Imputation of Extreme Data using GAM with assesement by extremograms

Jairo Cugliari[1]

ERIC EA 3083, Université de Lyon, Lyon 2

26 de Febrero 2019 @ SIESTA

Joint work with: Santiago de Mello (UdelaR)
Madeleine Renom (UdelaR, Inumet)
José G. Gómez (Univ. de Caen Normandie)

Context

- Relevance of extreme temperatures recently highlighted
  - Intergovernmental Panel on Climate Change (IPCC, 2013)
  - Special Report on Extreme Events (SREX, 2011)
- global increment on the number of extreme temperature events ...
- ... but lack of information on Africa and South America.
- Rusticucci (2012) shows from a set of works about extreme events on South America, that the geographical heterogeneity of the results is mainly due to the absence of data.
- A special case is Uruguay, were long daily time series are available
- However, the series are not complete with missing records.

Aim Obtain an operational methodology for imputation of daily extreme temperatures in Uruguay from 11 meteorological stations.

Outline

Plan

Data

Imputation via GAM

Extremograms with missing data

Data

- Daily minimum records of 11 measurement stations in Uruguay.
- For this work, data ranges from 1950 until the end of 2014.
- Different patterns of missing records

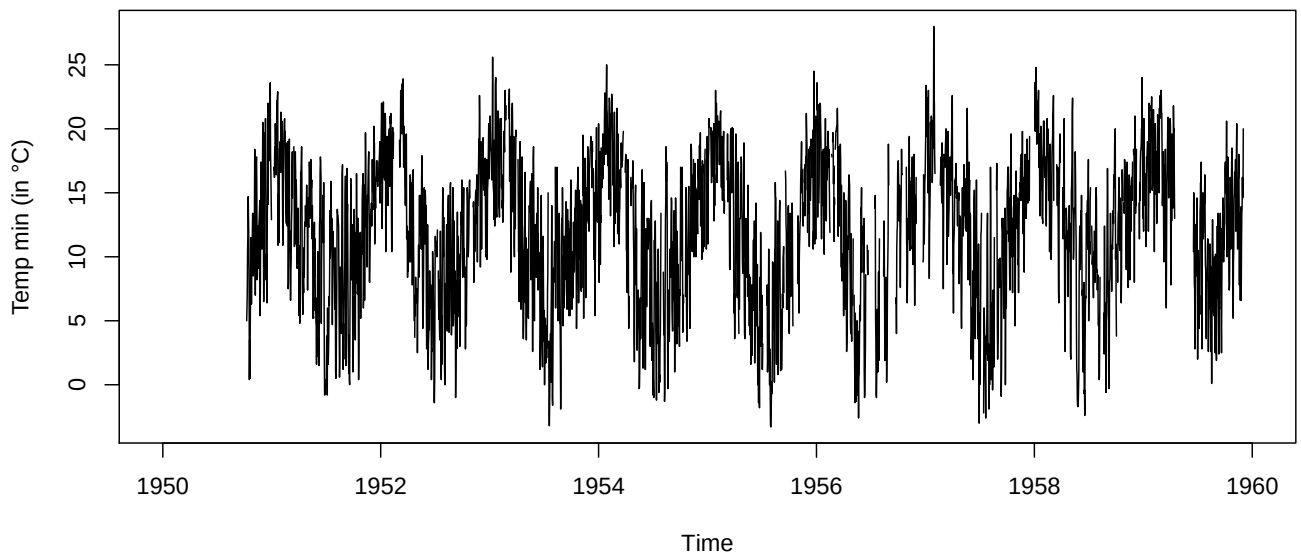


Figure 1: Part of the daily minimum temperature records for station in Paso de los Toros with missing records on the beginning of the sample period and close to year 1959.

Spatial dependence

- Distance between stations x and y:

$$d_c(x, y) = \sqrt{2(1 - \text{corr}(x, y)^2)},$$

where $\text{corr}(x, y)$ (linear) correlation coefficient (LCC).

- Multidimensional scaling of the distance matrix induced by $d_c(x, y)$.

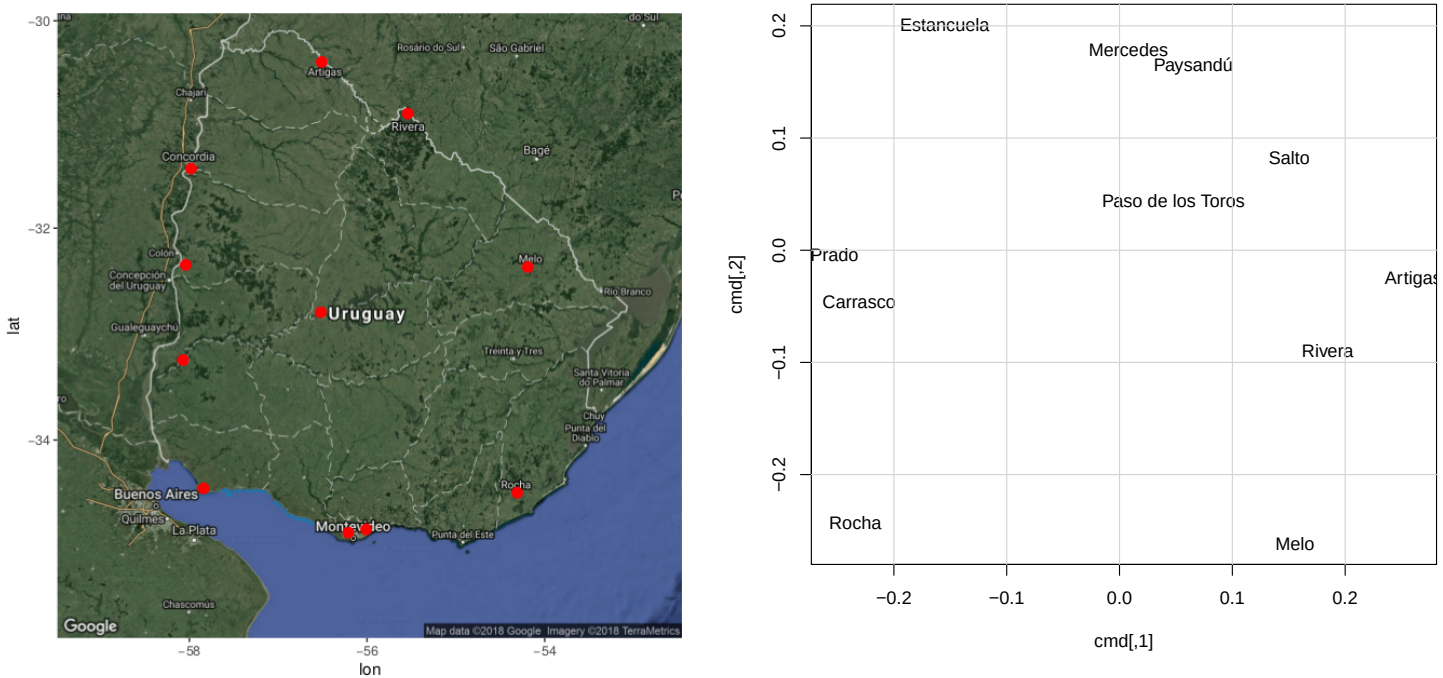


Figure 2: Meteorological stations on the map (left) and recovered relative positions from the distance $d_c(x, y)$.

Mean imputation (and problems)

- Fit a reasonable linear model (i.e. with several relevant covariates)
- Check residuals (for instance graphically)

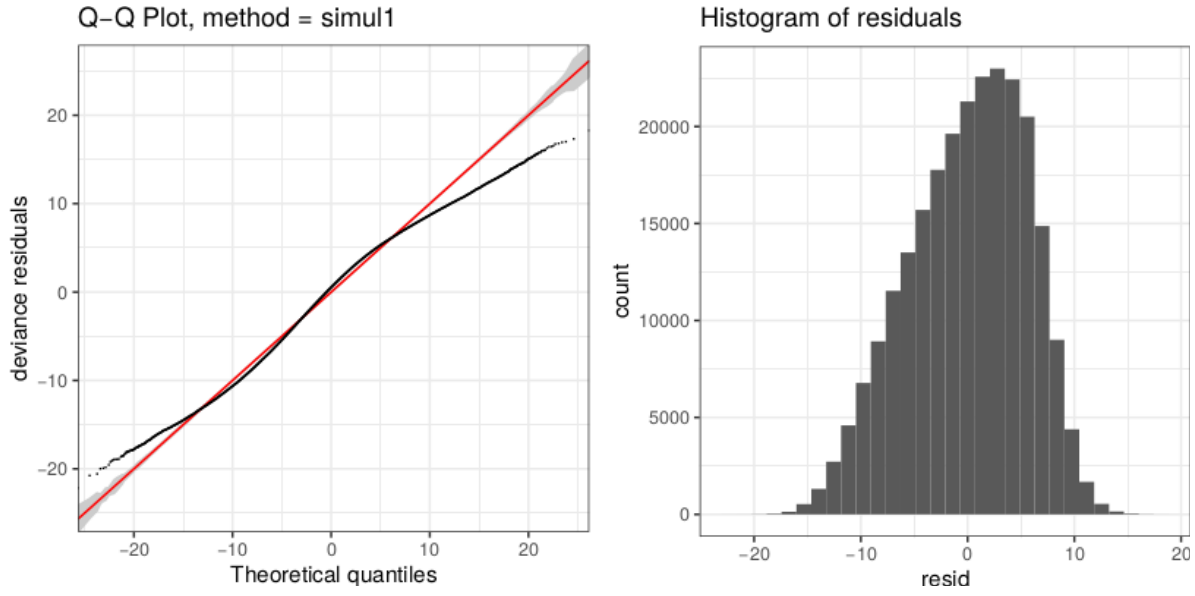


Figure 3: Residual analysis of a first model for daily minimum temperature

Plan

Data

Imputation via GAM

Extremograms with missing data

Generalized Additive Models

- Response, y_i, predictors x_{ji}, model

$$y_i \sim \pi(\mu, \boldsymbol{\theta}), \quad \text{where} \quad g(\mu_i) = \mathbf{A_i}\boldsymbol{\gamma} + \sum_j f_j(x_{ji}).$$

  - π a distribution, location parameter μ and other ones $\boldsymbol{\theta}$
  - f_j are smooth functions (to be estimated)
  - $\mathbf{A}$ is a know design matrix and $\boldsymbol{\gamma}$ the associated parameters
  - g is a know link function (e.g. identity, log, ...)
- Estimation:
  - expand f_j over a spline basis with coefficient to be estimated
  - smoothness is controlled by penalization
  - (more technicalities: identifiability, estimation of smoothing parameters, ...)

S. Wood, Generalized Additive Models: An Introduction with R, 2nd ed. (2017)

Imputación con extremos generalizados

Generalized Extreme Values (GEV)

- Let $Y_k = \{Z_1^{(k)}, \dots Z_m^{(k)}\}$ be the block maxima
  (for us m is the number of measurement on day, Y_i is the daily max)

- An interesting class of limiting distributions is

$$G(y) = \exp\left\{ -\left[1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\},$$

  with parameters μ (location), $\sigma > 0$ (scale) and ξ (shape)
- GEV is connected to Gumbel, Fréchet and Weibull

- In GAM they can be estimated thanks to the `gevlss` family of link

GAM for extremes data

Distributional regression

(or GAM for location, scale and shape (GAMLSS))

$$y_i \sim \pi(\theta_{1i}, \theta_{2i}, ...)$$

with $g_j(\theta_{ji}) = \sum f_j$.

Generalized Extreme Values (GEV)

- An interesting class of limiting distributions is

$$G(y) = \exp\left\{ -\left[1 + \xi\left(\frac{y-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\},$$

  with parameters μ (location), $\sigma > 0$ (scale) and ξ (shape)

GAM for extremes is $\pi = G$ (up to reparametrizations).

In practice

- R package `mgcv` allows one to fit a location-scale-shape model (also `VGAM`)
- $\mu, \log \sigma, \xi$ depend on smooth linear predictors
- ξ is constrained to $[-1, 0.5]$ to ensure both MLE consistency and finite variance.

- A good sense principle to follow : deeper the component, lesser the information

In our case, we enrich our data with the following informations

- target: minus the minimum temperature (`tempM`)
- locations of each measurement `station` (`lat, lon, h`)
- trend and annual cycle (`trend, cycle`)
- climatological variables (`mei`)

Some results: parametric and smooth terms

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
μ	-14.91	0.55	-27.149	$<$ 2e-16	***
clinland	0.33	0.04	7.436	1.04e-13	***
clnord	1.13	1.53	0.738	0.461	
$\log\sigma$	1.61	0.003	582.412	$<$ 2e-16	***
clinland.1	0.07	0.004	18.827	$<$ 2e-16	***
clnord.1	0.055	0.004	15.365	$<$ 2e-16	***
ξ	0.16	0.004	36.739	$< 2e-16$	***

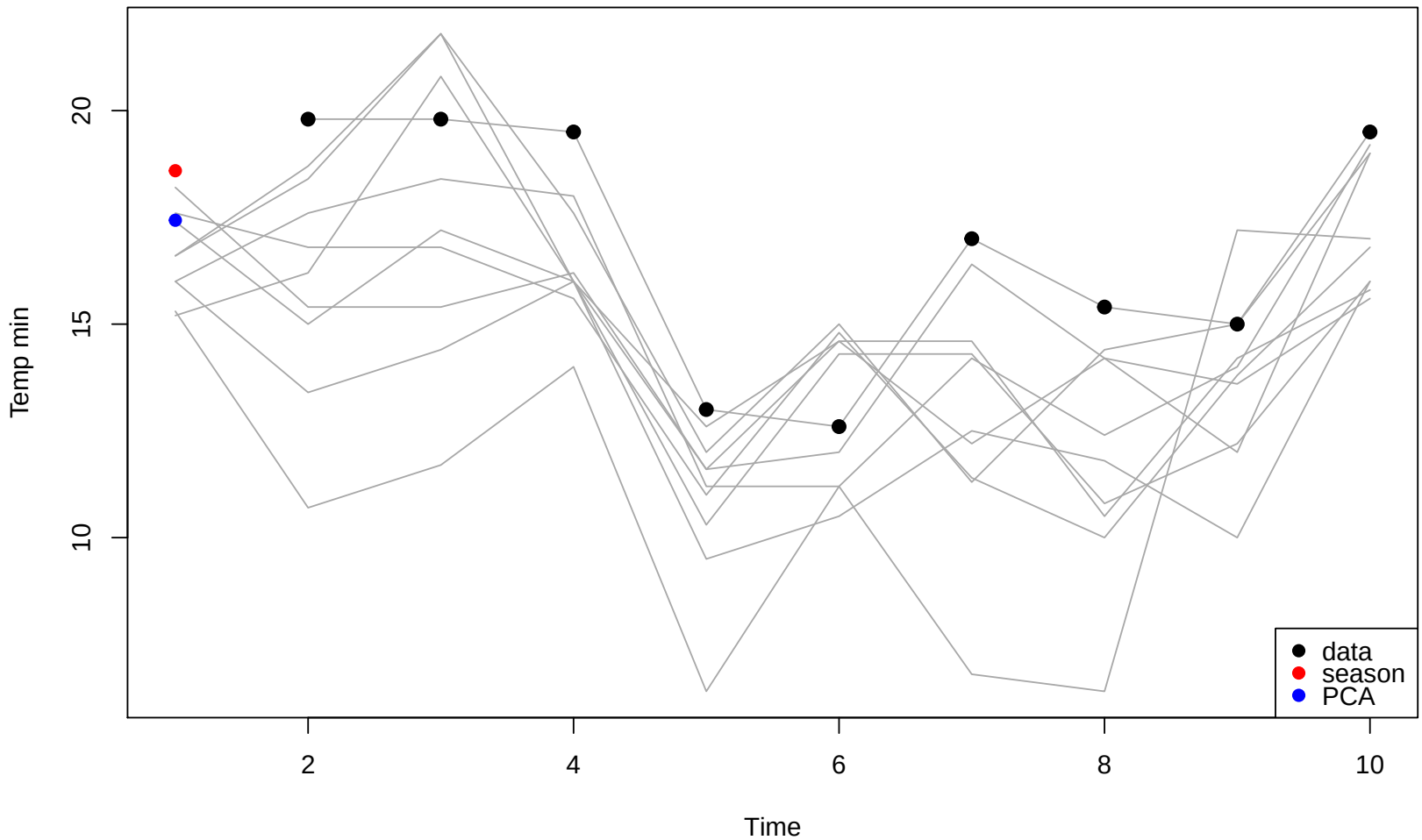
	edf	Ref.df	Chi.sq	p-value	
s(mei)	8.965	9.000	1826.9	$<$ 2e-16	***
s(cycle)	8.571	8.947	1556.5	$<$ 2e-16	***
s(h)	7.984	8.000	924.7	$<$ 2e-16	***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

- Fréchet kind of extreme ($\xi > 0$)
  - coefficient's signs seems to be reasonable
  - smooth terms are all significant

Algunos ejemplos de reconstrucción

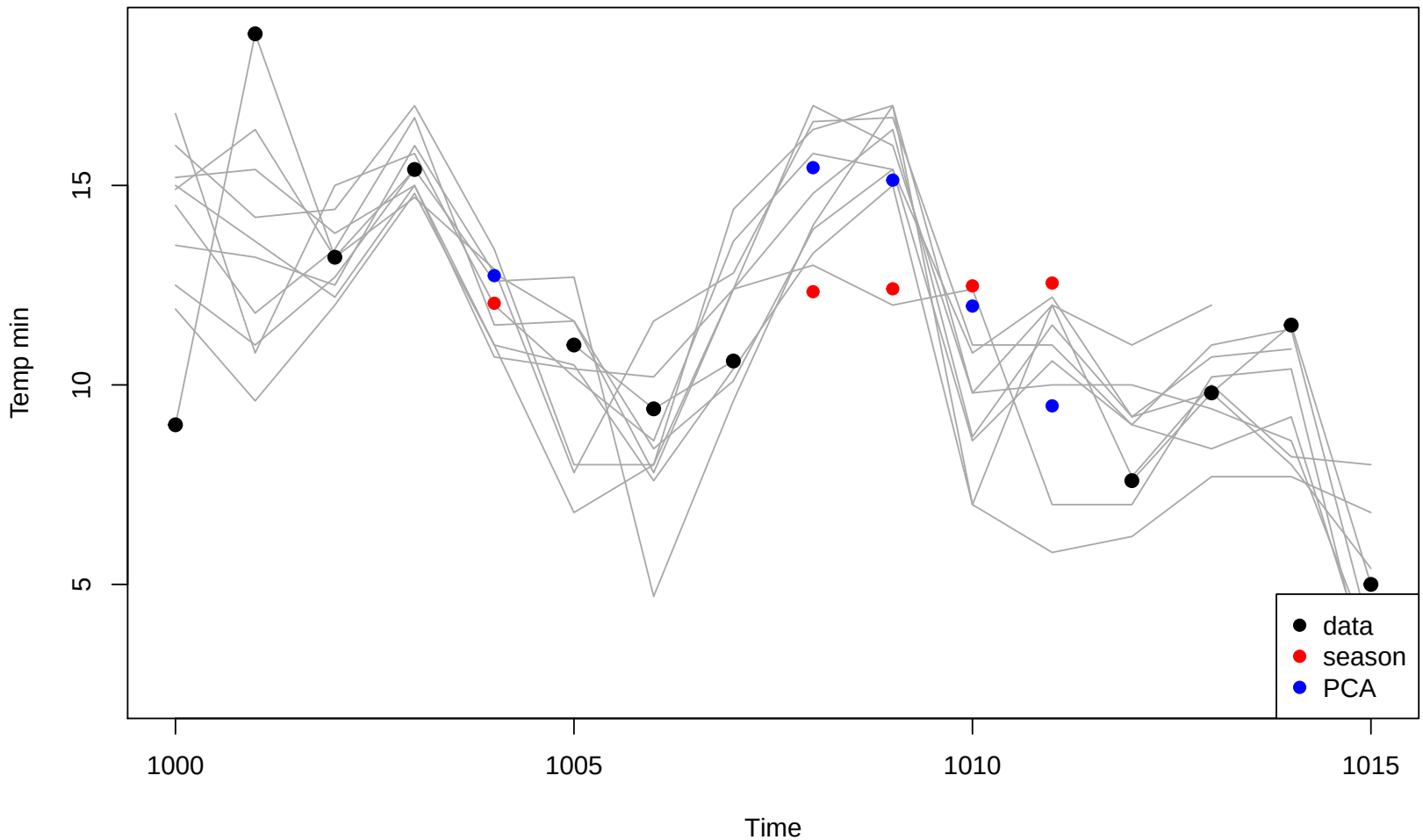
- Artigas, 1 dato faltante al inicio del periodo
- En rojo la construcción por ciclo anual, en azul la imputación PCA



Algunos ejemplos de reconstrucción

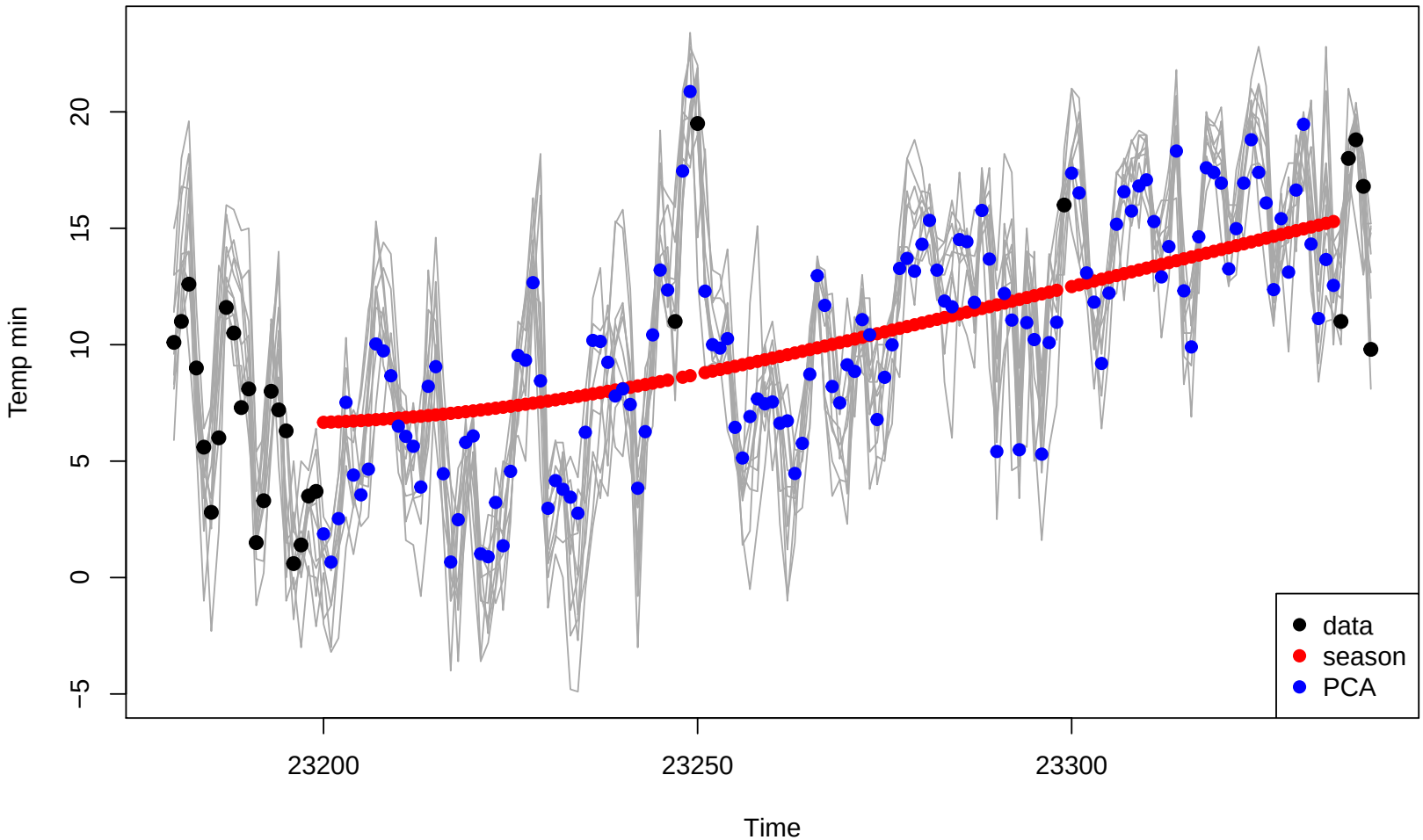
Artigas, 5 datos faltantes.

Los puntos rojos son casi constantes por la resolución diaria (el ciclo estacional es de muy baja frecuencia)



Algunos ejemplos de reconstrucción

Carrasco, muchos datos faltantes. La mejora de la imputación PCA es clara respecto al ciclo estacional. Los puntos azules siguen los picos y los valles.



Plan

Data

Imputation via GAM

Extremograms with missing data

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

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

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].

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

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

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

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

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

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

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

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

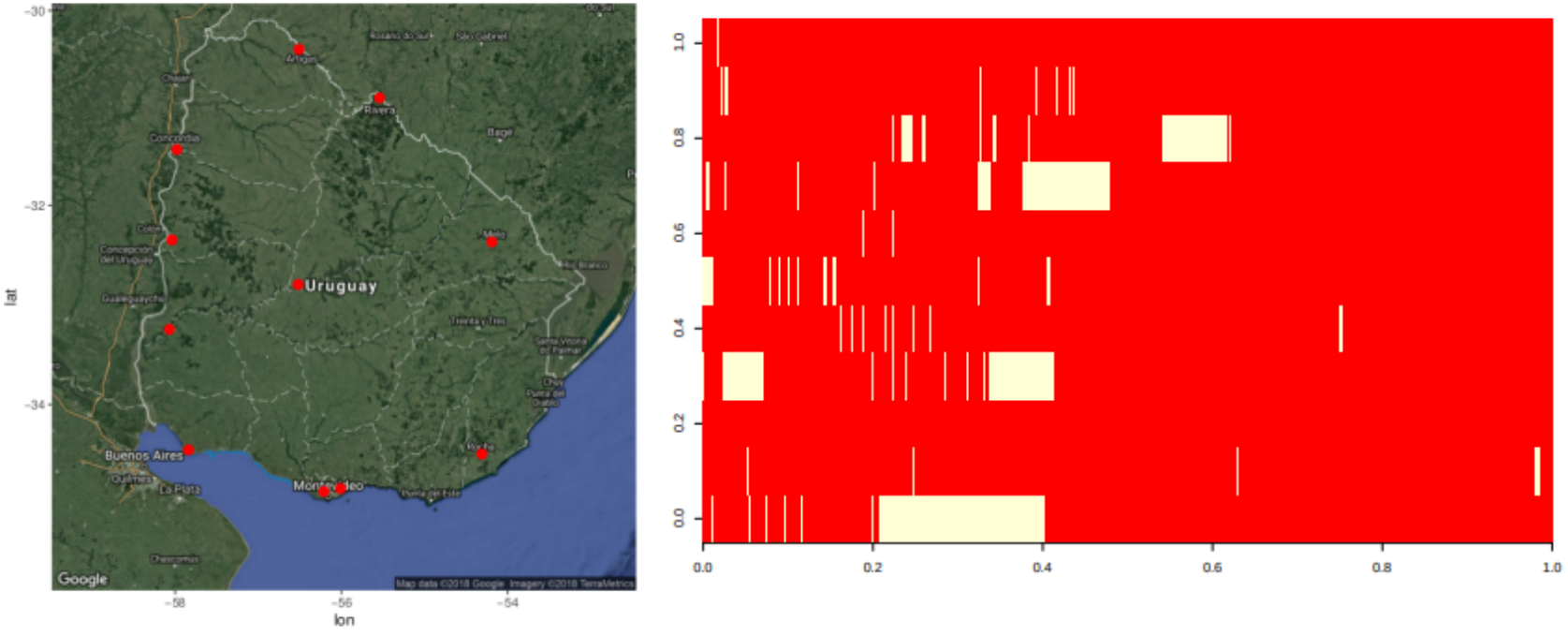


Figure 1: Geographical location of measuring stations (*l.*) and missing values patterns (*r.*).

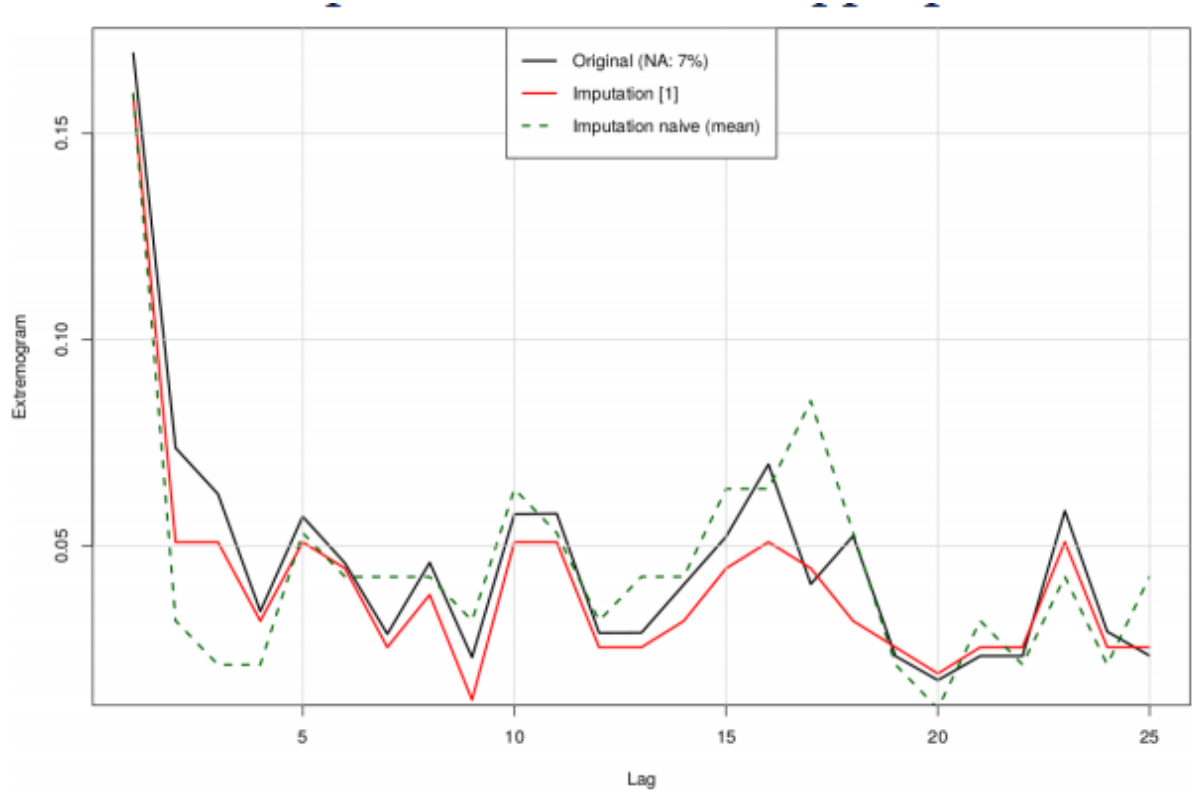


Figure 2: Extremogram for the original and imputed series, both by the method proposed in [1] and using a naive approach.

Conclusion

- GAM with GEV provides with a flexible framework for modelling extreme data
- Data imputation is done by predicting values of unobserved temperatures

- Some clues on how to asses the quality of the reconstruction