

## L'étude des risques extrêmes avec l'EVT (Extreme Value Theory)

L'étude des valeurs extrêmes est un problème couramment rencontré en ingénierie hydraulique comme l'illustre l'exemple suivant.

En février 1953, à la suite de l'émotion soulevée par les dépassements de digues et les violentes inondations sur les côtes des Pays-Bas (plus de 1800 morts), le gouvernement néerlandais met en place un groupe d'experts à qui il confie la tâche de répondre à la question suivante : quelle doit être la hauteur des digues pour éviter qu'une telle catastrophe ne se reproduise ? Plus précisément, la demande est de couvrir le risque d'inondation avec un niveau de probabilité correspondant à un temps de retour d'une fois tous les 10 000 ans.

Le système de mesure utilisé exprime la hauteur d'eau en mètres par rapport à un point fixe : le NAP, niveau normal à Amsterdam. Les informations disponibles sont les maxima annuels enregistrés sur 166 années. Le quantile cherché est donc bien au-delà de l'étendue des données. La catastrophe de 1953 correspond au niveau NAP+3,85, le précédent record connu étant NAP+4, en 1570. Les conclusions du groupe d'experts ont abouti à recommander une couverture au niveau NAP+5,14.

Dans le domaine de la gestion des risques, la prise en compte des événements extrêmes fait naturellement l'objet d'une attention particulière. La principale difficulté dans leur étude est évidemment qu'il est difficile d'obtenir une précision statistique satisfaisante sur des phénomènes par définition très rares ! Le plus souvent, en pratique, le quantile qui nous intéresse se situe même bien au-delà de l'étendue des observations disponibles et le calcul du quantile empirique est impossible. Les données seules ne suffisent pas à traiter le problème et l'adoption d'un minimum d'hypothèses et de modélisation est inévitable.

Cet article s'intéresse aux techniques dérivées de la statistique des échantillons ordonnés, usuellement désignées par l'acronyme EVT (de l'anglais *Extreme Value Theory*). Elles sont d'utilisation relativement ancienne en hydrologie pour représenter les phénomènes de crue, ou plus récente pour la modélisation du trafic dans les réseaux informatiques, par exemple. D'une portée assez générale, l'exposé doit être perçu dans le contexte de la finance et de l'assurance et correspond à une mise en œuvre opérationnelle de l'EVT sur les données de marchés boursiers européens. Ces travaux ont été réalisés en [collaboration](#) avec la Direction des Risques de LCH.Clearnet, en vue de déterminer le niveau des garanties déposées par les membres du marché.

Une première approche naturelle consiste à essayer d'ajuster une loi de probabilité sur les données, puis à en étudier les queues de distribution. Identifier et caler une distribution

satisfaisante dans cette perspective n'est en rien trivial, notamment si on se limite aux lois les plus usuelles. Un rapide calcul permet par exemple de se convaincre qu'une distribution gaussienne est totalement inapte à rendre compte des grandes variations de prix observées sur les marchés financiers : avec une telle distribution, un événement comme le krach boursier d'octobre 1987 présente un temps de retour de l'ordre de l'âge de l'univers. Dans le contexte boursier, le recours à des lois de probabilité présentant des queues de distribution plus fortes que celles de la loi gaussienne remonte au début des années 60.

Une solution pragmatique et couramment utilisée consiste à renoncer à probabiliser et à réaliser des stress-tests, c'est-à-dire à simuler les conséquences de scénarios extrêmes fixés de façon plus ou moins subjective. Il s'agit moins de chercher à évaluer précisément

leur probabilité d'occurrence que de prendre conscience de ce qui peut arriver de «pire». La démarche qui suit vise au contraire à rester dans un cadre probabiliste.

On note  $X_1, X_2, \dots, X_i, \dots$  la série de données dont on cherche à représenter les extrêmes (par exemple des rendements boursiers quotidiens) et que l'on considère comme un échantillon de variables aléatoires de fonction de répartition  $F(x) = P(X_i \leq x)$ . Dans un premier temps, on suppose les données indépendantes. Cette hypothèse, assez forte en générale, n'est cependant pas cruciale dans la mesure où, comme on le verra plus bas, elle peut être affaiblie sans remettre en cause la démarche. L'objectif poursuivi est l'estimation de quantiles élevés, c'est-à-dire de quantités  $x_q = F^{-1}(q)$ , pour  $q$  «proche» de 1.

L'approche par EVT ne vise pas à modéliser ou estimer la fonction de répartition inconnue  $F$  dans son ensemble, mais uniquement ses queues de distribution, qui sont seules utiles à la représentation des extrêmes. Deux approches sont possibles : la première, dénommée *block maxima method* s'appuie sur un découpage des données en blocs, dont les maxima sont supposés distribués selon une loi d'une famille connue, en vertu du résultat asymptotique exposé plus bas. La seconde approche est désignée sous le terme *Peaks-Over-Threshold (POT) method* et modélise la distribution des valeurs dépassant un seuil donné. Ces deux méthodes reposent sur des théorèmes de convergence mathématiquement équivalents. En pratique, on se ramène dans les deux cas à un cadre de statistique paramétrique et d'ajustement de paramètres sur les données. Selon le contexte, l'une ou l'autre des approches peut se révéler mieux adaptée, mais il est le plus souvent utile de les mettre toutes deux en œuvre afin d'en comparer les résultats.

### **Block maxima**

$M_n = \text{Max}(X_1, \dots, X_n)$  désigne par la suite le maximum sur un bloc de  $n$  observations consécutives. Le théorème de Fisher-Tippett est le résultat central de la démarche et établit que s'il existe deux suites de normalisation  $\alpha_n > 0$  et  $\beta_n$  telles que la suite de variables

aléatoires normalisées  $\frac{M_n - \beta_n}{\alpha_n}$  converge en loi quand la taille des blocs  $n$  croît, alors la loi

limite est nécessairement une distribution de la famille GEV (pour *Generalized Extreme Value*), caractérisée par un paramètre de queue  $\tau$  (*tail index*) intrinsèque à la loi  $F$  et en particulier indépendant de  $n$ .

Pour une taille des blocs donnée  $n$ , l'échantillon d'origine est réduit aux maxima par blocs et une distribution GEV est ensuite ajustée par maximum de vraisemblance sur cet échantillon réduit. On obtient ainsi des estimations des paramètres  $\hat{\alpha}_n$ ,  $\hat{\beta}_n$  et  $\hat{\tau}$ , ainsi que des erreurs statistiques d'échantillonnage. L'expression analytique de la fonction de répartition de la famille GEV étant connue, on peut en déduire une formulation explicite du  $q$ -quantile de  $F$  et l'estimateur correspondant.

Tout le problème pratique réside dans le choix de la taille des blocs, qui doit être suffisamment large pour pouvoir s'appuyer sur un résultat asymptotique. Or dans le même temps, si l'on dispose au départ d'un échantillon de taille  $N = nk$ , l'échantillon réduit aux maxima, de taille  $k$ , doit aussi être le plus grand possible pour limiter l'erreur d'échantillonnage dans l'estimation des paramètres  $\hat{\alpha}_n$ ,  $\hat{\beta}_n$  et  $\hat{\tau}$ . Il y donc un compromis à trouver et une difficulté que l'on rencontre, sous une forme ou sous une autre, dans toutes les approches.

Les résultats sont ensuite directement exploitables. A titre d'exemple, si l'on s'intéresse à des rendements boursiers journaliers négatifs et en relevant les maxima trimestriels, le quantile à 99 % de la distribution ainsi ajustée donne le niveau de baisse qu'il faut s'attendre à enregistrer en moyenne tous les 100 trimestres. Autrement dit, on obtient ainsi un scénario de baisse sur une journée correspondant à un temps de retour de 25 ans.

### **Peaks-over-threshold**

Cette approche s'appuie également sur un théorème limite (analogue du théorème de Fisher-Tippett), qui établit que  $F_u$ , fonction de répartition des dépassements d'un seuil  $u$ , converge quand  $u$  croît, vers celle d'une distribution de la famille GPD (pour Generalized Pareto Distribution). Cette distribution limite est paramétrée par un facteur d'échelle  $\sigma_u$  dépendant de  $u$  et un paramètre de queue  $\tau$  intrinsèque à  $F$ .

La fonction de répartition des dépassements du seuil  $u$  est une distribution conditionnelle représentant la probabilité que la variable étudiée dépasse  $u$  d'au plus  $y$  conditionnellement au dépassement de  $u$ . Elle est donc définie par la relation :

$$F_u(y) = P[X_i \leq u + y \mid X_i > u].$$

Pour un seuil  $u$  donné, on note  $N_u$  le nombre de points de l'échantillon  $X_1, X_2, \dots, X_N$  dépassant  $u$ . L'ajustement sur ces  $N_u$  points d'une loi de la famille GPD permet d'obtenir des estimations des paramètres  $\hat{\sigma}_u$  et  $\hat{\tau}$ . Dans ce cadre également, on obtient ainsi des estimateurs explicites des quantiles.

De manière comparable à l'approche précédente, la difficulté se situe ici dans le choix du seuil  $u$ . Le compromis est à trouver entre un seuil suffisamment élevé pour justifier le recours au résultat précédent et un nombre d'excès  $N_u$  suffisamment grand pour limiter l'erreur d'échantillonnage dans l'estimation des paramètres  $\hat{\sigma}_u$  et  $\hat{\tau}$ .

Les théorèmes de base précédents se placent dans le cadre i.i.d. (données indépendantes et identiquement distribuées). Cette hypothèse est souvent discutable, en particulier pour des données de marché qui présentent des périodes alternées de forte et de faible volatilité. Des travaux récents se sont concentrés sur cette forme de dépendance et conduisent à introduire un paramètre supplémentaire  $\theta$  (*extremal index* compris entre 0 et 1) rendant compte du degré de regroupement dans le temps des extrêmes. Une propriété importante en pratique est que l'on peut d'abord ajuster une distribution GEV ou GPD comme si les données étaient i.i.d., puis estimer  $\theta$  ensuite. Une fois le paramètre  $\theta$  estimé, les estimateurs des quantiles sont ajustés d'une façon qui rend compte de la propriété suivante : avec des données corrélées, le comportement des extrêmes dans un échantillon de taille  $N$  est comparable à celui des extrêmes d'un échantillon i.i.d. de taille réduite  $N\theta$ .

Les résultats précédents fournissent un cadre mathématique à l'analyse des extrêmes. Leur forme et leur rôle sont analogues à ceux du théorème de la limite centrale, lorsqu'il s'agit d'étudier les propriétés du maximum (ou minimum) sur un échantillon plutôt que celles de la valeur moyenne. Pour autant, leur utilité pratique n'est pas immédiate. Comme souligné précédemment, toutes les approches se heurtent au compromis à trouver entre biais et précision des estimations. Avec une procédure automatisée, ce choix peut cependant être guidé par l'ajustement de distributions GEV et GPD correspondant à différentes tailles de blocs et valeurs de seuil, puis la minimisation d'une distance entre distribution empirique et distribution estimée pour identifier le meilleur ajustement.