Rough data or preprocessing for extremes of temperatures : A stochastic process approach

Didier Dacunha-Castelle Université Paris-Sud Orsay

Joint work with Sylvie Parey (EDF) and Thi Tu Hoang (Orsay-EDF)

Sylvie Parey on **Wednesday** will present applications of the first part to series of observations, or produced by reanalysis or numerical models

Preprocessing or rough data : how to choose

ROUGH APPROACH $\longrightarrow$ no preprocessing : example : non stationary extremes GEV models : $G(\mu(t),\sigma(t),\xi(t))$
Problems well known : quality of probability asymptotics : Parametric or non parametric; quite small samples for GEV or POT. Model choices for parametric, seasonality smoothness for non parametric
PREPROCESSING APPROACH $\longrightarrow$ try to let the stochastic part of the signal as stationary and simple as possible. Basically $X_t = T_t + V_t Y_t$ *and* $T_t = m_t + S_t$ and $V_t = v_t s$
Separation (when justified?) low frequency and seasonality
REMARK Justification of any treatment of rough data needs analysis of stochastic properties

Preprocessing and reduced stochastic process

Remark: very important statistical pre-processing (non parametric techniques as loess, lasso, wavelets) **depends on properties of *Y***: for instance of the global level of correlation

Ex: control of the global correlation of the process *Y*

$$\Gamma_N = \sum_i \sum_j \gamma(i,j)$$

basic for objective smooth parameter tuning (cross validation)

(Thi Tu Hoang thesis Orsay 2010 and forthcoming paper)

the **same** remark for the use of **rough** data and interpretation of statistical results

Main goals for statistical studies of temperature

Preprocessing:

Non stationarities 1- trends low frequency smoothness : mean;variance; extremes 2-seasonalities 3-links seasonalties low frequency

Modelisation, fit, validation Analysis of the stochastic part dynamics extremes how are extremes produced

Models of simulations with "right representation of extremes : complex events (extremes…)

Comparison between series covariables attribution and causality : distressing polemics on climate science : scientific part almost always on time series statistical problems ex France :sun activity versus temperatures

Cyclo-stationarity of the reduced process $Y_t = \frac{X_t - T_t}{\sigma_t}$

Tests of (cyclo)stationarity

$\longrightarrow$For **correlations** or functional of correlations ex: mean time equilibrium return

$\longrightarrow$for extremes ***extY*** of Y: **K** hypothesis ***extY*** is stationary GEV(μ,σ,ξ) to test against **H** alternative ***ext Y*** non stationary GEV ($\mu(t),\sigma(t),\xi(t)$)

Let Δ **a** **distance** **between the models estimated under H and the model estimated under K** (for instance: L^2 or Kullback distances)

Tables by bootstrap, power test computed

General conclusion (Sylvie Parey): for the amount of observed data: K cannot be rejected on almost all parts of Europe

Stochastic modelisation : results and work program

Temperature reduced process has complex properties
Obviously a continuous time process with continuous trajectories
Evident biperiodicity day and year, the two periodicities are linked
What about the memory : there are physical reasons to think that continuous time process has Markov property
What about discrete time observed subprocesses : markovianity can be tested. For instance : at fixed hour every this properties remains
Series of max or min have Markov properties
One can check what theory predicts mean temperature are not Markovian (for instance the mean memory is about 3 for day scale)
The continuous process is thus a bicyclic stationary diffusion if the **first preprocessing has eliminated low frequency** , if not stationary is not too far

Stochastic analysis for extremes : the continuous time

Let X_t be a recurrent diffusion with values in the open interval (r_1 , r_2) the endpoints r_1 , r_2 being inaccessible

$$dX_t = b(X_t)dt + a(X_t)dW_t$$

where b is the drift and a is the diffusion coefficient of the diffusion process. W is a Brownian motion.
Let s be the scale function of the process:

$$s(x) = \int^x e^{\int^u -2\frac{b(v)}{a^2(v)}dv} du$$

Maximum of a stationary diffusion
Berman result

Let M$_T$ maximum of a stationary ergodic diffusion on 0,T)

Theorem : *If there exists two sequences of real numbers such that* $\frac{M_T - A_t}{B_T} \rightarrow G$ *in distribution then G is a* GEV *distribution*, G(μ,σ,ξ) and G is also the max limit of a sequence of independent equidistributed r.v of distribution **F** linked to the diffusion by

$$\log F = -\frac{1}{s} \qquad \frac{s''}{s} = -2\frac{b}{a}$$

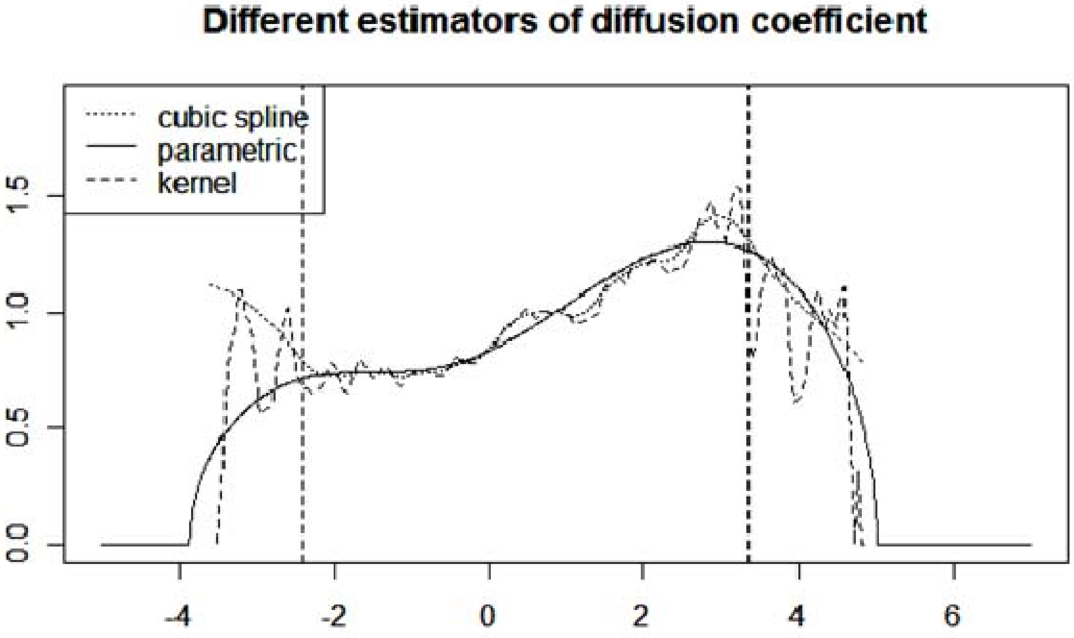
Basic result

Lemma: *Suppose that F is in the extreme domain of attraction of some GEV distribution G with shape parameter* $\xi <0$, *let* r_S *the common upper bound of F and G.*

We have the following behavior of a near the upper bound r_s ,

$$a^2(t) \approx \frac{-2b(r_s)(r_s - t)}{1 - \frac{1}{\xi}}$$

July Bordeaux Vertical lines 1% and 99% quantiles r= μ-σ/ξ



Density transition for the diffusion skeleton (DDC 80)

P(x,y)=A(x,y)L(x,y) with

$$\text{A(x,y)=}\frac{1}{a(y)\sqrt{2\pi}}\,\text{exp}-\frac{1}{2}((V(y)-V(x)^{2}+H(y)-H(x))$$

$$\text{V(y)= }\int_{0}^{y}\frac{du}{a(u)} \qquad \text{C(V(y))=}\frac{b(y)}{a(y)}-a'(y) \qquad \text{H(y)= }\int_{0}^{V(y)}C(u)du$$

$$L(x,y)=E\int_{O}^{1}g((1-u)s(x)+us(y))+B(u))du$$

B brownian bridge and $g=-\frac{1}{2}C^{2}+C$

Bivariate distribution and transition density

- In the previous formula the behaviour of the **density transition**, the **invariant marginal density** and so the **bivariate distribution** can be obtained as **x and y tend to r** , only the term in **H** is important and the transition satisfies following formula and allows to study asymptotic independence ans index of clusterisation (the marginal invariant distribution is given by that, well known , of the diffusion process)

$$p(x,y) \approx (y-x)^{-\frac{1}{\xi}}$$

Summary of the use of continuous time process

Statistics (blocks and GEV, treshold POT) $\longrightarrow \xi < 0$ $\longleftrightarrow$ **boundness** $\longleftrightarrow$ for continuous time diffusion 1)**marginals and transitions** have the same shape parameter and 2)

$a(x) \approx \dfrac{b(r)\sqrt{r-x)}}{(1-\frac{1}{\xi})}$ as x tends to r this implies $\longrightarrow$ for the

discrete observed Markov chain : the tail (asymptotic) of the transition and of the marginal are known (i.e the bivariate distribution)

$p(x,y) \approx (y-x)^{-\frac{1}{\xi}}$ as y and x tend to the boundary r with y>x

This implies the possibility of a study for asymptotic independence index extreme etc

Second approximation : Euler scheme

- The skeleton even with the previous approximation is **difficult to manage in statistics** and not usefull for **simulation**
- First order scheme is a **FARCH** process: it is a stationary process geometrically ergodic (need some care) where Δ (here 1) is the mesh of observations

$$Y_{k\Delta} = b(X_{k\Delta}) + a(X_{k\Delta})\varepsilon_k$$

$$p(x, y)dy = \frac{1}{\sqrt{2\pi a(x)}} \exp\left(-\frac{1}{2}\frac{(y-b(x))^2}{a(x)^2}\right)dy \text{ for } x \in J$$

$$P(x, b(x)) = 1 \text{ for } x \in J^C \text{ and } P(x, y) = 0 \text{ for } x \in J^C \text{ and } y \neq b(x)$$

Discrete approximations as misspecifications

FARCH approximated discretization ⇔ misspecification

is cyclo stationary , geometrically ergodic, no density for $a(.,j)$ =0

out of a bounded interval.

- Misspecified process : marginals support whole **R**

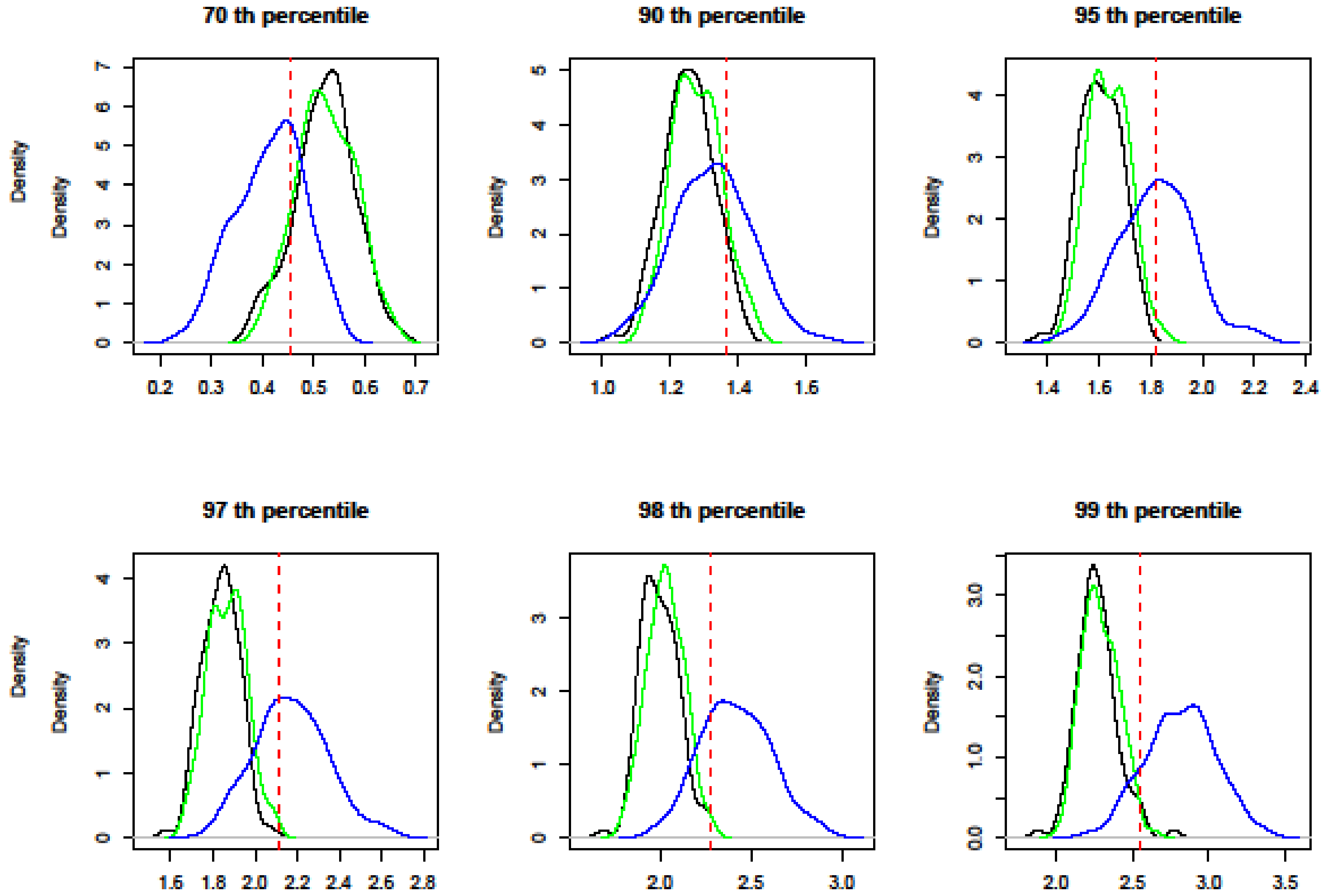
The observed distributions are supported by (r_1, r_2)

Exact data ⇔ conditional estimation ⇔ THE OBSERVATIONS ARE

IN THE FINITE SUPPORT (r_1, r_2) OF $\boldsymbol{a}$ ⇔ Estimation made with all data in (r_1, r_2) ⇔ $a(y) = 0$ if $y \notin (r_1, r_2)$

Simulation model with Gaussian noise has a weak percentage ($<10^{-3}$) out of I for 50 years ⇔ Probability of large excursion

Proof: m.l.e for misspecified models



Quantiles (red vertical lines) for July of Y_t and their distributions built from the simulations of different models: in black, model with constant a, in green, model with $a = f(t)$, in blue, model with $a(t, Y_{t-1})$

Embedding and seasonalities

The reduced process Y has 0 mean and variance 1;
nevertheless it remains stochastic periodicity
Let us look only to the year seasonality.
The drift ***b(x)*** is very close to linearity ***bx*** even in the extreme part and slowly varying with the season
The diffusion coefficient ***a(x)*** is ***0*** out of and interval slowly varying with the months but it is quite linear between the quantiles 2% and 98% and of course taken positive its slope is positive in summer, negative in winter and important. The slope is weak in spring and autumn
The shape coefficient has slow variations
It is not possible to do a complete "deseasonalisation" of a periodic dynamic
To use previous results for stationary process and specifically foe extremes, the best is to use an **imbedding** of the discrete time chain of observation in order to use the previous results;
This is always possible and can have a physical interpretation
Thus problems of seasonality are difficult to take in account for extremes (as well for rough treatment as for preprocessing).

Conclusion

Use of extremes is based on probability approximations intrinsically difficult
Time climatic series require high level of care. Stochastic analysis after statistical pre processing is an interesting framework. **Local variance is depends on the state and drives the extremes behaviour**. The shape of the conditional mean is no important. We think that even when a direct treatment is done for extremes, it is necessary to study the reduced process to « qualify the direct work. »
Statistical evidence depends on (the amount) data, often for extremes behaviour and specifically for that of reduced variables it seems depends on" feelings"