Modélisation statistique d'évènements extrêmes

Application multivariée 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

Remerciments

Je tiens à remercier Marie-Pier Côté pour la supervisation de mon projet de recherche et pour le temps qu'elle a investi dans celui-ci.

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 :

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

Théorème 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 :

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

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.

Théorème 1.2. *Nous savons déjà que* $\Pr\{\max\{X_1, \ldots, X_n\} \leq z\} \approx G(z)$ *où*

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

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

$$H(y) = 1 - \Big(1 + \frac{\xi y}{\tilde{\sigma}}\Big)^{-1/\xi}, \ \{y : \ y > 0 \ \ et \ \ (1 + \xi y/\tilde{\sigma}) > 0\}$$

où

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

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

Nous avons donc comme théorème que si les maximums ont comme distribution approximative G, les excès de seuil ont comme distribution approximative un membre de la famille Pareto généralisée. De plus, les paramètres de H sont uniquement déterminés par ceux de la distribution *GEV*. ξ conserve la même valeur, par exemple. La valeur de n influence les paramètres de G, mais pas ceux de H.

Démonstration du théorème 1.2. tiré de Coles et collab. (2001)

$$F^n(z) \approx \exp\left\{ -\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi} \right\}$$

$$\Rightarrow n \log F(z) \approx -\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}$$

Pour n assez grand, un développement de Taylor donne :

$$\log F(z) \approx -(1 - F(z))$$

$$\Rightarrow 1 - F(u) \approx \frac{1}{n}\Big[1 + \xi\Big(\frac{u-\mu}{\sigma}\Big)\Big]^{-1/\xi}$$

$$\Rightarrow 1 - F(u+y) \approx \frac{1}{n}\Big[1 + \xi\Big(\frac{u+y-\mu}{\sigma}\Big)\Big]^{-1/\xi}$$

$$\Rightarrow \Pr\{X > u + y \mid X > u\} \approx \frac{n^{-1}[1 + \xi(u + y - \mu)/\sigma]^{-1/\xi}}{n^{-1}[1 + \xi(u - \mu)/\sigma]^{-1/\xi}}$$
$$= \left[1 + \frac{\xi(u + y - \mu)/\sigma}{1 + \xi(u - \mu)/\sigma}\right]^{-1/\xi}$$
$$= \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}, \quad \tilde{\sigma} = \sigma + \xi(u - \mu)$$

□

On propose le cadre suivant pour la modélisation de valeurs extrêmes : les données brutes représentent une séquence de valeurs indépendantes et identiquement distribuées $x_1, \ldots, x_n$ et les valeurs extrêmes sont identifiées en définissant un seuil de grande valeur u. Les observations qui dépassent ce seuil sont définies par $x_{(1)}, \ldots, x_{(k)}$ et les excès de seuil par $y_j = x_{(j)} - u,\ j = 1, \ldots, k$. Ensuite, 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.

Un des défis de cette démarche est le choix de la valeur de u, où il faut trouver une bonne balance entre la variance et le biais du modèle. La pratique standard est de choisir le plus petit u possible qui respecte les hypothèses du modèle. Une première méthode est basée sur la moyenne de la distribution Pareto généralisée :

$$\mathrm{E}(Y) = \frac{\sigma}{1 - \xi}, \quad \xi < 1 \tag{1.2.2}$$

Si la distribution est appropriée pour modéliser les excès d'un seuil u_0 :

$$\mathrm{E}(X - u_0 \mid X > u_0) = \frac{\sigma_{u_0}}{1 - \xi} \tag{1.2.3}$$

où nous utilisons σ_{u_0} pour le paramètre d'échelle des excès du seuil u_0. De plus, si la distribution est adéquate pour u_0, elle l'est également pour tous les seuils $u > u_0$ avec un changement approprié du paramètre d'échelle :

$$\begin{aligned} \mathrm{E}(X - u \mid X > u) &= \frac{\sigma_u}{1 - \xi} \\ &= \frac{\sigma_{u_0} + \xi(u - u_0)}{1 - \xi} \end{aligned} \tag{1.2.4}$$

Donc, $\mathrm{E}(X - u \mid X > u)$ est une fonction linéaire de u et est également la moyenne des excès du seuil u. Cette moyenne peut être empiriquement approximée pour les données disponibles. Nous savons que ces approximés devraient changer linéairement avec u quand la valeur de u est appropriée pour le modèle Pareto généralisé.

Tout ceci mène finalement à la première procédure : nous considérons les points suivants :

$$\left\{ \left(u,\ \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\} \tag{1.2.5}$$

Cette séquence de points correspond au *mean residual plot*. Après un seuil u_0 pour lequel la distribution Pareto généralisée fournie une approximation adéquate pour les excès, le graphique devrait être linéaire en u. Toutefois, il peut parfois être difficile d'interpréter ce type de graphique en pratique.

La deuxième méthode propose d'estimer le modèle pour différents seuils. Après un seuil u_0 pour lequel la distribution Pareto généralisée fournie une approximation adéquate pour les excès, les estimés du paramètre de forme ξ devraient rester constants et les estimés de σ_u devraient être linéaires en u à moins que $\xi = 0$, car $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$. Nous pouvons également modifier le paramètre d'échelle de la distribution : $\sigma^* = \sigma_u - \xi u$. Avec cette paramétrisation, σ^* et ξ devraient rester constants au-delà de u_0. Nous pouvons donc tracer les graphiques de $\hat{\sigma}^*$ et $\hat{\xi}$ par rapport à u et sélectionner u_0 comme la plus petite valeur of u pour laquelle les estimés sont presque constants.

Après avoir choisi un seuil, les paramètres de la distribution Pareto généralisée peuvent être estimés avec la méthode du maximum de vraisemblance. Définissons $y_1, \ldots, y_k$ comme les k excès du seuil u. Pour $\xi \neq 0$, la log-vraisemblance est :

$$\ell(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi y_i/\sigma), \quad (1 + \xi y_i/\sigma) > 0 \tag{1.2.6}$$

Si $\xi = 0$:

$$\ell(\sigma) = -k \log \sigma - (1/\sigma) \sum_{i=1}^{k} y_i \tag{1.2.7}$$

Tout comme avec les maximums, des techniques numériques sont nécessaires pour la maximisation.

Pour les niveaux de retour, nous avons :

$$\begin{aligned} \Pr\{X > x \mid X > u\} &= \Big[1 + \xi\Big(\frac{x-u}{\sigma}\Big)\Big]^{-1/\xi} \\ \Rightarrow \Pr\{X > x\} &= \zeta_u \Big[1 + \xi\Big(\frac{x-u}{\sigma}\Big)\Big]^{-1/\xi} \end{aligned} \tag{1.2.8}$$

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 :

$$\begin{aligned} \frac{1}{m} &= \zeta_u \Big[1 + \xi\Big(\frac{x-u}{\sigma}\Big)\Big]^{-1/\xi} \\ \Rightarrow x_m &= u + \frac{\sigma}{\xi}\Big[(m\zeta_u)^{\xi} - 1\Big], \quad x_m > u \end{aligned} \tag{1.2.9}$$

Si $\xi = 0$:

$$x_m = u + \sigma \log(m\zeta_u) \tag{1.2.10}$$

Il est souvent plus utile d'obtenir le niveau qui est dépassé en moyenne une fois chaque N années. S'il y a n_y observations par années, le niveau de retour *N-années* est défini comme suit :

$$z_N = u + \frac{\sigma}{\xi}\Big[(N n_y \zeta_u)^{\xi} - 1\Big] \tag{1.2.11}$$

Si $\xi = 0$:

$$z_N = u + \sigma \log(N n_y \zeta_u) \tag{1.2.12}$$

Pour estimer les niveaux de retour, il suffit de remplacer les paramètres par leur estimation. Pour ξ et σ, il s'agit tout simplement de $\hat{\xi}$ et $\hat{\sigma}$ obtenus avec la méthode du maximum de vraisemblance et $\hat{\zeta}_u = k/n$, la proportion des données observées qui dépassent u. De plus, $\mathrm{Var}(\hat{\zeta}_u) \approx \hat{\zeta}_u/(1-\hat{\zeta}_u)$, car $k \sim \mathrm{Bin}(n, \zeta_u)$.

En appliquant la méthode delta :

$$\mathrm{Var}(\hat{x}_m) \approx \nabla x_m^\top V \nabla x_m \tag{1.2.13}$$

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

$$\begin{aligned} \nabla x_m^\top &= \left[\frac{\partial x_m}{\partial \zeta_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi}\right] \\ &= \left[\sigma \xi \zeta_u^{\xi-1},\ \xi^{-1}\{(m\zeta_u)^\xi - 1\},\ -\sigma\xi^{-2}\{(m\zeta_u)^\xi - 1\} + \sigma\xi^{-1}(m\zeta_u)^\xi \log(m\zeta_u)\right]\Big|_{(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})} \end{aligned} \tag{1.2.14}$$

De plus, pour la deuxième méthode de sélection de seuil, nous avons :

$$\mathrm{Var}(\sigma^*) \approx \nabla \sigma^{*\top} V \nabla \sigma^* \tag{1.2.15}$$

et

$$\begin{aligned} \nabla \sigma^{*\top} &= \left[\frac{\partial \sigma^*}{\partial \sigma_u}, \frac{\partial \sigma^*}{\partial \xi}\right] \\ &= [1, -u] \end{aligned} \tag{1.2.16}$$

Tout comme avec les maximums, ces quatre graphiques sont utiles pour la validation du modèle :

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

Le graphique *P-P* est constitué des points :

$$\{((i/(k+1), \hat{H}(y_{(i)}));\ i = 1, \ldots, k\}$$

où

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

Le graphique *Q-Q* est constitué des points :

$$\{(\hat{H}^{-1}(i/(k+1)), y_{(i)});\ i = 1, \ldots, k\}$$

où

$$\hat{H}^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}\Big[y^{-\hat{\xi}} - 1\Big]$$

Le graphique du niveau de retour est constitué des points :

$$\{(m, \hat{x}_m)\}$$

où

$$\hat{x}_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\Big[(m\hat{\zeta}_u)^{\hat{\xi}} - 1\Big]$$

1.3 Modèles additifs généralisés et splines

À compléter...

1.4 Techniques de bootstrap

À compléter...

Chapitre 2

Article étudié

2.1 Introduction

Dans le but d'aller au-delà des notions préliminaires du chapitre 1, un article en particuler a été étudié dans le cadre du projet de recherche. Plusieurs articles ont été brièvement explorés pour finalement choisir l'article qui répondait le plus aux bessoins du présent projet. L'article choisi (Chavez-Demoulin et collab. (2016)) est tout de même récent, est de longueur et difficulté convenables et se penche sur un sujet spécifique intéressant en plus de fournir un implémentation en R de la méthode proposée. À noter que si le temps l'avait permis, Thiombiano et collab. (2017) aurait également été un article intéressant à étudier.

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. Plus précisément, ce sont les paramètres du modèle de valeurs extrêmes choisi qui dépendront des covariables disponibles. C'est à dire que contrairement à ce qui a été vu aux sections 1.1 et 1.2 où les paramètres des modèles sont uniques étant donné qu'il sont obtenus par la méthode du maximum de vraisemblance avec l'ensemble des données, les auteurs proposent, ici, d'avoir des paramètres différents pour chaque niveaux de covariables. Cette approche vise à améliorer la calibration des lois aux données et du même coup rendre l'inférence plus crédible et précise. La méthode décrite propose de modéliser la fréquence de pertes avec un processus de Poisson non-homogène dont le paramètre dépend des covariables. Cette démarche implique l'utilisation des modèles additifs généralisés, des splines ainsi que la méthode du maximum de vraisemblance pénalisé. D'un autre côté, la sévérité est modélisée avec une distribution Pareto généralisée non-stationnaire dont les paramètre dépendent des variables exogènes. [1]. Ici, les splines de lissage ne peuvent pas être directement appliqués et l'article propose un algorithme basé sur les paramètres orthogonaux pour palier à la situation. Pour finir, les auteurs appliquent la méthodologie proposée à des données de pertes de risques opérationnels ainsi qu'à des données simulées pour valider la proposition.

La méthodologie proposée dans l'article peut s'appliquer à plusieurs contextes et types de données qui respectent certains 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.

1. L'article présente également la méthode avec la famille *GEV*, mais brièvement.

— Les données contiennent des évènement (valeurs) extrêmes.

— Les données contiennet des covariables significatives.

Il est important de noter que l'article utilise la méthodologie proposée avec des pertes, mais travailler avec des gains serait également possible. Même chose par rapport à des très grandes valeurs ou des très petites valeurs. Les critères mentionnés sont seulement générales et représentent une situation idéale. Quelqu'un pourrait très bien tenter d'appliquer la méthode même si un ou plusiers critères ne sont pas respectés et obtenir des résultats concluants.

À noter que les sections de l'article qui présentent les techniques de modélisation classiques ne seront pas discutés dans le présente section étant donné que l'ensemble de cette information peut être trouvé dans les sections 1.1 et 1.2. La brève section sur l'approche dynamique avec les maximums ne sera également pas élaborée, car le focus de l'article ainsi que celui du chapitre 3 est vraiment l'approche *POT*.

À noter également que le paramètre d'échelle σ sera noté β pour la suite des choses pour suivre la notation des auteurs de l'article.

2.2 Approche dynamique de la théorie des valeurs extrêmes

Comme mentionnée à la section 2.1, la nouvelle méthode proposée laisse les paramètres du modèle dépendent des covariables. Cette dépendance peut être paramétrique, non-paramétrique ou bien semi-paramétrique et elle peut également inclure des intéractions entre les variables. L'idée générale du modèle de dépendance entre les paramètres et les covariables peut être représenté par l'équation suivante :

$$g_k(\theta_k) = f_k(x) + h_k(x), \quad k \in \{1, \ldots, p\} \tag{2.2.1}$$

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.

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.

Ensuite, le vecteur des paramètres du modèle θ peut être estimé en maximisant la log-vraisemblance pénalisée :

$$\ell(\theta; \cdot) - \sum_{k=1}^{p} \left(\gamma_k \int_A h_k''(t)^2 \, dt \right) \tag{2.2.2}$$

où $\ell(\theta; \cdot)$ est tout simplement la log-vraisemblance du modèle, *POT* ici. Le terme de pénalité est une technique répandue qui a pour but d'éviter du surapprentissage des données. À noter que plus γ_k est grand, plus la courbe obtenue est lisse et vice-versa.

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

$$\lambda = \lambda(x,t) = \exp(f_\lambda(x) + f_\lambda(t)) \tag{2.2.3}$$

où f_λ est une fonction pour les différents niveaux d'une variable catégorique x et h_λ est une fonction générale qui ne dépend pas de paramètres spécifiques comme c'est le cas pour f_λ. Ensuite, l'application d'un modèle additif généralisé mène à un estimé $\hat{\lambda}$ qui sera utile pour le reste de la modélisation. Cette étape peut être complétée avec l'aide du logiciel R et du paquetage `mgcv`.

L'approche est très semblable à l'équation 2.2.3 pour ce qui est de la modélisation de la sévérité de la perte. Cependant, une étape supplémentaire est nécessaire pour 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. De ce fait, β est reparamétrisé comme suit :

$$\nu = \log((1+\xi)\beta) \;\; \xi > -1 \Rightarrow \beta = \frac{\exp(\nu)}{1+\xi} \tag{2.2.4}$$

ce qui est orthognal à ξ. Voir Cox et Reid (1987) pour plus de détails. En suit la log-vraisemblance reparamétrisée :

$$\ell(\theta;\cdot) \longrightarrow \ell^r(\xi,\nu;y) = \ell\Big(\xi, \frac{\exp(\nu)}{1+\xi}; y\Big) \tag{2.2.5}$$

où Y représente le vecteur des excès de seuil u.

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

$$\xi = \xi(x,t) = f_\xi(x) + f_\xi(t) \tag{2.2.6}$$

$$\nu = \nu(x,t) = f_\nu(x) + f_\nu(t) \tag{2.2.7}$$

$$\Rightarrow \beta = \beta(x,t) = \frac{\exp(\nu(x,t))}{1+\xi(x,t)} \tag{2.2.8}$$

Les équations 2.2.6 et 2.2.7 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.

Contrairement à la méthode présentée pour la fréquence (équation 2.2.3), un modèle additif généralisé n'est pas suffissant ici et un algorithme est donc développé pour être en mesure d'obtenir les paramètres estimés. Pour être en mesure d'ajuster des fonctions h_ξ et h_ν qui sont assez lisses ont vecteurs observés z_i, la log-vraisemblance de l'équation 2.2.2 est utilisée :

$$\ell^p(f_\xi, h_\xi, f_\nu, h_\nu; z_1,\ldots,z_n) = \ell^r(\xi,\nu;y) - \gamma_\xi \int_0^T h_\xi''(t)^2\,dt - \gamma_\nu \int_0^T h_\nu''(t)^2\,dt \tag{2.2.9}$$

où

— $\gamma_\xi, \gamma_\nu \geq 0$ sont les paramètres de lissage.
— $y = (y_{t1}, \ldots, y_{tn})$.
— $\ell^r(\xi, \nu; y) = \sum_{i=1}^n \ell^r(\xi_i, \nu_i; y_{ti}) = \sum_{i=1}^n \ell^r(\xi_i, \frac{\exp(\nu_i)}{1+\xi_i}; y_i)$

En combinant les conditions énumérés à la section 1.3 et (Green et Silverman, 1993, p. 13), pour une spline naturelle cubique h ayant comme noeuds $s_1, \ldots, s_m$:

$$\int_0^T h''(t)^2 dt = h^\top K h \tag{2.2.10}$$

où

— $h = (h_{s1}, \ldots, h_{sn}) = (h(s_1), \ldots, h(s_n))$,
— K est une matrice symétrique de m colonnes de rang $m-2$ qui dépend des noeuds $s_1, \ldots, s_m$

Du coup, l'équation 2.2.9 peut être réécrite comme suit :

$$\ell^p(f_\xi, h_\xi, f_\nu, h_\nu; z_1, \ldots, z_n) = \ell^r(\xi, \nu; y) - \gamma_\xi {h_\xi}^\top K h_\xi - \gamma_\nu {h_\nu}^\top K h_\nu \tag{2.2.11}$$

C'est avec l'obtention de cette formule que devient possible le développement d'un algorithme permettant d'obtenir $\hat{\xi}$ et $\hat{\beta}$. Des intervalles de confiance pour les paramètres sont également obtenus avec l'aide d'une technique de bootstrap. Cet algorithme est présenté à la section 2.5.

Après avoir obtenu les paramètres estimés du modèle utilisé et avant d'appliquer tout genre d'inférence avec le modèle obtenu, il est nécessaire d'évaluer la qualité de celui-ci. La méthode proposée ici repose sur le fait que si les excès Y_{ti} suivent une distribution Pareto généralisée avec comme paramètre ξ et β, $R_i = 1 - G_{\xi_i, \beta_i}(Y_{ti})$, $i \in \{1, \ldots, n\}$ forme approximativement un échantillon aléatoire d'une loi uniforme $U(0, 1)$. Ainsi, il est possible de vérifier si les valeurs r_i se comportent comme des variables aléatoires indépendantes suivant une loi exponentielle standard où

$$r_i = -\log(1 - G_{\hat{\xi}_i, \hat{\beta}_i}(y_{ti})), \quad i \in \{1, \ldots, n\} \tag{2.2.12}$$

Finalement, lorsque tous les paramètres sont obtenus et que l'adéquation du modèle est validé, l'inférence et plus particulièrement le calcul de mesures de risques pour des covariables et un point dans le temps en particulier est possible :

$$\widehat{\mathrm{VaR}_\alpha} = u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{1-\alpha}{\hat{\lambda}/n'} \right)^{-\hat{\xi}} - 1 \right) \tag{2.2.13}$$

$$\widehat{\mathrm{ES}_\alpha} = \begin{cases} \frac{\widehat{\mathrm{VaR}_\alpha} + \hat{\beta} + u\hat{\xi}}{1-\hat{\xi}}, & \hat{\xi} \in (0, 1) \\ \infty, & \hat{\xi} \geq 1 \end{cases} \tag{2.2.14}$$

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 avec une régression logistique. Des intervalles de confiance sont obtenus pour les mesures de risque avec une technique de bootstrap.

Un dernier court point à couvrir est le choix du paramètre de lissage souvent reconnu comme le nombre de degrés de liberté. L'utilisation du critère d'information d'Akaike (AIC) est la technique proposée par les auteurs.

$$\mathrm{AIC} \propto -2\ell(\theta; \cdot) + 2\mathrm{Df} \tag{2.2.15}$$

2.3 Application à des données réelles

La présente section discute des fait saillants de l'application de la méthodologie à des données réelles faite par les auteurs de l'article. Voir **?** pour de plus amples détails.

2.4 Discussion

2.5 Algorithmes

Chapitre 3

Application à des données réelles

3.1 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 deux jeux de données contiennent également différentes statistiques de ces catastrophes. Voici une liste des variables disponibles :

- `Year` : l'année d'occurence de la catastrophe.
- `Quarter` : le trimestre d'occurence de la catastrophne.
- `Date` : la date d'occurence complète de la catastrophe.
- `FirstDay` : la date de la première journée d'occurence de la catastrophe.
- `LastDay` : la date de la dernière journée de la catastrophe (seulement disponible pour l'Australie).
- `Event` : une description de la catastrophe.
- `Type` : le type de catastrophe.
- `Location` : une description du lieu de la catastrophe.
- `OriginalCost` : coût original de la catastrophe en millions de AUD ou NZD.
- `NormCost2011` : coût normalisé en millions de dollars de 2011 (inflation, richesse et population).
- `NormCost2014` : coût normalisé en millions de dollars de 2014 (inflation, richesse et population).

Les tables 3.1 et 3.2 contiennent un résumé statistique des données australiennes. Les tables 3.3 et 3.4 contiennent le même type de résumé pour la Nouvelle-Zélande.

	Mean	Std.Dev	Min	Median	Max	N.Valid	Pct.Valid
NormCost2011	254	589	2	66	4296	190	92
NormCost2014	288	639	2	77	4606	206	100
OriginalCost	104	295	1	15	2388	206	100
Quarter	2	1	1	2	4	206	100
Year	1995	12	1967	1998	2014	206	100

TABLE 3.1 – Résumé statistique des variables numériques pour l'Australie

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Bushfire	26	13	13	13	13
Cyclone	33	16	29	16	29
Earthquake	4	2	31	2	31
Flood	25	12	43	12	43
Flood, Storm	27	13	56	13	56
Hailstorm	33	16	72	16	72
Other	3	1	73	1	73
Storm	54	26	100	26	100
Tornado	1	0	100	0	100
<NA>	0			0	100
Total	206	100	100	100	100

TABLE 3.2 – Distribution du Type de catastrophe pour l'Australie

	Mean	Std.Dev	Min	Median	Max	N.Valid	Pct.Valid
NormCost2011	17.169	40.885	0.010	5.270	371.120	129.000	88.966
NormCost2014	137.777	1437.470	0.000	5.870	17320.680	145.000	100.000
OriginalCost	124.591	1369.525	0.000	3.500	16500.000	145.000	100.000
Quarter	2.262	1.143	1.000	2.000	4.000	145.000	100.000
Year	1999.448	10.674	1968.000	2001.000	2014.000	145.000	100.000

TABLE 3.3 – Résumé statistique des variables numériques pour la Nouvelle-Zélande

Pour chaque jeu de données, une nouvelle variable du montant de catastrophe est créée à partir de la variable `NormCost2014` (`CostUS2019`). Cette nouvelle variable représente le coût ajusté au niveau du 30 juin 2019 en considérant l'indice du prix à la consommation de chaque pays et le coût est également converti en dollar américain. La table 3.5 contient les chiffres utilisés [1].

Étant donné que les deux jeux de données sont très semblables, ceux-ci sont regroupés en un seul jeu de données et une variable `Country` est rajoutée. Étant donné le nombre limité de données disponibles et pour rendre l'analyse pertinente et possible, seulement les variables `CostUS2019`, `Year`, `Country` et `Type` sont conservées. Les figures 3.1, 3.2, 3.3 ainsi que la table 3.6 résument bien le jeu de données final utilisé pour commencer l'analyse.

1. https ://www.rateinflation.com/consumer-price-index
https ://www.exchange-rates.org/Rate/AUD/USD/6-30-2019

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Cyclone	4	3	3	3	3
Earthquake	7	5	8	5	8
Flood	58	40	48	40	48
Flood, Storm	9	6	54	6	54
Hailstorm	8	6	59	6	59
Other	10	7	66	7	66
Power outage	2	1	68	1	68
Storm	32	22	90	22	90
Tornado	11	8	97	8	97
Weather	4	3	100	3	100
<NA>	0			0	100
Total	145	100	100	100	100

TABLE 3.4 – Distribution du Type de catastrophe pour la Nouvelle-Zélande

	IPC 2014	IPC 2019	Taux de change 2019
AUS	105.9000	114.8000	0.7031
NZ	974.7000	1039.0000	0.6722

TABLE 3.5 – Valeurs utilisées pour CostUS2019

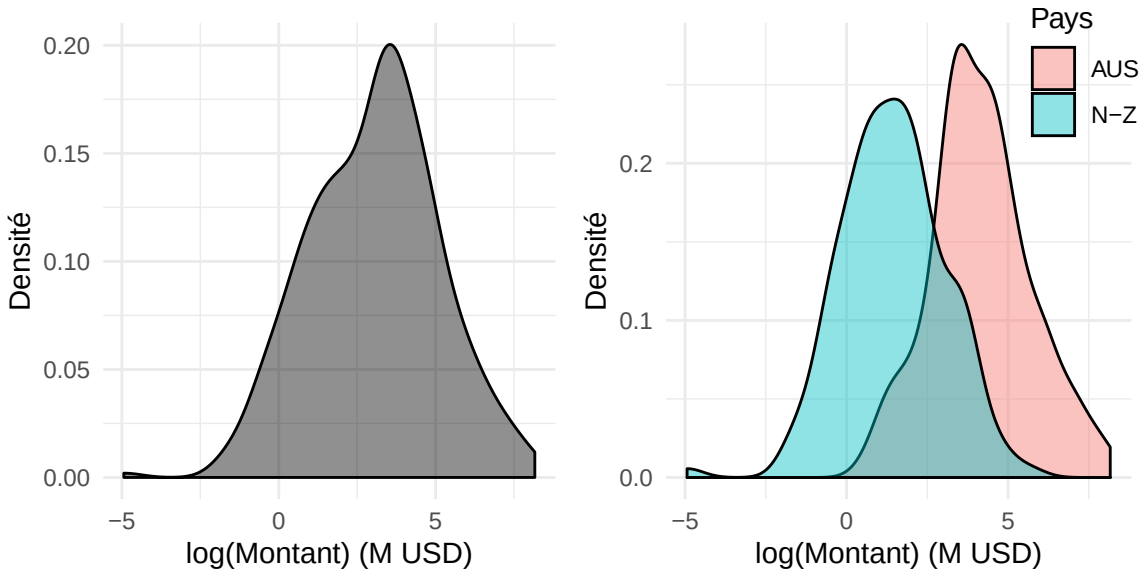

FIGURE 3.1 – Densité du logarithme du montant des catastrophes

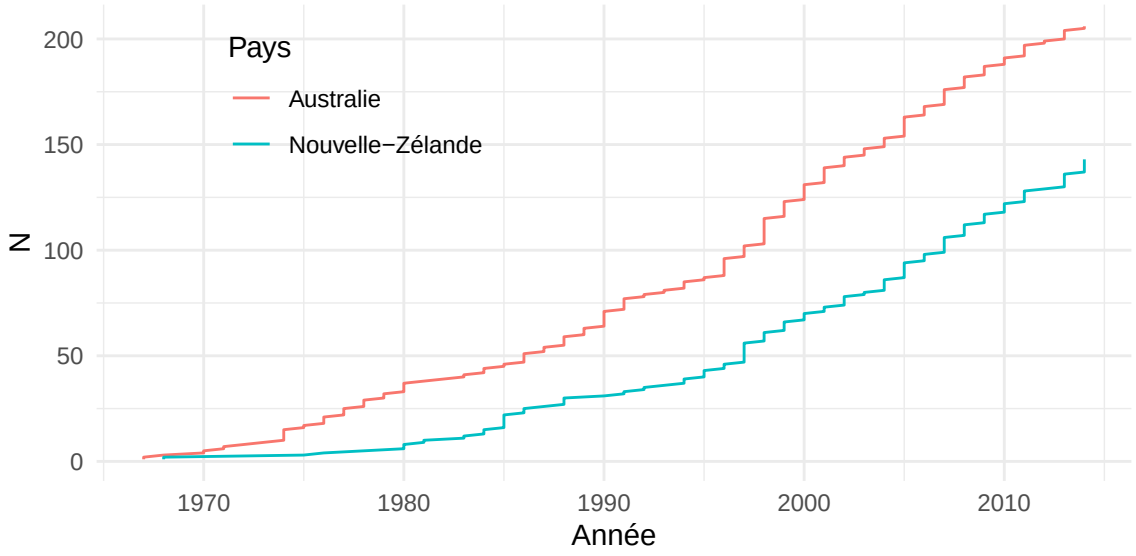

FIGURE 3.2 – Nombre cumulatif de catastrophes par pays

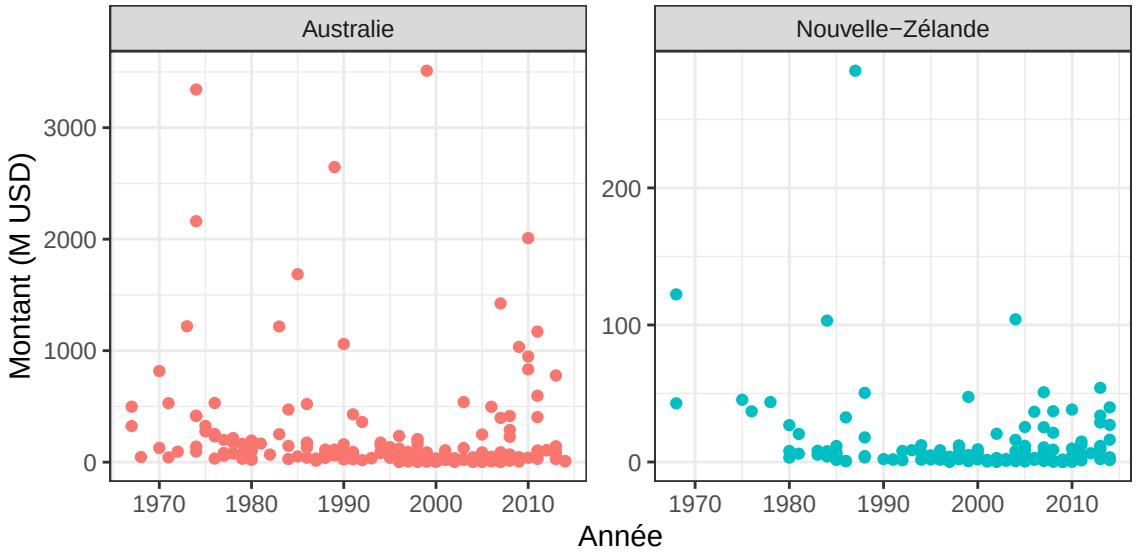

FIGURE 3.3 – Évolution des catastrophes par pays

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	10	315	824	0	28	82	2648
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 3.6 – Résumé statistique des montants de catastrophes par Type

3.2 Approche classique

Avant d'essayer la nouvelle méthode proposée, les méthodes classiques présentées à au chapitre 1 sont appliquées pour être en mesure d'évaluer la valeur de la nouvelle méthode. Plus spécifiquement, l'approche *POT* présentée à la section 1.2 est appliquée dans la présente section étant donné que c'est ce type de modèle qui est appliqué dans les prochaines sections. À noter que les différents calculs de cette section nécessitant des techniques numériques sont faits avec le paquetage `ismev`. L'approche ici est donc de prendre l'ensemble des données, de choisir un seuil u approprié et d'estimer les paramètres de la loi Pareto généralisée avec les excès de seuil, le tout en tenant compte d'aucunes variables exogènes. Comme mentionné, la première étape de cette méthodologie est le choix de la valeur du seuil u. La première méthode demande de tracer le *mean residual plot* dont les points mentionnés dans l'équation 1.2.5. Après un certaine valeur u pour lequel la distribution est appropriée, le graphique devrait être linéaire en u. La figure 3.4 montre que déjà à partir de petites valeurs, cette conditon est respectée, il est ensuite difficile, voire subjectif de choisir une valeur précise. Ici, le graphique suggère de sélectionner une valeur entre 0 et 10.

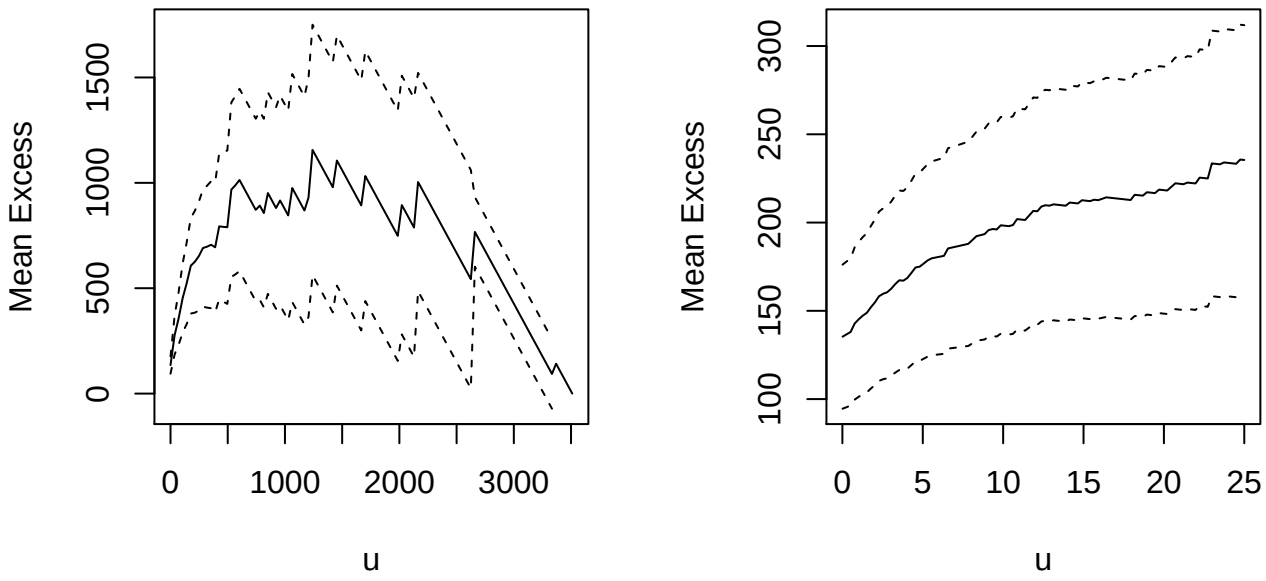

FIGURE 3.4 – *Mean residual plot*

La deuxième méthode mentionnée à la section 1.2 propose d'estimer les paramètres du modèle pour différentes valeurs de u. Comme expliqué dans cette section, σ^* et ξ devraient rester constants au-delà de u_0. La figure 3.5 montre des résulats plus concluants que la figure 3.4. En effet, les valeurs de σ^* et de ξ deviennent constantes environ à $u = 10$. L'analyse des deux différentes techniques de sélection de seuil mènent donc à une sélection de $u = 10$. Environ 64% des données dépassent ce seuil, un bon nombre de données est donc utile pour la modélisation des excès.
Après avoir sélectionné le seuil, il faut ensuite estimer les paramètres de la loi Pareto généralisée avec la méthode du maximum de vraisemblance. Les différents détails de ce calcul se trouvent dans

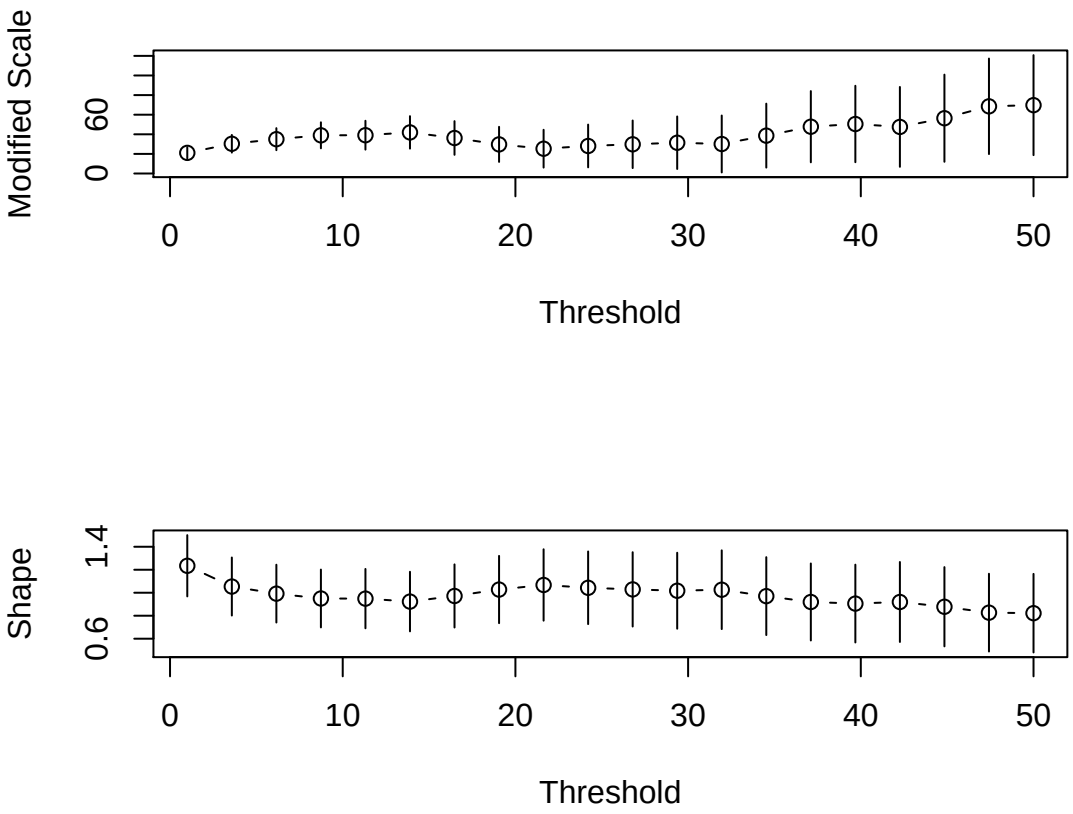

FIGURE 3.5 – Estimation des paramètres du modèle pour différents seuils

les équations 1.2.6 et 1.2.7. Voici les résultats obtenus avec un intarvalle de confiance de 95 %.

$$(\hat{\sigma}, \hat{\xi}) = (49.329951;\ 0.941132)$$
$$\ell(\hat{\sigma}, \hat{\xi}) = -1308.013$$
$$\mathrm{Var}(\hat{\sigma}) = 41.9446912$$
$$\mathrm{Var}(\hat{\xi}) = 0.01673518$$

$$\hat{\zeta}_u = 0.6418338$$
$$\mathrm{Var}(\hat{\zeta}_u) = 0.000658691$$

$$IC_{\hat{\sigma}} = (36.6362993;\ 62.023603)$$
$$IC_{\hat{\xi}} = (0.6875822;\ 1.194682)$$
$$IC_{\hat{\zeta}_u} = (0.5915314;\ 0.6921362)$$

Suite aux résultats obtenus, les graphiques de validation mentionnés à la section 1.2 peuvent être tracés pour juger de la qualité du modèle. La figure 3.6 montre des résultats qui ne sont pas désastreux, mais qui sont loin d'être concluants en ce qui concerne l'ajustement des données au modèle proposé dans la présente section. Le graphique P-P est tout de même adéquat, mais les trois autres graphiques sont loins de l'être. Les graphiques Q-Q et celui de la densité obtenue montrent que le modèle représentent mal les données utilisées et le graphique des niveaux de retour montre que dès que la période de retour est un peu élevé, le niveau obtenu est extrêmement volatile. Pour conclure, le modèle proposé dans cette section représente un modèle de base qui considère toutes les données de la manière sans égard aux informations supplémentaires disponibles par rapport aux montants des catastrophes. Il n'est donc pas surprenant de voir que cette aproche n'est pas tout à fait adéquate dans le cas présent, mais reste que cette approche est viable, rapide et peut être une bonne solution lorsque seulement les montants sont disponibles. Il est également important de savoir que le modèle testé dans cette section est la base de tous autres modèles plus avancés, tous comme les modèles qui seront testés aux prochaines sections.

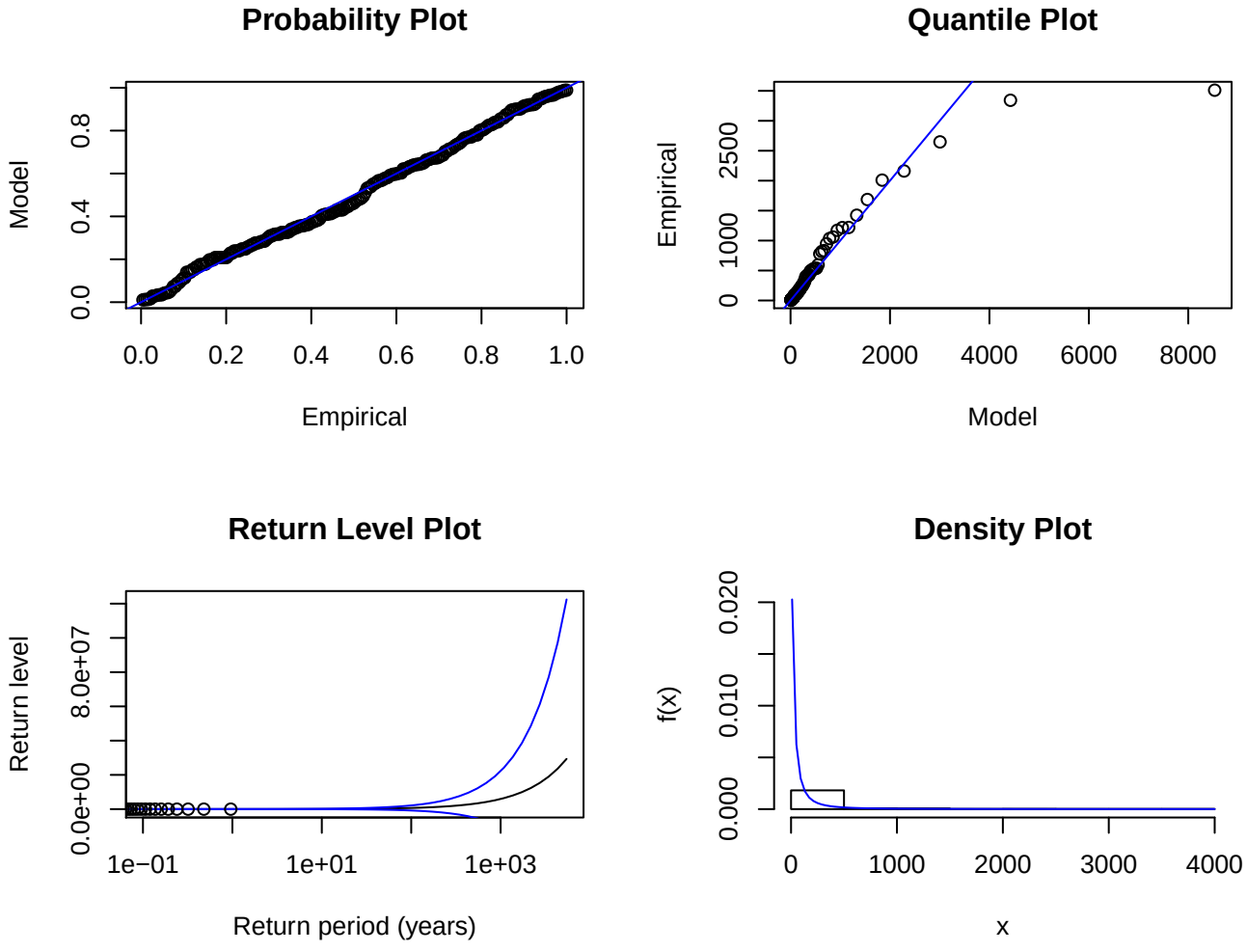

FIGURE 3.6 – Graphiques de validation du modèle

3.3 Approche dynamique à deux variables

Cette section applique la méthodologie et le modèle proposés au chapitre 2 avec deux variables, l'année et le pays. On considère donc une variable numérique de temps ainsi qu'une variable catégorique à deux niveaux. Seulement deux variables sont initialemment utilisées étant donné que c'est ce que l'article étudié utilise pour ilustrer la méthode. De plus, étant donné le nombre limité de données disponibles, il est préférable de ne pas commencer avec un nombre de sous-groupes élevés. Une troisième variable sera considérée à la section 3.4.

3.4 Approche dynamique à trois variables

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.

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.