Assessing Imputation of Extreme Data on Climatological Time Series

Jairo Cugliari, José G Gómez

Assessing Imputation of Extreme Data on Climatological Time Series

Jairo Cugliari[1], José G. Gómez [2]

1 Jairo.Cugliari@univ-lyon2.fr, ERIC EA 3083, Université de Lyon, Lyon 2
2 jose-gregorio.gomez-garcia@unicaen.fr, LMNO,UMR CNRS 6139, Université de Caen Normandie

Abstract

Techniques for imputation of missing data are generally well known. However, they are often applied and validated on non-extreme data. Even for specific approaches specialized on extreme data, the question raises on how to assess the imputation. Because of this, we propose to use known estimators of extreme value theory, such as the extremogram, which we generalize to time series with the presence of missing data.

Introduction

Recently extreme temperature events took important relevance in global scale studies. However, there is a lack of information and/or studies on Africa and South America where the geographical inhomogeneity of the results is caused mainly due to the absence of data. A special case is Uruguay, where long time series are available, but incomplete, with missing records on several consecutive years.
On the other hand, climate change, annual seasonality and exogenous factors are some of the well known non stationary patterns present in climatological records. We use a flexible non parametric model to extract these factors from a non stationary time series. The remaining noise will then be assumed as a strictly stationary $\mathbb{R}^d$-valued time series, denoted by $X := (X_i)_{i\in\mathbb{Z}}$.

Preliminaries : The extremogram [2]

Def. 1. The *extremogram* of X, for two sets A and B bounded away from zero, is defined (provided the limit exists) as

$$\rho^X_{A,B}(h) := \lim_{x\to\infty} \mathbb{P}\left(x^{-1}X_h \in B | x^{-1}X_0 \in A\right), \qquad h = 0, 1, \ldots. \tag{1}$$

For the observations $X_1, \ldots, X_n$, the sample extremogram is given by

$$\bar\rho^X_{A,B,n}(h) := \frac{\sum_{i=1}^{n-h} \mathbb{I}_{\{u_m^{-1}X_{i+h}\in B, u_m^{-1}X_i\in A\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{u_m^{-1}X_i\in A\}}}, \tag{2}$$

where u_m is the $(1-1/m)$-quantile of the distribution of $|X_0|$.
In order to have a consistent result, we require $m = m_n \longrightarrow \infty$ with $m = o(n)$ as $n \to \infty$, and u_m is the sequence used in the definition of regularly varying time series in [2, § 1.2].

Our framework

The aim of this work is to study the extremogram of X when in the time series X some observations might be missed. For this, we suppose that the observations of the data $(X_i)_{i=1,\ldots,n}$ occur at times $1 = i_1 < i_2 < \cdots < i_m = n$. From [4], we use the amplitude modulated observations of X, which is defined by

$$Y_i := b_i X_i, \qquad i = 1, \ldots, n \tag{3}$$

where $b_i = 1$ if X_i is observed and $b_i = 0$ if X_i is missing. Moreover, we will assume that $(X_i)_i$ and $(b_i)_i$ are independent sequences.

Assumptions for the process (b_i)

(A.1) (b_i) has finite second moments and, for $h = 0, 1, \ldots$, $\gamma_b(h) := \mathrm{Cov}\,(b_i, b_{i+h})$ is independent of $i \in \mathbb{Z}$.
(A.2) $\bar\mu_{b,n} := \frac{1}{n}\sum_{i=1}^n b_i \xrightarrow{a.s.} \mu_b := \mathbb{E}b_0 = \mathbb{P}(b_0 = 1)$
(A.3) $\bar\nu_{b,n}(h) := \frac{1}{n-h}\sum_{i=1}^{n-h} b_i b_{i+h} \xrightarrow{a.s.} \nu_b(h) := \mathbb{E}b_0 b_h, \quad h = 0, 1, \ldots.$
(A.4) $\mu_b \neq 0$ and $\nu_b(h) \neq 0$ for each $h = 0, 1, \ldots$.

Assumptions for the process (X_i)

(D) $(X_i)_{i\in\mathbb{Z}}$ is α-mixing with rate function $(\alpha_l)_{l\in\mathbb{N}}$. Moreover, there exist $m = m_n$ and $r = r_n \longrightarrow \infty$ with $m/n \longrightarrow 0$ and $r/m \longrightarrow 0$, such that

$$\lim_{n\to\infty} m \sum_{l=r}^{\infty} \alpha_l = 0, \tag{4}$$

and $$\lim_{k\to\infty} \limsup_{n\to\infty} m \sum_{h=k}^{r} \mathbb{P}\left(|X_h| > \epsilon a_m, |X_0| > \epsilon a_m\right) = 0, \quad \forall \epsilon > 0. \tag{5}$$

Results

Proposition 1 Suppose that $(b_i)_i$ satisfy the condition (A.1) and that X is regularly varying. Then, for two sets A and B bounded away zero, $\rho^X_{A,B}(h)$ and $\rho^Y_{A,B}(h)$ exist for $h = 0, 1, 2, \ldots$, and the following equality holds

$$\rho^Y_{A,B}(h) = \rho_b(h)\, \rho^X_{A,B}(h), \tag{6}$$

for each $h = 0, 1, 2, \ldots$, where $\rho_b(h) := \mathbb{P}(b_h = 1 | b_0 = 1) = \nu_b(h)/\mu_b$.

From Proposition 1, we can naturally provide an estimator of the extremogram of the sequence $X = (X_i)_i$ through its amplitude modulated version $(Y_i)_i$ and the sequence $(b_i)_i$ as follows:

$$\bar\rho^X_{A,B,n}(h) = \frac{\bar\rho^Y_{A,B,n}(h)}{\bar\rho_{b,n}(h)},$$

provided that $\bar\rho_{b,n}(h) \neq 0$, where $\bar\rho_{b,n}(h) := \bar\nu_{b,n}(h)/\bar\mu_b$.

Theorem 1 Suppose that $(b_i)_i$ satisfies the conditions (A.1)-(A.4) and assume that $X = (X_i)_i$ is regularly varying with index $\alpha > 0$. If condition (D) holds, $\alpha_r = o(m/n)$ and $m = o(n^{1/3})$, then

$$\left(\frac{n}{m}\right)^{1/2} \left[\bar\rho^X_{A,B,n}(h) - \rho^X_{A,B,n}(h)\right]_{h=0,1,\ldots,H} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{A,B}), \tag{7}$$

where the asymptotic covariance is defined in [2, Corollary 3.3] and $\rho^X_{A,B,n}(h) := \mathbb{P}(u_m^{-1}X_h \in B | u_m^{-1}X_0 \in A)$.

Application on real data

Data. Daily minimum records of 11 measurement stations in Uruguay. (cf. Fig.) with different patterns of missing records.

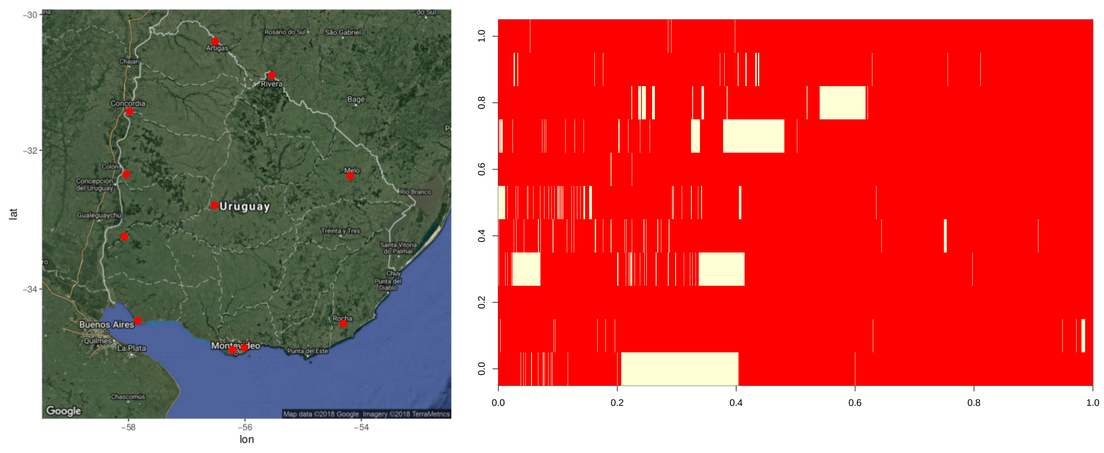

Figure 1: Geographical location of measuring stations (*l.*) and missing values patterns (*r.*).

After following the imputation strategy presented in [1], we obtain the extremogram for the original and imputed data. The graphic of this object is in Figure 2. Both lines (black and red) follow a reasonable similar pattern. Then, both series present the same extreme behavior and thus we conclude that the imputation scheme is appropriate.

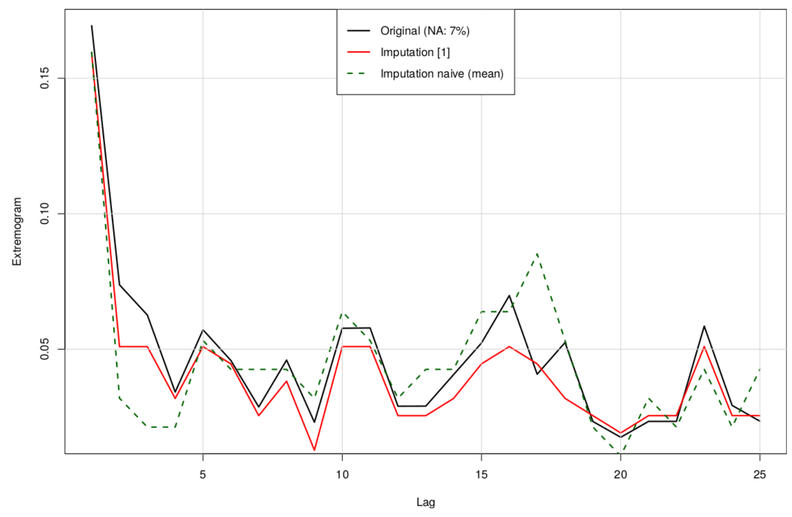

Figure 2: Extremogram for the original and imputed series, both by the method proposed in [1] and using a naive approach.

Conclusion & future work

- Use measures of extremes constructed from cluster functionals [3].
- Adaptation for usage on climatological random fields.
- Modeling the absence/presence of missing data through a 2-state Markov process.
- Assume only local stationarity on the X processes.

References

[1] Cugliari, J., De Mello, S., Renom, M. *Imputation of Temperature Extremes using Generalized Additive Models*, JDS 2018, France.

[2] Davis, R.A. & Mikosch, T. *The extremogram: a correlogram for extreme events*. Bernoulli. 2009; 15: 977–1009.

[3] Gómez, J.G. *Dependent Lindeberg central limit theorem for the fidis of empirical processes of cluster functionals*. Statistics : A Journal of Theoretical and Applied Statistics (2018).

[4] Parzen, E. *On spectral analysis with missing observations and amplitude modulation*. Sankhya Ser. 25:383-392 (1963).