Modelling Dependence in Space and Time with Vine Copulas

Benedikt Gräler, Edzer Pebesma

Abstract We utilize the concept of *Vine Copulas* to build multi-dimensional copulas out of bivariate ones, as bivariate copulas are quite well understood and easy to estimate. The basis of our multidimensional copula is a *bivariate spatio-temporal copula* varying over space and time. The *spatio-temporal vine copula* models the underlying spatio-temporal random field for local neighbourhoods in a fully probabilistic manner.

Focusing on the interpolation of spatially under-sampled but temporally rich random fields, we apply this newly developed approach to a large data set of daily mean PM_{10} measurements over Europe during 2005. A cross-validation study is conducted to asses the power and quality of this approach.

1 Introduction

Copulas are capable of modelling any kind of dependence between random variables detached from their margins. The ability to capture the dependencies of extreme values made them popular in finance. Extreme values can also be found in many spatial datasets and their non-Gaussian dependence structures can easily be captured with copulas. Exploiting copulas potentially improves the interpolation of skewed and heavy tailed data.

The concept of *Vine Copulas* allows us to build multi-dimensional copulas out of bivariate ones. As bivariate copulas are quite well understood and easy to estimate, vine copulas are a promising tool to model multivariate distributions. The basis of our multidimensional copula is a bivariate spatio-temporal copula varying over

Benedikt Gräler
Institute for Geoinformatics, Weseler Str. 253, 48151 Münster
e-mail: ben.graeler@uni-muenster.de

Edzer Pebesma
Institute for Geoinformatics, Weseler Str. 253, 48151 Münster

Ninth International Geostatistics Congress, Oslo, Norway, June 11. – 15., 2012

space and time. The vine copula is fitted to local neighbourhoods and used to derive estimates from the neighbourhood's multivariate distribution.

Focusing on the interpolation of spatially under-sampled but temporally rich random fields, we apply this newly developed approach to a large data set of daily mean PM_{10} measurements over Europe during 2005. To asses the goodness of this model, we conduct a cross validation on the PM_{10} measurements predicting the stations mean, median and 95%-quantile.

In the following, we will give a brief introduction to copulas and extend this concept to spatial and spatio-temporal bivariate copulas and later to spatio-temporal vine copulas. In Section 3, the new approach is applied to a one year series of daily PM_{10} measurements in Europe followed by a discussion in Section 4. In the closing section, we conclude and point to further directions of this work.

2 Copulas

Copulas are a probabilistic tool that allow to model altering dependencies across the full range of multivariate distributions. Following Sklar's theorem (see e.g. [7], as well for a detailed introduction), any d-variate distribution H can be decomposed into its marginal cumulative distribution functions $F_1, \ldots, F_d$ and its copula C by:

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

The copula C can be seen as a d-variate uniform distribution function over the hyper unit-cube $[0,1]^d$. Following the above decomposition, allows to build a vast set of multivariate distributions out of desired margins and a dedicated dependence structure.

Unfortunately, as flexibility increases with the dimension, so does the effort to estimate an appropriate copula. Quite many copula families have been discussed for the bivariate case, of which only few can easily be extended to the multivariate case without loosing the necessary flexibility. One possible approximation of multivariate copulas is obtained by *vine copulas* [1, 2, 5]. Vine copulas decompose a multivariate copula into a set of (conditional) bivariate ones. Any of these bivariate building blocks can be modelled by the best suitable copula without any restriction. This is advantageous (1) as it allows for a huge degree of flexibility and (2) as established estimation routines for the bivariate case can be used. The complete d-dimensional density of this copula is given as the product of all involved $\frac{1}{2}d(d-1)$ bivariate copulas and corresponding conditional cumulative distribution functions.

Naturally, the decomposition of a multivariate copula is not unique and a different ordering of the variables might lead to a different estimate. The two basic concepts of decomposition are called canonical vines (C-vines) and D-vines [1] where in the first approach, the first variable is used as conditioning variable for the following ones, and in the latter approach, the conditioning is done sequentially. More general decompositions are referred to as regular vines (R-vines). In this work, we will build

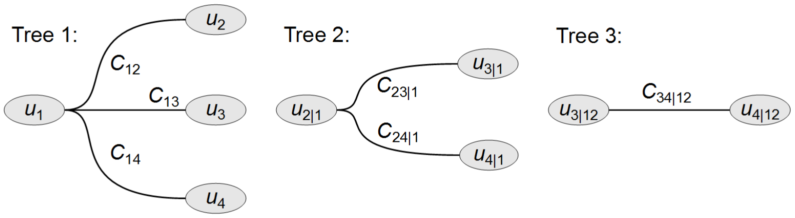

Fig. 1 Structure of a 4-dimensional C-vine. The conditioned variables $u_{i,\nu} := F_{i|\nu}(i|\nu)$, with $i \in \{2,3,4\}$ and $\nu \in \{\{1\},\{1,2\}\}$ can be derived from the bivariate copulas of the preceding tree as illustrated in equation 1.

on the canonical vine structure exemplary depicted in Figure 1 for a 4-dimensional copula. The full density of this C-vine copula is given by:

$$\begin{aligned} c(u_1,\ldots,u_4) = {} & c_{34|12}(u_{3|12},u_{4|12}) \\ & \cdot c_{23|1}(u_{2|1},u_{3|1}) \cdot c_{24|1}(u_{2|1},u_{4|1}) \\ & \cdot c_{12}(u_1,u_2) \cdot c_{13}(u_1,u_3) \cdot c_{14}(u_1,u_4) \end{aligned}$$

The conditioned variables $u_{i|\nu}$, $i \in \{2,3,4\}$ and $\nu \in \{\{1\},\{1,2\}\}$, are derived through the copulas in the preceding tree (e.g. from tree 1):

$$u_{i|1} := F_{i|1}(u_i|u_1) = \left. \frac{\partial C_{1i}(u_1,u_i)}{\partial u_1} \right|_{u_1}, \; i \in \{2,3,4\} \tag{1}$$

A similar equation holds for the higher order trees.

2.1 Spatial and Spatio-Temporal Bivariate Copulas

In the domain of geosciences, one typically deals with spatially or spatio-temporally spread data. The locations of measurements in space and time can usually be used to derive relationships of the variables. In the following, we will assume a stationary and isotropic spatial (or spatio-temporal) random field. That is, we assume for any location $s \in \mathscr{R}$ in our spatial (spatio-temporal) region $\mathscr{R}$ the random filed Z to take the same random variable $X = Z(s)$ and the dependence between two random variables $X_1 := Z(s_1)$ and $X_2 := Z(s_2)$ is a function of the separating Euclidean distance $h := ||s_1 - s_2||$ only.

Considering the task of interpolating a spatial random field, for instance, the locations (or distances between locations) are used to derive the covariance matrix for the kriging predictor. We will follow a similar avenue and define a *spatial bivariate copula* as a bivariate copula taking the distance $h \in \mathbb{R}_{\geq 0}$ as parameter with the property that for $h \to \infty$ the copula tends to the product copula $\Pi(u,v) = u \cdot v$ denoting independence. Typically, the spatial bivariate copula will tend to the upper Fréchet-Hoeffding bound $M(u,v) := \min(u,v)$ denoting perfect positive dependence as h approaches 0. However, due to missing information on the very short distance variation of the phenomenon, this bound does not have to be reached (similar as the nugget effect in kriging). In this work, the spatial copula is given as a convex linear combination of bivariate copulas where the mixing parameter function $\lambda : \mathbb{R}_{\geq 0} \to [0,1]$ and the copulas depend on the separating distance h of two locations s_1, s_2:

$$C_h(u_1,u_2) := \lambda(h) \cdot C_i(u_1,u_2) + \big(1-\lambda(h)\big) \cdot C_j(u_1,u_2),\ (i,j) := I(h)$$

Where I denotes a set of paired indicators separating the spatial range r_S of the model into a set of disjoint intervals (lags) and $I(h)$ provides the one pair of indexes (i,j) with respect to the distance h and corresponding copulas C_i and C_j. The copulas C_i and C_j denote the boundary conditions. Any distance larger than the spatial range r_S is modelled with the product copula Π. Due to the convex combination of copulas, the spatial bivariate copula will again be a copula for any distance h.

We define a *spatio-temporal bivariate copula* as a bivariate copula taking two parameters, the spatial and temporal separating distances h and t fading towards the product copula Π if h or t tend to infinity. This could for instance be realized as a convex combination of spatial bivariate copulas. For now, we consider only discrete points in time. Typically, this corresponds to the temporal resolution of measurements or aggregates thereof. Thus, the spatio-temporal bivariate copula can be defined as a set of spatial bivariate copulas indexed by the temporal gaps $1,\ldots,r_T$ investigated:

$$C_{h,t}(u_1,u_2) := \begin{cases} C_h^1(u_1,u_2) & ,t=1 \\ \vdots & \vdots \\ C_h^{r_T}(u_1,u_2) & ,t=r_T \end{cases}$$

2.2 Spatio-Temporal Vine Copulas

A *spatio-temporal vine copula* models a neighbourhood of a spatio-temporal random field of size $k+1$. This neighbourhood is composed of one central location and its k-neighbours in space and time. The first tree of the vine is realized by spatio-temporal bivariate copulas, reflecting the fact that the dependence structure changes over space and time. The remainder of the vine, i.e. the vine of the variables conditioned under the value of the central location, is modelled as some k-dimensional R-vine (a C-vine in our case). The structure of the first tree of a spatio-temporal vine copula is illustrated in Figure 2. Every curved connection is modelled by the same spatio-temporal copula $C_{h,t}$ but with different spatial and temporal distances h and t derived from the spatio-temporal locations involved. Combining the multivariate copula with the margins, which are assumed to be stationary, yields a full multivariate distribution of the neighbourhood of the spatio-temporal random field. This distribution can then be used to simulate, predict or analyse the observed phenomenon. Quantile predictions can be made for any fraction $p \in (0,1)$ through equation 2. The

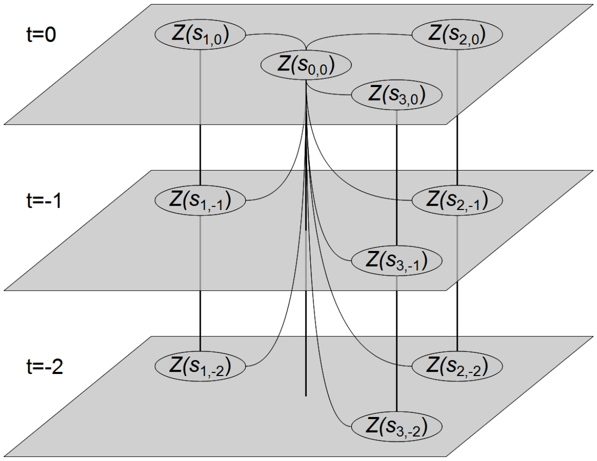

Fig. 2 The first tree of a spatio-temporal vine copula for a neighbourhood of size 10 with 3 neighbours in space and 3 instances in time (the same moment, one and two time instances before indicated with 0,-1 and -2 respectively). Every curved connection is modelled by the same spatio-temporal copula $C_{h,t}$ but with different spatial and temporal distances h and t derived from the indicated locations $s_{i,j}, i \in \{0,1,2,3\}, j \in \{0,-1,-2\}$. The remaining trees follow a 9-dimensional C-vine.

expected value of the conditioned distribution can be calculated by equation 3. The two predictors are given by

$$\widehat{Z}_p(s_0) = F^{-1}\big(C^{-1}(p|F\big(Z(s_1)\big),\ldots,F\big(Z(s_k)\big)\big) \tag{2}$$

$$\widehat{Z}_m(s_0) = \int\limits_{[0,1]} F^{-1}(u)\; c\big(u|F\big(Z(s_1)\big),\ldots,F\big(Z(s_k)\big)\big)\;\mathrm{d}u \tag{3}$$

where F denotes the cumulative distribution function of the stationary random field and s_1, ..., s_k are the spatio-temporal neighbours of s_0. Even though not explicitly stated, the copula C and its density c depend on the spatial and temporal distances between s_0 and its k-neighbours through the spatio-temporal bivariate copula in the first tree.

3 Application to daily PM$_{10}$ concentrations

The vine copula method is applied to a sample data set of daily mean PM_{10} measurements across Europe, using rural background stations from 2005. The data is publicly available through the AirBase[1] database hosted by the European Environmental Agency (EEA). Typically, the marginal distributions are unknown and have

[1] http://www.eea.europa.eu/themes/air/airbase

to be estimated as well. In order to not affect the copula by this estimation, we use rank-order transformed observations (i.e. $u_i := \text{rank}(\mathrm{x_i})/(n+1)$, where n is the length of the sample). To infer on the spatial dependencies, a set of 40 spatial lag classes is derived. Separate lag classes are filled with rank transformed data pairs from the same day, the one and two preceding days resulting in a set of 120 spatio-temporal lag classes. For every spatio-temporal lag, the best fitting copula from several copula families (elliptical, Archimedean, copulas with cubic-quadratic sections including an asymmetric one) is selected based on their log-likelihood values. These estimated copulas are then combined in a spatio-temporal bivariate copula as described in Section 2.1. As to be expected, the spatial bivariate copula fitted to rank transformed pairs measured the same day does not show any asymmetric dependencies. However, the asymmetric copula family is preferred over the other investigated families for some lags with pairs one day apart and dominates the convex linear combination for data pairs two days apart.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
$\Delta = 0$	t	F	t	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	Q	Q	Q	F
$\Delta = -1$	F	F	F	F	F	F	F	F	F	F	F	F	F	A	A	A	A	F	F	F	F	A	A	A	A	A
$\Delta = -2$	F	F	F	F	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Q	Q	Q	Q	Q

Table 1 The copula families with the highest log-likelihood values for the first 26 spatial lag classes corresponding to distances up to 1000 km and three time instances. Abbreviations are as follows: t =Student, F = Frank, A = asymmetric, Q = cubic-quadratic-sections. The vertical lines indicate 100 km, 250 km, 500 km breaks

To build multiple samples of the stationary neighbourhoods, the data was arranged as spatial neighbourhoods and a random sample of 90 days for every station was included in the following analysis to reduce unwanted autocorrelation effects. This data is grouped in spatio-temporal neighbourhoods building the basis of the 10-dimensional vine copula. The fitted spatio-temporal bivariate copula is used in the first tree (see Figure 1) to derive the 9 dimensional data set conditioned under the one central location $s_{0,0}$ (see Figure 2). The remaining trees consist of 36 bivariate copulas and are iteratively estimated based on their maximum log-likelihood values. To assess the quality of our fit, we calculated the overall log-likelihood and compared it against simpler approaches. The log-likelihood value of our spatio-temporal vine copula (72709) is about 35% larger than the fit of a Gaussian copula (53305) which included 45 covariance parameters. Furthermore, the Gaussian copula does not allow for asymmetric dependencies opposed to the vine copula including asymmetric copulas.

To extend the validation of the fit beyond the log-likelihoods, we perform a cross validation leaving the full time series of one station out after another and predicting the expected value (equation 3), median and 95% quantile (equation 2 for $p = 0.5$ and $p = 0.95$) for every day during the year based on the conditional distribution from the three spatial neighbours and their three temporal instances. Thus, the cross-validation relies purely on spatial and spatio-temporal dependencies. To fully estimate the desired indicators, the marginal distribution has to be fitted. The best fit is achieved for a generalized extreme value distribution $GEV(\mu, \sigma, \xi)$ with its parameters location, scale and shape set to $\mu = 13.94$, $\sigma = 8.54$ and $\xi = 0.20$ respectively (following the notation of the R-package `evd` [11]).

The full study is performed in the statistical computing environment and language `R` [9] using the package `spcopula` [4] that connects and extends the packages `spacetime` [8], `copula` [12, 6] and `CDVine` [10]. The `R`-scripts are available on request from the authors.

4 Discussion

In the following, we will compare the copula approach with a spatio-temporal interpolation procedure described in the recent ETC/ACM Technical Paper 2011/10 [3]. The method therein performing best, applies residual kriging assuming a metric spatio-temporal covariance model (where 1 day $\approx$ 120 km) following a log-transformation of the original measurements and detrending by a linear regression with altitude and daily EMEP model predictions (Further details can be found in [3].).

Table 2 Cross validation results for the expected value and median estimates following the vine copula approach and the best performing method in [3] for comparison.

	expec. value	median	metric cov. kriging
root mean sq. er.	11.2	12.08	9.84
bias	-0.73	1.94	-0.24
mean abs. er.	6.95	6.87	5.66

Cross validation indicators for the predictions based on the conditional expected value and the median following the vine copula approach are shown in Table 2 a long with the values of the best performing method from [3]. Based on these numbers, no improvement could be achieved. However, the errors of the estimates based on the conditional expected value are of the same order of magnitude. It has to be noted that the neighbourhoods of the metric covariance model relying on the 100 nearest neighbours in a metric spatio-temporal space differs from the one underlying the vine copula approach building on the three nearest neighbours in space and three instances in time. The overall reproduction of the data set by the copula interpolation is rather good. The predicted median and 95%-quantile are almost precisely exceeded by 50% and 5% of the original observations respectively. Looking into the predictions of single stations reveals cases where the copula approach outperforms kriging, but also versa. Two extreme scenarios are discussed in the following.

In Figure 3, a time series plot of a Finnish station roughly 600 km apart from any other station is shown. In our application, the copula prediction (drawn in magenta) is far above the real observations (drawn in red) while the metric covariance kriging prediction (drawn in green) seems to represent the mean process. An ex-

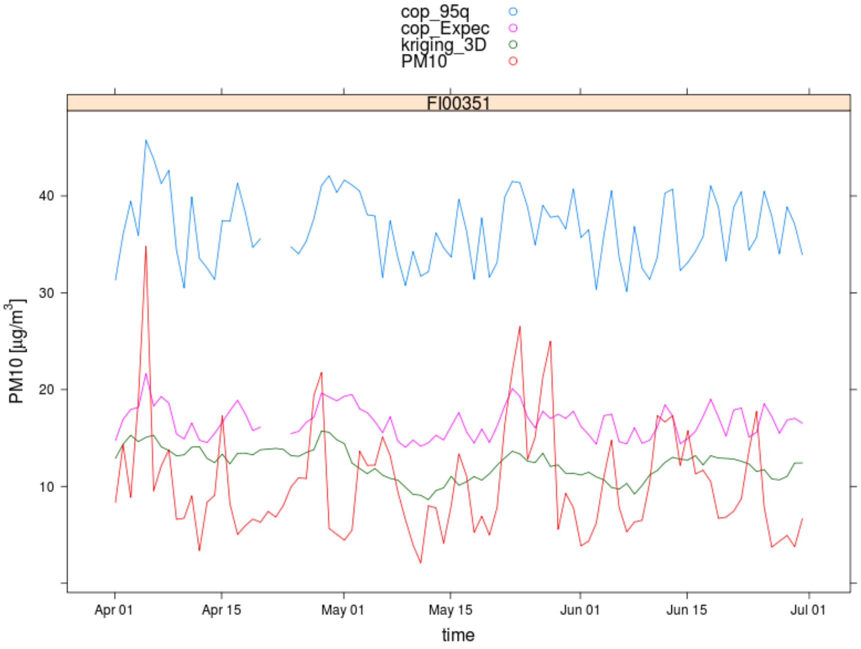

Fig. 3 A time series plot of a Finish station showing the 95%-quantile of the copula prediction (cop_95q, blue), the conditional expected value estimate (cop_Expec, magenta), the metric covariance kriging predictions (kriging_3D, green) and the original observed PM_{10} concentrations (PM10, red).

planation might be given by the fact that the metric kriging model relies on the nearest stations in time and space. In this specific neighbourhood and a temporal scaling of 1 day≈120 km, the nearest neighbours are roughly dominated by a factor of 10 through temporal instances. Thus, the predictions are similar to a temporal moving window average of a very few spatial neighbours. In the copula approach, we always rely on three nearest spatial neighbours and three instances in time. For larger distances, the conditioning influence of these neighbours is rather weak and the prediction value tends towards the expected value of the marginal distribution (21.0 $\mu g/m^3$ in our case) as the conditional density approximates a uniform distribution.

Another extreme case is shown in Figure 4. Here, the vine copula approach outperforms the 3D kriging approach for instance with respect to the station-wise root mean squared error with 4.0 $\mu g/m^3$ opposed to 5.4 $\mu g/m^3$. Even though both predictors follow the shape of the observations, the kriging estimate is often considerably above the observed concentrations. As this German measurement station is situated within a rather dense and dominating network, it closely follows the marginal distribution and the copula seems to well capture the spatio-temporal dependencies.

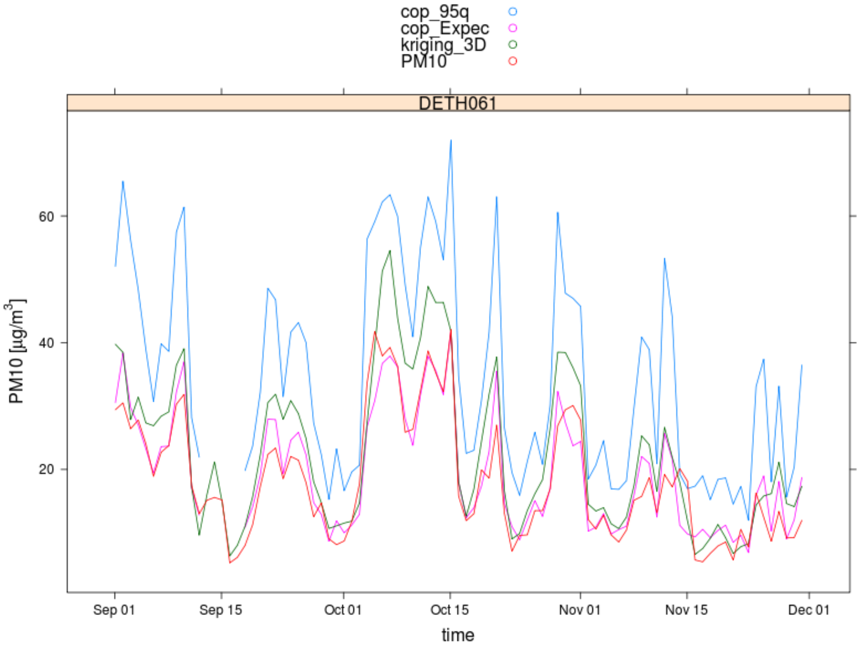

Fig. 4 A time series plot of a German station showing the 95%-quantile of the copula prediction (cop_95q, blue), the conditional expected value estimate (cop_Expec, magenta), the metric covariance kriging predictions (kriging_3D, green) and the original observed PM_{10} concentrations (PM10, red).

A potential advantage of the copula approach is the ease and flexibility with which one can predict quantiles of the distribution. In general, the conditional distributions at unobserved locations derived from the vine copula are not restricted to any specific distribution opposed to the kriging approach where every location is assumed to follow a Gaussian distribution. The blue lines in Figures 3 and 4 showing the 95%-quantile of the copula prediction are estimated with the same general equation 2 as the median. Deriving quantiles for the complete modelling approach in [3] including log-transformations and detrending would require a simulation procedure.

5 Conclusion & Outlook

This paper reports on a early stage attempt to model spatio-temporal dependencies with copulas, and exposes cases where predictions based on copulas are worse than following a residual kriging approach as well as where the flexibility of the copula's dependence structures seems beneficial. The fully probabilistic nature of copulas allows to predict different kinds of statistical values. Mean estimates can immediately be given alongside with quantile estimates. However, further research will be necessary to fully identify the strengths and weaknesses of the vine copula approach in comparison to kriging. The presented study includes effects of log-transformations and detrending (altitude, EMEP model predictions) in the metric covariance kirging procedure. Thus, the kriging approach relies on more information and their effect on the predictions is hard to assess.

As illustrated in Figure 3, the assumption that the PM_{10} concentrations across Europe follow the same distribution, i.e. that the random process has the same mean at any location, seem not very well supported by the data. Further work will be needed to address the issue of non-stationarity. Even though building higher dimensional distributions as in the presented approach remains a challenge, theoretical and software tools evolve to tackle these issues. Further developments will ease the estimation and application of spatial and spatio-temporal copulas.

References

1. Aas K, Czado C, Frigessi A, Bakken H (2009) The pair-copula constructions of multiple dependence. Insurance: Mathematics and economics 44:182–198
2. Bedford T., Cooke R. (2001) Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. Annals of Mathematics and Artificial Intelligence 32:245–268, doi: 10.1023/A:1016725902970
3. 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
4. Gräler B (2012) spcopula: copula driven spatial analysis. R package version 1.0.57. `http://r-forge.r-project.org/projects/spcopula/`
5. Hobæk Haff I, Aas K, Frigessi A (2010) On the simplified pair-copuila construction – Simply useful or too simplistic? Journal of Multivariate Analysis 101:1296–1310
6. Kojadinovic I, Yan J (2010) Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. Journal of Statistical Software, 34(9):1–20
7. Nelsen R (2006) An Introduction to Copulas, 2nd edition. In: Springer Series in Statistics. Springer, New York
8. Edzer Pebesma (2012) spacetime: classes and methods for spatio-temporal data. R package version 0.6-0. `http://CRAN.R-project.org/package=spacetime`
9. R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`
10. Schepsmeier U, Brechmann EC (2011) CDVine: Statistical inference of C- and D-vine copulas. R package version 1.1-5. `http://CRAN.R-project.org/package=CDVine`
11. Stephenson AG (2002) evd: Extreme Value Distributions. R News, 2(2). `http://CRAN.R-project.org/doc/Rnews/`
12. Yan J (2007) Enjoy the Joy of Copulas: With a Package copula. Journal of Statistical Software, 21(4):1–21