Cluster analysis of time series via Kendall distribution

ROBERTA PAPPADÀ

rpappada@units.it

Department of Economics, Business, Mathematics and Statistics
University of Trieste

Joint work with F. DURANTE (FUB)

Workshop on *Clustering methods and their applications*
Free University of Bozen-Bolzano

November 28, 2014

1. Introduction
    - Motivation
    - Why Clustering?
    - Clustering time series
2. Detecting dependence via Kendall distribution
    - The Kendall distribution function
    - Kendall distribution function and multivariate Return Period
3. Clustering time series via Kendall distribution
    - The clustering procedure
    - An empirical case study
4. Conclusions

Motivation

"Impacts on the physical environment are often the result of compound events."

IPCC Report, 2012

"Dependencies change the risks. It is possible to calculate the consequences of individual events, such as an extreme tide, heavy rainfall and key workers being absent. However, if the events are interrelated, (for example a storm causes a high tide, or heavy rain prevents workers from accessing the site) then the probability of their co-occurrence is much higher than might be expected."

Sutherland et al., Science, 2013

Motivation

- Recent years have seen a growing need for new tools in the analysis of the regional variability of rainfall extremes useful in environmental monitoring
- Hydrological phenomena are often multidimensional and require the joint modeling of several random variables (Genest, Favre, 2007)
- Many research efforts have remarked on the usefulness of extreme value theory and copula functions in assessing climate changes and detecting spatial clusters

Why Clustering?

- Informally, clustering consists in finding natural groupings among objects
- The identification of different groups in a set of climate time series is relevant to identify subgroups characterized by similar behavior in order to adopt specific risk management strategies
- Clustering techniques can be used to find some dependence information, which is a key tool in geosciences and hydrology
- The detection of spatial clusters can help in summarizing available data, extracting useful information

Clustering time series

- A widely used approach to measure similarity is to consider a Pearson-correlation based distance metric
- Many studies have underlined that classical correlation measures are often inadequate to capture the real dependence structure between individual risk factors (Embrechts, McNeil, Straumann, 2002)
- Recent approaches in time series clustering adopt a suitable copula-based dissimilarity measure (Durante, Pappadà, Torelli, 2014) or combine extreme value theory and classification techniques for assessing the spatial distribution of extremes (Scotto, Alonso, Barbosa, 2010)

Clustering time series

The Problem

A time-series clustering procedure allows to group together series exhibiting common trends occurring at different times or similar sub-patterns in the data.

Choice of a proximity measure: A similarity (proximity) measure is defined to measure the “closeness” of the observations.

Choice of group-building algorithm: On the basis of the proximity measures the objects are assigned to groups to obtain

- high intra-cluster similarity
- low inter-cluster similarity

What is similarity?

A quality that makes one person or thing like another;
the quality or state of being similar: resemblance;
a comparable aspect: correspondence.

Merriam-Webster's Dictionary

Clustering time series

The notion of dissimilarity

A dissimilarity function is usually understood to measure some kind of distance between objects.

The dissimilarity $\delta(x_i, x_j)$ between x_i, x_j satisfies:

1. non-negativity: $\delta(x_i, x_j) \geq 0$
2. identity: $\delta(x_i, x_i) = 0$
3. symmetry: $\delta(x_i, x_j) = \delta(x_j, x_i)$

Clustering time series

The notion of dissimilarity

A dissimilarity function is usually understood to measure some kind of distance between objects.
The dissimilarity $\delta(x_i, x_j)$ between x_i, x_j satisfies:

1. non-negativity: $\delta(x_i, x_j) \geq 0$
2. identity: $\delta(x_i, x_i) = 0$
3. symmetry: $\delta(x_i, x_j) = \delta(x_j, x_i)$

- if the triangle inequality holds, δ is a distance measure;
- many clustering methods use distance measures to define the dissimilarity between any pair of objects;
- common dissimilarity measures can be obtained as a suitable transformation of Pearson correlation coefficient, rank correlation coefficients (Kendall's τ, Spearman's ρ), Hoeffding's D statistic, etc.

Kendall d.f.

(Bivariate) copula.

A bivariate copula (or a 2-copula) is a 2-dimensional distribution function whose univariate marginals are uniformly distributed on $[0, 1]$.

Kendall distribution function

Let $\boldsymbol{X}$ be a continuous random vector on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose distribution function is equal to H.

- The Probability Integral Transform (PIT) of $\boldsymbol{X}$ is the random variable $W = H(\boldsymbol{X})$.
- The distribution function K of W is called Kendall distribution function associated with $\boldsymbol{X}$

Kendall d.f.

The calculation of K depends only on the copula C of $\boldsymbol{X}$ and does not involve the knowledge the marginal distributions.

Specifically, for every $t \in [0,1]$

$$\begin{aligned} K(t) &= \mathbb{P}(W \leq t) \\ &= \mathbb{P}(H(\boldsymbol{X}) \leq t) \\ &= \mu_H(\{\boldsymbol{x} \in \mathbb{R}^2 : H(\boldsymbol{x}) \leq t\}) \\ &= \mu_C(\{\boldsymbol{u} \in [0,1]^2 : C(\boldsymbol{u}) \leq t\}) \end{aligned}$$

where μ_H, μ_C are the measures induced by the distribution function H and the copula C on $\mathbb{R}^2$, respectively.

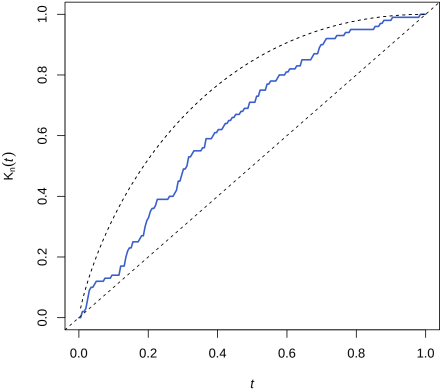
Kendall d.f. and dependence

- For a random pair (X, Y), $K(t)$ represents the probability of the event $\{H(X, Y) \leq t\}$ thus the distribution puts no mass outside the interval $[0, 1]$.
- For a given pair (X, Y) distributed as H,

$$\tau(X, Y) = 4E\{H(X, Y)\} - 1 = 3 - 4 \int_0^1 K(t)dt.$$

- $K(t) = K_M(t) = t$ for all $0 \leq t \leq 1$ if and only if $C = M$, where M is the the Fréchet-Hoeffding upper bound copula, denoting *perfect positive dependence*.
- $K(t) = t - t \log(t)$, $0 \leq t \leq 1$ if and only if $C = \Pi$, where Π is the product copula, denoting *independence*.
- For each Kendall distribution K, one has the lower bound $K \geq K_M$ on $[0, 1]$.

Kendall d.f.: illustration



Empirical Kendall d.f. obtained from a random sample of size $n = 100$ points from a bivariate normal vector with Kendall's tau equal to 0.5.

Estimation of the Kendall distribution

Suppose that $(X_{11}, X_{12}), \ldots, (X_{T1}, X_{T2})$ is a random sample from a distribution H with copula C.

The empirical Kendall distribution function K_T is given, for all $q \in [0, 1]$, by

$$K_T(q) = \frac{1}{T} \sum_{j=1}^{T} \mathbf{1}(W_j \leq q),$$

where, for each $j \in \{1, \ldots, T\}$,

$$W_j = \frac{1}{T+1} \sum_{t=1}^{T} \mathbf{1}(X_{t1} < X_{j1}, X_{t2} < X_{j2}).$$

The empirical process $\sqrt{T}(K_T - K)$ converges in law to a centered Gaussian limit under mild regularity conditions.

[Genest, Něslehová, Ziegel, 2011]

The multivariate RP

The notion of Return Period (RP) is frequently used in environmental sciences for the identification of dangerous events, and provides a means for rational decision making and risk assessment

Informally, the RP is defined as the average time elapsing between two successive realizations of a prescribed event.

The calculation of the RP is strictly related to the notion of copula and the Kendall distribution function turns out to be a fundamental tool for calculating a copula-based RP for multivariate events.

[Salvadori, De Michele, Durante, 2011]

The multivariate RP

Preliminaries

Consider a d-dimensional random vector $\boldsymbol{X}$ with d.f. F, describing the phenomenon under investigation, with suitable marginals F_i's and d-copula C. Assume F continuous and strictly increasing in each marginal. By virtue of Sklar's Theorem

$$F = C(F_1, \ldots, F_d).$$

Given any $t \in (0, 1)$ define the region

$$\mathcal{R}_t^> = \{\boldsymbol{x} \in \mathbb{R}^d : F(\boldsymbol{x}) > t\}.$$

called the super-critical region.

Remark Realizations lying over the same *critical layer* $\mathcal{L}_t^F = \{\boldsymbol{x} \in \mathbb{R}^d : F(\boldsymbol{x}) = t\}$ do always identify the same "dangerous" region (i.e. $\mathcal{R}_t^>$).

The multivariate RP

Definition

Kendall RP

Let $\boldsymbol{X}$ be a multivariate r.v. with d.f. $F = C(F_1, \ldots, F_d)$. Let $\mathcal{L}_t^F$ be the critical layer supporting a realization $\boldsymbol{x}$ of $\boldsymbol{X}$. Then, the RP associated with $\boldsymbol{x}$ is defined as

$$T_{\boldsymbol{x}}^{>} = \frac{\mu}{\mathbb{P}(\boldsymbol{X} \in \mathcal{R}_t^{>})} = \frac{\mu}{1 - K_C(t)}$$

where μ is the average time elapsing between $\boldsymbol{X}_i$ and $\boldsymbol{X}_{i+1}$ and K_C is the Kendall distribution function associated with C.

The goal

We would like to use the Kendall distribution function in order to develop a novel clustering procedure for grouping random vectors:

- a time-series clustering procedure can be used in order to obtain a description of the relationship between measurements at different sites;
- we aim at grouping the climate time series according to the strength of their inter-dependence;
- a copula-approach allows to detect the presence of clusters of the analysed sites on the basis of the componentwise maxima.

The clustering procedure

Consider an iid sample $X_1^t, \dots, X_n^t$ from a given r.v. $\boldsymbol{X}$ corresponding to n different measurements collected at time $t \in \{1, \dots, T\}$.

1. Calculate the Kendall distribution function K^{ij} for (X_i, X_j)
2. Define the metrics
$$d_2(K, K_M) = \int_0^1 (q - K(q))^2 dq$$
$$d_\infty(K, K_M) = \sup_{q \in [0,1]} |q - K(q)| dq$$
where $K = K^{ij}$ and $K_M(t) = t$ is the Kendall distribution of comonotone random variables
3. Create a dissimilarity matrix $D := (\delta_{ij})$, $i, j = 1, \dots, n$, where, for instance, $\delta_{ij} = d_2(K^{ij}, K_M)$
4. Apply classical clustering algorithms like hierarchical clustering (`hclust`) or *fuzzy* clustering (`fanny`).

The clustering procedure

Hierarchical Agglomerative Clustering

hierarchical Create a hierarchical decomposition of the set of objects using some criterion.

The clustering procedure

Hierarchical Agglomerative Clustering

hierarchical Create a hierarchical decomposition of the set of objects using some criterion.

agglomerative Starting with each item in its own cluster, find the best pair to merge into a new cluster and repeat until all clusters are fused together.

The clustering procedure

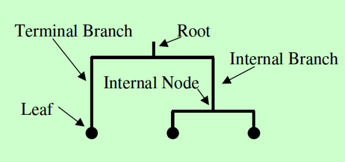
Hierarchical Agglomerative Clustering

hierarchical Create a hierarchical decomposition of the set of objects using some criterion.

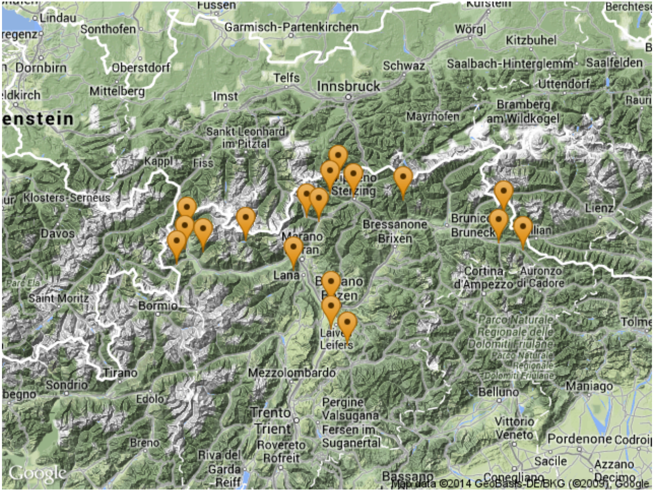
agglomerative Starting with each item in its own cluster, find the best pair to merge into a new cluster and repeat until all clusters are fused together.

A useful tool for summarizing similarity measurements

The similarity between two objects is represented in a dendrogram as the height of the lowest internal node they share.



An empirical case study



Map of the rainfall measurement stations in the province of Bolzano-Bozen (North-Eastern, Italy)

The data

We consider daily rainfall measurements recorded at 18 gauge stations spread across the province of Bolzano-Bozen in the North-Eastern Italy from 1961 to 2010. This data consists of $d = 18$ time series originally formed by $T = 18262$ observations.

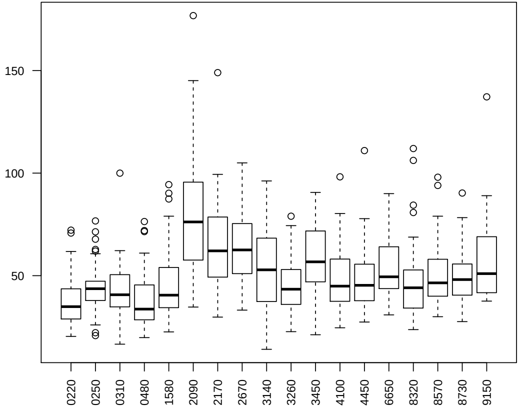
We extract annual maxima at each spatial location and obtain a 50×18 matrix of time series observations $\tilde{X}_i^m$, $m \in \{1, \ldots, 50\}$, $i \in \{1, \ldots, 18\}$

$$M = \begin{pmatrix} \tilde{X}_1^1 & \tilde{X}_2^1 & \cdots & \tilde{X}_{18}^1 \\ \tilde{X}_1^2 & \tilde{X}_2^2 & \cdots & \tilde{X}_{18}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{X}_1^{50} & \tilde{X}_2^{50} & \cdots & \tilde{X}_{18}^{50} \end{pmatrix}$$

The data

Code	Station	Longitude	Latitude	Height (m)
0220	S.VALENTINO ALLA MUTA	10.5277	46.7745	1520
0310	TUBRE	10.4775	46.6503	1119
2090	PLATA	11.1783	46.8225	1147
3140	FLERES	11.3477	46.9639	1246
3260	VIPITENO-CONVENTO	11.4295	46.8978	948
8320	BOLZANO	11.3127	46.4976	254
9150	SESTO	12.3477	46.7035	1310
0250	MONTE MARIA	10.5213	46.7057	1310
0480	MAZIA	10.6175	46.6943	1570
1580	VERNAGO	10.8493	46.7357	1700
2170	S.LEONARDO PASSIARIA	11.2471	46.8091	644
2670	PAVICOLO	11.1093	46.6278	1400
3450	RIDANNA	11.3068	46.9091	1350
4450	S.MADDALENA IN CASIES	12.2427	46.8353	1398
6650	FUNDRES	11.7029	46.8872	1159
8570	BRONZOLO	11.3111	46.4065	226
8730	REDAGNO	11.3968	46.3465	1562
9100	ANTERIVO	11.3678	46.2773	1209

The data



Box plots of annual maxima at each station from 1961 to 2010. On the *y*-axis the amount of rainfall is measured in millimeters.

The dissimilarity measure

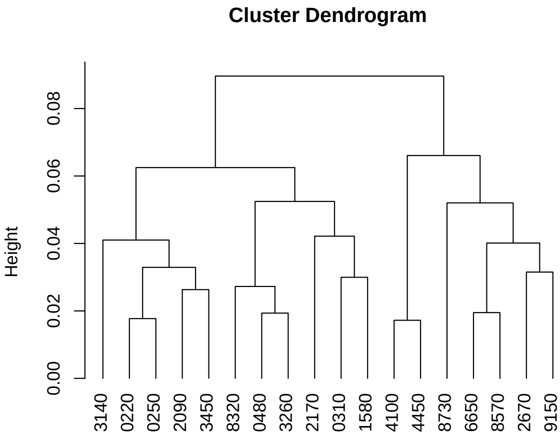
The choice of the dissimilarity measure has to reflect the final goal of the clustering procedure: two strongly dependent time series correspond to an extremely low value of their dissimilarity.

For $i, j = 1, \ldots, 18$, the dissimilarity between two time series is computed as

$$\delta_{ij} = d_2(\hat{K}^{ij}, K_M) = \int_0^1 (q - \hat{K}^{ij}(q))^2 dq,$$

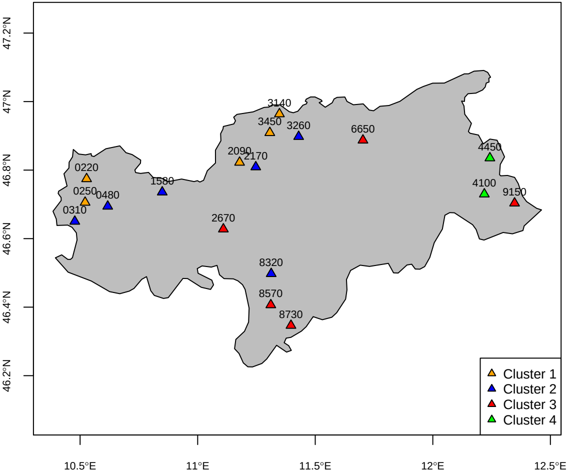
where $\hat{K}^{ij}$ is the non-parametrically estimated Kendall distribution function based on each pair of maxima observations $(\tilde{X}_i^m, \tilde{X}_j^m)$, $m \in \{1, \ldots, 50\}$.

Clusters visualization



Dendrogram for the 18 rainfall measurement stations based on hierarchical clustering with complete linkage method.

Clusters visualization



Map of the rainfall measurement stations marked according the the 4-clusters solution

Conclusions

- A procedure for grouping time series according to a copula-based dependence function has been presented.
- A criterion to measure dissimilarity is defined on the basis of the Kendall distribution associated to two continuous random variables, since such a function provides useful information in terms of environmental risk.
- Homogeneity in the sense of Kendall's distance implies homogeneity in the sense of return period, a notion frequently used in environmental sciences for the identification of dangerous events and risk assessment.
- A case study with environmental data illustrates the potential of the presented methodology.

Main Bibliography

- Durante, F., Pappadà, R., Cluster analysis of time series via Kendall distribution. In P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*, Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3–25
- Durante, F., Pappadà, R., Torelli, N., Clustering of time series via non-parametric tail dependence estimation. Stat Papers, DOI: 10.1007/s00362-014-0605-7
- Durante, F., Pappadà, R., Torelli, N., Clustering of financial time series in risky scenarios. Adv Data Anal Classif, 8(4): 359-376 (2014)
- Embrechts, P., McNeil, A., Straumann, D., Correlation and dependence in risk management: properties and pitfalls. In M.A.H. Dempster ed., Risk Management: Value at Risk and Beyond, 176-223 (2002)
- Genest, C., Favre, A.-C., Everything you always wanted to know about copula modeling but were afraid to ask. J. Hydrologic Eng., 12(4): 347-368 (2007)
- Salvadori, G., De Michele, C., Durante, F., On the return period and design in a multivariate framework. Hydrol. Earth Syst. Sci., 15: 3293-3305 (2011)
- Scotto, M.G., Alonso, A.M., Barbosa, S.M., Clustering time series of sea levels: Extreme value approach. J. Waterway, Port, Coastal, and Ocean Engrg., 136: 215- 225 (2010)

Thank You for Your attention!