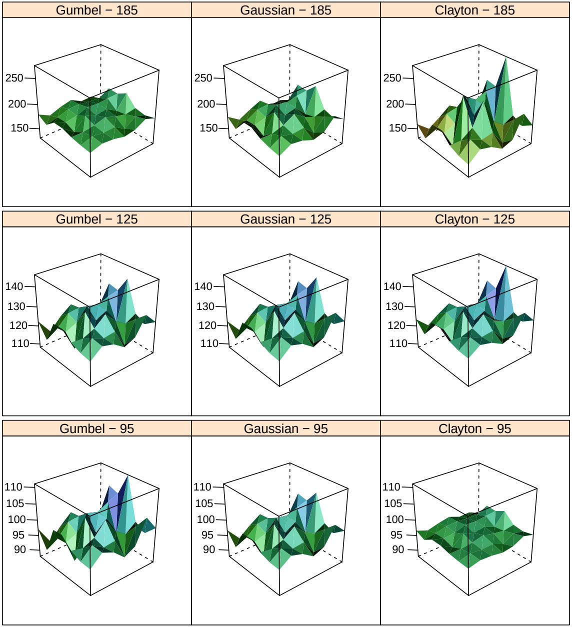
Chapter 1
Copulas in Spatial Statistics

Summer School *GEOSTAT 2014*,
Spatio-Temporal Geostatistics,
2014-06-19

Benedikt Gräler
http://ifgi.de/graeler
Institute for Geoinformatics
University of Muenster

ifgi
Institute for Geoinformatics
University of Münster

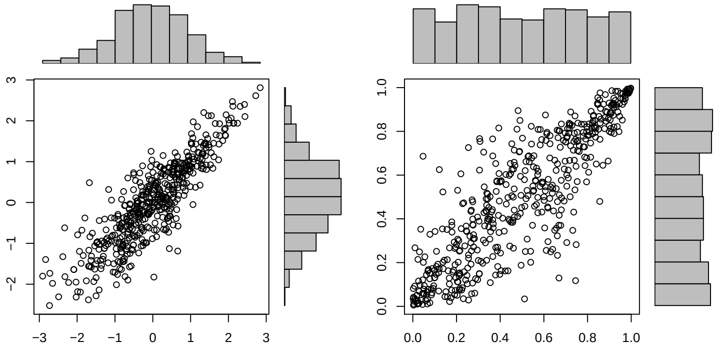
What if the world happens to be non-Gaussian?



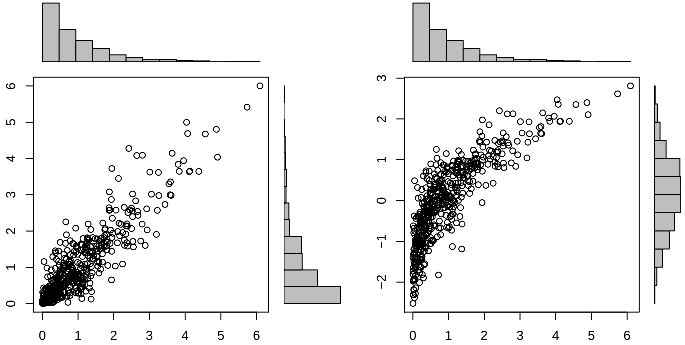
Bivariate Copulas I

Copulas allow to model dependencies much more detailed than a typical correlation value.

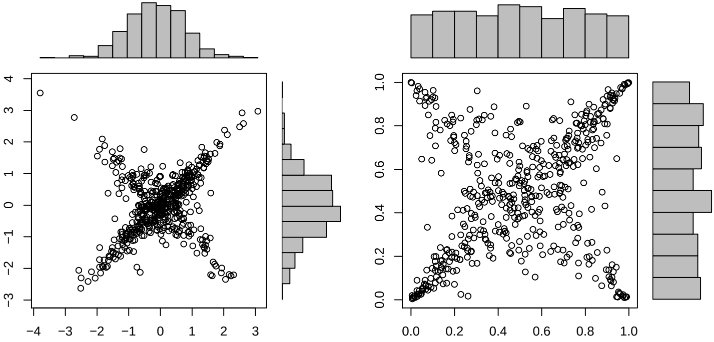
Instead of a single value, a full distribution is fitted describing dependence.



Bivariate Copulas II



Bivariate Copulas III



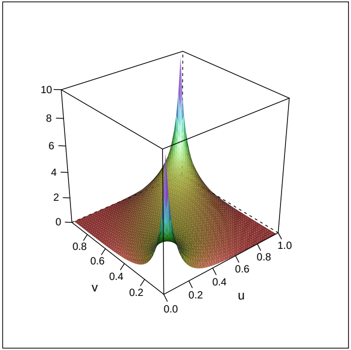
See the copulatheque for further interactive examples.

Behind the scenes - Sklar's Theorem

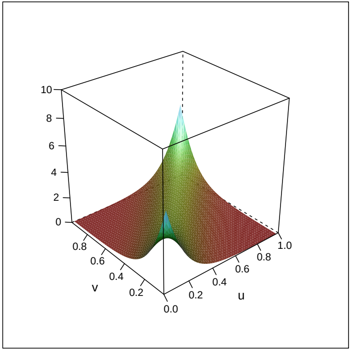
Every bivariate distribution H is composed out of some *copula* C and marginal distributions F_1 and F_2:

$$H(x, y) = C\big(F_1(x), F_2(y)\big)$$

Gaussian copula density



Frank copula density



Why is this useful?

Sklar's theorem allows us to model any multivariate distribution in two steps:

1. find marginal distribution functions using your favourite estimation technique that suite the data
2. find a copula that describes the dependence

This allows for a huge flexibility and a clear outline how to proceed.

See the interactive copulatheque on my website for more copula families.

Accounting for distance

Thinking of pairs of locations (s_i, s_j) we assume ...

- **distance** has a strong influence on the strength of dependence
- **dependence structure** is identical for all neighbours, but might change with distance
- **stationarity** and build k bins by spatial distance and estimate a bivariate copula $c_j(u, v)$ for each bin $\{[0, l_1),[l_1, l_2), \ldots, [l_{k-1}, l_k)\}$

Accounting for distance

Thinking of pairs of locations (s_i, s_j) we assume ...

distance has a strong influence on the strength of dependence

dependence structure is identical for all neighbours, but might change with distance

stationarity and build k bins by spatial distance and estimate a bivariate copula $c_j(u, v)$ for each bin $\{[0, l_1), [l_1, l_2), \ldots, [l_{k-1}, l_k)\}$

Accounting for distance

Thinking of pairs of locations (s_i, s_j) we assume ...

distance has a strong influence on the strength of dependence

dependence structure is identical for all neighbours, but might change with distance

stationarity and build k bins by spatial distance and estimate a bivariate copula $c_j(u, v)$ for each bin $\{[0, l_1), [l_1, l_2), \ldots, [l_{k-1}, l_k)\}$

Density of the bivariate spatial copula

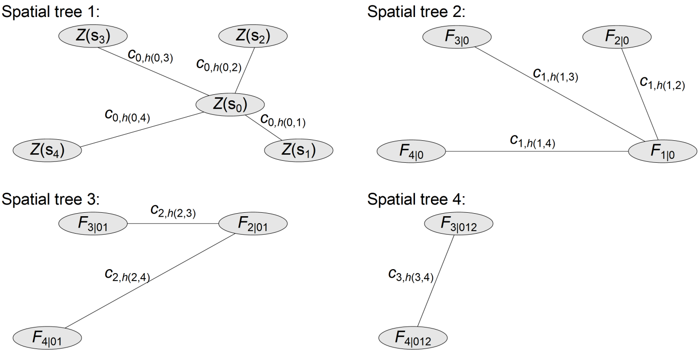
The density of the *bivariate spatial copula* is given by a convex combination of bivariate copula densities:

$$c_h(u,v) := \begin{cases} c_1(u,v) & , 0 \leq h < l_1 \\ (1-\lambda_2)c_1(u,v) + \lambda_2 c_2(u,v) & , l_1 \leq h < l_2 \\ \vdots & \vdots \\ (1-\lambda_k)c_{k-1}(u,v) + \lambda_k \cdot 1 & , l_{k-1} \leq h < l_k \\ 1 & , l_k \leq h \end{cases}$$

where $\lambda_j := \frac{h-l_{j-1}}{l_j - l_{j-1}}$. Each tree has its own bivariate spatial copula where distance h relates the involved pairs of locations.

ifgi
Institute for Geoinformatics
University of Münster

The spatial neighbourhood



The upper trees

Using the bivariate spatial copula on the first tree, the sample conditioned on s_0 is obtained.

The next bivariate spatial copula uses the distances between locations (s_1, s_2), (s_1, s_3), ..., (s_1, s_d).

A spatial binning allows to estimate the next bivariate spatial copula to generate the sample conditioned on s_0 and s_1.

$\vdots$

The full density

We get the full copula density as a product of all involved bivariate densities:

$$c_{\mathbf{h}}(u_0,\ldots,u_d)$$
$$= \prod_{i=1}^{d} c_{h_0(i)}(u_0,u_i) \cdot \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{h_j(j+i)}\left(u_{j|0,\ldots,j-1}, u_{j+i|0,\ldots,j-1}\right)$$

where $u_i = F_i\big(Z(s_i)\big)$ for $0 \leq i \leq d$ and

$$\begin{aligned} u_{j+i|0,\ldots,j-1} &= F_{h_{j-1}(j+i)}\left(u_{j+i}|u_0,\ldots,u_{j-1}\right) \\ &= \frac{\partial C_{h_{j-1}(j+i)}\left(u_{j-1|0,\ldots j-2}, u_{j+i|0,\ldots j-2}\right)}{\partial u_{j-1|0,\ldots j-2}} \end{aligned}$$

Spatial vine copula interpolation

The estimate can be obtained as the expected value

$$\widehat{Z}_m(s_0) = \int\limits_{[0,1]} F^{-1}(u)\ c_{\mathbf{h}}\big(u|u_1,\dots,u_d\big)\ \mathrm{d}u$$

or by calculating any percentile p (i.e. the median)

$$\widehat{Z}_p(s_0) = F^{-1}\big(C_{\mathbf{h}}^{-1}(p|u_1,\dots,u_d)\big)$$

with the conditional density

$$c_{\mathbf{h}}(u|u_1,\dots,u_d) := \frac{c_{\mathbf{h}}(u,u_1,\dots,u_d)}{\int_0^1 c_{\mathbf{h}}(v,u_1,\dots,u_d)\mathrm{d}v}$$

and $u_i = F\big(Z(s_i)\big)$ as before.

R-package spcopula

The developed methods are implemented as R-scripts and are bundled in the package spcopula available at R-Forge.

The package spcopula extends and combines the R-packages VineCopula, spacetime and copula.

Simulated nuclear radiation

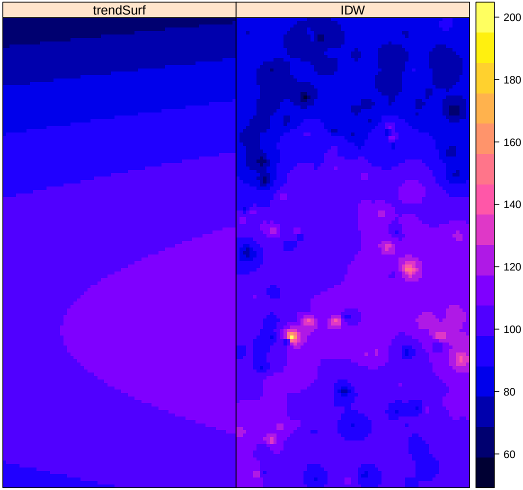
The new method is applied to simulated nuclear radiation data mimicking an emergency scenario.

The data has been generated for the spatial interpolation comparison 2004 (SIC2004, [?]) with 200 data and 808 validation locations.

To better approximate stationarity, a quadratic trend surface has been fitted excluding the (two) extremes.

This has been published in Journal of Spatial Statistics [1].

The trend surface



The marginal distribution

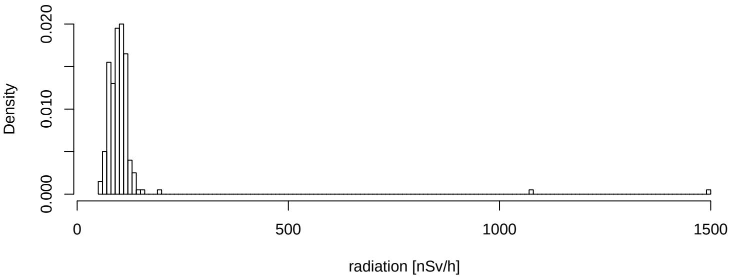
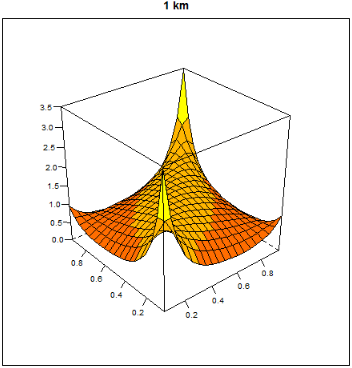
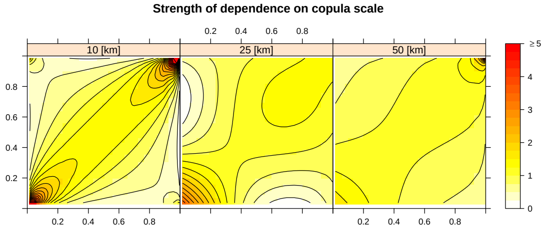


Figure: Histogram of the "observed" radiation values.

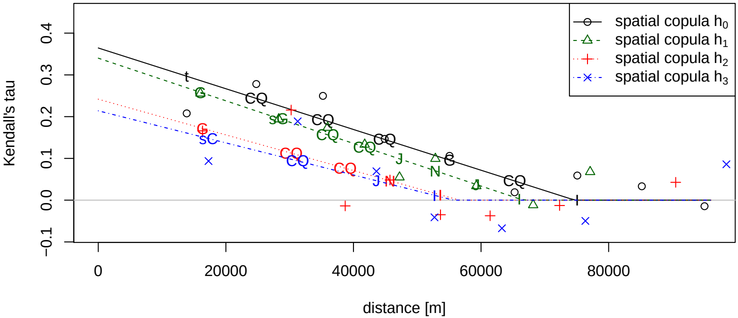
The empirical marginal distribution function has been used.

The bivariate spatial copulas I

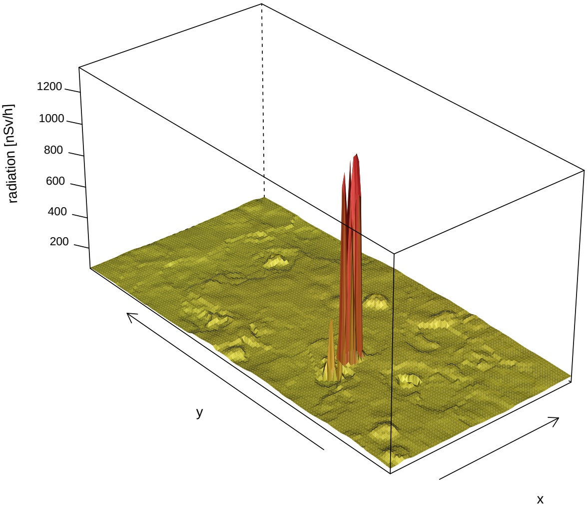




The bivariate spatial copulas II



Interpolated grid

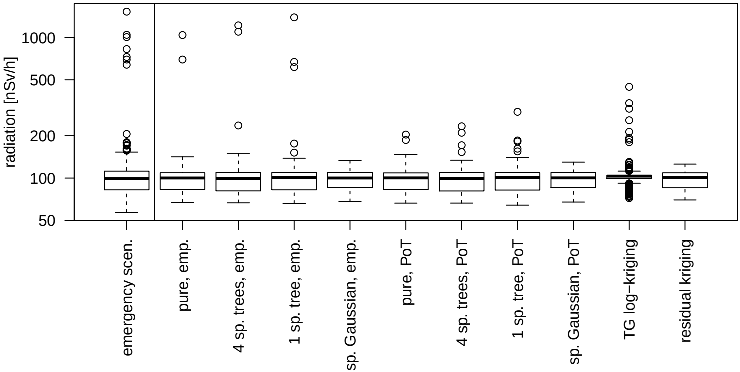


Validation data set

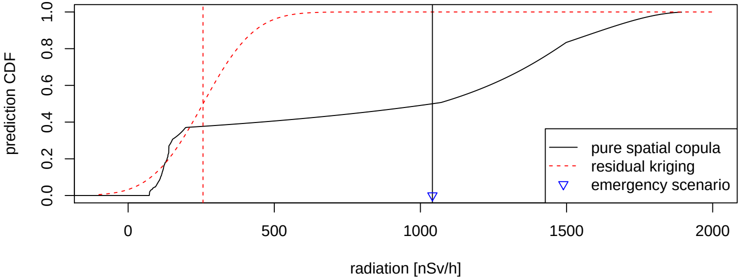
808 locations have been hold back to assess the prediction quality.

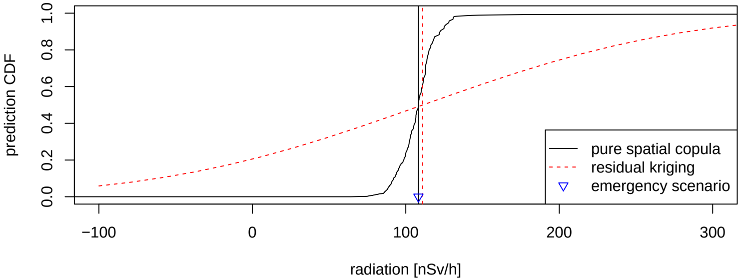
approach	MAE	RMSE	ME	COR
spatial vine copula	*14.5*	67.6	-6.1	0.60
TG log-kriging	20.8	78.2	-2.1	0.39
residual kriging	21.1	75.6	5.2	0.43
best one in SIC2004	14.9	*45.5*	*-0.5*	*0.84*

Reproduction of margins



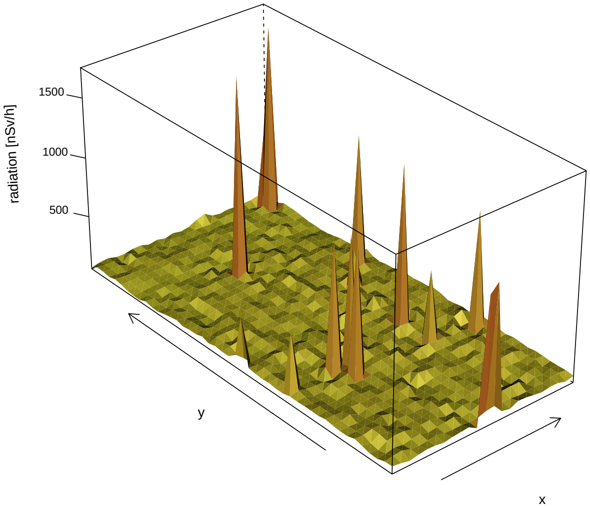
Uncertainty assessment





Simulation

Predicting random quantiles from the spatial vine copula.



Benefits

richer flexibility due to the various dependence structures

asymmetric dependence structures become possible (temporal direction)

probabilistic advantage sophisticated uncertainty analysis, drawing random samples, ...

Benefits

richer flexibility due to the various dependence structures

asymmetric dependence structures become possible (temporal direction)

probabilistic advantage sophisticated uncertainty analysis, drawing random samples, ...

Benefits

richer flexibility due to the various dependence structures

asymmetric dependence structures become possible (temporal direction)

probabilistic advantage sophisticated uncertainty analysis, drawing random samples, ...

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, . . .)
- flexible/complex neighbourhoods
  (e.g. by spatial direction, . . .)
- larger neighbourhoods possibly using vine truncation techniques
- include further copula families
- improve performance

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, ...)
- flexible/complex neighbourhoods
  (e.g. by spatial direction, ...)
- larger neighbourhoods possibly using vine truncation techniques
- include further copula families
- improve performance

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, . . .)
- flexible/complex neighbourhoods
  (e.g. by spatial direction, . . .)
- larger neighbourhoods possibly using vine truncation techniques
- include further copula families
- improve performance

Further extensions

- including covariates (e.g. altitude, population, EMEP, ...)
- flexible/complex neighbourhoods (e.g. by spatial direction, ...)
- larger neighbourhoods possibly using vine truncation techniques
- include further copula families
- improve performance

ifgi
Institute for Geoinformatics
University of Münster

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, ...)
- flexible/complex neighbourhoods
  (e.g. by spatial direction, ...)
- larger neighbourhoods possibly using vine truncation techniques
- include further copula families
- improve performance

References I

[1] Benedikt Gräler. Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, 2014. available online, in pess.