

RELATÓRIOS COPPEAD

434

Fevereiro 2019

SPATIAL COPULA  
MODELING OF  
EXTREME CROP  
INSURANCE CLAIMS  
IN BRAZIL

Beatriz V. M. Mendes  
Eduardo F. L. de Melo

Relatórios COPPEAD é uma publicação do Instituto COPPEAD de Administração da Universidade Federal do Rio de Janeiro (UFRJ)

**Editora**

Leticia Casotti

**Editoração**

Lucilia Silva

**Ficha Catalográfica**

Anderson Luiz Cardoso Rodrigues

M538	Mendes, Beatriz V. M. Spatial copula modeling of extreme crop insurance claims in Brazil / Beatriz V. M. Mendes e Eduardo F. L. de Melo. -- Rio de Janeiro, 2019. 16 p.; 27 cm. -- (Relatórios COPPEAD; 434) ISBN: 978-85-7508-121-1 ISSN: 1518-3335 1. Desastre natural. 2. Seguro agrícola - Reclamação. 3. Modelo de valor extremo. 4. Estatística robusta.
------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

# Spatial copula modeling of extreme crop insurance claims in Brazil

Beatriz V. M. Mendes <sup>1a,b</sup>, Eduardo F. L. de Melo<sup>a</sup>

<sup>a</sup>COPPEAD Graduate School of Business, UFRJ, Brazil. <sup>b</sup>IM/UFRJ, Brazil

## Abstract

We use robustly estimated spatial R-vine copula models to assess spatial dependencies among extreme crop insurance claims. A truthful predictive model for simultaneous extreme losses is derived based on the linear structure found between copula parameters and distances between groups. Findings are compared to those from classical estimation of pair-copulas. Univariate fits of the excess-losses are based on the Generalized Pareto distribution. The dependence implied by the spatial component is captured by the Gumbel copulas in Tree 1, whereas a few atypical points are handled by robust inference which reveals that the influence of joint multivariate extreme outliers can not be neglected. Our findings are useful for crop insurance firms as well as for local authorities trying to minimize the effects of the natural disasters.

*Key words:* Natural disasters; Excess-claims; Insurance; Spatial pair-copulas; Robust statistics; Extreme value models.

## 1 Introduction

Relevant players in the Brazilian crop insurance industry frequently face huge losses originated from natural catastrophic disasters occurring in the agricultural south region of Brazil. Extremes are usually associated with loss of lives, property destruction, with remarkable effects on insurance markets, sometimes interrupting the continuity of business. The worst scenarios are expected when catastrophic events are correlated.

Typically, these extreme dependent losses are caused by the geographical concentration of policies sold and the intrinsic features of a physical phenomenon: a number of claims may arise at close locations due to a single natural disaster. The data in this study provide an example where a collection of important spatially correlated losses originated from hail were observed on two neighborhood coastal areas.

From the insurance standpoint is much more simple to deal with a large number of independent losses following some well known pattern than with a small number of correlated losses caused by some natural disaster. All the activities and numbers involved are magnified, including the amount of financial resources to cover the losses, the settlement of disputes under law, and so on. For example, the Hurricane Katrina, an extremely destructive storm hit the Gulf Coast of the US in 2005, and it was the costliest natural

---

<sup>1</sup>E-mail: beatriz@im.ufrj.br

disaster and one of the five deadliest hurricanes in the history of the US. *Total property damage was estimated at \$108 billion (2005 USD), roughly four times the damage wrought by Hurricane Andrew in 1992* (Source: Wikipedia). As a consequence, a major insurer, the Allstate, exited several coastal states.

The effects of the global warming and climate changes can be detected everywhere, in particular, on the crop insurance number and magnitude of claims. New records suggest that the extreme tails of claims joint distributions may be changing, as well as the strength of the (non-Gaussian) dependence structure. Therefore, in recent years, major crop insurance companies in Brazil have identified the need of more sophisticated models for correctly estimating the joint risks, a strategic tool for quantifying financial reserves (provisioning), for pricing insurance premiums, and also for designing local alert systems.

In their simplest version, actuarial models assume independence among claim sizes. Alternative (spatial) models usually assume elliptical distributions, in particular the multivariate Gaussian. However, this assumption implies that all margins are Gaussian and that the only possible relationship among them is given by the linear correlation coefficient. This is a serious restriction since we are ruling out the worst scenarios based on tail dependence between components.

In this study we investigate the spatial linear and non-linear (tail) dependence between extreme crop insurance claims occurring at several locations in the south region of Brazil (2005-2017) using pair-copula models. The major appealing characteristic of pair-copula models is their flexibility since the multivariate distribution is constructed based just on bivariate copulas. Moreover, the copula families and marginal distributions may vary freely (see, for instance, Aas et al. (2009) and Dißmann et al. (2013)).

To minimize the influence of atypical data points on a low volume of multivariate data we apply pair-copula robust estimation. We apply Weighted Maximum Likelihood method (WMLE) proposed initially for bivariate copulas in Mendes et al. (2007). The linear relationship between copula parameters and distances is assessed and used in a prediction model for future locations.

We focus on the long right tail of the claims distribution, location of rare events. This approach reduces the dimension of the relevant data and calls for univariate models with strong theoretical support (Pickands (1975), Resnick (1987), de Haan and Ferreira (2006)). Results from the Extreme Value Theory (EVT) demonstrate that the univariate models for the excess-claims beyond a high threshold should be based on the Generalized Pareto Distribution (GPD). We also consider the influence of covariates on the GPD parameters (Coles (2001), Embrechts et al. (1997)).

The EVT approach to model (univariate) extremes is very popular. Bermudez et al. (2009) carry on univariate analysis of spatial and temporal patterns of large fires in Portugal using the GPD. Born and Viscusi (2006) examine how catastrophic events affect the performance of the market of homeowners' insurance through multiple linear regressions.

Working on spatial multivariate extremes pose very special and interesting difficulties.

A comprehensive review on different spatio-temporal modeling approaches is found in Cressie and Wikle (2011). Gräler and Pebesma (2011) developed a spatial pair-copula based interpolation method, deriving a convex combination of copulas based on two limit copulas (perfect dependence and independence) for different distances.

Davison et al. (2012) review recent modeling strategies for spatial extremes, showing that copulas and spatial max-stable processes are the most successful models (with an application to a dataset on rainfall in Switzerland). Cooley et al. (2012) complement the Davison et al. (2012) study and survey the current methodologies for analyzing spatial extreme data, including copula approaches for modeling residual spatial dependence after marginal fits. They review the steps for combining EVT models (Generalized Extreme Value distribution) and copulas but does not use pair-copulas.

Erhardt et al. (2015) apply an extension of R-vine models to allow for spatial dependencies. Jane et al. (2016) method for estimating the significant wave height at a coastal location based upon spatial correlations combine EVT models and t-student and Gaussian copulas.

The joint modeling of (crop) insurance excess-claims using GPD models and pair-copulas estimated via the WMLE robust method are the novelties of this paper. At the best of our knowledge, this paper is the first to apply robust estimation of pair-copulas combined with the characterization of the tails of claims through GPD fits to investigate the relationships among crop insurance extreme-claims reported at several locations. Using the estimated spatial model we are able to obtain more accurate estimates of the (small) probabilities associated with joint extreme risks. The findings of this paper will contribute to local public sectors and insurance companies to design better local alert systems which may help insurers to predict and extend their models to close locations adapting their policies to these catastrophic risks. At the event of unexpected catastrophes this can prevent the exit of firms from the region. We recall that small size insurance firms are the most adversely affected by extreme events.

In Section 2 we introduce the data and briefly review the models and estimation methods used. In Section 3 we analyze the data and the results obtained through classical and robust estimation. Section 4 discusses the findings.

## 2 Data and methodologies

### 2.1 Data

The data are composed by crop insurance claims related to natural disasters occurred in the south region of Brazil from January/2005 through September/2017 and kindly provided by **provided by??** Figure 1 shows the study area ranging from latitude 22°S up to 32.56°S, and longitude 57°E to 48°E.

Cropland areas represent 31.1% of Brazil's total area. However, the vast majority of



Figure 1: Map of the area under study (south of Brazil).

cultivated lands are geographically located at the bottom-half of the country. The main products are sugarcane, coffee, corn, and soybeans. Permanent crops represent 2.7% of the agricultural land. Although possessing an important role in Brazil's economy and showing an important contribution to the GDP, Agriculture in Brazil still faces many problems and challenges.

Among these challenges we would highlight a better understanding of extreme climate effects on neighboring croplands in order to set systematic monitoring and warning policies, while gathering reliable relevant data set.

Information in the original data set include, among others, the claim value, two off-set variables, and event type recorded at non-regularly spaced dates. There are 906 cities reporting the claims, identified by their geographical coordinates, latitude ( $La$ ) and longitude ( $Lo$ ), spreaded over three states Paraná (PR), Santa Catarina (SC) and Rio Grande do Sul (RS). There are nine event types: fire, waterspout, strong or cold wind, gale, hail, heavy rain, drought, rime, flooding. The five largest claims came from hail or rime.

Due to the very large number of locations and the large sparsity of the data matrix, we use the variables  $La$  and  $Lo$  to make a partition of the 906 cities into a smaller number of groups using the k-means clustering technique which minimizes the within-cluster sum of squares. Then, for fixed event  $e$  and date  $t_i$ , we compute the aggregate process  $\{Y_{t_i}^{(e)}(G_k)\}$  for the location (Group)  $G_k$ ,  $k = 1, \dots, 6$ , as the sum of claim values of cities in the group. The decision for 6 groups was based on the rate of decrease of the within groups sum of squares. Each group  $k$  is geographically identified by the coordinates of its centroid, see

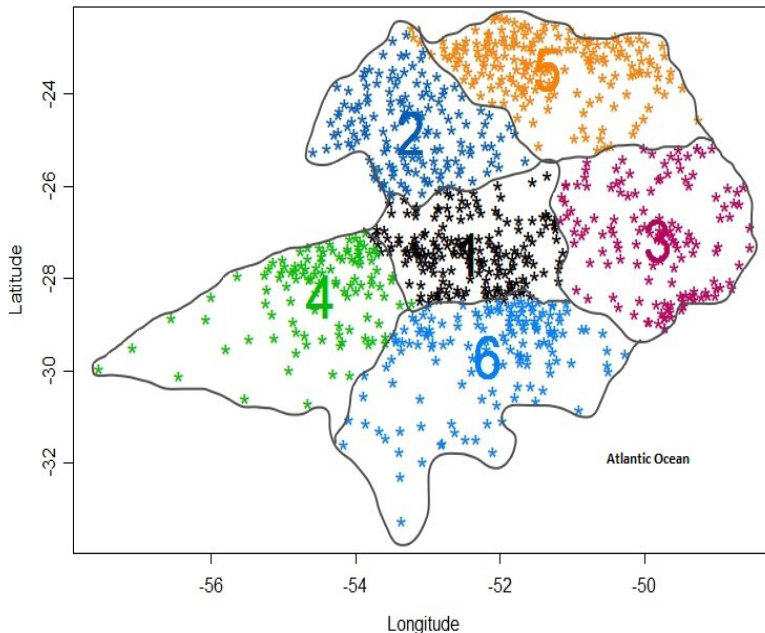


Figure 2: *The 6 groups.*

Figure 2. The four largest claims are due to rime (frost) and came from group 1, 4, and 5. The largest losses from the coast groups 3 and 6 are due to hail.

## 2.2 Methodologies

In summary, the multivariate dependence among the 6 groups resulting from natural catastrophic events will be modeled using spatial pair-copulas. Initially, we select a high threshold for the claim values at each location, estimate the marginal distributions of the excesses using results from the EVT, and obtain the copula multivariate data. The pair-copula model is estimated and inference at locations not present in the data will be assessed based on the spatial information.

More formally, let  $X_k$  with unconditional cumulative distribution function (cdf)  $F_k$  represent the claim size for group  $G_k$ ,  $k = 1, \dots, 6$ . Assume that  $F_k$  is in the max-domain of attraction (MDA) of one of the three extreme value distributions  $H_\xi$ , that is,  $F_k \in MDA(H_\xi)$ ,  $\xi \in \mathfrak{R}$  (de Haan, 1984). Since we focus on the tails of claims' distributions, consider a high threshold  $u_k$  in margin  $k$  and the excess claims  $Y_k = (X_k - u_k)\mathbf{I}_{(X_k > u_k)}$ , where  $\mathbf{I}_{(\cdot)}$  is the indicator function.

Pickands (1975) established the adequacy of the Generalized Pareto distribution (GPD) as the asymptotic distribution of positive excesses above a high threshold. The tail of the

conditional distribution of  $Y_k$ ,  $\bar{F}_{u_k}(y) = P(X_k - u_k > y | X_k > u_k)$ , may be modeled by the GPD. The result holds also for non-i.i.d. processes, see proof in Leadbetter et al. (1983).

The standard GPD distribution function  $P_\xi$  is given by

$$P_\xi(y) = \begin{cases} 1 - (1 + \xi y)^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - e^{-y}, & \text{if } \xi = 0 \end{cases} \quad (1)$$

where  $y \geq 0$  if  $\xi \geq 0$ , and  $0 \leq y \leq -1/\xi$  if  $\xi < 0$ . The scale family is obtained by introducing the scale parameter  $\psi$ , which depends on the threshold. For most applications in Actuarial Science the shape parameter is positive.

Focusing on the tail of the claim size distribution allows one to accurately estimate claim values associated with very small probabilities of occurrence, or much more precisely estimate the probability of occurrence of an extreme claim value. This follows from

$$\bar{F}_{u_k}(y) = P(X_k - u_k > y | X_k > u_k) = \frac{P(X_k > y + u_k)}{P(X_k > u_k)}. \quad (2)$$

For instance, let  $\alpha = P(X_k > y + u_k) = \bar{F}_k(y + u_k)$  be a very small exceedance probability, and let the observed proportion  $p^*$  of data above  $u_k$  be an estimate of  $\bar{F}_k(u_k)$ . Then, for  $\alpha < p^*$  the extreme claim size  $(y + u_k)$  with risk  $\alpha$  is estimated using

$$\alpha = \bar{F}_{u_k}(y) * p^*. \quad (3)$$

It is clear from (3) that more accurate estimates are obtained from the GPD fit,  $\bar{F}_{u_k}(y)$ . For details see Embrechts, Klüppelberg and Mikosch (1997).

The GPD is the only continuous POT-stable distribution. The shape parameter  $\xi$  does not depend on the threshold, a property of great interest in Actuarial Science. There is a trade-off when choosing the threshold, which should be as high as possible to follow the theoretical assumptions and at the same time low enough to provide enough data for estimation. Several graphical and analytical proposals for choosing  $u_k$  are available in the literature (Pickands (1994), Smith (1987), among many others). Here we use a high percentile of the series, chosen such that the empirical distribution of the excesses indicate the strictly decreasing shape of the GPD, along with goodness of fit tests.

Estimation usually applies either the maximum likelihood method or the probability weighted moments (Hosking and Wallis, 1987). As suggested in Coles (2001), covariates may be incorporated in the GPD model.

We model the spatial dependence between the different locations  $G_k$  using pair-copula models. Copulas have been widely used in Actuarial Sciences, and one of the reasons for this popularity is the inadequacy of the multivariate normal distribution for the vast majority of data sets.

We briefly review the definitions, and to simplify the notation, assume  $k = 2$ . Let  $(Y_1, Y_2)$  be a continuous random variable (rv) in  $\mathfrak{R}^2$  with joint cdf  $H$  and margins  $G_k$ ,



$k = 1, 2$ . Consider the probability integral transformation of  $Y_1$  and  $Y_2$  to uniformly distributed rvs on  $[0, 1]$ , that is,  $(U_1, U_2) = (G_1(Y_1), G_2(Y_2))$ . The copula  $C$  corresponding to  $H$  is the joint cdf of  $(U_1, U_2)$  (Nelsen (2006), Joe (1997)). As multivariate distributions with Uniform  $[0, 1]$  margins, copulas provide very convenient models for studying dependence structure with tools that are scale-free.

From Sklar's theorem (Sklar, 1959) we know that for continuous rvs, there exists a unique 2-dimensional copula  $C$  such that for all  $(y_1, y_2) \in [-\infty, \infty]^2$ ,

$$H(y_1, y_2) = C(G_1(y_1), G_2(y_2)). \quad (4)$$

The Sklar theorem allows one to construct multivariate distributions by simply choosing a copula family and marginal distributions.

To measure monotone dependence, one may use the population version of the measure of association known as Kendall's  $\tau$ . Kendall's  $\tau$  does not depend upon the marginal distributions and is given in terms of the copula. However, the (Pearson) product-moment (linear) correlation coefficient  $\rho$  is not a copula based measure.

To measure upper tail dependence one may use the upper tail dependence coefficient  $\lambda_U$  which also may be expressed using the copula corresponding to  $H$ :

$$\lambda_U = \lim_{u \uparrow 1} \frac{\bar{C}(u, u)}{1 - u}, \text{ where } \bar{C}(u_1, u_2) = P\{U_1 > u_1, U_2 > u_2\}, \text{ and } \lambda_L = \lim_{u \downarrow 0} \frac{C(u, u)}{u}, \quad (5)$$

if these limits exist. The measures  $\lambda_U \in (0, 1]$  (or  $\lambda_L \in (0, 1]$ ) quantify the amount of extremal dependence within the class of asymptotically dependent distributions. If  $\lambda_U = 0$  ( $\lambda_L = 0$ ) the two variables  $Y_1$  and  $Y_2$  are said to be asymptotically independent in the upper (lower) tail.

More flexibility may be gained by considering pair-copulas models. Pair-copulas is an hierarquichal decomposition of a  $d$ -dimensional copula into a cascade of  $d(d - 1)/2$  potentially different bivariate copulas. It was originally proposed by Joe (1997), and later discussed in detail by Bedford and Cooke (2001, 2002), Kurowicka and Cooke (2006) and Aas, Czado, Frigessi, and Bakken (2007). The method of construction is hierarchical, and variables are sequentially incorporated into the conditioning sets as one moves from level 1 (Tree 1, denoted by  $\mathcal{T}_1$ ) to tree  $d - 1$ . The composing bivariate copulas may vary freely, with respect to choice of the parametric families and parameter values. Therefore, all types and strengths of pair-wise dependence can be captured.

Consider the  $d$ -dimensional joint distribution cdf  $H$  and density  $h$  of the co-excesses with strictly continuous marginal cdf's  $G_1, \dots, G_d$  with densities  $g_k$ . The multivariate density function may be uniquely decomposed as

$$h(y_1, \dots, y_d) = g_d(y_d) \cdot g(y_{d-1}|y_d) \cdot g(y_{d-2}|y_{d-1}, y_d) \cdots g(y_1|y_2, \dots, y_d). \quad (6)$$

The conditional densities in Equation (6) may be written as functions of the corresponding copula densities. That is, for every  $j$

$$g(y | v_1, v_2, \dots, v_d) = c_{y v_j | \mathbf{v}_{-j}}(G(y | \mathbf{v}_{-j}), G(v_j | \mathbf{v}_{-j})) \cdot g(y | \mathbf{v}_{-j}), \quad (7)$$

where  $\mathbf{v}_{-j}$  denotes the  $d$ -dimensional vector  $\mathbf{v}$  excluding the  $j$ th component. Note that  $c_{y_j|\mathbf{v}_{-j}}(\cdot, \cdot)$  is a *bivariate* marginal copula density.

Decomposition (6) together with (7) was described in Czado:2010 and in Aas et al. (2009). Expressing all conditional densities in Equation (6) by means of Equation (7), we derive a decomposition for  $h(y_1, \dots, y_d)$  that consists of only univariate marginal distributions and bivariate copulas. Then, a factorization of the  $d$ -dimensional copula density  $c(G_1(x_1), \dots, G_d(x_d))$  is obtained based only in bivariate copulas, the *pair-copula decomposition*. This is a very flexible and natural way of constructing a higher dimensional copula. Note that, given a specific factorization, there are many possible reparametrizations.

For large  $d$ , the number of possible pair-copula constructions is very large. Bedford and Cooke (2001) introduced a systematic way to obtain the decompositions, the so called regular vines (R-vines). These graphical models help understanding the conditional specifications made for the joint distribution. Two special cases are the canonical vines (C-vines) and the D-vines. C-vines (D-vines) possess star (path) structures in their tree sequence.

Given a parametric copula family, to estimate the copula parameter  $\delta$  (it may be a vector), in this paper we use the sequential approach proposed in Aas et al. (2009) and applied in AasBerg:2009 in which the estimates from the previous tree are used to transform the data in the current tree. Precise recursions for sequential estimation in C-vine models were given in Czado et al. (2012). The MLE is the classical estimation method.

Bayesian methods have also been applied to pair-copulas. In Dalla Valle (2009), Bayesian inference based on MCMC is proposed for multivariate elliptical copulas using the inverse Wishart distribution as the prior distribution for the correlation matrix. Min and Czado (2010) also developed a Markov chain Monte Carlo algorithm that provides credibility intervals. Turkman et al. (2010) model Portuguese wildfires using Bayesian hierarchical models. In Czado et al. (2012), a very interesting data driven sequential selection procedure is proposed or jointly choosing the C-vine structure and the pair-copula families. A sequential estimation procedure for copula parameters in a previously specified C-vine was also developed and implemented.

However, occasional atypical points may be present and they may corrupt the classical estimates of the dependence structure. In this paper we apply the Weighted Minimum Distance (WMDE) and the Weighted Maximum Likelihood (WMLE) robust estimates proposed in Mendes, Nelsen and Melo (2009). They are based on either a redescending weight function or on a hard rejection rule applied to one or several outliers occurring anywhere in the data.

The WMLE result from a two-step procedure. In the first step, outlying data points are identified by a robust covariance estimator and receive zero weights, and in the second step the copula MLE are computed for the reduced data. In the first step we are not concerned with efficiency. The goal is to identify outlying points by computing the Mahalanobis distance with as a cutoff point the 0.975 quantile of a chi-square distribution with one degree of freedom. There are many high breakdown point estimators of multivariate location and

scatter that could be used in this preliminary phase. We use the robust Stahel-Donoho (SD) estimator based on projections (Stahel, 1981 and Donoho, 1982) which is implemented in the free  $R$  software. For every copula family there is a specific weighted minimum distance estimator able to downweight the influence of contaminating points which does not depend on the sample size. For details about the robust estimates see also Mendes and Accioly (2011).

The WMDE minimize some selected goodness of fit statistics. Copula measures of goodness of fit may be obtained by computing some *distance* between the empirical copula  $\tilde{C}$  and the parametric copula  $\hat{C} = C_{\hat{\delta}}$  fitted to the data. The WMDE estimate for  $\delta$  is the solution  $\delta^*$  which minimizes over all  $\delta$  in  $\Delta$ , the selected empirical copula based goodness of fit statistics.

Many discrete norms may be defined. For each copula family there is a WMDE as good as the MLE *under no contamination* as measured by by the mean squared error. In summary, under no contamination the WMDE and the MLE are equivalent. Under contaminations best solutions are provided by the WMLE and the WMDE. In this paper we compute and compare the classical and the robust estimates of the pair-copula models applied to the extreme crop claims.

## 3 Empirical analyses

### 3.1 Univariate excess-data

We start by taking a look at the behavior of the extreme claims at each location. For each group  $k$  we collect the excess-data based on threshold values selected as a high percentile (94%) of the series of size 5731, resulting in 344 observations in each margin. In Table 1 we provide some descriptive statistics of the excess-data: their minimum, medium, and maximum values along with the group center coordinates  $La(k)$  and  $Lo(k)$  and threshold values. All series showed a long right tail characteristic of the Pareto distribution with positive shape parameter. All excess data came from the natural events hail, heavy rain, drought, and rime.

After checking for and not finding any short, long, or seasonal serial dependence in the excess-data we proceed fitting by maximum likelihood the GPD model. There exist many exogenous risk factors besides climate and geographical variables, that may affect the outcomes of crop insurance claims' size and number. To eliminate the impact of the variables not included in the model but influencing the results, the actuarial modelers usually make use of some measure of exposure, an offset variable. The exposure might be the number of people who contract the insurer, insured area, value insured, and so on. Here we allow both the the scale parameter  $\psi$  and the shape parameter  $\xi$  to depend on a covariate (insured value) correcting for exposure through a linear model.

All maximum likelihood estimates are statistically significant and a goodness of fit

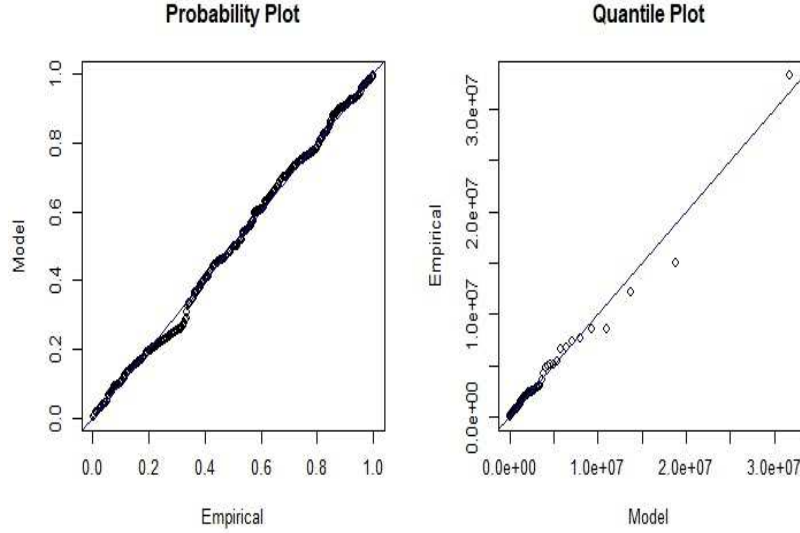


Figure 3: *PP-plot and QQ-plot from the GPD fit for Group 1.*

test accepted the null for all groups. Figure 3 shows the pp-plot and qq-plot for Group 1. Estimates(standard errors) of the shape parameter are 0.7937(0.0806), 0.6616(0.0756), 0.8893(0.0903), 0.5764(0.0708), 0.7030(0.0788), and 0.9064(0.0903), respectively for groups 1,  $\dots$ , 6. All groups have finite mean and infinite variance.

### 3.2 Exploring close neighborhoods

The proportion of common exceedances (denoted by  $\pi$ ) between two groups is an interesting empirical measure of dependence. Here, for two independent groups this proportion would be estimated as  $(344/5731)^2 = 0.0036$ , but we found  $\pi$  greater than that for all pairs. The smallest value, 0.0037, was observed for the pair composed by the up-country group 5 and

Table 1: Descriptive statistics of the excess-data: their minimum, medium, and maximum values along with the group center coordinates  $La(k)$  and  $Lo(k)$  and threshold values.

Group	Latitude	Longitude	Threshold	Excess claims		
				Minimum	Medium	Maximum
Group 1	-27.64692	-52.43124	59181	51	97226	23871328
Group 2	-25.01310	-53.23867	209712	551	259651	14031941
Group 3	-27.35435	-49.82260	155416	131	280601	20591557
Group 4	-28.54984	-54.48945	129409	241	205916	28818966
Group 5	-23.64538	-51.33528	220994	798	265568	33783266
Group 6	-29.87321	-52.16852	108980	27	153933	20664311

the south-coast group 6, and the largest one, 0.0234, came from group 5 and its closest neighbor country group 2 (see Figure 2).

This measure  $\pi$  may be used to define the ordering of the unconditional copulas composing the  $\mathcal{T}_1$  in a D-vine. The groups' ordering with the highest numbers of common exceedances is: 5-2-1-4-6-3. As we will see in the next subsection, this ordering agrees with the one suggested by the Kendall's monotone correlation coefficient  $\tau$  computed on the 6-dimensional space of the pair-copula.

It is worth investigating whether or not there is a relationship between  $\pi$  and the distance  $\mathcal{D}$ , the usual geostatistical approach for spatial modeling. A least squares fit of a linear model having  $\pi$  as explanatory variable and  $\mathcal{D}$  a response resulted in estimates 1% statistically significant and a  $R^2$  of 70% (negative slope). On the left hand side of Figure 4 we observe that the two groups 5 (up-country) and 6 (south-coastal) providing the smallest  $\pi$  also provided the largest  $\mathcal{D}$ .

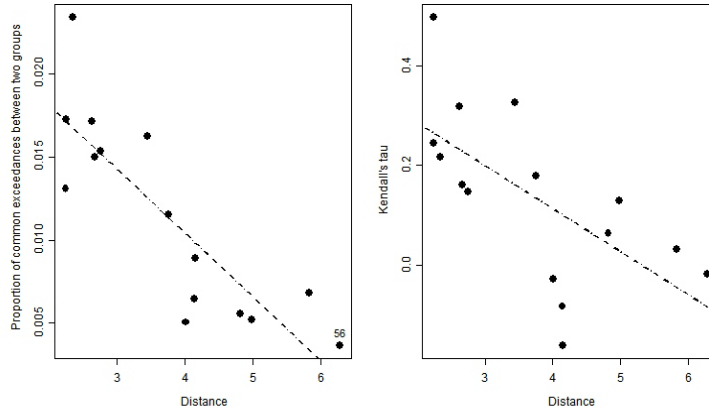


Figure 4: *On the left hand side the observed relation between groups distances  $\mathcal{D}$  and proportions of pair-wise common exceedances  $\pi$  along with the least squares fit. On the right hand side the observed relation between groups distances  $\mathcal{D}$  and the correlation coefficient  $\tau$ .*

### 3.3 Spatial pair-copulas

Once the marginal effects have been accounted for and the univariate series have been transformed to be Uniform[0, 1] through the probability integral transform applied to the GPD fit, we model the spatial dependence between the different locations  $G_k$  using pair-copula models. As stressed in several papers the success of the copula approach and of

course of the predictions made relies on the correct specification of the marginal models. Goodness of fit tests were carried on and confirmed the good quality of the GPD fits.

Several regular vines may be fitted. The D-vine modeling starts with the specification of the ordering of groups in  $\mathcal{T}_1$ . The ordering defined by  $\tau$  is  $5 - 2 - 4 - 1 - 6 - 3$ , and for this data set it coincides with the one implied by the linear correlation coefficient  $\rho$ . Figure 5 shows the support set of the empirical copulas in  $\mathcal{T}_1$ . Table 2 provides the  $\tau$  values.

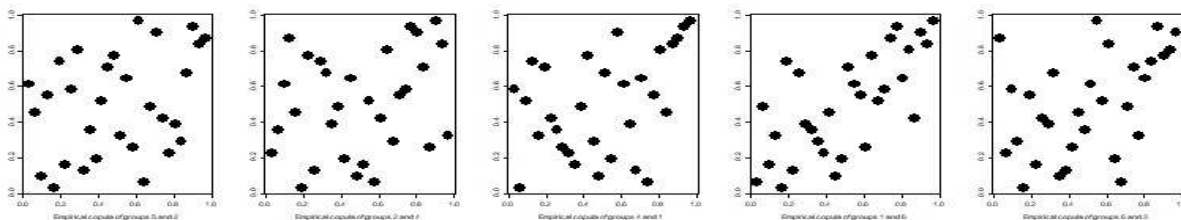


Figure 5: *The support set of the empirical copulas in Tree 1.*

The bivariate copulas chosen to compose the D-vine are: Gaussian, T-student, Gumbel, Clayton, Galambos, bb7, Frank, product and Tawn. They cover all types of extremal dependence and include the asymmetric case. The Akaike criterion is used to choose the best copula fit at each bivariate building block. The maximum likelihood estimates based on the sequential approach were computed using the SPlus and the R packages. Our final model is a D-vine, although the R package also found a R-vine specification as good as the D-vine chosen.

All five copulas in  $\mathcal{T}_1$  are Gumbel, an extreme value copula. The Gaussian and Frank copulas compose the following trees. Upper tail dependence is only observed in  $\mathcal{T}_1$ , with weaker (non-tail) dependencies captured on the remaining trees which showed correlations varying between  $-0.17$  and  $+0.12$ .

Classical estimation provided a total log-likelihood of 30.3293, whereas the robust WMLE estimation provided a larger value (33.5963) even though a smaller data set was used (two outliers were detected by the robust procedure). Stronger dependence was captured by the robust estimation resulting in larger tail dependence coefficients ( $\lambda_U$ ). Table 2 also shows the classical and robust estimates (standard errors) of the copulas' parameters in  $\mathcal{T}_1$  along with the corresponding  $\lambda_U$ . Although the GOF test accepted the null, we observe large standard errors, probably explained by the very small data set (only 30 6-dimensional joint observations).

As expected, the inspection of the support set of the empirical copulas for all fifteen pairs showed stronger dependence between locations which are close to each other. A linear regression between  $\tau$  and the distances  $\mathcal{D}$  provided both estimates statistically significant (less than 1%), an adjusted  $R^2$  of 38%, and a negative slope coefficient (-0.0865). For example, groups 5 and 6 showed negative dependence (-0.0184) and the highest distance (6.2833). The right hand side of Figure 4 shows the linear relationship between  $\tau$  and  $\mathcal{D}$ .

The functional relation between  $\tau$  and  $\mathcal{D}$  suggests that it is worth to assess the spatial dependence between groups by exploring the relationship between the copula parameters and the  $\log(\mathcal{D})$ . Recall that for each copula family there is a specific functional relationship between the copula parameter  $\delta$  and  $\tau$  and also  $\lambda_U$ . In the case of the Gumbel copula we have

$$\lambda_U = g_U(\delta) = 2 - 2^{1/\delta} \quad \text{and} \quad \tau = g_\tau(\delta) = \frac{\delta - 1}{\delta}.$$

Figure 6 shows the scatter plot of log-distances (and distances) and the  $\delta$  robust estimates for  $\mathcal{T}_1$  along with the robust line ( $MM$ -estimator). As expected, a negative slope. This structure may be used for predicting crop claims at unobserved locations (or missing data) based on the spatial variable:

$$\delta_{ij} = \beta_0 + \beta_1 \log(\mathcal{D}_{ij}) + \epsilon \tag{8}$$

for all pairs of groups  $i, j$  in  $\mathcal{T}_1$ . This spatial model may also be used to predict the strength of tail dependence  $\lambda_U$  as well as the correlation  $\tau$  between a new group and any other one from the original 6 groups. Whenever a new group is defined and data are not available, it should be allocated as a new leaf in  $\mathcal{T}_1$  joined to the group providing the smallest  $\mathcal{D}$  value, and the model (8) is used to predict  $\delta$ . Erhardt et al. (2015) used a similar but more complex relationship to reduce the number of parameters of a regular vine.

We optimize the full likelihood of the R-vine using as starting values those provided by the robust regression ( $MM$ -estimates). The final solutions and (standard errors) are  $(\beta_0, \beta_1) = (1.78(0.0356), -0.37(0.0307))$ .

In summary, the results show that the association between crop insurance claims from groups spatially spreaded over the south region of Brazil might be modeled by a spatial

Table 2: Some results for  $\mathcal{T}_1$ : The pairs of groups, the distance between them, and corresponding  $\tau$ . The classical and robust estimates (standard errors), and  $\lambda_U$ .

			Classical Fit			Robust Fit		
Pair	Distance	$\tau$	Copula	Estimate	$\lambda_U$	Copula	Estimate	$\lambda_U$
52	2.34	0.22	Gumbel	1.34(0.17)	0.32	Gumbel	1.39(0.19)	0.35
24	3.75	0.18	Gumbel	1.31(0.16)	0.30	Gumbel	1.30(0.17)	0.29
41	2.25	0.24	Gumbel	1.60(0.24)	0.46	Gumbel	1.61(0.25)	0.46
16	2.24	0.50	Gumbel	2.16(0.32)	0.62	Gumbel	2.22(0.34)	0.63
63	3.44	0.33	Gumbel	1.48(0.21)	0.40	Gumbel	1.63(0.24)	0.47

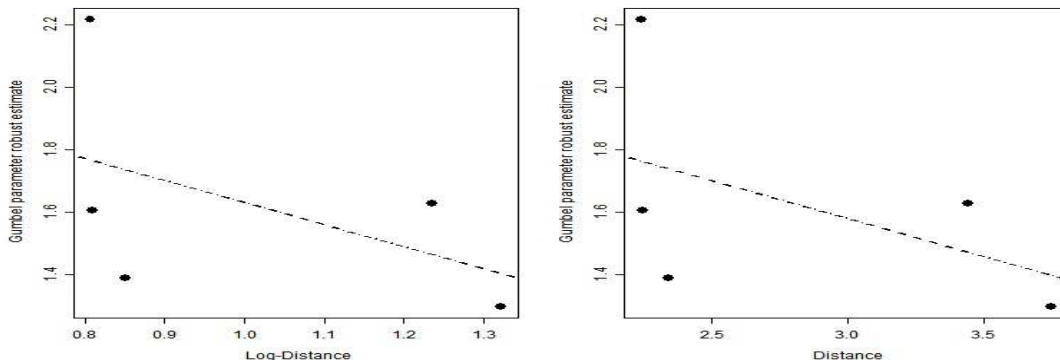


Figure 6: Robust line representing the association between (log)-distances and Gumbel- $\delta$  robust estimates for Tree 1.

R-vine model where all unconditional copulas are Gumbel with parameter  $\delta$  following a linear model having the log-distance as regressor.

The model may be simulated to estimate with accuracy any quantity of interest such as extreme joint claim sizes associated with very small exceedances probabilities  $\alpha$  of occurrence. To illustrate, consider the 0.95 quantile of the claim size (in each margin) which is computed as the 0.1667 quantile of the GPD distribution (see 3), since the threshold represents 6% of data. Under independence the joint probability of the six groups jointly exceed these claim values is  $0.05^6 = 0.00000015625$ . This joint probability estimated via simulations of the spatial pair-copula model is 0.0252!

## 4 Discussions

Insurance companies are usually prepared for huge losses. Even though, whenever a collection of important spatially correlated losses arise from close locations, there is a chance of a complete or partial interruption of the continuity of their businesses. Natural events frequently give rise to dependent losses. Note that a single natural catastrophic event may result in huge correlated losses.

The global warming has exacerbated all numbers related to extreme events: strength, speed, coverage, duration, volume, frequency, and so on. This calls for more sophisticated models for correctly estimating the joint risks and for quantifying financial reserves, pricing insurance premiums, and also for designing local alert systems.

The data analyzed in this paper are losses caused by natural extreme events. We modeled the spatial linear and non-linear dependence between extreme crop insurance claims occurring at several locations in the south region of Brazil using a spatial D-vine



model where all unconditional copulas in tree 1 are Gumbel. The strenght of dependence between two groups may be measured by the value of the parameter of the corresponding Gumbel copula. We found a linear relationship between the copula parameter and the geographical distance between two locations. This structure allows for predicting the degree of dependence between the original groups and any new one and estimate the dependence coefficients.

Results from the classical estimation was compared to the robust ones. Stronger dependence was captured by the robust estimates resulting in larger tail dependence coefficients. The model was simulated to estimate with accuracy quantities of interest such as extreme joint claim sizes associated with very small exceedances probabilities of occurrence.

It is very difficult to get assess to insurance data. It requires confidentiality, and usually just part of the data is released. In a further extension when more data are available we intend to separate the claims according to the event type and proceed with the multivariate spatial analysis of the dependence structure.

## References

- K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2009.
- J. Dißmann, E. C. Brechmann, C. Czado, and D. Kurowickaand. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, (59):52–69, 2013.
- B. V. M. Mendes, E. F. L. Melo, and R. B. Nelsen. Robust fits for copula models. *Communication in Statistics*, 36:997–1017, 2007.
- J. III Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, (3):119–131, 1975.
- S. I. Resnick. *Extreme values, regular variation, and point processes*. Springer, New York, 1987.
- L. de Haan, Laurens and A. F. Ferreira. *Extreme Value Theory, An Introduction*. Springer-Verlag, New York, 2006.
- S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Science & Business Media, 2001.
- P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer-Verlag, Berlin, 1997.

- P.Z. Bermudez, J. Mendes, J.M.C. Pereira, K.F. Turkman, and M.J.P. Vasconcelos. Spatial and temporal extremes of wildfire sizes in Portugal. *International J. of Wildland Fire*, 18: 983–991, 2009.
- P. Born and W. K. Viscusi. The catastrophic effects of natural disasters on insurance markets. *Journal of Risk and Uncertainty*, (33):55–72, 2006.
- N. Cressie and C. K. Wikle. *Extreme Value Theory, An Introduction*. Wiley & Sons, 2011.
- B. Gräler and E. Pebesma. The pair-copula construction for spatial data: a new approach to model spatial dependency. *Procedia Environmental Sciences*, 7:206–211, 2011.
- A.C. Davison, S. Padoan, and M. Ribatet. Statistical modeling of spatial extremes (with discussion). *Stat. Science*, 27(2):161–186, 2012.
- D. Cooley, J. Cisewski, R.J. Erhardt, S. Jeon, E. Mannshardt, B.O. Omolo, and Y. Sun. A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *REVSTAT – Statistical Journal*, 10(1):135–165, 2012.
- T. M. Erhardt, C. Czado, and U. Schepsmeier. R-vine models for spatial time series with an application to daily mean temperature. *Biometrics*, 71(2):323–332, 2015.
- R. Jane, L. Dalla Valle, D. . Simmonds, and A. Raby. A copula-based approach for the estimation of wave height records through spatial correlation. *Coastal Engineering*, (117):1–18, 2016.
- C. Czado, U. Schepsmeier, and A. Min. Maximum likelihood estimation of mixed c-vines with application to exchange rates. *Statistical Modelling*, 3(12):229–255, 2012.
- L. Dalla Valle. Bayesian copulae distributions with application to operational risk management. *Methodology and Computing in Applied Probability*, 11(95–115), 2009.
- A. Min and C. Czado. Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8(4):511–546, 2010.
- K.F. Turkman, M.A.A. Turkman, and J.M. Pereira. Asymptotic models and inference for extremes of spatio-temporal data. *Extremes*, 13:375–397, 2010.

