spcopula: Modelling Spatial and Spatio-Temporal Dependence with Copulas in **R**

Benedikt Gräler

Abstract

The **spcopula** R package provides tools to model spatial and spatio-temporal phenomena with *spatial* and *spatio-temporal vine copulas*. Copulas allow us to flexibly build multivariate distributions with mixed margins where the copula describes the multivariate dependence structure coupling the margins. In classical geostatistics, a multivariate Gaussian distribution is typically assumed and dependence is summarized in a covariance matrix implying limitations like elliptical symmetry in the strength of dependence. Copulas allow for dependence structures beyond the Gaussian one, being for instance asymmetric. We developed the spatio-temporal vine copulas such that the bivariate copula families in the lower trees may change with distance across space and time allowing not only for a varying strength of dependence but also for a changing dependence structure. These spatio-temporal distributions are used to predict values at unobserved locations, assess risk, or run simulations. Based on the concept of vine copulas, the **spcopula** package provides a large set of multivariate distributions. As bivariate spatial copulas do not have any probabilistic restrictions, the spatial vine copula is a powerful approach for modelling skewed or heavy tailed data with complex and potentially asymmetric dependence structures in the spatial and spatio-temporal domain.

1. Introduction

Interpolation of spatial random fields is a common task in geostatistics. Simple approaches like inverse distance weighted predictions or the well known kriging procedures have routinely been applied for many years. However, when the underlying assumptions (i.e., a multivariate Gaussian distribution, possibly after transformation) of these approaches are hard to be fulfilled, alternatives are needed. Copulas have been used in some applications in the domain of spatial statistics. Bárdossy (2006) was one of the first to apply copulas in a geostatistical context. Some recent advances incorporating copulas in this field have for instance been published by Kazianka and Pilz (2011, 2010), Bárdossy (2011), Bárdossy and Pegram (2009) or Bárdossy and Li (2008). They use a comparatively small set of copula families to model spatial processes.

The spatio-temporal domain rises in interest since several years, and several extensions of the spatial approaches to spatio-temporal ones have been developed (see e.g., Cressie and Wikle 2011). A major challenge with extending spatial kriging to spatio-temporal kriging is to build and fit well defined spatio-temporal covariance functions. The approach presented here differs from the classical geostatistical ones by using *spatio-temporal vine copulas* to build

a spatio-temporal distribution that does not rely on the Gaussian assumption, nor involves a covariance matrix. It extends the spatial vine copula (Gräler 2014) to the spatio-temporal context. Similar to co-kriging, we will introduce a spatio-temporal vine copula approach incorporating covariates.

The advantage of the *spatio-temporal vine copula* is its flexibility in the selection of copula families through *bivariate spatio-temporal copulas*. Bivariate spatio-temporal copulas are a convex combination of different copula families that are parameterised by spatial and temporal distance (Equation 1 in Section 2). This changing dependence structure allows for instance to preserve spatial symmetry within each time step while allowing for a directional effect across time. The introduction of a *bivariate spatial copula* into a vine copula for interpolation has been described by Gräler (2014). The bivariate spatial or spatio-temporal copulas are combined into a *vine copula* (initially called *pair-copula construction* by Aas, Czado, Frigessi, and Bakken 2009; Bedford and Cooke 2002) for a local spatial or spatio-temporal neighbourhood. A first approach to extend the spatial to the spatio-temporal approach has been presented in Gräler and Pebesma (2012). This paper describes a more flexible spatio-temporal neighbourhood structure and the introduction of a covariate to improve the prediction. Adding marginal distributions to the spatial or spatio-temporal vine copula yields a full multivariate distribution describing a local distribution of the observed phenomenon.

The **spcopula** R package provides functions and classes to model spatial and spatio-temporal phenomena by vine copulas. Different tools have been implemented to fit spatial and spatio-temporal vine copulas to a data set, to interpolate the random field, and to predict quantiles from it. The package extends the **copula** R package (Kojadinovic and Yan 2010; Yan 2007) and provides additional copula families. Wrapper classes following the **copula** design to the copula families available in **VineCopula** (Schepsmeier, Stoeber, and Brechmann 2013) that have been implemented in **spcopula** are now directly available in **VineCopula**. The functionality for non-spatial vine copulas relies on **VineCopula**. For handling spatial and spatio-temporal data, **spcopula** builds on the R packages **sp** (Pebesma and Bivand 2005) and **spacetime** (Pebesma 2012). A more detailed overview of core dependencies and contributions of **spcopula** is provided in Table 1.

The remainder of this paper is organized as follows. The theoretical background of copulas, bivariate spatial copulas, bivariate spatio-temporal copulas and vine copulas are addressed in the next section. A strategy to estimate spatio-temporal vine copulas is given in Section 3. Section 4 discusses different applications of the multivariate distribution such as the possibility to predict values at unobserved locations or simulate from the spatial or spatio-temporal random field. One application is illustrated in Section 5, where daily mean PM_{10} concentrations (particulate matter less than 10 μm in diameter) across Europe observed throughout the year 2005 are interpolated including an additional covariate. Results are discussed in Section 6. Conclusions are drawn in Section 7.

Package	**spcopula** reuses and extends	**spcopula** adds to the functionality
copula	• S4-class definition `copula` • methods `fitCopula`, `dCopula`, `pCopula`, `rCopula`	• bivariate copula families `asCopula`, `cqsCopula` and `tawn3pCopula` • empirical and analytical tail dependence functions `empTailDepFun` and `tailDepFun` • partial derivatives via methods `dduCopula` and `ddvCopula` • inverse of partial derivatives via methods `invdduCopula` and `invddvCopula` • inverse of bivariate copulas for a given u or v via methods `qCopula_u` and `qCopula_v`
VineCopula	• function `BiCopSelect` • S4-class wrapper `vineCopula`	-/-
sp	• abstract S4-class definition `Spatial` • function `spDists`	• nearest spatial neighbour calculation via function `getNeighbours`
spacetime	• abstract S4-class definition `ST`	• nearest spatio-temporal neighbour calculation via function `getStNeighbours`

Table 1: Overview of core dependencies and contributions of **spcopula**.

2. Spatio-temporal vine copulas

In the following, we assume a spatio-temporal random field $Z : \Omega \times \mathcal{S} \times \mathcal{T} \to \mathbb{R}$ defined over some spatial domain $\mathcal{S}$ and temporal domain $\mathcal{T}$ of interest and an underlying probability space Ω. Typically, a sample $\mathbf{Z} = \big(z(s_1, t_1), \ldots, z(s_n, t_n)\big)$ has been observed at a set of distinct spatio-temporal locations $(s_1, t_1), \ldots, (s_n, t_n) \in \mathcal{S} \times \mathcal{T}$ where in general some spatial locations have been sampled at multiple time instances. Often, one is interested in modelling Z from the sample $\mathbf{Z}$ in order to predict at unobserved locations in space and time or simulate from the distribution. The spatio-temporal random field might as well be accompanied by a co-variate Y leading to a bivariate spatio-temporal random field: $(Z, Y) : \Omega \times \mathcal{S} \times \mathcal{T} \to \mathbb{R}^2$.

Copulas describe the dependence between the margins of multivariate distributions. Sklar (1959) proved that any n-variate distribution H can be split into its margins $F_1, \ldots, F_n$ and the copula C which couples the margins with a given dependence structure: $H(x_1, \ldots, x_n) = C_n\big(F_1(x_1), \ldots, F_n(x_n)\big)$. A copula can be imagined as a multivariate cumulative density distribution with uniform margins where its density reflects the strength of dependence between the margins. Many different parametric copula families exist allowing for very different dependence structures including certain symmetry properties but as well asymmetric, directional influences. In a bivariate symmetric case, the strength of dependence, i.e., the copula density (denoted by c), of a pair $(u, v) \in [0, 1]^2$ does not depend on the order, i.e., $c(u, v) = c(v, u)$, $\forall (u, v) \in [0, 1]^2$ while this does not hold for an asymmetric copula. Thinking of u and v as cumulative distribution values of two consecutive time steps, this allows to model a sudden rise in value with a different strength of dependence than a sudden drop. For further details we refer to the introductory book by Nelsen (2006).

Sklar's Theorem is true for any dimension $d \geq 2$, but we will at first only consider bivariate copulas $C : [0, 1]^2 \to [0, 1]$. The density of a copula expresses the strength of dependence which changes over the range of the marginal distributions. The only copula exhibiting a constant strength of dependence across its margins is the product copula Π describing independence.
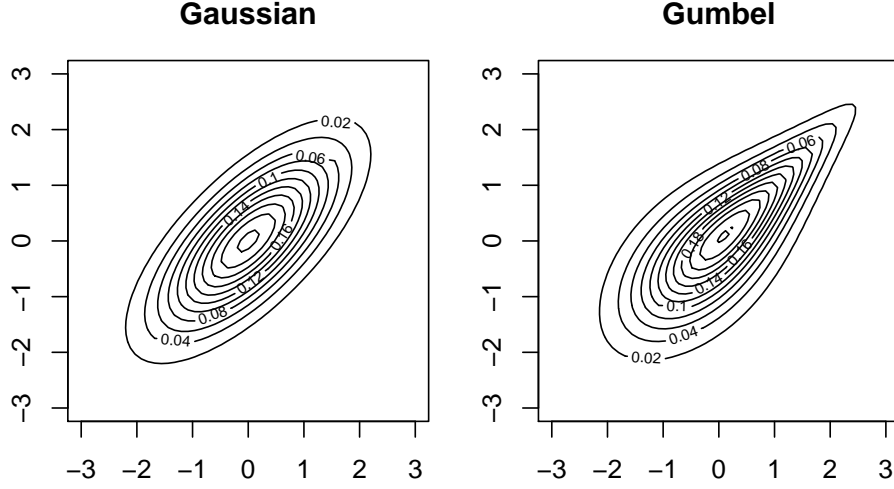
Figure 1: Copula densities of the Gaussian and Gumbel copulas. Both copulas are shown with standard normal distributed margins and a Kendall's tau correlation of 0.5.

Commonly, strength of dependence in a bivariate setting is measured by the correlation (or covariance) between two random variables and a Gaussian distribution is typically assumed implicitly. As a Gaussian distribution can be decomposed into a Gaussian copula with Gaussian margins, one imposes the Gaussian dependence structure which is elliptically symmetric (following the notion of elliptical contours of the bivariate Gaussian distribution). Hence, by only investigating the correlation of two variables, potential deviations of dependence from the Gaussian elliptical model are neglected. Different copulas might reflect samples having identical correlation, but different dependence structure (Figure 1). The same applies to the spatial and spatio-temporal domain where kriging implicitly assumes a Gaussian dependence structure. However, looking into different data sets and investigating pairwise scatter plots reveals non-Gaussian dependence structures. Especially the correlation structure of pairs spanning across time may exhibit an asymmetric (directional) dependence. Such structures can be captured by copulas.

Copulas allow to model the dependence structure of a multivariate distribution disjoint from the marginals. This introduces a huge flexibility and eases the estimation at the same time. As the analytically known multivariate copula families are rather limited in their flexibility, we use vine copulas that allow to flexibly build multivariate copulas by any combination of bivariate ones. For a successful model it is important to obtain good fits of both, margins and copula. The fitting of the margins can be carried out with any approach available in the literature. For the subsequent development of spatio-temporal vine copulas, we assume to have marginal distributions $F_{q,r}$ of Z and $G_{q,r}$ of Y for any location $(s_q, t_r) \in \mathcal{S} \times \mathcal{T}$.

We briefly introduce *bivariate spatial copulas* as in Gräler (2014) by incorporating distance as the only parameter but utilizing the flexibility of many bivariate copula families. For pairs of locations we assume that the separation distance of these is the driving parameter determining the dependence. Hence, pairs of locations very close to each other are likely to exhibit a dependence structure close to perfect dependence where noise might reduce the strength of dependence to some degree (analogous to the nugget effect in kriging). For large

distances, the pairs will tend to be independent and are modelled by the product copula Π. The approaches by Bárdossy (2011) and Kazianka and Pilz (2010) allow only for a single multivariate copula family. The bivariate spatial copula $c_h(u, v)$ recalled here is designed as a convex combination of bivariate copulas (in terms of their densities) that is not limited to a single family (Equation 1). Hence, we allow not only for a varying strength of dependence but also for a dependence structure changing with distance:

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

where $\lambda_j := \frac{h-l_{j-1}}{l_j-l_{j-1}}$ is a linearly interpolated weight, h denotes the separating distance and $l_1, \ldots, l_k$ denote chosen distances of spatial bins (e.g., midpoint or mean distance of all point pairs involved in the estimation). The parameters of the copulas $c_{i,h}$ in the convex combination may as well depend on the distance h. With the help of the marginal CDF or a rank order transformation, the arguments u and v are the values of the pairs of locations transformed to $[0, 1]$. Inspecting Equation 1 reveals that different choices of bins will in general yield different approximations to the underlying spatial dependence structure. This binning faces the same balancing issue as a classical variogram estimation where many bins allow for a flexible model but too few observations per bin and conversely few but well filled bins reduce the flexibility. It is important to maintain enough pairs per bin to allow for a sensible copula family selection throughout the estimation process.

The temporal extension of the bivariate spatial copula is yet another convex combination of bivariate spatial copulas c_h^Δ at different time lags Δ. In the case where one does not want to predict the spatio-temporal random field between time steps, the bivariate spatio-temporal copula can be reduced to a piecewise defined copula where the temporal lag between both spatio-temporal locations selects the bivariate spatial copula to be used.

Concentrating on a local neighbourhood of d spatio-temporal neighbours (Figure 2), we now model the pair-wise dependence between locations through a bivariate spatio-temporal copula. However, these copulas need to be joined to benefit from the full d-dimensional distribution of the neighbourhood. A technique to combine bivariate copulas into multivariate copulas has been introduced by Aas *et al.* (2009) building on work from Bedford and Cooke (2002). This approach has first been introduced as the *pair-copula construction* and the resulting copulas are now known as *vine copulas* in the literature.

Vine copulas approximate multivariate copulas through bivariate building blocks (Figure 3). The joint density is obtained as the product of all bivariate copula densities involved. In the special case of spatio-temporal vine copulas with an additional covariate $Y : \mathcal{S} \times \mathcal{T} \to \mathbb{R}$, we model the first tree by bivariate spatio-temporal copulas c_h^Δ and c_{ZY}, the copula describing the dependence between Z and Y. The remaining upper trees are modeled by $c_{j,j+i|0,\ldots,j-1}$ fixed over space and time:
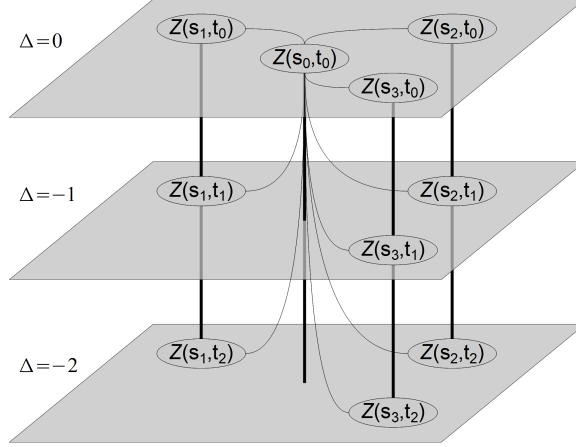
Figure 2: A spatio-temporal neighbourhood including the three spatially closest neighbours at three different time lags.

$$c_{\mathbf{h}}^{\mathbf{\Delta}}(u_0, v_0, u_1, \ldots, u_d)$$

$$=c_{ZY}(u_0, v_0) \cdot \prod_{i=1}^{d} c_{h(0,i)}^{\Delta}(u_0, u_i) \cdot \prod_{i=1}^{d} c_{Y,i|0}(u_{Y|0}, u_{i|0})$$

$$\cdot \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|Y,0,\ldots,j-1}(u_{j|Y,0,\ldots,j-1}, u_{j+i|Y,0,\ldots,j-1}) \tag{2}$$

where $v_0 = G_0\big(Y(s_0, t_0)\big)$ with G_0 being the co-variate's Y marginal cumulative distribution function at (s_0, t_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 closest neighbor of (s_0, t_0) with marginal cumulative distribution function $F_i = F_{q,r}$. For the spatio-temporally fixed upper part of the vine it is

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

$$u_{i|0} = F_{i|0}(u_i|u_0) = \frac{\partial C_{h(0,i)}^{\Delta}(u_0, u_i)}{\partial u_0} \tag{3}$$

$$\tag{4}$$

$$u_{j+i|Y,0,\ldots,j-1} = F_{j+i|Y,0,\ldots j-1}(u_{j+i}|v_0, u_0, \ldots, u_{j-1})$$

$$= \frac{\partial C_{j-1,j+i|Y,0,\ldots j-2}(u_{j-1|Y,0,\ldots j-2}, u_{j+i|Y,0,\ldots j-2})}{\partial u_{j-1|Y,0,\ldots j-2}}$$

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

In general, different decompositions of a multivariate copula exist, refereed to as regular vines, but in the spatial or spatio-temporal interpolation where a central element is naturally identified, we use a canonical vine where all initial dependencies are with respect to the central location. In the spatio-temporal tree (first tree in Figure 3) of the spatio-temporal vine, all
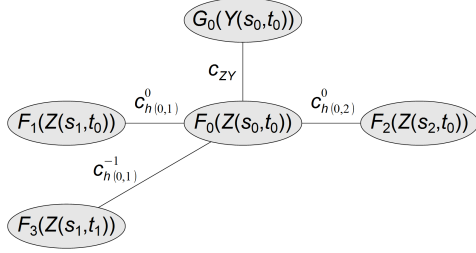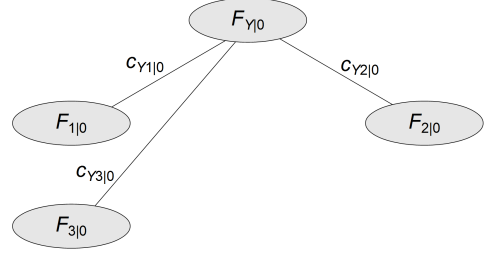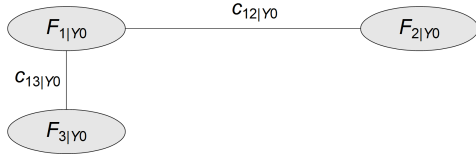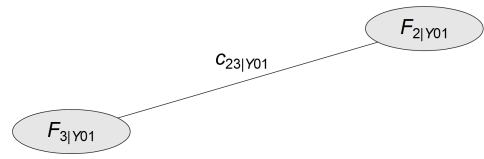
Figure 3: Graphical representation of a 5-dimensional local spatio-temporal vine copula with covariate Y, reoccurring location s_1 at the current and one preceding time slice and location s_2 at the current time slice. The notation follows the one introduced in Equation 2.

edges connecting different spatio-temporal neighbours are modelled through a spatio-temporal copula $c_{h(0,q)}^{\Delta}$ parametrized by the spatial distance $h(0, q)$ and temporal lag $\Delta = t_0 - t_r$ between the data locations (s_0, t_0) and (s_q, t_r). The edge connecting the central location with its co-located covariate $Y(s_0, t_0)$ is represented by the best fitting bivariate copula c_{ZY}. In general, the dependency between the variable of interest and its covariate might as well change over space and time. All consecutive upper trees are modelled through spatio-temporally constant copulas. The upper vine structure does not impose any restriction on the bivariate copulas involved and are kept fixed no matter how the neighbourhood might be organized. The conditional distribution functions given in the above equations can immediately be obtained as partial derivatives of the already modelled copulas.

To achieve a full distribution describing the local behaviour of the spatio-temporal random field Z, margins need to be joined with the spatio-temporal vine copula. Depending on the properties of the phenomenon to be modelled, one might use a single margin for all locations (in case the random field can be assumed to be stationary) or several margins incorporating some trend that is based for example on location, elevation or additional covariates. The density of the full distribution is obtained by multiplying the copula's density with the marginal densities and the variables are mapped to the copula scale through the marginal cumulative distribution functions G_0 for the co-variate Y at (s_0, t_0) and $F_i = F_{q,r}$ of Z with (s_q, t_r) being the i-th neighbour of (s_0, t_0):

$$f_{\mathbf{h}}^{\mathbf{\Delta}}(z_0, v_0, z_1, \ldots, z_d) = g_0(y_0) \cdot \prod_{i=0}^{d} f_i(z_i) \cdot c_{\mathbf{h}}^{\mathbf{\Delta}}\big(F_0(z_0), G_0(y_0), F_1(z_1), \ldots, F_d(z_d)\big) \qquad (5)$$

where the z_i are representations of the random field Z at (s_i, t_i), the i-th neighbour of (s_0, t_0).

3. Spatio-temporal vine copula estimation

We introduce an estimation procedure for the spatio-temporal vine copula that borrows ideas from classical geostatistical approaches. To estimate the bivariate spatio-temporal copula, the data is grouped into spatial bins for each temporal lag. Kendall's tau correlation measure is marginally independent and thus represents the correlation at the copula level. This makes it very useful in the application of copulas and some one-parameter copula families even exhibit a one-to-one relationship between Kendall's tau and their parameter. The correlogram, using Kendall's tau, is calculated for the binned data. For each bin, several copula families are fitted to the transformed data (using a rank-order transformation or the fitted cumulative distribution functions of the margins) and the best fitting family (based on e.g., likelihood, AIC or BIC) is selected. When one restricts the set of copula families to those exhibiting a direct link between Kendall's tau and their parameter, one might fit functions to the afore obtained correlograms. These functions then relate separating distance through Kendall's tau to a parameter estimate for the copulas involved in the convex combination for each temporal lag. This way, the bivariate spatio-temporal copula will exactly reproduce Kendall's tau for any distance as modelled through the function from the correlogram. In case several best fitting families cannot be parametrized through Kendall's tau, one representative fit for each bin is obtained and combined as given in Equation 1. Using these static representations in the convex combination of copulas produces Kendall's tau values as a piecewise linear interpolation of the values obtained in the correlogram.

For further processing, the data needs to be re-arranged in spatio-temporal neighbourhoods of central locations and their spatio-temporally closest neighbours. Typically, the complex dependence structure over space and time does not relate to an easily obtained Euclidean distance measure in $\mathcal{S} \times \mathcal{T}$. To select the most correlated neighbours, a considerably larger spatio-temporal block neighbourhood (Figure 2) as the target dimension of the spatio-temporal vine copula is investigated and the d neighbours having the strongest correlation using the fitted correlogram functions are selected. This introduces some additional flexibility compared to the approach described by Gräler and Pebesma (2012) as the neighbourhood does not depend on a fixed spatio-temporal block size and missing values may easily be replaced by the next strongest correlated locations. These neighbourhoods generate a $d + 1$-dimensional dataset with approximately uniform on $[0, 1]$ distributed margins. The bivariate spatio-temporal copula c_h^{Δ} on the first tree can now be used to derive the conditional sample of dimension d (conditioned to the value at the central location (s_0, t_0)). The spatio-temporally conditioned data is combined with data conditioned on the covariate and used for the remainder of the vine (Figure 3). The spatio-temporally fixed vine copula estimation proceeds sequentially by using the best fitting copula per bivariate pair. Details on the upper vine estimation are provided in Aas *et al.* (2009), Czado, Schepsmeier, and Min (2012) and Dissmann, Brechmann, Czado, and Kurowicka (2013).

The joint copula density $c_{\mathbf{h}}^{\boldsymbol{\Delta}}$ can then be obtained through Equation 2 where the first sequence of products reflects the spatio-temporal tree. The remaining spatio-temporally constant trees appear in the second and third product sequences. Fitting the marginal distributions, following generally any approach available in the literature, yields a full distribution through Equation 5 describing the local behaviour of the spatio-temporal random field $Z \times Y$.

4. Prediction of the spatio-temporal random field

The local representation of the random field Z can be used for different purposes. A typical task is prediction of the modelled phenomenon at unobserved locations in space and time. To produce such predictions from a local neighbourhood, every unobserved location needs to be grouped with its d closest, i.e., strongest correlated, observed neighbours. Conditioning the $d+1$-dimensional copula $c_{\mathbf{h}}^{\boldsymbol{\Delta}}$ on the observed values, yields a 1-dimensional distribution of the phenomenon. This conditional distribution can then be used to calculate the expected value (Equation 6), median or any other desired quantile (Equation 7) denoting for instance confidence intervals. At any location $s_0 \in \mathcal{S}$, the predictors for the mean value $\widehat{Z}_m$ and quantile values $\widehat{Z}_p$ for any $p \in (0, 1)$ are:

$$\widehat{Z}_m(s_0) = \int_{\mathbb{R}} z \cdot f_{\mathbf{h}}^{\boldsymbol{\Delta}}(z|y_0, z_1, \ldots, z_d) \, \mathrm{d}z$$

$$= \int_{[0,1]} F_0^{-1}(u) \, c_{\mathbf{h}}^{\boldsymbol{\Delta}}(u|v_0, u_1, \ldots, u_d) \, \mathrm{d}u \tag{6}$$

$$\widehat{Z}_p(s_0) = F_0^{-1}\big(C_{\mathbf{h}}^{\boldsymbol{\Delta}^{-1}}(p|v_0, u_1, \ldots, u_d)\big) \tag{7}$$

where $u_i = F_i(z_i) = F_i\big(Z(s_i, t_i)\big)$ for $1 \le i \le d$ and $v_0 = G_0(y_0)$ as before. The equality for $\widehat{Z}_m$ is based on a probability integral transform. An advantage of this approach is that the conditional distribution describing the random field at the unobserved location may take any form. This is different from kriging, where every predictive distribution is again a normal distribution. This richer flexibility has the potential to provide more realistic uncertainty estimates. Another advantage that is immediate from Equation 6 and Equation 7 is that the only information on the marginals needed is their quantile function. This allows for instance to use approximations derived from the empirical cumulative distribution function without the knowledge of any explicitly known form of the family's density. However, the empirical cumulative distribution function is typically limited to the domain defined by the smallest and largest observation.

5. Application

The following calculations have been made using R 3.0.2 (R Core Team 2013) and can be reproduced with the **spcopula** R package. The demo `stCoVarVineCop` runs the estimation of the spatio-temporal vine copula as described below (on a 2 month subset of the full dataset).

Data

The dataset used in this application was obtained from the openly accessible AirBase[1], an European air quality data base maintained by the European Environmental Agency (EEA). We consider daily mean rural background PM_{10} concentrations (particulate matter smaller than 10 μm in diameter) in μg/m^3 across Europe for the entire year 2005. The data consists of 194 rural background stations with some missing observations at random points in time. A histogram of the skewed distribution is depicted in Figure 4. Preliminary results of this study excluding covariates and station wise marginal distributions were presented by Gräler and

---

[1]available from http://acm.eionet.europa.eu/databases/airbase/
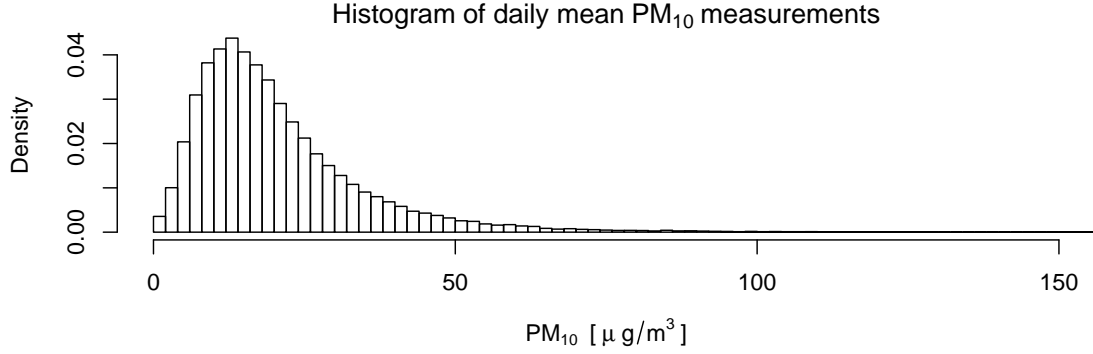
Histogram of daily mean PM$_{10}$ measurements

Figure 4: Histogram of the daily mean PM$_{10}$ rural background concentrations across Europe during the year 2005. 60 observations extend beyond the plot up to approximately 400 μg/m^3.

Pebesma (2012) and the same data set has been analysed by Gräler, Gerharz, and Pebesma (2012) using spatio-temporal approaches based on kriging. As a covariate, we included daily mean PM$_{10}$ concentrations derived from model driven estimates by the European Monitoring and Evaluation Programme (EMEP 2007). The 50 km EMEP Polar Stereographic grid is converted and projected to match the 10 km gridded interpolation domain in the standard EEA ETRS89-LAEA5210 projection.

Marginal distributions F_s and G_s

We fit marginal distributions for each location $s \in \mathcal{S}$ based on the time series leading to margins F_s and G_s for daily mean PM$_{10}$ measurements and EMEP model estimates respectively. This leads to a slightly less general set-up as introduced earlier where the marginal distributions might as well change over time. Using the **evd** R package (Stephenson 2002), a generalized extreme value distribution is automatically fitted to each station's time series of PM$_{10}$ and the EMEP model estimates. The assumptions of a single marginal distribution describing all stations could not be verified for either variable, as too many stations rejected these distributions in a Kolmogorov–Smirnov test. Obviously, these marginal distributions can only directly be fitted where we observed data. For an interpolation scenario, the marginals need to be extended towards prediction locations as well. Assuming that the marginal distributions change rather smoothly over space, we use two different models based on spatial proximity. One relies on a linear model incorporating the locations' coordinates and altitude followed by an inverse distance weighted interpolation of the residuals and the second one uses only inverse distance weighted means of the local neighbourhood's marginal parameters as estimates. Both approaches only use the spatially closest 9 locations for the inverse distance weighted means. Even though the spatio-temporal field we are modelling is not assumed to be stationary, we assume that the local dependence structure is the same all over Europe and can sufficiently be described by a station's 9 strongest correlated neighbours from up to 4 preceding time steps. In the following code snippets, we assume `EU_RB_2005` to be a spatio-temporal full data frame (class `STFDF`, see the documentation of **spacetime** (Pebesma 2012) for details) that holds the already station-wise marginal transformed data and model estimates denoted as `marPM10` and `marEMEP` respectively.
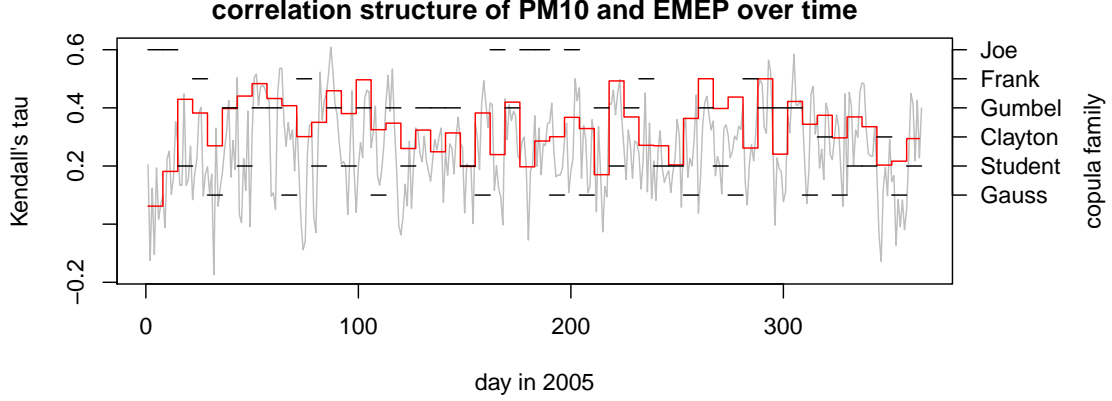
Figure 5: Correlation structure of daily mean PM_{10} measurements and EMEP predictions over time. The grey line indicates daily Kendall's tau values while the red step function describes weekly empirical Kendall's tau values. The black line segments denote the copula family best describing the dependence structure during each week for the given empirical correlation (without any particular ordering).

Covariate copula C_{ZY}

Starting with the covariate copula c_{ZY}, we investigate the correlation structure of the daily mean PM_{10} measurements and EMEP model estimates over time. Figure 5 illustrates how the strength of correlation and the dependence structure changes throughout the year 2005. We use weekly correlations (red line in Figure 5) averaging out a great deal of variation, but largely maintaining the changes over time. The copula families are selected among the elliptical Gaussian and Student and the Archimedean Clayton, Gumbel, Frank (Nelsen 2006) and Joe (Joe 1997) copula families as indicated by the black line segments in Figure 5. The marginal independence of Kendall's tau ensures that this strength of dependence is the same for the marginal transformed as well as the raw data. This temporally changing covariate copula needs to be encoded as function taking the current spatio-temporal indices and returning a copula object:

```
R> library("spcopula")
R> coVarCop <- function(stInd) {
+    week <- min(ceiling(stInd[2] / 7), 52)
+    copulaFromFamilyIndex(weekCop[[week]]$family, weekCop[[week]]$par,
+                          weekCop[[week]]$par2)
+ }
```

Spatio-temporal bivariate copula

For the estimation of the spatio-temporal bivariate copula, we follow the suggested procedure from Section 3 and start by grouping the data into spatial bins for five temporal lags (i.e., the same and first to fourth preceding day). For each spatio-temporal lag, the mean distance of all involved pairs and their Kendall's tau are calculated by:

```
R> stBins <- calcBins(EU_RB_2005, "marPM10", nbins = 40, tlags = -(0:4))
```
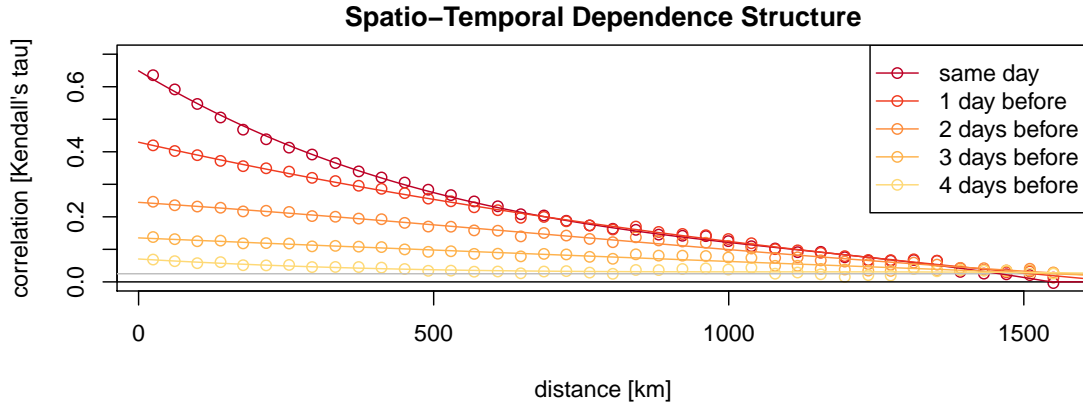
Figure 6: Empirical and modelled values of Kendall's tau for the bivariate spatio-temporal copula over five temporal lags.

The resulting object `stBins` is of type `list` with entries `meanDists` denoting the mean spatial distance of spatio-temporal bins, `lagCor` holding the correlation values per spatial bin and temporal lag and `lags` providing spatial and temporal indices to access the data from the underlying `STFDF`. Correlation functions (here five polynomials of degree three each) can directly be fitted to the above output and are joint in a spatio-temporal dependence function (`stDepFun`) through:

```
R> stDepFun <- fitCorFun(stBins, rep(3, 5), tlags = -(0:4))
```

Figure 6 illustrates the empirical values of Kendall's tau per spatial bin and temporal lag alongside with the corresponding polynomial fits. These polynomials describe how the strength of dependence changes with spatial and temporal distance. Following the estimation procedure, we need to investigate how the dependence structure, i.e., the copula families change for different spatial bins and temporal lags. The polynomials describing Kendall's tau in terms of distance are used to derive the parameter of a set of copula families and their log-likelihood is calculated per spatial bin and temporal lag. In case the relationship between Kendall's tau and the copula parameters is not unique, the parameter is optimised based on a log-likelihood approach restricted under the desired value of Kendall's tau. The function `loglikByCopulasStLags` calculates the log-likelihoods per spatial bin and temporal lag additionally returning the evaluated copula:

```
R> families <- c(normalCopula(0), tCopula(0),
+                claytonCopula(0), frankCopula(1), gumbelCopula(1),
+                joeBiCopula())

R> loglikTau <- loglikByCopulasStLags(stBins, EU_RB_2005, families,
+                                     stDepFun)
```

Copula families considered for the bivariate spatio-temporal copula (`families`) include the elliptical Gaussian and Student copulas, the Archimedean Clayton, Frank, Gumbel (Nelsen 2006) and Joe (Joe 1997) copulas. The best fitting copula family is selected based on the highest log-likelihood.

	Spatial lag														
ID	1	2	3	5	6	7	22	23	25	26	27	28	29	30	33
mean dist. [km]	25	61	99	177	216	255	843	881	961	999	1038	1079	1117	1156	1274
$\Delta = 0$	t	G	t	…	t	G			…		G	F	N	F	N	…
$\Delta = -1$	G			…				G	F		…	F	N	F	N	N
$\Delta = -2$	G				…				G	N			…		
$\Delta = -3$	G					…				G	N		…		
$\Delta = -4$	G	…	G	J			…				J	G	G	N	…

Table 2: Spatio-temporal bivariate copula family configuration for the first 33 spatial lags as suggested by the highest log-likelihoods. Δ indicates the time lag and the copula families are abbreviated as follows: N = Gaussian, t = Student, C = Clayton, F = Frank, G = Gumbel and J = Joe.

```
R> bestFitTau <- lapply(loglikTau,
+                     function(x) apply(apply(x$loglik, 1, rank),
+                                       2, which.max))
```

In this application, the copula families change rather little and the Gumbel copula family (compare Figure 1) dominates the dependence structure. The spatio-temporal bivariate copula configuration is listed in Table 2. Using the earlier fitted polynomials and this selection of copula families, the convex combination of copulas (Equation 1 and the following paragraph) can now be composed to a spatio-temporal bivariate copula. The selected copula fits `listCops` and the representative distances `listDists` are provided as lists with one entry for each temporal lag. Each of this entries contains a list of copulas in spatially ascending order:

```
R> distSelect <- function(x) {
+    stBins$meanDists[sort(unique(c(which(diff(x) != 0),
+                                   which(diff(x) != 0) + 1, 1, 40)))]
+ }
R> listDists <- lapply(bestFitTau, distSelect)

R> famSelect <- function(x) {
+    families[x[sort(unique(c(which(diff(x) != 0),
+                     which(diff(x) != 0) + 1, 1, 40)))]]
+ }
R> listCops <- lapply(bestFitTau, famSelect)
```

As the corresponding Kendall's tau value used to tune the bivariate copula's parameter is calculated for each spatio-temporal distance, the above components and the spatio-temporal dependence function define the bivariate spatio-temporal copula `stBiCop`:

```
R> stBiCop <- stCopula(components = listCops, distances = listDists,
+                   tlags = -c(0:4), stDepFun = stDepFun)
```

Joining vine copula

For the further processing, the data needs to be rearranged in local neighbourhoods. In our application, we are interested in the nine strongest correlated neighbours. Typically, there

is no easily identified metric selecting these and we start with a larger static neighbourhood structure following Figure 2. The function `reduceNeighbours` selects the n strongest correlated neighbours as given in the third argument. To accomplish this, the spatio-temporal distances of the larger cubic neighbourhood are passed to the spatio-temporal dependence function `stDepFun` and the ones with the highest values are selected:

```
R> stNeigh <- getStNeighbours(EU_RB_2005, var = "marPM10", spSize = 9,
+                             tlags = -(0:4), timeSteps = 90, min.dist = 10)
R> stRedNeigh <- reduceNeighbours(stNeigh, stDepFun, 9)
```

This approach allows to drop missing values and to select the next best (i.e., next strongest correlated) available neighbour. The argument `spSize` denotes the number of locations at the current level including the central location. By setting `timeSteps` to 90, every location will only be used at 90 randomly assigned time steps as a centre of the neighbourhood. This way, only temporal chunks of five consecutive days are used reducing unwanted autocorrelation in the time series but reflecting the local temporal structure of the phenomenon. With `min.dist` one ensures that any pair of neighbours has a minimum spatial separation distance (here 10 m).

The conditioning on the covariate EMEP and the evaluation of the spatio-temporal bivariate copula on the neighbours can be done separately. The spatio-temporal bivariate copula returns pseudo observations that are the values of the neighbours conditioned on the value of the central location ($u_{i|0}$ from Equation 3) incorporating the spatio-temporal distances between these by

```
R> condData <- dropStTree(stRedNeigh, EU_RB_2005, stBiCop)
```

In a loop, the weekly varying copulas as depicted in Figure 5 and encoded as `coVarCop` are used to relate the marginal transformed variable PM_{10} with its covariate EMEP. In order to select the appropriate copula, the spatio-temporal indices need to be retrieved from the larger neighbourhood:

```
R> condCoVa <- condCovariate(stNeigh, coVarCop)
```

Binding this conditioned column `condCoVa` with the conditioned data from the spatio-temporal bivariate copula `condData` yields the data set for the upper vine estimation through the generic function `fitCopula`:

```
R> secTreeData <- cbind(condCoVa, as.matrix(condData@data))
R> vineFit <- fitCopula(vineCopula(10L), secTreeData)
```

Following the initial definition of the function `fitCopula`, the copula family that should be fitted needs to be provided as copula object. The call of the constructor `vineCopula` with an integer as argument (denoting the dimension) generates a canonical vine with independence copulas. The fitting routine sequentially selects the best fitting copula from a versatile set of copulas (see the documentation of **VineCopula**). The final spatio-temporal covariate vine copula is defined by:

```
R> stCVVC <- stCoVarVineCopula(coVarCop, stBiCop, vineFit@copula)
```

The cross-validation carried out to assess the goodness of fit of this spatio-temporal vine copula drops the complete time series of each location in turn and predicts this time series based on the neighbouring data. In this application, we predict the expected value as given in Equation 6 of the time series for each moment in time. Prediction for a spatio-temporal target geometry `targetGeom` from the spatio-temporal vine copula is obtained by:

```
R> predNeigh <- getStNeighbours(EU_RB_2005, targetGeom,
+                                spSize = 9, tlags = -(0:4),
+                                var = "marPM10", coVar = "marEMEP",
+                                prediction = TRUE, min.dist = 10)
R> predNeigh <- reduceNeighbours(predNeigh, stDepFun, 9)
R> stVinePred <- stCopPredict(predNeigh, stCVVC, list(q = qFun),
+                             method = "expectation")
```

Where the argument `method` selects from the prediction methods `expectation` and `quantile` as given in Equation 6 and Equation 7 respectively. The default quantile is the median, but any fraction between 0 and 1 can be assigned to an argument `p`. This can also be used for simulation purposes, as simulated values can be obtained through uniform distributed fractions assigned to `p`.

6. Results and discussion

Performing a cross-validation by leaving one complete station time series out in turn, the performance of this approach is assessed. Table 3 lists the results of a cross-validation using the expectation predictors for the newly presented spatio-temporal covariate vine copula approach (STCV) for different margins, a STCV solely composed out of Gaussian copulas and the simpler spatio-temporal copula presented in Gräler and Pebesma (2012). Additionally, results of a cross-validation for an approach based on spatio-temporal metric residual kriging with an underlying log-linear regression of the same data set but incorporating 100 nearest neighbours (Gräler *et al.* 2012) is presented. In the case where the marginals refer to *local GEV*, the station-wise estimates are used. This is only possible in the case of a cross-validation as in a typical prediction or simulation application one would not have this extra knowledge about the margins. This can be seen as a very good model of the marginals across space as the distributions of the margins are still not known, but extra data is used to estimate them. Hence, in a realistic cross-validation scenario, an additional model on the marginal distributions' parameters needs to be used. Here, we fitted two models for each of the three parameters. One using a linear model including coordinates and the station's altitude and performing a inverse distance weighted interpolation on the residuals of the spatially closest 9 neighbours (denoted *lm+IDW*) and another using only an inverse distance weighted interpolation of the spatially closest 9 neighbours (denoted *IDW*). The major benefit in Table 3 is due to the additional knowledge on the marginal distribution functions. A smaller improvement could be made for the STCV using the lm+IDW marginal distributions but these statistics are within the same order of magnitude as the earlier presented spatio-temporal vine copula or the kriging predictor. These results underline how important it is to obtain good fits of both, the copula and the marginal distributions.

Besides looking at pure cross-validation statistics, it is important to consider the reproduction of the full distribution. Figure 7 shows boxplots of the observed versus modelled time series at

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

Table 3: Cross-validation results comparing the expectation spatio-temporal estimators for the newly presented spatio-temporal covariate vine copula with different margins, the spatio-temporal vine copula as in Gräler and Pebesma (2012) and results from an earlier study by Gräler *et al.* (2012) using kriging based approaches assuming a metric spatio-temporal covariance structure of the residuals of log-linear regression.
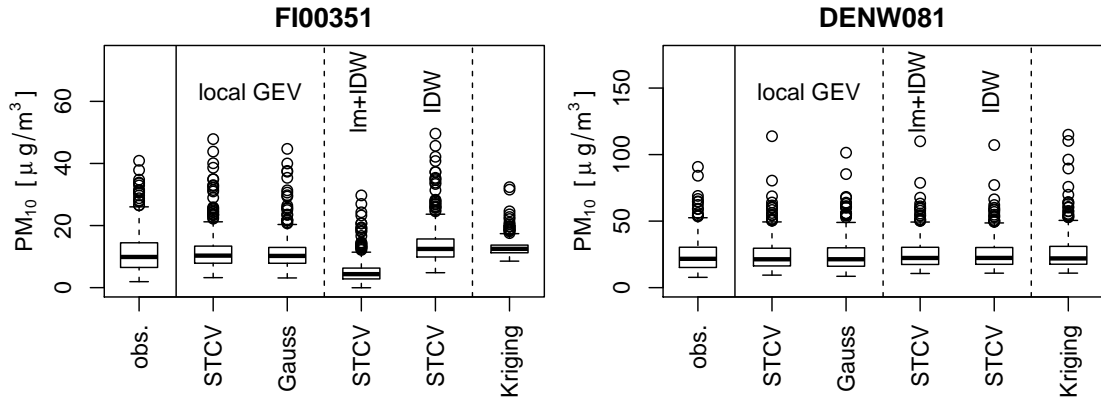

Figure 7: Boxplots comparing the different predictors with the observed values at two stations. The Finish one (left) is an isolated location while the German one (right) is situated in a rather dense network.
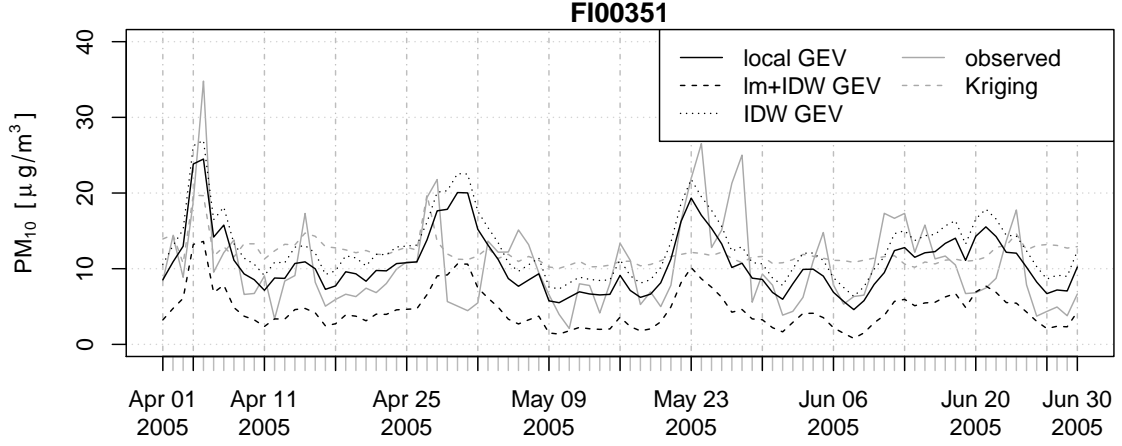
Figure 8: A subset of the time series at a Finish station comparing different predictors.

two exemplary stations in Finland (FI00351) and Germany (DENW081). The Finish one is far away from any other station while the German one is situated in a rather dense network. Looking at many more plots of this kind for several stations (not shown) reveals that the copula based methods are rather close to each other and represent the original data quite well. Within the copula approaches, the full cross-validations *lm+IDW* and *IDW* have the largest deviations. The kriging based approach turns out to predict too large as well as too small ranges of values depending on the single station with a notable shift of the median at some stations. The issue of failing to reproduce the time series at isolated locations detected in the first application of a simpler spatio-temporal vine copula (Gräler and Pebesma 2012) could be overcome with the local GEV margins and considerably improved with *lm+IDW* and *IDW*. In Figure 8 the same temporal subset is plotted as in Figure 3 of Gräler and Pebesma (2012) but the copula approaches are now able to follow the time series more closely than the kriging based predictor.

An advantageous feature of the copula approaches is their ability to provide potentially more reliable uncertainty assessments. Different from kriging, the prediction variance does not only depend on the spatio-temporal configuration of the locations, but also on the predicted value. Due to the nature of kriging, every conditional distribution is again a normal distribution. This is different for the copula approaches where the predictive density can take any form and the fitted marginal distribution functions ensure a reasonable range of possible values. Figure 9 illustrates the predictive densities at two different days at location DENW081. Note that besides the position also the shape of the prediction CDFs changes.

Even though simulating from a spatio-temporal random field modelled by a spatio-temporal covariate vine copula has not been shown in this application, it is readily done by not predicting for a constant cumulative distribution value but a random p in $\widehat{Z}_p(s,t)$ for each location (s,t). In a conditional simulation, it is a modellers choice to which degree already sampled and closer locations are preferred over conditioning but further apart locations in the local neighbourhood. Modelling the spatio-temporal random field only locally requires a simulation along a random path. To avoid clustering effects along this path, one might start with simulations on a regular coarse subset of the target locations and simulate subsequently on finer regular subsets (Gómez-Hernández 1991).
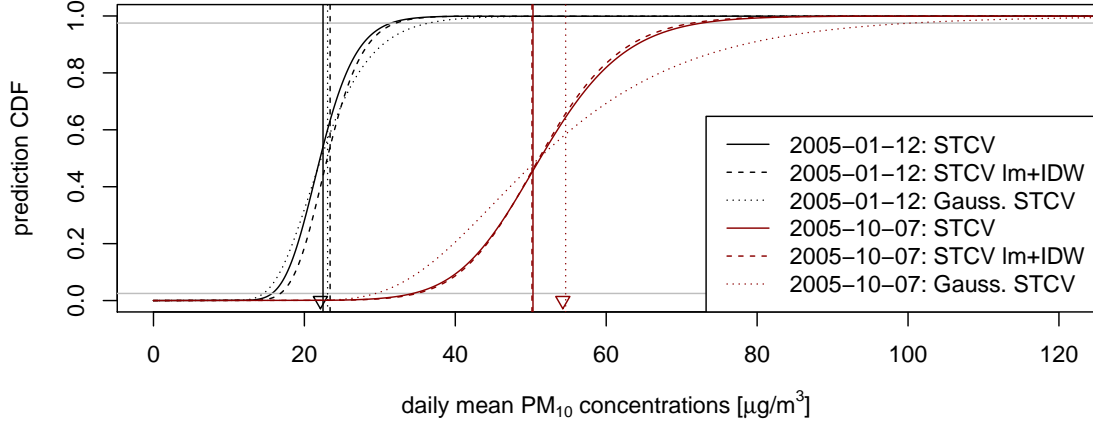
Figure 9: Prediction CDFs for station DENW081 at two different days in 2005. The vertical lines denote the predicted while the symbols on the x-axis indicate the observed value. The two horizontal lines denote the cumulative probabilities 0.025 and 0.975 easing the assessment of the 95%-confidence intervals.

Computational cost of the vine copula approaches might be a burden where results need to be generated very fast. The prediction of approximately 70000 values in the presented cross-validation takes a bit more than a day on a common laptop. No efforts have been made to speed up the execution for example by making use of parallel evaluation.

7. Conclusions

The **spcopula** package allows the modelling of spatial and spatio-temporal random fields by vine copulas. An earlier study (Gräler 2014) demonstrated the potential of spatial vine copulas for heavily skewed spatial random fields. The newly presented spatio-temporal covariate vine copula improves the interpolation of particulate matter compared to earlier spatio-temporal vine copulas (Gräler *et al.* 2012) and linear geostatistical approaches. Nevertheless, the current bottleneck of the presented application is the flexible modelling of marginal distributions. This became evident by the presented cross-validation using extra knowledge about the margins.

The bivariate spatio-temporal copulas on the first tree mainly incorporate a Gumbel copula indicating a stronger dependence in the tails, a feature that is not available in a Gaussian set-up. Each spatio-temporal location has its own individual conditional distribution that is used for prediction. Different from kriging where each predictive distribution is again Gaussian, the conditional distributions of the STCV can take any form, potentially providing more realistic uncertainty estimates. Simulating from the modelled random field is possible and has been implemented in the presented package. A disjoint modelling of margins and dependence structure introduces a large flexibility to define a random fields distribution. Nevertheless, it is very important to obtain good models for both components for a successful application.

Acknowledgements

References

Aas K, Czado C, Frigessi A, Bakken H (2009). "Pair-Copula Constructions of Multiple Dependence." *Insurance: Mathematics and Economics*, **44**, 182–198. `doi:10.1016/j.insmatheco.2007.02.001`.

Bárdossy A (2006). "Copula-Based Geostatistical Models for Groundwater Quality Parameters." *Water Resources Research*, **42**(11), W11416. `doi:10.1029/2005WR004754`.

Bárdossy A (2011). "Interpolation of Groundwater Quality Parameters with Some Values Below the Detection Limit." *Hydrology and Earth System Sciences*, **15**(9), 2763–2775. `doi:10.5194/hess-15-2763-2011`.

Bárdossy A, Li J (2008). "Geostatistical Interpolation Using Copulas." *Water Resources Research*, **44**(7), W07412. `doi:10.1029/2007WR006115`.

Bárdossy A, Pegram GGS (2009). "Copula Based Multisite Model for Daily Precipitation Simulation." *Hydrology and Earth System Sciences*, **13**(12), 2299–2314. `doi:10.5194/hess-13-2299-2009`.

Bedford T, Cooke RM (2002). "Vines – a New Graphical Model for Dependent Random Variables." *The Annals of Statistics*, **30**(4), 1031–1068. `doi:10.1214/aos/1031689016`.

Cressie N, Wikle CK (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Czado C, Schepsmeier U, Min A (2012). "Maximum Likelihood Estimation of Mixed C-Vines with Application to Exchange Rates." *Statistical Modelling*, **12**(3), 229–255. `doi:10.1177/1471082X1101200302`.

Dissmann JF, Brechmann EC, Czado C, Kurowicka D (2013). "Selecting and Estimating Regular Vine Copulae and Application to Financial Returns." *Computational Statistics & Data Analysis*, **59**(1), 52 – 69. `doi:10.1016/j.csda.2012.08.010`.

EMEP (2007). *Transboundary Particulate Matter in Europe Status Report 4/2007*. Norwegian Institute for Air Research (EMEP Report 4/2007), Kjeller. URL `http://www.nilu.no/projects/ccc/reports/emep4-2007.pdf`.

Gómez-Hernández J (1991). *A Stochastic Approach to the Simulation of Block Conductivity Fields Conditioned upon Data Measured at a Smaller Scale*. Ph.D. thesis, Stanford University, Stanford, CA.

Gräler B (2014). "Modelling Skewed Spatial Random Fields Through the Spatial Vine Copula." *Spatial Statistics*. ISSN 2211-6753. `doi:10.1016/j.spasta.2014.01.001`. Available online, in pess.

Gräler B, Gerharz LE, Pebesma EJ (2012). "Spatio-Temporal Analysis and Interpolation of PM10 Measurements in Europe." *Technical report*, ETC/ACM. URL http://acm.eionet.europa.eu/reports/ETCACM_TP_2011_10_spatio-temp_AQinterpolation.

Gräler B, Pebesma EJ (2012). "Modelling Dependence in Space and Time with Vine Copulas." In *Expanded Abstract Collection from Ninth International Geostatistics Congress, Oslo, Norway June 11 – 15, 2012*. International Geostatistics Congress. URL http://geostats2012.nr.no/1742830.html.

Joe H (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

Kazianka H, Pilz J (2010). "Copula-Based Geostatistical Modeling of Continuous and Discrete Data Including Covariates." *Stochastic Environmental Research and Risk Assessment*, **24**(5), 661–673. doi:10.1007/s00477-009-0353-8.

Kazianka H, Pilz J (2011). "Bayesian Spatial Modeling and Interpolation Using Copulas." *Computers & Geosciences*, **37**(3), 310–319. ISSN 0098-3004. doi:10.1016/j.cageo.2010.06.005.

Kojadinovic I, Yan J (2010). "Modeling Multivariate Distributions with Continuous Margins Using the copula R Package." *Journal of Statistical Software*, **34**(9), 1–20. URL http://www.jstatsoft.org/v34/i09.

Nelsen RB (2006). *An Introduction to Copulas*. Second edition. Springer-Verlag, New York.

Pebesma E (2012). "**spacetime**: Spatio-Temporal Data in R." *Journal of Statistical Software*, **51**(7), 1–30. URL http://www.jstatsoft.org/v51/i07/.

Pebesma EJ, Bivand RS (2005). "Classes and Methods for Spatial Data in R." *R News*, **5**(2), 9–13. URL http://CRAN.R-project.org/doc/Rnews/.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schepsmeier U, Stoeber J, Brechmann EC (2013). ***VineCopula**: Statistical Inference of Vine Copulas*. R package version 1.1-2.

Sklar A (1959). "Fonctions de répartition á n dimensions et leurs marges." *Publ. Inst. Statist. Univ. Paris*, **8**, 229 – 231.

Stephenson AG (2002). "evd: Extreme Value Distributions." *R News*, **2**(2), 31–32. URL http://CRAN.R-project.org/doc/Rnews/.

Yan J (2007). "Enjoy the Joy of Copulas: With a Package **copula**." *Journal of Statistical Software*, **21**(4), 1–21. ISSN 1548-7660. URL http://www.jstatsoft.org/v21/i04.

Affiliation:

Benedikt Gräler
Institute for Geoinformatics

University of Münster
Heisenbergstr. 2
48149 Münster, Germany
E-mail: ben@graeler.org
URL: http://ben.graeler.org