

---

# Spatial Interpolation Using Copula-Based Geostatistical Models

Hannes Kazianka and Jürgen Pilz

Institute of Statistics, University of Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria

`hannes.kazianka@uni-klu.ac.at`

`juergen.pilz@uni-klu.ac.at`

**Abstract.** It is common practice in geostatistics to use the variogram to describe the spatial dependence structure of the underlying random field. The variogram is not only sensitive to outlying observations but its estimation is also based on the Gaussian assumption which holds in hardly any practical application. As an alternative to spatial modelling using the variogram we consider describing the spatial correlation by means of copula functions. We present three methods for performing spatial interpolation using copulas. The first method is disjunctive kriging, which uses the relationship between bivariate copulas and indicator variograms. As a second method we propose a simple kriging of the rank-transformed data. The third method is a plug-in Bayes predictor, where the predictive distribution is calculated using the conditional copula given the observed data and the model parameters. We show that the latter approach generalizes the frequently applied trans-Gaussian kriging. Finally, we report on the results obtained for the Joker data set from the spatial interpolation comparison SIC2004.

## 1 Introduction

Copulas describe the dependence between random variables independently from their marginal distributions. They are commonly used in financial and actuarial statistics, however, they are just beginning to become popular in geostatistics. Spatial dependence is traditionally described using the variogram which is strongly influenced by the univariate distribution of the random field. Extreme outlying observations adversely affect the empirical and theoretical variogram estimates. Moreover, parametric variogram fitting methods rely on the Gaussian assumption which is hardly fulfilled for environmental processes. To circumvent these disadvantages Bardossy [1] proposed the use of copulas to describe the spatial variability. In the following we adopt this methodology and use it not only for spatial modelling of dependence structures but also for spatial interpolation. We present three methods for estimating the values of the random field at unknown locations. The first method we suggest is

indicator and disjunctive kriging. The second method is rank-order kriging, originally proposed by Journel and Deutsch [5], where we calculate the covariance function through its relationship to the Spearman rank correlation. The third method is a plug-in Bayes predictor and can be used if all multivariate distributions of the random field are modelled using the copula.

The paper is organized as follows. Section 2 reviews the basic properties of copulas, while Sect. 3 briefly describes the spatial copula methodology. In Sect. 4 the spatial interpolation techniques using copulas are presented and in Sect. 5 they are used to analyze the Joker data set from the spatial interpolation comparison SIC2004 [3]. Section 6 is devoted to conclusions.

## 2 Copulas

The word “copula” was first used by Sklar [10] to describe distribution functions on the  $n$ -dimensional unit cube,  $\mathbf{I}^n$ , that link multivariate distributions to their one-dimensional margins. To be precise, an  $n$ -dimensional copula (or  $n$ -copula) is an  $n$ -dimensional real function  $C : \mathbf{I}^n \rightarrow \mathbf{I}$  which satisfies the following properties:

1. For every  $\mathbf{u} \in \mathbf{I}^n$

$$C(\mathbf{u}) = 0 \text{ if at least one coordinate of } \mathbf{u} \text{ equals } 0,$$

and

$$C(\mathbf{u}) = u_k \text{ if all coordinates of } \mathbf{u} \text{ are } 1 \text{ except } u_k.$$

2. For every  $\mathbf{a}, \mathbf{b} \in \mathbf{I}^n$  with  $\mathbf{a} \leq \mathbf{b}$

$$V_C([\mathbf{a}, \mathbf{b}]) \geq 0,$$

where  $V_C([\mathbf{a}, \mathbf{b}]) = \Delta_{a_n}^{b_n} \Delta_{a_{n-1}}^{b_{n-1}} \dots \Delta_{a_1}^{b_1} C(\mathbf{u})$  is the  $n$ -th order difference of  $C$  on  $[\mathbf{a}, \mathbf{b}]$  and a first order difference is defined as  $\Delta_{a_k}^{b_k} C(\mathbf{u}) = C(u_1, \dots, u_{k-1}, b_k, u_{k+1}, \dots, u_n) - C(u_1, \dots, u_{k-1}, a_k, u_{k+1}, \dots, u_n)$ .

From this definition it is clear that a copula is a distribution function on the  $n$ -dimensional unit cube with uniformly distributed margins. The most important theoretical result about copulas was proved by Sklar [10] and expresses the ability of copulas to describe the dependence between random variables without information about their marginal distributions: If  $H$  denotes an  $n$ -dimensional distribution function with margins  $F_1, \dots, F_n$ , then there exists an  $n$ -copula  $C$  such that for all  $\mathbf{x} \in \overline{\mathbb{R}}^n$ ,

$$H(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)). \quad (1)$$

If  $F_1, \dots, F_n$  are all continuous, then  $C$  is unique. Conversely, if  $C$  is an  $n$ -copula and  $F_1, \dots, F_n$  are distribution functions, then the function  $H$  is an

$n$ -dimensional distribution function with margins  $F_1, \dots, F_n$ . Furthermore, if  $F_1^{-1}, \dots, F_n^{-1}$  are the inverse distribution functions of  $F_1, \dots, F_n$ , we get

$$C(u_1, \dots, u_n) = H(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)). \quad (2)$$

If  $C$  is an absolutely continuous copula, its density can be written as

$$c(u_1, \dots, u_n) = \frac{\partial^n C(u_1, \dots, u_n)}{\partial u_1 \dots \partial u_n} = \frac{h(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n))}{\prod_{i=1}^n f_i(F_i^{-1}(u_i))}, \quad (3)$$

where  $h$  denotes the density of  $H$  and the  $f_i$  denote the densities of  $F_i$ . One of the advantages of working with copulas is that they are invariant under almost surely increasing transformations of the random variables. Therefore, typical data transformation methods, such as taking the logarithm or performing a Box-Cox transformation, do not have an impact on the copula. Bivariate copulas are directly linked to the scale free measure of association known as Spearman's rho. The Spearman rank correlation between two random variables  $X_1$  and  $X_2$  with copula  $C$  can be calculated as

$$\rho_{X_1, X_2} = 12 \int \int_{I^2} u_1 u_2 dC(u_1, u_2) - 3 = 12 \int \int_{I^2} C(u_1, u_2) du_1 du_2 - 3. \quad (4)$$

For a thorough introduction to copulas the reader is referred to Nelsen [8].

### 3 Spatial Modelling Using Copulas

Although copulas are widely used for describing the dependence between random variables, for example in financial statistics, there are only two papers about incorporating copulas into the geostatistical framework so far. Suppose that we have a second-order stationary random field  $\{Z(\mathbf{x}) \mid \mathbf{x} \in \mathcal{S}\}$ , where  $\mathcal{S} \subseteq \mathbb{R}^2$  is the area of interest. Bondar et al. [2] use copulas for describing the dependence between  $X = \|\mathbf{h}\|$  and  $Y = \frac{|Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})|^2}{0.91}$ , where  $\mathbf{h}$  denotes a vector separating two points. A median regression of  $Y$  on  $X$  yields a robust semivariogram estimator without having used any frequently applied variogram model.

#### 3.1 Describing the Random Field Using Copulas

Bardossy [1] presented a different method for spatial modelling using copulas that generalizes the concept of the variogram. Let  $F_Z$  denote the univariate distribution of the random process which is the same for each location  $\mathbf{x}$  due to stationarity. All multivariate distributions of the random field are described using multivariate copulas with the help of Sklar's theorem (1). For example, the relation between two locations separated by the vector  $\mathbf{h}$  is characterized by the bivariate distribution

$$P(Z(\mathbf{x}) \leq z_1, Z(\mathbf{x} + \mathbf{h}) \leq z_2) = C_{\mathbf{h}}(F_Z(z_1), F_Z(z_2)), \quad (5)$$

whose dependence structure is described by the copula  $C_{\mathbf{h}}$ . The copula becomes a function of the separating vector  $\mathbf{h}$  (or the separating distance  $h := \|\mathbf{h}\|$  if the random field is isotropic) and does not depend on the location  $\mathbf{x}$ . Hence, the spatial copula describes the dependence over the whole range of quantiles and not only the mean dependence as the variogram does. Every spatial copula is symmetric by definition. This means that  $C_{\mathbf{h}}(u_1, \dots, u_n) = C_{\mathbf{h}}(u_{\pi(1)}, \dots, u_{\pi(n)})$  for an arbitrary permutation  $\pi$  and  $n \geq 2$ . Moreover, we want to add the following two restrictions:  $\|\mathbf{h}\| \rightarrow \infty$  implies  $C_{\mathbf{h}}(\mathbf{u}) \rightarrow \prod_{i=1}^n u_i$  and  $\|\mathbf{h}\| \rightarrow 0$  implies  $C_{\mathbf{h}}(\mathbf{u}) \rightarrow \min_i u_i$ . These restrictions ensure that far distant observations have almost no dependence and observations that are very close to each other have a strong dependence. The special case of a Gaussian random field, where the copula can be written as  $C(u_1, \dots, u_n) = \Phi_{\mathbf{0}, \mathbf{A}}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n))$  with  $\mathbf{A} = \text{diag}(1, \dots, 1)$  and the marginal distribution is  $F_Z = \Phi_{m, \sigma^2}$ , is included in this model. Here,  $\Phi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$  denotes the distribution function of the multivariate Gaussian distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ . The Gaussian copula becomes a function of  $\mathbf{h}$  by assuming that its correlation function follows one of the well-known parametric geostatistical models, e.g. the Matern model. However the Gaussian copula, as well as the Student-t copula, does not only express a symmetric but also a radially symmetric dependence,  $C(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$ . This means that high and low values of the distribution have equal dependence properties. To allow for more flexibility Bardossy [1] introduced a non-Gaussian copula family which is constructed from a multivariate non-central  $\chi^2$ -distribution. Squaring the entries of a Gaussian random vector  $\mathbf{Y} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ , where  $\mathbf{m} = (m, \dots, m)$  and  $\boldsymbol{\Sigma}$  denote the mean vector and the correlation matrix respectively, leads to a multivariate distribution with margins having a non-central  $\chi^2$ -distribution with 1 degree of freedom and non-centrality parameter  $\lambda = m^2$ . The distribution function  $D$  and density  $d$  can be calculated as

$$D(z_1, \dots, z_n) = \sum_{i=0}^{2^n - 1} (-1)^{\sum_{j=1}^n i_j} \Phi_{\mathbf{m}, \boldsymbol{\Sigma}}(\boldsymbol{\varepsilon}_i),$$

$$d(z_1, \dots, z_n) = \frac{\sum_{i=0}^{2^n - 1} \phi_{\mathbf{m}, \boldsymbol{\Sigma}}(\boldsymbol{\varepsilon}_i)}{2^n \sqrt{\prod_{i=1}^n z_i}},$$

where  $i_j \in \{0, 1\}$ ,  $i = \sum_{j=1}^n i_j 2^{j-1}$ ,  $\boldsymbol{\varepsilon}_i = \left( (-1)^{i_1} \sqrt{z_1}, \dots, (-1)^{i_n} \sqrt{z_n} \right)$  and  $\phi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$  denotes the Gaussian density function. Using (2) and (3) the copula and its density can be evaluated.

### 3.2 Parameter Estimation

Inference for the copula parameters and the parameters of the correlation function can be based on the maximum likelihood approach. In the case where

the copula density can be evaluated for all  $n \geq 2$  dimensions maximization of the likelihood is not difficult. However, as is the case for the non-central  $\chi^2$ -copula, it may occur that calculation of the copula density is infeasible for higher dimensions. Here we proceed to perform maximum likelihood estimation only with the bivariate copula densities. Under the assumption that different pairs of observations are treated as independent we have to maximize

$$l(\boldsymbol{\theta}; Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)) = \prod_{i,j \in \{1, \dots, N\}, i \neq j} c_{\mathbf{h}, \boldsymbol{\theta}}(F_Z(Z(\mathbf{x}_i)), F_Z(Z(\mathbf{x}_j))),$$

where  $\boldsymbol{\theta}$  is the parameter vector and  $c_{\mathbf{h}, \boldsymbol{\theta}}$  is the copula density. The procedure works well as long as one has no intention to estimate geometric anisotropy. An advantage of working with copulas that are constructed from elliptical distributions is that the correlation matrix explicitly appears in their analytical expression. If we parameterize the correlation matrix using a geostatistical covariance model, we no longer need to estimate a sill since it is equal to 1. The reason for this is that the overall variance of the random field is a property of the marginal distribution and the copula describes the dependence structure without information about the margins.

### 3.3 Goodness-of-Fit Testing for Spatial Copulas

For selecting a spatial copula model that suits to the given data we have to perform a goodness-of-fit test. We use a blanket test recently presented and validated by Genest and Remillard [6] and apply it to the different lag classes  $h_1, \dots, h_r$ . Although the test is designed for  $n$ -copulas we recommend to work only with bivariate copulas for simplicity. The test is based on a bootstrapping procedure and makes use of the Kolmogorov-Smirnov statistic,  $T_n$ , or the Cramer-von Mises statistic,  $S_n$ :

$$S_n = \int_{[0,1]^2} \mathbb{C}_n(\mathbf{u})^2 dC_n(\mathbf{u}) \quad \text{and} \quad T_n = \sup_{\mathbf{u} \in [0,1]^2} |\mathbb{C}_n(\mathbf{u})|,$$

where  $\mathbb{C}_n = \sqrt{n}(C_n - C_\theta)$ ,  $C_n$  is the empirical copula calculated using the  $n$  data points and  $C_\theta$  is the estimation under the null hypothesis. The steps of the algorithm are as follows:

1. For each of the lags  $h_1, \dots, h_r$  compute the empirical copula  $C_n^{h_1}, \dots, C_n^{h_r}$ .
2. Estimate the theoretical copula, for example using the maximum likelihood approach. Denote the estimated parameters by  $\boldsymbol{\theta}$ . For every lag class there is a corresponding theoretical copula  $C_\theta^{h_1}, \dots, C_\theta^{h_r}$ .
3. Calculate the Cramer-von Mises or the Kolmogorov-Smirnov statistic for every lag class,  $T_n^{h_1}, \dots, T_n^{h_r}$  or  $S_n^{h_1}, \dots, S_n^{h_r}$ .
4. For a large integer  $N$ , repeat the following steps for every  $k \in \{1, \dots, N\}$ 
  - a) Simulate a random field whose copula is exactly the estimated theoretical copula from step 2.

- b) Compute the empirical copula for every lag class,  $C_{n,k}^{h_1}, \dots, C_{n,k}^{h_r}$ .
  - c) Estimate the theoretical copula of the simulated field and denote the estimated parameters by  $\theta_k$ . For every lag class there is a corresponding theoretical copula,  $C_{\theta_k}^{h_1}, \dots, C_{\theta_k}^{h_r}$ .
  - d) Evaluate the test statistics  $T_{n,k}^{h_1}, \dots, T_{n,k}^{h_r}$  or  $S_{n,k}^{h_1}, \dots, S_{n,k}^{h_r}$ .
5. An approximate p-value for every lag class  $h_1, \dots, h_r$  is given by

$$p_{h_j} = \frac{1}{N} \sum_{k=1}^N I(S_{n,k}^{h_j} > S_n^{h_j}) \quad \text{or} \quad p_{h_j} = \frac{1}{N} \sum_{k=1}^N I(T_{n,k}^{h_j} > T_n^{h_j}),$$

where  $I(\cdot)$  is an indicator function and  $j = 1, \dots, r$ .

In the case where the spatial copula is constructed from a multivariate distribution, simulation of a random field with a predefined copula means simply simulating from the multivariate distribution.

## 4 Spatial Interpolation Using Copulas

After having modelled the spatial data we are interested in predicting the values of the random field at unknown locations. In the following we propose three different methods for performing spatial interpolation using copulas.

### 4.1 Indicator Kriging and Disjunctive Kriging

Indicator kriging is used to estimate the conditional distribution of the random field given the data. This is done by cokriging of indicator variables  $I(Z(x_i) \leq z_j)$ , where  $i = 1, \dots, n$  and the  $z_j$  are certain thresholds, e.g. quantiles. Bardossy [1] showed that bivariate copulas are related to indicator variograms and cross-variograms via

$$\begin{aligned} \gamma_{z_j}(h) &= F_Z(z_j) - C_h(F_Z(z_j), F_Z(z_j)), \\ \gamma_{z_j, z_k}(h) &= \min\{F_Z(z_j), F_Z(z_k)\} - C_h(F_Z(z_j), F_Z(z_k)). \end{aligned} \quad (6)$$

Plugging these relationships in the cokriging procedure, we arrive at an indicator kriging that is based on the spatial copula model. The fact that only bivariate copulas are needed makes it possible to use one of the numerous flexible copulas that do not have multivariate extensions or that have too few parameters for using them in a multivariate approach. The Gumbel-Hougaard extreme value copulas would be an example.

Indicators are not only used in indicator kriging but also in non-linear geostatistics. If the random field is discretized and takes only a finite number of values, say 1 to  $m$ , every function of  $Z(\mathbf{x})$  can be written as

$$f(Z(\mathbf{x})) = f_1 I(Z(\mathbf{x}) \leq 1) + \dots + f_m I(Z(\mathbf{x}) \leq m).$$

The disjunctive kriging estimator is calculated by cokriging of the indicators

$$[f(Z(\mathbf{x}))]^{DK} = f_1 [I(Z(\mathbf{x}) \leq 1)]^{CK} + \dots + f_m [I(Z(\mathbf{x}) \leq m)]^{CK}.$$

Again, the relationships (6) are used in the cokriging system. Rivoirard [9] argued that "in the same way that kriging is based on the variogram, so disjunctive kriging is based on the bivariate distributions". In our case the bivariate distribution of the random field is defined in terms of the bivariate copula and so is disjunctive kriging.

## 4.2 Rank-Order Kriging

Assume we have an isotropic random field with known univariate distribution  $F_Z$  and the bivariate distributions can be described by the copula  $C_h$ . Furthermore, we have a realization  $\{Z(\mathbf{x}_i) \mid \mathbf{x}_i \in \mathbf{S}\}$  of the random field and we want to predict the values at  $\mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+m}$ . Applying (4) we can calculate the Spearman rank correlation curve  $\rho$  as a function of  $h$  which is exactly the correlation function for the rank-transformed variable  $V(\mathbf{x}) = F_Z(Z(\mathbf{x}))$ . Since  $V(\mathbf{x})$  is a uniform distribution on  $[0, 1]$ ,  $\frac{\rho}{12}$  gives the corresponding covariance function. Journel and Deutsch [5] proposed to apply a simple kriging of ranks, where the linear predictor at the unknown locations  $\mathbf{x}_j, j = n+1, \dots, n+m$ , is given by

$$v^*(\mathbf{x}_j) = \sum_{i=1}^n \lambda_i v(\mathbf{x}_i) + \frac{1}{2} \left( 1 - \sum_{i=1}^n \lambda_i \right). \quad (7)$$

Since back-transforming  $v^*(\mathbf{x}_j)$  using  $F_Z^{-1}$  would lead to a biased estimate for  $Z(\mathbf{x}_j)$ , a bias correction is introduced

$$z^*(\mathbf{x}_j) = F_Z^{-1}(v^*(\mathbf{x}_j)) + \lambda(\mathbf{x}_j) [F_Z^{-1}(L(v^*(\mathbf{x}_j))) - F_Z^{-1}(v^*(\mathbf{x}_j))], \quad (8)$$

where  $L(\cdot)$  is the distribution function of all kriged values  $v^*(\mathbf{x}_j)$  and  $\lambda(\mathbf{x}_0) = \left( \frac{\sigma_K^2(\mathbf{x}_j)}{\sigma_{max}^2} \right)^\omega$  with  $\sigma_K^2$  being the kriging variance,  $\sigma_{max}^2$  being the maximal kriging variance of all estimations and  $\omega > 0$  a correction level parameter. Although this method reproduces the original distribution of the data and  $z^*$  is an unbiased estimate, the covariance structures of  $V(\mathbf{x})$  and  $Z(\mathbf{x})$  are not reproduced. Another disadvantage of rank-order kriging is that there is no guarantee for the estimated ranks to be in the interval  $[0, 1]$ . To ensure that all estimates are between 0 and 1 it is sufficient to force all kriging weights to be non-negative, however, this is accompanied by a loss in accuracy. Moreover,  $z^*$  has no minimum kriging variance, only  $v^*$  has that property.

To partially overcome these drawbacks a direct sequential simulation of the ranks at the kriging locations  $\mathbf{x}_j, j = n+1, \dots, n+m$ , is proposed. The simulated ranks are drawn from a uniform distribution with mean equal to the kriging predictor and variance equal to the kriging variance,  $[v^*(\mathbf{x}_j) - \sqrt{3}\sigma_K(\mathbf{x}_j), v^*(\mathbf{x}_j) + \sqrt{3}\sigma_K(\mathbf{x}_j)]$ . At each step, the kriging system consists of the original data and the previously sampled data. It may

occur that the endpoints of the uniform distribution are outside the  $[0, 1]$  interval leading to simulated ranks outside  $[0, 1]$ . In this case they have to be set to 0 or 1, depending on whether they are  $< 0$  or  $> 1$ . After simulation the bias correction described in (8) is applied to the estimated ranks. For a large number  $N$  the sequential simulation is repeated  $N$  times and the resulting predictors are back-transformed using  $F_Z^{-1}$  and averaged. This procedure yields estimates that are exact at known data locations, unbiased, follow the univariate distribution  $F_Z$  and reproduce the covariance of the random field. Sequential simulation is a time-consuming method for large data sets. Therefore, we adapt a method proposed by Saito and Goovaerts [7] who used it in the case of a normal-score transformation. Again, the simple kriging predictor,  $v^*(\mathbf{x}_j)$ , and the simple kriging variance,  $\sigma_K^2(\mathbf{x}_j)$ , are calculated. The conditional distribution of  $V(\mathbf{x}_j)$  given the data is modelled as a uniform distribution with mean equal to  $v^*(\mathbf{x}_j)$  and variance equal to  $\sigma_K^2(\mathbf{x}_j)$ . If the endpoints  $a$  and  $b$  of the uniform distribution are outside the  $[0, 1]$  interval, they are reset to 0 and 1, respectively, and the density of the local distribution changes to

$$d(x) = \begin{cases} \min \left\{ \frac{1}{2(v^*(\mathbf{x}_j) - a)}, \frac{1}{2v^*(\mathbf{x}_j)} \right\}, & \text{if } x \in [\max\{0, a\}, v^*(\mathbf{x}_j)], \\ \max \left\{ \frac{1}{2(b - v^*(\mathbf{x}_j))}, \frac{1}{2(1 - v^*(\mathbf{x}_j))} \right\}, & \text{if } x \in [v^*(\mathbf{x}_j), \min\{b, 1\}]. \end{cases}$$

The 100 percentiles,  $v_p(\mathbf{x}_j)$ , of this local distribution are calculated, where  $p = \frac{k}{100} - \frac{0.5}{100}$  and  $k = 1, \dots, 100$ . After back-transformation,  $z_p(\mathbf{x}_j) = F_Z^{-1}(v_p(\mathbf{x}_j))$ , the  $z_p(\mathbf{x}_j)$  are unbiased estimators for the quantiles of the local distribution of  $Z(\mathbf{x})$ . Their average is an unbiased estimator for the mean, hence, the kriging estimate is defined as

$$z(\mathbf{x}_j) = \frac{1}{100} \sum_{k=1}^{100} z_p(\mathbf{x}_j) \quad \text{with } p = \frac{k}{100} - \frac{0.5}{100}.$$

### 4.3 Plug-In Bayes Estimation

The copula enters the rank order kriging procedure only through the Spearman rank correlation. Furthermore, both rank order kriging and disjunctive kriging only use bivariate copulas. On the one hand these facts may be useful since flexible bivariate copula families can be applied, but on the other hand these methods do not fully exploit the spatial copula model presented in Sect. 3.1. When we go the Bayesian way, we can take account of the uncertainty of parameter estimation. Moreover, there is a predictive distribution for every rank-transformed variable  $V(\mathbf{x}_0)$  at an unknown location  $\mathbf{x}_0$ ,

$$p(v(\mathbf{x}_0) | \mathcal{D}) = \int p(v(\mathbf{x}_0) | \theta, \mathcal{D}) p(\theta | \mathcal{D}) d\theta, \quad (9)$$

where  $\mathcal{D} = \{Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n)\}$  denotes the set of all  $n$  known data values. When we (falsely) assume that the maximum likelihood estimates,  $\hat{\theta}$ , of the

copula parameters and the correlation function parameters are the true values, we get that  $p(v(\mathbf{x}_0) | \mathcal{D}) = p(v(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D})$ . In the spatial copula model this is exactly the density of the conditional copula of  $v(\mathbf{x}_0)$  given the rank-transformed data and the estimated parameters

$$p(v(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D}) = c_h(v(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D}) = \frac{c_h(v(\mathbf{x}_0), v(\mathbf{x}_1), \dots, v(\mathbf{x}_n) | \hat{\boldsymbol{\theta}})}{c_h(v(\mathbf{x}_1), \dots, v(\mathbf{x}_n) | \hat{\boldsymbol{\theta}})},$$

where  $v(\mathbf{x}_i) = F_Z(Z(\mathbf{x}_i))$  and  $i = 1, \dots, n$ . If the copula is constructed from a multivariate distribution with univariate conditional density  $d$  and univariate marginal distribution  $F$  with density  $f$ , (3) tells us that the conditional copula can be written as

$$c_h(v(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D}) = \frac{d(F^{-1}(v(\mathbf{x}_0)) | \hat{\boldsymbol{\theta}}, \mathcal{D})}{f(F^{-1}(v(\mathbf{x}_0)))}.$$

In the case of a Gaussian copula  $F = \Phi$ ,  $f = \phi$  and  $d = \phi_{\mu, \sigma^2}$  is a Gaussian density with mean  $\boldsymbol{\mu} = \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \mathbf{a}$  and variance  $\sigma^2 = 1 - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$ , where  $\mathbf{a} = (\Phi^{-1}(v(\mathbf{x}_1)), \dots, \Phi^{-1}(v(\mathbf{x}_n)))^T$ ,  $\boldsymbol{\Sigma}_{22}$  is the correlation matrix of the known locations and  $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T$  is the vector of correlations between the known locations and the location where prediction should take place.

Since the predictive density of  $V(\mathbf{x}_0)$  is defined on the unit interval, we avoid estimated ranks outside  $[0, 1]$ . Furthermore, the predictive density of  $Z(\mathbf{x}_0)$  can be calculated by just using a Jacobian transformation. To get back from the ranks to the original scale the transformation is  $F_Z^{-1}$ . The corresponding Jacobian determinant is exactly the density  $f_Z$ . Hence,

$$p(z(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D}) = c_h(F_Z(z(\mathbf{x}_0)) | \hat{\boldsymbol{\theta}}, \mathcal{D}) f_Z(z(\mathbf{x}_0)). \quad (10)$$

The Bayes estimator for  $Z(\mathbf{x}_0)$  under the quadratic loss is the mean of the predictive distribution,  $E(Z(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D})$ . Of course it can be evaluated using (10), but with the help of an integral transformation we can also derive an estimator similar to the one by Saito and Goovearts [7] which we have already used for rank order kriging:

$$\begin{aligned} E(Z(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D}) &= \int_{-\infty}^{\infty} z(\mathbf{x}_0) c_h(F_Z(z(\mathbf{x}_0)) | \hat{\boldsymbol{\theta}}, \mathcal{D}) f_Z(z(\mathbf{x}_0)) dz(\mathbf{x}_0) \\ &= \int_0^1 F_Z^{-1}(v(\mathbf{x}_0)) c_h(v(\mathbf{x}_0) | \hat{\boldsymbol{\theta}}, \mathcal{D}) dv(\mathbf{x}_0). \end{aligned}$$

Similarly to copula kriging the frequently applied trans-Gaussian kriging [4] also works with a marginal transformation of the random field. The aim of trans-Gaussian kriging is to deal with non-Gaussian random fields by assuming that the transformed random field,  $Y(\mathbf{x}) = g(Z(\mathbf{x}))$ , is Gaussian and  $g$

is a suitable transformation that has to be determined. In most applications the transformation  $g$  is chosen from the Box-Cox family of transformations. In the following we show that there is a direct relationship between the trans-Gaussian kriging model and the spatial copula model.

**Theorem 1.** *The trans-Gaussian kriging model using an almost surely strictly monotone transformation is equivalent to the Gaussian spatial copula model.*

*Proof.* Assume we have a trans-Gaussian random field with an almost surely strictly monotone transformation  $g$ . Hence,  $Y(\mathbf{x}) = g(Z(\mathbf{x})) \sim \mathcal{N}(\mu, \sigma^2)$ . From the invariance theorem mentioned in Sect. 2 we get that the copula corresponding to the multivariate distributions of  $Z(\mathbf{x})$  must be the Gaussian copula corresponding to  $Y(\mathbf{x})$ . Using  $Z(\mathbf{x}) = g^{-1}(Y(\mathbf{x}))$  we obtain the univariate marginal distribution of  $Z(\mathbf{x})$  as

$$F_Z(z) = \int_{-\infty}^z \phi_{\mu, \sigma^2}(g(t)) |g'(t)| dt,$$

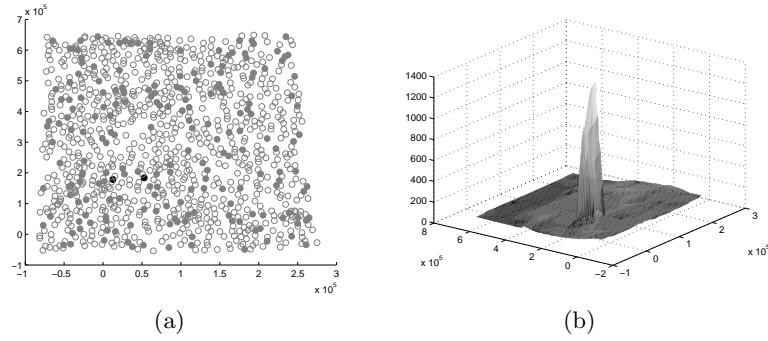
and the Gaussian spatial copula model is fully determined. If we assume that the random field follows the Gaussian spatial copula model with known  $F_Z$ , then  $g(z) := \Phi^{-1}(F_Z(z))$  is a suitable transformation.

Since we can also use any other copula different from the Gaussian copula in our approach, we get that the spatial copula model is a generalization of the trans-Gaussian model. Even if we want to stay within the Gaussian framework it is more convenient to use the copula methodology because it is easier to specify the univariate distribution of the random field than to determine a suitable transformation function. Especially when we work with multimodal or extreme value data this fact gets obvious.

The only minor drawback of the plug-in Bayes approach is that for certain copula families not all data values can be used to build the predictive distribution. In the case of the non-central  $\chi^2$ -copula mentioned in Sect. 3.1 this happens because one would need to evaluate  $2^n$  terms for the calculation of the conditional copula. Therefore, we propose to make a local prediction and to use the 15 data points that are closest to the location where prediction takes place.

## 5 Application: SIC2004 Joker Data

In this section we want to test our methodology by means of the Joker data set, which was investigated in detail during the spatial interpolation comparison SIC2004 [3]. This extreme value data set simulates an accidental release of radioactivity using a dispersion process. The 200 training points have a mean of 108.99, a standard deviation of 121.96 and a skewness of 9.92. Figure 1(a) displays the training data as gray dots and the 808 test data as gray circles.



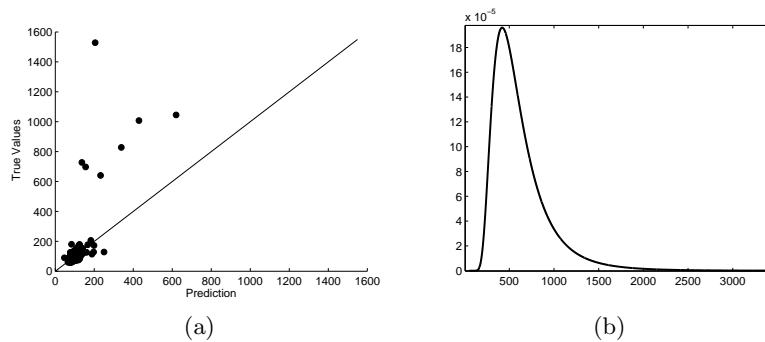
**Fig. 1.** The locations of the Joker training (dots) and test data (circles) are displayed in (a). A surface plot of the Joker data is shown in (b).

The two extreme observations (1070.4 and 1499) are indicated by the black dots.

At first we fit a quadratic trend surface model to the data. The residuals follow a generalized extreme value distribution. Using the goodness-of-fit test described in Sect. 3.3 we find out that it is sufficient to work with the Gaussian spatial copula. The correlation matrix of the Gaussian distribution is parameterized (cf. Sect. 3.1) by a mixture of a Gaussian and an exponential correlation model. Geometric anisotropy is considered by a  $2 \times 2$  transformation matrix, where one entry is fixed to avoid any interference with the ranges of the correlation models. All parameters, including the nugget, two ranges, one mixing parameter and three anisotropy parameters, are estimated using the maximum likelihood approach. Note that there is no need to estimate a sill. Prediction is performed using the plug-in Bayes approach. The predicted values are plotted against the true values in Fig. 2(a). The predictive density at a hotspot is visualized in Fig. 2(b) and it shows that values around 1500 have still some probability. The results for the test data are: RMSE=65.87, MAE=16.22, ME=-2.58 and Pearson-r=0.71. These results would be the third best in terms of RMSE and Pearson correlation and the second best in terms of MAE when compared to the methods applied in the SIC2004 exercise.

## 6 Conclusion

Copulas can be used to describe spatial dependence and in this way generalize the concept of the variogram. Moreover, the spatial copula model can be used to perform spatial interpolation. It generalizes the trans-Gaussian kriging method and is therefore a flexible tool for working with non-Gaussian, multimodal and extreme value data. Specifying the univariate distribution of the random field is easier than finding a suitable transformation for trans-Gaussian kriging, which makes the spatial copula model attractive even if



**Fig. 2.** The predicted values of the test set are plotted against the true values in (a). The predictive density at a hotspot is displayed in (b).

the Gaussian copula is used. Another advantage is that the sill need not to be estimated and, hence, the model contains one parameter less. Results on the SIC2004 Joker data demonstrate that the presented approach could be applied for emergency monitoring and estimating exceedance probabilities for certain emergency thresholds in environmental monitoring systems.

## References

1. Bardossy, A. (2006), Copula-Based Geostatistical Models for Groundwater Quality Parameters, *Water Resources Research*, 42, W11416
2. Bondar, I., McLaughlin, K. & Israelsson, H. (2005), Improved Event Location Uncertainty Estimates, *27th Seismic Research Review*, 299-307
3. Dubois, G. (2005), Automatic Mapping Algorithms for Routine and Emergency Monitoring Data - Spatial Interpolation Comparison 2004, European Commission Joint Research Centre, Belgium
4. Diggle, P. & Ribeiro, P. (2007), *Model-Based Geostatistics*, Springer, New York
5. Journel, A. & Deutsch, C. (1996), Rank Order Geostatistics, In: *Geostatistics Wollongong '96* (E. Baafi and N. Schofield, Eds.), Kluwer, Dordrecht, 174-187
6. Genest, C. & Remillard, B. (2008), Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models, *Annales de l'Institut Henri Poincaré: Probabilités et statistiques*, 44, (in press)
7. Saito, H. & Goovaerts, P. (2000), Geostatistical Interpolation of Positively Skewed and Censored Data in a Dioxin-Contaminated Site, *Environmental Science & Technology*, 34, 4228-4235
8. Nelsen, R. (2006), *An Introduction to Copulas*, Springer, New York
9. Rivoirard, J. (1994), *Introduction to Disjunctive Kriging and Non-Linear Geostatistics*, Oxford University Press, Oxford
10. Sklar, A. (1959), Fonctions de repartition a  $n$  dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, 8, 229-231