Etude de l'erreur relative d'approximation des quantiles extrêmes.

Clément ALBERT [1], Anne DUTFOY [2] & Stéphane GIRARD [1]

[1] *Equipe MISTIS, Inria Grenoble Rhône-Alpes & Laboratoire Jean Kuntzmann, 655, avenue de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.*
[2] *EDF R&D département Management des Risques Industriels, 7 boulevard Gaspard Monge - 91120 Palaiseau*
clement.albert@inria.fr, anne.dutfoy@edf.fr, stephane.girard@inria.fr

Résumé. Une problématique générale en analyse des valeurs extrêmes consiste, à partir d'une loi de valeurs extrêmes ajustée sur des données, à déterminer les quantiles extrêmes de période de retour centenale ou millénale. Dans cette communication, nous quantifions les limites d'extrapolation associées à ces estimations de quantiles extrêmes. Pour ce faire, nous nous plaçons dans le cadre de la méthode des maxima par blocs et nous étudions le comportement de l'erreur relative d'approximation associée à un estimateur des quantiles dédié au domaine d'attraction de Gumbel. Nous donnons des conditions nécessaires et suffisantes à la convergence de l'erreur vers zéro et le cas échéant un équivalent de cette dernière. Nous montrons que la qualité de l'extrapolation est grandement dépendante de la loi dont sont issues les données.

Mots-clés. Théorie des valeurs extrêmes, estimation de quantiles extrêmes, propriétés asymptotiques, risques environnementaux.

Abstract. In the risk management context, the extreme value methodology consists in estimating extreme quantiles - one hundred years return period or more - from an extreme-value distribution adjusted on data. In this communication, we quantify the extrapolation limits associated with these extreme quantile estimations. To this end, we focus on the framework of the block maxima method and we study the behaviour of the relative approximation error of a quantile estimate dedicated to the Gumbel attraction domain. We give necessary and sufficient conditions for the error to converge towards zero and, if so, we provide a first order approximation of the latter. We show that extrapolations can be greatly limited depending on the data distribution.

Keywords. Extreme-value theory, extreme quantiles estimation, asymptotic properties, environmental risks.

1 Introduction

La R&D d'EDF utilise la théorie des valeurs extrêmes pour effectuer de nombreuses études statistiques d'évènements extrêmes à partir de relevés de variables météorologiques (température, débit, vitesse de vent, ...). Ces études servent à dimensionner les ouvrages face aux agressions météorologiques, de type inondation, tempête ou encore sécheresse. Elles consistent, à partir d'une loi de valeurs extrêmes ajustée sur des données, à déterminer les quantiles extrêmes de période de retour centenale ou millénale. Mais quelle crédibilité accorder à ces extrapolations ? Si certains travaux proposent des règles empiriques d'extrapolation limite [2], nous nous proposons dans cette communication, de définir des règles mathématiques issues de l'analyse de la convergence vers les lois de valeurs extrêmes.

Les travaux que nous présentons se placent dans le cadre où les données sont issues d'une loi F appartenant au domaine d'attraction de Gumbel ($F \in DA(Gumbel)$).

Dans un premier temps, nous exposons un résultat qui établit la consistance relative de l'estimateur d'un quantile extrême issu de ce domaine d'attraction. Dans le cas où la consistance est vérifiée, un second résultat établit un équivalent de l'erreur asymptotique d'approximation associée à l'estimateur en question.

2 Estimation de quantiles extrêmes dans le cadre d'une approche de type maxima par blocs.

Notre étude se place dans le cadre de l'estimation de quantiles extrêmes issus d'une loi inconnue $F \in DA(Gumbel)$. Etant donné un n-échantillon, un quantile extrême est le $(1-p)$-ème quantile x_p de la loi F, avec $p \to 0$ quand $n \to +\infty$.

La méthode des maxima par blocs consiste à diviser la période d'observation des données en blocs de taille égale m et à se focaliser sur l'étude des maxima associés à chaque bloc. Les nouvelles observations créées de la sorte suivent - sous des conditions de domaine d'attraction - approximativement une loi des valeurs extrêmes G. Des méthodes statistiques paramétriques sont ensuite appliquées à cet ensemble de maxima.

Cette méthode repose sur le théorème suivant :

Théorème (Fisher and Tippett, 1928) : Considérons $X_1, X_2, ..., X_m$ une suite de variables aléatoires indépendantes suivant une loi $F \in DA(Gumbel)$ et posons $X_{m,m} = max\,(X_1, X_2, ..., X_m)$. Il existe des suites $\{a_m > 0\}$ et $\{b_m\}$ telles que :

$$\mathbb{P}\left(\frac{X_{m,m} - b_m}{a_m} \leq x\right) \to G(x), \quad m \to \infty,$$

avec

$$G(x) = \exp\left\{-\exp\left(-x\right)\right\}, \; x \in \mathbb{R}. \tag{1}$$

Interpréter la limite dans ce théorème comme une approximation pour de grandes valeurs de m suggère l'utilisation de G pour modéliser la loi des maxima d'une longue séquence.

Une approximation du quantile extrême x_{p_m} de F est alors obtenue par une approximation d'ordre un du logarithme en inversant l'équation (1) et en utilisant le fait que $\mathbb{P}(X_{m,m} <$

$x) = F^m(x) \approx G\left(\dfrac{x-b_m}{a_m}\right)$:

$$\tilde{x}_{p_m} = b_m - a_m \log(mp_m).$$

Par la suite, on s'intéresse à l'erreur relative d'approximation ϵ_{app_m} de x_{p_m} par $\tilde{x}_{p_m}$:

$$\epsilon_{app_m} = \frac{x_{p_m} - \tilde{x}_{p_m}}{x_{p_m}}.$$

3 Résultats asymptotiques

Nous définissons le taux de hasard cumulé $H(x) = -\log(1 - F(x))$ et nous supposons que H est une fonction strictement croissante et que $H \in C^2(\mathbb{R})$. De plus, nous considérons p_m de telle sorte que $\dfrac{\log(1/p_m)}{\log m} = \tau_m$, avec $\tau_m \to \tau \geq 1$, ou de manière équivalente $p_m = m^{-\tau_m}$.

Nous définissons également les fonctions $K_i(x) = \dfrac{x^i \left(H^{-1}\right)^{(i)}(x)}{H^{-1}(x)}$, $i \in \{1, 2\}$ ainsi que $l_1 = \lim\limits_{x \to +\infty} K_1(x)$.

Le premier résultat donne une condition nécessaire et suffisante pour que l'erreur relative d'approximation tende vers zéro. Cette condition lie le comportement de la queue de la loi (via K_2) avec la vitesse à laquelle p_m tend vers zéro (via la différence $\tau_m - 1$).

Théorème 1 : Supposons que K_1 soit à variation régulière d'indice θ_1 ($K_1 \in RV_{\theta_1}$), $\theta_1 \leq 1$, $K_2 \in RV_{\theta_2}$, $\theta_2 \leq 2$ et $\tau \geq 1$. Alors :

$$\epsilon_{app_m} \overset{m\to+\infty}{\longrightarrow} 0 \iff (\tau_m - 1)^2 K_2(\log m) \overset{m\to+\infty}{\longrightarrow} 0.$$

L'hypothèse de variation régulière $K_1 \in RV_{\theta_1}$ a été introduite initialement par [1]. Une discussion sur les types de lois satisfaisant cette hypothèse peut également y être trouvée.

Notre second résultat donne un équivalent de l'erreur d'approximation relative en fonction de la limite l_1 de K_1.

Théorème 2 : Sous les mêmes hypothèses que le Théorème 1, on a, quand $m \to +\infty$:

1. si $l_1 \in \{0, 1\}$ alors $K_2(\log m) \to 0$, $\theta_2 \leq 0$ et
$$\epsilon_{app_m} \sim (\tau_m - 1)^2 K_2(\log m) \tau^{-l_1} g(\tau, \theta_2 + l_1) \to 0, \ \forall \tau \geq 1.$$
2. si $l_1 \in [0, +\infty[$, $l_1 \notin \{0, 1\}$, alors $K_2(\log m) \to l_1(l_1 - 1)$, $\theta_2 = 0$ et
$$\epsilon_{app_m} \sim (\tau_m - 1)^2 l_1(l_1 - 1) \tau^{-l_1} g(\tau, l_1) \to 0 \ ssi \ \tau = 1.$$

3. si $l_1 = +\infty$ alors $K_2(\log m) \to +\infty$ et

$$\epsilon_{app_m} \sim (\tau_m - 1)^2 K_2(\log m)\Psi\left((\tau_m - 1)\sqrt{K_2(\log m)}\right) \to 0 \ ssi \ (\tau_m - 1)^2 K_2(\log m) \to 0.$$

On a défini

$$\Psi(t) := \int_0^1 x e^{-tx}\,\mathrm{d}x \quad et \quad g(\tau,\theta) := \frac{\tau^\theta - \theta(\tau-1) - 1}{\theta(\theta-1)(\tau-1)^2}, \quad \theta \notin \{0,1\}, \ \tau > 1, \ t \geq 0.$$

Ce deuxième Théorème nous permet de distinguer 3 familles de lois. En premier lieu, les lois vérifiant $l_1 \in \{0,1\}$, pour lesquelles il y a convergence vers zéro de l'erreur sans qu'aucune condition ne soit requise sur τ. Puis viennent les lois vérifiant $l_1 \in [0,+\infty[$, $l_1 \notin \{0,1\}$. Pour ces lois, la convergence vers zéro nécessite $\tau = 1$. Enfin le cas le plus restrictif est le cas où $l_1 = +\infty$. Dans ce dernier cas, τ doit être égal à 1, mais on a en plus besoin d'une condition quant à la vitesse de convergence de τ vers 1. En effet, la condition $(\tau_m - 1)^2 K_2(\log m) \to 0$ implique $\tau = 1$. Si cette condition tient, alors $\epsilon_{app_m} \sim \dfrac{1}{2}(\tau_m - 1)^2 K_2(\log m)$.

Ces résultats seront illustrés numériquement lors de l'exposé. Dans un premier temps, l'exactitude des équivalents sera évaluée sur des simulations. Dans un second temps, nous montrerons comment utiliser ces résultats en pratique pour contrôler l'erreur relative d'extrapolation.

Références

[1] de Valk, C. (2016). Approximation of high quantiles from intermediate quantiles. *Extremes*, 19(4), 661-686.

[2] MétéoFrance, Direction de la Climatologie (2014). Durées de retour de précipitations extrêmes. `http://pluiesextremes.meteo.fr/media/doc/Cartes_reseau/Fiche_methode_durees_retour.pdf`.