Modélisation statistique d'évènements extrêmes

Application multivariée dynamique aux catastrophes naturelles de l'Océanie

Marc-Olivier Ricard

Sous la supervision de
Marie-Pier Côté

4 décembre 2019

UNIVERSITÉ LAVAL

Plan de la présentation

1. Mise en contexte
2. Notions préliminaires
   - Théorie classique des valeurs extrêmes
   - Méthode par l'approche POT
3. Article étudié
   - Introduction
   - Approche dynamique de la théorie des valeurs extrêmes
4. Application à des données réelles
   - Analyse de données
   - Approche classique
5. Conclusion

UNIVERSITÉ LAVAL

Motivations

- Intérêt recherche
- Stage été 2019
- Maîtrise
- Problématiques de l'industrie
- Black Swans

UNIVERSITÉ LAVAL

Théorie classique des valeurs extrêmes

Soit $M_n = \max\{X_1, \ldots, X_n\}$, le maximum des n observations indépendantes et de distribution commune. S'il existe des séquences de constantes $\{a_n > 0\}$ et $\{b_n\}$ tel que

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

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^*.

Une meilleure analyse est possible en combinant en une seule distribution les distributions Gumbel, Fréchet et Weibull :

Famille de distributions d'extremum généralisée (GEV)

$$G(z) = \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \right\} \tag{1}$$

Comme on peut voir, le modèle possède trois paramètres : μ (paramètre de position), σ (paramètre d'échelle), ξ (paramètre de forme).

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

Méthode par l'approche POT

Soit $X_1, \ldots, X_n$, 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.

Nous savons déjà que $\Pr\{\max\{X_1, \ldots, X_n\} \leq z\} \approx G(z)$ où

$$G(z) = \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \right\}$$

Ensuite, pour u assez grand, la fonction de répartition de $X - u \mid X > u$ est environ :

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{2}$$

où

$$\tilde{\sigma} = \sigma + \xi(u - \mu)$$

Il s'agit ici de la famille de distributions Pareto généralisée.

Les paramètres de H sont uniquement déterminés par ceux de la distribution *GEV*. ξ conserve la même valeur, par exemple.

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

- Les données brutes représentent une séquence de valeurs iid $x_1, \ldots, x_n$
- Les valeurs extrêmes sont identifiées en définissant un seuil de grande valeur u
- Les observations qui dépassent u sont définies par $x_{(1)}, \ldots, x_{(k)}$ et les excès de seuil par $y_j = x_{(j)} - u$, $j = 1, \ldots, k$
- Les y_j sont reconnues comme des réalisations indépendantes d'une variable aléatoire dont la distribution peut être approximée par un membre de la famille Pareto généralisée

UNIVERSITÉ LAVAL

Un des défis de cette procédure est le choix de la valeur du seuil, où il faut trouver une bonne balance entre la variance et le biais du modèle.

Deux méthodes :

- *Mean residual plot*
- Estimer les paramètres du modèle pour différents seuils

Niveaux de retour :

$$\begin{aligned} \Pr\{X > x \mid X > u\} &= \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \\ \Rightarrow \Pr\{X > x\} &= \zeta_u\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \end{aligned} \tag{3}$$

où $\zeta_u = \Pr\{X > u\}$. Ensuite, le niveau x_m qui est dépassé en moyenne une fois chaque m observations est obtenu comme suit :

$$\begin{aligned} \frac{1}{m} &= \zeta_u\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \\ \Rightarrow x_m &= u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right], \quad x_m > u \end{aligned} \tag{4}$$

Article étudié

Article choisi : Chavez-Demoulin et collab. (2016)

- Tout de même récent
- Longueur et difficulté convenables
- Sujet spécifique intéressant
- Implémentation en R

UNIVERSITÉ LAVAL

Introduction

L'article choisi présente une méthodologie dans laquelle la modélisation des valeurs extrêmes dépend de variables exogènes associées à la variable réponse qui est modélisée.

- Ce sont les paramètres du modèle de valeurs extrêmes choisi qui dépendront des covariables disponibles.
- Vise à améliorer la calibration des lois aux données et du même coup rendre l'inférence plus crédible et précise.

UNIVERSITÉ LAVAL

- Processus de Poisson non-homogène
- Modèles additifs généralisés, splines de lissage et méthode du maximum de vraisemblance pénalisé
- Pareto généralisée non-stationnaire dont les paramètre dépendent des variables exogènes
- Algorithme basé sur les paramètres orthogonaux
- L'idée de dépendance entre les paramètres du modèle et des covariables a déjà élaborée dans Coles et collab. (2001), ce qui est réellement nouveau dans l'approche proposée est la dépendance semi-paramétrique avec les splines de lissage

UNIVERSITÉ LAVAL

Critères :

- Les données représentent des pertes aléatoires encourus à des temps aléatoires.
- Le but de l'analyse est d'estimer la distribution de la perte globale à travers le temps.
- Les données contiennent un nombre suffisant d'observations.
- Les données contiennent des évènement extrêmes.
- Les données contiennent des covariables significatives.

Notes :

- Gains
- β

UNIVERSITÉ LAVAL

Approche dynamique de la théorie des valeurs extrêmes

La dépendance entre les paramètres et les covariables peut être paramétrique, non-paramétrique ou bien semi-paramétrique et elle peut également inclure des intéractions entre les variables. L'idée générale du modèle peut être représentée par l'équation suivante :

$$g_k(\theta_k) = f_k(x) + h_k(t), \quad k \in \{1, \ldots, p\} \tag{5}$$

où :

- θ est le vecteur des p paramètres du modèle.
- g_k est une fonction de lien.
- f_k est la fonction pour les différents niveaux d'une variable catégorique.
- h_k est soit une fonction linéaire paramétrique ou bien une fonction lisse non-paramétrique de t.

UNIVERSITÉ LAVAL

On assume que le nombre d'excès du seuil u sélectionné suit un processus de Poisson non-homogène avec comme taux :

$$\lambda = \lambda(x, t) = \exp(f_\lambda(x) + h_\lambda(t)) \tag{6}$$

L'application d'un modèle additif généralisé mène à un estimé $\hat{\lambda}$.

UNIVERSITÉ LAVAL

Il faut s'assurer que les procédures de calibration des paramètres ξ et β aux données convergent. Pour ce faire, il faut absolument que ces deux paramètres soient orthogonaux par rapport à l'information de Fisher.

$$\begin{aligned} \nu &= \log((1+\xi)\beta),\ \xi > -1 \\ \Rightarrow \beta &= \frac{\exp(\nu)}{1+\xi} \end{aligned} \tag{7}$$

Dans la même ordre d'idée que pour λ, on définit ξ et ν comme suit :

$$\xi = \xi(x,t) = f_\xi(x) + h_\xi(t) \tag{8}$$

$$\nu = \nu(x,t) = f_\nu(x) + h_\nu(t) \tag{9}$$

$$\Rightarrow \beta = \beta(x,t) = \frac{\exp(\nu(x,t))}{1+\xi(x,t)} \tag{10}$$

Les équations 8 et 9 sont donc celles qui seront estimées avec les excès de seuil disponibles pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$. Ces estimateurs sont donc des estimateurs basés sur les estimateurs $\hat{f}_\xi, \hat{h}_\xi, \hat{f}_\nu$ et $\hat{h}_\nu$. Ces derniers estimateurs sont obtenus à partir des vecteurs observés $z_i = (t_i, x_i, y_{t_i})$ où :

- $i \in \{1, \ldots, n\}$.
- $0 \leq t_1 \leq \cdots \leq t_n \leq T$ représente les temps d'excès de seuil.
- x_i représente le vecteur des covariables.
- y_{t_i} représente les réalisations de la variable aléatoire Y_{t_i}.
- Y_{t_i} représente les excès de seuil u.

$$\widehat{\mathrm{VaR}_\alpha} = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{1-\alpha}{\hat{\lambda}/n'}\right)^{-\hat{\xi}} - 1\right) \tag{11}$$

$$\widehat{\mathrm{ES}_\alpha} = \begin{cases} \frac{\widehat{\mathrm{VaR}_\alpha}+\hat{\beta}+u\hat{\xi}}{1-\hat{\xi}}, & \hat{\xi} \in (0,1) \\ \infty, & \hat{\xi} \geq 1 \end{cases} \tag{12}$$

où $n' = n'(x,t)$ représente le nombre total de pertes pour les covariables x et pour le temps t.

Cependant, en pratique, il est plus utile de modéliser directement
$\rho = \rho(x,t) = \lambda(x,t)/n'(x,t)$ qui représente le taux d'excès de seuil pour x et t.

Analyse de données

Pour mettre en application l'approche proposée par Chavez-Demoulin et collab. (2016), deux jeux de données du paquetage `CASdatasets` sont utilisés. Soit `auscathist` et `nzcathist`.

Ces données représentent respectivement l'historique des catastrophes naturelles pour l'Australie ainsi que pour la Nouvelle-Zélande. Les prochaines diapositives présentent une analyse globale des données utilisées.

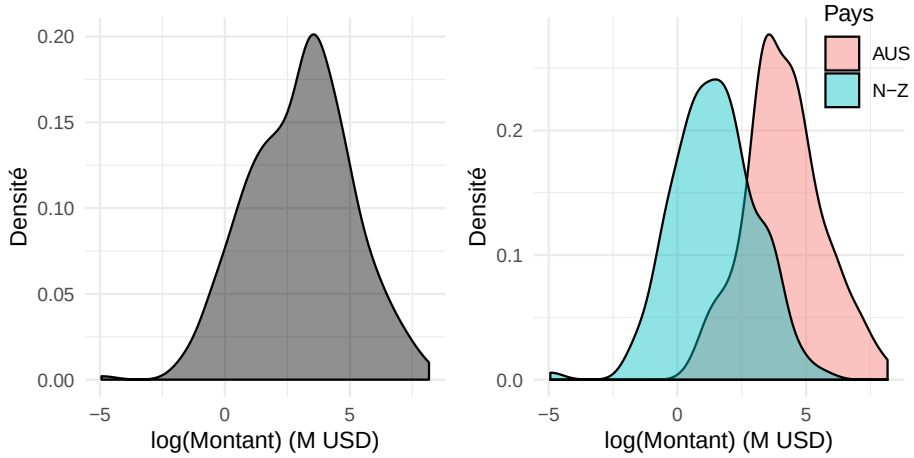

FIGURE 1 – Densité du logarithme du montant des catastrophes

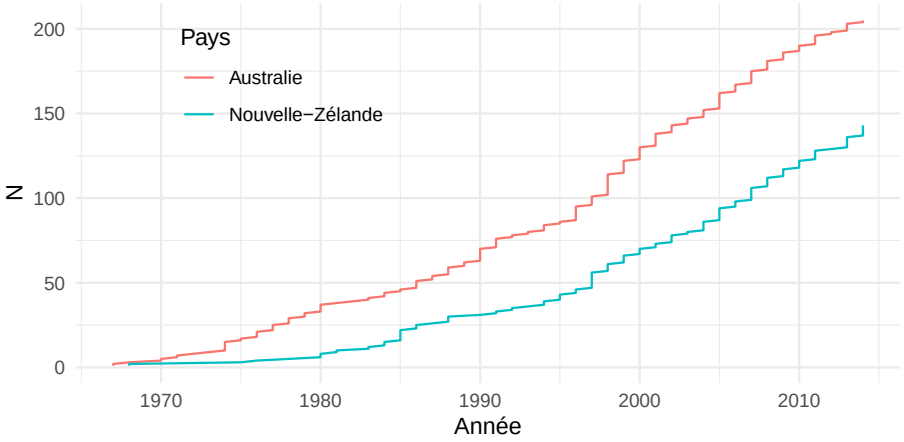

FIGURE 2 – Nombre cumulatif de catastrophes par pays

UNIVERSITÉ LAVAL

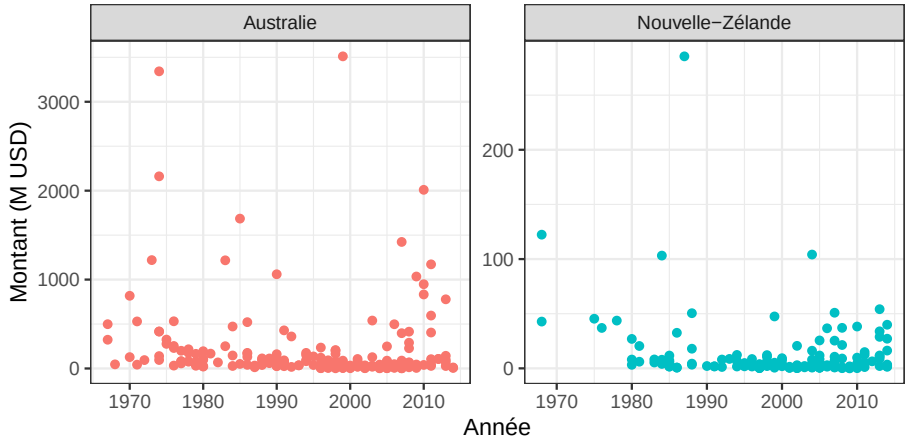

FIGURE 3 – Évolution des catastrophes par pays

Type	N	Moyenne	Écart	Minimum	Médiane	Q3	Maximum
Bushfire	26	143	259	7	39	134	1217
Cyclone	37	325	675	2	88	217	3343
Earthquake	9	56	91	0	25	47	286
Flood	83	62	230	0	5	39	2010
Flood, Storm	36	54	79	1	24	71	414
Hailstorm	41	274	606	0	75	235	3511
Other	12	121	296	2	6	50	1035
Power outage	2	6	6	1	6	8	11
Storm	86	97	228	0	28	57	1424
Tornado	12	7	13	0	3	7	47
Weather	4	11	11	2	6	13	27

TABLE 1 – Résumé statistique des montants (M USD) de catastrophes par Type

- Par souci de manque de temps, les résultats obtenus avec l'approche classique ne seront pas montrés de façon détaillée. On conlut que cette approche n'est pas tout à fait adéquate dans le cas présent.
- Cette approche reste viable, rapide et peut être une bonne solution lorsque seulement les montants sont disponibles.
- Un seuil de 10 M USD fut sélectionné

Bibliographie

Chavez-Demoulin, V., P. Embrechts et M. Hofert. 2016, «An extreme value approach for modeling operational risk losses depending on covariates», *Journal of Risk and Insurance*, vol. 83, n° 3, p. 735–776.

Coles, S., J. Bawa, L. Trenner et P. Dorazio. 2001, *An introduction to statistical modeling of extreme values*, vol. 208, Springer.

Cox, D. R. et N. Reid. 1987, «Parameter orthogonality and approximate conditional inference», *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 49, n° 1, p. 1–18.

Green, P. J. et B. W. Silverman. 1993, *Nonparametric regression and generalized linear models : a roughness penalty approach*, Chapman and Hall/CRC.

UNIVERSITÉ LAVAL

Bibliographie (suite)

Thiombiano, A. N., S. El Adlouni, A. St-Hilaire, T. B. Ouarda et N. El-Jabi. 2017, «Nonstationary frequency analysis of extreme daily precipitation amounts in southeastern canada using a peaks-over-threshold approach», *Theoretical and Applied Climatology*, vol. 129, n° 1-2, p. 413–426.

Wood, S. N. 2017, *Generalized additive models : an introduction with R*, Chapman and Hall/CRC.

UNIVERSITÉ LAVAL