ESTIMATION ET COMPARAISON DE NIVEAUX DE RETOUR POUR LES VITESSES EXTREMES DES VENTS

Henri Klajnmic

EDF R & D
Département ICAME
Groupe Statistique et Outils d'Aide à la Décision
1, avenue du Général de Gaulle
92141 Clamart Cedex
Henri.Klajnmic@edf.fr

Résumé

Plusieurs modélisations des valeurs extrêmes (ici la vitesse maximum journalière du vent) sont possibles :

- par les lois des valeurs extrêmes généralisées (Fréchet, Gumbel et Weibull) dont on peut estimer les paramètres par maximum de vraisemblance ou par moments pondérés. Les hypothèses d'application de ces modèles ne sont pas toujours vérifiées. En dehors de la loi de Weibull, estimer des niveaux de retour à plus de 50 ans donne des valeurs et des intervalles de confiance inexploitables. Selon les stations météo, les lois obtenues peuvent être différentes.
- par la méthode des dépassements de seuil (POT) et la loi de Pareto généralisée. Le choix du seuil n'est pas simple et les hypothèses contraignantes.

On a pu constater des résultats différents selon les méthodes employées pour le même phénomène. Cela montre que cette modélisation, nécessaire en raison du faible nombre de données extrêmes, implique une mise en oeuvre délicate et une grande prudence dans l'exploitation des résultats, une validation exacte ne semblant pas possible.

Mots-clés : valeurs extrêmes, lois de Fréchet, Gumbel, Weibull, maximum de vraisemblance, moments pondérés, dépassement de seuil, loi de Pareto généralisée, vraisemblance profil, méthode delta.

Abstract

Several modelling of extreme values (here maximum daily wind speed) are possible:

- Generalized Extreme Value Distributions (Fréchet, Gumbel or Weibull) whose parameters can be estimated either by maximum likelihood or by probability weighted moments. Hypotheses required for this model are not always verified. Except for the Weibull distribution, the estimation of return levels above 50 years gives unexploitable confidence intervals. Depending on the meteorological stations, the distributions may be different.
- Peak Over Threshold method and Generalized Pareto Distribution. But to determine the threshold is not so easy and the required hypotheses restricting.

Depending on the methods different results are noticed for the same phenomenon. This shows that this modelling, imperative due to very little extreme data, implies delicate statistical investigations and to be careful when using the results, an exact validation seeming not possible.

Keywords: extreme values, Fréchet distribution, Gumbel distribution, Weibull distribution, Maximum Likelihood, Probability Weighted Moments, Peak Over Threshold, Generalized Pareto Distribution, profile likelihood, delta method

Introduction

La théorie moderne des valeurs extrêmes est apparue entre 1920 et 1940, due à M. Fréchet (1927), R. A. Fisher et L. H. C. Tippett (1928), E. J. Gumbel (1935) et enfin B. V. Gnedenko (1943). On peut citer aussi les travaux antérieurs de W. E. Fuller (1914), A. A. Griffiths (1920) et L. von Bortkiewicz (1922). Les problèmes de valeurs extrêmes ont d'abord concerné les hauteurs des crues ainsi que le génie civil.

On notera :

- $M_n = \max\{X_1, \cdots, X_n\}$
- $(X_i)_{i \in I}$ sont des variables aléatoires indépendantes et identiquement distribuées

La vitesse maximum du vent est déterminée sur un petit intervalle de temps et on en recueille le maximum journalier, donc peut être considérée comme « un maximum de quelque chose ».

On peut établir le théorème (fondamental, dû à Gnedenko) suivant :

S'il existe des suites de constantes $\{a_n > 0\}$ et $\{b_n\}$ telles que $\Pr\{(M_n - b_n)/a_n \le z\} \to G(z)$, G étant une fonction de répartition non dégénérée, alors G ne peut appartenir qu'à l'une des trois lois (dites GEV) : Weibull (à support borné supérieurement), Gumbel et Fréchet (à support non borné).

Les lois de Weibull, Gumbel et Fréchet ont respectivement pour fonctions de répartition,

- $a > 0, \alpha > 0, -\infty < b < +\infty$
- $G(z) = \begin{cases} \exp\left\{-\left[-\dfrac{(z-b)}{a}\right]^{\alpha}\right\}, z < b \\ 1, z \ge b \end{cases}$
- $G(z) = \exp\left\{-\exp\left[-\dfrac{(z-b)}{a}\right]\right\}, -\infty < z < +\infty$
- $G(z) = \begin{cases} 0, z < b \\ \exp\left\{-\left[\dfrac{(z-b)}{a}\right]^{-\alpha}\right\}, z \ge b \end{cases}$

Le premier objectif, lorsque l'on étudie une série de maximums, est de déterminer la loi et d'en estimer les paramètres. On se pose alors la question suivante : quelle vitesse de vent revient tous les dix (décennale), vingt (vicennale), cinquante ans (« cinquantennale »), cent ans (centennale) …? Ou bien, ayant observé x m/s, une telle vitesse revient tous les … ?

Le niveau de retour $\mathbf{z_p}$ d'un phénomène extrême est défini comme le quantile de la distribution $G(z_p) = P\{Z \le z_p\} = 1 - p$. Cette valeur est donc dépassée lors d'une année quelconque avec une (petite) probabilité p. En utilisant la loi géométrique (voir annexe), ceci s'interprète comme la

valeur d'une variable (hauteur d'une crue, vitesse de vent ...), qui revient en moyenne tous les $x = 1/p$ années, par exemple le niveau de retour à 50 ans que nous utiliserons dans cet article.

Nous avons étudié la vitesse maximale journalière du vent entre 1981 et 2000 sur 32 stations synoptiques de Météo-France.

En utilisant les lois des valeurs extrêmes, on peut estimer les paramètres de ces lois soit par maximum de vraisemblance (Coles (2001)), soit par les moments pondérés (Hosking (1985)). Les estimations obtenues sont voisines. Il est plus facile de calculer les intervalles de confiance sur les paramètres et sur les niveaux de retour avec le maximum de vraisemblance (méthode delta) qu'avec les moments pondérés. On peut utiliser de plus la vraisemblance profil, qui donne des intervalles non symétriques.

Avec nos 32 stations météo, on a constaté que l'on obtenait les trois types de lois. Lorsque l'on a une loi de Gumbel et surtout de Fréchet, les niveaux de retour à horizon supérieur à 50 ans deviennent très grands et inexploitables. La méthode est acceptable avec la loi de Weibull à support borné supérieurement. Une autre méthode de modélisation est possible : le dépassement de seuil.

Cette dernière technique a été développée par J. Pickands, et A.C. Davison et R. L. Smith (Coles (2001), Davison (2003), Embrechts (2001)) : pour un seuil assez grand, la loi conditionnelle des dépassements du seuil sachant que l'on est au-dessus du seuil est une loi de Pareto généralisée (GPD), toujours en supposant des hypothèses d'indépendance. On a le résultat suivant :

Pour un seuil u assez grand, sous les hypothèses du théorème sur les lois de valeurs extrêmes, alors la fonction de répartition de $X - u$ sachant $X > u$ (loi conditionnelle) est approximativement :

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \tilde{\sigma} = \sigma + \xi(u - \mu), \{y : y > 0, (1 + \xi\, y/\tilde{\sigma}) > 0\}$$

Le paramètre ξ est le même que celui de la modélisation par les lois des valeurs extrêmes. Le principal problème est de choisir le seuil et les méthodes proposées (Coles (2001)) ne sont pas simples à mettre en oeuvre. On a constaté que les niveaux de retour obtenus ne sont pas les mêmes selon la modélisation employée (GEV ou GPD), et que la vraisemblance profil n'est pas toujours applicable.

Enfin, tout ce que nous utilisons pour des maximums pourrait être appliqué de manière similaire pour des minimums (températures, sécheresse …).

Les données

Nous disposons de vingt ans de vitesses maximales journalières de vent (du 1er janvier 1981 au 31 décembre 2000) sur 32 stations synoptiques de Météo-France ne constituant pas un échantillon ressemblant à l'ensemble des stations. Il y a une grande diversité de situations : on sait que selon les régions les conditions de vent et les vitesses peuvent être très différentes. Les vitesses maximales peuvent varier du simple au double (60 m/s à la Pointe du Raz, 55 m/s à la Pointe de Chassiron, contre 29 m/s à Ambérieu, 31 m/s à Besançon). 60 m/s représente 60x3.6=216 km/h (!) et 29 m/s 104.4 km/h. Il y a peu de valeurs manquantes (de 0 à 5 % des données). Il y a rarement des vitesses journalières nulles, ce qui aurait pu être une erreur de mesure. Les coefficients d'asymétrie sont tous positifs de 0.55 à 2.25. Les coefficients d'aplatissement sont positifs et vont de 0.32 à plus de 7... Enfin, la tempête de fin décembre 1999 ne correspond pas nécessairement à la plus forte valeur observée : il y a eu des événements plus extrêmes en 1987 et 1990 et les résultats varient selon la

région. Dans la suite, nous nous limiterons à trois stations que nous appellerons A, C1 et C2.

Logiciels utilisés

Il existe quelques logiciels pour pouvoir travailler sur les valeurs extrêmes sans tout programmer soi-même. Nous avons utilisé les fonctions proposées par S. Coles (2001) (ismev) et celles très voisines proposées par A. McNeil (evis4). Ces fonctions, initialement écrites en S (S-Plus est un produit industriel diffusé par Insightful, dont nous avons utilisé la version 6.2), ont été portées sous R (logiciel libre ressemblant beaucoup à S : http://www.r-project.org/) par A. Stephenson (evir et ismev). Les documentations sont assez succinctes et il faut parfois regarder le code pour bien les utiliser, mais il ne semble pas y avoir pour l'instant de solution plus commode. SAS, par exemple, ne fournit pas de procédure pour travailler sur les valeurs extrêmes.

Modélisation par les valeurs extrêmes

Loi des valeurs extrêmes

Soit $M_n = \max\{X_1, \cdots, X_n\}$, ensemble de n variables aléatoires indépendantes de même fonction de répartition F et cherchons la distribution de probabilité de ce maximum. Le théorème suivant constitue la base de la modélisation :

S'il existe deux suites de constantes réelles $(a_n > 0)$ et (b_n) telles que : $P\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z)$, (convergence faible), G étant une **fonction de répartition non dégénérée**, alors G **ne peut appartenir qu'à l'une des trois lois** suivantes : Weibull, Gumbel ou Fréchet. Ces lois peuvent s'exprimer sous une forme **unifiée (dite de von Mises-Jenkinson)** :

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \quad \text{pour} \quad \left\{z : 1+\xi\left(\frac{z-\mu}{\sigma}\right) > 0\right\} \text{avec} \quad \xi \quad \text{(paramètre de forme)},$$

$\sigma > 0$ (paramètre d'échelle) et μ réel (paramètre de position). La loi de Gumbel s'obtient en faisant $\xi \to 0$.

L'inconvénient pour nos données est qu'il peut y avoir des épisodes venteux (donc l'indépendance n'est pas vraiment justifiée), présence de saisonnalité (mauvais temps en hiver par exemple et les vitesses ne suivraient pas la même loi). Regrouper par semaine, mois, trimestre ou année fait perdre de l'information. Dans le cas des vitesses de vent, c'est la répétition de vents forts qui peut causer des dégâts aux bâtiments et le maximum dans la semaine ou le mois fait perdre cette notion.

Niveau de retour

On définit le **niveau de retour** z_p par : $G(z_p) = 1 - p = \exp\left\{-\left[1+\xi\left(\frac{z_p-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$

ce qui nous donne, en inversant G : $z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right], \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, \xi = 0 \end{cases}$.

L'estimateur au maximum de vraisemblance de z_p est obtenu en substituant dans la formule les estimateurs au maximum de vraisemblance des trois paramètres des lois ξ , μ et σ (propriété d'invariance du maximum de vraisemblance).

Intervalle de confiance par la méthode delta

Le niveau de retour est une fonction (réelle) des trois variables ξ , μ et σ. On note ∇z_p le vecteur (gradient) des dérivées partielles de z_p par rapport à ξ , μ et σ :

$$\left[\frac{\partial z_p}{\partial\mu}, \frac{\partial z_p}{\partial\sigma}, \frac{\partial z_p}{\partial\xi}\right]^T = \left[1, -\xi^{-1}\left(1-y_p^{-\xi}\right), \sigma\xi^{-2}\left(1-y_p^{-\xi}\right) - \sigma\xi^{-1}y_p^{-\xi}\log y_p\right],$$

en posant $y_p = -\log(1-p)$.

Alors on a une expression approchée pour la variance du niveau de retour :

$Var(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p$, V étant la matrice des variances-covariances des paramètres. On utilise ensuite la loi normale pour avoir un intervalle de confiance, par exemple $\left[\hat{z}_p \pm 1.96\sqrt{Var(\hat{z}_p)}\right]$.

Intervalle de confiance par vraisemblance profil

Cette méthode consiste à choisir un des paramètres (ici z_p) et à considérer les autres comme des paramètres de nuisance. Considérons un ensemble de valeurs du paramètre d'intérêt (le niveau de retour). Pour chacune de ces valeurs, fixée, on calcule le maximum de vraisemblance (par rapport aux autres paramètres) et on trace la courbe obtenue. L'horizontale située en dessous du maximum de cette courbe à une distance de $0.5\chi_1^2(1-\alpha)$ (quantile à $(1-\alpha)$% du χ^2 à un degré de liberté pour un paramètre à une dimension) coupe la courbe en deux points qui sont les extrémités de l'intervalle recherché. On peut obtenir des intervalles de confiance non symétriques, contrairement à la méthode delta.

Moments pondérés

Il y a une alternative au maximum de vraisemblance : les moments pondérés. S'ils permettent l'estimation des paramètres, avoir des intervalles de confiance exige des calculs compliqués.

Pour $r = 0,1,2,\cdots$, définissons $\beta_r = E\left(X\{F(X)\}^r\right)$. Si $r = 0$, c'est la moyenne de la variable aléatoire X de fonction de répartition $F(x)$. En utilisant l'échantillon x trié par ordre croissant, on montre que $b_r = n^{-1}\sum_{j=1}^{n}\frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)}x_j$ est un estimateur sans biais de β_r.

Si la fonction de répartition est une loi GEV de la forme :

$F(x) = \exp\left[-\{1+\xi(x-\mu)/\sigma\}^{-1/\xi}\right]$ si $\xi \neq 0$ et $F(x) = \exp[-\exp\{-(x-\mu)/\sigma\}]$ si $\xi = 0$, on démontre

que l'on peut estimer les paramètres de cette loi par les formules suivantes (Hosking (1985)) :

$c = \frac{2b_1 - b_0}{3b_2 - b_0} - \frac{\log 2}{\log 3}$, $\hat{\xi} = -\left(7.8590c + 2.9554c^2\right)$,

$\hat{\sigma} = -\frac{(2b_1 - b_0)\hat{\xi}}{\Gamma\left(1-\hat{\xi}\right)\left(1-2^{\hat{\xi}}\right)}$, $\hat{\mu} = b_0 - \hat{\sigma}\frac{\left\{\Gamma\left(1-\hat{\xi}\right)-1\right\}}{\hat{\xi}}$.

Γ est la fonction gamma d'Euler $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,dt, x > 0$.

Le niveau de retour à N années s'écrit : $\hat{z}_N = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \left(-\log\left(1 - \frac{1}{365N}\right)\right)^{-\hat{\xi}}\right]$.

Nous prendrons $N = 50$. Les résultats sont proches de ceux au maximum de vraisemblance.

Résultats obtenus

Comme les vitesses de vent sont exprimées en nombres entiers de m/s, elles ont été « jittérisées » en ajoutant un nombre aléatoire suivant une loi uniforme sur [-0.5 +0.5]. Cette petite perturbation évite les nombreux ex æquo lorsque l'on utilisera les méthodes à dépassement de seuil. La nature de la loi change selon les stations. Les intervalles de confiance seront toujours à 95%.

Dans la démonstration du théorème sur la loi des valeurs extrêmes, on découpe les données en blocs de longueur fixe, c'est cette idée que nous retenons en ajustant les modèles sur les maximums mensuels, trimestriels ou en blocs de longueur quelconque. Nous allons voir que l'on obtient parfois des résultats assez surprenants.

Station C1

On obtient la loi de Weibull, la plus intéressante car elle suppose l'existence d'une vitesse limite $z_0 = \mu - \sigma/\xi$ que l'on ne peut dépasser. Pour la série initiale (une seule valeur manquante, avant jittérisation) : $(\mu = 12.18, \sigma = 3.74, \xi = -0.118)$. Les niveaux de retour à 50 ans sont : 33.3 [32.3 34.3] par la méthode delta et par vraisemblance profil : [33.1 35].

Considérons maintenant la série jittérisée. On obtient : $(\mu = 12.18, \sigma = 3.76, \xi = -0.119)$. Le niveau de retour à 50 ans est 33.9 et l'intervalle de confiance par la méthode delta [33 34.8]. La vitesse limite est estimée à 43.6. Il y a peu de différence avec la série initiale.

Considérons maintenant la série des maximums mensuels (20 ans : 240 mois). On obtient $(\mu = 21.45, \sigma = 2.81, \xi = -0.031)$, ξ n'est pas significativement différent de zéro. Avec la série trimestrielle $(\mu = 24.43, \sigma = 2.86, \xi = -0.065)$, ξ n'est pas significativement différent de zéro.

En utilisant la loi de Gumbel, on obtient pour la série mensuelle $(\mu = 21.40, \sigma = 2.78)$ avec pour niveau de retour à 50 ans 39.2 (intervalle de confiance par la méthode delta : [37.2 41.1]. Pour la série trimestrielle $(\mu = 24.33, \sigma = 2.82)$ avec pour niveau de retour à 50 ans 39.3 (intervalle de confiance par la méthode delta : [36.5 42.1]). La modélisation de la série annuelle donne la loi de Gumbel $(\mu = 28.57, \sigma = 2.02)$ avec pour niveau de retour à 50 ans 36.5 (intervalle de confiance par la méthode delta : [33.1 39.8]).

On notera que le niveau de retour obtenu avec la série journalière est très différent des niveaux de retour obtenus avec les séries mensuelle et trimestrielle. La série annuelle donne encore un résultat

différent.

Station A

Si l'on modélise la série journalière jittérisée (une seule valeur manquante), on trouve : $(\mu = 7.43, \sigma = 2.00, \xi = 0.227)$, c'est une loi de Fréchet (donc à support non borné vers les valeurs positives) car ξ est significativement différent de zéro. Le niveau de retour à 50 ans est 80.2 et l'intervalle de confiance par la méthode delta [71.7 88.7], ce qui paraît très élevé et peu réaliste. La cause de la loi de Fréchet est la présence de quelques vitesses journalières (moins de 10) très au-dessus des autres. La queue de distribution s'aplatit donc très lentement, les dix valeurs les plus élevées étant 28.45, 28.52, 28.67, 29.25, 29.78, 29.93, 30.95, 30.98, 31, 34.32.

Modélisons la série des maximums mensuels. On trouve $(\mu = 17.20, \sigma = 4.51, \xi = -0.162)$, on a une loi de Weibull et le niveau de retour à 50 ans devient 35.16 et l'intervalle de confiance par la méthode delta [31.3 39]. En utilisant la vraisemblance profil : [32.5 40.7]. Par la méthode des moments pondérés $(\mu = 17.16, \sigma = 4.60, \xi = -0.155)$ et le niveau de retour à 50 ans : 35.8. On peut voir en annexe la vraisemblance profil.

Modélisons la série des maximums trimestriels. On trouve $(\mu = 21.79, \sigma = 3.90, \xi = -0.225)$, on a une loi de Weibull et le niveau de retour à 50 ans devient 33.85 et l'intervalle de confiance par la méthode delta [31.3 36.6]. En utilisant la vraisemblance profil : [32.1 38.6]. Par la méthode des moments pondérés $(\mu = 21.80, \sigma = 3.99, \xi = -0.238)$ et le niveau de retour à 50 ans : 33.80.

Si maintenant, on utilise la série des 20 maximums annuels, on trouve la loi de Gumbel : $(\mu = 26.50, \sigma = 2.54)$ et le niveau de retour à 50 ans est 36.4 et un intervalle de confiance par la méthode delta [32.5 40.2].

On doit noter le changement de nature de la loi (on peut vérifier qu'il se produit avec une taille de bloc de l'ordre de 20) dû sans doute au manque de robustesse du modèle aux hypothèses. Les valeurs des niveaux de retour selon le degré d'agrégation sont voisines, mais il n'est pas sûr que l'on réponde vraiment à la question du niveau de retour à 50 ans d'un maximum journalier.

Station C2

La série comporte huit valeurs manquantes alors qu'il n'y en avait qu'une pour A et C1. Avec 7305 valeurs, 1% de valeurs manquantes représente 73 jours. Cela peut poser problème s'il manque un mois ou plus. C'est pour cette raison que nous exposons seulement le traitement sur trois stations. La vitesse maximum observée pour C2 est 43.89 m/s.

La modélisation de la série journalière jittérisée donne la loi de Gumbel $(\mu = 8.32, \sigma = 3.12)$, le niveau de retour à 50 ans est 38.92 (intervalle de confiance par la méthode delta [38.3 39.5]).

Pour la série mensuelle, on a également la loi de Gumbel $(\mu = 17.11, \sigma = 3.82)$, le niveau de retour à 50 ans passe à 41.5 (intervalle de confiance par la méthode delta [38.9 44.2]).

Avec la série trimestrielle, on a la loi de Gumbel $(\mu = 20.53, \sigma = 4.22)$, le niveau de retour à 50 ans passe à 42.9 (intervalle de confiance par la méthode delta [38.6 47.2]).

Enfin, avec la série annuelle, on a la loi de Gumbel $(\mu = 26.92, \sigma = 4.13)$, le niveau de retour à 50 ans passe à 43 (intervalle de confiance par la méthode delta [38.7 49.3]).

Conclusions

La modélisation donne des résultats pas toujours cohérents et on obtient des niveaux de retour parfois plutôt différents. On se rend compte également que selon les stations (et les régions), les situations météorologiques peuvent être différentes en raison d'une part des épisodes venteux et d'autre part du caractère « très extrême » de certaines tempêtes. Le fait d'agréger par mois etc...change la nature de l'estimation : estimer à partir d'observations journalières des niveaux de retour à 50 ans. Le maximum de vraisemblance semble plus commode d'application que les moments pondérés. Il serait plus judicieux d'utiliser la vraisemblance profil, les intervalles de confiance semblant asymétriques. Dans la suite, nous allons utilser les techniques plus récentes de dépassement de seuil et de processus ponctuel.

Résumé

Station C1, niveau de retour à 50 ans

33.3 (série initiale journalière, loi de Weibull), 33.9 (série journalière jittérisée, loi de Weibull), 39.2 (série mensuelle, loi de Gumbel) 39.3 (série trimestrielle, loi de Gumbel).

Station A, niveau de retour à 50 ans

80.2 (série journalière jittérisée, loi de Fréchet), 35.16 (série mensuelle, loi de Weibull, maximum de vraisemblance), 35.8 (série mensuelle, loi de Weibull, moments pondérés), 33.85 (série trimestrielle, loi de Weibull, maximum de vraisemblance), 33.8 (série trimestrielle, loi de Weibull, moments pondérés), 36.4 (série annuelle, loi de Gumbel, maximum de vraisemeblance).

Station C2, niveau de retour à 50 ans

38.92 (série journalière jittérisée, loi de Gumbel), 41.5 (série mensuelle, loi de Gumbel), 42.9 (série trimestrielle, loi de Gumbel), 43 (série annuelle, loi de Gumbel).

Méthode de dépassement de seuil

Introduction

Cette méthode, basée sur les travaux de J. Pickands, A.C. Davison et R. L. Smith est plus récente. Elle va impliquer le choix d'un seuil au-delà duquel le modèle peut être appliqué. Elle est sensible aux perturbations sur les données et aux hypothèses d'indépendance.
On montre que la loi conditionnelle de $X-u$ sachant que $X>u$ est approximativement : $H(y)=1-\left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}$ sur $\{y : y>0, (1+\xi y/\tilde{\sigma})>0\}$, $y=x-u$ et $\tilde{\sigma}=\sigma+\xi(u-\mu)$ si $\xi \neq 0$.

Dans le cas où $\xi=0$, on a $H(y)=1-\exp\left(-\frac{y}{\tilde{\sigma}}\right)$ avec $y>0$. Le paramètre de forme ξ est le même que celui des lois GEV du paragraphe précédent. Cette famille de lois constitue les lois de Pareto généralisées (GPD en anglais).

Comment estimer des niveaux de retour ?

Supposons que le modèle de dépassement au-dessus d'un seuil u soit de ce type, alors on a la relation :

$P\{X > x | X > u\} = \left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$. D'où $P\{X > x\} = P\{X > u\}\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$. Et on estime $p\{x > u\} = \varsigma_u$ par la fréquence relative des observations au-dessus de u. Ce modèle est sensible aux perturbations des observations, car un certain nombre de celles-ci peuvent passer de part et d'autre du seuil, modifiant ς_u. Observer un phénomène pendant 50 ans avec des observations journalières représente $m = 365 \times 50 = 18250$ mesures.

La « clustérisation » exprime le fait que des phénomènes extrêmes se produisent souvent de manière non indépendantes et forment des agrégats (clusters). Pour le dépassement de seuil, l'hypothèse d'indépendance est contraignante. Il faut donc envisager la « déclustérisation » : lorsque l'on dépasse le seuil, on va exiger un certain intervalle de temps (i.e. un certain nombre de valeurs dans la série) en-dessous du seuil pour dire que les dépassements sont indépendants. Sinon, ils seront regroupés au sein du même agrégat et on prendra le maximum de l'agrégat.

Niveau de retour

Le niveau de retour z_N dépassé en moyenne toutes les N années (m observations) est donné par la relation : $\varsigma_u\left[1+\xi\left(\frac{z_N-u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}$.

Selon les valeurs de ξ on a les formules : $\begin{cases} z_N = u + \sigma\log(m\varsigma_u), \xi = 0 \\ z_N = u + \frac{\sigma}{\xi}\left[(m\varsigma_u)^{\xi} - 1\right], \xi \neq 0 \end{cases}$

Résultats

Station A

Tout d'abord, nous allons modéliser la série mensuelle avec un modèle à seuil. On va admettre qu'il y a indépendance entre les maximums mensuels, mais il faudrait vérifier qu'il n'y a pas d'épisodes venteux à cheval sur deux mois consécutifs.

Rappelons que nous avons légèrement perturbé les données (nombres entiers de m/s) de façon à éviter les problèmes d'ex æquo. Comment déterminer un seuil ? On peut démontrer (Coles, 2001) que :

Si u_0 est une valeur au delà de laquelle le modèle Pareto s'applique, alors, pour $u > u_0$, $E(X - u | X > u)$ est une fonction linéaire de u.

On tracera donc $\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}\left(x_{(i)} - u\right)\right) : u < x_{\max}\right\}$ avec $x_{\max} = \max(X_i)$ et $x_{(1)}, \cdots, x_{(n_u)}$, les observations qui dépassent u. Il faut faire un compromis sur la valeur du seuil, trop bas, les hypothèses asymptotiques ne sont pas valides, trop élevé, il n'y a pas assez de valeurs pour une estimation fiable.

Avec la série mensuelle, on va choisir un seuil de 25 m/s (voir figure 2 en annexe) et on obtient $(\xi = -0.194, \tilde{\sigma} = 2.89)$, le niveau de retour à 50 ans est alors 33.7 et comme intervalle de confiance par la méthode delta [31.1 36.3]. Avec un seuil de 26, on a un niveau de retour de 33.8 avec un

intervalle de confiance [30.6 37]. Les résultats sont proches, mais on n'a pas pu utiliser la vraisemblance profil et le paramètre ξ n'est pas significativement différent de zéro.

Station C1

Pour C1, sur la série mensuelle, avec un seuil de 26 m/s, on trouve $(\xi = 0.098, \tilde{\sigma} = 2.07)$, ξ n'est pas significativement différent de zéro et le niveau de retour à 50 ans en imposant ce paramètre à zéro est de 35.8 [31.5 40.1] par la méthode delta. On peut voir en annexe (figure 3) la forme curieuse de la vraisemblance profil, très plate et l'intervalle de confiance qui « s'allonge vers les fortes valeurs ».

Station C2

Avec un seuil de 25 m/s, on trouve $(\xi = 0.192, \tilde{\sigma} = 2.72)$, mais ξ n'est pas significativement différent de zéro. Le niveau de retour en imposant ce coefficient à zéro est 37.1 avec un intervalle de confiance [30.8 43.3]. La vraisemblance profil ne semble pas applicable.

Déclustérisation

Pour finir, nous allons modéliser la série journalière pour la station A avec la méthode de dépassement de seuil, mais en prenant en compte la présence d'agrégats.

Si on fixe un seuil à 25 et on impose deux journées en-dessous du seuil pour séparer les clusters, on trouve $\xi = -0.25(0.14)$ et $\tilde{\sigma} = 3.25$ et avec un seuil de 26 et deux journées en-dessous du seuil : $\xi = -0.14(0.20)$ et $\tilde{\sigma} = 2.48$. Si on utilise le modèle sans déclustériser la série, on trouve $\xi = -0.19(0.14)$ et $\tilde{\sigma} = 2.89$ avec 37 dépassements au lieu de 33 en déclustérisant.

Résumé

Station C1, niveau de retour à 50 ans

35.8 (série mensuelle, seuil 26, paramètre ξ contraint à zéro)

Station A, niveau de retour à 50 ans

33.7 (série mensuelle, seuil 25), 33.8 (série mensuelle, seuil 26) ; vraisemblance profil non utilisable.

Station C2, niveau de retour à 50 ans

37.1 (série mensuelle, seuil 25, paramètre ξ contraint à zéro) ; vraisemblance profil non utilisable.

Conclusions

On constate que les phénomènes de vents extrêmes ne se modélisent pas de la même façon selon les stations. On peut remarquer des valeurs un peu détachées de l'ensemble des observations. On peut se poser la question de les retirer. Mais il y a un paradoxe à retirer des valeurs atypiques lorsque l'on cherche justement à les modéliser, on doit donc tout prendre en compte, en espérant que la mesure reste valide.

Travailler sur les valeurs extrêmes nécessite l'emploi de modèles car on va procéder à une extrapolation. Les conditions d'application de ces modèles paraissent bien plus contraignantes qu'on ne le pense (indépendance, loi asymptotique …). Ces techniques paraissent peu robustes.

Selon le niveau d'agrégation, les résultats peuvent varier. Le maximum de vraisemblance semble plus commode d'emploi.

Nous avons disposé de deux logiciels libres dédiés à ces méthodes. Ceux-ci sont parfois peu documentés et il faut donc regarder de près la théorie tout en dépouillant les résultats. L'usage de la vraisemblance profil reste délicat si l'on ne dispose pas de fonctions déjà écrites.

Des idées à envisager seraient l'étude et l'estimation de l'indice extrémal et la prise en compte de connaissances a priori (techniques bayésiennes).

Bibliographie

[1] Coles S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Editions Springer

[2] A. C. Davison (2003) *Statistical Models*, Cambridge University Press

[3] Embrechts P., Klüppelberg C., Mikosch T., (2001) *Modelling Extremal Events for Insurance and Finance*, Editions Springer

[4] E. J. Gumbel (1958), *Statistics of extremes*, Columbia University Press

[5] Hosking J. R. M., Wallis J. R., Wood E. F. (1985), *Estimation of the generalized extreme-value distribution by the method of probability-weighted moments*, Technometrics, vol. 27, no. 3, pages 251-261

[6] Resnick S. I. (1987), *Extreme Values, Regular Variation, and Point Processes*, Editions Springer

Annexes

Loi géométrique

Considérons une suite d'essais indépendants où à chaque essai, un événement a une probabilité donnée constante $0 < p < 1$ de se produire. Le numéro de l'essai où cet événement se produit pour la première fois suit une loi géométrique : $P\{X = x\} = p(1-p)^{x-1}$ pour $x = 1,2,3,\cdots$. On démontre que $E(X) = \sum_{x=1}^{\infty} x(1-p)^{x-1}p = 1/p$ (dériver la série $g(p) = \sum_{x=0}^{+\infty}(1-p)^x$). Ce résultat est utilisé pour interpréter les niveaux de retour.

Figures

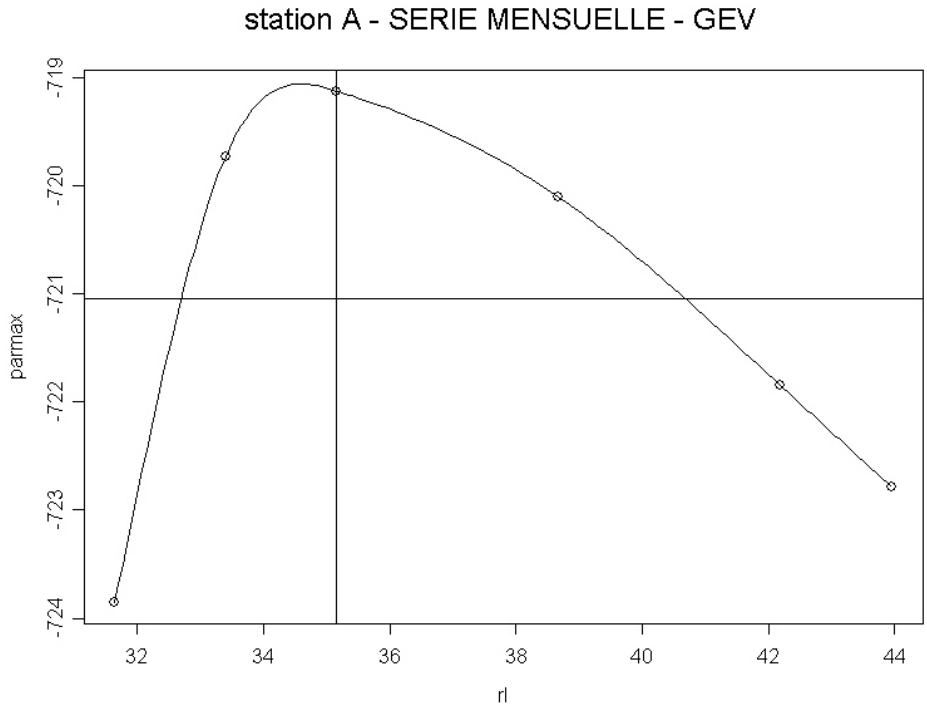

figure 1 : exemple de vraisemblance profil

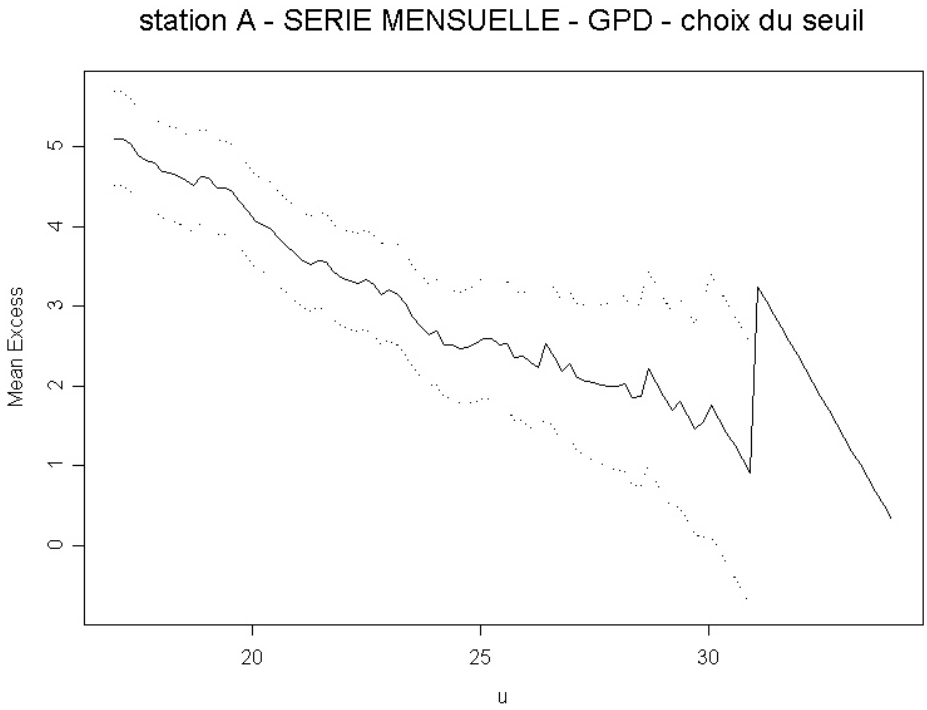

figure 2 : exemple de choix du seuil

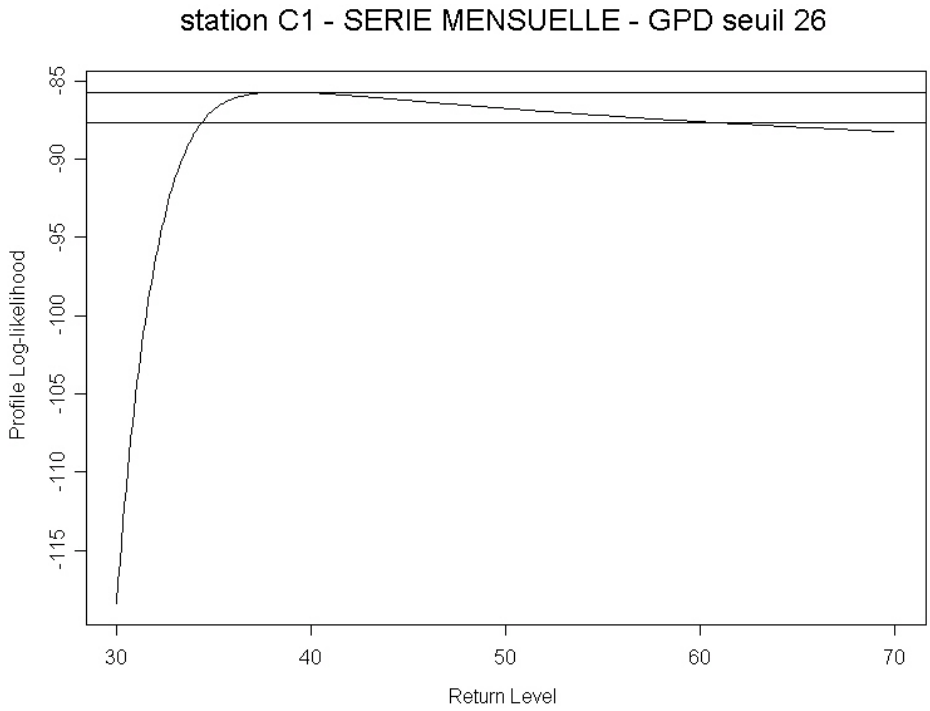

figure 3 : modèle à dépassement de seuil : vraisemblance profil