

# Automatic Mapping in the Presence of Substitutive Errors: A Robust Kriging Approach

Baptiste FOURNIER

*Chair of Applied Statistics*  
*Swiss Federal Institute of Technology*  
*Lausanne, Switzerland*  
[baptiste.fournier@epfl.ch](mailto:baptiste.fournier@epfl.ch)

Reinhard FURRER

*Geophysical Statistics Project*  
*National Center for Atmospheric Research*  
*Boulder, Colorado*  
[furrer@ucar.edu](mailto:furrer@ucar.edu)

Interpolation of a spatially correlated random process is used in many scientific domains. The best unbiased linear predictor (BLUP), often called kriging predictor in geostatistical science, is sensitive to outliers. There are a few attempts to robustify the kriging predictor, however none of them is completely satisfactory. In this article, we present a new robust linear predictor for a substitutive error model. First, we derive a BLUP which is computationally very expensive even for moderate sample sizes. A forward search type algorithm is used to derive the predictor resulting in a linear likelihood-weighted mean predictor that is robust with respect to substitutive errors. Monte Carlo simulations support the theoretical results. The new predictor is applied to the two SIC 2004 data sets and is evaluated with respect to automatic interpolation and monitoring.

*Keywords: Substitutive errors; Universal kriging; Robustness; Forward Search; Likelihood.*

## 1 Introduction

The need for automatic interpolation or mapping algorithms of monitoring networks has greatly increased over the last few years. On the one hand, there are many automatic monitoring networks providing a steady flow of data. On the other hand, it is imperative to react in the shortest amount of time possible in case of anomalies.

Such anomalies, “outliers” or values that are not in agreement with the remaining part of the data, essentially have two sources. They might be erroneous measurements coming from a defective device, bad data transcription, etc. Or, they are real, “true” values observed from a different physical system, such as values measured after accidents releasing radioactivity.

An ideal automatic mapping algorithm has to address both types of anomalies and goes far beyond standard spatial outlier detection (*e.g.* Haining, 1990; Haslett *et al.*, 1991). In this article we propose a new method which copes with the first type of outliers. The goal is to develop a new automatic mapping algorithm that is insensitive to outliers and produces maps reflecting as close as possible the true but unknown process. As a byproduct, the algorithm identifies possible outliers or anomalies. Knowing the locations of outliers and the values of what should normally be observed, decision makers have to evaluate and distinguish between possible reasons for the observed anomaly. It would be far too risky to allow a simple algorithm to decide if a value is wrongly reported or if there are actual accidents involved. Recall that the late discovery of the ozone depletion over Antarctica was mainly due to a computer program designed to discard sudden, large drops in ozone concentrations as “errors”<sup>1</sup>.

---

<sup>1</sup><http://www.nas.nasa.gov/About/Education/Ozone/history.html>

In this article, we assume that the observations are realizations of an underlying physical process  $Y(\cdot)$  over some domain  $\mathcal{D}$ , for which we suppose an additive structure:

$$Y(s) = T(s) + Z(s), \quad s \in \mathcal{D}, \quad (1)$$

where  $T(\cdot)$  is a large scale variation (trend or drift) and  $Z(\cdot)$  is a stationary second-order process (Cressie, 1993, pages 112-113). The process  $Z(\cdot)$  is assumed to be a Gaussian spatially correlated process with substitutive outliers (see next section for more details). We will first estimate the mean structure and then predict the field according the second-order residual process. The robust kriging approach presented here is concerned with the second step. Of course we need to assure that trend and correlation estimation are insensitive to outliers.

This article is structured as follows. Section 2 introduces the substitutive error model and derives the robust kriging predictor in the case of constant mean structure. We shortly discuss the robust techniques used for general trend functions and unknown covariance structures in Section 3. We illustrate the performance of the new robust predictor with a simulation study in Section 4. Section 5 applies the algorithm to the SIC2004 data and compares the performance of the presented approach with some other interpolators. In Section 6 we discuss the proposed robust kriging predictor and outline possible extensions, which are part of our current research. Given the context of SIC2004 we refer to Dubois and Galmarini (2005) for a discussion of the data sets.

## 2 Robust Kriging for a Substitutive Error Model

This section presents the mathematical development of our method. Basic knowledge in statistics allows any reader to understand the main idea as we essentially only use Gaussian multivariate distributions and conditional expectations. To simplify the development of the proposed robust predictor, we assume for the time being a second order stationary process with a known constant mean and known covariance structure. In this case it is not necessary to explicitly introduce an additive structure as in (1). It is sufficient to assume that  $Z(\cdot)$  is a second order stationary random process with

$$E(Z(s)) = \mu, \quad \forall s \in \mathcal{D}$$

and

$$\text{Cov}(Z(s_1), Z(s_2)) = \mathcal{C}(s_1 - s_2), \quad \forall s_1, s_2 \in \mathcal{D}.$$

However, we will also work with the variogram

$$2\gamma(s_1 - s_2) = \text{Var}(Z(s_1) - Z(s_2)) = 2\mathcal{C}(0) - 2\mathcal{C}(s_1 - s_2), \quad \forall s_1, s_2 \in \mathcal{D}.$$

### 2.1 Substitutive Error Model

The main idea for the construction of our model is that we do not observe the process of interest  $Z(\cdot)$  directly, but a contaminated version thereof. The observed process is given by:

$$X(s) = (1 - B(s)) Z(s) + B(s) C(s).$$

The process  $C(\cdot)$  corresponds to the contamination process and the process  $B(\cdot) \in \{0, 1\}$  represents what we call the contamination scenario process, which indicates wheter a given site is contaminated or not. In what follows, denote by  $\mathbf{Z}$  the  $n$ -vector containing  $Z(s_1), \dots, Z(s_n)$  and similarly for the other processes. We assume a Gaussian process, *i.e.*

$$\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with  $\boldsymbol{\mu}$  the  $n$ -vector containing  $\mu$ ,  $\boldsymbol{\Sigma}_{ij} = \mathcal{C}(s_i - s_j)$  and  $\boldsymbol{\Sigma}_{ii} = \sigma^2 = \mathcal{C}(0)$ ,  $\forall i = 1, \dots, n$ , and

$$\mathbf{C} \sim \mathcal{N}(\boldsymbol{\mu}, k^2 \sigma^2 \mathbf{I}),$$

where  $k^2 \gg 1$ ,  $\mathbf{I}$  is the identity matrix in  $\mathbb{R}^n$ . Further, we suppose that  $Z(\cdot)$  and  $C(\cdot)$  are orthogonal. Note that the expectations of the two processes are assumed to be identical and that the contamination scale factor  $k$  guarantees that the tails of  $C(\cdot)$  are wider (but still normal) than those of  $Z(\cdot)$ .

The contamination scenario process  $B(\cdot)$  is also supposed to be orthogonal to  $Z(\cdot)$  and  $C(\cdot)$ . Its distribution is Bernoulli:

$$B(s) \sim \text{Bernoulli}(\epsilon),$$

where  $\epsilon$  is the contamination rate of our model. Note that we do not assume any spatial dependence for the process  $B(\cdot)$ , that the contamination is noncontagious across space.

## 2.2 Linear Predictor

If we want to predict the value of the process of interest  $Z(\cdot)$  at a new site  $s_0$  based on what we observe at sites  $s_1, \dots, s_n \in \mathcal{D}$  and if the considered loss function is quadratic, it is well known that the optimal predictor is given by the conditional expectation:

$$\mathbb{E}(Z(s_0) | \mathbf{X}).$$

In order to evaluate this expectation, consider

$$\mathbf{W} = \mathbf{X} | \mathbf{B} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}).$$

This process is by construction also Gaussian, with the same expectation  $\mu$ . Its variance is given by

$$\boldsymbol{\Omega} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A} + k^2 \sigma^2 (\mathbf{I} - \mathbf{A}),$$

where  $\mathbf{A} = \mathbf{I} - \text{diag}(\mathbf{B})$ . It follows that

$$Z(s_0) | \mathbf{W} \sim \mathcal{N}(\mu + \boldsymbol{\nu}^t \mathbf{A} \boldsymbol{\Omega}^{-1} (\mathbf{X} - \boldsymbol{\mu}), \sigma^2 - \boldsymbol{\nu}^t \mathbf{A} \boldsymbol{\Omega}^{-1} \mathbf{A} \boldsymbol{\nu}),$$

where the superscript  $t$  denotes the transpose and  $\boldsymbol{\nu} = \text{Cov}(Z(s_0), \mathbf{Z}) = (\mathcal{C}(s_0 - s_1), \dots, \mathcal{C}(s_0 - s_n))^t$ . The optimal predictor is then given by

$$\begin{aligned} \mathbb{E}(Z(s_0) | \mathbf{X}) &= \mathbb{E}(\mathbb{E}(Z(s_0) | \mathbf{X}, \mathbf{B})) \\ &= \mathbb{E}(\mathbb{E}(Z(s_0) | \mathbf{W})) \\ &= \mathbb{E}(\mu + \boldsymbol{\nu}^t \mathbf{A} \boldsymbol{\Omega}^{-1} (\mathbf{X} - \boldsymbol{\mu})) \\ &= \mu + \sum_{\mathbf{b} \in \{0,1\}^n} \omega_{\mathbf{b}} \boldsymbol{\nu}^t \mathbf{A}_{\mathbf{b}} \boldsymbol{\Omega}_{\mathbf{b}}^{-1} (\mathbf{X} - \boldsymbol{\mu}), \end{aligned} \tag{2}$$

where the weights  $\omega_{\mathbf{b}}$  are of the form  $\omega_{\mathbf{b}} = \text{P}(\mathbf{B} = \mathbf{b})$ . For the derivation of (2) we used the definition of conditional expectation and substituted with the expectations obtained above. Note that summation in (2) is over all possible  $2^n$  configurations  $\mathbf{b}$  of the Bernoulli process  $\mathbf{B}$ . A configuration  $\mathbf{b}$ , a  $n$ -vector containing zeros and ones, will be called a contamination scenario in the following. Even for moderately large  $n$ , the exact computation of this conditional expectation is unfeasible in a reasonable time frames. Therefore, we consider the predictor

$$\widehat{Z}(s_0) = \mu + \sum_{\mathbf{b} \in \mathcal{S}} \tilde{\omega}_{\mathbf{b}} \boldsymbol{\nu}^t \mathbf{A}_{\mathbf{b}} \boldsymbol{\Omega}_{\mathbf{b}}^{-1} (\mathbf{X} - \boldsymbol{\mu}), \tag{3}$$

where  $S$  is a suitable subset of contamination scenarios and the weights  $\tilde{\omega}_{\mathbf{b}}$  are based on the likelihood:

$$\tilde{\omega}_{\mathbf{b}} \propto P(\mathbf{B} = \mathbf{b})P(\mathbf{X} = \mathbf{x}), \quad (4)$$

where  $\mathbf{x}$  is the observation vector. Among all possible  $2^n$  weights  $\tilde{\omega}_{\mathbf{b}}$ , a few are much bigger than the remaining bulk of weights. In other words

$$1 \approx \sum_{\mathbf{b} \in S} \tilde{\omega}_{\mathbf{b}} \gg \sum_{\mathbf{b} \notin S} \tilde{\omega}_{\mathbf{b}} \approx 0. \quad (5)$$

In order to obtain an unbiased estimator, we normalize the likelihood weights  $\tilde{\omega}_{\mathbf{b}}$  such that  $\sum_{\mathbf{b} \in S} \tilde{\omega}_{\mathbf{b}} = 1$ .

The crucial point in our approach is therefore to determine a suitable subset  $S$  of the most likely contamination scenarios corresponding to the largest weights.

### 2.3 Determination of the Subset $S$

In order to determine the subset  $S$  of contamination scenarios we proceed in two steps:

**Step 1:** identify the most likely (initial) contamination scenario,

**Step 2:** find a subset of contamination scenarios “close” to the scenario of Step 1.

For the first step, we use the method proposed by [Cerioli and Riani \(1999\)](#), which is based on a forward search algorithm. This technique orders the observations beginning with those most in agreement with the specified autocorrelation model to those least in agreement with it. At each step of the algorithm, we predict at unordered sites, based on the already ordered observations. The observation at the site with the smallest resulting standardized prediction residual joins the already ordered observations. At the end, we obtain an order of all observations.

Now, we have to decide which portion of the ordered observations are contaminated. In other words, from which point on, do we declare the observations as being contaminated. To this end, [Cerioli and Riani \(1999\)](#) consider plots of functions of the ordered prediction residuals. The first outlier in the sample should be preceded by a peak in this plot. Since we have to produce an automatic algorithm, it is necessary to add an automatic threshold decision procedure to the forward search algorithm. We consider the same series of ordered prediction residuals as [Cerioli and Riani \(1999\)](#) and differentiate it, applying a back-shift operator. Based on the series constructed in this way, which we call  $(e_m)$ ,  $m < n$ , we use the two following rules to decide if an observation is contaminated or not.

**Rule I:** The observation which corresponds to  $e_m$  is suspicious if

$$e_m > q\hat{F}_{(e_1, \dots, e_{m-1})}(1 - \alpha) \quad \text{for a well chosen } \alpha,$$

where  $\hat{F}_{(e_1, \dots, e_{m-1})}$  is the empirical distribution function.

**Rule II:** The first suspicious observation (in the forward search order) and all the following observations are declared outliers.

We proceed to Step 2 of our search for a suitable subset  $S$  using the initial scenario of contamination found by the forward search procedure. All possible scenarios can be seen as vertices of the  $n$ -dimensional unit hypercube  $\{0, 1\}^n$ , each of which has an associated weight given by (4). Our initial contamination scenario is one of these vertices and the first element of the subset  $S$ . We compute the weights associated to its  $n$  first order neighbor vertices and add to  $S$  the vertices having bigger or equal weights than the vertex corresponding to the initial scenario. We repeat the same procedure for all

vertices contained in  $S$ . For each selected vertex, we consider its  $n$  first order neighbor vertices adding those neighbor vertices to  $S$ , which have bigger or equal weights than the selected vertex. This operation is repeated until the cardinality of the subset  $S$  is stabilized, *i.e.* all neighbor vertices have smaller associated weights.

## 2.4 Prediction Intervals

The linear predictor (3) has a Gaussian distribution. Therefore, it is straightforward to determine prediction intervals:

$$\left[ \pm q\mathcal{N}(1 - \alpha)\widehat{\text{MSE}}(\hat{Z}(s_0))^{1/2} \right], \quad 0 < \alpha < 1, \quad (6)$$

where  $\widehat{\text{MSE}}(\hat{Z}(s_0))$  is the estimation of the mean squared prediction error (or kriging variance) given below. To simplify the notation, suppose that the process  $Z(\cdot)$  has a zero mean structure, then (3) can be written as

$$\hat{Z}(s_0) = \sum_{b \in S} \lambda_b^t \mathbf{X}, \quad (7)$$

where  $\lambda_b^t = \tilde{\omega}_b \boldsymbol{\nu}^t \mathbf{A}_b \boldsymbol{\Omega}_b^{-1}$ . Denoting  $\boldsymbol{\lambda} = \sum_{b \in S} \lambda_b$ , we have

$$\begin{aligned} \text{MSE}(\hat{Z}(s_0)) &= \text{E}((\boldsymbol{\lambda}^t \mathbf{X} - Z(s_0))^2) \\ &= \boldsymbol{\lambda}^t \text{Var}(\mathbf{X}) \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^t \text{Cov}(Z(s_0), \mathbf{X}) + \text{Var}(Z(s_0)) \\ &= \boldsymbol{\lambda}^t \left( \sum_{b \in \{0,1\}^n} \omega_b \boldsymbol{\Omega}_b \right) \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^t \left( \sum_{b \in \{0,1\}^n} \omega_b \mathbf{A}_b \right) \boldsymbol{\nu} + \sigma^2. \end{aligned}$$

As for the exact predictor itself, the exact mean squared prediction error is computationally too expensive to evaluate. The weights corresponding to the unselected scenarios are assumed to be negligible compared to those contained in  $S$ , *cf.* equation (5). Therefore, the mean squared prediction error can be approximated by

$$\widehat{\text{MSE}}(\hat{Z}(s_0)) = \boldsymbol{\lambda}^t \left( \sum_{b \in S} \omega_b \boldsymbol{\Omega}_b \right) \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^t \left( \sum_{b \in S} \omega_b \mathbf{A}_b \right) \boldsymbol{\nu} + \sigma^2. \quad (8)$$

## 3 General Mean Trend and Unknown Covariance Structure

The kriging predictor was developed under the hypotheses of a known constant mean and known covariance structure. In reality however, this is rarely the case. Assuming an additive structure of the mean as in (1) and a second order stationary residual process, we will proceed as follows. We fit the trend surface with a robust method. The resulting residuals are used to estimate and fit the variogram. In the context of SIC2004, the prior information is used to tune the remaining parameters, such as the variogram structure and the initial starting values of various minimization algorithms.

### 3.1 Trend estimation

To estimate the mean structure  $T(\cdot)$  of the additive model (1), we fitted a local regression model (Cleveland, 1993) to the observations. This regression model allows for the common situation where the errors have a distribution with tails that are stretched out compared with normal tails. The general form of the function  $T(\cdot)$  is polynomial. The explanatory variables are the coordinates of the sites. We impose a conditionally linear

surface, meaning that given one of the two coordinates, the surface is a linear function of the remaining coordinate. Based on this, we obtain a smooth surface estimate.

Note that the correlation structure in  $Z(\cdot)$  is not taken into account when estimating the trend surface. As the mean captures the large scale variation this should not influence the estimation. One could also envision an iterative approach but we believe that there is little gain in doing so.

### 3.2 Variogram Estimation and Fitting

Based on the residuals of the trend estimation we estimate and fit a valid variogram function. To remain in the robust framework, we estimate the dependence structure with the method proposed by [Genton \(1998\)](#), based on the robust  $Q_h$  scale estimator ([Rousseeuw and Croux, 1993](#)). For a given lag  $h$ , define the process of differences  $V(h) = Z(s+h) - Z(s)$ . Since we do not have gridded locations we use standard binning to obtain the sequences  $V_1(h), \dots, V_{N_h}(h)$  for each specific lag  $h$ . The empirical variogram at lag  $h$  is defined as the following scaled  $\ell$ th order statistic

$$(2\hat{\gamma}(h))^{1/2} = Q_h = 2.2191 \cdot \{ |V_i(h) - V_j(h)| ; i < j \}_{(\ell)},$$

where  $\ell = \binom{m}{2}$  with  $m = \lfloor N_h/2 \rfloor + 1$  (see also [Genton and Furrer, 2003](#)). The factor 2.2191 guarantees consistency for the Gaussian distribution. This highly robust scale estimator has a 50% breakdown-point even for asymmetric distributions and a smooth influence function ([Rousseeuw and Croux, 1993](#)).

Finally, sill, nugget effect and range can be obtained from the empirical variogram with a weighted least squares technique ([Cressie, 1985](#)).

## 4 Simulation

In order to assess the quality of the new robust predictor, we compare its performance with several other predictors. We assume a Gaussian process with underlying spherical covariance structure with sill 1 and nugget effect 0.05 (*e.g.* [Cressie, 1993](#)). We vary the range between 4, 6 and 8. The contamination rate  $\epsilon$  varies between 0%, 5%, 15% or 25%, with contamination scale factor  $k^2 = 10$ . We consider a data set of  $n = 100$  regular locations within the square  $\mathcal{D} = [1, 10]^2$ . For each sample, we randomly select a location of the grid for the prediction. The variogram is estimated with the robust estimator introduced by [Cressie and Hawkins \(1980\)](#) and fitted with least squares ([Cressie, 1985](#)). For 200 samples, we calculate the mean squared error between  $\hat{Z}(s_0)$  and  $Z(s_0)$ .

The performance of our predictor, denoted with RK, is compared with five other methods:

FSK : our method without the neighborhood search of Step 2, *i.e.* considering the scenario found with the forward search algorithm and the automatic threshold procedure (Rule I and II);

HCK : robust kriging as developed by [Hawkins and Cressie \(1984\)](#);

WLD : a weighted least absolute deviation, where the weights are given by the covariance function  $\mathcal{C}(\cdot)$ ;

LMN : a local mean of the fourth nearest neighbors;

LMD : a local median of the fourth nearest neighbors.

Table 1 summarizes the mean squared error of the six interpolation techniques. The smallest mean squared error of each simulation setup is in bold and the second smallest in italics. Overall, the presented predictor performs well, especially for contamination ratios smaller than 25%. This is partially due to the fact that the influence of the outliers on the estimated variograms is too important otherwise.

We performed a similar simulation study on a naively small grid of size 9. For this case, we can evaluate the exact predictor (2) as the enumeration of all contamination scenarios is feasible. The exact predictor had the smallest mean squared error for all cases considered here.

## 5 Results with SIC2004 Data Sets

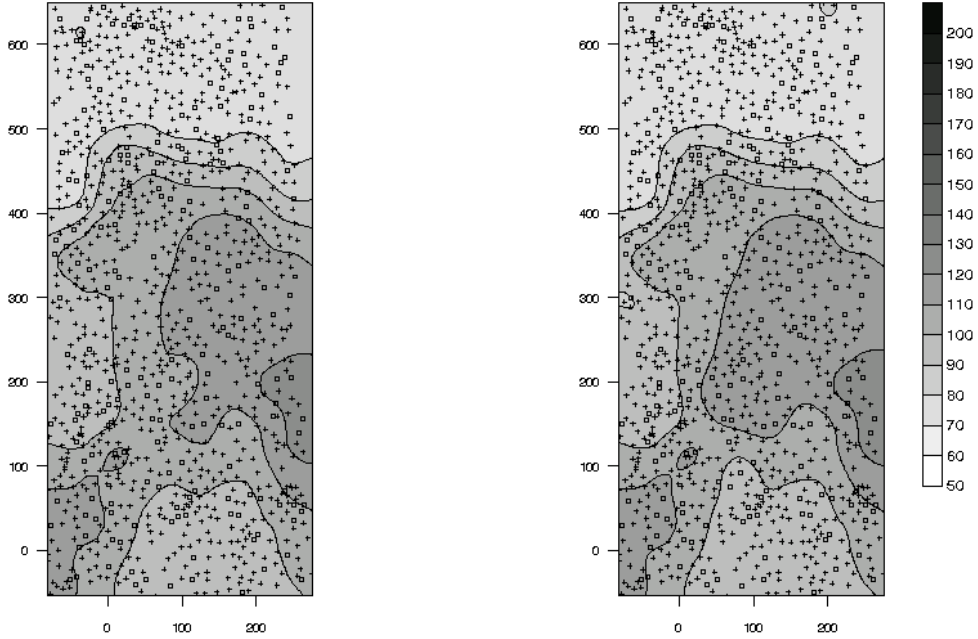
The robust kriging algorithm has been programmed with the freely available computer software R (Ihaka and Gentleman, 1996; R Development Core Team, 2004). R (“GNU S”) is similar to the S system, which was developed at Bell Laboratories by John Chambers and coauthors. It provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.). The source code used in this analysis can be obtained from the authors.

### 5.1 Robust Kriging Results

In the following, we discuss the results of the robust kriging predictor applied to the SIC2004 data sets. The ten training data sets serve as prior information to set several model parameters and initial values for minimization algorithms. For the trend estimation we use the R function `loess` with smoothing parameter 0.75. The function does not extrapolate outside an axis-aligned hypercube enclosing the original data. For the 17 points concerned by this, the fitted trend is a local mean obtained from the three nearest neighbors. The residuals of the training data indicate a slight anisotropy in the east-west direction. Therefore, we scale the east-west coordinate with a factor of 1.1. An exponential variogram seems to be adequate for all training data sets and we impose

			Methods					
			FSK	RK	HCK	WLD	LMN	LMD
Range	4	Contami. 0%	0.34	<b>0.32</b>	0.35	0.54	<i>0.34</i>	0.38
		Contami. 5%	<i>0.41</i>	<b>0.40</b>	0.41	0.58	0.42	0.41
		Contami. 15%	0.50	0.48	<i>0.43</i>	0.65	0.56	<b>0.42</b>
		Contami. 25%	0.80	0.67	<i>0.63</i>	<b>0.62</b>	0.96	0.75
	6	Contami. 0%	0.26	<b>0.25</b>	0.28	0.56	<i>0.25</i>	0.30
		Contami. 5%	<i>0.26</i>	<b>0.25</b>	0.27	0.76	0.44	0.31
		Contami. 15%	0.43	<i>0.34</i>	<b>0.33</b>	0.50	0.61	0.35
		Contami. 25%	0.58	0.49	<b>0.43</b>	1.03	0.76	<i>0.43</i>
	8	Contami. 0%	0.22	<i>0.21</i>	0.23	0.53	<b>0.21</b>	0.23
		Contami. 5%	<i>0.26</i>	<b>0.25</b>	0.27	0.63	0.30	<i>0.26</i>
		Contami. 15%	0.34	0.29	<b>0.25</b>	0.50	0.52	<i>0.28</i>
		Contami. 25%	0.54	0.39	<b>0.33</b>	0.73	0.70	<i>0.38</i>

**Table 1:** Comparison of the mean squared errors of the different predictors. See text for detailed simulation setup and description of the predictors. The smallest value of each simulation setup is in bold and the second smallest in italic.



**Figure 1:** *Isoline levels ( $nSv/h$ ) for the first data set (left) and the second data set (right).*

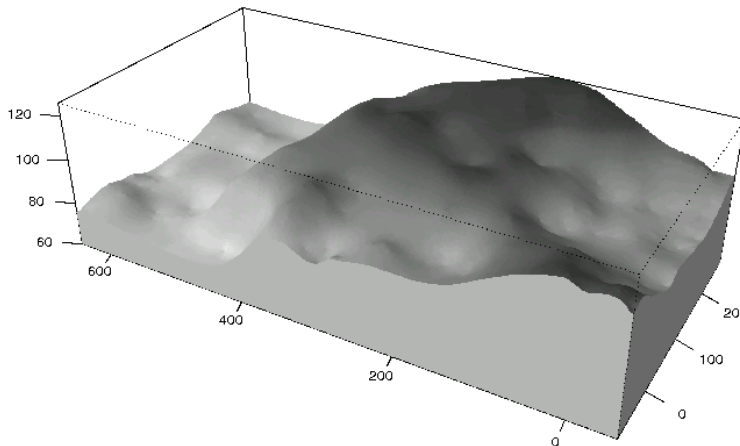
this variogram structure for both data sets in the comparison. One might argue, that a Matérn structure might be more appropriate (Matérn, 1960, 1986; Handcock and Wallis, 1994; Stein, 1999). However, fitting its smoothness parameter is chronically problematic and does often require human intervention, which is incompatible with our objectives. The contamination scale factor is set to  $k^2 = 9$ . Finally, the level of the empirical quantile considered for the automatic threshold procedure of contaminated observations detection (Rule I) is set to  $1 - \alpha = 0.9$ .

Tables 2 and 3 compare the observations, the predicted/estimated values and the errors. The predicted values for the first and second data set are very similar. The minor difference are in part due to differences in the trend estimation. The robust trend and variogram estimation are slightly affected by the large values of the second data set. The fitted variogram parameters for the first and second data sets are (range, sill, nugget) = (41908.5, 90.5, 54.6) and (52319.3, 112.4, 74.7), respectively. By nature of the robust predictor, the two large values of the second data set are classified as outliers and thus the peak is not reproduced, clearly reflected in the error statistics of Table 3. This is also shown by the isoline map of the predicted values in Figure 1 or by Figure 2.

The mean squared prediction error approximation (8) is an indication of the associated prediction uncertainty. The associated isoline maps are shown in Figure 3. The larger nugget effect and sill for the second data set translates into slightly larger uncertainties.

$N = 808$	Min.	Max.	Mean	Median	Std. Dev.
Observed (first data set)	57.00	180.00	98.02	98.80	20.02
Estimates (first data set)	70.74	124.60	96.70	99.54	14.30
Observed (second data set)	57.00	1528.00	105.40	98.95	83.71
Estimates (second data set)	70.33	125.30	96.83	99.48	14.58

**Table 2:** *Comparison of the estimated and measured values ( $nSv/h$ ).*



**Figure 2:** 3D map showing the predicted values of the second data set (vertical scale is in nSv/h).

The exercise was carried out on a Linux powered 2.6 GHz Xeon processor with 2 Gbytes RAM. The total calculation time was 662 seconds for the first data set and 218 seconds for the second one. The considerable difference between the processing times is due to the fact that the outliers are more easily detectable in the second data set and the subset of considered contamination scenarios is of smaller size. Hence, we have a mean time of 440 seconds, or seven minutes and twenty seconds. The trend estimation, the estimation and fitting of the variogram represent a small part of the processing time. They are almost instantaneous compared to the robust kriging estimator developed in Section 2. Determining an initial contamination scenario with the forward search algorithm and the threshold decision procedure represents 20% of the time, determining the subset of contamination scenarios 69% and predicting 10%. Although being computationally expensive the robust kriging approach can be carried out on any reasonable sized desktop computer. See also Table 6 in the next section.

The algorithm should provide an estimated value at locations with observations. If the corresponding prediction/confidence interval based on (6) does not contain the observation a risk manager needs to evaluate the validity of the observation.

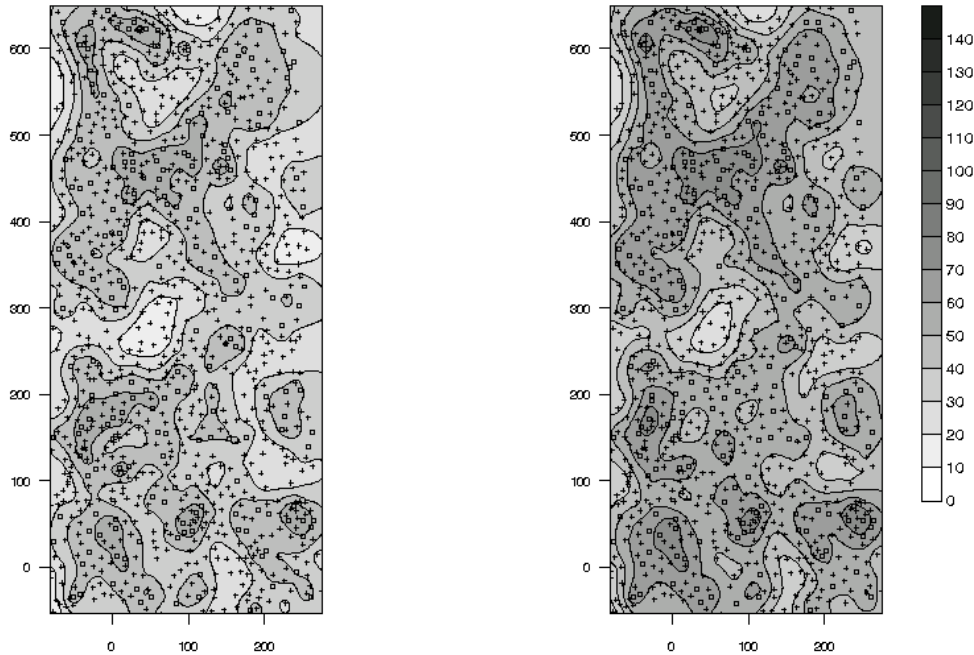
## 5.2 Comparison with Other Methods

Applying the robust methods HCK, WLD or LMD to the two data sets, we obtain similar prediction and error summaries as presented in Tables 2 and 3. The key ingredients are robust techniques to estimate the trend and the variogram parameters. If any of these two are exchanged with classical approaches, prediction results would be significantly different.

Not only robust prediction procedures but also some classical interpolation methods suffer from the drawback of being smoothing methods when trying to predict the anomaly introduced in the second data set. They smooth the anomaly out as illustrated

Data sets	MAE	ME	Pearson's $r$	RMSE
First data set	9.06	-1.32	0.79	12.43
Second data set	16.22	-8.58	0.27	81.44

**Table 3:** Comparison of the errors.



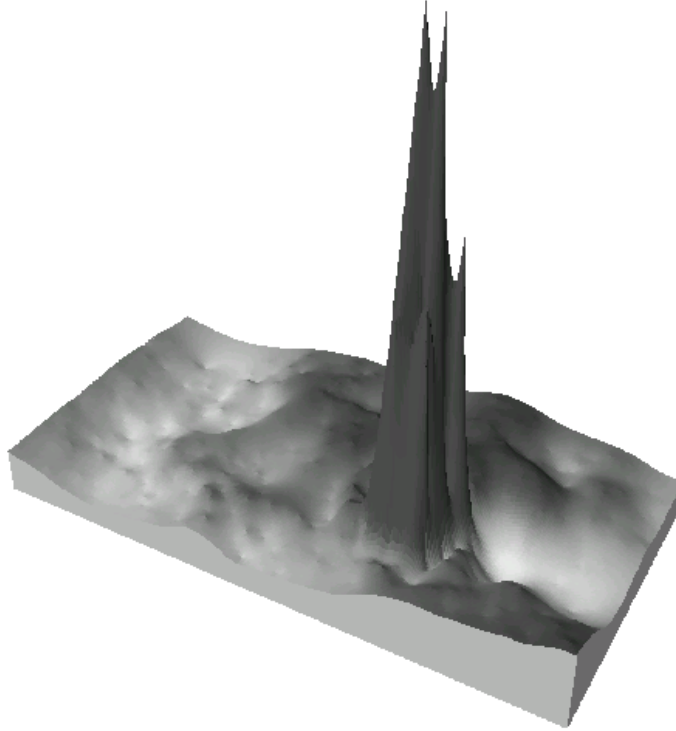
**Figure 3:** Isoline levels showing the uncertainty associated to the estimations obtained for the first data set (left) and the second data set (right).

in Figure 2. To reproduce the anomaly peak a local non-robust interpolation technique needs to be used. For illustration, we compare the results of our robust kriging approach with two local predictors.

The first is a nearest neighbor kriging technique (NNK) (Cressie, 1993, page 158) applied to the same residuals as for the robust predictor and the same covariance structure (*i.e.* we use the same prior information). For simplicity we fix the number of nearest neighbors used to eight. Furrer *et al.* (2005) show that this small number is sufficient with only a negligible loss of optimality. On a regular grid, considering eight neighbors means taking into account all first order and some of the second order neighbors. By increasing the number of neighbors we would get similar results as obtained with robust kriging. Associated prediction intervals are straightforward to obtain. Anomalies can be addressed by decision makers in a similar manner as discussed in the previous section. Nearest neighbor kriging has appealing properties, especially if the number of observations are large and not many points have to be predicted, but has several drawbacks as pointed out in Cressie (1993).

The second method is a local mean interpolator (LMN), also called global measure of central tendency (Cressie, 1993, page 370). A predefined adjacency matrix is constructed, where for each point to predict we determine the four nearest observations. The adjacency matrix is stored and considered as the prior knowledge. To predict at a site  $s_0$  we simply take the mean of the four nearest observations. Thus, prediction at a site conditional on the given neighbors has operation count four additions and one division. This predictor is not based on a stochastic model and requires no knowledge of spatial model parameters. In general, however, it is biased. Conservative prediction intervals are obtained by assuming Gaussianity and independence of the observations. Anomalies can be detected by comparing the observations with the ten observations from the prior data sets.

Tables 4 and 5 summarize the predicted values and errors. Note the striking similarity



**Figure 4:** 3D map showing extreme values found in the second data set with the local mean (LMN) approach (vertical scale is in  $nSv/h$ ).

for the first data set among the methods. The NNK approach reproduces partially the anomaly of the second data set. Although LMN captures most of the peak (see Figure 4), the RMSE is only reduced by 12% compared with RK. Figure 5 shows the associated uncertainty. Compared to the robust case, the MSE is considerably higher which is not surprising since the model does not take any spatial correlation into account and since the mean is taken over only four points.

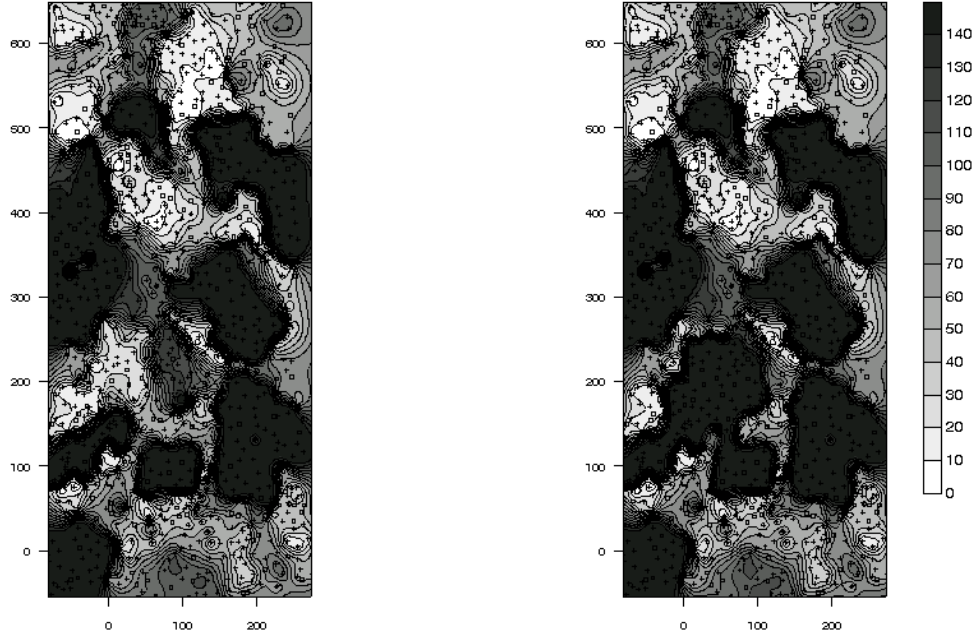
The three methods, *i.e.* robust kriging (RK), nearest neighbor kriging (NNK) and local mean (LMN) represent approaches with decreasing theoretical and computational complexity, reflected by the associated computing times given in Table 6. Given the fundamentally different nature of the predictors the choice of the prediction method should probably be made on the basis of reasonable model characteristics and not based on resulted error statistics as presented here.

Method	Min.	Max.	Mean	Median	Std. Dev.
RK	70.74	124.60	96.70	99.54	14.30
NNK	70.19	126.53	97.13	99.25	14.40
LMN	67.06	127.40	96.90	98.56	14.86
RK	70.33	125.30	96.83	99.48	14.58
NNK	70.14	553.32	105.20	99.81	46.60
LMN	66.50	718.38	105.30	99.00	57.07

**Table 4:** Comparison of the predicted values with robust kriging (RK), nearest neighbor kriging (NNK) and local mean (LMN) for the first (top) and second (bottom) data set ( $nSv/h$ ).

Method	MAE	ME	Pearson's $r$	RMSE
RK	9.06	-1.32	0.79	12.43
NNK	9.22	-0.89	0.78	12.51
LMN	9.29	-1.12	0.78	12.56
RK	16.22	-8.58	0.27	81.44
NNK	19.43	-0.22	0.48	73.50
LMN	19.44	-0.12	0.53	71.87

**Table 5:** Comparison of the errors with robust kriging (RK), nearest neighbor kriging (NNK) and local mean (LMN) for the first (top) and second (bottom) data set.



**Figure 5:** Isoline levels showing the uncertainty associated to the estimations obtained for the first data set (left) and the second data set (right) with the local mean (LMN) approach.

Action	RK	NNK	LMN	
Initializing, loading data	0.64	0.46	0.53	0.01
Trend estimation	0.04	0.04	0.06	—
Variogram estimation	0.52	0.51	0.48	—
Variogram fitting	0.05	0.05	0.06	—
Forward search	110.85	110.71	—	—
Contamin. scenarios	471.87	73.92	—	—
Prediction	74.73	29.25	—	—
Kriging total/Prediction	661.44	217.85	8.35	0.06
Finalizing	0.07	0.07	0.06	0.02
Total	662.76	218.98	9.50	0.09

**Table 6:** Computation time in seconds for robust kriging (RK), nearest neighbor kriging (NNK) and local mean (LMN) interpolator. For NNK and LMN only the time for the second data set is given (Linux, 2.6 GHz Xeon processor with 2 Gbytes RAM).

## 6 Discussion and Outlook

This paper introduces a new robust kriging approach for substitutive error processes. We derived the best linear unbiased predictor for a substitutive error model from the conditional expectation. Although being optimal, the resulting predictor is computationally unfeasible and we propose an appropriate approximation. Simulations confirm that the predictor is competitive among other robust methods.

Coded in the high-level programming language R, it takes a few minutes on a reasonable sized desktop computer to predict a spatial field with several hundred locations. Optimizing the programming code would result in considerable improvement of computation speed.

The robust kriging predictor is a fully automatic approach: no human intervention or decision is required for the mapping. However, decision makers are required to classify outliers as measurement errors or as genuine anomalies. According to the nature of the outliers corresponding decisions need to be made.

Given the first and second order structure of the process, the robust prediction algorithm has two parameters: the contamination scale factor  $k$  and the level of the empirical quantile  $1 - \alpha$ . The predictant is not sensitive to these values and we therefore used standard values (e.g. Huber, 1981). A cross-validation technique for the selection of the parameters is possible, but leads to a substantial increase of computing time.

When constructing the subset  $S$ , the criterion for neighbor vertices to be added could be relaxed to considering only the weight of the initial scenario. The result is a speed up of the algorithm but has substantial change with respect to the final prediction.

For completeness it should be mentioned that we could not find a simple relationship between the computing time for the contamination scenarios determination and the amount of outliers in the dataset.

The new predictor is not capable — by construction — to distinguish between an outlier and a large “true” value. Current research consists in expanding the substitutive error model to clustered contamination scenario processes and contamination processes with a spatial structure. Such a model should be able to detect contaminated regions and predict fields such as the second data set more accurately.

## Acknowledgments

The authors greatly acknowledge the help of Eva Maria Restle in the preparation of the manuscript. The first author would like to thank his thesis advisor Professor Stephan Morgenthaler for his support. The research of the second author was supported in part by the Geophysical Statistics Project at the National Center for Atmospheric Research under the National Science Foundation grants DMS-9815344 and DMS-0355474.

## References

- Cerioni, A. and Riani, M. (1999). The ordering of spatial data and the detection of multiple outliers. *Journal of Computational and Graphical Statistics*, **8**, 239–258. 4
- Cleveland, W. S. (1993). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836. 5
- Cressie, N. A. C. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, **17**, 563–586. 6
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons Inc., New York, revised reprint. 2, 6, 10

- Cressie, N. A. C. and Hawkins, D. M. (1980). Robust estimation of the variogram. I. *Journal of the International Association for Mathematical Geology*, **12**, 115–125. 6
- Dubois, G. and Galmarini, S. (2005). Introduction to the spatial interpolation comparison (SIC) 2004 exercise and presentation of the data sets. This Volume. 2
- Furrer, R., Genton, M. G., and Nychka, D. (2005). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, under revision. 10
- Genton, M. G. (1998). Highly robust variogram estimation. *Mathematical Geology*, **30**, 213–221. 6
- Genton, M. G. and Furrer, R. (2003). Analysis of rainfall data by robust spatial statistics using S+SpatialStats. In: “*Mapping Radioactivity in the environment. Spatial Interpolation Comparison 1997*”, Dubois, G., Malczewski, J., and De Cort, M. (Eds), pp. 118–129. EUR 20667 EN, Office for Official Publications of the European Communities, Luxembourg. 6
- Haining, R. P. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge. 1
- Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, **89**, 368–390. 8
- Haslett, J., Bradley, R., Craig, P., Unwin, A., and Wills, G. (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, **45**, 234–242. 1
- Hawkins, D. M. and Cressie, N. A. C. (1984). Robust kriging — a proposal. *Journal of the International Association for Mathematical Geology*, **16**, 3–18. 6
- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York. 13
- Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314. 7
- Matérn, B. (1960). *Spatial Variation: Stochastic Models and their Application to some Problems in Forest Surveys and other Sampling Investigations*. Meddelanden Fran Statens Skogsforskningsinstitut, Band 49, Nr. 5, Stockholm. 8
- Matérn, B. (1986). *Spatial Variation*. Springer-Verlag, Berlin, second edition. 8
- R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>. 7
- Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273–1283. 6
- Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer-Verlag, New York. 8