Covariates in single tree Spatio-Temporal Vine Copulas

Spatial Copula Workshop 2014
September 22, 2014

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 bivariate spatio-temporal random field $(Z,Y) : \Omega \times S \times T \to \mathbb{R}^2$ modelling the process at any location (s,t) in space and time.

Here, we look at daily fine dust concentrations across Europe (PM_{10}) measured at stations (Z) and modelled (Y) through the European Monitoring and Evaluation Programme (EMEP).

basic set-up & assumptions

We assume

1. that the marginal distribution can be parametrized by location $s \in S$: F_s and G_s
2. stationarity on the copula scale
3. that the dependence does not change within the study area.

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

The classical geostatistical approach

A multivariate Gaussian distribution is assumed where

- a variogram function can be used to parametrize the (large) covariance matrix by distance

ifgi
Institute for Geoinformatics
University of Münster

The classical geostatistical approach

A multivariate Gaussian distribution is assumed where

- a variogram function can be used to parametrize the (large) covariance matrix by distance
- the mean vector is set to the observed values

ifgi
Institute for Geoinformatics
University of Münster

Our approach

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

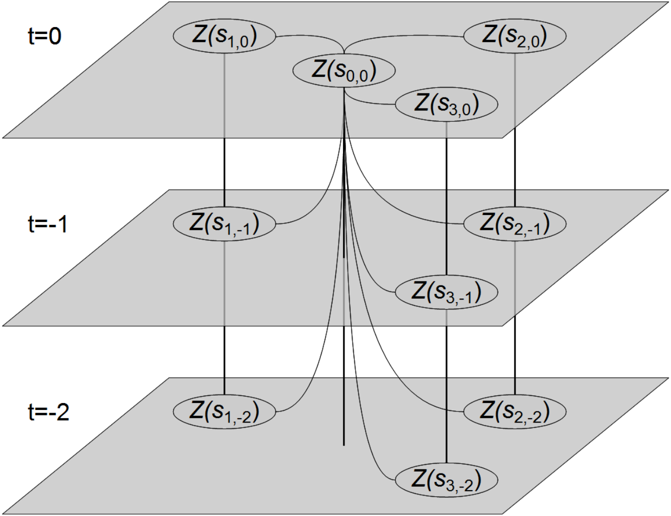
Our approach

- 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 nine strongest correlated neighbours.

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 nine strongest correlated neighbours.
- an estimate is calculated from the conditional distribution at an unobserved location conditioned under the values of its spatio-temporal neighbourhood and covariate yielding an elven dimensional distribution.

A metric spatio-temporal neighbourhood



The spatio-temporal covariate vine-copula

We decompose the eleven dimensional distribution into it's marginal distribution F (identical for all 10 margins), the marginal distribution of the covariate G and a vine copula:

On the first tree, we use a spatio-temporal bivariate copula accounting for spatial and temporal distance and the copula relating the variable of interest and its covariate.

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

Accounting for spatial 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 spatial 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 spatial 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
: on the copula scale 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:

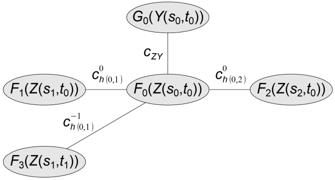
$$c_h^{\Delta}(u,v) := \begin{cases} c_{1,h}(u,v) & ,0 \leq h < l_1 \\ (1-\lambda_2)c_{1,h}(u,v) + \lambda_2 c_{2,h}(u,v) & ,l_1 \leq h < l_2 \\ \vdots & \vdots \\ (1-\lambda_k)c_{k-1,h}(u,v) + \lambda_k \cdot 1 & ,l_{k-1} \leq h < l_k \\ 1 & ,l_k \leq h \end{cases}$$

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

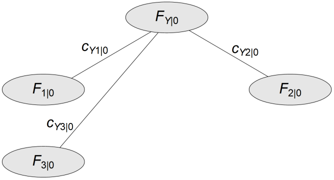
The density of the *bivariate spatio-temporal copula* $c_{h,\Delta}(u,v)$ is then given by a convex combination of bivariate spatial copula densities in an analogous manner.

Adding the Covariate

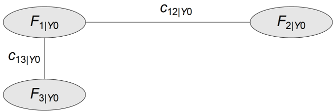
Tree 1 (Spatio-Temporal Tree):



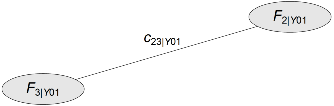
Tree 2:



Tree 3:



Tree 4:



The full density I

We get the full 11-dim copula density as a product of all involved bivariate densities:

$$
\begin{aligned}
&c^{\boldsymbol{\Delta}}_{\mathbf{h}}(u_0, v_0, u_1, \dots, u_d) \\
=&c_{ZY}(u_0, v_0) \cdot \prod_{i=1}^{d} c^{\Delta}_{h(0,i)}(u_0, u_i) \cdot \prod_{i=1}^{d} c_{Y,i|0}(u_{Y|0}, u_{i|0}) \\
&\cdot \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|Y,0,\dots,j-1}(u_{j|Y,0,\dots,j-1}, u_{j+i|Y,0,\dots,j-1})
\end{aligned}
$$

where $v_0 = G_0\big(Y(s_0, t_0)\big)$ with G_0, $u_i = F_i\big(Z(s_q, t_r)\big)$ for $0 \leq i \leq d$ with (s_q, t_r) denoting the i-th strongest correlated neighbour of (s_0, t_0) with $F_i = F_{q,r}$ and ...

The full density II

$$u_{Y|0} = F_{Y|0}(v_0|u_0) = \frac{\partial C_{Z,Y}(u_0, v_0)}{\partial u_0}$$

$$u_{i|0} = F_{i|0}(u_i|u_0) = \frac{\partial C^{\Delta}_{h(0,i)}(u_0, u_i)}{\partial u_0}$$

and

$$\begin{aligned} u_{j+i|Y,0,\ldots,j-1} &= F_{j+i|Y,0,\ldots j-1}(u_{j+i}|v_0, u_0, \ldots, u_{j-1}) \\ &= \frac{\partial C_{j-1,j+i|Y,0,\ldots j-2}(u_{j-1|Y,0,\ldots j-2}, u_{j+i|Y,0,\ldots j-2})}{\partial u_{j-1|Y,0,\ldots j-2}} \end{aligned}$$

for $1 \leq j < d$ and $0 \leq i \leq d - j$.

Spatio-Temporal Covariate Vine-Copula interpolation

The estimate can be obtained as the expected value

$$\widehat{Z}_m(s_0) = \int_{\mathbb{R}} z \cdot f_{\mathbf{h}}^{\mathbf{\Delta}}(z|y_0, z_1, \ldots, z_d) \, \mathrm{d}z$$
$$= \int_{[0,1]} F_0^{-1}(u) \; c_{\mathbf{h}}^{\mathbf{\Delta}}\big(u|v_0, u_1, \ldots, u_d\big) \; \mathrm{d}u$$

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

$$\widehat{Z}_p(s_0) = F_0^{-1}\big({C_{\mathbf{h}}^{\mathbf{\Delta}}}^{-1}(p|v_0, u_1, \ldots, u_d)\big)$$

R-package spcopula

The developed methods are implemented as R-scripts and are bundled in the package spcopula available at R-Forge (briefly presented later today).

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

ifgi
Institute for Geoinformatics
University of Münster

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.

As covariate, daily mean PM_{10} concentrations derived from the EMEP model are included.

ifgi
Institute for Geoinformatics
University of Münster

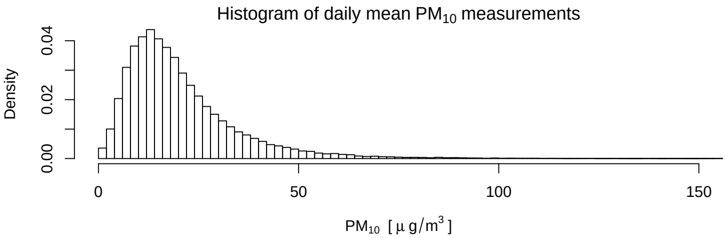
The marginal distributions

We fit extreme value distributions for each location $s \in \mathcal{S}$ based on the time series leading to margins F_s and G_s.

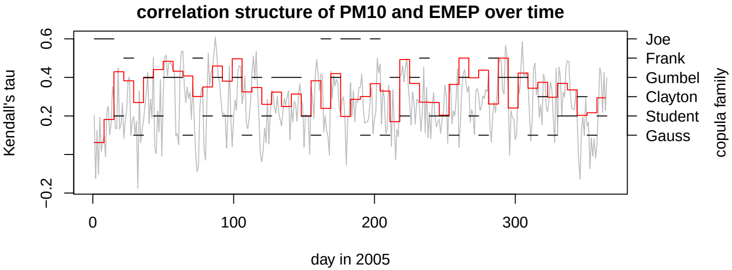
For the interpolation scenario we use

1. a linear model incorporating the locations' coordinates and altitude followed by an inverse distance weighted interpolation of the residuals ...
2. inverse distance weighted interpolation ...

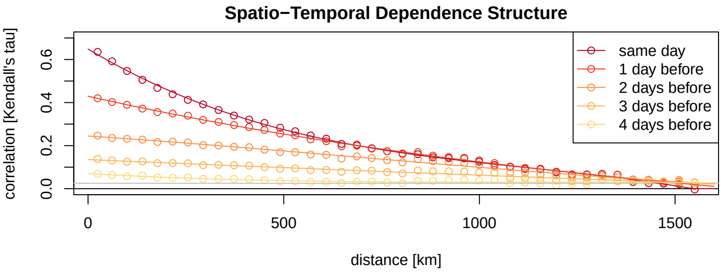
of the local neighbourhood's marginal parameters.



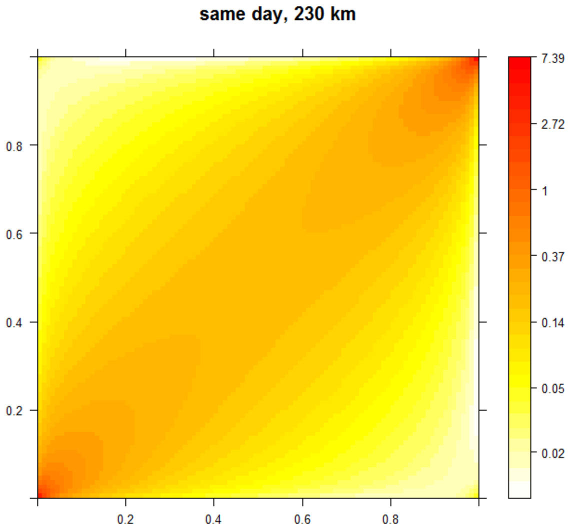
The covariate copula



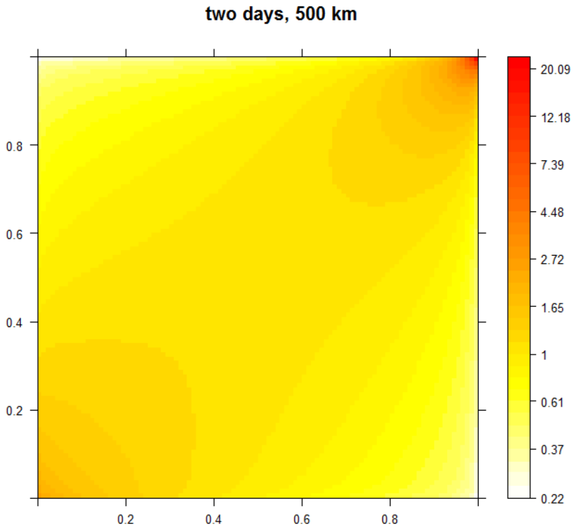
A look inside the spatio-temporal copula I



A look inside the spatio-temporal copula II



A look inside the spatio-temporal copula III



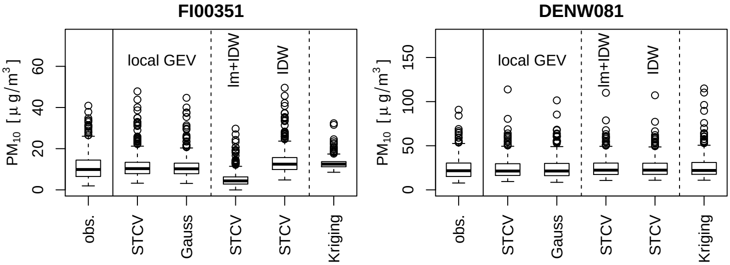
A look inside the spatio-temporal copula IV

	Spatial lag											
ID	1	2	3	5	6	7	22	23	25	26	27	28
mean dist. [km]	25	61	99	177	216	255	843	881	961	999	1038	1079
$\Delta = 0$	t	G	t	...	t	G		...		G	F	N
$\Delta = -1$	G			...			G	F	...		F	N
$\Delta = -2$	G				...				G	N	...	
$\Delta = -3$	G						...					G
$\Delta = -4$	G	...		G	J		...			J	G	G

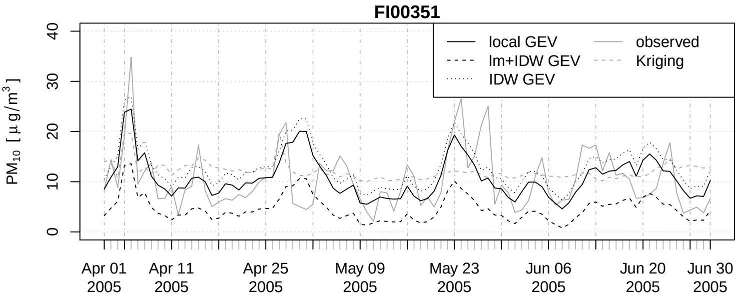
Cross validation

Dependence model	Margin	RMSE	MAE	ME	COR
STCV $\widehat{Z}_m$	local GEV	8.53	4.61	-0.05	0.84
Gaussian STCV $\widehat{Z}_m$	local GEV	8.65	4.59	0.08	0.83
STCV $\widehat{Z}_m$	lm+IDW GEV	10.12	5.79	0.17	0.76
STCV $\widehat{Z}_m$	IDW GEV	10.82	6.26	0.14	0.72
metric res. kriging	log linear reg.	10.67	6.16	0.47	0.74
sp.-temp. vine $\widehat{Z}_m$	global GEV	11.20	6.95	-0.73	NA

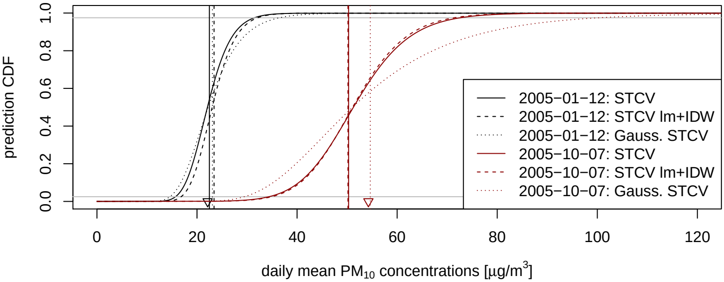
Box Plots of marginal reproduction



A station in Finland



Conditional distribution functions



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

- larger neighbourhoods possibly using vine truncation techniques

Further extensions

- larger neighbourhoods possibly using vine truncation techniques
- include further copula families

Further extensions

- larger neighbourhoods possibly using vine truncation techniques
- include further copula families
- improve performance