Modélisation statistique d'évènements extrêmes

Application dynamique aux catastrophes naturelles de l'Océanie

Rapport soumis dans le cadre du cours ACT-2101 du
Baccalauréat en actuariat

Par
Marc-Olivier Ricard

Présenté à
Marie-Pier Côté

École d'actuariat
Université Laval

Décembre 2019

Table des matières

Résumé

Introduction

L'objectif principal d'une analyse de valeur extrême est d'être capable de quantifier le comportement stochastique d'un processus à des niveaux exceptionnellement élevés ou bas. De plus, ce type d'analyse a souvent comme but d'estimer la probabilité de réalisation d'évènements qui sont encore plus extrêmes que n'importe quel évènement passé. La théorie des valeurs extrêmes rend ce genre d'extrapolation possible.

Chapitre 1

Notions préliminaires

1.1 Théorie classique des valeurs extrêmes

Posons $M_n = \max\{X_1, \ldots, X_n\}$, le maximum des n observations indépendantes et de distribution commune. Dans le cas où le comportement des X_i serait connu, il serait facile d'obtenir le comportement exact de M_n, mais, en pratique, cette situation est très rare. Par contre, sous des hypothèses appropriées et pour $n \to \infty$, il est possible d'approximer le comportement de M_n et d'obtenir une famille de modèles qui peuvent être ajustés avec les différentes valeurs observées. On appelle cette approche le paradigme des valeurs extrêmes. Il faut ensuite être en mesure d'estimer les différents paramètres du modèle, de quantifier l'incertitude, d'évaluer le modèle et finalement de maximiser l'utilisation de l'information disponible.

Dans cette section, on étudie le modèle qui s'intéresse au comportement statistique de M_n. Nous pourrions utiliser la distribution théorique :

$$\begin{aligned} P\{M_n \leq z\} &= P\{X_1 \leq z, \ldots, X_n \leq z\} \\ &= P\{X_1 \leq z\} \times \cdots \times P\{X_n \leq z\} \\ &= \{F(z)\}^n \end{aligned} \tag{1.1.1}$$

Par contre, en pratique, F est inconnu et donc, cette approche n'est pas utile. Il serait possible d'estimer F avec les valeurs observées, mais la moindre erreur dans l'estimation pourrait mener à une très grande erreur pour F^n. L'approche alternative est d'approximer directement F^n avec seulement les valeurs extrêmes. Étant donné que la distribution de M_n est dégénératrice à un certain point, nous normalisons M_n :

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{1.1.2}$$

Theorème 1.1 *S'il existe des séquences de constantes* $\{a_n > 0\}$ *et* $\{b_n\}$ *tel que*

$$P\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z), \quad \textit{quand} \;\; n \to \infty$$

où G *est une fonction de répartition non-dégénératrice. Alors,* G *appartient à une des distributions de valeur extrême, soit les lois Gumbel, Fréchet et Weibull. Donc, peu importe la distribution* $F_{X\,i}$*, les trois dernières lois sont les seules distributions limites pour* M_n^**.*

Malgré qu'il pourrait sembler logique de choisir une des trois distributions et d'estimer ses paramètres, cette piste possède deux faiblesses : une technique est nécessaire pour choisir la distribution la plus appropriée et dès que cette décision est prise, les inférences qui suivent présument que le bon choix a été fait. Une meilleure analyse est possible en combinant en une seule distribution les distributions Gumbel, Fréchet et Weibull :

$$G(z) = \exp\left\{ -\left[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\right]^{-1/\xi}\right\}, \tag{1.1.3}$$
$$\{z : (1 + \xi(z-\mu)/\sigma) > 0\},\ -\infty < \mu > \infty,\ \ \sigma > 0,\ -\infty < \xi > \infty$$

Ceci est la famille de distributions d'extremum généralisée (GEV). Comme on peut voir, le modèle possède trois paramètres : μ (paramètre de position), σ (paramètre d'échelle), ξ (paramètre de forme).

Si nous revenons au Théorème 1.1, une autre difficulté est le fait que, en pratique, les constantes de normalisation a_n et b_n sont inconnus. Ce problème est facilement résolu :

$$\text{Nous savons déjà que } P\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \approx G(z), \quad \text{quand } n \to \infty$$

$$\Rightarrow P\{M_n \leq z\} \approx G\left\{\frac{z - b_n}{a_n}\right\} = G^*(z),$$

où $G^*(z)$ est simplement un autre membre de la famille *GEV*. Étant donné qu'en pratique, les paramètres doivent être estimés, ceci ne change rien au modèle proposé.

Tout ceci mène à une première méthodologie pour modéliser les valeurs extrêmes :

- Les données sont groupées en séquence de longueur n
- Le maximum de chaque séquence est calculé
- Une distribution *GEV* est calibrée à ces maximums
- La distribution peut être manipulée pour obtenir différentes statistiques

La distribution permet, par exemple, d'obtenir les très grands quantiles et ceux-ci sont obtenus comme ceci :

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\Big[1 - \{\log(1-p)\}^{-\xi}\Big], & \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0 \end{cases} \tag{1.1.4}$$

où :

- $G(z_p) = 1 - p$
- z_p est le niveau de retour correspondant à la période de retour $1/p$
- On s'attend à ce que le niveau z_p soit dépassé en moyenne une fois chaque $1/p$ année.
- Chaque année, le niveau z_p est dépassé avec probabilité p

Nous pouvons également poser $y_p = -\log(1-p)$ pour obtenir :

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\Big[1 - y_p{}^{-\xi}\Big], & \xi \neq 0 \\ \mu - \sigma\log y_p, & \xi = 0 \end{cases} \tag{1.1.5}$$

Nous simplifions maintenant la notation en dénotant les maximums par $Z_1, \ldots, Z_m$, on assume que ce sont des variables indépendantes d'une distribution GEV dont il faut estimer les paramètres. À noter, que même si les X_i sont dépendants, il peut être raisonnable d'assumer que les Z_i sont indépendants.
La méthode la plus populaire pour l'estimation des paramètres est la méthode du maximum de vraisemblance. À noter, que si $\xi \leq -0.5$, il est fort probable qu'il sera impossible d'obtenir des estimateurs valides. Cependant, en pratique, cette situation est plutôt rare.

La log-vraisemblance va comme suit dans le cas où $\xi \neq 0$:

$$\begin{gathered}\ell(\mu, \sigma, \xi) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big] - \sum_{i=1}^{m} \Big[1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big)\Big]^{-1/\xi}, \\ 1 + \xi\Big(\frac{z_i - \mu}{\sigma}\Big) > 0, \quad i = 1, \ldots, m\end{gathered} \tag{1.1.6}$$

Dans le cas où $\xi = 0$, il faut utiliser la limite Gumbel de la distribution :

$$\ell(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \Big(\frac{z_i - \mu}{\sigma}\Big) - \sum_{i=1}^{m} \exp\Big[-\Big(\frac{z_i - \mu}{\sigma}\Big)\Big] \tag{1.1.7}$$

Après l'estimation des paramètres, nous pouvons estimer différents niveaux de retour :

$$\hat{z_p} = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\Big[1 - y_p{}^{-\hat{\xi}}\Big], & \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma} \log y_p, & \hat{\xi} = 0 \end{cases} \tag{1.1.8}$$

$$\mathrm{Var}(\hat{z_p}) \approx \nabla z_p^T V \nabla z_p \tag{1.1.9}$$

où V correspond à la matrice variance-covariance de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ et

$$\begin{aligned}\nabla z_p^T &= \Big[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\Big] \\ &= [1,\ -\xi^{-1}(1 - y_p^{-\xi}),\ \sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1} y_p^{-\xi} \log y_p]\Big|_{(\hat{\mu}, \hat{\sigma}, \hat{\xi})}\end{aligned} \tag{1.1.10}$$

Même s'il est impossible de valider l'extrapolation faite par le modèle, on peut tout de même vérifier la qualité du modèle avec les données observées. Ces quatre graphiques sont utiles à cet effet :

- Graphique P-P
- Graphique Q-Q
- Graphique du niveau de retour
- Histogramme des données avec la densité prédite par le modèle

1.2 Méthode par l'approche POT (*Peaks over threshold*)

Un des inconvénients de la modélisation avec les maximums est qu'il y a potentiellement des données utiles qui ne sont pas utilisées étant donné que celles-ci ne sont pas le maximum de leur séquence, mais qui auraient pu être celui d'une autre séquence. La méthode présentée dans cette section fait une meilleure utilisation des données.

Soit $X_1, X_2, \ldots$, une séquence de variables indépendantes, identiquement distribuées et avec fonction de répartition marginale F. Nous considérons comme évènements extrêmes ceux qui dépassent un certain seuil u. Le comportement stochastique d'un évènement extrême peut donc être décrit comme suit :

$$\Pr\{X > u + y \mid X > u\} = \frac{1 - F(u+y)}{1 - F(u)} \tag{1.2.1}$$

Si le comportement exact de F était connu, la distribution de l'excès de seuil serait également connue. Cependant, en pratique, cette situation est rare et des approximations sont alors applicables lorsque u est assez grand.

1.3 Modèles additifs généralisés et splines

1.4 Techniques de bootstrap

Chapitre 2

Article étudié

Chapitre 3

Application à des données réelles

Conclusion