UBC Spatial Stats Course VI

Jürgen Pilz

Institut für Statistik
Universität Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
juergen.pilz@uni-klu.ac.at

November 30, 2010 / UBC Vancouver

2nd Extension to Non-Gaussianity: GLGM

- Diggle and Ribeiro (2007):

  GLGM = Generalized Linear Geostatistical Model

  $g(\mu_i) = c_i^T \beta + S(x_i)$, where

  $S(x_i)$ = random spatial effects

- R-Software: library(**geoR**), library(**geoRglm**)

- Bayesian Hierarchical GLM framework in Banerjee et al. (2004) **BHGLM**

Going beyond GLM and BHGLM

- How to proceed with distributions not covered by the GLM-framework?
- How to deal with extreme events coming from a different random process as opposed to the „background“ process?

Modelling of heavy-tailed/**extreme-value** distributions:
These are distributions which have a much slower decay of probability in the tails

Remember: Exponential prob. decay for normal distribution

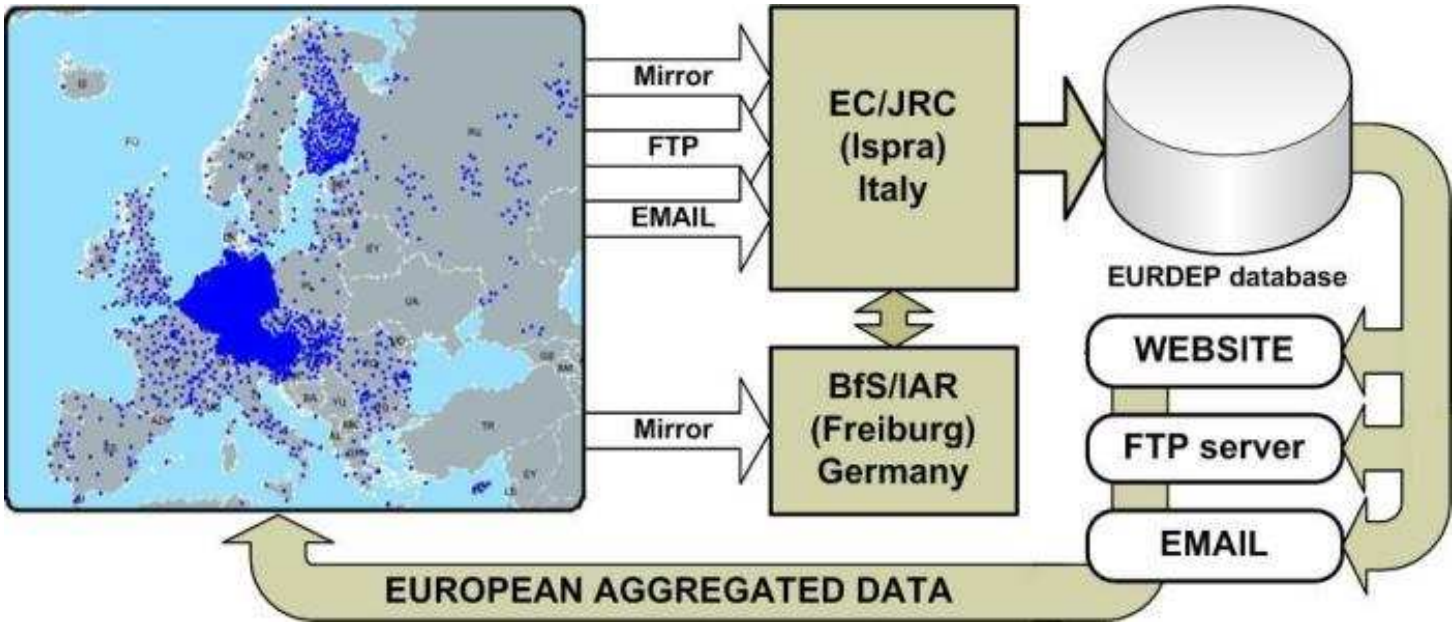
GEV: $Z = Z(x)$ has c.d.f.

$$F(z; \mu, \sigma, \tau) = \exp\left(-\left[1 - \tau\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\tau}\right)$$

Main challenges in the EU-Project INTAMAP

- Development of new methodology for Bayesian spatial interpolation for non-Gaussian observations, with specific attention to extreme-value distributions
- Development of new methods for characterizing spatial dependence structures (based on copulas)
- Implementation (existing + new methods) in case of non-Gaussian, skewed and heavy-tailed observations:

  library(intamap)
  Library of R-functions


Mirror
FTP
EMAIL
EC/JRC
(Ispra)
Italy
EURDEP database
WEBSITE
FTP server
EMAIL
Mirror
BfS/IAR
(Freiburg)
Germany
EUROPEAN AGGREGATED DATA

Briefly on Copulas

Copulas: distribution functions on the unit cube $[0,1]^n$ with uniformly distributed margins, introduced by Abe Sklar.

Sklar's Theorem: Let H be an n-dimensional distribution function with margins $F_1, \ldots, F_n$. Then there exists an n-dimensional copula C such that for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in R^n$

$$H(x_1, \ldots, x_n) = C\left(F_1(x_1), \ldots, F_n(x_n)\right).$$

If $F_1, \ldots, F_n$ are all continuous, then C is unique. Conversely, if C is an n-copula and $F_1, \ldots, F_n$ are d. f.s, then H is an n-dimensional d. f. with margins $F_1, \ldots F_n$, and

$$C\left(u_1, \ldots, u_n\right) = H\left(F_1^{-1}\left(u_1\right), \ldots, F_n^{-1}\left(u_n\right)\right).$$

Main Properties of Copulas

- Copulas describe the dependence between the quantiles of random variables
- Copulas are **invariant** under strictly increasing transformations of the marginals: Thus, frequently applied data transformations (e.g. square root and log transformations) do not change the copula.
- Random variables $X_1, \ldots, X_n$ are stochastically independent if and only if their copula is the product copula

$$\Pi^n(\boldsymbol{u}) = \prod_{i=1}^{n} u_i.$$

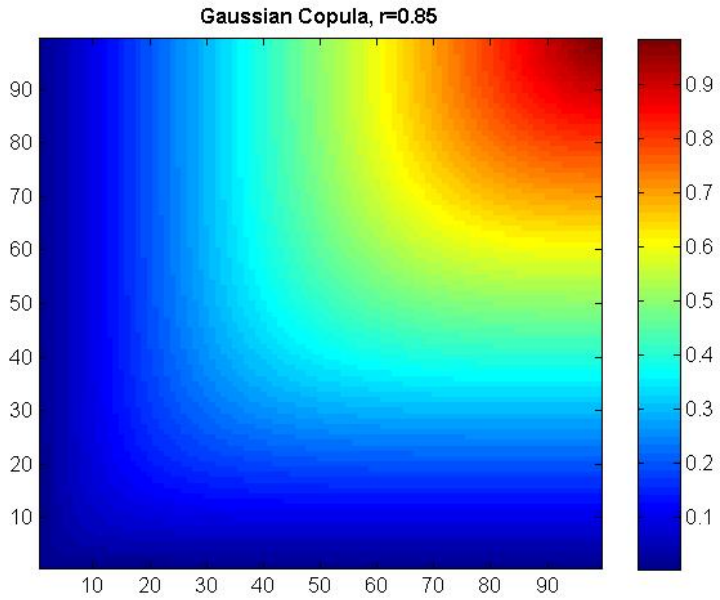
- Furthermore: (conditional) copula densities available

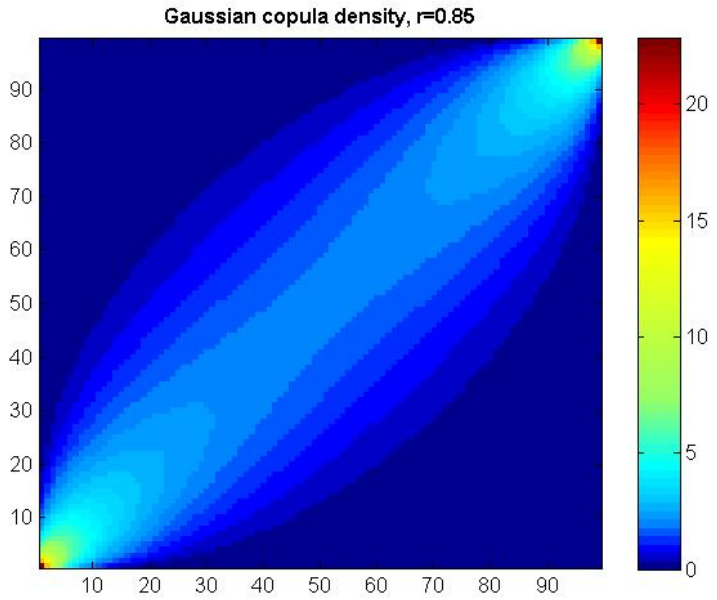
The Gaussian Copula

Sklar's Theorem provides a simple way of constructing copulas from multivariate distributions: Suppose $H = \mathbf{\Phi_{0,\Sigma}}$ and $F_1 = \cdots = F_n = \Phi$, then

$$C^G_{\mathbf{\Sigma}}(u_1, \ldots, u_n) = \mathbf{\Phi_{0,\Sigma}}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\right)$$

is called the Gaussian copula.





Spatial Modeling using Copulas

- Bardossy (2006) was the first to present a method for spatial modeling using copulas
- The relation between two locations separated by the vector $\boldsymbol{h}$ is characterized by the bivariate distribution

$$P\left(Z\left(\boldsymbol{x}\right) \leq z_1, Z\left(\boldsymbol{x}+\boldsymbol{h}\right) \leq z_2\right) = C_{\boldsymbol{h}}\left(F_Z\left(z_1\right), F_Z\left(z_2\right)\right)$$

  The copula thus becomes a function of the separating vector $\boldsymbol{h}$
- Spatial copulas describe spatial dependence over the whole range of quantiles for a given separating vector $\boldsymbol{h}$, and not only the mean dependence as the variogram does.
- H. Kazianka made enormous progress in this direction

Incorporating Spatial Trend

- Natural extension to incorporate trend: parameterize F_Z
- If F_Z belongs to the class of exponential dispersion models, $ED(\mu, \sigma)$, the mean μ appears explicitly inside the analytical expression of the density f_Z and we can set

$$g(\mu_i) = \eta(\boldsymbol{c}_i) = \boldsymbol{c}_i^T \beta, \quad i = 1, \ldots, n$$

  where g is the **link function**, $\eta(\boldsymbol{c}_i)$ is the response surface, $\boldsymbol{c}_i$ is the vector of **covariates** corresponding to location $\boldsymbol{x}_i$ and β is the vector of regression parameters.
- For other models, e.g. the GEV, parameterize the location parameter μ of the distribution in the same way.

Relation to Diggle's GLGM

- Copula-based spatial modeling approach is an alternative to the model-based approach of Diggle (1998, 2007)
- Their approach can be easily extended to include mixed linear effects:

$$g(\mu_i) = \boldsymbol{c}_i^T \boldsymbol{\beta} + S(\boldsymbol{x}_i), \quad i = 1, \ldots, n$$

  where $S(\cdot)$ is a stationary Gaussian process with mean zero, variance σ^2 and given correlation function $\rho(\cdot)$
- The main difference: in GLGM only μ_i is modeled while we build a complete multivariate distribution for the response variables $Z(\boldsymbol{x}_i); i = 1, \ldots, n$

Parameter Estimation for Cont. Margins

- Parameter vector $\mathbf{\Theta} = (\theta, \boldsymbol{\lambda}, \boldsymbol{\eta})$, where
  θ = correlation function parameters, $\boldsymbol{\lambda}$ = copula parameters and
  $\boldsymbol{\eta}$ = parameters of the known fam. of univ. distributions F
- Likelihood of the data $Z = (Z(x_1), \ldots, Z(x_n))^T$:

$$l(\mathbf{\Theta}; Z) = c_{\theta,\boldsymbol{\lambda}}\left(F_{\boldsymbol{\eta}}\left(Z\left(x_1\right)\right), \ldots, F_{\boldsymbol{\eta}}\left(Z\left(x_n\right)\right)\right) \prod_{i=1}^{n} f_{\boldsymbol{\eta}}\left(Z\left(x_i\right)\right)$$

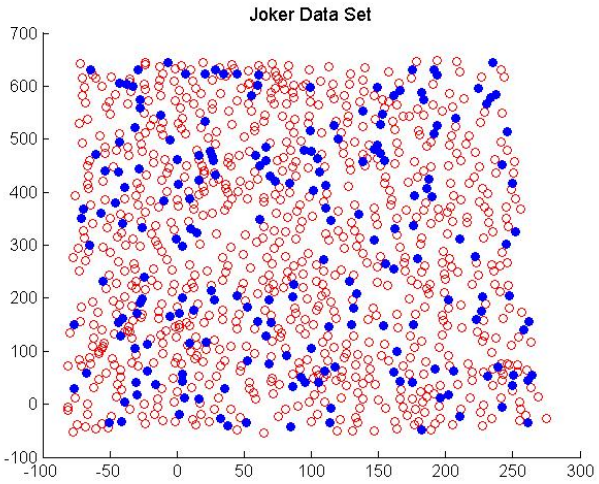
- For copulas different from the Gaussian c. the evaluation of the density is infeasible in higher dimensions. Here we proceed with **composite ml** using the bivariate copula densities, assuming that different pairs of observations can be treated as independent.

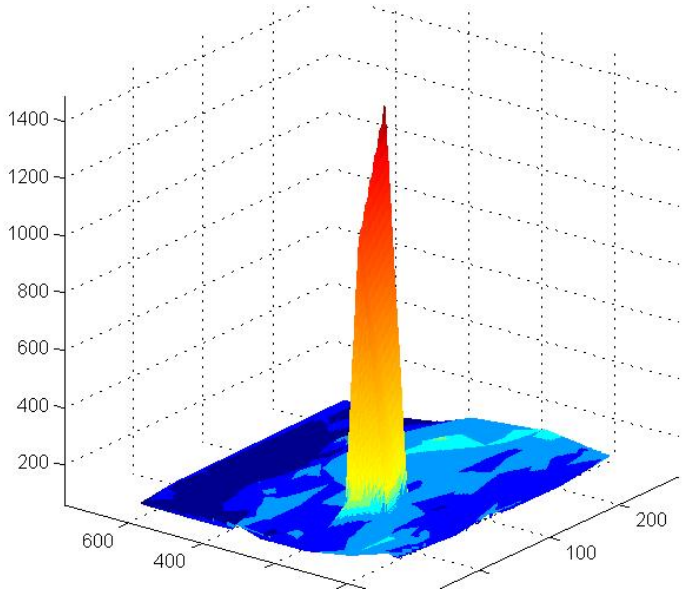
Bayesian Spatial Modeling

- Use a proper prior for $\boldsymbol{\Theta}$ to assure that the posterior distribution, $p(\boldsymbol{\Theta} \mid Z)$, is proper too. Otherwise, one has to prove the propriety of the posterior, which can be a difficult task
- For computation of the posterior d. of $\boldsymbol{\Theta} = (\theta, \boldsymbol{\lambda}, \boldsymbol{\eta})$, we use an MH-Algorithm
- Let $\hat{\boldsymbol{\Theta}}^{(i)} = \left(\hat{\theta}^{(i)}, \hat{\boldsymbol{\lambda}}^{(i)}, \hat{\boldsymbol{\eta}}^{(i)}\right)$ denote the i-th sample from the posterior. The approximation to the predictive density is

$$p(Z(x_0) \mid Z) \approx \frac{1}{N} \sum_{i=1}^{N} p(Z(x_0) \mid Z, \hat{\boldsymbol{\Theta}}^{(i)})$$

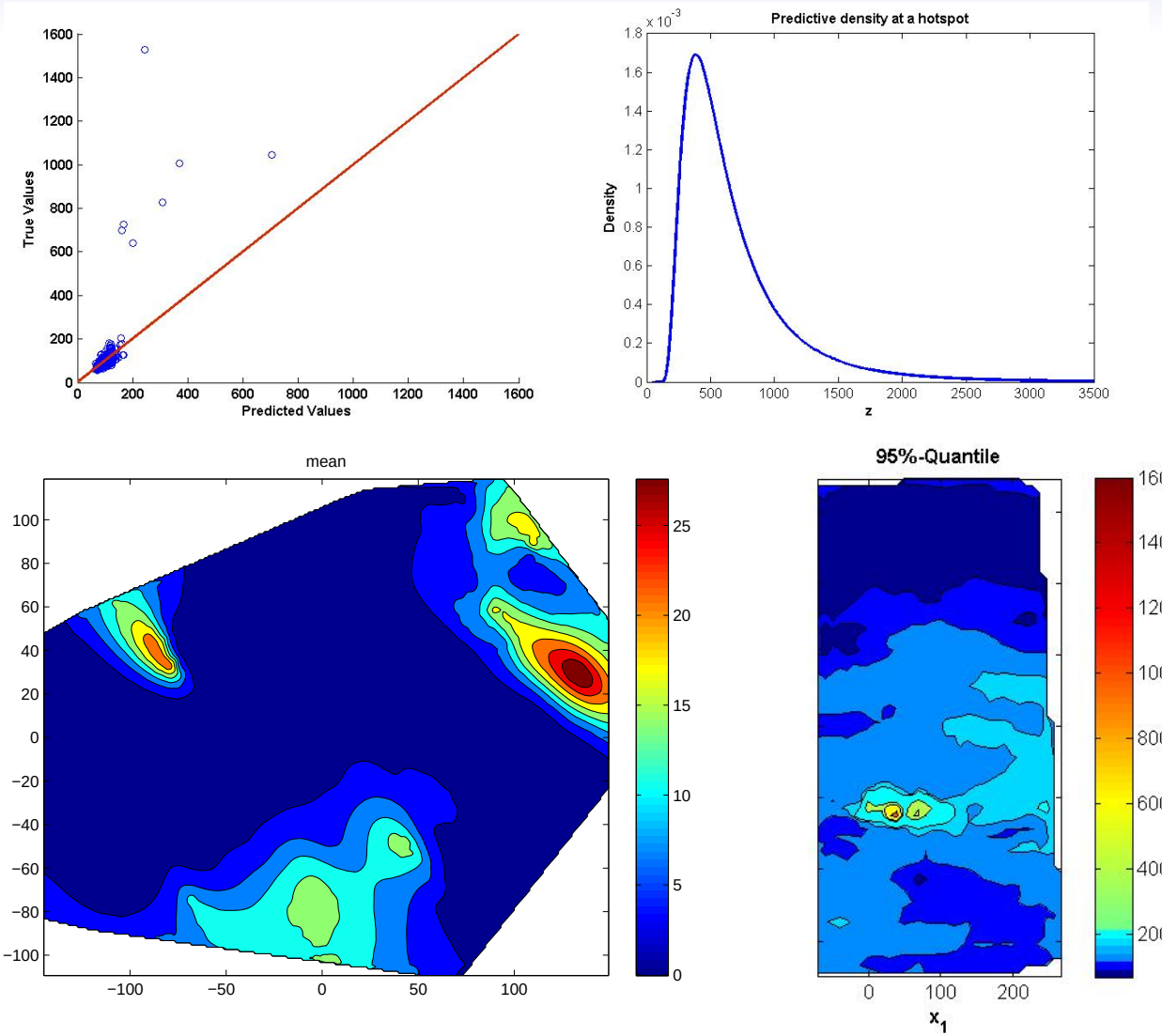
SIC2004 Joker Data Set





- Extreme value data set from the Spatial Interpolation Contest 2004 (Dubois)
- 200 training data and 808 test data. The training data include two very large observations (1070.4 and 1499) and have a sample **skewness** of **9.92**
- We assume a **GEV** d. as the marginal d.
- The marginal location parameters are specified as $\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$.
- Fit a Gaussian copula with correlation function = mixture of a Gaussian and an exponential model.
  Geometric anisotropy is also considered.
- RMSE=64.47, MAE=15.83, ME=-2.56, Pearson r=0.72
  Better results than with trans-Gaussian kriging using Box-Cox or Log-Log transformation and other classical geostatistical applications.

Spatial Interpolation of SIC2004 Joker Data



Conclusions

- Slight to moderate deviations from Gaussianity can be modelled with (Bayesian) Trans-Gaussian Kriging, using Box-cox-transf., GLM-framework of Diggle & Ribeiro (geoRglm) or Aston's **library(psgp)** or using Gaussian copulas (copula package)
- Heavy-tailed/Extreme value distributions we propose to model with function **copulaEstimation** in **library(intamap)**
- Code for estimation of parameters and prediction at unobserved locations in the case of the Gaussian spatial copula model is part of this library: **spatialPredict**
- Automatic choice of marginal distribution, correlation function model, anisotropy and starting values by using certain heuristics.
- Copula package not yet ready for **real**-time monitoring of extreme events
  (it took 10 years to have current version of geoRglm!)

Bibliography

- Kazianka, H. & Pilz, J. (2010) Spatial interpolation using copula-based geostatistical models. In: GeoENV VII Geostatistics for Environmental Applications (P. Atkinson and C. Lloyd, eds), Springer, Berlin 2010, 307-319
- Kazianka, H. & Pilz, J. (2010) Copula-based geostatistical modeling of continuous and discrete data including covariates. Stochastic Environmental Research and Risk Assessment 24 (2010), 661–673
- http://www.intamap.org