Chapter 1
Copulas as a tool

Seminar *GeoChange in the Brazilian Amazon*,
2. November, 2011

Benedikt Gräler
http://ifgi.uni-muenster.de/graeler
Institute for Geoinformatics
University of Muenster

Dependence Measures

We are always interested in the dependence of two or more measures/random variables X and Y or samples $\mathbf{X}$ and $\mathbf{Y}$.

covariance

$$\mathrm{Cov}(X, Y) := \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big)$$

correlation Pearson's correlation coefficient is defined by

$$\mathrm{Cor}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Rank Based Dependence

Rank based correlation measures compare the pair-wise order of bivariate samples $\mathbf{X}$ and $\mathbf{Y}$.

Spearman's rho

$$\rho(\mathbf{X}, \mathbf{Y}) := \mathrm{Cor}(\mathrm{rank}(\mathbf{X}), \mathrm{rank}(\mathbf{Y}))$$

Kendall's tau

$$\tau(\mathbf{X}, \mathbf{Y}) := 2 \cdot \frac{n_c - n_d}{n(n-1)}$$

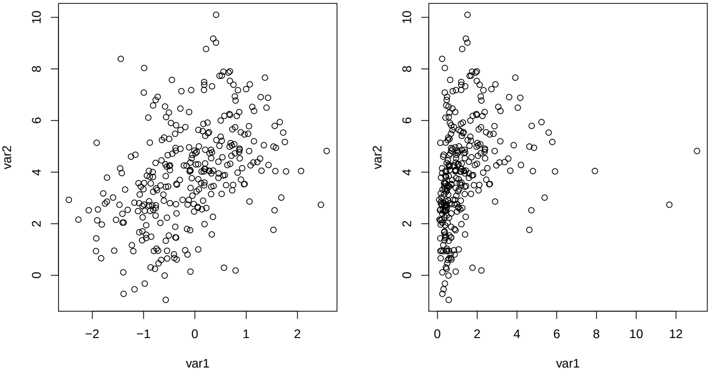
where n_c denotes the number of concordant pairs and n_d the number of discordant pairs in the sample of length n. Correction terms for ties have to be applied.

Dependence Measures in R

```
> cor(X, Y)
[1] 0.4739112
> cor(X, Y, method = "kendall")
[1] 0.3264214
> cor(X, Y, method = "spearman")
[1] 0.4842952
```

Scatterplots

Typically, we look at scatterplots to visually investigate the dependence structure.



Pearson's correlation coefficient for both plots evaluates to:

```
[1] 0.4198233
```

```
[1] 0.2619756
```

Bivariate Distributions

A scatterplot can be seen as a sample of a bivariate random variable. Every dot is one realisation of a pair (x, y) in the plain from a random variable $Z : \Omega \to \mathbb{R}^2$.
For univariate samples, a histogram illustrates the empirical likelihood distribution of a sample.
For bivariate samples, the empirical likelihood distribution is a set of pillars on the scatterplot. The elevation of each pillar corresponds to the number of points beneath it.

Example (Students 1)

Every student has a certain body weight and body height. Randomly selecting students from the institute will give us pairs of kg and cm measurements following some positive dependence.

ifgi
Institute for Geoinformatics
University of Münster

Margins I

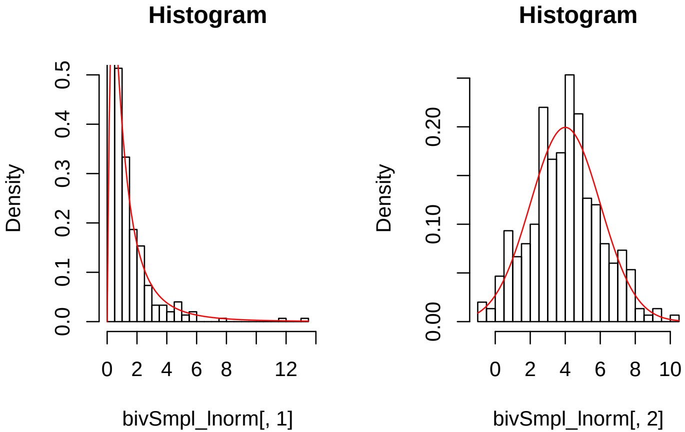
Every multivariate distribution can be seen from certain "directions", focussing on the *marginal* distributions.

Example (Students 2)

The two distributions of body weights and body heights are the two *margins* of their joint distribution.

Margins II

```
> hist(bivSmpl_lnorm[, 1], n = 20, freq = F)
> curve(dlnorm(x), col = "red", add = T)
> hist(bivSmpl_lnorm[, 2], n = 20, freq = F)
> curve(dnorm(x, 4, 2), col = "red", add = T)
```

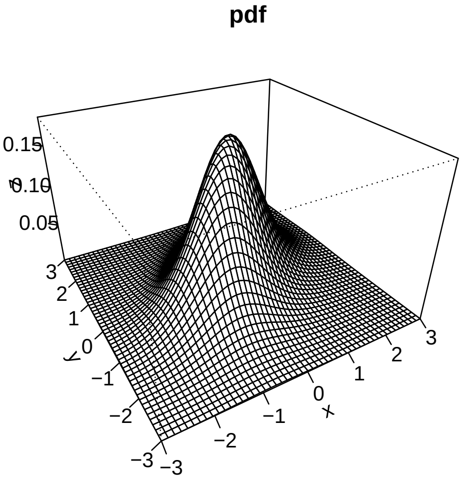


pdf and cdf I

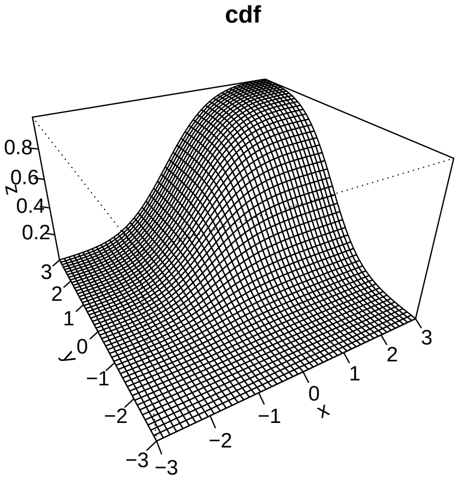
ifgi
Institute for Geoinformatics
University of Münster

As in the univariate case, multivariate random variables can be characterized by their *probability distribution function* (pdf, or density) and *cumulative distribution function* (cdf).

pdf and cdf II



pdf and cdf III



Sklar's Theorem (1959)

Theorem (Sklar)

Let H be a joint cdf with margins F and G. Then there exists a copula *such that for all $x, y \in \mathbb{R}$*

$$H(x, y) = C\big(F(x), G(y)\big).$$

If F and G are continuous, than is C uniquely defined (see e.g. Theorem 2.3.3 in [Nelsen]).

Thus, *every* multivariate distribution can be decomposed into its margins and its copula (determining the dependence of the multivariate distribution).

Copulas

Definition (Copula)

A (bivariate) *copula* is a function from $\mathbf{I}^2$ to $\mathbf{I}$, $C : \mathbf{I}^2 \to \mathbf{I}$, with

1. $C(u,0) = 0 = C(0,v), \forall u,v \in \mathbf{I}$
2. $C(u,1) = u$ and $C(1,v) = v, \forall u,v \in \mathbf{I}$
3. For every $u_1, u_2, v_1, v_2 \in \mathbf{I}$, with $u_1 \leq u_2$ and $v_1 \leq v_2$ holds:

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$$

A copula can be seen as a bivariate cdf with uniform distributed margins (condition 2).

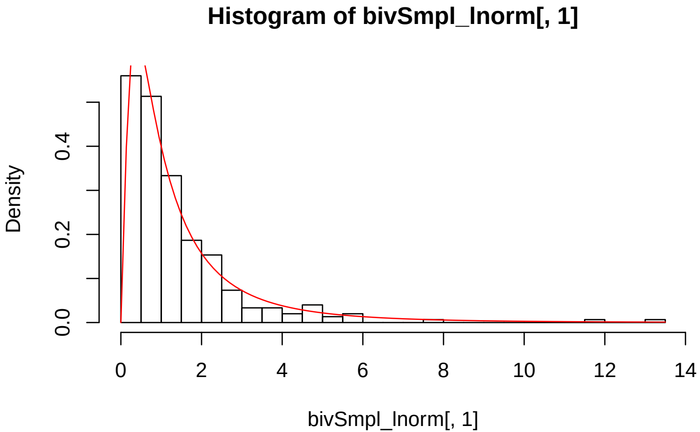
rank transformation I

This allows us to "strip" the margins and to only study the dependence structure in a common way for *all* multivariate distributions.
Any continuous margin can be transformed to be uniform distributed using a rank transformation: $u := \frac{\text{rank}(\mathbf{X})}{(n+1)}$.

obs	rank	transform
0.467	3	0.5
0.535	4	0.667
0.055	1	0.167
0.836	5	0.833
0.458	2	0.333

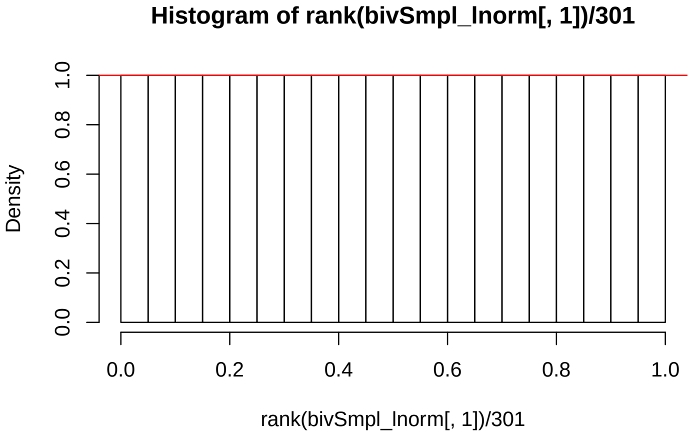
rank transformation II

```
> hist(bivSmpl_lnorm[, 1], n = 20, freq = F)
> curve(dlnorm(x), col = "red", add = T)
```

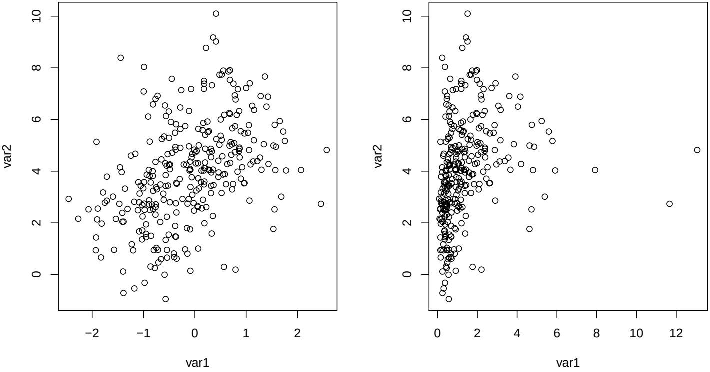


rank transformation III

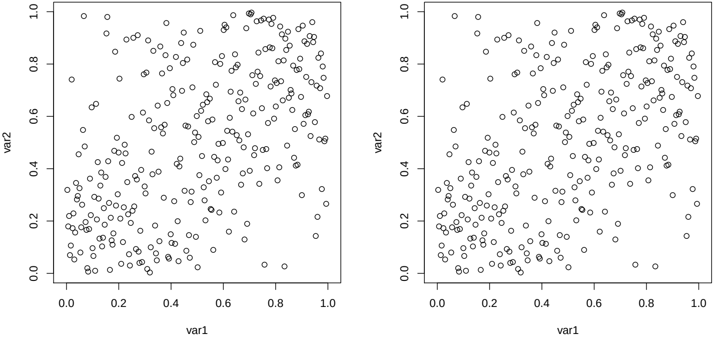
```
> hist(rank(bivSmpl_lnorm[, 1])/301, n = 20, freq = F)
> abline(h = 1, col = "red")
```



Revisiting scatterplots



scatterplots with uniform margins



Now, both scatterplots show the identical distribution of points. Therefore, they exhibit an identical dependence structure. Recall that the correlation measures differed.

Correlation under rank transformation

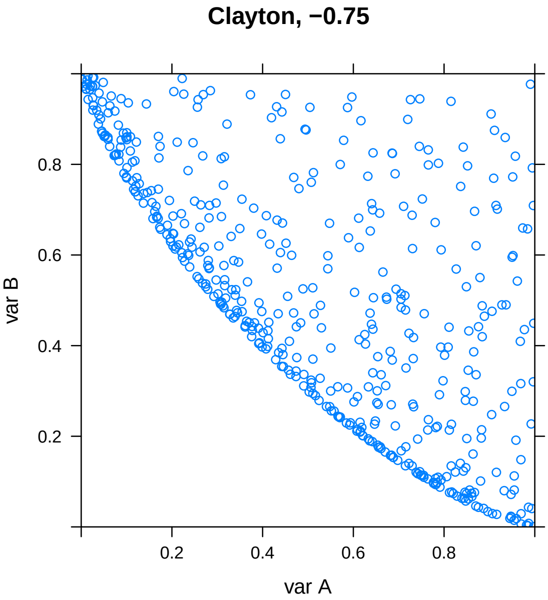
```
> cor(bivSmpl_norm)[1, 2]
[1] 0.4198233
> cor(bivSmpl_lnorm)[1, 2]
[1] 0.2619756
> cor(bivSmpl_norm, method = "kendall")[1, 2]
[1] 0.3264214
> cor(bivSmpl_lnorm, method = "kendall")[1, 2]
[1] 0.3264214
> library(spcopula)
> cor(rankTransform(bivSmpl_lnorm), method = "kendall")[1, 2]
[1] 0.3264214
```

Correlation and Copulas

Many copulas exhibit a 1-1 relation with Kendall's tau and/or Spearman's rho.
Thus, these measures can be used to estimate the copula parameter from the data set.
Finally, only the margins have to be estimated to build the bivariate distribution, but this is a one-dimensional task and a usual job in statistics.

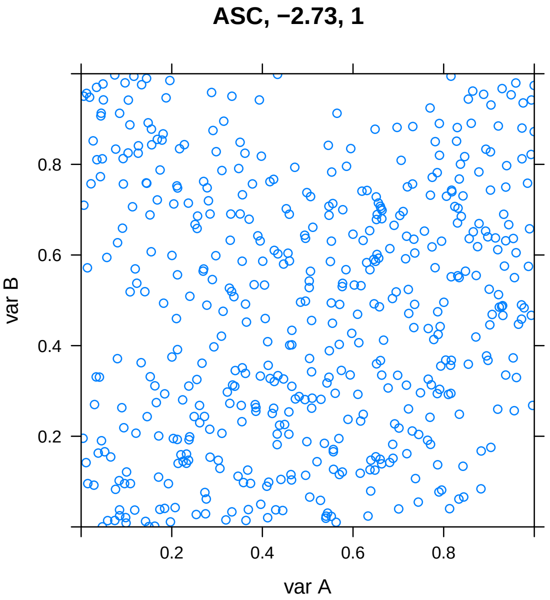
Sampling from copulas

```
> clayCop2nd <- rcopula(claytonCopula(-0.75), 500)
```



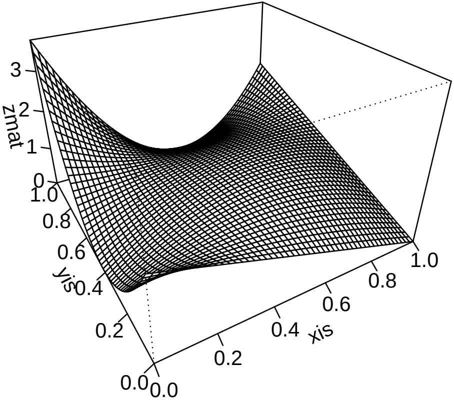
Sampling from copulas

```
> bivASC2Cop <- rcopula(asCopula(c(-2.73, 1)), 500)
```



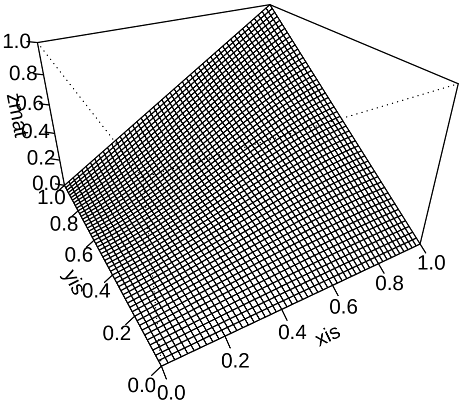
one asymmetric copula family - density

```
> persp(asCopula(c(-2.73, 1)), dcopula, ticktype = "detail")
```



one asymmetric copula family - jcdf

```
> persp(asCopula(c(-2.73, 1)), pcopula, ticktype = "detail")
```



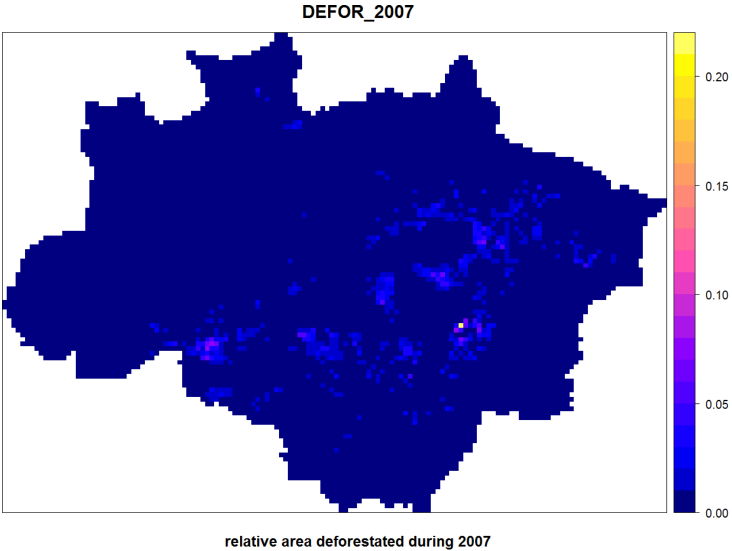
Deforestation of the Amazon

The following study is published in [Gräler et al. 2010] and was presented at the Research Symposium GIScience for Environmental Change, November 27, 2010, Campos do Jordão (São Paulo), Brazil.

Deforestation in the Amazon

The data I

The amount of yearly deforested area per raster cell is calculated by INPE, Brazil.



The data II

Additional variables present are e.g.

- demographic information
- altitude
- preserved ares
- price of forest land
- area of sugarcane or soy beans

and many more!

We will investigate the three dimensional random process given by:

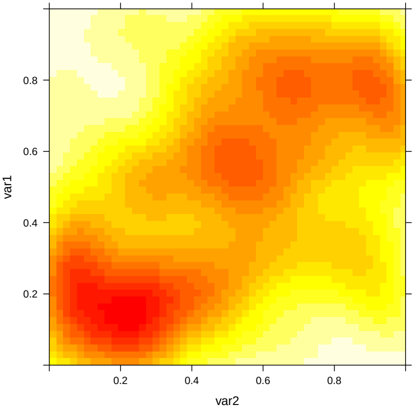
$$\text{defores. } 2006 \approx \text{defores. } 2007 \approx \text{price of forest } 2007$$

ifgi
Institute for Geoinformatics
University of Münster

studying dependence non-parametrically

Copulas can as well be used to study the dependence of different variables in a visual way. As simple scatterplots might be false leading, coloured density plots are helpful tools.

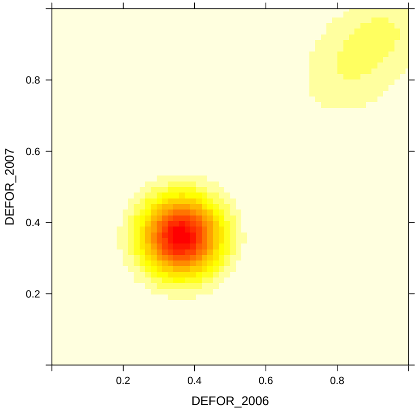
```
> dependencePlot(var1 ~ var2, bivSmpl)
```



pair-wise dependencies

Looking into pair-wise dependencies ...

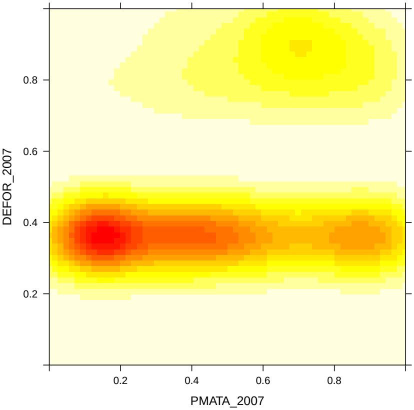
```
> dependencePlot(DEFOR_2007 ~ DEFOR_2006, defor2006)
```



pair-wise dependencies

Looking into pair-wise dependencies . . .

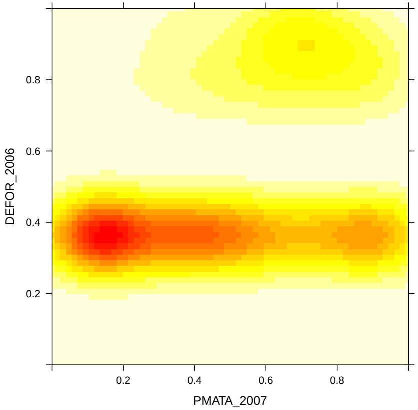
```
> dependencePlot(DEFOR_2007 ~ PMATA_2007, defor2006)
```



pair-wise dependencies

Looking into pair-wise dependencies . . .

```
> dependencePlot(DEFOR_2006 ~ PMATA_2007, defor2006)
```



Zero inflated data I

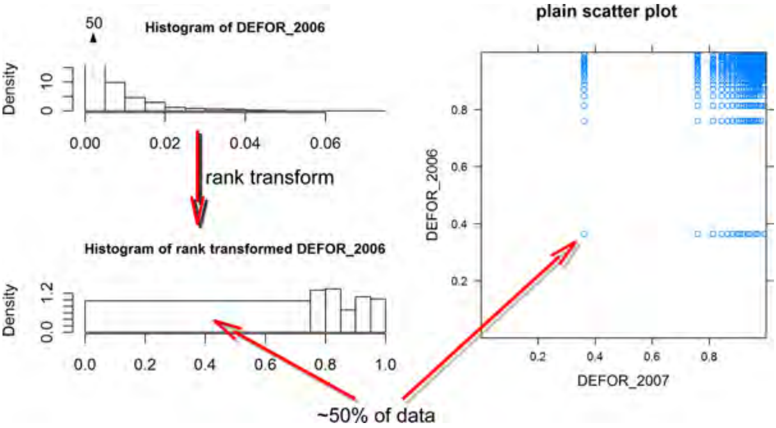
In several applications one will find a huge amount of 0s (or very small values) in a sample. This is the case for example for

- rainfall data
- nuclear radiation
- deforestation

This leads to scatter plots where a large quantity of observations is concentrated in a single point or line.

But, copulas assume continuous, equally spread data instead.

Zero inflated data II



copulas for zero inflated data - TMC I

An approach to solve this problem is by introducing *truly mixed copulas* (*TMC*) [Gräler et al. 2010].

The unit square is broken up into four areas: the lower left rectangle denoting the zero-zero pairs, the top left and lower right rectangles denoting the zero-non-zero and non-zero-zero pairs and the top right corner which can be rescaled and modelled as a copula.

ifgi
Institute for Geoinformatics
University of Münster

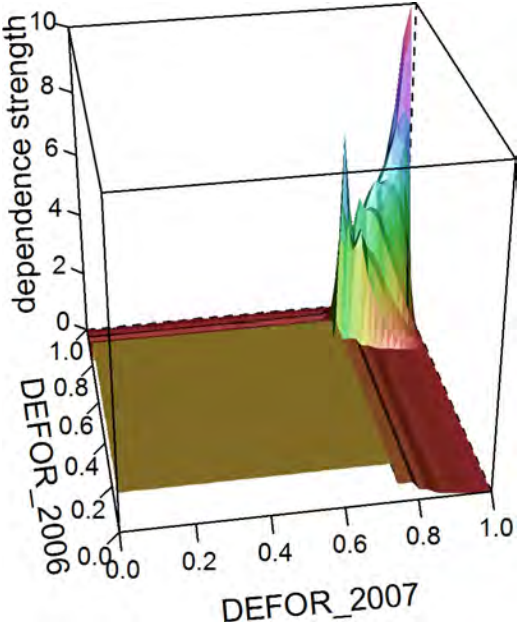
copulas for zero inflated data - TMC II

The rescaling is done in a way that the joint bivariate function is a copula again maintaining the mass relations and copula properties:

To achieve this, we need to estimate inner marginal functions and counter parts such that both add up to a constant 1.

A truly mixed copula density might look like:

copulas for zero inflated data - TMC III



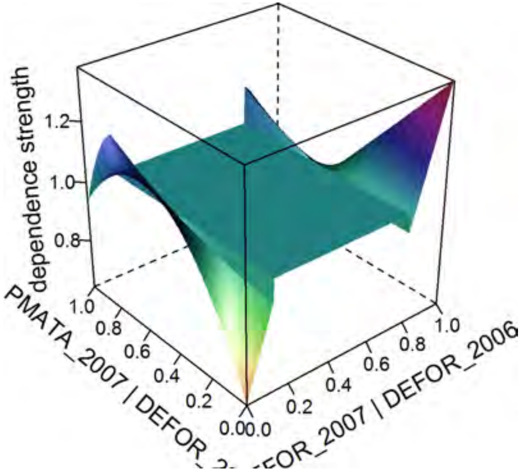
cutted copulas I

Instead of in one corner one might find a big bunch of values some where in the middle of the unit interval.

This part can be cut out according to its mass and inserted after the estimation process [Gräler et al. 2010].

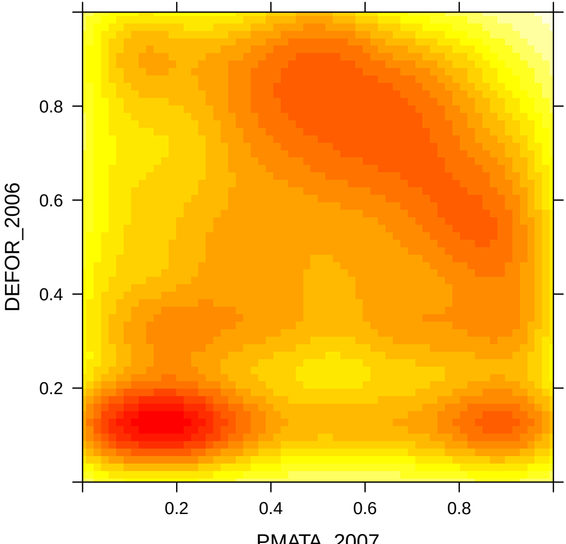
Depending on the distribution of this cut-out, a distribution function might be necessary. The cut copula looks like

cutted copulas II

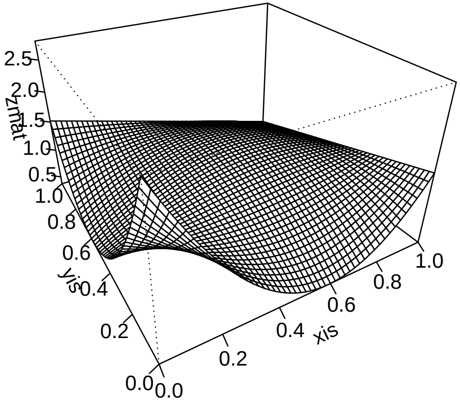


the empirical copula: DEFOR 2006 - PMATA 2007

The top right part is rescaled to uniform distributed margins and a copula is fitted:

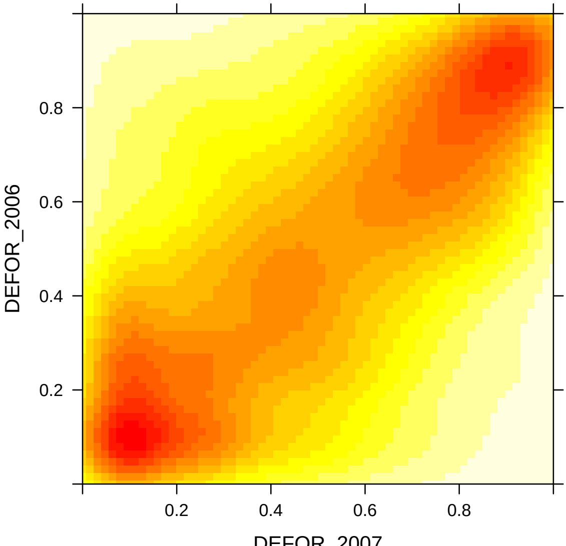


the fit's density

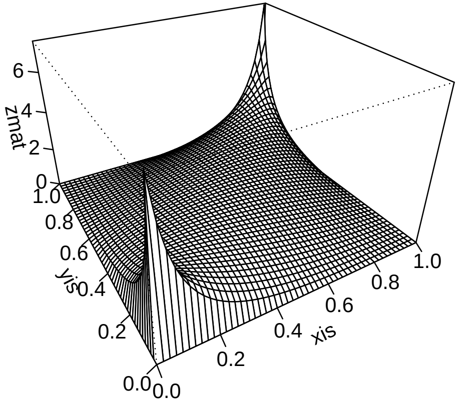


the empirical copula: DEFOR 2007 - DEFOR 2006

The top right part is rescaled to uniform distributed margins and a copula is fitted:



the fit's density



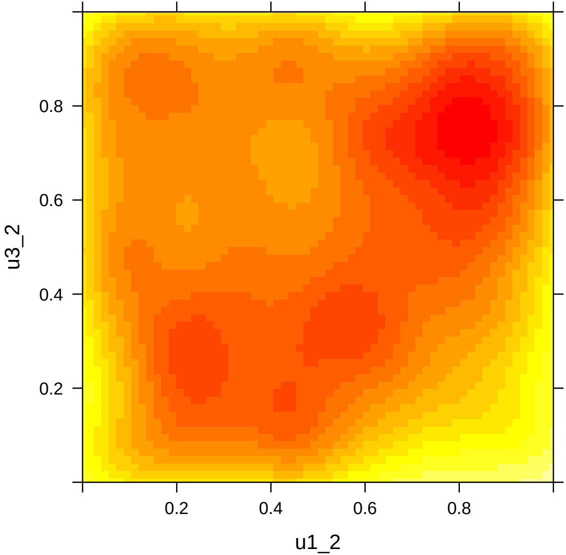
ifgi
Institute for Geoinformatics
University of Münster

the cutted copula I

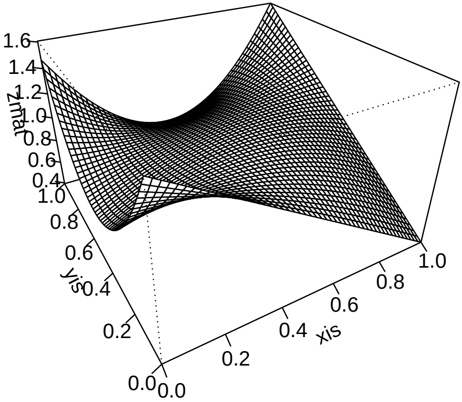
After the transformation of the data under the conditional distribution there is a second value which takes a massive mass.

the cutted copula II

The remaining copula is:

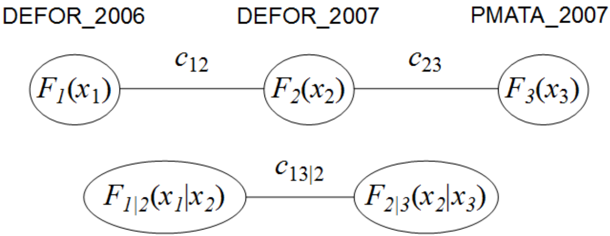


its density



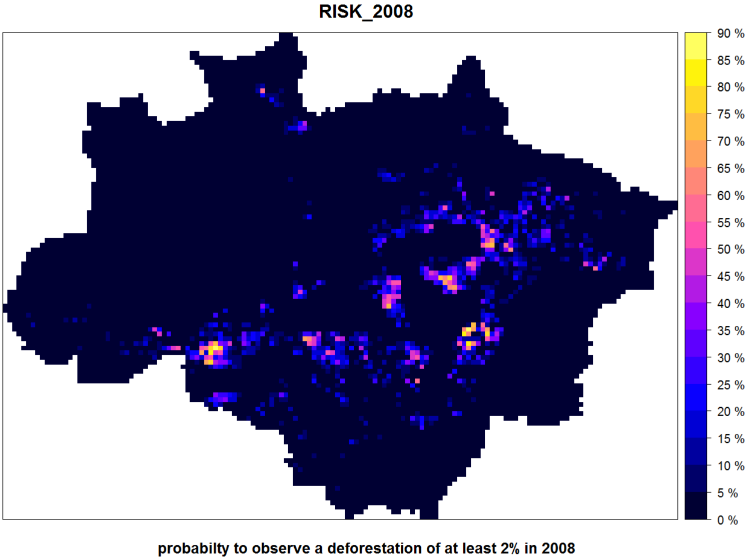
the joined copula

After we estimated the three pieces we can put them together:



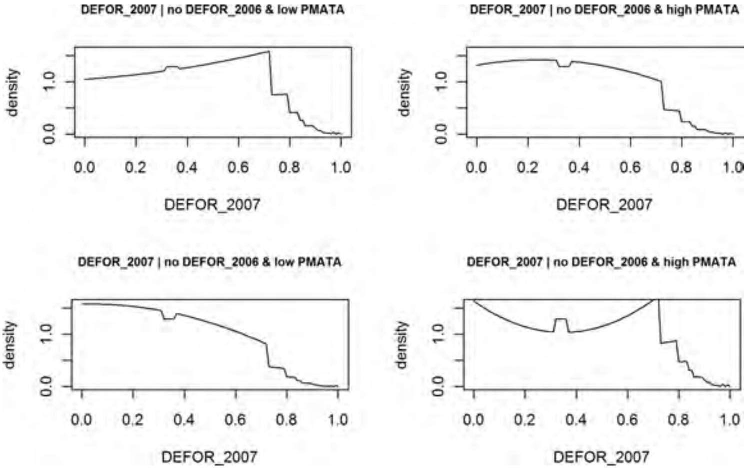
Risk

Assuming temporal stationarity lets us calculate a risk map of deforestation for a given threshold

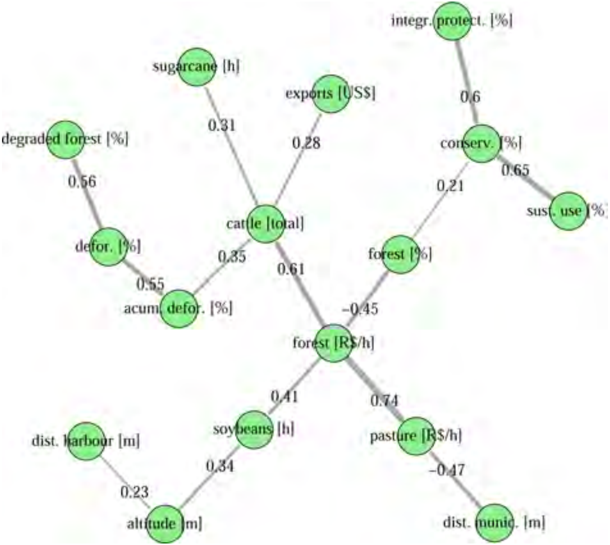


inspecting conditional densities for different copulas

Substituting the CQSec copula (top row) C_{23} with the best Gaussian (bottom row) has a visible impact:



the maximum spanning tree



- Gräler, Benedikt (2009), 'Copulas for Spatio-Temporal Random Fields', Diploma thesis at the Institute of Mathematical Statistics and Institue for Geoinformatics, University of Muenster.
- Gräler, B., H. Kazianka & G. M. de Espindola (2010): "Copulas, a novel approach to model spatial and spatio-temporal dependence". In K. Henneböhl, L. Vinhas, E. Pebesma, & G. Cãmara (Eds.), GIScience for Environmental Change Symposium Proceedings, ifgiprints (Vol. 40, pp. 49-54). Presented at the GIScience for Environmental Change, November 27, 2010, Campos do Jordão (São Paulo), Brazil: AKA Verlag.

- Gräler, Benedikt & Edzer Pebesma (2011): The pair-copula construction for spatial data: a new approach to model spatial dependency. *Procedia Environmental Sciences* (Vol. 7, pp. 206-211). Poster at: Spatial Statistics 2011 - Mapping global change. Enschede, The Netherlands, 23-25 March 2011.
- Nelsen, Roger B. (2006): An Introduction to Copulas. Springer Series in Statistics, second edition.
- Salvadori, G., C. De Michele, N. T. Kottegoda & R. Rosso (2007): Extremes in Nature, An Approach Using Copulas. Water Science and Technology Library, Springer.