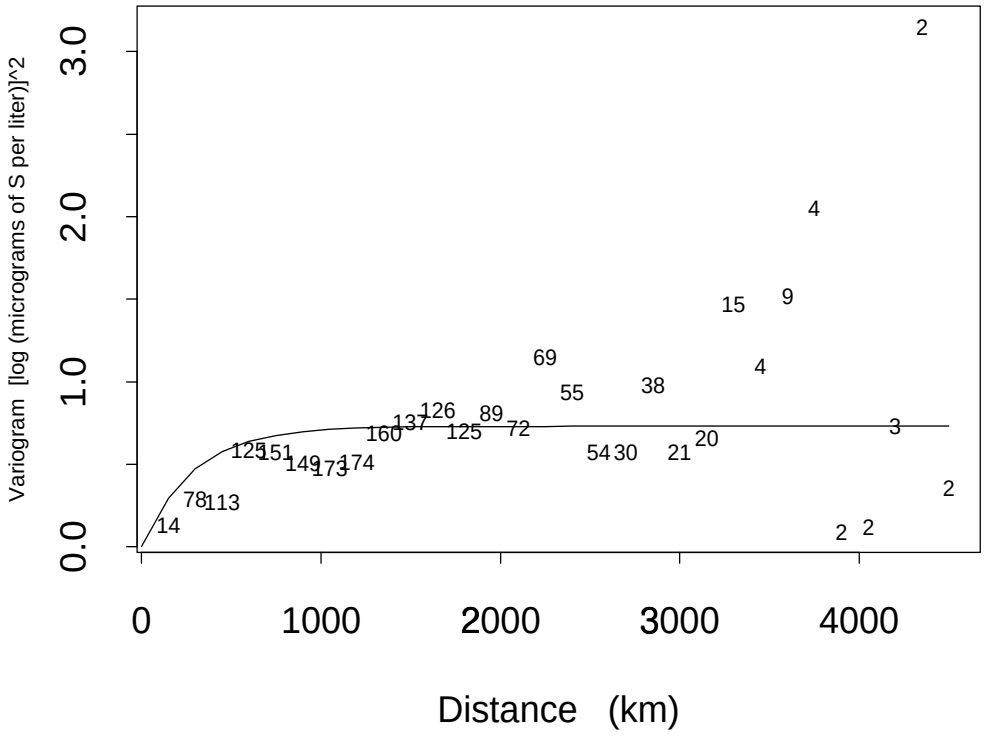

Figure 8: *Fitted variogram and lag-averaged squared differences for the proposed model*

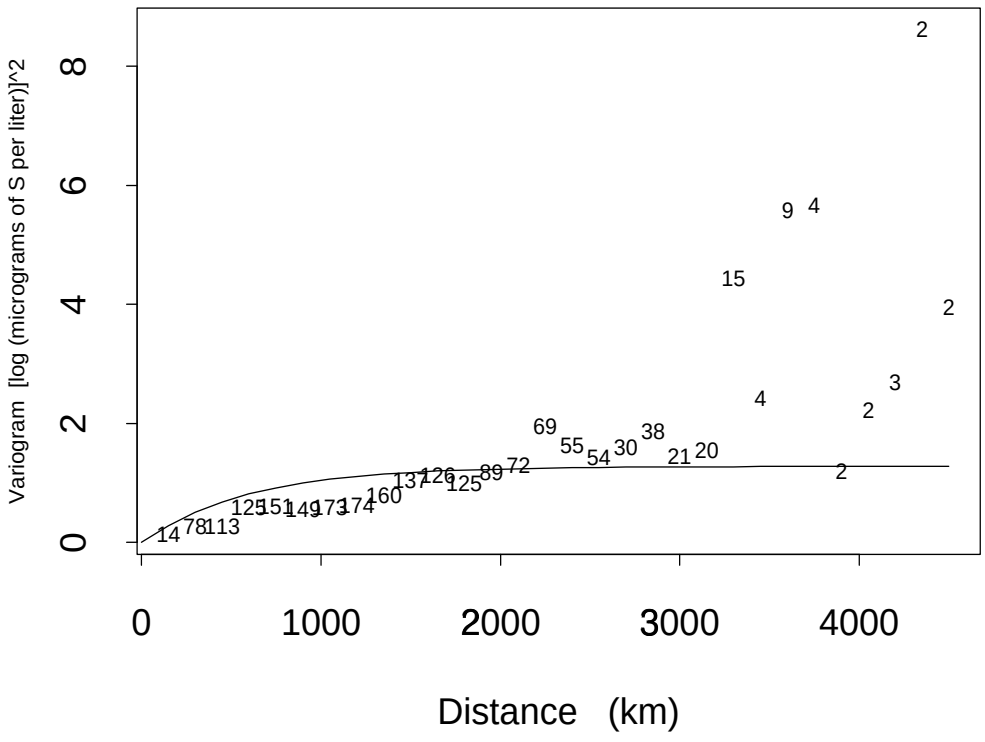

Figure 9: *Fitted variogram and lag-averaged squared differences for Kriging*

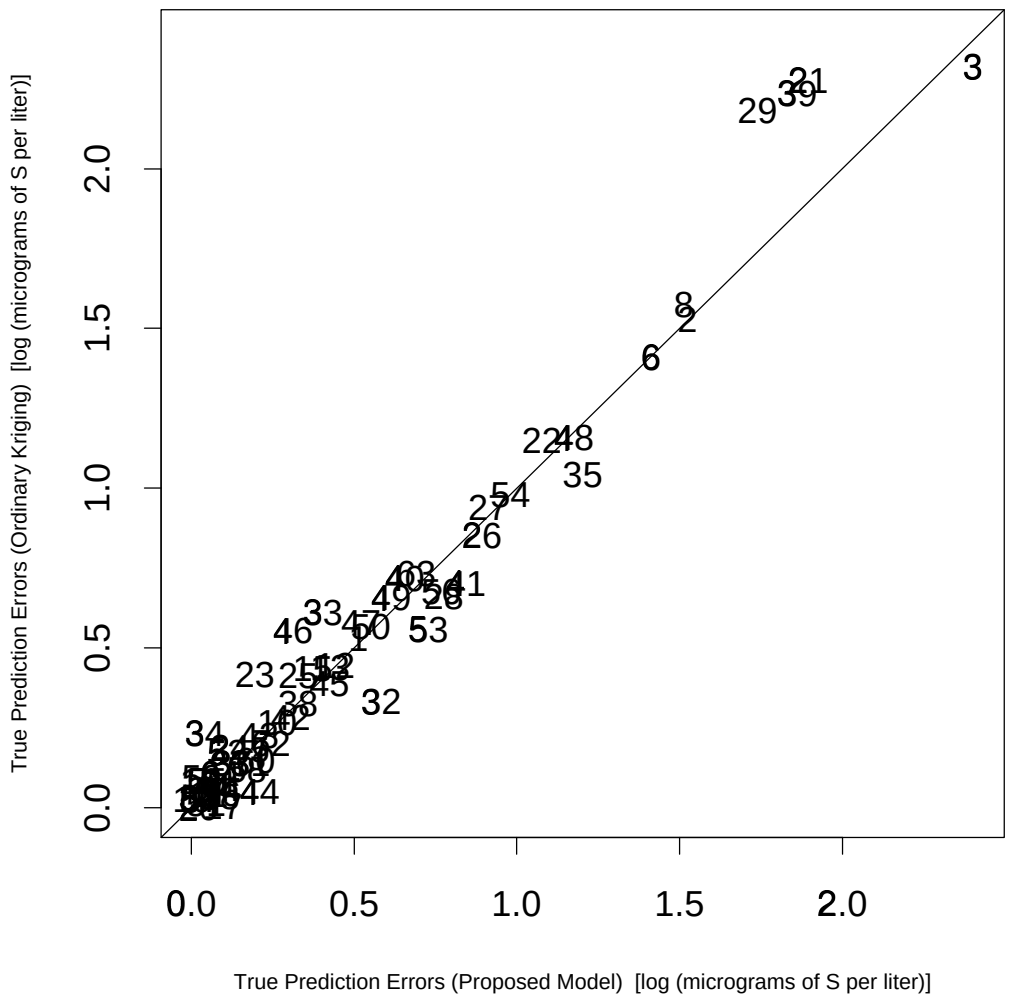

Figure 10: *True prediction errors for the proposed method and for the Kriging approach (The numbers in the figure refers to observation sites)*