Copulas, a novel approach to model spatial and spatio-temporal dependence

Benedikt Gräler[1], Hannes Kazianka[2], Giovana Mira de Espindola[3]

[1]Institute for Geoinformatics, University of Münster, Germany
ben.graeler@uni-muenster.de

[2]Vienna University of Technology, Austria
hannes.kazianka@fam.tuwien.ac.at

[3]National Institute for Space Research (INPE), Brazil
giovana@dpi.inpe.br

Abstract. Copulas are a statistical concept which allows for a novel approach to model dependencies of spatial and spatio-temporal variables. They capture the dependence structure of a multivariate distribution over its whole range detached from its specific margins. In contrast to a single measure of association, this allows for varying strength of dependence throughout the multivariate distribution. Thus, copulas are capable of capturing many different (i.e. non-Gaussian) dependence structures and allow for asymmetric dependencies which can be found in many natural processes. We applied this approach to data from the deforestation survey of the Brazilian Amazon in order to capture the dependence of deforestation on a selection of variables.

1 Introduction

Most techniques in classical geostatistics rely on the assumption of an underlying Gaussian process. The dependence of two or more variables is then usually modelled by a covariance matrix which reduces the dependencies of pairs of margins to a single value. However, one can easily think of situations in which the dependence strength changes for different quantiles of a distribution. Imagine, for instance, Ozone measurements in a city. Moderate values are usually due to local conditions and will occur quite loosely across the city area whereas very high values typically depend on regional phenomena like very intense solar radiation or inversion. Reducing the dependence of two measurement stations to their covariance and assuming the process to be Gaussian introduces errors which might have severe effects.

Figure 1 illustrates three different dependence structures of bivariate samples which all have standard normal margins and exhibit a covariance of about 0.7. The concept of copulas is capable of capturing these non-Gaussian dependence structures and resolves the Gaussian assumption.

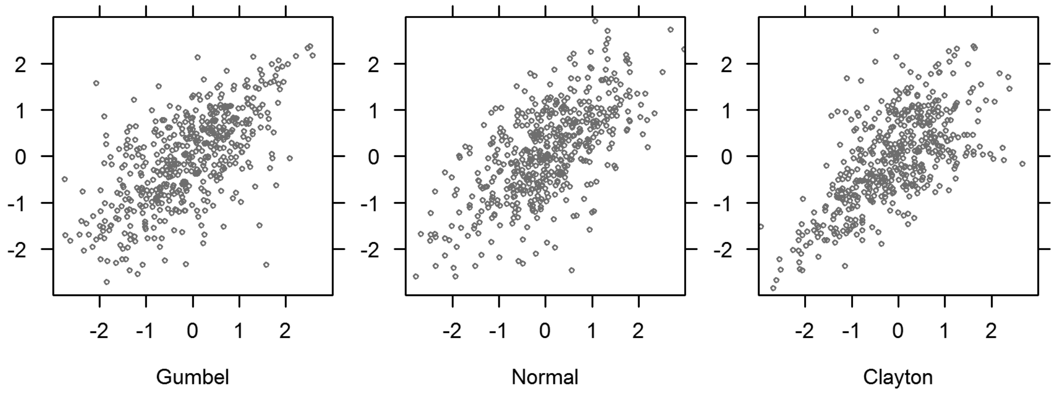

Figure 1: A comparison of different dependence structures with identical covariance (0.7) and standard normal distributed margins. In the *Gumbel* case the scatter plot shows higher dependence in the upper-right corner in contrast to the *Clayton* plot which gets tighter in the lower-left corner. The plot of the bivariate *normal* distribution exhibits the characteristic elliptical shape.

2 Copulas

The theory of copulas grounds on the theorem by Sklar (Nelsen 2006, Sec. 2.3) stating that any n-variate joint cumulative distribution function $H(x_1, ..., x_n)$ can be described by a copula C and the cumulative distribution functions of the margins $F_1, ..., F_n$ through

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)).$$

Thus, a copula captures the dependence structure detached from the margins and can be understood as cumulated distribution function in the unit hypercube.

The set of known two dimensional families of copulas is vast and includes well established distributions. The bivariate Student and Normal distributions, for instance, belong to the family of elliptical copulas. When copulas are estimated there is no need to estimate the univariate distributions of the margins beforehand. The estimation procedure can simply be based on the rank transformed observations of each variable because a copula is invariant under strictly increasing transformations of its margins. The copula densities corresponding to the scatter plots in Figure 1 are drawn in Figure 2. The densities can be understood as strength of dependence. It is clearly visible that the three plots considerably differ in range and shape.

The Gaussian assumption and the use of the covariance as a single measure of association have another drawback. The covariance and the Gaussian dependence structure are by definition symmetric in the sense that the variable X depends on Y in the same way as vice versa. Natural phenomena in contrast (typically including the temporal domain) do show asymmetric dependencies. For instance, monitoring data of the development of a forest over time will show asymmetries as trees are much faster cut down than grown. Typically, the exposure of toxics is as well an asymmetric process. The release of some toxic occurs usually all out of a sudden and reaches very fast high values in contrast to the decay that is often time consuming. Purely spatial data may exhibit asymmetric dependencies as well. These are often due to some spatial trend as prevailing air or water currents. Gneiting, Genton & Guttorp (2007) show that more realistic models, in particular non-symmetric covariance models, yield more reliable results. The class of asymmetric copulas is designed to capture these kinds of dependencies.

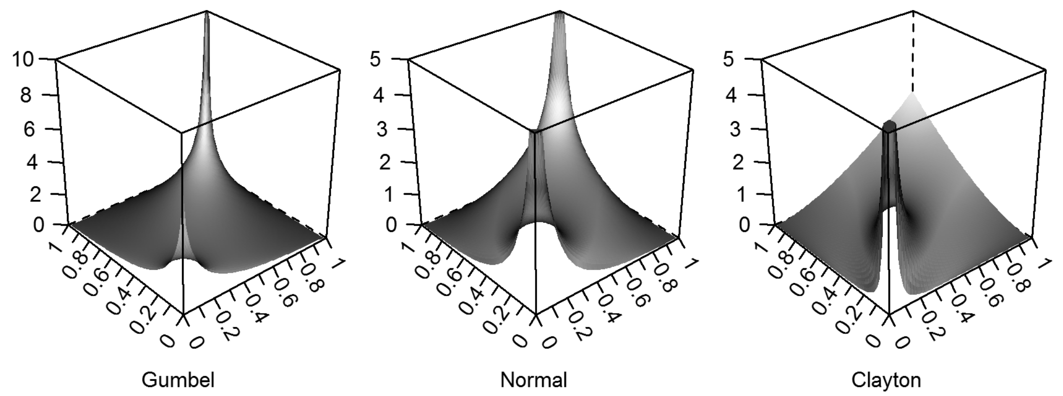

Figure 2: The densities of the underlying copulas of Figure 1.

Figure 3 illustrates a small pseudo sample and a scatter plot of the rank transformed variables with a temporal offset of one day. The distribution of points in the scatter plot is captured by a copula's density. The scatter plot of two rank transformed variables which exhibit a symmetric dependence would be symmetric with respect to the principal diagonal. In our case, the black and grey shapes on top of the plots relate the prominent asymmetric regions to each other. The black elliptical regions have very low probability to see any point where as their symmetric counterpart, the grey ellipses, enclose some points. The two tetragons oppose as well two symmetric counterparts which exhibit different frequencies. Thus, this bivariate sample would be best approximated by the asymmetric copula whose density is drawn as a filled contour plot on the right side of Figure 3.

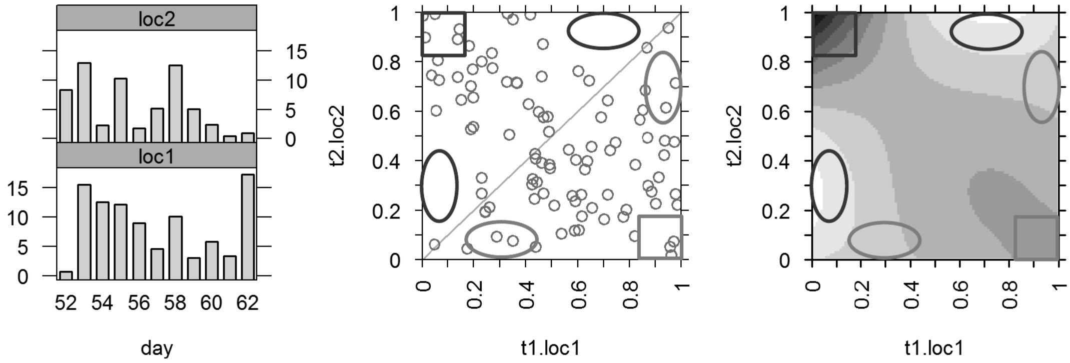

Figure 3: An eleven day subset of a spatio-temporal pseudo sample of length 100 (left) and its corresponding rank transformed scatter plot of pairs between location 1 and location 2 with a temporal offset of one day (middle). The density plot (right) describes the asymmetric copula which fits the data best.

A third issue that can be resolved when copulas are used to model data concerns the dependence of extreme values. The Gaussian dependence structure does not allow any tail dependence. That is, in the bivariate case, the conditional probability to observe an extremely small value in X under the assumption that Y is extremely small tends to zero as Y approaches its minimum:

$$\lim_{t \searrow 0} \mathbb{P}\left(X < F^{-}(t) | Y < G^{-}(t)\right)$$

where F^{-} and G^{-} are the pseudo inverse functions of the cumulative distribution functions of X and Y respectively. The above limit inspects the *lower tail dependence* and can be defined analogously for the *upper tail dependence* (Nelsen 2006, Sec. 5.4). This restriction may be too limiting for data related to natural catastrophes or other extreme events. In this case, it is advisable to use copulas that do exhibit tail dependence according to the data. Candidates are, for instance, copulas taken from the Gumbel or Clayton family (drawn in Figure 1 and Figure 2). Furthermore, the class of extreme value copulas is capable of capturing multivariate max-stable processes. Risk assessment and interpolation for data exhibiting extremes will highly benefit from a flexible copula instead of a steady Gaussian approach.

3 Copulas in Geostatistics

In the field of geostatistics we deal with data observed in time and space and are interested in modelling natural processes. One common application is interpolation of data. In this case, we desire the multivariate distribution that describes how one unobserved location depends on known values in its neighbourhood. The dependence structure of this distribution can be captured by a multivariate copula. Bárdossy & Li (2008) and Kazianka & Pilz (2010) successfully exploit copulas for interpolation of spatial data using a comparatively small set of copulas. A promising approach exploiting the simplicity of bivariate copulas and generating a huge and very flexible set of multivariate copulas is the pair-copula construction. Therein, the desired multivariate copula is decomposed into a set of bivariate copulas incorporating the conditional and unconditional distribution functions of the variables. This procedure was introduced by Aas, Czado, Frigessi & Bakken (2009) in the financial sector. However, adapting the pair-copula construction to the spatio-temporal case is not straightforward.

We applied the copula approach to data on the deforestation survey of the Brazilian Amazon in order to investigate underlying dependencies. The Amazon rainforest is a tropical moist broadleaf forest that represents about half of the Earth's remaining rainforest in the world. However, deforestation is a major issue and its underlying processes need to be understood. The dependence of deforestation on multiple parameters is complex and ranges from environmental to socio-economic factors. Their dependence structure changes for different quantiles and shows asymmetric non-Gaussian dependencies (Figure 4).

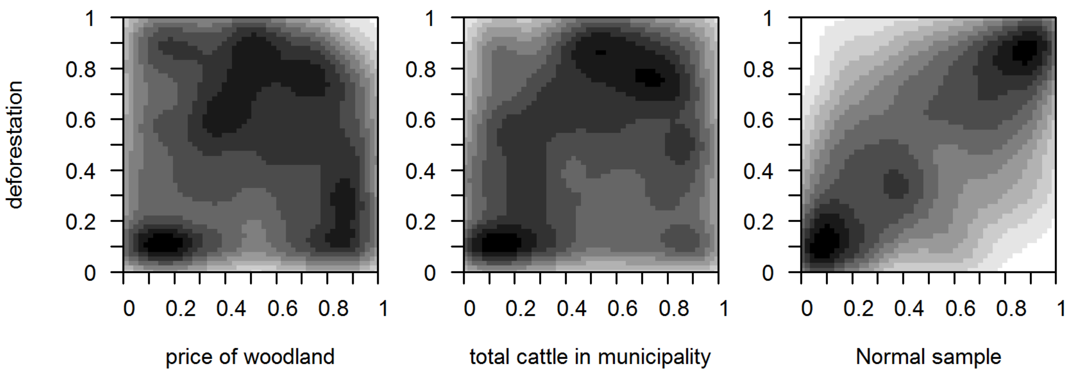

Figure 4: Three empirical copula densities (filled contour) illustrating the different dependence structures of *deforestation* on the *price of woodland* and the *number of total cattle in municipality* in comparison to a sample of a Normal copula.

Comparing filled contour plots of the empirical copula densities in Figure 4 makes the non-Gaussian structure obvious. The two left plots are based on data from the Amazon and the right plot on a sample drawn from a Normal copula. The typical assumption of an underlying Gaussian process restricts the phenomena to exhibit elliptical patterns of dependence as in the right plot. The two plots based on data from the Amazon show shapes following an arc running from the top-left to the bottom-right corner and from the bottom-left to the top-right corner respectively. Thus, the Gaussian model cannot capture these dependencies. Incorporating asymmetric copulas allows us to model these skewed shapes in a statistical set-up.

4 Conclusion

Copulas can be used in many fields of geostatistics and allow for new modelling approaches neglecting the common assumptions of symmetric dependencies and an underlying Gaussian process. Detaching the dependence structure from the margins allows us to easily incorporate many different variables regardless of their domain. These new approaches will especially be beneficial for skewed data or data exhibiting extremes. Furthermore, procedures based on copulas can be applied identically to purely spatial as well as spatio-temporal data. Developing spatio-temporal copulas based on the pair-copula construction will enable us to model multivariate data over space and time in a very flexible and sophisticated way. In the light of copulas it might be worth to revisit older scatter plots which did not show any reasonable correlation in order to apply a rank transformation to their margins and to inspect their copulas. This might reveal undiscovered but meaningful dependence structures of the underlying processes.

5 References

Aas, K., C. Czado, et al. (2009). Pair-copula constructions of multiple dependence, Insurance: Mathematics and Economics 44, 182 – 198.

Bárdossy, A. & J. Li (2008), Geostatistical interpolation using copulas, Water Resources Research 44.

Gneiting, T., M. G. Genton & P. Guttorp (2007), Geostatistical space-time models, stationarity, separability and full symmetry, in B.Finkenstaedt, L.Held & V.Isham, eds, `Statistics of Spatio-Temporal Systems', Chapman & Hall/CRC Press, 151 – 175.

Kazianka, H. & J. Pilz (2010), Copula-based geostatistical modeling of continuous and discrete data including covariates, Stochastic Environmental Research and Risk Assessment 24(5), 661 – 673.

Nelsen, R. B. (2006), An Introduction to Copulas, second edn, Springer Science+Business, New York.