Geostatistical modeling using non-gaussian copulas

Hannes Kazianka
Department of Business Mathematics
Vienna University of Technology
Vienna, Austria
hannes.kazianka@tuwien.ac.at

Jürgen Pilz
Department of Statistics
Alpen-Adria University of Klagenfurt
Klagenfurt, Austria
juergen.pilz@uni-klu.ac.at

Abstract—**Copula-based spatial models have recently attracted much attention and are used as a flexible tool for spatial interpolation. For computational reasons, in most applications only the radially symmetric Gaussian copula is employed. However, radial asymmetry is a property often observed in environmental data i.e. high values of the data have a stronger spatial dependence than low values. This paper presents a case study where radiological measurements have been taken in the region of Gomel, near Tschernobyl. We show that copula models that are based on the radially asymmetric chi-squared-copula outperform the Gaussian copula models and classical interpolation methods like ordinary kriging.**

Keywords: copula; spatial interpolation; radial asymmetry; Bayesian prediction; predictive distribution

I. INTRODUCTION

Spatial interpolation methods which are based on the Gaussian assumption lead to unreliable results when applied to environmental data that exhibit a non-Gaussian dependence structure or a non-Gaussian and possibly extreme value univariate marginal distribution. Specifically, they report wrong estimates about the uncertainty associated with the interpolation. A number of methods have been proposed for the analysis of non-Gaussian data. Transformed-Gaussian kriging (Pilz and Spöck 2008), for example, addresses the case of a non-Gaussian univariate marginal distribution by applying a Box-Cox transformation. The dependence structure, however, remains unchanged by this transformation and shows a radially symmetric behavior i.e. high and low quantiles of the distribution have equal dependence properties.

The use of copula functions provides a flexible way of separately specifying both the dependence structure and the univariate marginals of any multivariate distribution. Copula-based spatial modeling for isotropic random fields with continuous univariate marginals was first proposed by Bardossy (2006). His work was extended to spatial interpolation by Bardossy and Li (2008) and Kazianka and Pilz (2010). The latter explain how to jointly estimate the model parameters and present an approach that works with both continuous and discrete margins. Moreover, they show how to include covariates e.g. a spatial trend or elevation. Kazianka and Pilz (in press) incorporate the copula-based spatial methodology into the Bayesian framework and use an MCMC algorithm to sample from the posterior distribution.

Due to the model assumptions, only a limited number of copula families have been proposed for the use in the spatial approach. To date, only Bardossy and Li (2008) have applied a non-Gaussian copula for spatial interpolation, namely the noncentral chi-squared copula. However, in their comparative study the advantage of using the radially asymmetric non-Gaussian copula instead of the radially symmetric Gaussian copula seems to be insignificant to us. In this paper we review the processes of parameter estimation and spatial prediction in the copula-based spatial model. Special attention is drawn to the case of non-Gaussian copulas. Furthermore, we present a case study in which models based on the chi-squared copula outperform the Gaussian copula models and we provide empirical evidence why the non-Gaussian copula is superior.

This work was partially funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811with DG INFSO, Action Line IST-2005-2.5.12 for Environmental Risk Management. The views expressed herein are those of the authors and not necessarily those of the European Commission.

II. THE SPATIAL COPULA MODEL

A. Copulas

Copula functions are distribution functions on the n-dimensional unit cube with uniformly distributed univariate marginals. Sklar's Theorem states that for any multivariate distribution, H, with univariate marginals $F_1,\ldots,F_n$ there exists a copula, C, such that

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

If $F_1^{-1},\ldots,F_n^{-1}$ denote the inverse distribution functions the copula may be written as

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

In case of an absolutely continuous copula, its density can be expressed in terms of the density of H, here denoted as h, and the densities of the univariate marginals, $f_1,\ldots,f_n$,

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

The Gaussian copula emerges when $F_1=\ldots=F_n=\Phi$, the distribution function of the univariate standard Gaussian distribution, and $h=\Phi_{0,R}$, where $\Phi_{0,R}$ denotes the multivariate Gaussian distribution with mean 0 and correlation matrix R:

$$C_R^G(u_1,\ldots,u_n) = \Phi_{0,R}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n)). \quad (4)$$

For a more elaborate introduction to copulas the interested reader is referred to Kazianka and Pilz (2010) or the book by Nelsen (2006).

B. Copula-based Spatial Modeling and Interpolation

The concept of specifying a multivariate distribution in terms of its copula and its univariate marginals can be beneficially used in spatial statistics. In the following we use the notation introduced by Kazianka and Pilz (2010) and describe how the methodology is implemented in the `R` statistical software as part of the `intamap` R-library which is freely available from www.r-project.org.

Assume that $\{Z(x)|x \in S\}$ is a weakly isotropic random field where S is the area of interest and let $x_1,\dots,x_n \in S$ be distinct observation locations. Recalling that a random field is uniquely determined by all multivariate distributions, Sklar's Theorem tells us that the relation between $Z(x_1),\dots,Z(x_n)$ is characterized by

$$P(Z(x_1) \le z_1,\dots,Z(x_n) \le z_n) = C_{\theta,\lambda}(F_\eta(z_1),\dots,F_\eta(z_n))$$

where $C_{\theta,\lambda}$ denotes an n-dimensional copula and the parameters θ and λ control its dependence structure. The parameters λ are copula-specific while the parameters θ control an isotropic correlation function model that makes the copula a function of the distances between the sampling locations. Moreover, F_η denotes a univariate distribution function with parameter vector η, which is the same at every observation location due to the stationarity.

The copula-based spatial model is a flexible extension of the classical Gaussian random field model in spatial statistics. Indeed, when we choose the copula $C_{\theta,\lambda}$ to be the Gaussian copula and $F_\eta = \Phi_{\mu,\sigma}$, the resulting random field is Gaussian. Here, μ and σ denote the mean and the standard deviation, respectively. The Gaussian copula is parameterized by assuming that its correlation function follows one of the classical isotropic models with parameters θ, e.g. an exponential or a Gaussian correlation model. Choosing the univariate marginal distribution to be non-Gaussian or the copula to be different from the Gaussian copula makes it possible to take account of non-Gaussian spatially varying phenomena.

In the case of continuous univariate marginals a maximum likelihood approach can be used to estimate the parameters $\Theta=(\theta,\lambda,\eta)$. Typically, numerical optimization routines are employed for maximizing

$$c_{\theta,\lambda}(F_\eta(Z(x_1)),\dots,F_\eta(Z(x_n)))\prod_{i=1}^{n} f_\eta(Z(x_i))$$

with respect to Θ. Another possibility for parameter inference would be to use the Inference Functions for Margins (IFM) method (Joe, 1997). This procedure is computationally simpler but is far less efficient. We recommend using the maximum likelihood approach whenever possible.

Plug-in spatial prediction is performed by evaluating the conditional distribution of $Z(x_0)$ given $\boldsymbol{Z}=(Z(x_1),\dots,Z(x_n))$ and the parameter estimates for Θ, where x_0 is the location of interest. The predictive density is given by

$$p(Z(x_0)|\hat{\Theta},\boldsymbol{Z}) = c_{\hat{\theta},\hat{\lambda}}(F_{\hat{\eta}}(Z(x_0))|\boldsymbol{Z}) f_{\hat{\eta}}(Z(x_0)),$$

where $c_{\hat{\theta},\hat{\lambda}}(F_{\hat{\eta}}(Z(x_0))|\boldsymbol{Z})$ denotes the conditional copula. If the copula is constructed from a multivariate distribution with conditional density d and marginal distribution F with density f, the conditional copula may be written as

$$c_{\hat{\theta},\hat{\lambda}}(F_{\hat{\eta}}(Z(x_0))|\boldsymbol{Z}) = \frac{d(F^{-1}(F_{\hat{\eta}}(Z(x_0)))|\hat{\theta},\hat{\lambda},\boldsymbol{Z})}{f(F^{-1}(F_{\hat{\eta}}(Z(x_0))))}.$$

Expected value and variance of the predictive distribution are obtained by numerical integration over the unit interval

$$\hat{Z}(x_0) = \int_0^1 F_{\hat{\eta}}^{-1}(u) c_{\hat{\theta},\hat{\lambda}}(u|\boldsymbol{Z})du,$$

$$\hat{\sigma}^2(x_0) = \int_0^1 (F_{\hat{\eta}}^{-1}(u) - \hat{Z}(x_0))^2 c_{\hat{\theta},\hat{\lambda}}(u|\boldsymbol{Z})du.$$

C. Working with Non-Gaussian Copulas

The Gaussian copula is radially symmetric, i.e. $c_\Sigma^G(u_1,u_2) = c_\Sigma^G(1-u_1,1-u_2)$. This restriction implies that high and low quantiles of any distribution possessing a Gaussian copula have equal dependence properties. For this reason non-Gaussian radially asymmetric copula families have been proposed for spatial modeling. Amongst them, the noncentral chi-squared copula (Bardossy, 2006) is most well-known. When $(\chi^2_{1,\lambda})^{-1}$ denotes the inverse distribution function of a noncentral chi-squared distribution with 1 degree of freedom and non-centrality parameter $\lambda>0$, $i_j \in \{0,1\}$, $i = \sum_{j=0}^{2^n-1} i_j 2^{j-1}$, and $\lambda^{0.5} = (\sqrt{\lambda},\dots,\sqrt{\lambda})$ $\varepsilon_i = \left((-1)^{i_1}\sqrt{(\chi^2_{1,\lambda})^{-1}(u_1)},\dots,(-1)^{i_n}\sqrt{(\chi^2_{1,\lambda})^{-1}(u_n)}\right)$ the non-central chi-squared copula is given by

$$C^{\chi^2}_{\lambda,\Psi}(u_1,\dots,u_n) = \sum_{i=0}^{2^n-1}(-1)^{\sum_{j=1}^n i_j}\Phi_{\lambda^{0.5},\Psi}(\varepsilon_i).$$

The additional parameter λ controls the degree of asymmetry. The smaller we choose λ, the more asymmetric the copula becomes. Kazianka and Pilz (in press) show that the noncentral chi-squared copula and the multiplied Gaussian copula (Kazianka and Pilz 2010) share the disadvantage of having zero tail dependence with the Gaussian copula. This means that "as we go far enough into the tail of the distribution, extreme events appear to occur independently in the marginals" (Embrechts et al. 2002). Therefore, neither of these copulas is particularly suitable for dealing with extreme data even when an extreme value marginal distribution is used.

If the number n of observations is large, the computation of the noncentral chi-squared copula density is impossible since we have to sum over 2^n terms. In this case we use a maximum composite likelihood approach for parameter estimation. Under mild conditions the estimates obtained from maximizing the pairwise likelihood

$$\prod_{i,j\in\{1,\ldots,n\},i\neq j} c_{\theta,\lambda}\left(F_\eta(Z(x_i)),F_\eta(Z(x_j))\right)\prod_{k\in\{i,j\}} f_\eta(Z(x_k))$$

are consistent and asymptotically normally distributed.

III. Analysis of the Gomel Data

In this section we analyze a data set here referred to as the Gomel data using the previously presented methodology. It contains Cs137 measurements, $Z(x_i)$, at 148 locations $x_i=(x_{1i},x_{2i})$, $i=1,\ldots,148$, in the region of Gomel, Belarus. The data were observed in 1996, ten years after the Chernobyl accident and have previously been used by Pilz and Spöck (2008) to illustrate Bayesian transformed-Gaussian kriging. Fig. 1 shows the observation locations as blue dots and a regular interpolation grid as red circles. The radii of the dots are proportional to the measured values. The arithmetic mean of the data is 4.32, the median is 1.65 and there is a skewness of 2.54. This already indicates that the distribution of the data is right-skewed and far from being Gaussian.

Methods for finding a suitable family of marginal distributions are, for example, predictive model checking and the likelihood ratio test (Kazianka and Pilz (in press)). For simplicity we use the latter approach and the Gaussian spatial copula model with exponential correlation function for determining an appropriate marginal distribution family. A likelihood ratio test picks the family that leads to the largest maximized likelihood value. When comparing the Gaussian, the log-Gaussian and the gamma distribution we obtain the following log-likelihood values: -285.38 for the log-Gaussian, -288.85 for the gamma and -432.30 for the Gaussian distribution. Therefore, the evidence that the log-Gaussian distribution fits best is strong (Robert 2007) and we use it in the subsequent analysis.

Exploratory data analysis must be based on bivariate copulas. Fig. 2 displays scatterplots of pairs of rank-transformed data for four different lag classes in the left column. As can be seen from these plots, the data exhibit radial asymmetry, especially for the lag classes h_2=10-20 and h_3=20-40. This already indicates that a model that is based on a non-Gaussian and radially asymmetric copula should fit better than the Gaussian copula model. The scatterplot for lag class h_1=0-10 shows that measurements from locations with a small separating distance have a strong dependence, as would be expected. Furthermore, pairs of values from lag class h_4=40-60 have almost no spatial dependence and are distributed uniformly over the unit square.

Using the maximum likelihood and the maximum composite likelihood approach we fit the Gaussian copula model with exponential correlation function and the noncentral chi-squared copula model with Gaussian correlation function, respectively. For each copula, we choose the correlation function model that maximizes the likelihood. The Gaussian, the exponential and the spherical model were tested. We investigate the two different copula models with and without taking account of geometric anisotropy. For the Gaussian spatial copula model we additionally perform a Bayesian analysis (Kazianka and Pilz (in press)) by running a Metropolis algorithm.

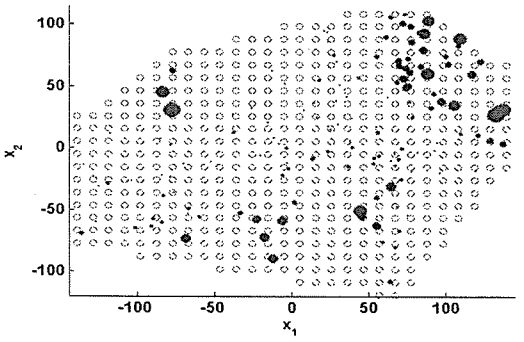


Figure 1. Observation locations (blue) and interpolation grid (red).

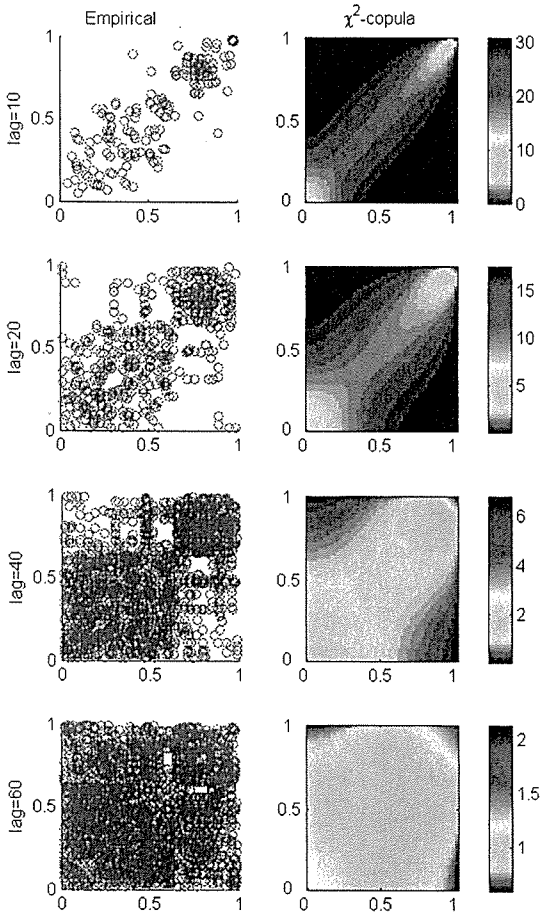
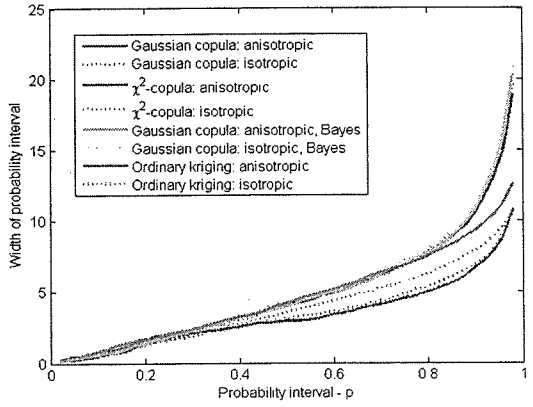
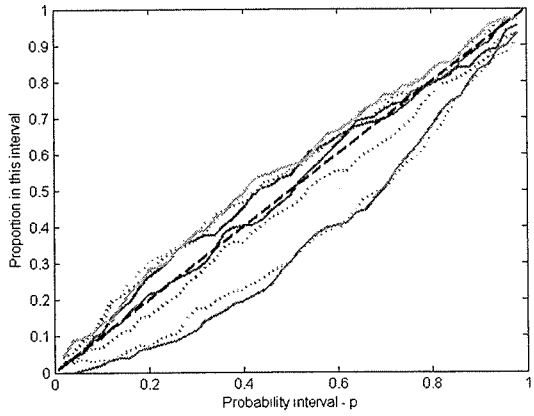


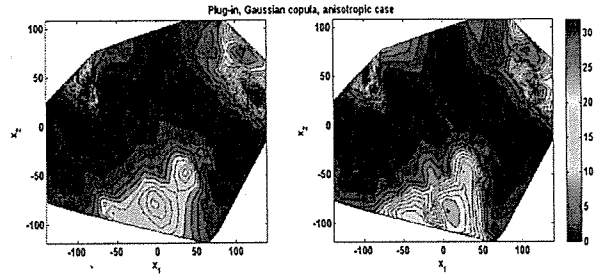
Figure 2. Scatterplot of rank-transformed pairs of observations for different lag distances (left) and density of the noncentral chi-square copula according to the spatial copula model (right).



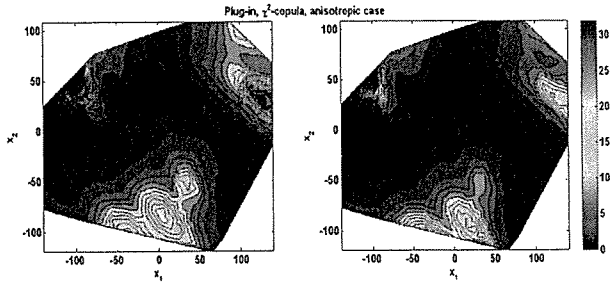
A



B



C



D

Figure 3. A) Width and B) coverage of cross-validation confidence intervals. Predictive mean and standard deviation for C) Gaussian and D) chi-squared model.

For a qualitative judgment of the predictive performance we examine the average width and coverage of median-centered cross-validation prediction intervals (Goovaerts 2001) bounded by the *(1-p)/2* and *(1+p)/2* quantiles of the predictive distribution. Fig. 3 (a)-(b) reveal that the chi-squared copula model including geometric anisotropy (red curve) leads to the smallest confidence intervals while being the only model with a percentage of coverage that is always very near to the nominal level. All models that are based on the Gaussian copula lead to too long confidence intervals and therefore overestimate the prediction uncertainty. This can also be seen when looking at the southern and north-eastern part of the study area in Fig. 3 (c)-(d). As one would expect for right-skewed data, ordinary kriging clearly performs worst and underestimates the prediction uncertainty.

Taking the mean of the predictive distribution as the predictor we calculate the median absolute deviation (MAD) and the median of the prediction errors (MPE) as robust measures of the overall precision and the bias, respectively. Table 1 points out that the chi-squared copula model performs best and that including anisotropy is beneficial.

TABLE I. QUANTITATIVE MEASURES OF PREDICTION PERFORMANCE

MAD/MPE	*χ^2-copula*	*Gaussian copula*	*Gaussian copula Bayes*	*Ordinary kriging*
Isotropic	0.70 / 0.06	0.87 / 0.43	0.87 / 0.48	2.23 / 2.00
Anisotropic	0.64 / 0.16	0.85 / 0.43	0.83 / 0.43	2.75 / 2.36

REFERENCES

Bardossy, A. (2006). Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*. 42 (W11416).

Bardossy, A., and Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research*. 44 (W07412).

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In DempsterM. (Ed) *Risk management: Value at risk and beyond* (pp. 176-223). Cambridge: Cambridge University Press.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma*. 103, 3-26.

Joe, H. (1997). *Multivariate models and dependence concepts*. Boca Raton: Chapman and Hall/CRC.

Kazianka, H., and Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*. Doi: 10.1007/s00477-009-0353-8.

Kazianka, H., and Pilz, J. (in press). Bayesian spatial modeling and interpolation using copulas. *Computers and Geosciences*.

Nelsen, R. (2006). *An introduction to copulas*. New York: Springer.

Pilz, J., and Spöck, G. (2008). Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*. 22, 621-632

Robert, C. (2007). *The bayesian choice*. New York: Springer