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

Plan de la présentation

UNIVERSITÉ LAVAL

Motivations

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

Théorie classique des valeurs extrêmes

Soit $X_1, \ldots, X_n$, une séquence de variables indépendantes, identiquement distribuées et avec fonction de répartition marginale F.

Théorie classique des valeurs extrêmes

Soit $X_1, \ldots, X_n$, une séquence de variables indépendantes, identiquement distribuées et avec fonction de répartition marginale F.

Soit $M_n = \max\{X_1, \ldots, X_n\}$, le maximum des n observations.

UNIVERSITÉ LAVAL

Théorie classique des valeurs extrêmes

Soit $X_1, \ldots, X_n$, une séquence de variables indépendantes, identiquement distribuées et avec fonction de répartition marginale F.

Soit $M_n = \max\{X_1, \ldots, X_n\}$, le maximum des n observations.

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

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

UNIVERSITÉ LAVAL

Théorie classique des valeurs extrêmes

Soit $X_1, \ldots, X_n$, une séquence de variables indépendantes, identiquement distribuées et avec fonction de répartition marginale F.

Soit $M_n = \max\{X_1, \ldots, X_n\}$, le maximum des n observations.

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

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

G appartient à une des distributions de valeur extrême.

UNIVERSITÉ LAVAL

Théorie classique des valeurs extrêmes

Les trois dernières lois sont les seules distributions limites pour M_n^*.

Théorie classique des valeurs extrêmes

Les trois dernières lois sont les seules distributions limites pour M_n^*.

Une meilleure analyse est possible :

Famille de distributions d'extremum généralisée (GEV)

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

Trois paramètres : μ (position), σ (échelle), ξ (forme).

UNIVERSITÉ LAVAL

Première méthodologie

- Les données sont groupées en séquence de longueur n.

Première méthodologie

- Les données sont groupées en séquence de longueur n.
- Le maximum de chaque séquence est calculé.

Première méthodologie

- Les données sont groupées en séquence de longueur n.
- Le maximum de chaque séquence est calculé.
- Une distribution *GEV* est calibrée à ces maximums.

Première méthodologie

- 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.

Méthode par l'approche POT

Soit $X_1, \ldots, X_n$. Nous considérons comme évènements extrêmes ceux qui dépassent un certain seuil u.

Nous savons déjà que $\Pr\{\max\{X_1, \ldots, X_n\} \leq z\} \approx G(z)$ (1).

UNIVERSITÉ LAVAL

Méthode par l'approche POT

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

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

Méthode par l'approche POT

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

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

Famille de distributions Pareto généralisée.

Paramètres de H.

Deuxième méthodologie

- Les données brutes représentent une séquence de valeurs iid $x_1, \ldots, x_n$.

Deuxième méthodologie

- Les données brutes représentent une séquence de valeurs iid $x_1, \ldots, x_n$.
- Les valeurs extrêmes sont identifiées en définissant un seuil u.

UNIVERSITÉ LAVAL

Deuxième méthodologie

- Les données brutes représentent une séquence de valeurs iid $x_1, \ldots, x_n$.
- Les valeurs extrêmes sont identifiées en définissant un seuil u.
- Les observations qui dépassent u sont définies par $x_{(1)}, \ldots, x_{(k)}$ et les excès de seuil par $y_j = x_{(j)} - u$.

UNIVERSITÉ LAVAL

Deuxième méthodologie

- Les données brutes représentent une séquence de valeurs iid $x_1, \ldots, x_n$.
- Les valeurs extrêmes sont identifiées en définissant un seuil u.
- Les observations qui dépassent u sont définies par $x_{(1)}, \ldots, x_{(k)}$ et les excès de seuil par $y_j = x_{(j)} - u$.
- Une distribution Pareto généralisée est calibrée avec les y_j.

UNIVERSITÉ LAVAL

Deuxième méthodologie

- Les données brutes représentent une séquence de valeurs iid $x_1, \ldots, x_n$.
- Les valeurs extrêmes sont identifiées en définissant un seuil u.
- Les observations qui dépassent u sont définies par $x_{(1)}, \ldots, x_{(k)}$ et les excès de seuil par $y_j = x_{(j)} - u$.
- Une distribution Pareto généralisée est calibrée avec les y_j.
- La distribution peut être manipulée.

UNIVERSITÉ LAVAL

Deuxième méthodologie

Défi : choix de la valeur du seuil.

Deuxième méthodologie

Défi : choix de la valeur du seuil.

Deux méthodes :

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

UNIVERSITÉ LAVAL

Article étudié

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

- Récent
- Longueur et difficulté convenables
- Sujet spécifique intéressant
- Implémentation en R

UNIVERSITÉ LAVAL

Introduction

L'article choisi présente une nouvelle méthodologie....

Introduction

L'article choisi présente une nouvelle méthodologie....

- Ce sont les paramètres du modèle choisi qui dépendront des covariables.

Introduction

L'article choisi présente une nouvelle méthodologie....

- Ce sont les paramètres du modèle choisi qui dépendront des covariables.
- Améliorer la calibration des lois.
- Inférence plus crédible et précise.

Introduction

- Processus de Poisson non-homogène.

Introduction

- Processus de Poisson non-homogène.
- Modèles additifs généralisés, splines de lissage et méthode du maximum de vraisemblance pénalisé.

Introduction

- Processus de Poisson non-homogène.
- Modèles additifs généralisés, splines de lissage et méthode du maximum de vraisemblance pénalisé.
- Pareto généralisée non-stationnaire.

Introduction

- Processus de Poisson non-homogène.
- Modèles additifs généralisés, splines de lissage et méthode du maximum de vraisemblance pénalisé.
- Pareto généralisée non-stationnaire.
- Coles et collab. (2001).

UNIVERSITÉ LAVAL

Critères

- Pertes aléatoires encourues à des temps aléatoires.

Critères

- Pertes aléatoires encourues à des temps aléatoires.
- But de l'analyse.

Critères

- Pertes aléatoires encourues à des temps aléatoires.
- But de l'analyse.
- Nombre suffisant d'observations.

Critères

- Pertes aléatoires encourues à des temps aléatoires.
- But de l'analyse.
- Nombre suffisant d'observations.
- Évènement extrêmes.

Critères

- Pertes aléatoires encourues à des temps aléatoires.
- But de l'analyse.
- Nombre suffisant d'observations.
- Évènement extrêmes.
- Covariables significatives.

Critères

- Pertes aléatoires encourues à des temps aléatoires.
- But de l'analyse.
- Nombre suffisant d'observations.
- Évènement extrêmes.
- Covariables significatives.
- Notes.

Approche dynamique de la théorie des valeurs extrêmes

Idée générale du modèle :

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

où :

- θ est le vecteur des p paramètres du modèle,

Approche dynamique de la théorie des valeurs extrêmes

Idée générale du modèle :

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

où :

- θ est le vecteur des p paramètres du modèle,
- g_k est une fonction de lien,

UNIVERSITÉ LAVAL

Approche dynamique de la théorie des valeurs extrêmes

Idée générale du modèle :

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

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,

Approche dynamique de la théorie des valeurs extrêmes

Idée générale du modèle :

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

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

Fréquence

Le nombre d'excès du seuil u suit un processus de Poisson non-homogène avec comme taux :

$$\lambda = \lambda(x, t) = \exp(f_\lambda(x) + h_\lambda(t)). \tag{3}$$

Reparamétrisation

S'assurer que les procédures de calibration des paramètres ξ et β aux données convergent.

Reparamétrisation

S'assurer que les procédures de calibration des paramètres ξ et β aux données convergent.

Deux paramètres orthogonaux.

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

Sévérité

Comme pour λ, on définit ξ et ν comme suit :

$$\xi = \xi(x,t) = f_\xi(x) + h_\xi(t) \tag{4}$$

$$\nu = \nu(x,t) = f_\nu(x) + h_\nu(t) \tag{5}$$

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

UNIVERSITÉ LAVAL

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Basés sur les estimateurs $\hat{f}_{\xi}, \hat{h}_{\xi}, \hat{f}_{\nu}$ et $\hat{h}_{\nu}$.

UNIVERSITÉ LAVAL

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Basés sur les estimateurs $\hat{f}_\xi, \hat{h}_\xi, \hat{f}_\nu$ et $\hat{h}_\nu$.

Obtenus à partir des vecteurs observés $z_i = (t_i, x_i, y_{t\,i})$ où :

- $i \in \{1, \ldots, n\}$,

UNIVERSITÉ LAVAL

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Basés sur les estimateurs $\hat{f}_\xi, \hat{h}_\xi, \hat{f}_\nu$ et $\hat{h}_\nu$.

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,

UNIVERSITÉ LAVAL

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Basés sur les estimateurs $\hat{f}_\xi, \hat{h}_\xi, \hat{f}_\nu$ et $\hat{h}_\nu$.

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,

UNIVERSITÉ LAVAL

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Basés sur les estimateurs $\hat{f}_\xi, \hat{h}_\xi, \hat{f}_\nu$ et $\hat{h}_\nu$.

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 $Y_{t\,i}$,

UNIVERSITÉ LAVAL

Estimation

Les équations (4) et (5) sont donc celles qui seront estimées pour être en mesure d'obtenir $\hat{\xi}$ et $\hat{\beta}$.

Basés sur les estimateurs $\hat{f}_\xi, \hat{h}_\xi, \hat{f}_\nu$ et $\hat{h}_\nu$.

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 $Y_{t\,i}$,
- $Y_{t\,i}$ représente les excès de seuil u.

UNIVERSITÉ LAVAL

Mesures de risque

$$\widehat{\mathrm{VaR}_\alpha} = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{1-\alpha}{\hat{\lambda}/n'}\right)^{-\hat{\xi}} - 1\right) \tag{6}$$

Mesures de risque

$$\widehat{\mathrm{VaR}_\alpha} = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{1-\alpha}{\hat{\lambda}/n'}\right)^{-\hat{\xi}} - 1\right) \tag{6}$$

$$\widehat{\mathrm{ES}_\alpha} = \begin{cases} \frac{\widehat{\mathrm{VaR}_\alpha}+\hat{\beta}+u\hat{\xi}}{1-\hat{\xi}}, & \hat{\xi} \in (0,1) \\ \infty, & \hat{\xi} \geq 1 \end{cases} \tag{7}$$

Mesures de risque

$$\widehat{\mathrm{VaR}_\alpha} = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{1-\alpha}{\hat{\lambda}/n'}\right)^{-\hat{\xi}} - 1\right) \tag{6}$$

$$\widehat{\mathrm{ES}_\alpha} = \begin{cases} \frac{\widehat{\mathrm{VaR}_\alpha}+\hat{\beta}+u\hat{\xi}}{1-\hat{\xi}}, & \hat{\xi} \in (0,1) \\ \infty, & \hat{\xi} \geq 1 \end{cases} \tag{7}$$

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

Deux jeux de données du paquetage `CASdatasets` sont utilisés. Soit `auscathist` et `nzcathist`.

Analyse de données

Deux jeux de données du paquetage `CASdatasets` sont utilisés. Soit `auscathist` et `nzcathist`.

Historique des catastrophes naturelles pour l'Australie ainsi que pour la Nouvelle-Zélande.

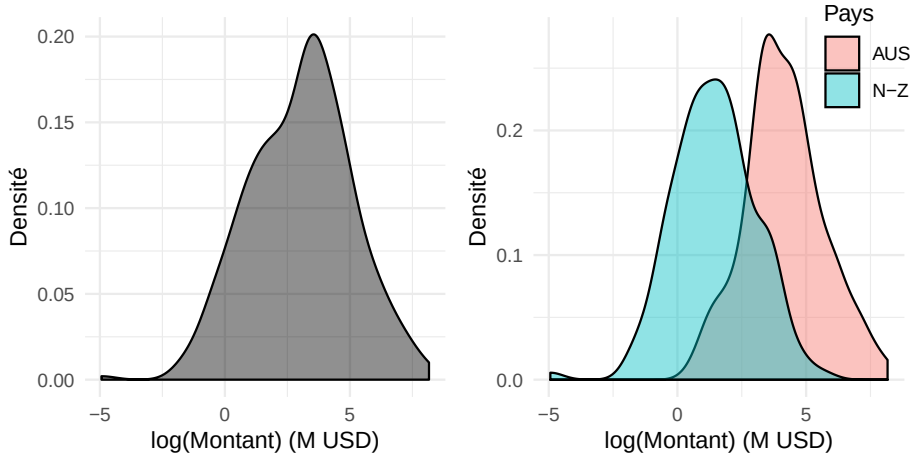

FIGURE – Densité du logarithme du montant des catastrophes

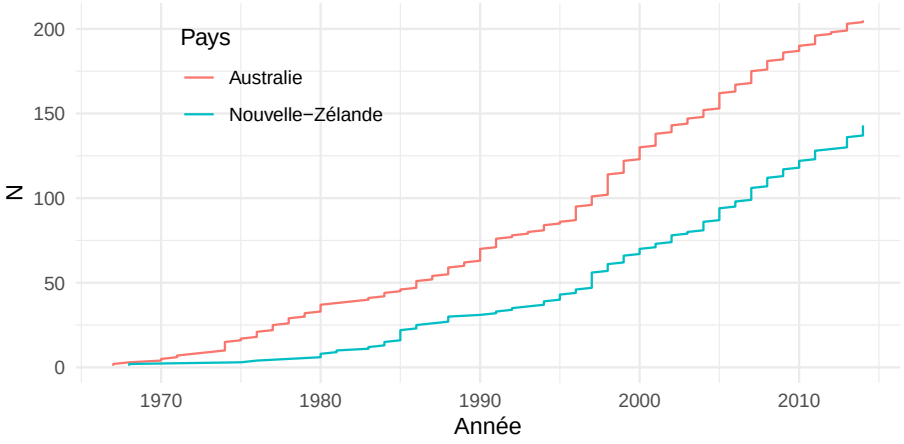

FIGURE – Nombre cumulatif de catastrophes par pays

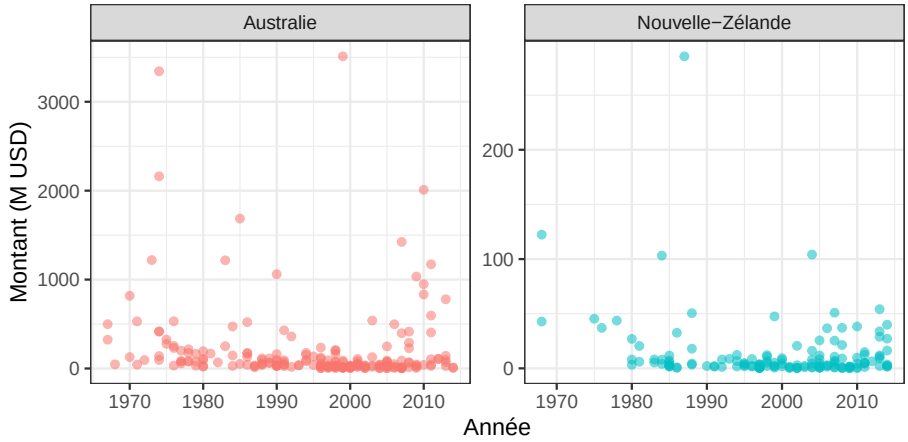

FIGURE – Évolution des catastrophes par pays

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

Tableau – Résumé statistique des montants (M USD) par type

Approche classique

- Résultats pas montrés de façon détaillée.

Approche classique

- Résultats pas montrés de façon détaillée.
- Approche pas tout à fait adéquate dans le cas présent.

Approche classique

- Résultats pas montrés de façon détaillée.
- Approche pas tout à fait adéquate dans le cas présent.
- Viable, rapide et peut être une bonne solution lorsque seulement les montants sont disponibles.

Approche classique

- Résultats pas montrés de façon détaillée.
- Approche pas tout à fait adéquate dans le cas présent.
- Viable, rapide et peut être une bonne solution lorsque seulement les montants sont disponibles.
- Un seuil de 10 M USD fut sélectionné.

Approche dynamique à deux variables

- Année et pays.

Approche dynamique à deux variables

- Année et pays.
- Variable numérique de temps et variable catégorique à deux niveaux.

Paramètre ρ

Le modèle sélectionné pour $\hat{\rho}$ dépend du pays et du temps :

$$\log\left(\frac{\hat{\rho}(x,t)}{1-\hat{\rho}(x,t)}\right) = \hat{f}_{\rho}(pays) + \hat{h}_{\rho}^{(2)}(annee),$$

où $h^{(\mathrm{Df})}$ représente une spline naturelle quadratique avec Df degrés de liberté.

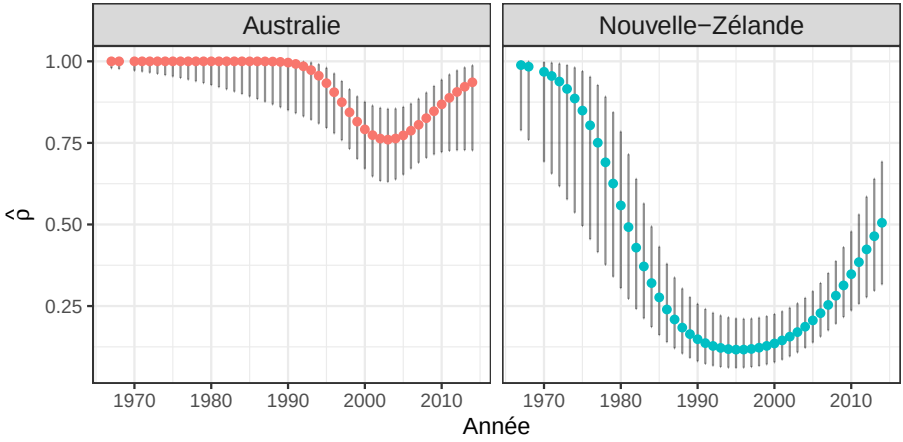

FIGURE – Prédictions du taux d'excès de seuil

Paramètres de la loi Pareto généralisée

Le modèle sélectionné pour $\hat{\xi}$ dépend du pays et non du temps :

$$\hat{\xi}(x,t) = \hat{f}_{\xi}(pays).$$

Paramètres de la loi Pareto généralisée

Le modèle sélectionné pour $\hat{\xi}$ dépend du pays et non du temps :

$$\hat{\xi}(x,t) = \hat{f}_{\xi}(pays).$$

Le modèle sélectionné pour $\hat{\beta}$ dépend du pays et du temps :

$$\hat{\beta}(x,t) = \hat{f}_{\beta}(pays) + \hat{h}_{\beta}^{(3)}(annee).$$

UNIVERSITÉ LAVAL

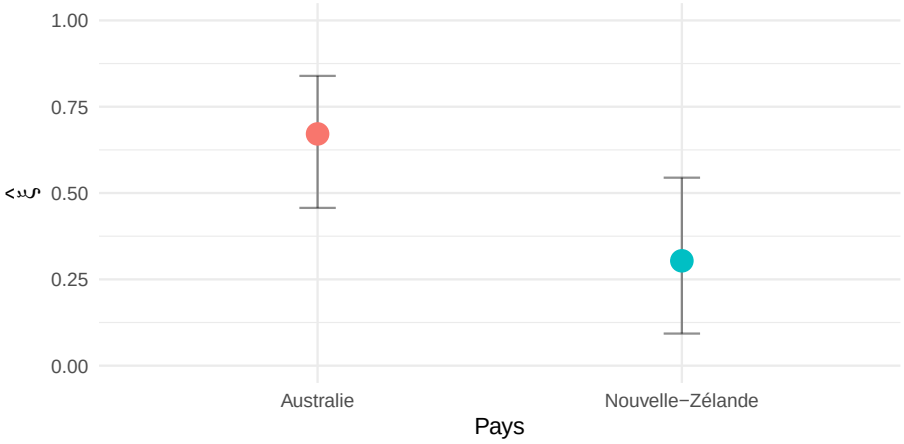

FIGURE – Prédictions du paramètre ξ

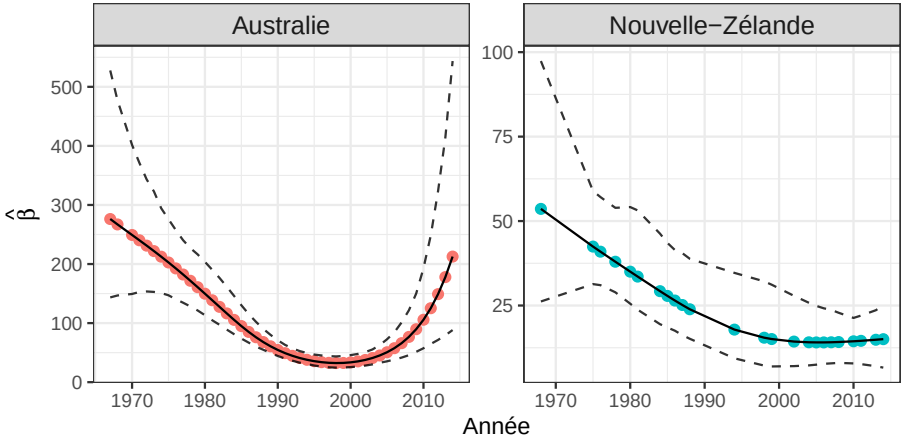

FIGURE – Prédictions du paramètre β

Validation du modèle

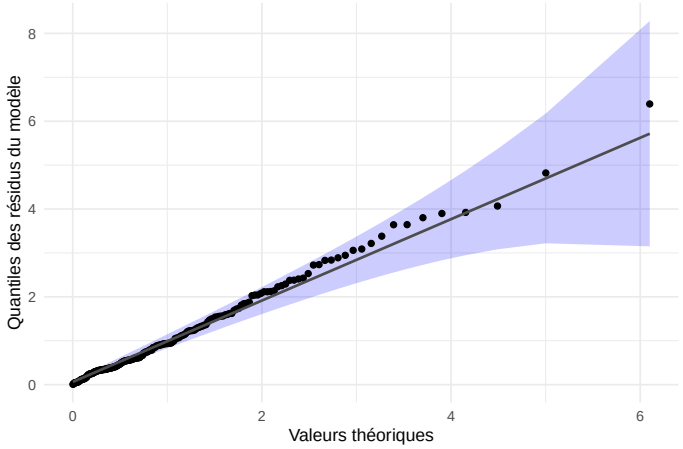

FIGURE – Graphique Q-Q de Exp(1)

Mesure VaR

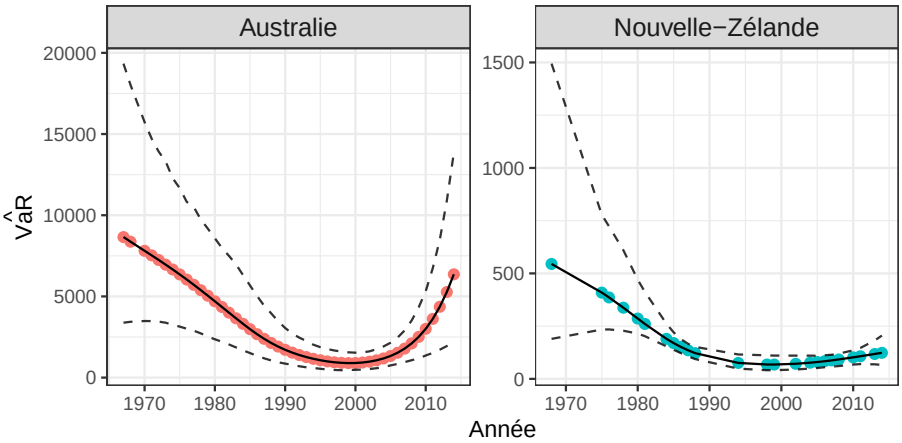

FIGURE – $\widehat{\text{VaR}}_{0.99}$

Résumé

- Méthodologie fonctionne.
- Donnent les résultats escomptés pour les variables.
- Amélioration du modèle classique.

Approche dynamique à trois variables

- Année, pays et type

Approche dynamique à trois variables

- Année, pays et type
- Regroupement des types.

Type	Type2
Bushfire	Bushfire
Hailstorm	Hailstorm
Cyclone	Cyclone
Flood	Flood, Storm
Flood, Storm	Flood, Storm
Storm	Flood, Storm
Other	Other
Tornado	Other
Earthquake	Other
Weather	Other
Power outage	Other

Tableau – Regroupement des types de catastrophes

Paramètre ρ

$$\log\left(\frac{\hat{\rho}(x,t)}{1-\hat{\rho}(x,t)}\right) = \hat{f}_\rho(pays) + \hat{f}_\rho(type) + \hat{h}_\rho^{(2)}(annee)$$

UNIVERSITÉ LAVAL

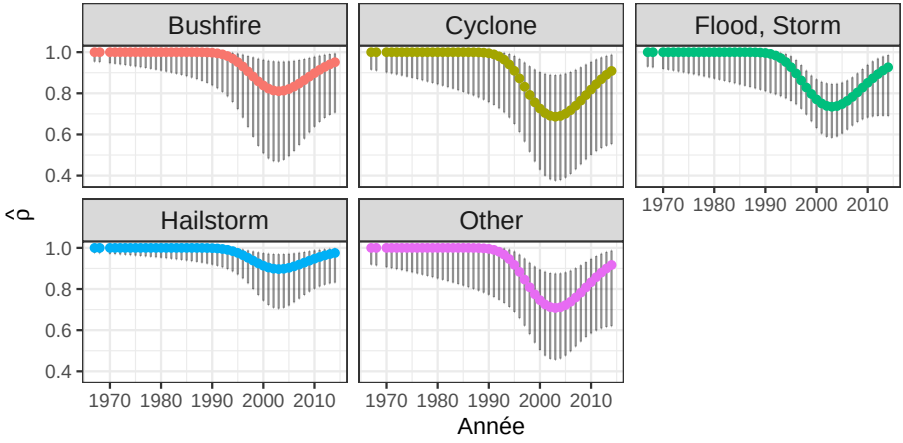

FIGURE – Prédictions du paramètre ρ pour l'Australie

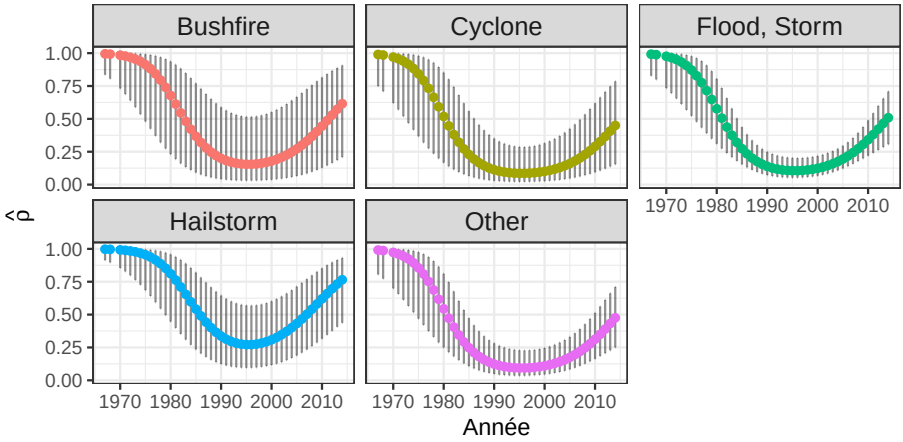

FIGURE – Prédictions du paramètre ρ pour la Nouvelle-Zélande

Paramètres de la loi Pareto généralisée

$$\hat{\xi}(x,t) = \hat{f}_{\xi}(pays) + \hat{f}_{\xi}(type)$$

$$\hat{\beta}(x,t) = \hat{f}_{\beta}(pays) + \hat{f}_{\xi}(type) + \hat{h}_{\beta}^{(4)}(annee)$$

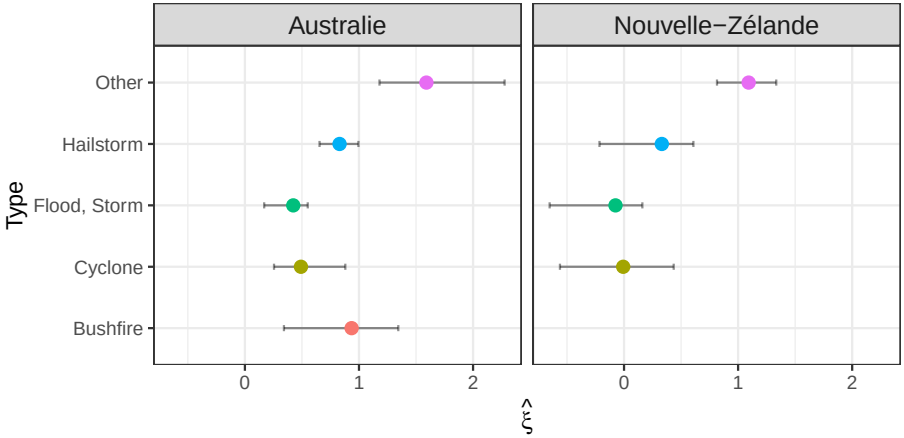

FIGURE – Prédictions du paramètre ξ

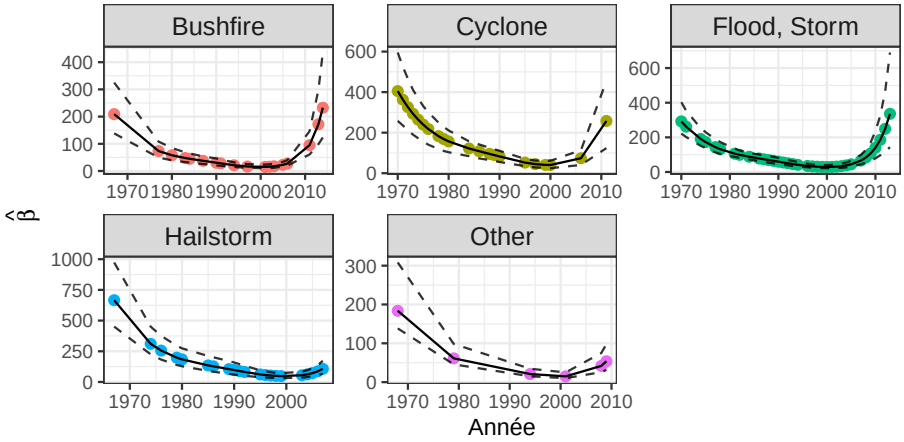

FIGURE – Prédictions du paramètre β pour l'Australie

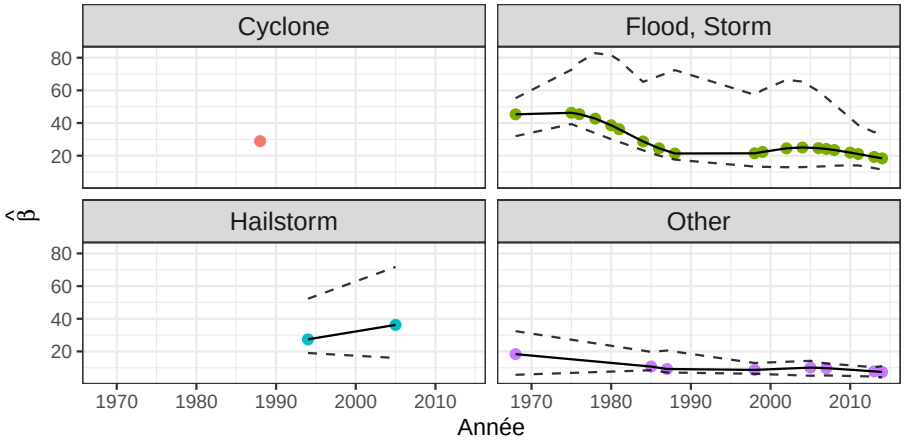

FIGURE – Prédictions du paramètre β pour la Nouvelle-Zélande

Validation du modèle

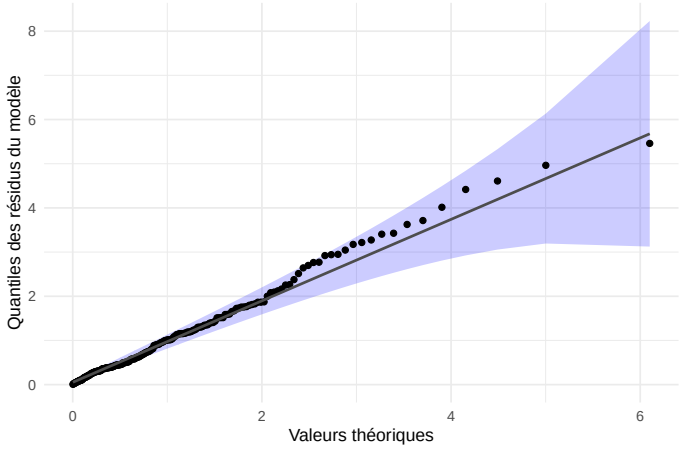

FIGURE – Graphique Q-Q de Exp(1)

Mesure VaR

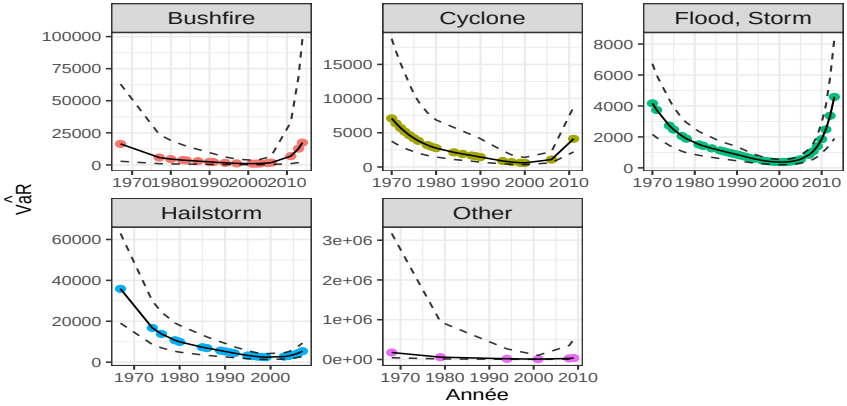

FIGURE – $\widehat{\text{VaR}}_{0.99}$ pour l'Australie

Mesure VaR

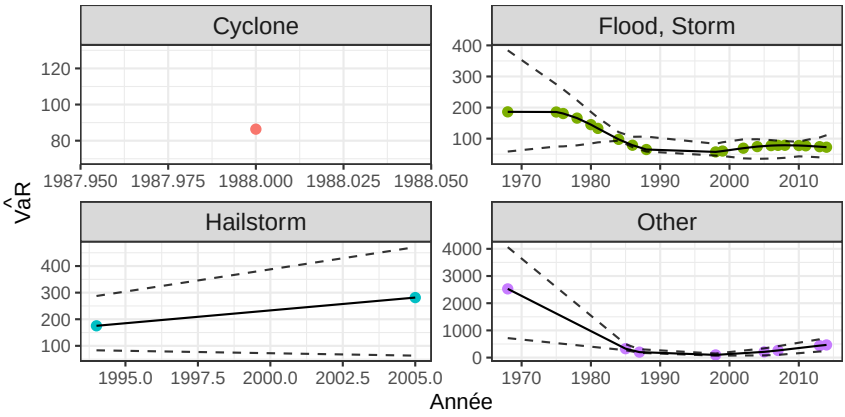

FIGURE – $\widehat{\text{VaR}}_{0.99}$ pour la Nouvelle-Zélande

Résumé

- Modèle adéquat.
- Aucun gain significatif.
- Paramètres et mesures de risques plus volatiles et moins crédibles.

Conclusion

- Méthode
- Appréciation

Questions ?

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.

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