
RAPPORT DE STAGE

STAGE À TITRE DE CHERCHEUR AU SEIN DE
L'INSTITUT NATIONALE DE RECHERCHE SCIENTIFIQUE

TRAVAIL PRÉSENTÉ À
M. ILIE RADU MITRIC

DANS LE CADRE DU COURS
TRAVAIL ACTUARIEL PRATIQUE EN ENTREPRISE
ACT-7005

PRÉPARÉ PAR
ALEXANDRE LEPAGE,
111 144 776

SUPERVISÉ PAR LE PROFESSEUR
FATEH CHEBANA, PHD

LE 14 AVRIL 2021



FACULTÉ DES SCIENCES ET DE GÉNIE
ÉCOLE D'ACTUARIAT
UNIVERSITÉ Laval

Table des matières

1 Sommaire	1
2 Description de l'entreprise	1
3 Attentes de l'étudiant	2
4 Attentes du superviseur	2
5 Mandat	2
5.1 Introduction au modèle étudié	3
5.2 Le modèle proposé	4
5.2.1 Description du modèle	4
5.2.2 Hypothèses du modèle	6
5.3 Méthodologie	6
5.3.1 Distributions marginales	7
5.3.2 Modélisation de la dépendance	8
5.3.3 Modélisation de la non-stationnarité	9
5.4 Études de cas	10
5.5 Résultats	11
5.6 Apprentissages	15
6 Évaluation du stage	15
6.1 Nature du travail	15
6.2 Environnement de travail	15
6.3 Préparation théorique à l'université	16
7 Conclusion	16
8 Exemple illustrant le modèle	19
A Illustrations	21
B Algorithme de simulation	23

1 Sommaire

2 Description de l'entreprise

L'Institut national de la recherche scientifique (INRS) est un centre de recherche universitaire composé de quatre établissements répartis dans les grands centres du Québec. Il est composé des établissements de recherche et de formation thématiques Armand-Frappier Santé Biotechnologie (Laval), Eau Terre Environnement (Québec), Énergie Matériaux Télécommunications (Montréal et Varennes) et Urbanisation Culture Société (Montréal et Québec).

Comme le superviseur de ce stage, le professeur Fateh Chebana¹, appartient au centre Eau Terre Environnement², c'est sur ce dernier que l'accent est porté. Ce centre de recherche est situé à Québec, au 490, rue de la Couronne. Il compte près de 36 professeurs répartis dans 6 programmes d'études dont les seuls en science de l'eau au Québec.

La mission de l'INRS est de promouvoir la recherche fondamentale et appliquée. Il valorise les études supérieures et la formation des chercheurs de demain. Dans l'atteinte de son objectif, l'institut doit orienter ses activités vers le développement économique, social et culturel du Québec, tout en assurant le transfert des connaissances et des technologies dans l'ensemble des secteurs où il œuvre.

Plus spécifiquement, le Centre Eau Terre Environnement a pour objectif d'**améliorer la protection, la conservation et la mise en valeur des ressources naturelles**. Les activités de recherche, quant à elles, sont concentrées dans quatre thématiques de recherche prioritaires : l'assainissement et la valorisation des résidus, la biogéochimie aquatique, l'hydrologie et les sciences de la Terre. Pour des exemples récents d'études réalisés par ce centre, on peut consulter le lien suivant : https://inrs.ca/wp-content/uploads/2021/01/4-Fiche-Hydrologie_ETE-2019-20.pdf.

Les valeurs véhiculées par cette institution sont l'excellence, l'interdisciplinarité, l'engagement, l'équité et l'intégrité. Autant en recherche que lors de la formation de nouveaux chercheurs, l'Institut cherche à tendre vers l'**excellence**. Pour ce faire, elle attend de ses membres qu'ils soient **intègres** dans leur travaille puisque cette valeur est un pilier de la recherche scientifique. Elle attend également de leur part qu'ils démontrent de l'**engagement** envers la mission de l'INRS. Avec ses quatre centres et ses nombreuses chaires de recherche³ dans des domaines tous plus variés les uns que les autres, l'INRS se définit à travers l'**interdisciplinarité**. C'est en unissant les forces de tous et chacun que l'on trouve des solutions durables et innovantes. Finalement, l'INRS propose un plan d'action⁴ en matière d'**équité**, de diversité et d'inclusion afin de valoriser la représentation des personnes des quatre groupes désignés (femmes, personnes en situation de handicap, autochtones, membres de minorités visibles) dans l'attribution de ces chaires.

1. <https://inrs.ca/la-recherche/professeurs/fateh-chebana/>

2. <https://inrs.ca/linrs/centres-de-recherche/centre-eau-terre-environnement/>

3. <https://inrs.ca/la-recherche/chaires-groupes-et-reseaux-de-recherche/>

4. https://inrs.ca/wp-content/uploads/2020/12/Plan_action_EDI_CRC_INRS_2020_12_07.pdf

3 Attentes de l'étudiant

Ce stage fait suite aux travaux réalisés lors de l'été 2020 sous la supervision des professeurs Étienne Marceau et Hélène Cossette. L'objet de ces travaux était de faire avancer la recherche actuarielle sur l'introduction d'une structure de dépendance dans les modèles de processus de renouvellement avec récompenses tels que définis par [Andersen, 1957] et décrit dans [Grimmett and Stirzaker, 2001] et [Gallager, 2013]. À la conclusion du mandat de l'été 2020, l'étape suivante était de trouver une application au modèle proposé. Comme plusieurs articles sur les copules et la modélisation des risques extrêmes sont en lien avec des contextes de climatologie ou, plus spécifiquement, en hydrologie (Voir p. ex. [Salvadori and De Michele, 2006], [Salvadori and De Michele, 2007] et [Vandenbergh et al., 2010]), ce domaine a suscité mon intérêt. Comme Mme Cossette et M. Marceau ont eu des étudiants qui ont fait des stages avec M. Chebana dans le passé et que ceux-ci se sont bien conclu, ce dernier m'a été référencé pour travailler sur un article collaboratif. L'idée étant que nous apportions le sujet et que celui-ci nous aidait à travers son expertise à trouver une application adéquate et à trouver des références intéressantes pour guider les recherches.

Pour aller plus en profondeur dans mes attentes, je souhaitais avoir beaucoup d'autonomie dans mes démarches. Malgré tout, une rencontre par semaine permet de tenir les collaborateurs au courant des avancements, de suggérer des lectures pertinentes, de partager des idées et de s'assurer que le projet va dans la direction que tout le monde s'était fixé. Pour cette raison, cette rencontre doit durer un temps suffisant et les différents collaborateurs doivent démontrer un intérêt commun pour le projet et s'impliquer.

4 Attentes du superviseur

5 Mandat

Comme il est mentionné dans la section 3, ce stage fait suite aux travaux réalisés lors de l'été 2020. L'idée était de trouver une application des processus de renouvellements avec récompenses lorsqu'il existe un lien de dépendance entre les temps inter-occurrences et la récompense (bonus ou malus, selon le contexte). Les premières semaines du stage ont donc été consacrées à trouver cette application. Celle-ci c'est présentée à la suite d'une présentation par M. Christian Genest⁵, professeur à l'université McGill, le 28 janvier 2021. Lors de cette présentation, M. Genest a présenté les résultats d'un article qu'il a publié en 2019, soit [Jalbert et al., 2019]. Le travail réalisé lors de ce stage reprend donc le contexte et s'inspire du modèle de cet article pour introduire le modèle développé lors de l'été 2020 dans la littérature en hydrologie.

La description du mandat est découpé comme suit : D'abord, le sujet de l'étude est introduit dans la sous-section 5.1. Puis, le modèle étudié est décrit de façon détaillé dans la section 5.2. Les résultats obtenus pour les deux études de cas réalisés sont présentés dans la sous-section 5.5. Finalement, la description du mandat est conclue avec la sous-section 5.6 les apprentissages réalisés dans le cadre de ce travail sont décrits.

5. <https://www.math.mcgill.ca/cgenest/>

5.1 Introduction au modèle étudié

Autant en assurance qu'en hydrologie, le sujet des inondations suscite l'intérêt des chercheurs qui tentent de modéliser ces événements afin de mieux se préparer à d'éventuelles catastrophes. Entre autres, le débordement du lac Champlain en 2011 a suscité l'intérêt de [Riboult and Brissette, 2016] qui a cherché à connaître les éléments déclencheurs d'une telle catastrophe. Lors de son étude, il arriva à la conclusion que, bien que la fonte des neiges soit une variable explicative importante, c'est l'accumulation de précipitations extrêmes dans la période de mars à juin qui a été la cause principale du désastre.

Suite à la publication de [Riboult and Brissette, 2016], [Jalbert et al., 2019] a cherché à prédire l'accumulation des pluies printanières du lac Champlain afin d'estimer l'amplitude maximale qu'un tel événement aurait pu avoir et de calculer l'espérance du temps qui s'écoulera avant qu'un incident d'une telle ampleur survienne à nouveau. Le modèle ainsi conçu sépare la modélisation des pluies en deux composantes. Une portion régulière qui représente la quantité de pluie totale tombée selon les normes saisonnières et une portion extrême où on considère les jours où la quantité de pluie tombée dépasse un certain seuil. Tandis que la première est bien modélisée avec une loi normale, la seconde nécessite plus de travail. En effet, afin de modéliser la quantité quotidienne de pluie tombée dans les cas extrêmes, [Jalbert et al., 2019] utilise la loi Pareto généralisée, aussi connue sous l'appellation *Peaks-Over-Threshold*(POT). Cependant, l'estimateur du paramètre de forme de la loi POT calculé sur les valeurs quotidiennes faisaient en sorte que la distribution avait un support fini. Se faisant, la probabilité qu'un événement de l'envergure de 2011 se produise était quasiment nulle. Conséquemment, [Jalbert et al., 2019] proposa une extension du modèle POT où il divise la quantité de pluie tombée lors d'une journée de pluie extrême par la proportion que cette quantité représente sur l'ensemble des précipitations tombées au cours d'une période de pluie continue. Grâce à ce stratagème, le modèle ainsi développé n'a plus de problème de distribution borné pour modéliser la sévérité des précipitations extrêmes.

Le modèle que nous proposons est une alternative à celui de [Jalbert et al., 2019]. Il reprend le concept de regroupement des jours en périodes de pluie continue, mais il adopte une approche légèrement plus intuitive en s'inspirant grandement de la méthodologie proposée dans [Zhang and Singh, 2019], [Salvadori and De Michele, 2006] et [Shaw et al., 2010]. Comme il est expliqué, entre autre, au chapitre 9.6 de [Shaw et al., 2010], les modèles de précipitations utilisés dans la pratique, communément appelés modèles DDF (Dept , Duration, Frequency), sous-tendent les variables de sévérité de durée et de fréquence, lesquelles ont toutes leur importance. Or, [Jalbert et al., 2019] se concentre principalement sur la sévérité. La fréquence est modélisée avec une loi de Poisson et l'aspect de la durée des événements n'est pas pris en compte. Notre approche, quant à elle, s'intéresse à toutes ces variables aléatoires et aux relations qui les unissent. Nous incorporons des notions des processus de renouvellement avec récompenses aux concepts communément utilisés en hydrologie afin d'améliorer la précision du modèle de fréquence et les liens de dépendance sont modélisés à l'aide de la théorie des copules introduite par [Joe, 1997]. Également, une attention particulière est apportée à la non-stationnarité des distributions de probabilité afin de prendre en compte le contexte des changements climatiques. Un modèle similaire est utilisé dans la modélisation des tempêtes par [Salvadori and De Michele, 2006] et [Salvadori and De Michele, 2007].

5.2 Le modèle proposé

L'intuition derrière le modèle proposé est de considérer les périodes de précipitation continue comme des événements uniques en agrégeant la quantité de pluie tombée lors de celles-ci. Se faisant, non seulement on considère les jours où le volume d'eau tombé sort de l'ordinaire, mais on considère également les séquences où le nombre de jours de pluie continue est anormalement élevé. Afin de modéliser la fréquence des événements extrêmes, les processus de renouvellement alternés tels que présentés dans [Small and Morgan, 1986] et [Salvadori and De Michele, 2006]) sont utilisés. Cette approche permet d'observer une structure de dépendance non négligeable entre la sévérité, la durée et le temps écoulé depuis le dernier événement extrême. Cette dépendance est modélisée à l'aide d'une copule en vigne (voir [Joe et al., 2010]). Puis, pour faire suite à [Jalbert et al., 2019] la théorie des valeurs extrêmes (voir [Hosking and Wallis, 1987]) est utilisée pour modéliser la sévérité des précipitations extrêmes avec la loi de Pareto généralisée (GP).

5.2.1 Description du modèle

Pour une année m , $m \geq 0$, on définit la suite de v.a. $\underline{Y}^{(m)} = \{Y_l^{(m)}, l \in \{1, \dots, 91\}\}$, où $Y_l^{(m)}$ représente la quantité, en mm de pluie, tombée lors des 91 jours du printemps, c.-à-d. du 1^{er} avril au 30 juin. Selon [Riboust and Brissette, 2016] et [Jalbert et al., 2019], cette période de temps est la plus critique lors des crues printanières.

Soit $\underline{\mathcal{C}}^{(m)} = \{\mathcal{C}_j^{(m)}, j \in \{1, \dots, n_{\mathcal{C}}^{(m)}\}\}$, la suite des $n_{\mathcal{C}}^{(m)}$ clusters regroupant les jours de pluie continue pour l'année m . En concordance avec [Jalbert et al., 2019], on définit un cluster comme une séquence de jours de pluie suivie par au moins une journée d'ensoleillement (0 mm de pluie). On définit la suite de v.a. $\underline{K}^{(m)} = \{K_j^{(m)}, j \in \{1, \dots, n_{\mathcal{C}}^{(m)}\}\}$, où $K_j^{(m)}$ représente la quantité de pluie totale tombée lors de la période $\mathcal{C}_j^{(m)}$ telle que $K_j^{(m)} = \sum_{l \in \mathcal{C}_j^{(m)}} Y_l^{(m)}$, $j = 1, \dots, n_{\mathcal{C}}^{(m)}$, $\forall m$.

Exemple 1. Supposons que, pour une année m , on ait $Y_1^{(m)} = 30$, $Y_2^{(m)} = 0$, $Y_3^{(m)} = 2$, $Y_4^{(m)} = 1$, $Y_5^{(m)} = 3$, $Y_6^{(m)} = 0$, $Y_7^{(m)} = 0$, $Y_8^{(m)} = 18$, $Y_9^{(m)} = 3$, $Y_{10}^{(m)} = 0$. Alors $\mathcal{C}_1 = \{1\}$, $\mathcal{C}_2 = \{3, 4, 5\}$, $\mathcal{C}_3 = \{8, 9\}$. De plus, $K_1^{(m)} = Y_1^{(m)} = 30$, $K_2^{(m)} = Y_3^{(m)} + Y_4^{(m)} + Y_5^{(m)} = 6$, $K_3^{(m)} = Y_8^{(m)} + Y_9^{(m)} = 21$.

Pour une année m et pour un seuil critique u , on sépare les événements en deux sous-ensembles $\mathcal{Z}^{(m)} = \{j : K_j^{(m)} < u\}$ et $\mathcal{X}^{(m)} = \{j : K_j^{(m)} \geq u\}$. On définira alors $\underline{X}^{(m)} = \{X_i^{(m)}, i \in \{1, \dots, \text{card}(\mathcal{X}^{(m)})\}\} = \{K_j^{(m)} : j \in \mathcal{X}^{(m)}\}$ comme étant une suite de v.a. où $X_i^{(m)}$ représente la quantité de pluie tombée lors du i -ème cluster extrême de l'année m .

Exemple 1, suite. On fixe un seuil $u = 18$. L'ensemble $\mathcal{X}^{(m)}$ correspond à $j \in \{1, 3\}$. On a donc $X_1^{(m)} = K_1^{(m)} = 30$, et $X_2^{(m)} = K_3^{(m)} = 21$. Puis, pour ce qui est des événements non-extrêmes, l'ensemble $\mathcal{Z}^{(m)}$ correspond au complément de l'ensemble $\mathcal{X}^{(m)}$. On a alors $\mathcal{Z}^{(m)} = \{2\}$.

Soit $\underline{T}^{(m)} = \left\{ T_i^{(m)}, i \in \{0, 1, \dots, \text{card}(\mathcal{X}^{(m)})\} \right\}$ une suite de v.a. où $T_i^{(m)}$ représente la première journée du cluster $\{C_j^{(m)}, j \in \mathcal{X}^{(m)}\}$ associé à $X_i^{(m)}$. On définit que $T_0^{(m)}$ correspond au 1^{er} avril de l'année m . Afin de calculer $T_i^{(m)} \forall i, m$, on définit une date d'origine qui correspond au jour zéro (p.ex. $T_0 = 1^{\text{er}} \text{ avril } 1900$), puis, on converti cette date en format numérique selon le nombre de jours écoulé depuis cette date de référence. Cette opération se réalise automatiquement en enchaînant les fonctions `as.Date` avec `as.numeric` dans le langage de programmation R.

Soit $\underline{D}^{(m)} = \left\{ D_i^{(m)}, i \in \{1, \dots, \text{card}(\mathcal{X}^{(m)})\} \right\}$, une suite de v.a. où $D_i^{(m)}$ représente la durée, en nombre de jours, du i -ème événement extrême de l'année m . Autrement dit, cela signifie que $\{D_i^{(m)}, i \in \{1, \dots, \text{card}(\mathcal{X}^{(m)})\}\} = \{\text{card}(C_j^{(m)}) : j \in \mathcal{X}^{(m)}\}$.

Soit $\underline{W}^{(m)} = \left\{ W_i^{(m)}, i \in \{1, \dots, \text{card}(\mathcal{X}^{(m)})\} \right\}$, une suite de v.a. où $W_i^{(m)}$ représente le temps écoulé depuis la fin du dernier événement extrême telle que

$$W_i^{(m)} := T_i^{(m)} - T_{i-1}^{(m)} - D_{i-1}^{(m)} = T_i^{(m)} - T_{i-1}^{\star(m)},$$

où $T_i^{\star(m)} := T_0^{(m)} + \sum_{k=0}^i W_k^{(m)} + D_k^{(m)}$. Par convention, on a $W_0^{(m)} = D_0^{(m)} = 0, \forall m$.

Exemple 1, suite. Si on attribut la valeur numérique 0 au 1^{er} avril de l'année m , alors on a $T_0^{(m)} = 0$, $T_1^{(m)} = T_0^{(m)} = 0$, $T_2^{(m)} = 7$. On a également $D_1^{(m)} = \text{card}(C_1^{(m)}) = 1$, $D_2^{(m)} = \text{card}(C_3^{(m)}) = 2$ et $W_1^{(m)} = T_1^{(0)} = 0$, $W_2^{(m)} = T_2^{(m)} - T_1^{(m)} - D_1^{(m)} = 7 - 0 - 1 = 6$.

Soit $\mathbf{N}^{(m)} = \{N_s^{(m)}(t), s, t \geq 0\} = \{N^{(m)}(s, s+t), s, t \geq 0\}$, un processus de renouvellement non-stationnaire alterné où un accroissement $N_s^{(m)}(t)$ permet de modéliser le nombre d'événements extrêmes ($\text{card}(\mathcal{X})$) survenus lors de l'intervalle de temps $[s, s+t]$. Dans le contexte des changement climatiques, aucune hypothèse n'est faite sur la stationnarité du processus. Par définition, on a

$$N_{T_0}^{(m)}(t) := \sum_{i=1}^{\infty} \mathbb{1}\{T_i^{\star(m)} \leq T_0^{(m)} + t\} = \sum_{i=1}^{\infty} \mathbb{1}\{T_i^{(m)} + D_i^{(m)} \leq T_0^{(m)} + t\} = \inf\{i : T_i^{(m)} + D_i^{(m)} \leq T_0^{(m)} + t\}. \quad (1)$$

Soit $\mathbf{V}^{(m)} = \{V_s^{(m)}(t), s, t \geq 0\} = \{V^{(m)}(s, s+t), s, t \geq 0\}$, un processus de renouvellement non-stationnaire alterné avec récompenses où un accroissement $V_s^{(m)}(t)$ permet de modéliser la quantité totale d'eau accumulée lors des événements extrêmes sur un intervalle de temps $[s, s+t]$. Ainsi, on a

$$V_{T_0}^{(m)}(t) := \sum_{i=1}^{N_{T_0}^{(m)}(t)} X_i^{(m)} = \sum_{i=1}^{\infty} X_i^{(m)} \mathbb{1}\{T_i^{(m)} + D_i^{(m)} \leq T_0^{(m)} + t\}. \quad (2)$$

Les processus de renouvellement alternés de même que leur homologue avec récompenses sont utilisés dans la littérature en hydrologie, notamment dans [Salvadori and De Michele, 2006], [Salvadori and

De Michele, 2007] et dans [Small and Morgan, 1986].

Soit $Z^{(m)}$, la v.a. de la quantité de pluie non-extrême totale tombée lors de la saison printanière de l'année m . On a $Z^{(m)} := \sum_{j \in \mathcal{Z}^{(m)}} K_j^{(m)}$. La suite de v.a. $\underline{Z} = \{Z^m, m \in \mathbb{N}\}$ n'est pas présumée stationnaire étant donné le contexte des changements climatiques. La quantité de pluie totale tombée au cours du printemps d'une année m est donc modélisée avec.

$$S_{T_0}^{(m)}(91) = Z^{(m)} + V_{T_0}^{(m)}(91). \quad (3)$$

Exemple 1, suite. Pour revenir à l'exemple 1, on a $N_0^{(m)}(10) = \text{card}(\mathcal{X}) = 2$ et $V_0^{(m)}(10) = X_1^{(m)} + X_2^{(m)} = 30 + 21 = 51$. Puis, $Z^m = K_2^{(m)} = 6$. Finalement $S_0^{(m)}(10) = 51 + 6 = 57$.

Pour un exemple plus complet utilisant des données réelles provenant de la rivière Clearwater en Alberta et contenant plus d'une année, voir l'annexe 8.

5.2.2 Hypothèses du modèle

En ce qui concerne l'hypothèse d'indépendance séquentielle, nous considérons qu'il est raisonnable de présumer que $Y_1^{(m)}$ est indépendante de $Y_{91}^{(m-1)}$. De ce fait, il faudrait réaliser le test d'indépendance proposé par [Genest and Rémillard, 2004] sur chacune des années de façon indépendante. Cependant, comme la méthode proposée réduit grandement le nombre d'observations disponibles pour chacune des années, on se retrouverait avec au plus une vingtaine d'observations par année pour les variables résultantes de l'agrégation. Conséquemment, les résultats du test sont instables et peu concluants. Pour cette raison et par souci de simplicité, l'hypothèse d'indépendance séquentielle est supposée pour chacune des v.a. du modèle.

Pour ce qui est de l'hypothèse de stationnarité, considérant le contexte des changements climatiques, aucune présomption n'est faite à ce sujet. Un test de Mann-Kendall est effectué sur chacune des v.a. afin de vérifier l'hypothèse. De plus, lors de la sélection des distributions marginales, l'AIC est calculée autant pour une version stationnaire de chaque loi, que pour une version non-stationnaire. Cette approche est suggérée dans [Chebana and Ouarda, 2021].

5.3 Méthodologie

Afin de pouvoir faire l'étude du processus (3), la méthodologie consiste à paramétrier les distributions marginales en premier. Puis, à l'aide des fonctions de répartition théoriques obtenus, d'identifier la structure de dépendance unissant ces v.a. pour choisir la(les) copule(s) appropriée(s). Une fois toutes ces composantes obtenues, il est possible d'étudier le comportement de (3) par simulation en faisant appel à l'algorithme 1 présenté en annexe.

5.3.1 Distributions marginales

Avec la théorie des valeurs extrêmes, comme on cherche à modéliser l'ensemble des précipitations qui sortent des normales saisonnières, la méthode POT est tout indiquée pour modéliser les excédents de seuil (voir p.ex. [Hosking and Wallis, 1987], [Klugman et al., 2013]). Alors on pose $(X_i - u | X_i \geq u) \sim \text{GPD}(\xi, \sigma)$ telle que la fonction de répartition s'exprime comme

$$F_u(x) := \begin{cases} 1 - (1 + \xi x / \sigma)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-x/\sigma), & \xi = 0, \end{cases} \quad (4)$$

où $x \geq 0$, $\sigma > 0$, $\xi \in \mathbb{R}$ et $1 + \xi x / \sigma > 0$, pour un seuil $u > 0$. À noter que la paramétrisation de la loi est grandement influencée par la valeur attribuée au paramètre u , lequel peut être obtenu par optimisation tel que présenté dans [Bader et al., 2016] et [Bader et al., 2018] ou encore avec [Northrop et al., 2015]. Pour faire suite à [Jalbert et al., 2019] et pour des fins de simplicité considérant le temps imparti pour ce stage, le seuil utilisé est fixe. Cependant, il est d'intérêt de mentionner que plusieurs auteurs tels que [Kysely et al., 2010], [Beguería et al., 2011] et [Cheng et al., 2014] préconisent pour un seuil non-stationnaire. Cette approche sera considérée dans une prochaine version de ce travail. Dans le cas stationnaire, la loi peut être paramétrée selon l'approche des moments ou celles des moments pondérés dépendamment de la forme que prend la loi (voir [Hosking and Wallis, 1987]). Cependant, dans le cas non-stationnaire, cela est impossible. la méthode utilisée devient alors celle du maximum de vraisemblance généralisée présentée dans [El Adlouni et al., 2007]. Comme cette approche est un dérivé de la paramétrisation bayésienne, l'algorithme MCMC (Simulation Monte-Carlo par chaînes de Markov) est utilisé et les paramètres de la loi doivent avoir des distributions *a priori*.

En ce qui concerne à la modélisation de la durée des périodes de pluie et les temps inter-occurrences, considérant que l'échelle de temps utilisée est en jours, les lois envisagées sont discrètes. De plus, considérant que le domaine des v.a. en jeux n'est pas fermé, la loi binomiale est écartée. Conséquemment, les lois retenues sont la loi binomiale négative, son homologue à un paramètre, la loi géométrique et la loi de Poisson. Également, étant donnée sa grande flexibilité, la loi Weibull discrétisée est également retenue. Finalement, dans le cas où la queue de la distribution serait très lourde, alors une version discrétisée de la loi GP est envisagée.

Afin de paramétriser la distribution de W , on pose la fonction de vraisemblance (5).

$$L(\boldsymbol{\theta} | \mathbf{w}, \mathbf{t}^*) = \prod_k f_W(w_k | \boldsymbol{\theta}, t_k^*), \quad (5)$$

où $\boldsymbol{\theta}$ correspond au vecteur des paramètres de la loi de W , $\mathbf{t}^* = \{t_k^*, k \in \{1, \dots, \text{card}(\mathbf{t})\}\}$ correspond à un vecteur d'observations où t_k^* est la k -ème observation de T^* dans les données disponibles. Puisque l'accroissement du processus de renouvellement est défini sur un intervalle de temps fini, il faut ajouter un terme à (5) pour tenir compte du temps qu'il reste à la fin du processus, pour chacune des années. Soit $w_{\text{res}}^{(m)} = 91 - t_N^{*(m)}$ où $t_N^{*(m)}$ correspond au temps de fin du dernier événement de l'année m et $w_{\text{res}}^{(m)}$ est le

temps résiduel du processus. Alors (5) devient

$$L'(\boldsymbol{\theta}|\mathbf{w}, \mathbf{t}^*) = L(\boldsymbol{\theta}|\mathbf{w}, \mathbf{t}^*) \times \prod_m (1 - F_W(w_{\text{res}}^{(m)}|\boldsymbol{\theta}, t_N^{*(m)})). \quad (6)$$

Dans le cas où f_W est la fonction de densité d'une loi GP, la méthode du maximum de vraisemblance telle que présentée dans [Hogg et al., 2005] ne peut être utilisée. Il faut plutôt utiliser la méthode du maximum de vraisemblance généralisé présenté dans [El Adlouni et al., 2007], laquelle nécessite l'algorithme MCMC pour réaliser l'optimisation.

Du point de vue des précipitations non-extrêmes, [Jalbert et al., 2019] recommande l'utilisation de la loi normale. Cependant, comme cette loi admet des valeurs négatives, on considérera aussi la loi gamma comme alternative.

5.3.2 Modélisation de la dépendance

La dépendance unissant les v.a. en jeu est modélisée avec la théorie des copules décrite dans [Joe, 1997] et [Nelsen, 2006]. Dans le contexte particulier des précipitations, on s'intéressera particulièrement à [Zhang and Singh, 2019] qui utilise les copules archimédiennes et elliptiques pour modéliser la dépendance entre la durée et l'intensité des précipitations. On s'intéresse également à [Salvadori and De Michele, 2006] qui définit un processus de renouvellement alterné avec structure de dépendance unissant les variables de durée des épisodes secs, de durée des périodes humides et d'une variable d'intensité pour modéliser les orages. Cette dépendance est modélisée à l'aide d'une copule en vigne tri-variée.

Dans notre cas, nous reprenons l'idée d'utiliser une copule en vigne pour modéliser la dépendance entre les v.a. de la sévérité X , de la durée D et des temps inter-occurrences W . La structure de vigne ainsi utilisée est présentée dans l'illustration 1.

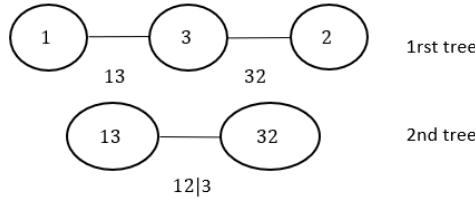


ILLUSTRATION 1 – Structure C-Vine de la copule tri-variée modélisant la dépendance entre (1) $u^{(X)}$, (2) $u^{(W)}$ et (3) $u^{(D)}$.

Dans le contexte des valeurs extrêmes, on s'intéresse à [Gudendorf and Segers, 2010] qui recommande, entre-autre, les copules de Gumbel, de Tawn et de Galambos lorsque les marginales impliquées sont de lois extrêmes. De plus, comme le suggère [Zhang and Singh, 2019], les copules archimédiennes sont aussi prises en compte. Parmi celles-ci, on retrouve les copules de Frank, de Clayton ainsi que les copules BB1-BB3 et BB6-BB7 décrites dans [Joe, 1997]. Finalement, on considère également des rotations à 180° de ces

copules ; dans ce cas, on parle alors de copules de survie. Les copules elliptiques sont appréciées dans la littérature en hydrologie étant donnée leur flexibilité et la facilité avec laquelle on peut les paramétriser. Cependant, dans le contexte de la modélisation de valeurs extrêmes, nous les écartons, conformément aux conclusions de [Renard and Lang, 2007], puisque cette famille de copule tend à sous-estimer la dépendance lorsque les marginales sont constituées à la fois de v.a. extrêmes et de v.a. non-extrêmes.

Soit $\mathbf{x} = \{(X_i^{(m)} - u|X_i^{(m)} \geq u), \forall i, m\}$, la suite des observations où $(X_i^{(m)} - u|X_i^{(m)} \geq u)$ représente l'excédent de seuil observé pour le i -ème événement de l'année m . Soit $\mathbf{w} = \{W_i^{(m)}, \forall i, m\}$, la suite des observations où $W_i^{(m)}$ représente le temps écoulé entre les i -ème et $(i-1)$ -ème événements de l'année m . Soit $\mathbf{d} = \{D_i^{(m)}, \forall i, m\}$, la suite des observations où $D_i^{(m)}$ représente la durée du i -ème événement extrême lors de l'année m . Soit $\mathbf{t} = \{T_i^{(m)}, \forall i, m\}$, la suite des observations où $T_i^{(m)}$ représente le temps de survenance du i -ème événement extrême lors de l'année m . Soit $u^{(X)} = \mathbb{P}(X - u \leq \mathbf{x}|X > u)$, le vecteur des uniformes générés en évaluant la fonction de répartition marginale estimée pour les excédents de seuil. Soit $u^{(W)} = \mathbb{P}(W \leq \mathbf{w}|\mathbf{t})$, le vecteur des uniformes générées en évaluant la fonction de répartition marginale estimée pour les temps inter-occurrences. Soit $u^{(D)} = \mathbb{P}(D \leq \mathbf{d}|\mathbf{t})$, le vecteur des uniformes générés en évaluant la fonction de répartition marginale estimée pour la durée de chacun des événements observés. Afin de calculer les mesures de corrélation et de paramétriser les copules de façon adéquate, il est important de considérer la proposition 1 et la remarque 1.

Proposition 1. *L'utilisation des fonctions de répartitions empiriques $F_n(x) = \text{rank}(x)/(n+1)$ sous-tend l'hypothèse que les observations disponibles sont représentatives des minimums et maximums des distributions réelles des données. Hors, comme la v.a. des excédents de seuil fait partie de la famille des lois à valeurs extrêmes, cette hypothèse doit être rejetée. C'est pourquoi, il est mieux d'utiliser les fonctions de répartitions estimées pour produire les pseudo-observations plutôt que d'utiliser une méthode basée sur les rangs comme le suggère généralement la littérature sur la théorie des copules (voir p.ex. [Genest and Favre, 2007]).*

Remarque 1. *Afin de tenir compte du fait que deux des trois v.a. impliquées ont un support discret, [Genest and Nešlehová, 2007] suggère de ne pas utiliser la méthode des moments pour paramétriser les copules puisque cela insérerait un biais. La méthode de paramétrisation s'appuie donc sur la méthode de la pseudo-vraisemblance (voir [Kim et al., 2007]). Comme il a été mentionné dans la remarque 1, dans le contexte des variables extrêmes, les pseudo-observations doivent être calculées à partir des distributions marginales théoriques plutôt qu'à partir des rangs.*

5.3.3 Modélisation de la non-stationnarité

Afin de tenir compte d'une éventuelle tendance dans la distribution des différentes v.a. en jeux, on considère que seule la moyenne varie dans le temps et que la variance est stable dans le temps. Cette hypothèse est appuyée par une analyse graphique des séries temporelles associées à chacune des variables. Ces graphiques ne sont pas inclus dans ce rapport par souci de concision. Ainsi, on pose $\mu(t) = a+bt$, $t \geq 0$, comme étant une fonction linéaire servant à modéliser l'espérance des lois. Puis, en posant le paramètre

d'échelle de chacune des lois égale à une fonction de cette moyenne, on arrive à paramétriser les lois en représentant adéquatement la tendance. Si, pour une loi donnée, la fonction de moyenne admet des valeurs négatives, alors une alternative est de considérer plutôt $\ln \mu(t) = a + bt, t \geq 0$. Cette approche s'inspire de celle proposée par [Khaliq et al., 2006].

Exemple 2. Soit la v.a. W telle que $W \sim \text{Geo}(p)$, $p \in [0, 1]$, avec espérance $\mathbb{E}[W] < \infty$ définie comme $\mathbb{E}[W] = 1/p$. Pour paramétriser la loi de W en tenant compte de la tendance, on pose $p(t) = 1/\mu(t) = (a + bt)^{-1}$, $a, b, t \in \mathbb{R}$.

Exemple 3. Soit la v.a. W telle que $W \sim \text{Weibull}(\alpha, \beta)$, $\alpha, \beta > 0$, avec espérance $\mathbb{E}[W] < \infty$ définie comme $\mathbb{E}[W] = \frac{1}{\beta}\Gamma(1 + 1/\alpha)$. Pour paramétriser la loi de W en tenant compte de la tendance, on pose $\beta(t) = \mu(t)/\Gamma(1 + 1/\alpha) = \frac{a+bt}{\Gamma(1+1/\alpha)}$, $\alpha > 0$, $a, b, t \in \mathbb{R}$.

5.4 Études de cas

Afin d'appliquer le modèle proposé dans la section 5.2, deux bases de données sont étudiées. La première fait suite à l'étude réalisée par [Jalbert et al., 2019] et concerne la ville de Burlington dans le Vermont. La seconde touche la rivière Clearwater en Alberta. Dans les deux cas, l'hypothèse de non-stationnarité est vérifiée pour chacune des v.a. en jeux. Différentes distributions sont testées afin de les modéliser et le critère de l'AIC est utilisé pour sélectionner les loi marginales offrant la meilleure adéquation. Cette approche s'appuie sur [Chebana and Ouarda, 2021] et sur [Khaliq et al., 2006]. Par la suite, les copules en vignes sont utilisées afin de modéliser la structure de dépendance unissant les différentes composantes du modèle proposé. La sélection de la copule la plus adéquate s'appuie sur [Dissmann et al., 2013] et les fonction R de la librairie VineCopula.

Pour la première étude de cas, les données utilisées proviennent du site de la National Oceanic and Atmospheric Administration (NOAA)⁶. La station USCO0431072 a débuté ses opérations en 1884. Cependant, elle comporte quelques données manquantes. Afin d'interpoler sur celles-ci, l'approche recommandée par [Shaw et al., 2010] est de moyenner les observations obtenues sur les stations avoisinantes. De plus, comme cette station a fermé le 3 juin 1943, nous avons compensé, comme indiqué par [Jalbert et al., 2019], avec la station de l'aéroport de Burlington, soit la station USW00014742, laquelle a commencé ses opérations le 1^{er} décembre 1940. Les données obtenues comportent de l'information jusqu'en 2020, soit 137 ans (12 467 observations printanières).

La rivière Clearwater se trouve en Alberta, au Canada. Les données utilisées proviennent de la station 07CD001 qui est située près de Fort McMurray, aux coordonnées GPS suivantes : (56°41'06" N, 111°15'18" W). L'illustration 4 présente l'emplacement de la station. Les données utilisées couvrent la période de 1960 à 2013 (53 ans ; 4823 observations printanières). Celles-ci sont disponibles sur le site des Relevés hydrologiques du Canada⁷. Aucune donnée manquante n'est recensée dans cette base de données.

6. <https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND>

7. https://eau.ec.gc.ca/search/historical_f.html

5.5 Résultats

Avec la méthodologie décrite dans la section 5.3, on trouve que, dans les deux études de cas, la loi de sévérité $X - u|X > u$ peut être considérée comme stationnaire. Avec un seuil u de 26.67mm de pluie (86.39^e percentile) pour le lac Champlain et de 18.29mm (88.74^e percentile) pour la rivière Clearwater, les estimateurs de la loi GP obtenus sont $\hat{\xi} = 0.08056533$, $\hat{\sigma} = 15.26847280$ pour le premier et $\hat{\xi} = 0.1317715$, $\hat{\sigma} = 14.2908019$ pour le deuxième. Conséquemment, dans les deux cas, la loi possède un domaine non-fini, ce qui constitue une amélioration à la première tentative de paramétriser un modèle POT réalisé par [Jalbert et al., 2019]. Également, bien que la méthode de sélection de seuil soit automatisé grâce à l'approche de [Bader et al., 2018], les deux études de cas démontrent des similitudes dans le seuil optimal de la loi GP et dans le paramètre d'échelle. Du point de vue de l'adéquation, les tests d'Anderson-Darling et de Cramer-Von Mises suggérés par [Choulakian and Stephens, 2001] offrent des *p-values* de 0.132 et de 0.13 pour la première étude de cas, ainsi que de 0.452 et de 0.725 pour la seconde. La loi GP non stationnaire s'agence donc mieux aux données de la rivière Clearwater qu'à ceux du la Champlain.

En ce qui a trait à la modélisation des temps inter-occurrences W , en comparant l'AIC des lois testées avec et sans tendance, comme il est expliqué dans les sections 5.3.1 et 5.3.3, on trouve que la loi qui s'agence le mieux aux données pour la base de données du lac Champlain est la loi géométrique avec les paramètres $\hat{a}^{(W)} = 32.48488$, $\hat{b}^{(W)} = -0.000410909$.

Pour ce qui est des données de la rivière Clearwater, comme la distribution empirique de la v.a. W possède une queue très lourde, c'est plutôt la loi GP qui modélise le mieux les temps inter-occurrences. Afin de paramétriser cette loi, on pose $\xi_t^{(W)} = \xi^{(W)}$ ainsi que $\sigma_t^{(W)} = \sigma^{(W)}$, $t \geq 0$. La distribution *a priori* de ces estimateurs est alors

$$\xi^{(W)} \sim \mathcal{N}(0, 0.25), \quad \sigma^{(W)} \sim \Gamma(5, 1). \quad (7)$$

Pour ce qui est du cas non-stationnaire, on pose $\xi_t^{(W)} = \xi^{(W)}$ ainsi que $\ln(\sigma_t^{(W)}) = a^{(W)} + b^{(W)} \times t$, $t \geq 0$. Puis, on pose

$$\xi^{(W)} \sim \mathcal{N}(0, 0.25), \quad a^{(W)} \sim \mathcal{N}(\tilde{a}, 3), \quad b^{(W)} \sim \mathcal{N}(0, 1), \quad (8)$$

où \tilde{a} correspond au paramètre initial de l'optimisation. Dans ce cas, ce paramètre initial correspond à la moyenne des données observées, soit 34.76. Dans ce cas, on trouve que la distribution GP stationnaire est celle qui représente le mieux les données. Ses paramètres estimés sont alors $\hat{\xi}^{(W)} = -0.5056498$ et $\hat{\sigma}^{(W)} = 58.5460087$. On peut donc voir que, comme le paramètre $\hat{\xi}^{(W)}$ est négatif, la loi GP prend la forme d'une loi Beta inversée qui possède un domaine fini sur l'intervalle $[0, -\xi/\sigma]$ (voir [Klugman et al., 2013]). Cependant, comme les données sont censurées, cela ne pose pas de problème du point de vue conceptuel.

On remarque que, pour la première étude de cas, la loi qui s'agençait le mieux était la loi géométrique qui possède une queue de distribution très fine. Tandis que, pour ce qui est de la seconde, c'est la loi GP qui offre la meilleure adéquation, laquelle possède une queue de distribution très lourde. Ce que l'on peut déduire de cette observation, c'est que la localisation de la station utilisée en Alberta possède un climat plus sec que celui de la rivière Champlain, du fait de la chaîne de montagne séparant la Colombie-Britannique de l'Alberta. Les précipitations extrêmes y sont donc moins fréquentes et surviennent à des

intervalles plus distancés.

Pour ce qui est de la modélisation de la v.a. de la durée des périodes de précipitation extrême D , le critère de l'AIC suggère que la loi Weibull avec tendance est celle qui s'agence le mieux aux données dans les deux scénarios. La paramétrisation alors obtenue est $\hat{\alpha}^{(D)} = 2.104619$, $\hat{a}^{(D)} = 3.94734$, $\hat{b}^{(D)} = 1.746127e - 05$, pour la première étude de cas, et $\hat{\alpha}^{(D)} = 2.2442893$, $\hat{a}^{(D)} = 1.0591309$, $\hat{b}^{(D)} = 0.7522425$, pour la seconde. On voit alors que le paramètre de forme des deux distributions est quasiment le même, mais que le phénomène de tendance semble beaucoup plus fort pour la base de données de la rivière Clearwater. Cette observation est cohérente avec le test de Mann-Kendall qui offre une *p-value* de 0.2366 pour le lac Champlain et de 0.0002 pour la rivière Clearwater ; ce qui laisse présager que la durée des précipitations extrêmes est peu influencée par les changements climatiques dans la région de Burlington, au Vermont, contrairement à la région de Fort McMurray, en Alberta. Comme cette dernière est située plus au nord, près de l'Alaska, peut-être que la fonte des glaciers peut expliquer en partie ce phénomène.

Au niveau des précipitations non-extrêmes Z , on trouve qu'il n'y a pas d'évidence contre l'hypothèse de stationnarité selon le test de Mann-Kendall. En effet, les *p-value* obtenues pour les deux études de cas correspondent à à 0.5124 et à 0.5357.

Du point de vue de la distribution, pour la première étude de cas, les données observées échouent le test de normalité de Shapiro-Wilk. La loi Gamma, en revanche offre une belle adéquation selon les tests de Kolmogorov-Smirnov et d'Anderson-Darling avec des *p-values* de 0.708 et de 0.817. Les paramètres alors estimés sont $\hat{\alpha}^{(Z)} = 14.47781$ et $\hat{\beta}^{(Z)} = 0.10900$, selon la paramétrisation de la loi gamma où l'espérance est calculée avec $\mathbb{E}[Z] = \alpha^{(Z)}/\beta^{(Z)}$.

Du côté des données de la rivière Clearwater, le test de Shapiro-Wilk offre une *p-value* de 0.53576, laissant présager que la loi normale s'agence bien aux données. Les tests d'Anderson-Darling et de Kolmogorov-Smirnov appuient cette observation avec des *p-values* de 0.8768 et de 0.9708. Les paramètres de la loi estimés sont $\hat{\mu}^{(Z)} = 63.54724$ et $\hat{\sigma}^{(Z)} = 19.46997$.

Une fois toutes les marginales paramétrées, il reste à modéliser la structure de dépendance unissant les v.a. W , D et X . Dans un premier temps, l'analyse de la dépendance passe par le calcul des coefficients de corrélations de Pearson sur les pseudo-observations calculés en tenant compte de la proposition 1. Se faisant, on trouve les matrices de corrélation (9) pour les données du Lac Champlain et (10) pour celles de la rivière Clearwater.

$$\boldsymbol{\rho}_P(u^{(X)}, u^{(W)}, u^{(D)}) = \begin{pmatrix} 1.0000000 & 0.13423607 & 0.36863770 \\ 0.1342361 & 1.00000000 & 0.01053063 \\ 0.3686377 & 0.01053063 & 1.00000000 \end{pmatrix}. \quad (9)$$

$$\boldsymbol{\rho}_P(u^{(X)}, u^{(W)}, u^{(D)}) = \begin{pmatrix} 1.0000000 & 0.1537633 & 0.4171039 \\ 0.1537633 & 1.0000000 & 0.2673687 \\ 0.4171039 & 0.2673687 & 1.0000000 \end{pmatrix}. \quad (10)$$

On remarque alors que seule la dépendance en W et D pour les données du lac Champlain est faible. Un test de Mantel-Haensel (voir [Mantel, 1963]) ne permet pas de rejeter l'hypothèse nulle d'indépendance

entre les variables D et W pour ce cas-ci. Cependant, considérant que la dépendance est significative pour les autres paires de v.a., alors on considère tout de même une copule tri-variée pour modéliser la dépendance de cette base de données.

Afin de visualiser la forme des copules bi-variée constituant la copule en vigne, les illustrations 5 et 6 permettent de visualiser les copules empiriques à partir des nuages de points pour les pseudo-estimateurs définis dans la section 5.3.2. Par ailleurs, il faut garder à l'esprit que les v.a. W et D sont toutes-deux discrètes. Or, comme l'explique [Genest and Nešlehová, 2007], dans ce cas, plus d'une copule peut s'agencer aux observations puisque celles-ci perdent leur propriété d'unicité dans ce contexte. Cette observation fait en sorte que la forme erratique de la relation entre W et D dans l'illustration 5 peut être expliquée par ce phénomène. De plus, il faut prendre l'asymétrie des copules empiriques avec un grain de sel dans ce contexte. Malgré tout, il serait intéressant de tester une copule asymétrique et e comparer les résultats obtenus avec ceux présentés dans ce rapport, mais cela est laissé à d'autres étant donné le manque de temps pour le stage.

De manière plus formelle, afin de sélectionner les copules bi-variées, [Genest and Favre, 2007] recommande de comparer les critères de l'AIC et du BIC. Les tableaux 1 et 2 synthétisent les résultats obtenus pour les copules candidates offrant la meilleure adéquation. Pour faire la sélection des copules, le critère

Variables	Copules	θ	log-vrais.	AIC	BIC
$(u^{(X)}, u^{(D)})$	Gumbel	1.29	24.77	-47.55	-43.72
	Frank	2.33	24.27	-46.55	-42.72
	BB8	[2.45, 0.75]	25.83	-47.65	-40.00
	Clayton (180°)	0.49	24.81	-47.61	-43.78
	BB1 (180°)	[0.41,1.05]	25.29	-46.58	-38.92
$(u^{(X)}, u^{(W)} u^{(D)})$	Indépendance	\emptyset	0.00	0.00	0.00
	Gumbel	1.1	2.12	-2.25	1.58
	Frank	0.81	2.80	-3.60	0.23
	Clayton (180°)	0.18	2.26	-2.52	1.31
	Tawn type 1 (180°)	[1.34, 0.20]	3.57	-3.13	4.52

TABLEAU 1 – Résultats de l'estimation des copules candidates pour la base de données du lac Champlain.

choisit est le BIC pour des raisons de simplicité du modèle. Ce faisant, les copules sélectionnées sont celles apparaissant en gras dans les tableaux 1 et 2.

Du point de vue de la dépendance entre le processus aléatoire \mathbf{V} et la v.a. Z , bien que le rho de Spearman soit significativement supérieur à zéro pour les deux études de cas, l'hypothèse d'indépendance est tout de même utilisée puisque le contraire ajouterais beaucoup de complexité au modèle. Il faudrait alors trouver des

L'illustration 2 présente les graphiques quantiles-quantiles des résultats de simulation obtenus avec les deux études de cas.

Variables	Copules	θ	log-vrais.	AIC	BIC
$(u^{(X)}, u^{(D)})$	Gumbel	1.36	11.62	-21.24	-18.63
	BB1	[0, 1.36]	11.62	-19.24	-14.03
	Clayton (180°)	0.62	10.99	-19.98	-17.38
	Tawn type 1 (180°)	[1.95, 0.48]	11.63	-19.26	-14.05
$(u^{(W)}, u^{(D)})$	Frank	1.55	3.57	-5.14	-2.53
	BB8	3.89	3.64	-3.29	1.93
	Clayton (180°)	0.31	2.81	-3.63	-1.02
	Tawn type 1 (180°)	[2.06, 0.07]	3.82	-3.64	1.57
$(u^{(X)}, u^{(W)} u^{(D)})$	Indépendance	\emptyset	0.00	0.00	0.00
	Frank	0.84	1.14	-0.28	2.33
	Joe	1.17	0.90	0.21	2.81
	Clayton (180°)	0.24	1.30	-0.61	2.00

TABLEAU 2 – Résultats de l'estimation des copules candidates pour la base de données de la rivière Clearwater.

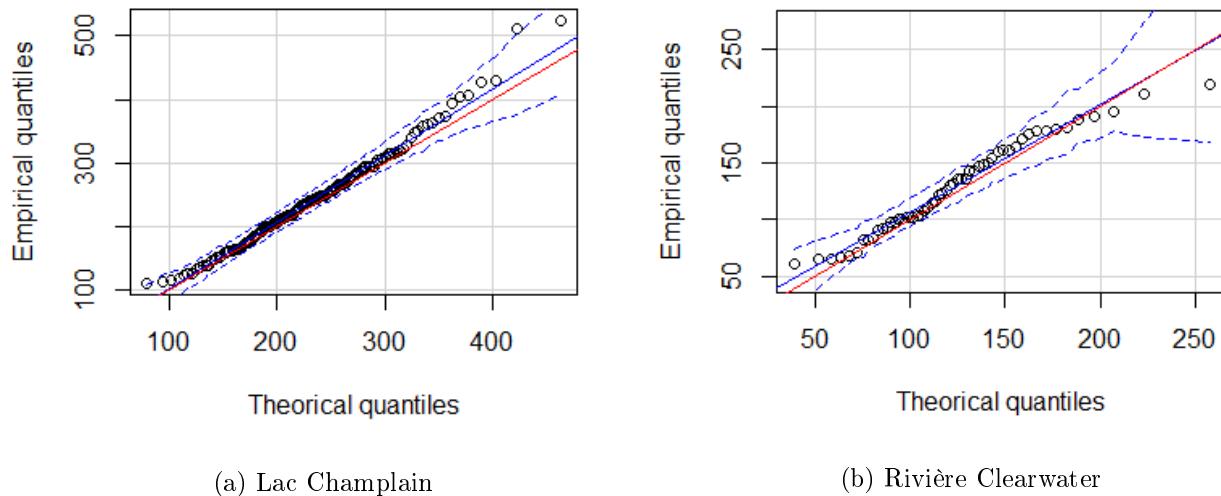


ILLUSTRATION 2 – Diagrammes quantiles-quantiles de $S_{T_0}^{(m)}(91)$: La ligne bleue présente la droite de tendance des quantiles. Les pointillés présentent l'intervalle de confiance au seuil de 5% pour cette droite et la ligne rouge est la diagonale d'adéquation parfaite. L'objectif est que la diagonale rouge se situe dans l'intervalle de confiance.

5.6 Apprentissages

Ce stage a permis de développer mes aptitudes à trouver de la

1. Copules en vigne
2. Méthode du maximum de vraisemblance généralisée
3. Algorithme MCMC et sa variante.
4. Travailler avec des processus de renouvellement alternés
5. Travailler dans un contexte de non-stationnarité. En dehors des processus de Poisson Non-homogène, qu'est-ce qu'il est possible de faire en termes de processus de renouvellement non-stationnaires
6. Comment tester l'indépendance séquentielle des données.
7. Travailler mes aptitudes en recherche de documentation, de rédaction et de programmation avec le langage R.

6 Évaluation du stage

6.1 Nature du travail

Ça a été long de trouver le sujet. Une fois le sujet trouvé, les choses se sont mises à filer. Par moment, la motivation était difficile à trouver :

1. Recherche de documentation laborieuse
2. apprentissage de nouvelles techniques telles que l'algorithme MCMC et les copules en vigne ont nécessité du temps.
3. Sensation d'être seul et de manquer de soutien quand les rencontres étaient trop espacées ou lorsque Fateh n'avait pas le temps de m'aider convenablement. Heureusement, Étienne et Hélène étaient là pour m'appuyer et m'aider quand Fateh ne le pouvait pas.
4. Sensation de fatigue accumulée au courant des 5 dernières années qui me rattrape de temps à autres.

La rédaction a été la partie la plus difficile. Particulièrement en matière de documentation.

J'ai travaillé à l'envers. J'ai construit un modèle au meilleur de mes connaissances acquises durant ma formation et avec les articles que j'ai lu jusqu'à ce jour. J'ai construit le modèle selon une approche logique et qui me semblait intuitive. Mais, par la suite, quant il est venu le temps d'appuyer les décisions prises avec la littérature, ça a été difficile de trouver des articles qui appuyaient mes décisions. J'aurais sans-doute mieux fait de lire d'avantage avant de construire le modèle et de commencer la programmation. Mais à ce moment là, je n'avais pas encore d'idée claire d'où je m'en allais et de ce qui devait être fait.

6.2 Environnement de travail

Contexte particulier de la pandémie. À domicile. Rencontres sur ZOOM ou sur Teams. Accès à distance aux ressources de documentation de l'université. Évaluation/appréciation ? Qu'est-ce qui aurait pu être

différent/amélioré ? J'aurais aimé que Fateh s'implique un peu plus dans le projet...surtout considérant qu'il voulait que son nom paraisse sur l'article. 15 minutes par semaine pour discuter de l'avancement et partager des idées, ce n'est pas assez. J'aurais aimé avoir plus d'aide pour naviguer dans la littérature en hydrologie et me faire davantage challenger sur mes idées.

6.3 Préparation théorique à l'université

1. Cours ACT-7000 : Modèles mathématiques en Actuariat :
 - (a) La théorie des valeurs extrêmes
 - (b) Comment paramétriser ce genre de modèle avec des méthodes basées sur les moments
2. ACT-2009 : Les processus de renouvellement stationnaires
3. ACT-3000 : Théorie du risque :
 - (a) Processus de renouvellement
 - (b) Théorie des copules
 - (c) Méthodes de simulation des processus de renouvellement.

7 Conclusion

En conclusion, je suis très fiers des résultats obtenus lors de ce stage. Bien que le contexte n'ait pas été idéal avec la pandémie,

1. Faire un rappel sur le projet, les contributions et les apprentissages.
2. Faire un rappel sur les points les plus importants de l'appréciation du stage.

Références

- [Andersen, 1957] Andersen, E. S. (1957). On the collective theory of risk in case of contagion between claims. *Bulletin of the Institute of Mathematics and its Applications*, 12(2) :275–279.
- [Bader et al., 2016] Bader, B., Yan, J., and Zhang, X. (2016). Automated threshold selection for extreme value analysis via goodness-of-fit tests with application to batched return level mapping. *arXiv preprint arXiv :1604.02024*.
- [Bader et al., 2018] Bader, B., Yan, J., Zhang, X., et al. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, 12(1) :310–329.
- [Beguería et al., 2011] Beguería, S., Angulo-Martínez, M., Vicente-Serrano, S. M., López-Moreno, J. I., and El-Kenawy, A. (2011). Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis : a case study in northeast spain from 1930 to 2006. *International Journal of Climatology*, 31(14) :2102–2114.

- [Chebana and Ouarda, 2021] Chebana, F. and Ouarda, T. B. (2021). Multivariate non-stationary hydrological frequency analysis. *Journal of Hydrology*, 593 :125907.
- [Cheng et al., 2014] Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Climatic change*, 127(2) :353–369.
- [Choulakian and Stephens, 2001] Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43(4) :478–484.
- [Dissmann et al., 2013] Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59 :52–69.
- [El Adlouni et al., 2007] El Adlouni, S., Ouarda, T. B., Zhang, X., Roy, R., and Bobée, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43(3).
- [Gallager, 2013] Gallager, R. G. (2013). *Stochastic processes : theory for applications*. Cambridge University Press.
- [Genest and Favre, 2007] Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4) :347–368.
- [Genest and Nešlehová, 2007] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin : The Journal of the IAA*, 37(2) :475–515.
- [Genest and Rémillard, 2004] Genest, C. and Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*, 13(2) :335–369.
- [Grimmett and Stirzaker, 2001] Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and random processes*. Oxford University Press.
- [Gudendorf and Segers, 2010] Gudendorf, G. and Segers, J. (2010). Extreme-value copulas. In *Copula theory and its applications*, pages 127–145. Springer.
- [Hogg et al., 2005] Hogg, R. V., McKean, J., and Craig, A. T. (2005). *Introduction to mathematical statistics*. Pearson Education.
- [Hosking and Wallis, 1987] Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3) :339–349.
- [Jalbert et al., 2019] Jalbert, J., Murphy, O. A., Genest, C., and Nešlehová, J. G. (2019). Modelling extreme rain accumulation with an application to the 2011 lake Champlain flood. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 68(4) :831–858.
- [Joe, 1997] Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.
- [Joe et al., 2010] Joe, H., Cooke, R. M., and Kurowicka, D. (2010). Regular vines : generation algorithm and number of equivalence classes. In *Dependence Modeling : Vine Copula Handbook*, pages 219–231. World Scientific.
- [Khaliq et al., 2006] Khaliq, M. N., Ouarda, T. B., Ondo, J.-C., Gachon, P., and Bobée, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations : A review. *Journal of hydrology*, 329(3-4) :534–552.

- [Kim et al., 2007] Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6) :2836–2850.
- [Klugman et al., 2013] Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2013). *Loss models : Further topics*. John Wiley & Sons.
- [Kysely et al., 2010] Kysely, J., Picek, J., and Beranová, R. (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change*, 72(1-2) :55–68.
- [Mantel, 1963] Mantel, N. (1963). Chi-square tests with one degree of freedom ; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303) :690–700.
- [Nelsen, 2006] Nelsen, R. B. (2006). An introduction to copulas. Springer, New York. *MR2197664*.
- [Northrop et al., 2015] Northrop, P., Attalides, N., and Jonathan, P. (2015). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *arXiv preprint arXiv:1504.06653*.
- [Renard and Lang, 2007] Renard, B. and Lang, M. (2007). Use of a gaussian copula for multivariate extreme value analysis : some case studies in hydrology. *Advances in Water Resources*, 30(4) :897–912.
- [Riboust and Brissette, 2016] Riboust, P. and Brissette, F. (2016). Analysis of lake Champlain/Richelieu river's historical 2011 flood. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, 41(1-2) :174–185.
- [Salvadori and De Michele, 2006] Salvadori, G. and De Michele, C. (2006). Statistical characterization of temporal structure of storms. *Advances in Water Resources*, 29(6) :827–842.
- [Salvadori and De Michele, 2007] Salvadori, G. and De Michele, C. (2007). On the use of copulas in hydrology : theory and practice. *Journal of Hydrologic Engineering*, 12(4) :369–380.
- [Shaw et al., 2010] Shaw, E. M., Beven, K. J., Chappell, N. A., and Lamb, R. (2010). *Hydrology in practice*. CRC press.
- [Small and Morgan, 1986] Small, M. J. and Morgan, D. J. (1986). The relationship between a continuous-time renewal model and a discrete markov chain model of precipitation occurrence. *Water Resources Research*, 22(10) :1422–1430.
- [Vandenbergh et al., 2010] Vandenbergh, S., Verhoest, N., and De Baets, B. (2010). Fitting bivariate copulas to the dependence structure between storm characteristics : A detailed analysis based on 105 year 10 min rainfall. *Water resources research*, 46(1).
- [Zhang and Singh, 2019] Zhang, L. and Singh, V. P. (2019). *Copulas and their applications in water resources engineering*. Cambridge University Press.

8 Exemple illustrant le modèle

Pour illustrer les différentes composantes du modèle décrit dans la section 5.2.1, le tableau 3 utilise les 15 premières observations printanières de la base de données Clearwater River, pour 5 années consécutives.

l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$Y_l^{(2009)}$	-	-	-	-	-	0.50	-	1.99	0.50	-	0.50	3.48	0.50	-	-
$Y_l^{(2010)}$	0.49	-	-	0.49	-	-	-	2.05	43.78	3.42	3.91	-	-	0.49	0.49
$Y_l^{(2011)}$	-	0.77	0.13	-	-	-	-	0.13	-	-	2.57	1.16	-	-	-
$Y_l^{(2012)}$	0.13	1.54	-	0.13	0.13	0.13	-	-	-	0.39	-	0.13	3.47	11.97	0.26
$Y_l^{(2013)}$	-	-	-	0.71	0.20	-	-	0.20	-	1.62	2.94	-	0.51	-	0.20

TABLEAU 3 – Observations de $Y_l^{(m)}$, pour $l = 1, \dots, 15$ et $m = 2009, \dots, 2013$, dans la base de données Clearwater River. Cette variable est représentée par la colonne Precip. (mm) et la période couvre du 1^{er} au 15 avril des années 2009 à 2013.

Avec les observations du tableau 3, on peut réaliser le clustering de manière à trouver les résultats du tableau 4. Puis, si on fixe un seuil $u = 1$, on obtient $\mathcal{X}^{(2009)} = \{2, 3\}$, $\mathcal{X}^{(2010)} = \{3\}$, $\mathcal{X}^{(2011)} = \{3\}$, $\mathcal{X}^{(2012)} = \{1, 4\}$, $\mathcal{X}^{(2013)} = \{3\}$. On trouve alors les résultats des tableaux 5, 6 et 7.

j	1	2	3	4	5
$C_j^{(2009)}$	6	8,9	11,12,13	-	-
$C_j^{(2010)}$	1	4	8,9,10,11	14,15	-
$C_j^{(2011)}$	2,3	8	11,12	-	-
$C_j^{(2012)}$	1,2	4,5,6	10	12,13,14,15	-
$C_j^{(2013)}$	4,5	8	10,11	13	15
$K_j^{(2009)}$	0.50	2.49	4.48	-	-
$K_j^{(2010)}$	0.49	0.49	53.2	0.98	-
$K_j^{(2011)}$	0.90	0.13	3.73	-	-
$K_j^{(2012)}$	1.67	0.39	0.39	15.83	-
$K_j^{(2013)}$	0.91	0.20	4.56	0.51	0.20

TABLEAU 4 – Clusterisation des observations du tableau 3.

i	1	2	3
X_i^{2009}	2.49	4.48	-
X_i^{2010}	53.2	-	-
X_i^{2011}	3.73	-	-
X_i^{2012}	1.67	15.83	-
X_i^{2013}	4.56	-	-
D_i^{2009}	2	3	-
D_i^{2010}	4	-	-
D_i^{2011}	2	-	-
D_i^{2012}	2	4	-
D_i^{2013}	2	-	-
W_i^{2009}	7	1	≥ 2
W_i^{2010}	7	≥ 4	-
W_i^{2011}	10	≥ 3	-
W_i^{2012}	0	9	-
W_i^{2013}	9	≥ 4	-

TABLEAU 5 – Construction des v.a. relatives aux événements excédents le seuil $u = 1$, à partir du tableau 4.

Avec les tableaux 4 et 5, on peut calculer les résultats présentés dans le tableau 7.

m	2009	2010	2011	2012	2013
$N_{T_0}^{(m)}(15)$	2	1	1	2	1
$V_{T_0}^{(m)}(15)$	6.97	53.2	3.73	17.5	4.56
$Z^{(m)}$	0.5	1.96	1.03	0.78	1.82
$S_{T_0}^{(m)}(15)$	7.47	55.16	4.76	18.28	6.38

TABLEAU 7 – Représentation de la quantité de pluie totale tombée pendant les 15 premiers jours du printemps des années 2009 à 2013 selon la notation présentée dans la section 5.2.

Format date			
i	0	1	2
$T_i^{(2009)}$	2009-04-01	2009-04-08	2009-04-11
$T_i^{(2010)}$	2010-04-01	2010-04-08	-
$T_i^{(2011)}$	2011-04-01	2011-04-11	-
$T_i^{(2012)}$	2012-04-01	2012-04-01	2012-04-12
$T_i^{(2013)}$	2013-04-01	2013-04-10	-

Format numérique			
i	0	1	2
$T_i^{(2009)}$	0	7	10
$T_i^{(2010)}$	365	372	-
$T_i^{(2011)}$	730	740	-
$T_i^{(2012)}$	1096	1096	1107
$T_i^{(2013)}$	1461	1470	-

TABLEAU 6 – Temps d'arrivée des événements décrits dans le tableau 5.

A Illustrations

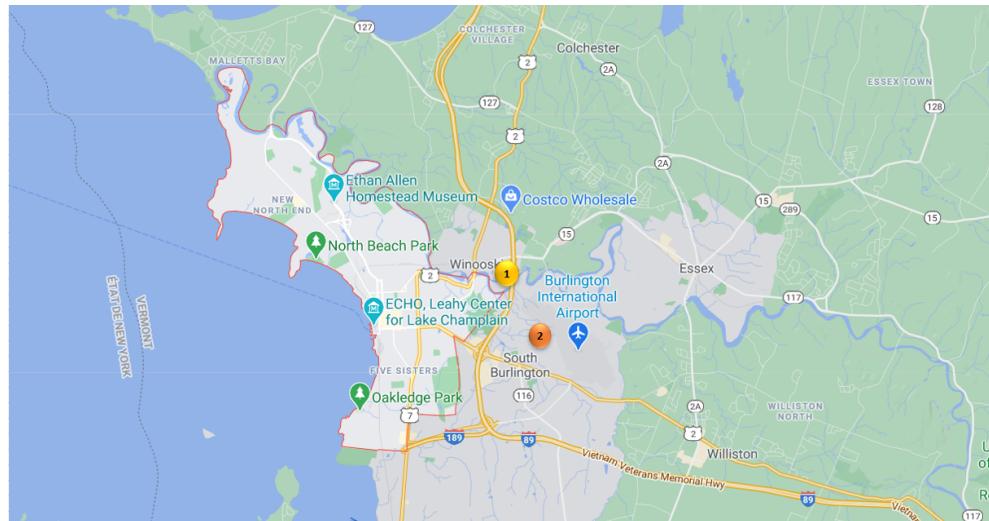


ILLUSTRATION 3 – Emplacement des stations (1) USC00431072 et (2) USW00014742 pour les données de Burlington, au Vermont, USA. (Image tirée de Google Map)

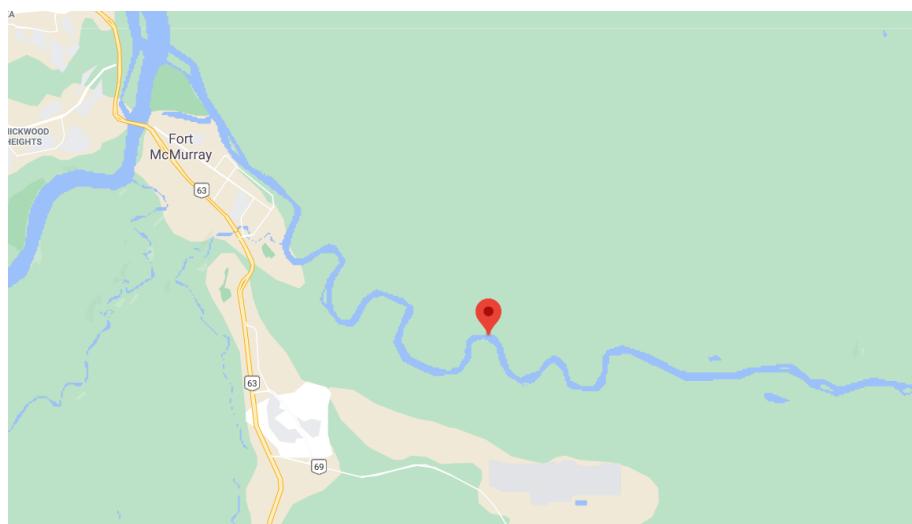


ILLUSTRATION 4 – Emplacement de la station 07CD001 où ont été collectées les données pour l'exemple de la rivière Clearwater. Les coordonnées de la stations sont Longitude -111.2554 et Latitude 56.68528. (Image tirée de Google Map)

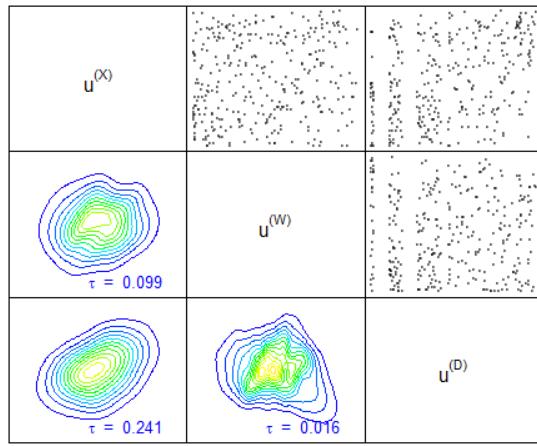


ILLUSTRATION 5 – Copules bi-variées empiriques pour les données du lac Champlain. Triangle supérieur de la matrice : Nuage de points des paires de v.a. Triangle inférieur de la matrice : graphiques de contours pour les estimations des copules empiriques selon une approche par noyau gaussien.

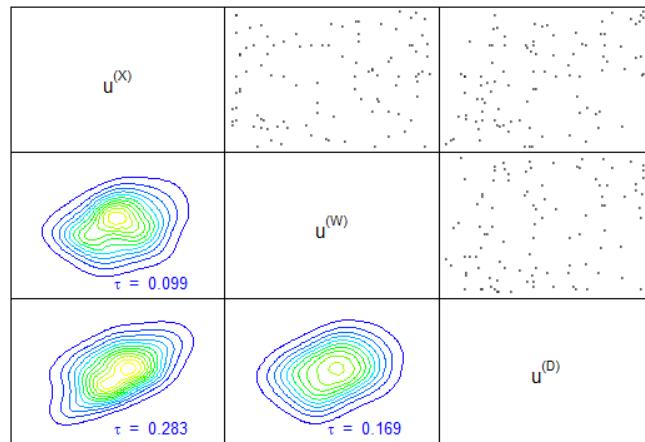


ILLUSTRATION 6 – Copules bi-variées empiriques pour les données de la rivière Clearwater. Triangle supérieur de la matrice : Nuage de points des paires de v.a. Triangle inférieur de la matrice : graphiques de contours pour les estimations des copules empiriques selon une approche par noyau gaussien.

B Algorithme de simulation

Soit $F_{X-u|X>u}$, F_W et F_D , les fonctions de répartition théoriques des v.a. $\{X-u|X > u\}$, W et D . Soit $F_{X-u|X>u}^{-1}$, F_W^{-1} et ${}^{-1}F_D$, les fonctions quantiles de ces mêmes v.a. La procédure de simulation permettant de générer des réalisations de $S(m)(91)$ est décrit dans l'algorithme 1.

Algorithm 1: Simuler un processus de renouvellement alterné avec récompense utilisant les copules pour modéliser la dépendance.

```

input :  $m$  (entiers) : vecteur des années pour lesquelles on désire simuler ;
         $t$  (entier) : la durée du processus (91 jours dans notre cas) ;
         $N_{\max}$  (entier) : Le nombre maximum d'événements attendu dans une année (p. ex. 100) ;
         $n$  (entier) : nb de réalisations de la simulation (p. ex.  $n = 10^3$ ).
output: matrice de dimension  $n \times \text{card}(m)$  correspondant à des réalisations du processus  $S_{T_0}^{(m)}(t)$ .
1 foreach  $m$  do
2   poser  $T_0 = \text{as.numeric(as.Date("m-04-01"))}$  ;
3   poser  $t_{\max} = T_0 + t$  ;
4   for  $j \leftarrow 1$  to  $n$  do
5     simuler  $(u_l^{(X)}, u_l^{(W)}, u_l^{(D)})$ ,  $l = 1, \dots, N_{\max}$ , des réalisations de la copule  $G$  ;
6     poser  $i = 0$  ;
7     do
8       incrémenter  $i = i + 1$  ;
9       calculer  $W_i = F_W^{-1}(u_i^{(W)} | T_{i-1})$ , une réalisation de  $W^{(m)}$  ;
10      calculer  $T_i = T_{i-1} + W_i$ , une réalisation de  $T^{(m)}$  ;
11      calculer  $D_i = F_D^{-1}(u_i^{(D)} | T_i)$ , une réalisation de  $D^{(m)}$  ;
12      calculer  $T_i = T_i + D_i$ , une réalisation de  $T^{\star(m)}$  ;
13     while  $T_i \leq t_{\max}$ 
14     calculer  $N = i - 1$ , une réalisation de  $N_{T_0}^{(m)}(t)$  ;
15     if  $N = 0$  then
16       poser  $V = 0$  ;
17     else
18       for  $i \leftarrow 1$  to  $N$  do
19         calculer  $X_i = u + F_{X-u|X>u}^{-1}(u_i^{(X)})$ , une réalisation de  $X^{(m)}$  ;
20       end
21       calculer  $V = \sum_{i=1}^N X_i$ , une réalisation de  $V_{T_0}^{(m)}(t)$  ;
22     end
23   end
24   simuler  $Z$ , une réalisation de  $Z^{(m)}$  ;
25   calculer  $S_{j,m} = V + Z$ , une réalisation de  $S_{T_0}^{(m)}(t)$  ;
26 end
27 end
28 return S

```
