Quantile regression: an approach based on GEV distribution and machine learning

Lucien M. Vidagbandji, Laurent Amanton, Alexandre Berred, Cyrille Bertelle

FRCCS 2023

French **R**egional **C**onference on **C**omplex **S**ystems

Le Havre, France

31 May - 02 June

Book of Abstracts

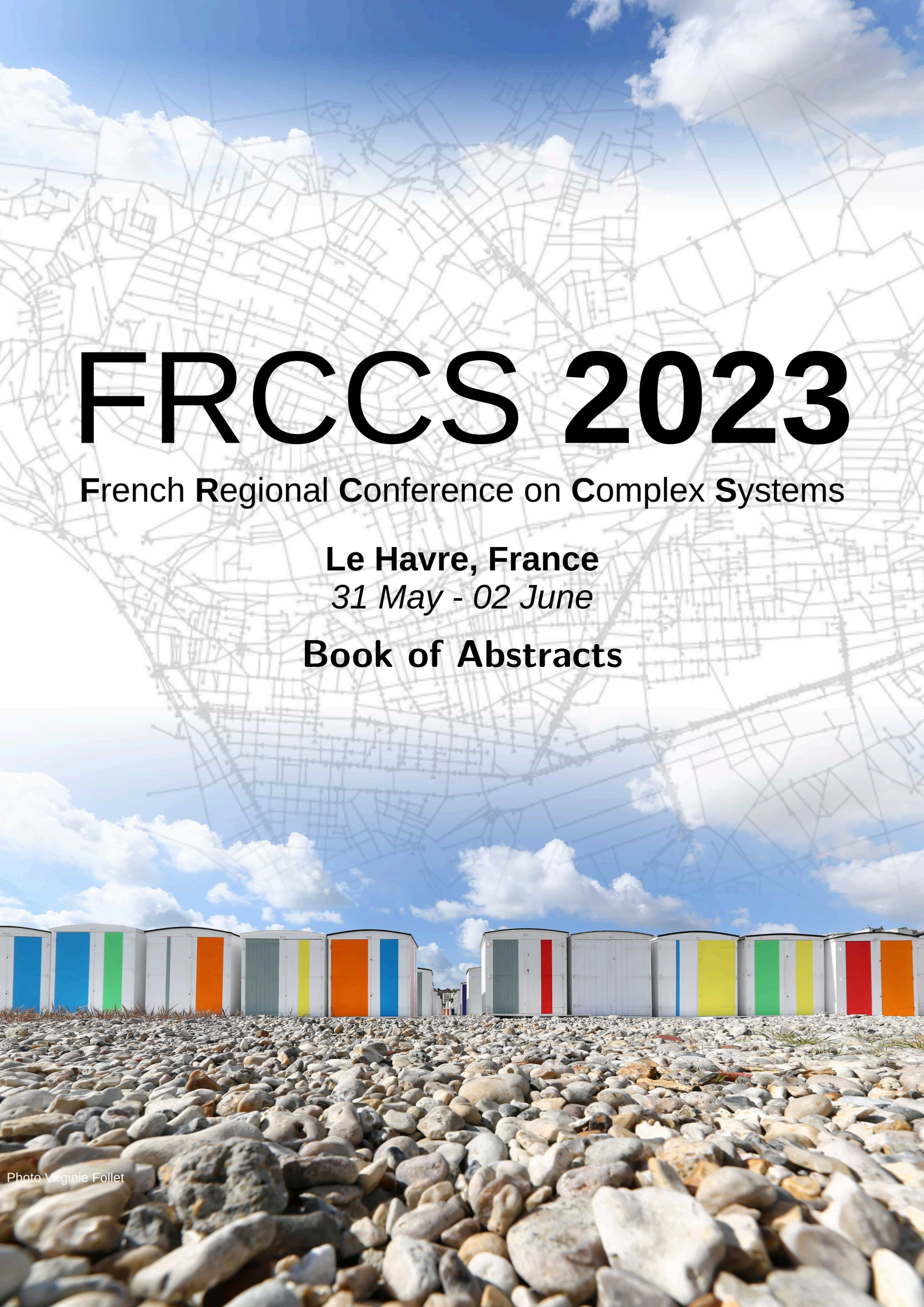

Photo Virginie Follet

Quantile regression: an approach based on GEV distribution and machine learning

Lucien M. Vidagbandji · Laurent Amanton · Alexandre Berred · Cyrille Bertelle

1 Introduction

Many complex natural or artificial systems are characterised by critical evolutions that are difficult to predict and control. It is therefore relevant to look at the risk assessment of extreme events, which requires an accurate estimation of high quantiles that can sometimes exceed the observation range. Thanks to the asymptotic results of the extreme value theory, extrapolation beyond the data range is possible. When the extreme event depends on some characteristic variables, quantile regression is a statistical tool allowing to have these extreme quantiles conditional to the dependent variables. The classical methods available for modelling conditional quantile regression fail mainly when the structure between the variable representing the extreme event and the characteristic variables is complex or when the size of the characteristics is large. Recent literature has seen the development of machine learning approaches for the analysis of extreme events. These algorithms are used to produce fast predictions in the context of extremes for high dimensional data, moreover they are able to capture much more complex structures in the data. In this work, statistical learning would be used in this context.

Two approaches are mainly used in extreme value theory: the Peaks-Over-Threshold (POT) approach and the block maxima (BM) approach. Existing methods combining extreme value theory and machine learning for modeling extreme quantile regression have used the POT approach. In contrast, the BM method is more suitable for modeling extremes in several domains, for example in hydrology and meteorology (Coles

L. M. Vidagbandji · A. Berred
Le Havre Normandy University, LMAH, Le Havre, France,
Tel.: +33773446863
E-mail: mahutin-lucien.vidagbandji@univ-lehavre.fr
alexandre.berred@univ-lehavre.fr

C.Bertelle · L. Amanton
Le Havre Normandy University, LITIS, Le Havre, France,
E-mail: laurent.amanton@univ-lehavre.fr
cyrille.bertelle@univ-lehavre.fr

et al. (2001)[2]) . Dombry (2013)[3] justified the use of the maximum likelihood method for the BM approach, and comparative studies of the two approaches are performed by Ana Ferreira and Dombry (2019)[4] and by Bücher *et al.* (2018)[1]. In this work we will propose a conditional extreme quantile regression model by combining the BM approach of extreme value theory and machine learning methods. According to the block maxima approach, we approximate the conditional extreme distribution by the generalized extreme value distribution (GEV) whose parameters depend on the features of the model. These parameters will be estimated using statistical learning methods.

2 Quantile regression

In classical regression, for example linear regression, we are looking for a relationship of the type $Y = f(X) + \varepsilon$ between a variable Y called response variable depending on certain characteristics $X \in \mathbb{R}^p$ and the goal of being able to estimate the value of y for a given x again. Classical regression predicting the conditional mean $E[Y|X = x]$ while the mean summarizes the behavior of $Y|X = x$ in the center of its distribution, applications in the field of risk assessment require knowledge of the tail of its distribution. Quantile regression provides information on this and allows a richer description since it is interested in all the conditional distributions of the variable of interest and not only in its average. It is shown that this model allows better estimation even in the presence of extreme values.

The conditional quantile function of order τ is defined by:

$$\mathscr{Q}_\tau(y|X) = \inf\{y : F_{Y|X}(y) \geq \tau\}$$

And the quantile regression model (Koenker (1978)[6]) is given by: $Y = \mathscr{Q}_\tau(y|X) + \varepsilon$ with:

$$\mathscr{Q}_\tau(y|X = x) = \arg\min_{q\in\mathbb{R}} E(\rho_\tau(Y - q)|X = x)$$

and $\rho_\tau(c) = c(\tau - \mathbb{I}_{c<0})$

The quantile regression is based on the estimation of:

$$\mathscr{Q}_\tau(y|X = x) = F^{-1}_{Y|X=x}(\tau) \tag{1}$$

for all $x \in \mathbb{R}^p$.

3 Quantile regression using the GEV approach

Several parametric, non parametric and machine learning based models are proposed for estimating this quantity, but these methods encounter difficulty mainly when the structure between Y and X is complex or the dimension of the dependent variable X is large. To circumvent these problems and perform modern applications with complex data, machine learning methods are useful because of their modeling flexibility and robustness in higher dimensions. Several methods have been proposed in the last few years using the combination of machine learning methods and extreme value theory

following the POT approach, we can cite the works of: Velthoen *et al.* (2021)[8] using the gradian boosting method, Gnecco *et al.* (2022) [5] using the random forest and Pasche *et al.* (2022)[7] using neural networks.

Since we are interested in extreme quantiles, we will use the block maxima method by approximating the conditional distribution $Y|X = x$ of the equation 1 by a generalized extreme value distribution (GEV). We will thus have a GEV distribution whose parameters depend on the characteristics x.

Setting $\Theta(x) = (\varepsilon(x), \sigma(x), \mu(x))$, this distribution is given by:

$$G(x;\Theta(x)) = \exp\left(-\left(1+\varepsilon(x)\frac{x-\mu(x)}{\sigma(x)}\right)_+^{-\frac{1}{\varepsilon(x)}}\right) \text{ if } \varepsilon(x) \neq 0$$

The conditional quantile of GEV is given for τ close to 1 by:

$$\mathscr{Q}_\tau(y|X=x) = \mu(x) + \frac{\sigma(x)}{\varepsilon(x)}\left(\left(\ln(\frac{1}{\tau})\right)^{-\varepsilon(x)} - 1\right) \text{ if } \varepsilon(x) \neq 0 \tag{2}$$

and

$$\mathscr{Q}_\tau(y|X=x) = \mu(x) + \sigma(x)\ln(-\ln(\tau)) \text{ if } \varepsilon(x) = 0 \tag{3}$$

The estimation of the conditional quantile, therefore, amounts to an estimation of $\Theta(x)$. Referring to the works of Gnecco *et al.* (2022)[5] , Pasche *et al.* (2022)[7] and Velthoen *et al.* (2021) [8], the order from which the quantile would be extreme would be considered as $\tau_0 = 0.8$ and so we consider $\tau \in [\tau_0, 1[$ in the expressions (2) and (3).

We will estimate the GEV parameter vector by $\hat{\Theta}(x) = (\hat{\varepsilon}(x), \hat{\sigma}(x), \hat{\mu}(x))$ using the maximum likelihood method. This choice in the BM approach is based on the work of Dombry (2013)[3], Ana Ferreira and Dombry (2019)[4] and Bücher *et al.* (2018)[1]. We have:

$$\hat{\Theta}(x) = \arg\max_{\varepsilon,\sigma,\gamma} L_n(x;\varepsilon,\sigma,\gamma) \tag{4}$$

where $L_n(x;\varepsilon,\sigma,\gamma)$ is the likelihood associated to the sample of maximums per block. The structure between Y and X being complex and the size of the features is large in our context, so the resolution of (4) will be performed using statistical learning algorithms and then the conditional quantile estimate is given by:

$$\hat{\mathscr{Q}}_\tau(y|X=x) = \hat{\mu}(x) + \frac{\hat{\sigma}(x)}{\hat{\varepsilon}(x)}\left(\left(\ln(\frac{1}{\tau})\right)^{-\hat{\varepsilon}(x)} - 1\right) \text{ if } \hat{\varepsilon}(x) \neq 0 \tag{5}$$

and

$$\hat{\mathscr{Q}}_\tau(y|X=x) = \hat{\mu}(x) + \hat{\sigma}(x)\ln(-\ln(\tau)) \text{ if } \hat{\varepsilon}(x) = 0 \tag{6}$$

This contribution questions the way to approach conditional quantile estimates characterising extreme events. Our work is part of a recent approach that links extreme value theory and machine learning. We do not yet address any application

to real data sets. We are exploring an application concerning the impact of climate change on port infrastructures. This topic requires careful handling of the data available to date, which is a work in progress.

References

1. Axel Bücher et Chen Zhou, A Horse Race between the Block Maxima Method and the Peak–over–Threshold Approach, Statistical Science, 36(3) (2021)
2. Stuart Coles, An Introduction to Statistical Modeling of Extreme Values. Springer, London (2001)
3. Clément Dombry, Maximum likelihood estimators for the extreme value index based on the block maxima method, arXiv preprint arXiv:1301.5611, (2013)
4. Clément Dombry et Ana Ferreira, Maximum likelihood estimators based on the block maxima method, The annals of Statistics (2019).
5. Nicola Gnecco, Edossa Merga Terefe et Sebastian Engelke, Extremal Random Forests, arXiv:2201.12865 [stat], (2022)
6. Roger Koenker et Gilbert Bassett, Regression Quantiles, Econometrica, 46(1):33 (1978)
7. Olivier C. Pasche et Sebastian Engelke, Neural Networks for Extreme Quantile Regression with an Application to Forecasting of Flood Risk, arXiv preprint arXiv:2208.07590, (2022).
8. Jasper Velthoen, Clément Dombry, Juan-Juan Cai et Sebastian Engelke, Gradient boosting for extreme quantile regression, arXiv preprint arXiv:2103.00808, (2021)