Modelling Dependence in Space and Time with Vine Copulas

Ninth International Geostatistics Congress
15.06.2012

Benedikt Gräler
Institute for Geoinformatics
University of Münster
http://ifgi.uni-muenster.de/graeler

spatio-temporal data

Typically, spatio-temporal data is given at a set of discrete locations $s_i \in S$ and time steps $t_j \in T$.

We desire a full spatio-temporal random field Z modelling the process at any location (s, t) in space and time.

Here, we look at daily fine dust concentrations across Europe (PM_{10}).

basic set-up & assumptions

We assume a stationary and isotropic process across the region S and the time frame T.

This allows us to consider lag classes by distances in space and time instead of single locations in space.

Our approach

- using the full distribution of the observed phenomenon of a local neighbourhood

Our approach

- using the full distribution of the observed phenomenon of a local neighbourhood
- each observed location $(s_0, t_0) \in (S, T)$, is grouped with its three nearest spatial neighbours $\{s_1, s_2, s_3\} \subset S$, the current (t_0) and two preceding time steps (t_{-1}, t_{-2}) of this neighbourhood generating a 10 dimensional sample: $(\mathbf{X}_{s_0,t_0}, \mathbf{X}_{s_1,t_0}, \ldots, \mathbf{X}_{s_3,t_{-2}})$

Our approach

- using the full distribution of the observed phenomenon of a local neighbourhood
- each observed location $(s_0, t_0) \in (S, T)$, is grouped with its three nearest spatial neighbours $\{s_1, s_2, s_3\} \subset S$, the current (t_0) and two preceding time steps (t_{-1}, t_{-2}) of this neighbourhood generating a 10 dimensional sample: $(\mathbf{X}_{s_0,t_0}, \mathbf{X}_{s_1,t_0}, \dots, \mathbf{X}_{s_3,t_{-2}})$
- an estimate is calculated from the conditional distribution at an unobserved location conditioned under the values of its spatio-temporal neighbourhood

The spatio-temporal neighbourhood

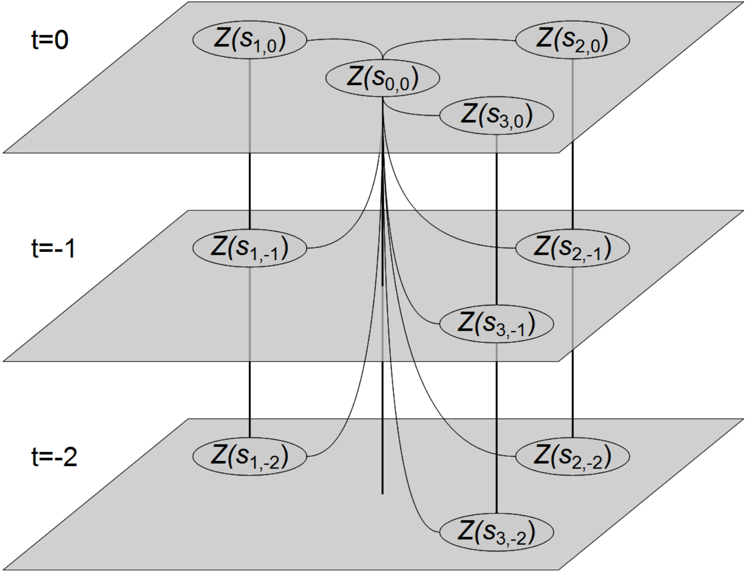


Figure: 10-dim spatio-temporal neighbourhood

Copulas

Copulas describe the dependence structure between the margins of a multivariate distribution. Sklar's Theorem states:

$$H(x_1, \dots, x_d) = C\big(F_1(x_1), \dots, F_d(x_d)\big)$$

where H is any multivariate CDF, F_1, ..., F_d are the corresponding marginal univariate CDFs and C is a suitable *copula* (uniquely determined in a continuous setting).

Since $F_i(X_i) \sim U(0,1)$, copulas can be thought of as CDFs on the unit (hyper-)cube.

Beyond Gaussian dependence structure

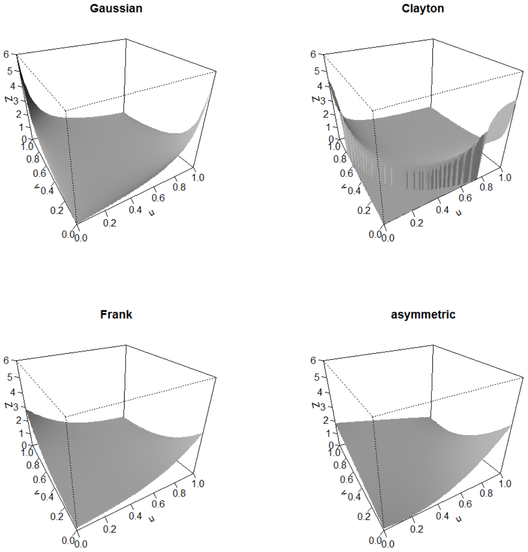


Figure: Copula density plots with rank correlation $\tau = -0.3$; density on Z-axis $\sim$ strength of dependence.

Vine-Copula

Bivariate copulas are pretty well understood and rather easy to estimate.

Unfortunately, most bivariate families do not nicely extend to a multivariate setting or lack the necessary flexibility.

Vine-Copulas allow to approximate multivariate copulas by mixing (conditional) bivariate copulas following a vine decomposition.

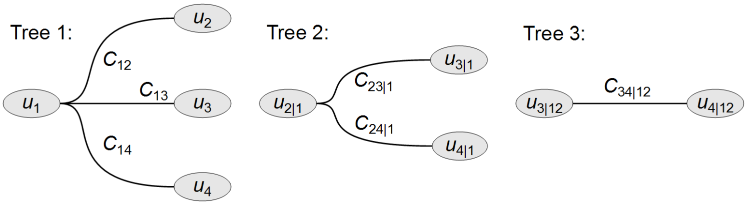


Figure: A 4-dim canonical vine decomposition.

Our spatio-temporal vine-copula

We decompose the ten dimensional distribution into its marginal distribution F (identical for all 10 margins) and a vine copula:

On the first tree, we use a spatio-temporal bivariate copula accounting for spatial and temporal distance.

The following trees are modelled from a wide set of "classical" bivariate copulas.

Accounting for distance

Thinking of pairs of locations (s_1, t_1), (s_2, t_2) we assume ...

distance has a strong influence on the strength of dependence

Accounting for distance

Thinking of pairs of locations (s_1, t_1), (s_2, t_2) we assume ...

distance
: has a strong influence on the strength of dependence

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

Accounting for distance

Thinking of pairs of locations (s_1, t_1), (s_2, t_2) 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 lag classes by spatial distance for each temporal distance Δ and estimate a bivariate copula $c_j^{\Delta}(u, v)$ for all lag classes $\{[0, l_1),[l_1, l_2), \ldots, [l_{k-1}, l_k)\} \times \{0, 1, 2\}$

Density of the spatial and spatio-temporal copula

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

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

where $\lambda_j := \frac{h-l_{j-1}}{l_j - l_{j-1}}$.

The density of the *bivariate spatio-temporal copula* $c_{h,\Delta}(u,v)$ is then given by a convex combination of bivariate spatial copula densities in a analogous manner.

The full density

The remaining copulas $c_{j,j+i|0,\dots j-1}$ with integers $1 \leq j \leq 9, 1 \leq i \leq 9-j$ are estimated over the nine dimensional conditional sample $\left(F_{\mathbf{h},\mathbf{\Delta}}(\mathbf{X}_{s_1,t_0}|\mathbf{X}_{s_0,t_0}), \dots, F_{\mathbf{h},\mathbf{\Delta}}(\mathbf{X}_{s_3,t_{-2}}|\mathbf{X}_{s_0,t_0})\right)$.

We get the full 10-dim copula density as a product of all involved bivariate densities:

$$\begin{aligned} & c_{\mathbf{h},\mathbf{\Delta}}(u_0, \dots, u_9) \\ = & \prod_{i=1}^{9} c_{h,\Delta}(u_0, u_i) \cdot \prod_{j=1}^{9-1} \prod_{i=1}^{9-j} c_{j,j+i|0,\dots,j-1}(u_{j|0,\dots,j-1}, u_{j+i|0,\dots,j-1}) \end{aligned}$$

Where $u_0 = F\big(Z(s_0,t_0)\big), \dots, u_9 = F\big(Z(s_3,t_{-2})\big)$,

$$u_{j|0} = F_{h,\Delta}(u_j|u_0) = \frac{\partial C_{h,\Delta}(u_0,u_j)}{\partial u_0} \text{ and}$$

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

Spatio-Temporal Vine-Copula interpolation

The estimate can be obtained as the expected value

$$\widehat{Z}_m(s_0, t_0) = \int\limits_{[0,1]} F^{-1}(u)\ c_{h,\Delta}\big(u|u_1, \dots, u_9\big)\ \mathrm{d}u$$

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

$$\widehat{Z}_p(s_0) = F^{-1}\big(C^{-1}_{h,\Delta}(p|u_1, \dots, u_9)\big)$$

with the conditional density

$$c_{h,\Delta}(u|u_1, \dots, u_9) := \frac{c_{h,\Delta}(u, u_1, \dots, u_9)}{\int_0^1 c_{h,\Delta}(v, u_1, \dots, u_9)\mathrm{d}v}$$

and $u_1 = F\big(Z(s_1, t_0)\big), \dots, u_9 = F\big(Z(s_3, t_{-2})\big)$ as before.

ifgi
Institute for Geoinformatics
University of Münster

R-package spcopula

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

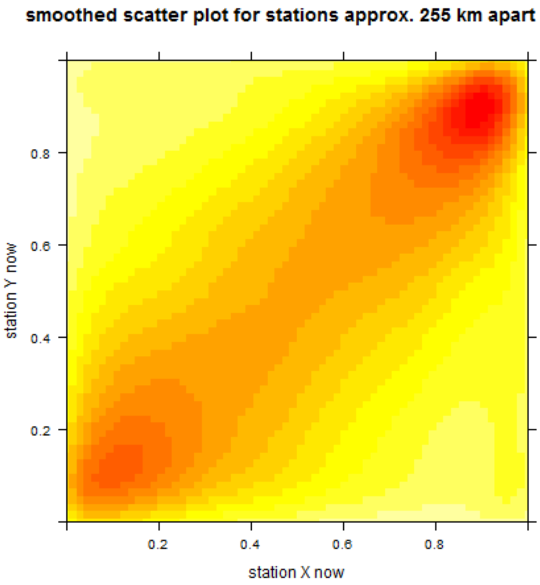
The package spcopula extends and combines the R-packages CDVine, spacetime and copula.

PM_{10} across Europe

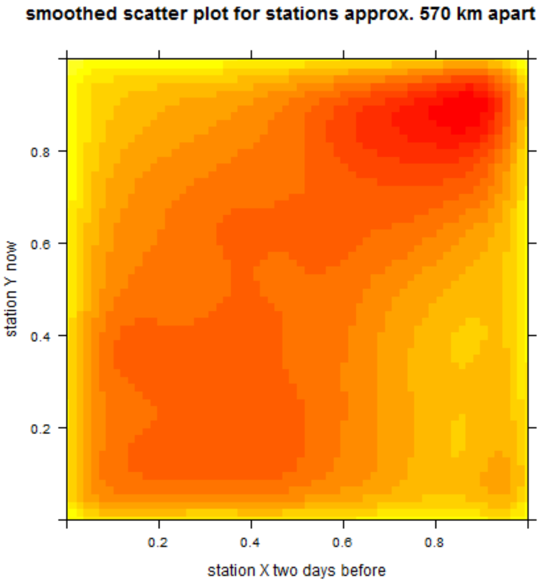
We applied our method to daily mean PM_{10} concentrations observed at 194 rural background stations for the year 2005 (70810 obs.).

The data is hosted by the European Environmental Agency (EEA) originally provided by the member states and freely available at http://www.eea.europa.eu/themes/air/airbase.

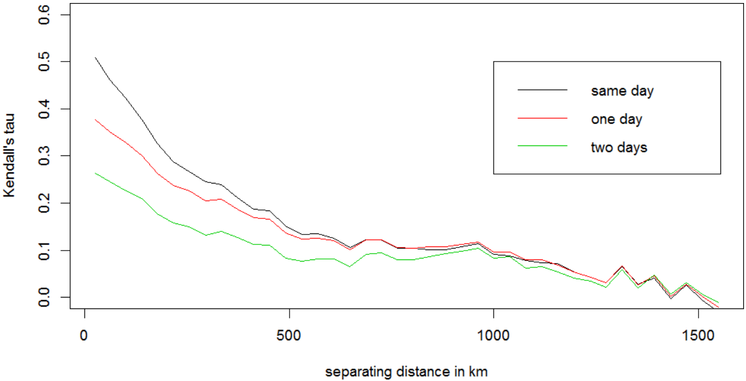
A look inside the spatio-temporal copula I



A look inside the spatio-temporal copula II



A look inside the spatio-temporal copula III



A look inside the spatio-temporal copula IV

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\Delta = 0$	t	F	t	F	F	F	*F*	F	F	F	F	F	F	F	F
$\Delta = -1$	F	F	F	F	F	F	F	F	F	F	F	F	F	A	A
$\Delta = -2$	F	F	F	F	A	A	A	A	A	A	A	A	A	A	*A*

Table: The selected copula families for the first 15 spatial lag classes (up to ≈ 600 km) and three time instances. (t =Student, F = Frank, A = asymmetric, the vertical lines indicate 100 km, 250 km, 500 km breaks)

comparing log-likelihoods

10-dim copula	log-likelihood
spatio-temporal vine-copula	72709
... backwards in time	63089
Gaussian	53305
Gumbel	35292
Clayton	30593

Table: Log-likelihoods for different copula families/parametrization.

The margins

The best fit to the marginal distribution could be achieved for generalized extreme value distribution (GEV):

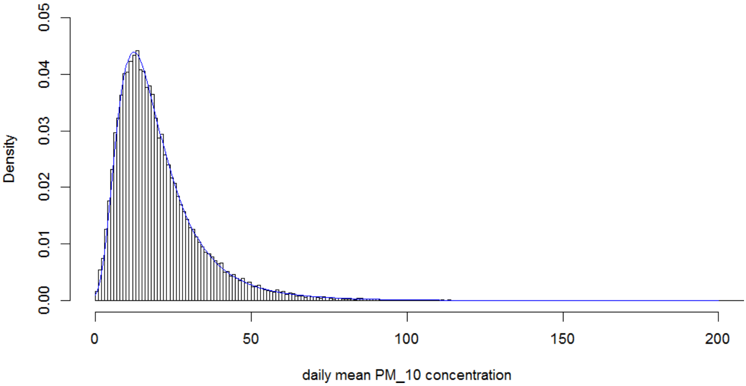


Figure: Histogram of all PM_{10} measurements and fitted density.

Cross validation

Leaving out complete time series of single stations one after another.

	RMSE	bias	MAE
expected value $\widehat{Z}_m$	11.20	-0.73	6.95
median $\widehat{Z}_{0.5}$	12.08	1.94	6.87
metric cov. kriging	10.69	-0.29	6.28
metric cov. res. kriging	9.84	-0.24	5.66

Table: Cross validation results for the expected value and median estimates following the vine copula approach and two methods from a recent comparison study [2] on spatio-temporal kriging approaches in PM_{10} mapping.

A station in Finland

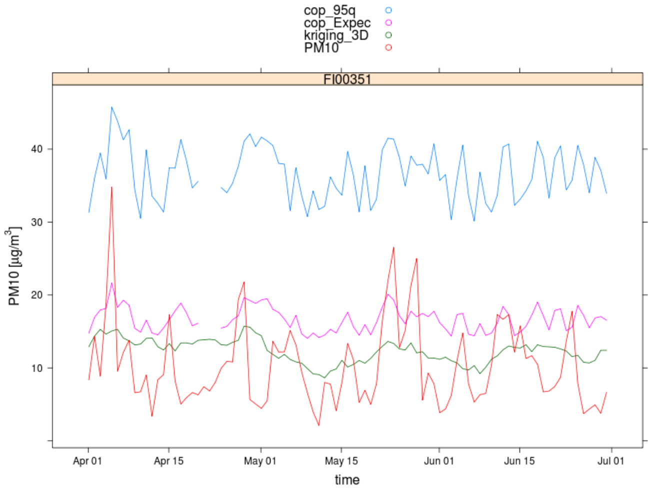


Figure: A Finish station roughly 600 km apart from any other station. The copula approach (magenta) is controlled by the global mean, kriging (green) reduces almost to a mowing window mean.

A station in Germany

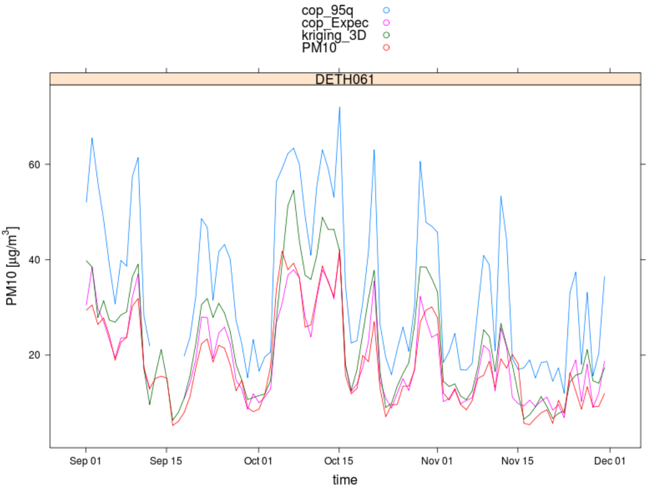


Figure: A German station within a rather dense network. The copula approach (magenta) reproduces the observed values (red) while kriging (green) overestimates several daily means.

Benefits

richer flexibility due to the various dependence structures

Benefits

richer flexibility due to the various dependence structures

asymmetric dependence structures become possible (temporal direction)

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, . . .)

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, . . .)
- complex neighbourhoods
  (e.g. by spatial direction, . . .)

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, . . .)
- complex neighbourhoods
  (e.g. by spatial direction, . . .)
- include further copula families

Further extensions

- including covariates
  (e.g. altitude, population, EMEP, ...)
- complex neighbourhoods
  (e.g. by spatial direction, ...)
- include further copula families
- improve performance

- Aas, Kjersti, Claudia Czado, Arnoldo Frigessi & Henrik Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics, 44*, 182 - 198.
- Gräler B, Gerharz LE, Pebesma E (2012) Spatio-temporal analysis and interpolation of PM10 measurements in Europe. ETC/ACM Technical Paper 2011/10, January 2012
- Nelsen, R. (2006). *An Introduction to Copulas*. 2nd ed. New York: Springer.