Hydrol. Earth Syst. Sci., 15, 2763–2775, 2011
www.hydrol-earth-syst-sci.net/15/2763/2011/
doi:10.5194/hess-15-2763-2011

Hydrology and Earth System Sciences

Interpolation of groundwater quality parameters with some values below the detection limit

A. Bárdossy

Institute of Hydraulic Engineering, University of Stuttgart, Stuttgart 70569, Germany

Received: 10 May 2011 – Published in Hydrol. Earth Syst. Sci. Discuss.: 26 May 2011
Revised: 19 August 2011 – Accepted: 22 August 2011 – Published: 1 September 2011

Abstract. For many environmental variables, measurements cannot deliver exact observation values as their concentration is below the sensitivity of the measuring device (detection limit). These observations provide useful information but cannot be treated in the same manner as the other measurements. In this paper a methodology for the spatial interpolation of these values is described. The method is based on spatial copulas. Here two copula models – the Gaussian and a non-Gaussian v-copula are used. First a mixed maximum likelihood approach is used to estimate the marginal distributions of the parameters. After removal of the marginal distributions the next step is the maximum likelihood estimation of the parameters of the spatial dependence including taking those values below the detection limit into account. Interpolation using copulas yields full conditional distributions for the unobserved sites and can be used to estimate confidence intervals, and provides a good basis for spatial simulation. The methodology is demonstrated on three different groundwater quality parameters, i.e. arsenic, chloride and deethylatrazin, measured at more than 2000 locations in South-West Germany. The chloride values are artificially censored at different levels in order to evaluate the procedures on a complete dataset by progressive decimation. Interpolation results are evaluated using a cross validation approach. The method is compared with ordinary kriging and indicator kriging. The uncertainty measures of the different approaches are also compared.

1 Introduction

There are many different chemicals which enter the groundwaterwater system through different mechanisms. Many of these compounds are known or suspected to have adverse effects on health and environment. Recent advances in laboratory techniques are providing improved capabilities for detecting large numbers of new and potentially harmful contaminants. The concentrations of different chemicals usually have strongly skewed distributions with a few very high values and a large number of low ones. Some of the low values are reported as non-detects due to the limited sensitivity of the laboratory instruments. The high skew and the occurrence of non-detects interpreted as values below a given threshold make the statistical and geostatistical analysis of these data unpleasant and complicated. The statistical treatment of censored data has a long history. Already Cohen (1959) had published a paper for the estimation of the normal distribution from censored data. Later work was performed for other distributions such as the 3 parameter lognormal distribution (Cohen, 1976). The first papers concentrated mainly on right censored (survival) data. In Helsel and Cohn (1988) left censored water quality data were analyzed. Despite recent works on the subject such as Shumway et al. (2002) the statistical treatment of censored environmental data is far less frequently applied as it could and should be Helsel (2005).

While the treatment of censored environmental data from the classical statistical viewpoint is reasonably well developed this is not the case in spatial statistics. Spatial mapping of variables with censored data is also of great interest and practical importance.

Correspondence to: A. Bárdossy
(bardossy@iws.uni-stuttgart.de)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Recently Sedda et al. (2010) presented a methodology to reflect censored data using a simulation approach. In Saito and Goovaerts (2000) the authors addressed the problem of censored and highly skewed variables, and showed that the indicator approach outperforms other geostatistical methods of interpolation.

Variables with non-detects are usually highly skewed which makes their interpolation even more difficult. The high skew of the distributions often leads to problems with the variogram or covariance function estimation. A few large values dominate the experimental curve, and outliers can lead to useless variograms. This problem is partly overcome by the use of indicator variables. However this approach suffers from other deficiencies as demonstrated in this paper.

The purpose of this paper is to develop a methodology to estimate spatial dependence structure from a mixed dataset containing differently censored data. The approach requires as a first step the estimation of the univariate distribution function of the variable under consideration. For this purpose a maximum likelihood method is used. In the next step the spatial dependence is described with the help of copulas, and the copula parameters are estimated using a maximum likelihood method. After this, the estimated dependence structure is used for the interpolation. Copulas are used in hydrology mainly for the analysis of extremes. In Keef et al. (2009) an interesting approach for treating missing data in a copula approach was presented.

The methodology is demonstrated using different water quality parameters obtained from large scale measurement campaigns in South-West Germany. Two highly censored parameters, namely arsenic and deethylatrazin are considered. In order to test the methodology a parameter with no censored data (chloride) is selected and subsequently artificially censored. The methodology is compared to ordinary and indicator kriging using different performance measures.

2 Methodology

2.1 Marginal distribution

The neagtive effects of highly skewed data on interpolation can be reduced by data transformations. Most frequently logariithmic or normal score transoformations are applied. However in the case of data below the detection limit these transormations cannot be used in a straightforward manner. The logarithm of the values below the detection limit requires an arbitrary choice. The normal score transformation suffers from the incomplete order of the data. For example if one has observations of $1.5\,\mathrm{mg\,l^{-1}}$ $3\,\mathrm{mg\,l^{-1}}$ and *below* $2\,\mathrm{mg\,l^{-1}}$ and *below* $1\,\mathrm{mg\,l^{-1}}$ then even the rank of the *exact* values cannot be determined. To treat this problem the distribution function of the studied variable $G(z)$ is estimated first.

Assume that there are n_d measurements with values below the detection limit d_i (note that the detection limits might differ from place to place, possibly due to different laboratory equippment used), and for n_z observations a measurement value z_j is given. The empirical distribution function of such observations can only be calculated for values above the largest detection limit. Due to the censoring the mean and the standard deviation cannot be calculated directly, thus the estimation of the parameters θ of a selected parametric distribution via method of moments is not possible. Instead a maximum likelihood method is required. Here one has two choices:

1. To assume a parametric distribution function stationary over the whole domain, and to assess the parameters via maximum likelihood
2. To assume a mixed distribution: for values below a threshold a parametric form is assumed and for those observations above the threshold, an empirical or a non parametric distribution is considered.

While the first approach is more or less straightforward, it has a few shortcomings. One of them is that outliers might have a very important influence on the parameters of the distribution; the other is that the underlying distribution could be bimodal.

In the first case the distribution parameters θ can be estimated using the likelihood function:

$$L_{\mathrm{low}}(\ldots,d_i,\ldots,z_i\ldots|\theta)=\prod_{i=1}^{n_d}F(d_i|\theta)\prod_{j=1}^{n_z}f(z_j|\theta) \tag{1}$$

where $F(.|\theta)$ is the distribution function applied to those values below the local detection limit and $f(.|\theta)$ the corresponding density with parameter θ.

In the case of the mixed approach we assume that the values below a given threshold z_{lim} (greater or equal than all d_is) follow a parametric distribution, while above z_{lim} the empirical distribution should be considered. Thus the estimation is restricted to those which are below z_{lim}. If a random variable with distribution function $F(z)$ is restricted to the interval $(-\infty,b)$ than the distribution function of the restricted variable is $F_R(z)=F(z)/F(b)$. This fact is used by estimating the restricted variable via maximum likelihood.

$$L_{\mathrm{low}}(\ldots,d_i,\ldots,z_i\ldots|\theta)=\frac{1}{F(z_{\mathrm{lim}}|\theta)}\prod_{i=1}^{n_d}F(d_i|\theta)\prod_{z_j<z_{\mathrm{lim}}}f(z_j|\theta) \tag{2}$$

Here a parametric distribution is fit to the observations below z_{lim}. This enables an extrapolation of the distribution function into the low value domain. By selecting an upper bound the negative effect of outliers is eliminated. In both cases the logarithm of the likelihood function can be maximized. Above the z_{lim} value a distribution $F_{\mathrm{lim}}(z)$ is assumed:

$$F_{\mathrm{lim}}(z)=\frac{1}{n_{\mathrm{lim}}+1}\sum_{i=1}^{n_d+n_z}1_{z_{\mathrm{lim}}<z<z_i} \tag{3}$$

where $n_{\lim}$ is the number of z_i greater than $z_{\lim}$.

The overall distribution function is:

$$G(z) = \begin{cases} \frac{\alpha_{\lim}}{F(z_{\lim}|\theta)} F(z|\theta) & \text{if } z \le z_{\lim} \\ \alpha_{\lim} + (1-\alpha_{\lim}) F_{\lim}(z) & \text{if } z > z_{\lim} \end{cases} \quad (4)$$

where

$$\alpha_{\lim} = \frac{n_{\lim}}{n_z + n_d + 1}$$

The introduction of $\alpha_{\lim}$ enables a continuous transition from the theoretical to the empirical part of the distribution.

Note that the limit $z_{\lim}$ is not estimated, but selected as a reasonable limit which is certainly below possible outlier observations. In order to obtain an estimated θ value it is important to have a sufficient number of exact observations below $z_{\lim}$. Possible candidates for the parametric part of the distribution are the Gamma (including exponential) and the Weibull distributions. Information criteria can be used for the choice of the appropriate distribution.

2.2 Spatial structure identification

The identification of the spatial structure from data with many non-detects is a difficult problem. Non detects and *exact* values carry very different information. An arbitrary setting of the non-detects leads to a reduction of the variance and to a false strong dependence between the low values. Neglecting them for the spatial variability estimation on the other hand usually leads to an overestimation of the variance and to an underestimation of the strength of the spatial dependence.

In order to deal with this problem a stochastic model is required. We assume that the variable of interest corresponds to the realization of a random function. For our study we restrict the random function $Z(x)$ to a spatial domain Ω. As only a single realization is observed on a limited number of points, further assumptions on the random function have to be made.

The spatial stationarity assumption is that for each set of points $\{x_1, \ldots, x_k\} \subset \Omega$ and vector h such that $\{x_1+h, \ldots, x_k+h\} \subset \Omega$ and for each set of possible values $w_1, \ldots, w_k$:

$$\begin{aligned} &P(Z(x_1) < w_1, \ldots, Z(x_n) < w_k) \quad (5) \\ &= P(Z(x_1+h) < w_1, \ldots, Z(x_k+h) < w_k) \end{aligned}$$

The spatial variability of a field is usually determined from exact observations. Variograms and covariance functions can be calculated from measured values directly, but even different measurement methods with different accuracies cause problems in the structure identification. Measurements with higher error variances lead to higher nugget values. The detection limit problem makes the assessment of the spatial structure extremely difficult. Setting the values below the detection limit d_j to either 0 or d_j leads to a false marginal distribution and to a false spatial dependence structure. The indicator approach provides a reasonable alternative solution, by calculating indicator variograms for a large number of cutoffs.

In this paper a copula based approach is taken as described in Bárdossy (2006). Two copula models, the Gaussian and the v-transformed normal copula are considered. The model is described in detail in Bárdossy and Li (2008).

Following Eq. (6) we assume that the random function Z is such that for each location $x \in \Omega$ the corresponding random variable $Z(x)$ has the same distribution function F_Z for each location x. The joint distribution can be written with the help of the copula:

$$F_{x_1,\ldots,x_k}(w_1, \ldots, w_k) = C_{x_1,\ldots,x_k}((F_Z(w_1), \ldots, F_Z(w_k)) \quad (6)$$

with $C_{x_1,\ldots,x_k}$ being the spatial copula corresponding to the locations $x_1, \ldots, x_k$. This approach allows us to investigate the new variable $U(x) = F(Z(x))$ which has a uniform marginal distribution.

Two copula models, the Gaussian (normal) and the v-transformed normal copula, are considered. The Gaussian copula is described by its correlation matrix Γ.

The v-transformed normal copula is parametrized by the transformation parameters m, k and the correlation matrix Γ which is likely to differ from the Gaussian one.

The v-transformed copula is defined using $\mathbf{Y}$ being an n dimensional normal random variable $N(\mathbf{0}, \Gamma)$. All marginals are supposed to have unit variance. Let $\mathbf{X}$ be defined for each coordinate $j = 1, \ldots, n$ as:

$$X_j = \begin{cases} k(Y_j - m) & \text{if } Y_j \ge m \\ m - Y_j & \text{if } Y_j < m \end{cases} \quad (7)$$

Where k is a positive constants and m is an arbitrary real number. When $k = 1$ this transformation leads to the multivariate non centered χ-square distribution. All one dimensional marginals of $\mathbf{X}$ are identical and have the same distribution function.

The parameters of the spatial copula are estimated using the maximum likelihood method.

For the Gaussian copula, as a consequence of the stationarity assumption, the correlations between any two points can be written as a function of the separating vector $\mathbf{h}$. Then for any set of observations $x_1, \ldots, x_n$ the correlation matrix Γ can be written as:

$$\Gamma = \left((\rho_{i,j})_{l,l}^{n,n} \right) \quad (8)$$

where $\rho_{i,j}$ only depends on the vector $\mathbf{h}$ separating the points x_i and x_j:

$$\rho_{i,j} = R(x_i - x_j) = R(\mathbf{h}_{i,j}) \quad (9)$$

For the estimation, the observed values are transformed to the standard normal distribution using:

$$y_k = \Phi_1^{-1}(G(z(x_k))) \quad k = 1, \ldots, n_z \quad (10)$$

$$y_j^d = \Phi_1^{-1}\left(G(d(x_j))\right) \quad j = 1, \ldots, n_d \quad (11)$$

Here $\Phi_1(.)$ is the distribution function of the standard normal distribution N(0,1).

The variable y is now a censored normal with data below the detection limits denoted by y_j^d. The correlation function $R(.,\beta)$ is assumed to have a parametric form with the parameter vector β. The likelihood function in this case can be written as:

$$L(\beta) = \prod_{(j,k)\in I_1} \phi_2\big(y_j, y_k, R(\mathbf{h}_{j,k},\beta)\big) \prod_{(j,k)\in I_2} \Phi_1 \left(\frac{y_j^d - y_k R(\mathbf{h}_{j,k},\beta),\beta)}{\sqrt{1-R(\mathbf{h}_{j,k},\beta)^2}}\right) \phi(y_k) \prod_{(j,k)\in I_3} \Phi_2\Big(y_j^d, y_k^d, R(\mathbf{h}_{j,k},\beta)\Big) \quad (12)$$

Here $\Phi_2(x,y,r)$ is the distribution function of the 2 dimensional normal distribution with correlation r and standard normal marginal distributions N(0,1) and $\phi_2(x,y,r)$ is its density function. The calculation of $\Phi_2(x,y,r)$ requires the numerical integration of the bivariate normal density. The likelihood function could also be written using multipoint configurations, but this would also lead to an increase of the complexity of the calculations. The set I_1 contains pairs of locations with both variables being measured exactly. In I_2 pairs are listed which consist of an exact observation and a below detection limit value. Finally, I_3 contains pairs with values below the detection limit. The logarithm of the likelihood function is maximized numerically.

The above procedure might require a lot of computation effort if the number of observations is large. Instead one can reduce the number of pairs considered in Eq. (12) by selecting different distance classes and taking each observation exactly M times as a member of a pair. This way one can avoid clustering effects.

A similar but slightly more complicated procedure has to be used for the estimation of the parameters of the v-copula. In this case the variable Z is first transformed to:

$$y_k = H_1^{-1}(G(z(x_k))) \quad k = 1,\ldots,n_z \quad (13)$$

$$y_j^d = H_1^{-1}\big(G(d(x_j))\big) \quad j = 1,\ldots,n_d \quad (14)$$

Here $H_1(.)$ is the univariate distribution function of the v-transformed normal distribution. This can be written as:

$$H_1(y) = \Phi_1\left(\left(\frac{y}{k}\right)+m\right) - \Phi_1(m-y) \quad (15)$$

The corresponding density is:

$$h_1(y) = \frac{1}{k}\phi_1\left(\left(\frac{y}{k}\right)+m\right) + \phi_1(m-y) \quad (16)$$

The likelihood function in this case is:

$$L(\beta) = \prod_{(j,l)\in I_1} h_2\big(y_j, y_l, \beta)\big) \prod_{(j,l)\in I_2} H_c\Big(y_j^d, y_l, \beta\Big) h_1(y_l) \prod_{(j,l)\in I_3} H_2\Big(y_j^d, y_l^d, \beta)\Big) \quad (17)$$

The sets I_1, I_2 and I_3 are defined as for the Gaussian case. $H_2(.,.)$ is the distribution function of the bivariate v-transformed distribution:

$$\begin{aligned} H_2(y_1,y_2,\beta) = {} & \Phi_2\left(\left(\frac{y_1}{k}\right)+m, \left(\frac{y_2}{k}\right)+m, R(\mathbf{h}_{1,2},\beta)\right) \\ & + \Phi_2\big(m-y_1, m-y_2, R(\mathbf{h}_{j,k},\beta)\big) \\ & - \Phi_2\left(\left(\frac{y_1}{k}\right)+m, m-y_2, R(\mathbf{h}_{1,2},\beta)\right) \\ & - \Phi_2\left(m-y_1, \left(\frac{y_2}{k}\right)+m, R(\mathbf{h}_{1,2},\beta)\right) \end{aligned} \quad (18)$$

The corresponding density function is:

$$\begin{aligned} h_2(y_1,y_2,\beta) = {} & \frac{1}{k^2}\phi_2\left(\left(\frac{y_1}{k}\right)+m, \left(\frac{y_2}{k}\right)+m, R(\mathbf{h}_{1,2},\beta)\right) \\ & + \phi_2\big(m-y_1, m-y_2, R(\mathbf{h}_{1,2},\beta)\big) \\ & + \frac{1}{k}\phi_2\left(\left(\frac{y_1}{k}\right)+m, m-y_2, R(\mathbf{h}_{1,2},\beta)\right) \\ & + \frac{1}{k}\phi_2\left(m-y_1, \left(\frac{y_2}{k}\right)+m, R(\mathbf{h}_{1,2},\beta)\right) \end{aligned} \quad (19)$$

Here $R(\mathbf{h}_{1,2},\beta)$ is the correlation function of the Gaussian variable Y and $h_2(.,.)$ is the density function corresponding to H_2. The mixed bivariate function $H_c(.,.)$ is obtained via integration of the density:

$$H_c(y_1,y_2,\beta) = \int_{-\infty}^{y_1} h_2(y,y_2)\,dy \quad (20)$$

As described in Eq. (19) the density h_2 is a weighted sum of normal densities, the corresponding integral can be calculated for each term separately, which is similar to the normal case.

Due to the complicated form of the overall likelihood function, a numerical optimization of the log-likelihood function is performed.

Different forms of the correlation function can be considered – such as the exponential with $\beta = (A,B)$:

$$R(\mathbf{h},A,B) = \begin{cases} 0 & \text{if } |\mathbf{h}| = 0 \\ B\exp\left(-\frac{|\mathbf{h}|}{A}\right) & \text{if } |\mathbf{h}| > 0 \end{cases} \quad (21)$$

where $0 \le B \le 1$ and $A > 0$.

3 Interpolation

Once the parameters of the correlation function (A,B) and for the v-transformed copula the parameters of the v-transformation (m,k) are estimated the interpolation can be carried out. In order to reduce the complexity of the problem, interpolation will be done using a limited number of neighboring observations. Due to an *umbrella* effect similar to ordinary kriging, observations which are *behind* other

observations have a minor influence on the conditional distribution. Further, this restriction to local neighborhoods relaxes the assumption of stationarity to a kind of local stationarity. An example in Bárdossy and Li (2008) demonstrates that this assumption does not significantly alter the results of interpolation.

The goal of interpolation is to find the density of the random variable $Z(x)$ conditioned on the available censored and uncensored observations. The conditional density $g_x(z)$ for location x can be written as:

$$\begin{aligned} g_x(z) &= P\left(Z(x)=z|Z(x_i)<d_i\ ,i;Z(x_j)=z_j\ ,j\right) \\ &= \frac{P\left(Z(x)=z,Z(x_i)<d_i\ i,Z(x_j)=z_j\ ,j\right)}{P\left(Z(x_i)<d_i\ i,Z(x_j)=z_j\ ,j\right)} = \\ &= \frac{P\left(Z(x_i)<d_i\ ,i|Z(x)=z\ ,Z(x_j)=z_j\ j\right)P\left(Z(x)=z,Z(x_j)=z_j\ j\right)}{P\left(Z(x_i)<d_i\ ,i=1,\ldots,n_d|Z(x_j)=z_j\ ,j\right)P\left(Z(x_j)=z_j\ ,j\right)} = \\ &= \frac{P\left(Z(x_i)<d_i\ ,i|Z(x)=z\ ,Z(x_j)=z_j\ j\right)}{P\left(Z(x_i)<d_i\ ,i=1,\ldots,n_d|Z(x_j)=z_j\ ,j\right)} \\ &\cdot P\left(Z(x)=z|Z(x_j)=z_j\ j\right)=g_x^d(z)g_x^e(z) \end{aligned} \tag{22}$$

The above equation shows that the final conditional density is composed of two terms. The first $g_x^d(z)$ is related to the non-detects the second multiplicative term $g_x^e(z)$ is the iterpolation (conditional density) obtained from the exact values. This term is the traditional interpolator itself as if there were no values below the detection limit. If there are no exact measurememnts in the neighborhood of x then the second term equals to the marginal density of the variable, which is modified by the non-detects in the neighborhood through $g_x^d(z)$ Both the numerator and the denumerator of the first part of the expression are conditional multivariate distribution function values which require integration of the corresponding multivariate densities in n_d dimensions.

For the normal copula case, Eq. (22) can be written with the help of the transformed variable Y for $z=G^{-1}(\Phi(y))$:

$$\begin{aligned} g_x(z) = &\frac{P\left(Y(x_i)<y_i^d\ ,i|Y(x)=y\ Y(x_j)=y_j\ ,j\right)}{P\left(Y(x_i)<y_i^d\ ,i|Y(x_j)=y_j\ ,j\right)} \\ &\cdot P\left(Y(x)=y|Y(x_j)=y_j\ ,j\right) \end{aligned} \tag{23}$$

The conditional distribution of a multivariate normal distribution is itself multivariate normal with expectation $\mu_{\mathbf{c}}^{\mathbf{0}}$ and covariance matrix $\Gamma_{\mathbf{c}}^{\mathbf{0}}$ with:

$$\Gamma_{\mathbf{c}}^{\mathbf{0}} = \Gamma_{\mathbf{00}} - \Gamma_{\mathbf{01}}{\Gamma_{\mathbf{11}}}^{-1}{\Gamma_{\mathbf{01}}}^{T} \tag{24}$$

The expected value of the conditional is:

$$\mu_{\mathbf{c}}^{\mathbf{0}} = \Gamma_{\mathbf{01}}{\Gamma_{\mathbf{11}}}^{-1}\mathbf{y} \tag{25}$$

$\mathbf{y}^T=(y,y_1,\ldots,y_{n_z})$.

The matrices $\Gamma_{\mathbf{00}}$ $\Gamma_{\mathbf{01}}$ and $\Gamma_{\mathbf{11}}$ are the correlation matrices corresponding to the pairs of observations with censored and uncensored data, calculated with the correlation function $R(h)$.

Thus the conditional probability in the numerator in Eq. (22) can be calculated as:

$$\begin{aligned} &P\left(Z(x_i^0)<d_i;i=1,\ldots,n_d|Z(x)=z;Z(x_j^1)=z_j;j=1,\ldots,n_z\right) \\ &\quad=\Phi_{\mu_{\mathbf{c}}^{\mathbf{0}},\Gamma_{\mathbf{c}}^{\mathbf{0}}}(y,y_1,\ldots,y_{n_z}) \end{aligned} \tag{26}$$

where $\Phi_{\mu_{\mathbf{c}}^{\mathbf{0}},\Gamma_{\mathbf{c}}^{\mathbf{0}}}$ is the distribution function of $N(\mu_{\mathbf{c}}^{\mathbf{0}},\Gamma_{\mathbf{c}}^{\mathbf{0}})$. Values of the multivariate normal distribution function can be calculated by numerical integration, for example using Genz and Bretz (2002). The denumerator in Eq. (22) requires the same type of calculation.

The denumerator is independent of the value z and can be calculated exactly as the numerator. Note that the point for which the interpolation has to be carried out is considered as a *pseudo* observation with the observed value z. Thus the numerator has to be evaluated for a number of possible z values to estimate the conditional density.

For the v-transformed copula the interpolation procedure is slightly more difficult, but as the n-dimensional density of the v-transformed variable is a weighted sum of 2^n normal densities the calculation procedure is similar. However, we will not go into further details here.

4 Application and results

The above described methodology was applied to a regional groundwater pollution investigation. Two censored variables and an artificially censored variable were used to demonstrate the methods, and to compare them to traditional interpolations.

4.1 Investigation area

An extensive dataset consisting of more than 2500 measurements of groundwater quality parameters of the near surface groundwater layer in Baden-Württemberg were used to illustrate the methodology. Three quality parameters namely deethylatrazine – degradation product of atrazine – arsenic and chloride were selected for this study. The measurements were carried out in the time period between 2007 and 2010.

While the first two parameters are heavily censored the chloride concentrations exceed the detection limit in 99.9 % of the cases. This variable is artificially censored using different thresholds in order to show the effectiveness of the method.

Table 1 shows the basic statistics for the selected data. Note the high positive skewness for all variables. This alone would lead to substantial difficulties in estimating spatial correlation functions, even in the case where most values had been above the detection limit.

Table 1. Basic statistics of the investigated variables mean, standard deviation and skewness are calculated from values above the detection limit.

	Number of observations	Number of above DL	Statistics of values > Detection limit			
			Mean	Standard deviation	Skewness	Maximum
Arsenic	2234	979	0.002733	0.007392	13.4	0.1618
deethylatrazine	2848	403	0.064243	0.068316	4.5	0.68
Chloride	2805	2801	39.9	165.8	30.3	6940.0

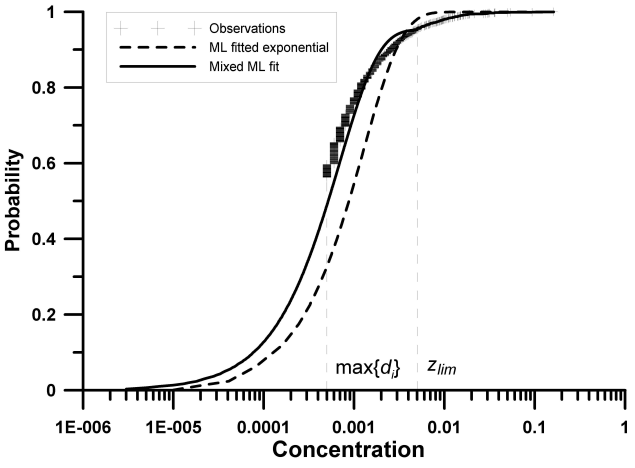

Fig. 1. The empiricel distribution of the observed arsenic concentrations (crosses) for the values above the highest detection limit and the distributions obtained fitted via maximum likelihood for the whole dataset (dashed line) and with setting $z_{\lim}$ to $0.005\,\mathrm{mg\,l^{-1}}$ (solid line).

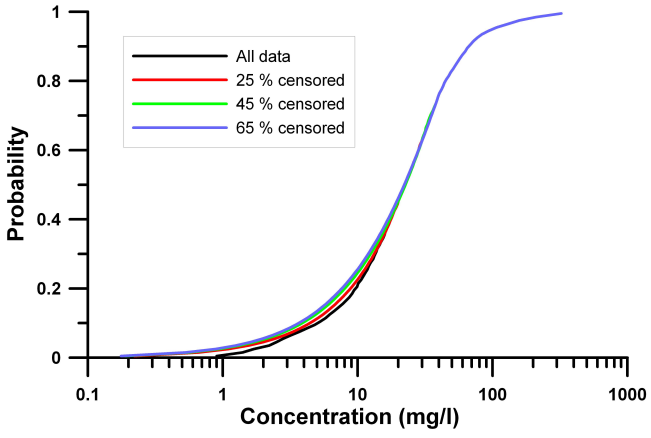

Fig. 2. The distribution of chloride concentrations and the estimated distributions corresponding to different degrees of censoring.

4.2 Parameter estimation

As a first step the marginal distributions were estimated using the approach described in Sect. 2.1. Figure 1 shows the distribution functions for arsenic. The estimation method was compared to the full maximum likelihood (Eq. 1 which would correspond to $z_{\lim} > \max(z_i, i = 1, \ldots, I$ in Eq. 2)). The empirical distribution function is only defined for values above the highest detection limit.The traditional maximum likelihood estimation is strongly influenced by outliers, leading to unrealistic, and unacceptable results. In contrast, setting $z_{\lim}$ such that $z_{\lim} > \max(d_j, j = 1, \ldots, J)$ and bearing in mind that there are at least a few (30 or more in our case) exact measurement values (z_is) below $z_{\lim}$, leads to a good fit of the observed values, but one can see a slight break in the distribution function at $z_{\lim}$.

In order to investigate the quality of the extension of the distribution to censored values the observed chloride concentration values were artificially censored. Detection limits were set to the 15, 25, 35, 45, 55, 65, 75 and 85 % quantiles of the distribution. Figure 2 shows distribution functions corresponding to different detection limits for chloride. Note that in order to see any differences the x-axis is shown on a logarithmic scale. All distribution functions are very similar, showing that the upper middle part of the distribution can be well used to extend it to low values.

The parameters of the spatial structure were estimated both for a normal and a v-transformed normal copula. An exponential spatial correlation function was assumed. Table 2 shows the parameters of the spatial copulas for the selected variables. The copula fits are very different. While for arsenic the correlation function of the normal copula has a high B value indicating a strong spatial structure, for the v-transformed copula the B is much lower. For deethylatrazine the situation is inverted: the v-transformed copula shows a strong spatial link and the normal nearly no spatial correlations.

Table 2. Parameters of the fitted copulas.

	Gauss copula		V-transformed copula			
	B	*A*	*B*	*A*	*m*	*k*
Arsenic	0.750	1325	0.810	49 000	1.78	0.376
deethylatrazine	0.030	669	0.579	35 000	0.29	2.469
Chloride	0.620	11 539	0.449	27 500	1.98	0.147

4.3 Interpolation

In order to illustrate the properties of the interpolation method illustrative examples are first considered. Assume that the value at the center of a square is to be estimated, with observations at the four corners. Four different configurations are considered:

1. Assume two corners on a diagonal have exact values equal to the 0.5 quantile of the distribution.
2. Assume two corners on a diagonal have exact values equal to the 0.5 quantile of the distribution and the two other corners have censored values with the same detection limit which is equal to the 0.5 quantile of the distribution.
3. All corners have exact values equal to the 0.5 quantile of the distribution.
4. Assume two corners on a diagonal have exact values equal to the 0.5 quantile of the distribution and the two other corners have censored values one with a detection limit equal to the 0.5 quantile the other equal to the 0.1 quantile of the distribution.

The spatial dependence structures of deethylatrazin were used for these examples. Figure 3 shows the conditional densities in the quantile space for the center of the square. Configuration one corresponds to the case if censored values are not considered corresponding to $g_x^e(z)$ in Eq. (22). This density is modified in configuration 2 – the two values below the 0.5 quantile lead to a higher density for lower values. Configuration 3 corresponds to the case when non-detects are set to the detection limit. This leads to an estimator with less uncertainty and with higher expectation than in configuration 2. Configuration 4 shows that a constraint corresponding to a low detection limit can substantially modify the density obtained by the interpolatioin.

Figure 4 shows the interpolated maps for chloride using all observations and three different maps using 25 %, 45 % and 65 % censoring. Note the high similarity between the maps. The pointwise correlation between the map based on all observations and the maps obtained after censoring was calculated and is shown on Fig. 5. The correlation is constant around 0.95 up to 65 %, and diminishes afterwards rapidly thereafter, reaching nearly 0 at 85 % censoring.

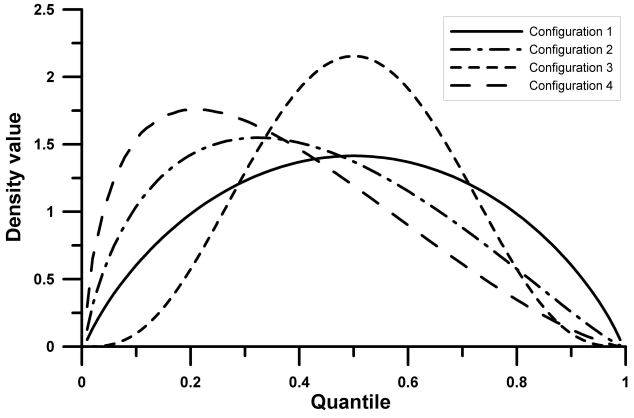

Fig. 3. Conditional densities obtained for the center of a square using different data at the corners.

- Configuration 1 two corners on a diagonal have exact values equal to the 0.5 quantile of the distribution.
- Configuration 2 two corners on a diagonal have exact values equal to the 0.5 quantile of the distribution and the two other corners have censored values with the same detection limit which is equal to the 0.5 quantile of the distribution.
- Configuration 3 all corners have exact values equal to the 0.5 quantile of the distribution.
- Configuration 4 two corners on a diagonal have exact values equal to the 0.5 quantile of the distribution and the two other corners have censored values one with a detection limit equal to the 0.5 quantile the other equal to the 0.1 quantile of the distribution

An advantage of the copula based approach is that it provides the full conditional distribution for each location. Thus confidence intervals can be calculated, which are more realistic than those obtained by kriging.

4.4 Comparison with other interpolation methods

As an alternative ordinary kriging (OK) was used for interpolation. Three different treatments of the values below the detection limit were considered:

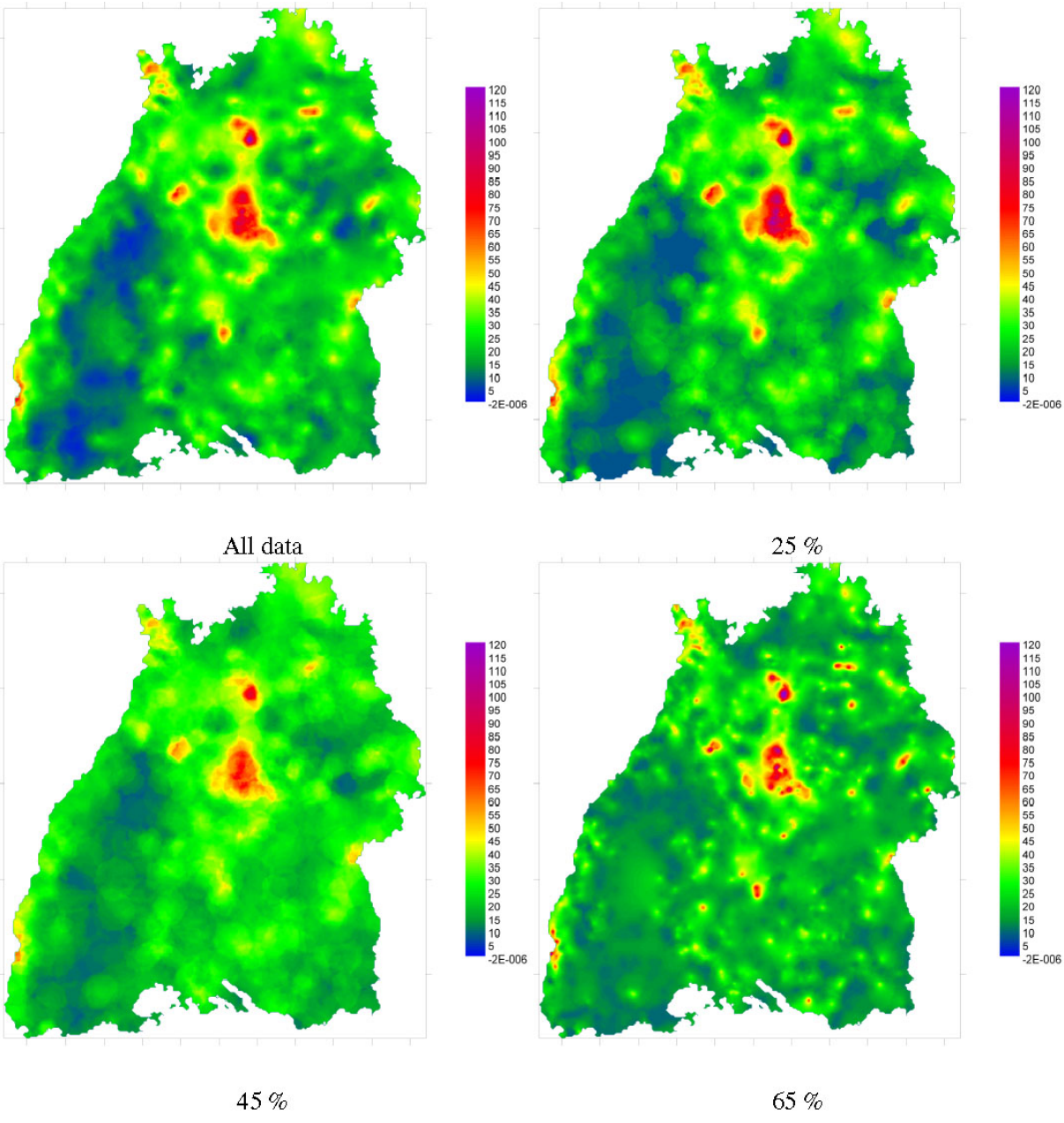

Fig. 4. Interpolated chloride concentrations for different grades of censoring.

1. All values below the detection limit were set to zero.
2. All values below the detection limit were set to half of the corresponding detection limit.
3. All values below the detection limit were set to the corresponding detection limit.

Empirical variograms were calculated for each case. Additionally the empirical variogram was calculated from the exact values only. Figure 6 shows the graph of these variograms for deethylatrazine. The exact values lead to a variogram without any structure and with the highest variance. The datasets with replaced values show a much lower variability and the replacement with zeros increases the variability only very slightly. These variograms do not show a spatial structure. Only after the removal of a few extremes, which were considered as outliers one could obtain a reasonable variogram. This example gives a good idea about the difficulties involved in the assessment of a reasonable variogram. The same procedure was carried out for arsenic and chloride. In the later case the variograms were calculated for different levels of censoring. A cross validation using OK was performed for each parameter and each censoring.

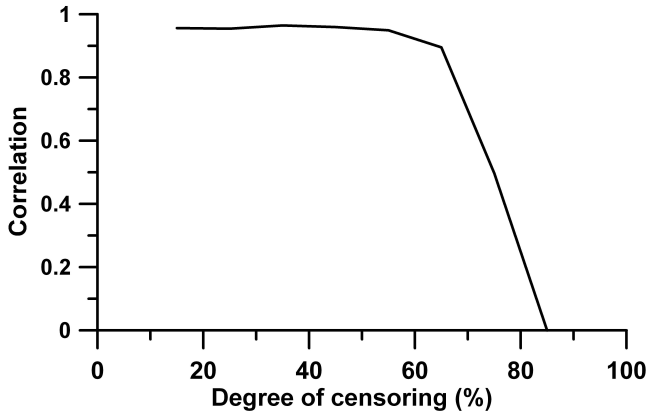

Fig. 5. Correlation between the interpolated map of Chloride and the maps interpolated from censored data.

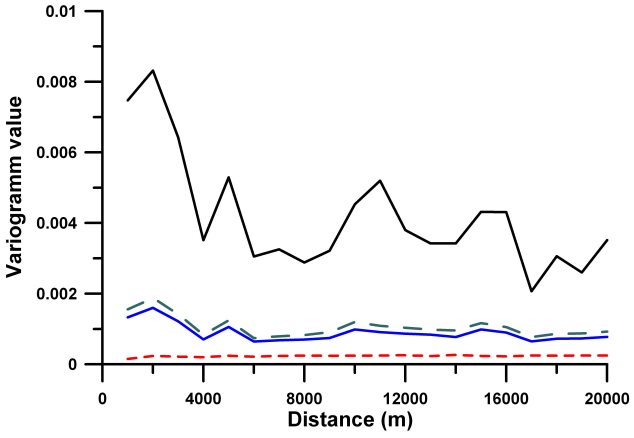

Fig. 6. Empirical variograms calculated for deethylatrasine, using exact data only (black solid), using nondetects replaced by zero (blue dashed) or by the detection limit (blue solid) and using nondetects replaced by zero and removal of outliers (red dashed).

Another popular method to treat highly skewed variables is indicator kriging (IK). The indicator corresponding to a cutoff value α is defined as:

$$I_\alpha(Z(x)) = \begin{cases} 0 \text{ if } Z(x) > \alpha \\ 1 \text{ if } Z(x) \leq \alpha \end{cases} \quad (27)$$

Indicator variograms are calculated for a set of α values. These do not suffer from the problem of outliers. A subsequent IK leads for each x and α to an estimated value which is usually interpreted as a probability of non-exceedance. The estimators corresponding to different α values are then assembled to a distribution function. The expected value can then be calculated for each location. Censored data can be treated with indicators, namely for α values below the detection limit the indicator remains undefined, while for above the indicator is 1. This is a correct treatment of the data, but leads to the problem that for each α below the lowest detection limit all indicator values equal zero. This means that the procedure is practically filling in the data with the detection limit, leading to similar biased estimators as OK. Figure 7 shows the graph of empirical indicator variograms for deethylatrazine. Note that in contrast to the empirical variograms of Fig. 6 these curves show a clear spatial dependence even without removing the outliers.

Lognormal kriging was not considered for this comparison, as it was reporeted the back transformation is very sensitive and might lead to problems with the estimator Roth (1998). Further the replacement of the non-detects would play a major role in the variogram estimation for this method.

Figure 8 shows the interpolated maps for deethylatrazine using the v-copula, IK and OK by setting all censored data equal to the corresponding detection limit. The OK maps show the typical problem the method has with skewed distributions. The high values have a large influence, and lead to an overestimation. The map obtained by IK is more realistic. However the overestimation is still a problem here,

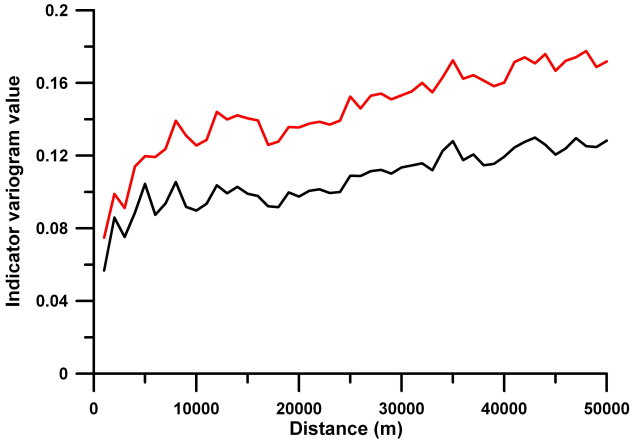

Fig. 7. Empirical indicator variograms calculated for deethylatrasine for the 85 % and 90 % values of the distribution.

as the values below the detection limit are practically set to the detection limit. The copula based interpolation allows interpolated values below the detection limit and, in doing so, leads to a plausible result.

The spatial means calculated chloride concentrations of the interpolated maps using different degrees of censoring are shown on Fig. 9. For IK and for OK using detection limit for censored values censoring leads to an increase of the spatial mean. Using zero for the censored data in OK results a decrease of the mean, while setting 50 % of the detection limit brings an increase only at high degrees of censoring. In contrast the copula approach shows only a slight decrease in the spatial mean. Note that the spatial mean is below the 55 % value of the distribution. Thus for the high levels of censoring the interpolated mean is below the lowest measured value.

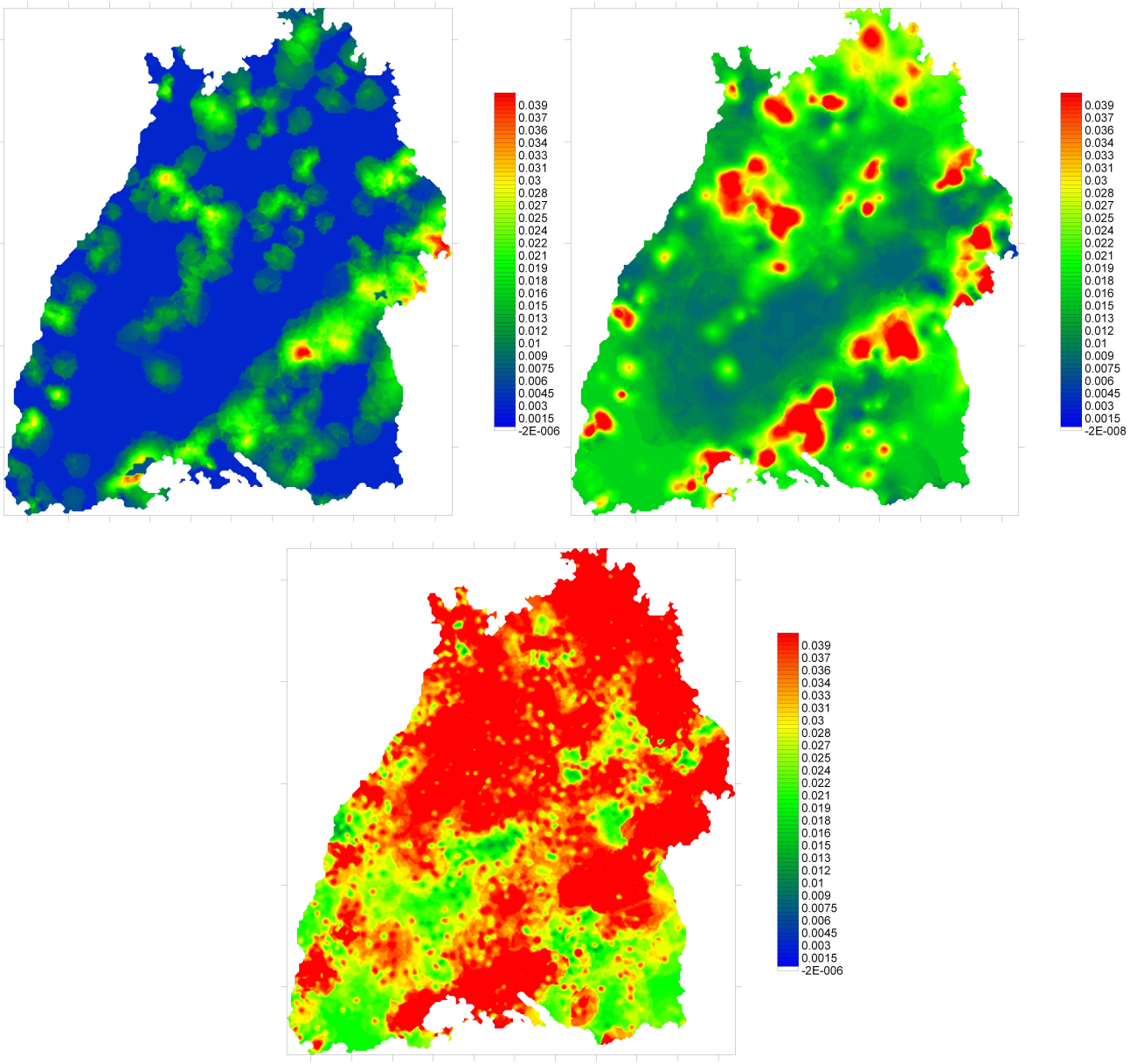

Fig. 8. Interpolated deethylatrazine concentrations using different interpolation methods. For OK the values were set to the detection limit.

Table 3. Cross validation results for Arsenic.

Measure	V-copula	Gauss-copula	Indicator Kriging	Ordinary Kriging 50 % of Detection limit
MSQE	$\mathbf{3.7 \times 10^{-6}}$	1.0×10^{-5}	5.3×10^{-5}	1.0×10^{-5}
Rank correlation	0.32	0.32	**0.33**	**0.33**
LEPS Score	**0.142**	0.154	**0.142**	0.159
Mean probability for $<$ DTL	**0.610**	0.559	0.042	0.437

Table 4. Cross validation results for deethylatrazin.

Measure	V-copula	Gauss-copula	Indicator Kriging	Ordinary Kriging 50 % of Detection limit
MSQE	$\mathbf{5.1 \times 10^{-4}}$	3.0×10^{-3}	5.0×10^{-3}	1.7×10^{-3}
Rank correlation	0.44	0.31	0.40	**0.48**
LEPS Score	0.168	0.311	**0.100**	0.110
Mean probability for < DTL	0.869	**0.888**	0.560	0.650

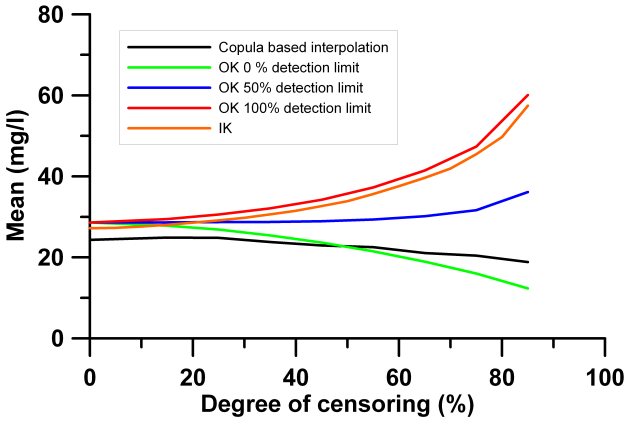

Fig. 9. Mean of the interpolated maps of Chloride for different degrees of censoring and different interpolations.

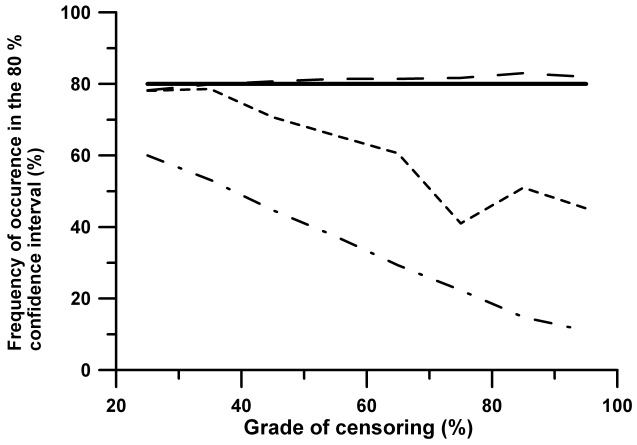

Fig. 10. Frequency of observations in the 80 % confidence interval for V-copula based interpolation (long dashes) and Gauss-copula based interpolation (short dashes) and indicator kriging (dashed dotted line) for different grades of censoring of Chloride.

As a next step for all three variables and all interpolation methods a cross validation was carried out. The evaluation of the cross validation results is not straightforward due to the censoring. The usual squared error is, even for the exact values, not appropriate as the distributions are highly skewed and some extreme outliers would dominate this measure. Instead this measure was calculated by leaving out the upper 1 % of the measured values, ensuring that outliers were not considered for the calculation. Further the rank correlation for the exact values was calculated. Additionally the LEPS score (linear error in probability space) Ward and Folland (1991) was calculated to evaluate the fit in the probability space.

$$\mathrm{LEPS} = \frac{1}{n}\sum_{i=1}^{n} |G_z(z(\mathbf{x_i})) - G_z(z^*(\mathbf{x_i}))| \tag{28}$$

Here $z^*(\mathbf{x_i})$ is the expected value of the iterpolation calculated from the density obtained in Eq. (22).

For the measurements below the detection limit the average of the probabilities to be below the detection limit was calculated.

Results for the two censored variables and for an artificially censored case (chloride) are displayed in Tables 3 and 4. As one can see the copula based approaches outperform the ordinary and the indicator kriging. Note that the mean squared error, the rank correlation and the LEPS score were all calculated for the exact measurements only. From the two copula models the v-copula allowing a non-symmetrical dependence is slightly better than the Gaussian.

For the artificially censored mean squared error, rank correlation and LEPS score were calculated using all data without considering the artificial censoring. Thus these measures represent a realistic measure of interpolation quality. The results are shown in Table 5. Note that ordinary kriging has a very high mean squared error. This is caused by the high skewness of the marginal distribution which had much less influence on the indicator and copula approaches. The evaluation of the procedures for values below the detection limit is rather difficult. The first three measures cannot be used for these observations. As all interpolation methods provide probability distributions a possible quality measure is to calculate the for each point the probability that the value is below the detection limit. Optimally this probability should be 1. For all three parameters the copula based approaches deliver the highest values. Indicator kriging is by far the weakest in this measure.

Table 5. Cross validation results for Chloride with 45 % artificial censoring.

Measure	V-copula	Gauss-copula	Indicator Kriging	Ordinary Kriging 50 % of Detection limit
MSQE	273.1	**251.8**	298.6	2922.5
Rank correlation	**0.61**	**0.61**	0.58	0.45
LEPS Score	0.186	0.174	0.191	**0.150**
Mean probability for < DTL	**0.593**	0.555	0.000	0.390

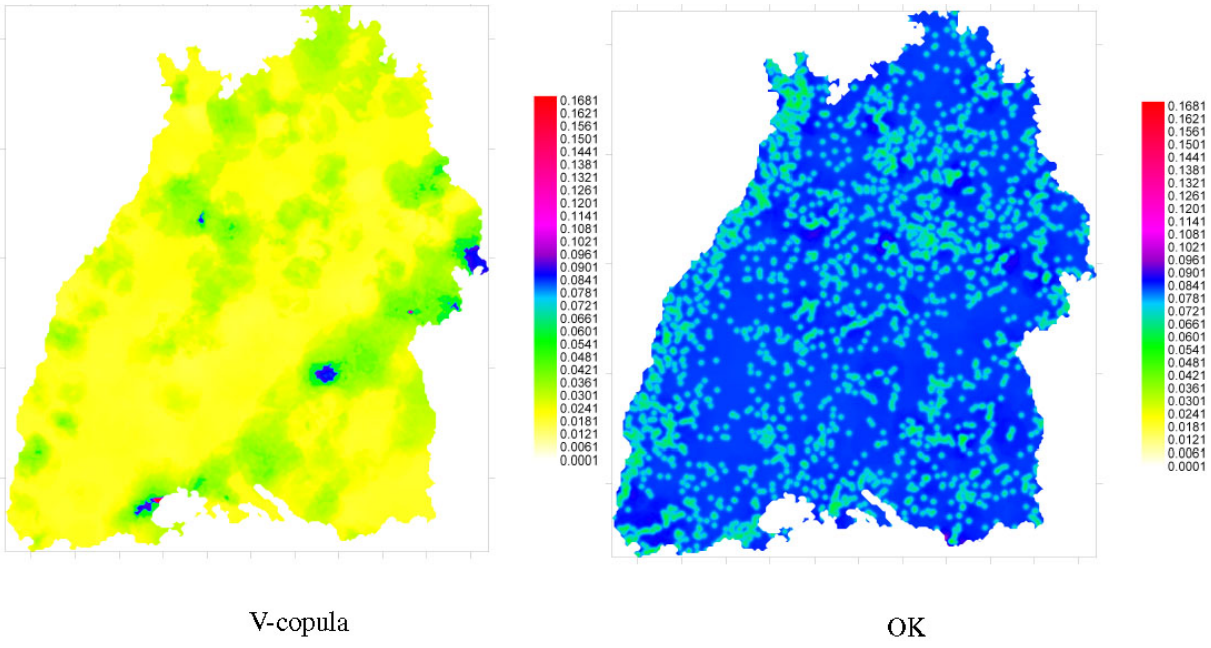

Fig. 11. Uncertainty maps for deethylatrazin: left the length of the 80 % confidence interval obtained via v-copula based interpolation, right the kriging standard deviation obtained by OK.

For interpolation and for possible random simulation of the fields a good measure of uncertainty is of great importance. As the kriging variance is only a good measure of uncertainty when the data follow a multivariate normal distribution. Else it is only a measure of data configuration, not data value dependent (especially for skewed distributions c.f. Journel, 1988) and it is not a good measure of uncertainty. The indicator approach provides estimates of the local conditional distribution functions. As it is not directly considering the estimation uncertainty (all indicator values are interpolated values with no uncertainty associated) it does not provide a good uncertainty measure. The copula approach yields full probability distributions for each location, thus arbitrary confidence intervals can be derived. Figure 11 shows the width of the 80 % confidence interval obtained using v-copula based interpolation and the width of the 80 % confidence interval obtained using ordinary kriging under the assumption of a normally distributed error for deethylatrazin. One can see that the estimation quality of the copula based interpolation is very heterogeneous over the whole domain. Regions with high observed values the confidence intervals are wide, in low areas narrow. For ordinary kriging the estimation error (kriging standard deviation) is small close to points with measured values, irrespective of the observed values.

In order to validate the confidence intervals the frequency of observations within the 80 % confidence interval (obtained from cross validation) was calculated. Figure 10 shows the percentage of chloride values falling into the 80 % confidence interval for different censoring levels obtained using the v-copula and the Gauss copula. As one can see for the v-copula the frequency is close to the target 80 % for all censoring levels while for the Gauss copula the confidence intervals become meaningless above 35 % censoring.

5 Conclusions

In this paper a methodology for the interpolation of variables with data below a detection limit was developed. As a first step the marginal distributions were estimated using a mixed approach which entailed a maximum likelihood method for the lower values and the empirical distribution for the high values. This procedure provides a robust estimator for the low concentrations without the negative influence of possible outliers. Using the fitted distributions the variables were transformed to the unit interval and their spatial copula was assessed, assuming spatial stationarity. Values below the detection limit are considered in a maximum likelihood estimation of the spatial copula parameters. Interpolation was done by calculating the conditional distributions for each location. The conditions include both the measurements as exact values and the below detection limit observations as inequality constraints.

The copula based interpolation is exact at the observation locations; the interpolated value equals the observed value. For locations with censored observations the method provides an updated distribution function which differs from the constrained marginal. Other procedures such as indicator kriging with inequality constraints do not update distributions at observation locations.

Investigations based on the artificially censored dataset show that the copula-based approaches remain unbiased even for large degrees of censoring. Among the kriging approaches only ordinary kriging with setting the censored values equal to the half of the corresponding detection limit did not show a systematic error for higher detection limits. This choice is clearly better than setting the values below the detection limit equal to the detection limit, or setting them all equal to zero, which both lead to systematic errors. Indicator kriging also shows a systematic bias increasing with the detection limit.

The copula-based approaches outperform ordinary and indicator kriging in their interpolation accuracy. Indicator kriging is only slightly worse than the copula based interpolation, while ordinary kriging with all different considerations of the values below detection limit are the poorest estimators.

The main advantage of the copula based approaches is in the estimation of the interpolation uncertainty. While ordinary kriging yields unrealistic estimation variances depending only on the configuration of the measurement locations, the copula-based interpolation yields reasonable confidence intervals. The v-copula based approach yields more realistic confidence intervals than the Gaussian alternative.

The suggested approach can be extended to handle any kind of inequality constraints both for spatial structure assessment and for interpolation.

The model can serve as a basis for conditional spatial simulation. It would be possible to extend the model to a Bayesian approach where prior distributions are assigned to individual locations.

Acknowledgements. Research leading to this paper was supported by the German Science Foundation (DFG), project number Ba-1150/12-2.

Edited by: H. Cloke

References

Bárdossy, A.: Copula-based geostatistical models for groundwater quality parameters., Water Resour. Res., 42, W11416, doi:10.1029/2005WR004754, 2006.

Bárdossy, A. and Li, J.: Geostatistical interpolation using copulas., Water Resour. Res., 44, W07412, doi:10.1029/2007WR006115, 2008.

Cohen, C.: Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated., Technometrics, 1, 217–237, 1959.

Cohen, C.: Progressively Censored Sampling in the Three Parameter Log-Normal Distribution., Technometrics, 18, 99–103, 1976.

Genz, A. and Bretz, F.: Comparison of Methods for the Computation of Multivariate t-Probabilities, J. Comp. Graph. Stat., 11, 950–971, 2002.

Helsel, D. R.: More than obvious: Better methods for interpreting nondetect data., Environmental Science and Technology, 39, 419A–423A, 2005.

Helsel, D. R. and Cohn, T. A.: Estimation of descriptive statistics for multiply censored water quality data., Water Resour. Res., 24, 1997–2004, 1988.

Journel, A. G.: New Distance Measures: The Route Toward Truly Non-Gaussian Geostatistics, Mathematical Geology, 20, 459–475, 1988.

Keef, C., Tawn, J., and Svensson, C.: Spatial risk assessment for extreme river flows, Journal of the Royal Statistical Society Series C Applied Statistics, 58, 601–618, 2009.

Roth, C.: Is lognormal kriging suitable for local estimation?, Mathematical Geology, 30, 999–1009, 1998.

Saito, H. and Goovaerts, P.: Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site., Environmental Science and Technology, 44, 4228–4235, 2000.

Sedda, L., Atkinson, P. M., Barca, E., and Passarella, G.: Imputing censored data with desirable spatial covariance function properties using simulated annealing., J. Geogr. Syst., 36, 3345–3353, 2010.

Shumway, R., Azari, R., and Kayhanian, M.: Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits., Environmental Science and Technology, 36, 3345–3353, 2002.

Ward, M. and Folland, C.: Prediction of seasonal rainfall in the Nordeste of Brazil using eigenvectors of sea-surface temperature., International Journal Climatology, 11, 711–743, 1991.