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

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 **??**, 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.

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

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

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 **??**.

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

Pour n assez grand, un développement de Taylor donne :

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

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

□

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

Chapitre 3

Application à des données réelles

Pour mettre en application l'approche proposée par *chavez2016extreme****, 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 **??** et **??** contiennent un résumé statistique des données australiennes :

	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

Les tables **??** et **??** contiennent le même type de résumé pour la Nouvelle-Zélande :

	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	41	0	5	371	129	89
NormCost2014	138	1437	0	6	17321	145	100
OriginalCost	125	1370	0	4	16500	145	100
Quarter	2	1	1	2	4	145	100
Year	1999	11	1968	2001	2014	145	100

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 **??** 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 **??**, **??**, **??** et **??** résument bien le jeu de données final utilisé pour 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

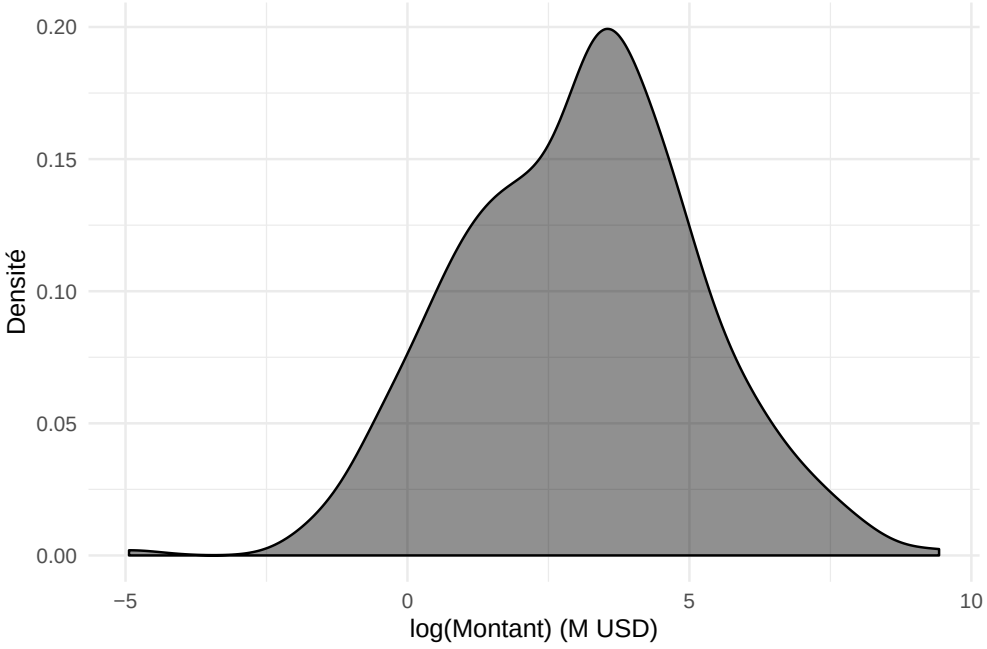

FIGURE 3.1 – Densité du logarithme du montant des catastrophes

Conclusion