UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Modélisation du risque climatique via les réseaux bayésiens et approches causales

Application au risk management des compagnies d'assurance

Auteur : **S**amir **D**ahche

Promoteur : **P**ierre **A**rs

Lecteur : **D**onatien **H**ainaut

Année académique 2023-2024

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute mon appréciation.

Je tiens en premier lieu à exprimer ma profonde reconnaissance à mon promoteur de mémoire, le professeur Pierre Ars. Sa patience, sa disponibilité et son écoute attentive ont été pour moi une source d'inspiration et de motivation tout au long de ce travail.

Je tiens également à exprimer ma gratitude envers tout le corps enseignants de la LSBA. La formation rigoureuse et diversifiée qu'ils m'ont apportée a joué un rôle essentiel dans ma formation académique et personnelle.

Enfin, je remercie chaleureusement ma famille. Ma mère Ajouz Nada, mon père Dahche Hadi, ainsi que ma sœur Dahche Nelly, pour le soutien indéfectible qu'ils m'ont apporté. Grâce à leurs encouragements, j'ai trouvé la force de persévérer et d'atteindre cet objectif.

Table des matières

1 Introduction. 3

2 Revue de littérature 5
2.1 Le risque climatique . . . 5
2.1.1 Impact sur les compagnies d'assurance . . . 5
2.1.2 Émergence du risque climatique . . . 6
2.2 Modèles de catastrophes . . . 7
2.2.1 Définition . . . 7
2.2.2 Composantes . . . 7
2.2.3 Types de résultats . . . 8
2.3 Réseaux Bayésiens et fondements théoriques . . . 9
2.3.1 Théorie des graphes . . . 9
2.3.1.1 Graphes, nœuds et arêtes . . . 9
2.3.1.2 Relations dépendantes . . . 9
2.3.1.3 Cyclicité . . . 10
2.3.1.4 Parenté . . . 10
2.3.2 Approche probabiliste . . . 11
2.3.2.1 Principes généraux . . . 11
2.3.2.2 Importance de l'indépendance . . . 13
2.3.2.3 L'indépendance conditionnelle . . . 13
2.4 Conception et paramétrisation des réseaux bayésiens . . . 15
2.4.1 Démarche suivie . . . 15
2.4.2 Conceptualisation . . . 16
2.4.3 Caractéristiques du modèle . . . 17
2.4.4 Apprentissage de la structure . . . 18
2.4.4.1 Structure basée sur la littérature et les experts . . . 18
2.4.4.2 Structure basée sur la contrainte . . . 19
2.4.4.3 Structure basée sur le score . . . 20
2.4.5 Apprentissage des paramètres . . . 20
2.5 Inférence bayésienne . . . 21
2.5.1 Types de preuves : Hard et Soft Evidence . . . 21
2.5.2 Conditional Probability Queries (CPQ) . . . 21
2.5.3 Maximum A Posteriori Queries (MAP) . . . 22
2.5.4 Belief Updating . . . 22
2.5.5 Algorithmes pour le Belief Updating . . . 23
2.5.5.1 Logic Sampling Algorithm . . . 23

2.5.5.2 Likelihood Weighting Algorithm . . . 23
2.6 Outils statistiques . . . 25
2.6.1 Distribution log-normale . . . 25
2.6.2 Test de normalité . . . 25
2.6.3 Test d'indépendance conditionnelle . . . 25
2.6.3.1 DBNs . . . 25
2.6.3.2 GBNs . . . 26
2.6.3.3 CLGBNs . . . 26
2.6.4 Score d'un réseau . . . 27
2.6.4.1 DBNs . . . 27
2.6.4.2 Vraisemblances pénalisées . . . 27
2.6.5 Discrétisation . . . 27
2.6.5.1 Discrétisation par intervalles égaux . . . 28
2.6.5.2 Discrétisation par quantiles . . . 28
2.6.5.3 Méthode de Hartemink . . . 28
2.6.6 Validation croisée . . . 28
2.6.7 Extreme Value Theory (EVT) . . . 29
2.6.7.1 Les deux grandes méthodes de l'EVT . . . 29
2.6.7.2 Peaks-Over-Threshold (POT) . . . 30
2.6.7.3 Distribution Pareto Généralisée (GPD) . . . 30
2.6.7.4 Estimateur de Hill . . . 30
2.6.7.5 Value at Risk . . . 30
2.6.7.6 Expected Shortfall . . . 31

3 Application **33**
3.1 Sources et présentation des données . . . 33
3.1.1 Préparation des données . . . 34
3.1.1.1 Horizon de temps . . . 34
3.1.1.2 Variables discrètes . . . 34
3.1.1.3 Variables continues et discrétisation . . . 34
3.1.1.4 Valeurs manquantes . . . 35
3.1.1.5 Hypothèses . . . 36
3.2 Apprentissage de la structure . . . 37
3.2.1 Par la littérature et l'expertise . . . 37
3.2.2 Par maximisation du score . . . 38
3.2.3 Par contrainte . . . 38
3.2.4 Choix . . . 39
3.3 Paramétrisation et validation des modèles . . . 39
3.4 Propriétés du modèle . . . 41
3.4.1 CPDs et CTPs . . . 41
3.4.1.1 Distribution de probabilité conditionnelle : Temperature | GtCO2 41
3.4.1.2 Table de probabilité marginale : Disaster_Subtype . . . 42
3.4.1.3 Table de probabilités conditionnelles : Vulnerability | ISO . . . 42
3.4.1.4 Table de probabilité marginale : ISO . . . 42
3.4.1.5 Distribution de probabilité conditionnelle : Damage | Temperature, Disaster Subtype, Vulnerability . . . 43
3.4.2 Analyse des résidus de *Damage* . . . 43

3.5 Prédiction 45
3.6 Scénario futur 45
3.7 Association avec un modèle de fréquence 47
3.7.1 Introduction 47
3.7.2 Exploration préliminaire 47
3.7.3 Ajustement et validation d'un modèle 47
3.7.4 Simulations des pertes agrégées 48
3.8 Application de la théorie des valeurs extrêmes 49
3.8.0.1 Sélection du seuil 49
3.8.0.2 Validation du modèle GPD 50
3.8.0.3 Calcul de la Value at Risk et de l'Expected Shortfall à 99,5% . 51
3.8.0.4 Conclusion 51

4 Limitations et développements futurs 53

5 Conclusion 55

Annexes 61

Abréviations

BNs	Réseaux Bayésiens
DBNs	Réseaux Bayésiens Discrets
GBNs	Réseaux Bayésiens Gaussiens
CLGBNs	Réseaux Bayésiens Gaussiens Linéaires Conditionnels
CPT	Table de probabilité conditionnelle
CDP	Distribution de probabilité conditionnelle

Chapitre 1

Introduction.

Le réchauffement climatique représente un défi sans précédent pour les modèles traditionnels de gestion des risques utilisés par les compagnies d'assurance. Ces modèles, historiquement basés sur l'observation de données passées, sont de plus en plus biaisés par les transformations rapides et imprévisibles du climat. Les événements extrêmes, tels que les inondations, tempêtes et vagues de chaleur, deviennent non seulement plus fréquents mais aussi plus intenses, rendant ces approches inadéquates.

Dans ce contexte, il devient impératif de développer des méthodes capables de modéliser les relations causales entre les facteurs climatiques et leurs impacts potentiels. C'est ici que les réseaux bayésiens (BN) entrent en jeu. Ces modèles graphiques probabilistes permettent de représenter les connaissances relatives à un domaine incertain et d'ajuster continuellement les probabilités en fonction des nouvelles données, offrant ainsi une méthode adaptative et réactive. De plus, ces réseaux facilitent l'analyse de scénarios "what-if", cruciale pour anticiper les impacts de futurs événements et maintenir la pertinence des modèles face à des environnements climatiques changeants.

Ce mémoire est structuré pour explorer d'abord le risque climatique et son impact sur les compagnies d'assurance, en exposant les défis posés par les événements climatiques extrêmes et en soulignant l'importance de développer des modèles robustes et adaptés à ces nouvelles réalités. Ensuite, il donne un aperçu de l'utilité des modèles de catastrophes, pour déboucher sur la théorie des graphes et des probabilités, menant à l'émergence des BN comme outil pour la modélisation des risques climatiques.

Une attention particulière est ensuite accordée à la conception d'un BN appliqué à cette problématique spécifique. Le mémoire détaille les étapes de conception du modèle, le choix des variables, et son implémentation pour simuler les pertes potentielles liées aux événements extrêmes. Cependant, les BN traditionnels présentent certaines limitations : la survenance des phénomènes météorologiques est souvent influencée par des processus complexes et chaotiques, rendant leur modélisation directe dans un BN particulièrement difficile sans une surabondance de variables et de données. Pour pallier cette difficulté, une approche complémentaire est proposée en associant un modèle de fréquence. Cette intégration permet d'enrichir les informations fournies par le BN et de proposer un modèle de catastrophe plus complet et adapté.

Chapitre 2

Revue de littérature

2.1 Le risque climatique

2.1.1 Impact sur les compagnies d'assurance

Dans le monde de l'assurance, nous distinguons principalement deux variétés de risques climatiques, chacun posant ses propres défis. Premièrement, il y a le risque physique, celui qui vient instinctivement à l'esprit lorsqu'on évoque les changements climatiques. Ce type de risque englobe une vaste gamme de phénomènes naturels – des inondations aux ouragans, en passant par les sécheresses, les vagues de chaleur, l'émergence de maladies, les pandémies, l'acidification des océans, et les risques de subsidence, entre autres. Par leur nature et leur intensification, ils contribuent à une hausse notable des sinistres et, par conséquent, des demandes d'indemnisation. La capacité à comprendre et à anticiper cette évolution est essentielle pour évaluer les pertes financières potentielles et ajuster en conséquence sa stratégie. [2] La seconde variété de risque climatique concerne le risque de transition. Dans une ère qui aspire de plus en plus à la durabilité, les régulateurs et les acteurs du marché imposent des exigences accrues en matière de transition vers une économie faiblement carbonée. L'incertitude plane sur le "quand" et le "comment" de l'implémentation de nouvelles réglementations, rendant la prudence indispensable. Pour une compagnie d'assurance, posséder un portefeuille d'investissements dominé par des entreprises fortement polluantes, notamment celles dépendantes du charbon, représente un risque de transition considérable. Car une nouvelle réglementation européenne visant à restreindre ou interdire l'utilisation du charbon affecterait sévèrement la valeur de ces investissements et la solvabilité de l'entreprise. La gestion de ce risque de transition nécessite donc une stratégie réfléchie et une adaptation proactive aux évolutions réglementaires et aux attentes du marché [2]. Dans ce travail, l'attention se portera avant tout sur le risque climatique physique, principalement abordé dans le secteur à travers l'utilisation de modèles prédictifs conçus pour estimer les potentielles pertes dues aux catastrophes, [4]. Les risques de transition, quant à eux, peuvent être analysés par d'autres méthodes, telles que le stress testing ou les modèles de gestion Actif-Passif [5].

2.1.2 Émergence du risque climatique

Nous vivons à une époque où le changement climatique, accéléré par des années d'activité industrielle, remodèle notre environnement de façon significative. Ce phénomène entraîne une augmentation des événements météorologiques extrêmes, qu'il s'agisse de situations déjà connues devenant plus fréquentes et plus intenses, ou de nouveaux risques émergents nécessitant une vigilance accrue. Traditionnellement, les compagnies d'assurance s'appuient sur des données historiques pour élaborer des modèles prédictifs destinés à évaluer les risques et à déterminer les coûts des primes d'assurance. Cependant, avec l'évolution constante du paysage, ces modèles basés sur le passé peuvent perdre en précision et ne plus être fiables pour prédire les risques futurs. Et pour certains risques émergents, les données sont tout simplement manquantes [3]. L'Agence européenne pour l'environnement illustre ce changement dans son rapport *Weather and Climate Extreme Events in a Changing Climate* [6], comme le montre l'infographie 2.1.1, qui présente à l'échelle mondiale la tendance à l'augmentation de la fréquence et de la gravité de trois aléas climatiques au cours des dernières décennies. [1]

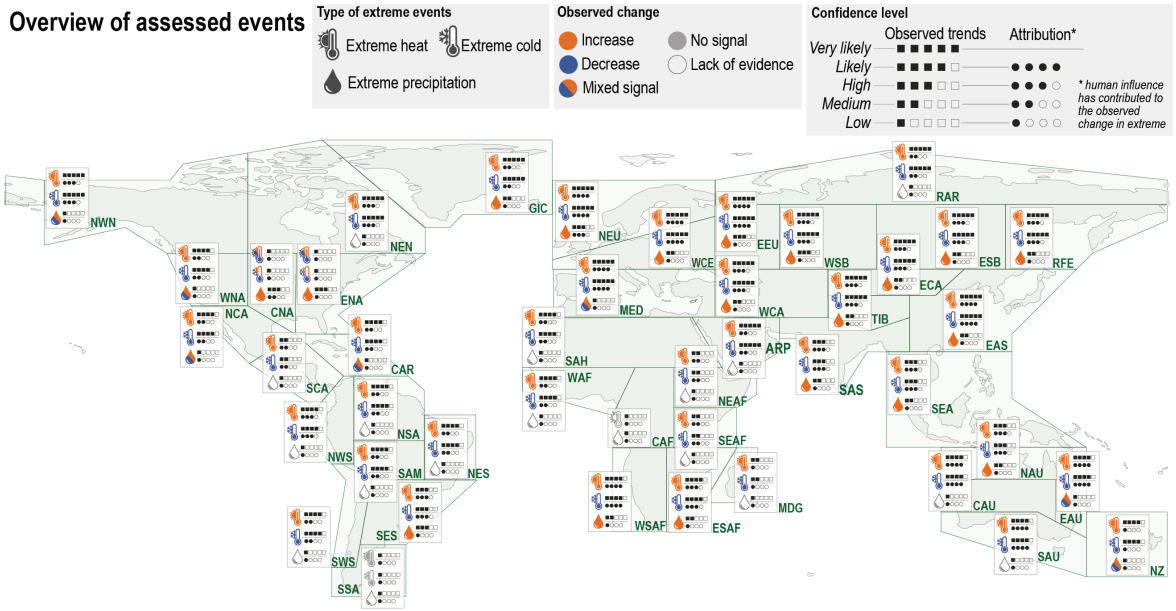


FIGURE 2.1.1 – Vue d'ensemble des changements observés pour les extrêmes de froid, de chaleur et d'humidité, ainsi que leur contribution humaine potentielle. [8]

1. Selon l'ORECC (Observatoire Régional des Effets du Changement Climatique), « l'aléa climatique est un évènement climatique ou d'origine climatique susceptible de se produire (avec une probabilité plus ou moins élevée) et pouvant entraîner des dommages sur les populations, les activités et les milieux. Les aléas peuvent être soit des évolutions tendancielles, soit des extrêmes climatiques. » (lien vers le glossaire).

2.2 Modèles de catastrophes

2.2.1 Définition

Comprendre les risques associés aux polices d'assurance et aux politiques de souscription est un aspect fondamental de la gestion d'une compagnie d'assurance. Dans cette optique, les modèles de catastrophes sont conçus pour évaluer les dommages potentiels causés par des catastrophes naturelles. Utilisés par les assureurs et réassureurs, ces modèles permettent d'identifier les portefeuilles à risque, de prendre des mesures de mitigation, et de déterminer les niveaux de capital nécessaires pour absorber les pertes potentielles sans compromettre la stabilité financière de l'entreprise. [33].

Un modèle de catastrophe aide à quantifier non seulement les pertes, mais aussi à déterminer la probabilité, les causes, et les localisations des risques en intégrant une expertise provenant de plusieurs disciplines, notamment la météorologie, la sismologie, l'hydrologie et l'ingénierie, offrant ainsi une évaluation complète du risque global [33].

2.2.2 Composantes

Un modèle de catastrophe se base sur trois composantes principales, chacune étant essentielle pour fournir une analyse complète du risque :

$$\text{Risque} = f(\text{Aléa}, \text{exposition}, \text{vulnérabilité})$$

FIGURE 2.2.1 – Noyau d'un modèle de catastrophe [2]

L'Aléa : L'aléa, aussi appelé péril, est un événement naturel pouvant causer des dommages importants aux biens, aux infrastructures, ou à l'environnement. Il s'agit d'une cause potentielle de sinistre, comme une inondation, un tremblement de terre, ou une tempête. Chaque péril est caractérisé par sa fréquence, sa sévérité, et sa distribution géographique. Les aléas constituent la première composante d'un modèle de catastrophe en apportant l'information sur l'intensité du danger dans un cadre géospatial [33].

Exposition : La composante d'exposition quantifie la valeur des biens exposés aux dangers. Cela inclut les propriétés immobilières, le contenu des bâtiments, les interruptions d'activité et les pertes humaines. L'exposition est généralement mesurée en termes de valeurs

2. Source : `https://content.naic.org/insurance-topics/catastrophe-models-(property)`

économiques ou assurées, et elle peut être segmentée selon différents critères, comme le type de biens ou leur localisation géographique [33].

Vulnérabilité : La vulnérabilité représente la susceptibilité des actifs exposés à subir des dommages lorsqu'un péril survient. Elle dépend fortement des caractéristiques de l'actif, comme la construction, l'emplacement, et la résilience face aux forces imposées par le péril [33].

2.2.3 Types de résultats

Avec ces trois composantes assemblées, les modèles de catastrophes permettent de générer plusieurs métriques financières importantes [33] :

- **Annual Average Loss (AAL)** : Représente la perte annuelle attendue. [35].
- **Standard Deviation (SD) autour de l'AAL** : Mesure la volatilité des pertes autour de l'AAL. Dans le cadre des dégâts catastrophes, cette SD est généralement élevée.
- **Exceedance Frequency (EF)** : Indique la fréquence annuelle des événements qui devraient entraîner des pertes supérieures à un certain montant. L'EF peut dépasser 1.0, ce qui indique qu'il est possible d'avoir plus d'un événement dépassant un certain seuil en une seule année, car il s'agit d'une fréquence cumulative et non d'une probabilité unique. La réciproque de l'EF est la période de retour d'un événement particulier. [3]
- **Occurrence Exceedance Probability (OEP)** : Probabilité que la perte maximale en une année dépasse un certain seuil.
- **Aggregate Exceedance Probability (AEP)** : Probabilité que la somme des pertes en une année dépasse un certain niveau.
- **Value at Risk (VaR)** : Mesure la perte à un quantile spécifique de la distribution des pertes. Par exemple, un VaR à 99,5% indique la perte à un niveau de 0,5% de la courbe AEP, soit une période de retour de 1 en 200 ans. Le VaR est crucial pour déterminer les réserves de capital nécessaires dictées par *Solvency II*. [4]

3. La période de retour mesure l'intervalle de temps moyen qui sépare deux événements d'une même intensité ou supérieure.

4. Cadre légal encadrant les activités de l'assurance et de la réassurance.

2.3 Réseaux Bayésiens et fondements théoriques

2.3.1 Théorie des graphes

La théorie des graphes, une branche fascinante des mathématiques, permet de créer des structures composées de nœuds, aussi appelés sommets, et d'arêtes, aussi appelées liens, qui connectent ces nœuds. Cela permet de dresser une cartographie visuelle des relations entre différents objets [9]. Ce concept est utilisé dans de nombreux domaines scientifiques. Par exemple, en médecine, il aide à représenter les interactions entre les protéines et à analyser les réseaux de neurones. En cryptographie, les graphes sont essentiels pour développer et évaluer des algorithmes sûrs. Dans le contexte des réseaux sociaux, ils permettent de mieux comprendre les relations entre les groupes, les influenceurs et leurs communautés. Aussi, en informatique, les graphes jouent un rôle crucial, notamment dans la création d'algorithmes performants, comme ceux utilisés par les GPS pour trouver les meilleurs itinéraires. [9] [10].

2.3.1.1 Graphes, nœuds et arêtes

La figure 2.3.1 illustre un graphe dans une de ses formes les plus élémentaires. Représenté par $G = (V, E)$, un graphe est composé d'un ensemble fini et non vide de sommets V et d'un ensemble non vide d'arêtes E. Ici, les points X, Y et Z représentent l'ensemble V de sommets, qui sont connectés par E les arêtes A et B. Ce graphe est considéré comme *incomplet* car il manque une arête pour établir une relation entre X et Z. À l'inverse, un graphe est qualifié de *complet* lorsque chaque paire de nœuds est reliée par une arête [10] [11].

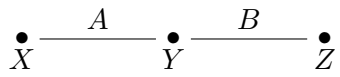


FIGURE 2.3.1 – Exemple de graphe linéaire avec nœuds et arêtes [11]

2.3.1.2 Relations dépendantes

Pour comprendre les chaînes de relations dans un graphe, l'ordre et la direction des connexions sont essentiels. Imaginons la cascade de relations suivante :

Émission de CO_2 → Absorption de ce CO_2 par les Océans → Formation d'acide carbonique

Tout d'abord, les émissions de dioxyde de carbone provenant des activités humaines augmentent les niveaux de CO_2 dans l'atmosphère. En réaction, les océans absorbent une partie de ce CO_2 excédentaire, un processus naturel qui contribue à équilibrer les niveaux de CO_2 dans l'atmosphère. Ensuite, une fois dissous dans l'eau de mer, le CO_2 réagit avec l'eau pour former de l'acide carbonique (H_2CO_3). Cet acide faible se dissocie en ions bicarbonate et hydrogène, augmentant ainsi l'acidité de l'eau de mer [12]. La figure 2.3.2 représente graphiquement cette réalité à l'aide de nœuds et d'arêtes. Un graphe directionnel est de cette manière obtenu : Émission de CO_2 (Nœud X) → Absorption de CO_2 par les Océans (Nœud Y) → Formation d'Acide Carbonique (Nœud Z). Lorsqu'un ordre est explicitement donné sur la paire de nœuds finaux (u, v), avec $(u, v \in V)$, cette paire (u, v) définit une arête directionnelle et nous disons que $G = (V, E)$ est dirigé [10] [11].

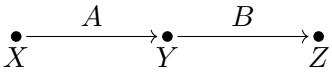


FIGURE 2.3.2 – Exemple de graphe linéaire avec nœuds et arêtes directionnelles [11]

2.3.1.3 Cyclicité

Un graphe est dit cyclique s'il existe un chemin permettant de partir d'un nœud, de suivre un certain nombre d'arêtes dans le graphe, et de revenir à ce même nœud. Sinon, il est acyclique. Par exemple, la figure 2.3.3 montre un graphe où les flèches vont de A à B, de B à C, et de C à A. Ce graphe est cyclique car on peut revenir à A en suivant les flèches. En revanche, si les flèches vont seulement de A à B, puis de B à C, sans revenir à A, alors le graphe est acyclique [11].

(a) Graphe acyclique (b) Graphe cyclique

FIGURE 2.3.3 – (a) Montre un graphe acyclique et (b) un graphe cyclique

2.3.1.4 Parenté

Un parent est une variable (ou nœud) qui influence directement une autre variable, appelée enfant. L'enfant dépend de ses parents de manière similaire à une relation hiérarchique dans un organigramme.

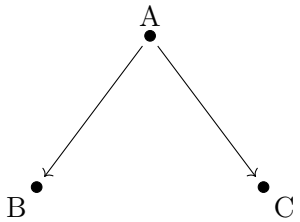


FIGURE 2.3.4 – Exemple de relation parent/enfants

2.3.2 Approche probabiliste

2.3.2.1 Principes généraux

Supposons qu'un actuaire souhaite évaluer le risque de dommages causés par une tempête de neige extrême dans une région en prenant en compte deux variables principales : si l'altitude de la région est adéquate (qui influence la quantité de neige) et si la température prévue est suffisamment basse (qui détermine si la neige va fondre ou s'accumuler). Quelle approche devrait-il prendre ?

- **Modélisation mécanistique** : Cette approche permet de créer des modèles basés sur une compréhension détaillée des mécanismes sous-jacents d'un système. Par exemple, avec l'usage d'équations différentielles et de connaissances en chimie, on peut prédire le comportement d'un système de solutions chimiques [36] [37].

  Cependant, cette méthode comporte deux limites majeures :
  - Si les mécanismes sont inconnus,
  - Ou si la problématique comporte trop de variables, la méthode n'est pas praticable.
- **Modélisation basée sur les données** : À partir des données récoltées, il est possible d'utiliser des algorithmes pour capturer des relations complexes sans nécessiter une compréhension des mécanismes sous-jacents [36] [37]. Devant un tableau de données météorologiques fourni, l'actuaire qui se demande si demain il y aura des dégâts dans une région où l'altitude et la température sont propices peut poser la question de la manière suivante :

  $$P(\mathbf{D}\text{ommages} = T \mid \mathbf{A}\text{ltitude} = T, \mathbf{T}\text{empérature} = T) = ?$$

  Et par application du théorème de Bayes [50] :

Définition 1 (Théorème de Bayes). Soit $A_1, \ldots, A_n$ un système complet d'événements, tous de probabilité non nulle. Soit B un événement. Alors, pour tout $j \in \{1, \ldots, n\}$, on a :

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}.$$

Notre problématique devient :

$$P(\mathrm{D} = T \mid \mathrm{A} = T, \mathrm{T} = T) = \frac{P(\mathrm{A} = T, \mathrm{T} = T \mid \mathrm{D} = T) \cdot P(\mathrm{D} = T)}{P(\mathrm{A} = T, \mathrm{T} = T)}$$

La clef pour résoudre ce type de problème est de trouver la densité de probabilité jointe $P(D, A, T)$, qui permet de dériver toutes les probabilités conditionnelles et marginales nécessaires. Passer par la densité de probabilité jointe permet d'économiser beaucoup d'espace et de temps. Plutôt que de créer de grandes tables pour chaque combinaison de variables, il suffit de connaître les probabilités conditionnelles et marginales individuelles.

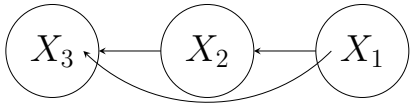
Dans la recherche de cette probabilité jointe, chaque variable est calculée conditionnellement à d'autres variables antérieures que l'on peut qualifier aussi de parents. Formellement, cela s'écrit comme suit :

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i \mid pa_i)$$

Et dans ce modèle à 3 variables, la densité de probabilité jointe sera de la forme suivante :

$$P(X_3, X_2, X_1) = P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

En indiquant la relation parent-enfant par des flèches, une représentation visuelle de ces distributions de probabilité conditionnelles (CPDs) peut être faite :



Cette représentation visuelle est communément appelée un réseau bayésien (BN), qui est en réalité la rencontre entre la théorie des graphes et la théorie des probabilités.

Remarquons que, dans un BN, les propriétés suivantes s'appliquent [11] :

- **Structure de graphe orienté et acyclique (DAG)** :
  - Orienté : Les arêtes ont toutes une direction, allant d'un nœud parent à un nœud enfant.
  - Acyclique : Il n'y a pas de cycles dans le graphe.
- **Nœuds** :
  - Chaque nœud du graphe représente une variable aléatoire.
  - Les nœuds peuvent représenter des variables observables ou latentes (non observables).
- **Arêtes** :
  - Une arête dirigée de X vers Y indique que X est un parent de Y et que Y est conditionnellement dépendant de X.
- **Table de probabilité conditionnelle (CPT)** :
  - Associée à chaque nœud, il y a une CPT qui spécifie la distribution de probabilité de la variable, étant donné les valeurs de ses parents.
  - Si un nœud n'a pas de parent, sa CPT est simplement sa distribution de probabilité marginale.

2.3.2.2 Importance de l'indépendance

Le problème à 3 variables binaires auquel l'actuaire fait face ne demande qu'une estimation de 8 paramètres.

Sol glissant	Pluie	Humidité	$P(S, P, H)$
T	T	T	p_1
T	T	F	p_2
T	F	F	p_3
T	F	T	p_4
F	T	T	p_5
F	F	T	p_6
F	T	F	p_7
F	F	F	p_8

TABLE 2.3.1 – Tableau de probabilités conjointes pour Sol glissant, Pluie, et Humidité

Dans des situations plus réalistes, notre actuaire ne fera pas face à seulement 3 variables binaires, mais souvent à un nombre beaucoup plus élevé de variables, dont certaines peuvent prendre plusieurs états différents. Par exemple, un simple modèle avec 10 variables binaires et 2 variables qui peuvent prendre 4 catégories demande déjà $2^{10} \times 4 \times 4 = 16384$ combinaisons possibles. Les calculs deviennent rapidement astronomiques, et il est nécessaire de recourir à des simplifications.

Pour un ensemble de variables $X_1, X_2, \ldots, X_n$, si chaque paire de variables est indépendante, la densité de probabilité jointe $P(X_1, X_2, \ldots, X_n)$ se décompose en la multiplication des densités marginales $P(X_1) \times P(X_2) \times \cdots \times P(X_n)$. Cela réduit considérablement la complexité des calculs, passant de l'évaluation d'une fonction à n dimensions à la simple multiplication de n fonctions unidimensionnelles, transformant une tâche potentiellement lourde en une opération plus gérable et efficiente.

2.3.2.3 L'indépendance conditionnelle

Dans l'exemple de l'actuaire qui souhaite évaluer le risque de dommages causés par une tempête de neige en sachant que la région est à une altitude élevée et que les températures sont basses, que se passerait-il si l'actuaire recevait une information météorologique plus complète, comme l'observation imminente d'une tempête de neige exceptionnelle ? À première vue, l'altitude élevée et les basses températures semblent pertinentes pour estimer le risque de dommages. Cependant, si l'actuaire apprend qu'une tempête de neige exceptionnelle est imminente, les informations sur la température et l'altitude deviennent superflues, et le risque de dommages devient alors indépendant de ces variables dès que l'observation de la tempête est relevée. La probabilité de distribution jointe de l'actuaire devient :

$$P(D, A, T, S) = P(D \mid S) \cdot P(S) \cdot P(A) \cdot P(T) \quad \text{où} \quad S \text{ est l'événement "tempête de neige"}$$

Cela signifie que pour calculer la probabilité des dommages D, on peut se concentrer uniquement sur $P(D \mid S)$ et $P(S)$, car l'information sur la tempête rend les variables A et T indépendantes de D.

La structure des relations entre parents et enfants dans un réseau bayésien capture ces indépendances conditionnelles entre les variables. Sur la base de ces constatations, Pearl, Glymour, et Jewell (2016) proposent trois règles à suivre pour les identifier [11] :

- **Règle 1 (Indépendance conditionnelle dans les chaînes)** : Deux variables, X et Y, sont conditionnellement indépendantes étant donné Z, s'il n'existe qu'un seul chemin unidirectionnel entre X et Y, et si Z est un ensemble de variables qui intercepte ce chemin.
- **Règle 2 (Indépendance conditionnelle dans les fourches)** : Si une variable X est la cause commune des variables Y et Z, et qu'il n'existe qu'un seul chemin entre Y et Z, alors Y et Z sont indépendants conditionnellement à X.
- **Règle 3 (Indépendance conditionnelle dans les collisions)** : Si une variable Z est le nœud de collision entre deux variables X et Y, et qu'il n'existe qu'un seul chemin entre X et Y, alors X et Y sont indépendants inconditionnellement, mais dépendent conditionnellement de Z et de tout descendant de Z.

Dérivée de ces règles, la notion de *Markov Blanket* (ou couverture de Markov), un terme inventé par Pearl en 1988 [41], permet de limiter les calculs des CPTs aux seules variables directement pertinentes. Cette couverture représente l'unique ensemble d'informations nécessaire pour prédire le comportement d'un nœud, ainsi :

Définition 2. La *Markov Blanket* d'un nœud $A \in \mathcal{V}$ est le sous-ensemble minimal $S \subset \mathcal{V}$ tel que :

$$A \perp (\mathcal{V} - S - A) \mid S$$

où $\perp$ désigne l'indépendance conditionnelle.

Dans la figure 2.3.5, la *Markov Blanket* est représentée à l'aide de pointillés pour le nœud entouré d'un cercle épais. Le nœud colorié en jaune, X, correspond aux données observées. Ce sont les informations déterminantes pour la CPT de chaque nœud.

Pour un nœud θ, la Markov Blanket comprend :

— Les parents de θ,

— Les enfants de θ,

— Les co-parents de θ (les autres parents des enfants de θ).

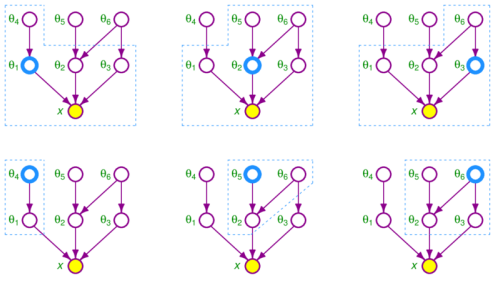


FIGURE 2.3.5 – Markov Blanket pour chacun des six paramètres θ dans l'exemple d'un graphe acyclique dirigé [25].

2.4 Conception et paramétrisation des réseaux bayésiens

2.4.1 Démarche suivie

Le protocole utilisé, de la conception à l'utilisation de notre modèle, s'inspire des bonnes pratiques de modélisation proposées par Jakeman et al. (2006) dans leur ouvrage *"Ten iterative steps in development and evaluation of environmental models"* [38]. Nous nous référons également au guide de Pollio et al. (2012) intitulé *"Good practice in Bayesian network modelling"* [39], ainsi qu'au rapport de la *Society Of Actuary "Climate Risk Assessment and Scenario Analysis : A Case Study on Climate Change Impact on Home Prices Utilizing Bayesian Network Qualitative to Quantitative Analysis"* [40]. L'implémentation de ces idées sera réalisée via le package *bnlearn* sur RStudio.

Voici les 10 étapes définies par Jakeman et al. (2006) qui seront suivies :

1. Définir l'objectif du modèle
2. Spécifier le contexte de modélisation (étendue et ressources)
3. Conceptualiser le système, spécifier les données et autres connaissances préalables
4. Sélectionner les caractéristiques et les familles de modèles
5. Décider comment déterminer la structure du modèle et les valeurs des paramètres
6. Sélectionner les critères et la technique de performance d'estimation
7. Identifier la structure et les paramètres du modèle
8. Vérification conditionnelle et tests de diagnostic
9. Quantifier l'incertitude
10. Évaluation et test du modèle

En suivant ces lignes directrices, il est important de comprendre comment ces étapes doivent être appliquées dans le contexte des BN. Les sections suivantes aborderont la conceptualisation du BN (étape 3), ses caractéristiques (étapes 1, 2, 4 et 5), sa paramétrisation (étapes 5 et 7), ainsi que son évaluation (étapes 6, 9 et 10) [39].

2.4.2 Conceptualisation

Avant de construire le BN, il est essentiel de synthétiser les connaissances existantes en un modèle conceptuel du système étudié, souvent représenté par un diagramme d'influence. Ce modèle doit être développé en tenant compte de l'objectif et de la portée du BN. Plusieurs modèles conceptuels peuvent être créés, par exemple, à différentes échelles, sous différentes perspectives ou avec divers niveaux de détail. L'objectif principal de cette étape est de fournir une vue d'ensemble visuelle des liens entre les différents moteurs (par exemple, le climat ou les politiques) [39] [40].

La première étape de cette synthétisation est cruciale : il s'agit de choisir une variable d'intérêt qui servira de nœud racine pour l'analyse et définira la direction de la conceptualisation [40].

Une méthode systématique pour élaborer ce modèle conceptuel est la construction d'un arbre, qui aide à organiser et à consolider les résultats de la recherche. Les étapes de cette méthode sont les suivantes [40] :

1. **Identification de la variable d'intérêt** : Cette variable devient le nœud racine de l'arbre et sert de point de départ pour l'ensemble du processus de conceptualisation.
2. **Analyse de la littérature existante** : Rechercher dans la littérature toutes les références à la variable d'intérêt. Les concepts ou variables pertinents identifiés dans cette littérature sont ajoutés comme nœuds enfants du nœud racine.
3. **Extension de l'arbre** : Répéter l'étape précédente pour chaque nœud enfant nouvellement ajouté, en continuant à explorer la littérature pour découvrir des concepts supplémentaires qui peuvent être liés. Ce processus se poursuit jusqu'à ce que la profondeur de l'arbre soit jugée satisfaisante.

Une fois l'arbre construit et les variables potentielles listées, la structure du BN peut être définie. Cela implique de placer chaque variable de l'arbre dans un graphe dirigé et acyclique via le *structure learning* [26] [40]. Cette étape nécessite d'abord de définir les caractéristiques des nœuds, en spécifiant la distribution de chaque variable (par exemple, discrète ou continue) et en déterminant les types de valeurs qu'elles peuvent prendre. Ensuite, il devient possible d'identifier les liens entre ces nœuds, c'est-à-dire quelles variables influencent directement d'autres variables, afin de structurer les relations causales dans le BN.

Une fois la structure du BN établie et les liens entre les nœuds définis, l'étape suivante consiste à paramétrer chaque nœud, ce qu'on appelle le *parameters learning*. Cela signifie attribuer à chaque nœud une distribution de probabilité conditionnelle, basée sur les variables parentales identifiées dans le graphe. Une fois les paramètres correctement estimés, le BN est prêt à être utilisé comme modèle prédictif [26].

2.4.3 Caractéristiques du modèle

Pour construire un BN, les deux éléments clés à prendre en compte dès le départ sont les suivants [39] :

— (i) ce que les nœuds doivent représenter ;

— (ii) l'état ou les valeurs qu'ils peuvent prendre.

En fonction de la nature des données qu'ils représentent, les nœuds peuvent prendre des valeurs discrètes ou continues. Les valeurs doivent être mutuellement exclusives et exhaustives, c'est-à-dire que chaque variable doit être dans un état unique à un moment donné. Les types de nœuds discrets couramment rencontrés dans la pratique incluent [40] :

— **Nœuds booléens** : prennent des valeurs binaires (vrai ou faux).

— **Nœuds classés** : peuvent prendre plusieurs états ordonnés (comme faible, moyen, élevé).

— **Nœuds continus** : prennent des valeurs d'une plage continue ou d'une distribution statistique.

Si une variable est continue et qu'elle suit une distribution statistique connue, comme une distribution normale (pour les tailles humaines), cette distribution est alors utilisée pour modéliser le nœud [40].

Cependant, il est parfois souhaitable de discrétiser une variable continue, c'est-à-dire de la convertir en une variable discrète prenant un nombre limité de valeurs ou d'états. La discrétisation simplifie le modèle, facilite les calculs, et rend l'interprétation des résultats plus intuitive. Par exemple, au lieu de traiter une plage continue de températures, on peut discrétiser cette variable en catégories telles que "basse", "moyenne" et "élevée" [26].

Historiquement, les nœuds étaient souvent discrétisés manuellement, ce qui simplifiait les calculs mais pouvait réduire la précision. Pour surmonter ces limitations, la discrétisation dynamique a été développée. Cette méthode moderne utilise des simulations pour créer automatiquement une représentation discrète qui capture fidèlement les valeurs continues originales dans une tolérance spécifiée [40] [43].

Une fois établi, en fonction des caractéristiques de l'ensemble des nœuds définis dans le BN, les réseaux bayésiens peuvent être classés en plusieurs types [42] :

— **BNs discrets** (DBNs), dans lesquels $\mathbf{X}$ et les $X_i \mid \Pi_{X_i}$ sont multinomiaux ;

— **BNs Gaussiens** (GBNs), dans lesquels $\mathbf{X}$ est multivariée normale et les $X_i \mid \Pi_{X_i}$ sont univariés normaux ;

— **BNs Gaussiens linéaires conditionnels** (CLGBNs), dans lesquels $\mathbf{X}$ est un mélange de normales multivariées et les $X_i \mid \Pi_{X_i}$ sont soit multinomiaux, univariés normaux, ou des mélanges de normales.

Les relations entre les nœuds sont quantifiées par des tables de probabilités conditionnelles (CPT). Ces tables spécifient la probabilité que chaque nœud prenne une valeur particulière, étant donné les valeurs des nœuds parents. Dans le cas des nœuds continus, il s'agit d'un ensemble de modèles de régression linéaire, un pour chaque configuration de parents discrets, les parents continus jouant le rôle de régresseurs [42].

Cette étape est très importante pour l'obtention d'un modèle exploitable, il est dans certains cas primordial de minimiser l'élicitation des probabilités [16].

2.4.4 Apprentissage de la structure

Une fois que les variables pertinentes capturant les aspects essentiels du problème à modéliser ont été sélectionnées, la première étape de la paramétrisation du modèle consiste à identifier la structure du BN [13]. Cette étape consiste à déterminer les relations de dépendance entre les différentes variables du réseau. La méthode la plus couramment utilisée repose sur les jugements d'experts, utilisant potentiellement des recherches antérieures ou de la littérature existante pour établir des relations probables entre les variables [14]. D'autres méthodes pour cette tâche incluent des approches basées sur des algorithmes de recherche de structure [14] [15], qui seront détaillées dans la prochaine section. Cette étape est cruciale pour obtenir un modèle exploitable, et dans certains cas, il est essentiel de minimiser l'élicitation des probabilités [16].

Pour rappel, la force des relations entre les nœuds d'un BN est quantifiée dans des (CPT) associées à chaque nœud. Pour les nœuds sans parents, on utilise des distributions de probabilité marginales. En revanche, pour chaque nœud enfant, des probabilités conditionnelles sont allouées pour chaque combinaison des états de leurs nœuds parents, comme le décrit la formule suivante [16] :

$$\text{taille(CPT)} = S \prod_{i=1}^{n} P_i \tag{2.4.1}$$

où S est le nombre d'états et P_i est le nombre d'états du i-ème nœud parent [18]. La taille des CPT peut ainsi augmenter de manière significative avec le nombre de parents, rendant le processus de peuplement des CPTs potentiellement intraitable. Pour alléger la complexité computationnelle des CPT, des bonnes pratiques et des outils algorithmiques peuvent être utilisés [18].

Tout d'abord, il est essentiel de reconsidérer la légitimité de chaque nœud sélectionné et des différents états qu'il peut prendre. Cain [17] conseille de ne pas surcharger le réseau d'informations superflues, en choisissant uniquement des états qui décrivent :

1. l'état actuel de la variable,
2. l'état potentiel sous des scénarios de changement,
3. seulement si nécessaire, tout état intermédiaire.

Une fois cette réflexion menée à bien, l'apprentissage de la structure du modèle peut commencer, avec plusieurs options offertes.

2.4.4.1 Structure basée sur la littérature et les experts

Lorsque la structure d'un réseau bayésien est définie par des experts, elle cherche généralement à refléter des relations causales déjà établies. L'article *Bayesian Networks in Environmental Risk Assessment : A Review* [44] examine l'utilisation et l'implémentation des réseaux bayésiens dans l'évaluation des risques environnementaux. Parmi les 72 études analysées, l'apprentissage de la structure basé sur la littérature est la méthode la plus couramment utilisée.

2.4.4.2 Structure basée sur la contrainte

Après avoir exploré l'approche basée sur la littérature et les experts, une autre méthode courante pour déterminer la structure d'un BN est basée sur des contraintes entre les variables. La paramétrisation par contraintes tire son nom du fait que les variables doivent satisfaire la contrainte suivante [45] :

$\Rightarrow\ P(C, M) \neq P(C)P(M)\ \Rightarrow\ C$ et M sont dépendants

Les approches fondées sur ce test d'indépendance s'inspirent des travaux pionniers de Pearl et de son algorithme d'induction causale (ICA) [46]. Cet algorithme fonctionne en trois étapes pour former la structure du DAG.

Algorithme 1 : Algorithme d'Induction Causale

Input : Ensemble de variables $\mathbf{X}$

Output : Graphe acyclique dirigé partiellement ou totalement

1. Pour chaque paire de variables A et B dans $\mathbf{X}$, rechercher un ensemble $S_{AB} \subseteq \mathbf{X}$ tel que A et B soient indépendants étant donné S_{AB} et $A, B \notin S_{AB}$. S'il n'existe pas un tel ensemble, placer une arête non orientée entre A et B.
2. Pour chaque paire de variables non adjacentes A et B avec un voisin commun C, vérifier si $C \in S_{AB}$. Si ce n'est pas le cas, orienter les arêtes $A - C$ et $C - B$ en $A \rightarrow C$ et $C \leftarrow B$.
3. Orienter les arêtes qui sont encore non orientées en appliquant récursivement les deux règles suivantes :
   (a) Si A est adjacent à B et qu'il existe un chemin strictement dirigé de A à B, alors orienter $A - B$ en $A \rightarrow B$.
   (b) Si A et B ne sont pas adjacents mais $A \rightarrow C$ et $C - B$, alors orienter ce dernier en $C \rightarrow B$.
4. Retourner le graphe acyclique dirigé (partiellement ou totalement).

Cependant, en raison du grand nombre de relations d'indépendance conditionnelle possibles, l'ICA présente des difficultés lors de l'application des deux premières étapes de l'algorithme dans des situations réelles. Cela a conduit au développement d'algorithmes alternatifs, tels que le *PC Algorithm*, *Grow-Shrink Algorithm*, *IAMB Algorithm*, *Fast-IAMB*, et *Inter-IAMB Algorithm*, qui tentent de surmonter ces difficultés. Une pratique commune à ces algorithmes, à l'exception du PC, est de commencer par identifier la couverture de Markov de chaque nœud du réseau, afin de réduire le nombre de tests d'indépendance conditionnelle effectués. Bien que ces solutions alternatives ne soient pas abordées dans ce travail, il est important de mentionner leur existence [26].

2.4.4.3 Structure basée sur le score

En utilisant des techniques d'optimisation heuristique associées à un calcul de score pour évaluer la qualité de l'ajustement de la structure, un meilleur candidat peut être trouvé par la maximisation de ce score. L'algorithme Hill-Climbing, largement utilisé dans ce contexte, cherche à améliorer le score en effectuant des modifications successives (ajout, suppression, ou inversion d'arêtes) jusqu'à ce qu'aucune amélioration supplémentaire ne soit possible [42].

Algorithme 2 : Algorithme de Hill-Climbing

Input : Ensemble de variables V

Output : Graphe acyclique dirigé

1. Choisir une structure de réseau G sur V, généralement (mais pas nécessairement) vide.
2. Calculer le score de G, noté $\text{Score}_G = \text{Score}(G)$.
3. Initialiser score_max $= \text{Score}_G$.
4. Répéter les étapes suivantes tant que score_max augmente :
   (a) Pour chaque ajout, suppression ou inversion d'arête possible ne résultant pas en un réseau cyclique :
      i. Calculer le score du réseau modifié G^*, $\text{Score}_{G^*} = \text{Score}(G^*)$.
      ii. Si $\text{Score}_{G^*} > \text{Score}_G$, définir $G = G^*$ et $\text{Score}_G = \text{Score}_{G^*}$.
   (b) Mettre à jour score_max avec la nouvelle valeur de Score_G.
5. Retourner le graphe acyclique dirigé G.

2.4.5 Apprentissage des paramètres

Une fois que la structure du DG est connue, l'estimation de la distribution jointe peut être résolue en estimant les paramètres des distributions locales, une à la fois, une tâche qui est facilitée par l'application des couvertures de Markov. Dans le cas de données complètes, il s'agit de maximiser l'estimateur de vraisemblance [42] [45] :

$$L(\Theta \mid \mathcal{D}) = \prod_{i=1}^{N} P(X_i = x_i \mid \text{Parents}(X_i), \Theta) \tag{2.4.2}$$

$$\hat{\Theta} = \arg\max_{\Theta} \log L(\Theta \mid \mathcal{D})$$

Dans le cas de données incomplètes, une approche couramment utilisée pour estimer les paramètres consiste à d'abord imputer les valeurs manquantes à l'aide de l'algorithme d'Expectation-Maximisation (EM), décrit ci-dessous.

Algorithme 3 : Algorithme d'Expectation-Maximisation (EM)

Input : Quantité statistique générique θ

1. Choisir une valeur initiale $\hat{\theta}_0$ pour θ.
2. Tant que $|\hat{\theta}_{j-1} - \hat{\theta}_j| < \epsilon$, augmenter j :
   (a) $\hat{\theta}_j = \hat{\theta}_{j-1}$
   (b) **Étape d'Expectation :** Calculer la distribution de probabilité sur les valeurs manquantes,

$$\mathrm{P}(X_i^{(M)} \mid X_i^{(O)}, \hat{\theta}_j) = \frac{\mathrm{P}(X_i^{(O)} \mid X_i^{(M)}, \hat{\theta}_j)\mathrm{P}(X_i^{(M)} \mid \hat{\theta}_j)}{\int_{X_i^{(M)}} \mathrm{P}(X_i^{(O)} \mid X_i^{(M)}, \hat{\theta}_j)\mathrm{P}(X_i^{(M)} \mid \hat{\theta}_j)}$$

   (c) **Étape de Maximisation :** Calculer la nouvelle estimation $\hat{\theta}_j$ étant donné

$$\mathrm{P}(X_i^{(M)} \mid X_i^{(O)}, \hat{\theta}_j)$$

3. Estimer θ avec la dernière valeur $\hat{\theta}_j$.

À noter que l'estimation des paramètres devient problématique, surtout lorsque la taille de l'échantillon n est bien inférieure au nombre de variables p dans le modèle. Dans ce contexte de "petit n, grand p", les estimations deviennent très variables [26].

2.5 Inférence bayésienne

Une suite logique à l'implémentation du réseau bayésien est de l'utiliser pour faire de l'inférence statistique et tirer des conclusions à partir de données observées, appelées *evidence*. Le réseau permet de répondre à deux types de requêtes principales : les *Conditional Probability Queries* (CPQ) et les *Maximum A Posteriori Queries* (MAP) [42].

2.5.1 Types de preuves : Hard et Soft Evidence

Les BN permettent d'incorporer différents types de preuves [26] :

— *Hard Evidence* : Correspond à des observations précises et déterminées, par exemple $\mathbf{E} = \{X_{i_1} = e_1, X_{i_2} = e_2, \ldots, X_{i_k} = e_k\}$.

— *Soft Evidence* : Représente une distribution probabiliste sur les valeurs possibles d'une variable, exprimée comme $\mathbf{E} = \{X_{i_1} \sim (\Theta_{X_{i_1}}), X_{i_2} \sim (\Theta_{X_{i_2}}), \ldots, X_{i_k} \sim (\Theta_{X_{i_k}})\}$.

2.5.2 Conditional Probability Queries (CPQ)

Les CPQ visent à déterminer la distribution d'un sous-ensemble de variables $\mathbf{Q} = \{X_{j_1}, \ldots, X_{j_i}\}$, en se basant sur des preuves $\mathbf{E} = \{X_{i_1} = e_1, X_{i_2} = e_2, \ldots, X_{i_k} = e_k\}$ observées sur un autre ensemble de variables $\mathbf{X}$. Ces requêtes calculent la probabilité conditionnelle marginale de $\mathbf{Q}$, notée [42] :

$$\mathrm{CPQ}(\mathbf{Q} \mid \mathbf{E}, \mathcal{B}) = P(\mathbf{Q} \mid \mathbf{E}, G, \Theta) = P(X_{j_1}, \ldots, X_{j_i} \mid \mathbf{E}, G, \Theta), \tag{2.5.1}$$

où $\mathcal{B} = \{G, \Theta\}$ désigne le réseau bayésien avec sa structure G et ses paramètres Θ.

Dans le cas d'une *hard evidence* $\mathbf{E}$, cette probabilité est la distribution marginale postérieure de $\mathbf{Q}$, calculée comme :

$$P(\mathbf{Q} \mid \mathbf{E}, G, \Theta) = \int P(\mathbf{X} \mid \mathbf{E}, G, \Theta)\, d(\mathbf{X} \setminus \mathbf{Q}). \tag{2.5.2}$$

2.5.3 Maximum A Posteriori Queries (MAP)

Les MAP se concentrent sur l'identification de la configuration la plus probable des variables $\mathbf{Q}$ en fonction des observations $\mathbf{E}$. Cette optimisation est obtenue en trouvant la valeur $\mathbf{q}^*$ qui maximise la probabilité postérieure [42] :

$$\text{MAP}(\mathbf{Q} \mid \mathbf{E}, \mathcal{B}) = \mathbf{q}^* = \arg\max_{\mathbf{q}} P(\mathbf{Q} = \mathbf{q} \mid \mathbf{E}, G, \Theta), \tag{2.5.3}$$

Dans le cas des GBNs et CLGBNs, il s'agit de maximiser la densité maximale a posteriori.

2.5.4 Belief Updating

Le *belief updating* consiste à intégrer de nouvelles preuves dans le réseau pour mettre à jour la distribution de probabilité postérieure des variables d'intérêt après l'apport des *evidence* [42]. Mathématiquement, cela se traduit par le calcul suivant :

$$P(\mathbf{Q} \mid \mathbf{E}, G, \Theta) = \int P(\mathbf{X} \mid \mathbf{E}, G, \Theta)\, d(\mathbf{X} \setminus \mathbf{Q}) \tag{2.5.4}$$

où $\mathbf{Q}$ représente les variables d'intérêt, $\mathbf{E}$ les preuves observées, G la structure du réseau, et Θ les paramètres. Ce calcul peut être reformulé et simplifié en utilisant des distributions locales pour chaque variable X_i :

$$P(\mathbf{Q} \mid \mathbf{E}, G, \Theta) = \prod_{i:X_i \in \mathbf{Q}} \int P(X_i \mid \mathbf{E}, \Pi_{X_i}, \Theta_{X_i})\, dX_i \tag{2.5.5}$$

où Π_{X_i} désigne les parents de X_i dans le réseau.

2.5.5 Algorithmes pour le Belief Updating

Les algorithmes de *belief updating* se classent en deux catégories : exacts et approximatifs. Les algorithmes exacts, tels que l'élimination de variables et les mises à jour basées sur les *junction trees 4*, utilisent des calculs locaux répétitifs pour obtenir une valeur précise de $P(\mathbf{Q} \mid \mathbf{E}, G, \Theta)$. En revanche, les algorithmes approximatifs, comme le *logic sampling 5* et le *likelihood weighting 6*, emploient des simulations de Monte Carlo pour estimer cette probabilité [42] [45].

Algorithme 4 : Junction Tree Clustering Algorithm

1. **Moraliser** : Créer le graphe moral du BN B.
2. **Trianguler** : Briser chaque cycle couvrant 4 nœuds ou plus en sous-cycles de 3 nœuds exactement.
3. **Cliques** : Identifier les cliques du graphe triangulé.
4. **Junction Tree** : Créer un arbre de jonction où chaque clique est un nœud, et les cliques adjacentes sont reliées par des arcs.
5. **Paramètres** : Utiliser les paramètres des distributions locales de B pour paramétrer les nœuds composés de l'arbre de jonction.

2.5.5.1 Logic Sampling Algorithm

Le *logic sampling* génère des échantillons en suivant l'ordre topologique des variables du BN, en commençant par les parents et avançant vers les enfants. Pour chaque échantillon, les valeurs des variables sont tirées selon les distributions conditionnelles du réseau. Si les échantillons incluent les *evidences*, ils sont comptabilisés pour estimer la probabilité conditionnelle recherchée, sinon ils sont éliminés [42].

Algorithme 5 : Logic Sampling Algorithm

1. Ordonner les variables dans $\mathbf{X}$ selon l'ordre partiel topologique induit par G.
2. Initialiser $n_E = 0$ et $n_{E,Q} = 0$.
3. Pour un nombre suffisant d'échantillons $\mathbf{x} = (x_1, \ldots, x_N)$:
   (a) Générer x_i pour $i = 1, \ldots, N$ à partir de $X_i \mid \Pi_{X_i}$.
   (b) Si $\mathbf{x}$ inclut $\mathbf{E}$, incrémenter n_E.
   (c) Si $\mathbf{x}$ inclut à la fois $\mathbf{Q}$ et $\mathbf{E}$, incrémenter $n_{E,Q}$.
4. Estimer $P(\mathbf{Q} \mid \mathbf{E}, G, \Theta)$ avec $\frac{n_{E,Q}}{n_E}$.

2.5.5.2 Likelihood Weighting Algorithm

L'algorithme de *likelihood weighting* est une amélioration du *logic sampling*. Il s'agit d'une forme d'échantillonnage d'importance où tous les échantillons générés incluent l'évidence $\mathbf{E}$ dès le départ. Il est plus efficace que le *logic sampling* car il réduit le nombre de rejets en garantissant que les échantillons sont conformes aux preuves observées [42].

Algorithme 6 : Likelihood Weighting Algorithm

1. Ordonner les variables dans $\mathbf{X}$ selon l'ordre partiel topologique induit par G.
2. Initialiser $w_E = 0$ et $w_{E,Q} = 0$.
3. Pour un nombre suffisant d'échantillons $\mathbf{x} = (x_1, \ldots, x_N)$:
   (a) Générer x_i pour $i = 1, \ldots, N$ à partir de $X_i \mid \Pi_{X_i}$ en utilisant les valeurs $e_1, \ldots, e_k$ spécifiées par l'évidence $\mathbf{E}$.
   (b) Calculer le poids $w_x = \prod P(X_i = e_i \mid \Pi_{X_i})$.
   (c) Mettre à jour $w_E = w_E + w_x$.
   (d) Si $\mathbf{x}$ inclut $\mathbf{Q} = q$ et $\mathbf{E}$, mettre à jour $w_{E,Q} = w_{E,Q} + w_x$.
4. Estimer $P(\mathbf{Q} \mid \mathbf{E}, G, \Theta)$ avec $\frac{w_{E,Q}}{w_E}$.

2.6 Outils statistiques

2.6.1 Distribution log-normale

La distribution log-normale est idéale pour modéliser les distributions qui sont positivement asymétriques. La fonction de densité de probabilité d'une distribution log-normale est donnée par :

$$f(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \tag{2.6.1}$$

où μ et $\sigma > 0$ sont respectivement la moyenne et l'écart-type de la distribution normale sous-jacente.

Pour transformer une variable X qui suit une distribution log-normale en une variable Y qui suit une distribution normale, on utilise la transformation logarithmique suivante :

$$Y = \log(X) \tag{2.6.2}$$

Ainsi, si $X \sim \text{Log-Normal}(\mu, \sigma^2)$, alors $Y = \log(X)$ suit $\mathcal{N}(\mu, \sigma^2)$.

2.6.2 Test de normalité

Le test de Shapiro-Wilk est utilisé pour vérifier si un ensemble de données suit une distribution normale en calculant une statistique W, qui compare les données observées à une distribution normale théorique [7].

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.6.3}$$

où $x_{(i)}$ sont les ordres statistiques des données, et a_i sont des coefficients constants dépendant de la taille de l'échantillon.

2.6.3 Test d'indépendance conditionnelle

2.6.3.1 DBNs

Les tests d'indépendance conditionnelle utilisés pour apprendre les DBNs reposent sur les fréquences observées $\{n_{ijk}\}$, où $i = 1, \ldots, R$, $j = 1, \ldots, C$, et $k = 1, \ldots, L$ représentent respectivement les différentes catégories des variables X, Y, et des configurations des variables de conditionnement Z. Parmi les méthodes classiques pour tester cette indépendance, on retrouve [26] :

- Ratio de vraisemblance : Cette mesure évalue l'information partagée entre deux variables X et Y, conditionnellement à un ensemble de variables Z. Elle quantifie combien savoir Z réduit l'incertitude sur la relation entre X et Y :

$$\text{MI}(X, Y \mid Z) = \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{k=1}^{L} \frac{n_{ijk}}{n} \log\left(\frac{n_{ijk} \cdot n_{++k}}{n_{i+k} \cdot n_{+jk}}\right) \tag{2.6.4}$$

  où n_{ijk} est la fréquence observée pour chaque combinaison des variables X, Y et Z.

- Le χ^2 de Pearson : Permet de vérifier l'hypothèse nulle selon laquelle les variables X et Y sont indépendantes, conditionnellement à Z. Il compare les fréquences observées avec les fréquences attendues pour calculer une statistique de test.

$$\chi^2(X, Y \mid Z) = \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{k=1}^{L} \frac{(n_{ijk} - m_{ijk})^2}{m_{ijk}}, \quad \text{où} \quad m_{ijk} = \frac{n_{i+k} \cdot n_{+jk}}{n_{++k}} \tag{2.6.5}$$

  où m_{ijk} est la fréquence attendue sous l'hypothèse d'indépendance conditionnelle.

2.6.3.2 GBNs

Les tests d'indépendance conditionnelle utilisés pour apprendre les GBNs sont basés sur les corrélations partielles $\rho_{XY|Z}$. Parmi les options couramment utilisées, on retrouve [26] [42] :

- Le test exact t pour le coefficient de corrélation de Pearson, défini comme :

$$t(X, Y \mid Z) = \rho_{XY|Z} \sqrt{\frac{n - |Z| - 2}{1 - \rho^2_{XY|Z}}} \tag{2.6.6}$$

  Ce test suit une distribution t de Student à $n - |Z| - 2$ degrés de liberté.

- Le test Z de Fisher, une transformation de $\rho_{XY|Z}$ avec une distribution normale asymptotique, défini comme :

$$Z(X, Y \mid Z) = \log\left(\frac{1 + \rho_{XY|Z}}{1 - \rho_{XY|Z}}\right) \sqrt{\frac{n - |Z| - 3}{2}} \tag{2.6.7}$$

  où n est le nombre d'observations et $|Z|$ représente le nombre de nœuds appartenant à Z.

2.6.3.3 CLGBNs

Dans le cas des CLGBNs, lorsque X est une variable catégorielle et Y suit une distribution gaussienne (ou inversement), le test le plus approprié est celui de l'information mutuelle [42] :

$$\propto \log \frac{\mathbb{P}(Y \mid X, Z)}{\mathbb{P}(Y \mid Z)} \tag{2.6.8}$$

où le numérateur et le dénominateur représentent des régressions linéaires.

Si X et Y sont tous deux catégoriels, et que $Z = \{Z_{c_1}, \ldots, Z_{c_l}, Z_{d_1}, \ldots, Z_{d_m}\}$ contient à la fois des variables catégorielles et gaussiennes, alors par plusieurs applications du théorème de Bayes et de la règle de chaîne, nous obtenons :

$$\begin{aligned} \frac{\mathbb{P}(X \mid Z_{d_1:d_m}, Z_{c_1:c_l})}{\mathbb{P}(X \mid Y, Z_{d_1:d_m}, Z_{c_1:c_l})} = & \frac{\prod_{i=1}^{l-1} \mathbb{P}(Z_{c_i} \mid Z_{c_{i+1}:c_l}, X, Z_{d_1:d_m}) \mathbb{P}(X, Z_{d_1:d_m})}{\prod_{i=1}^{l-1} \mathbb{P}(Z_{c_i} \mid Z_{c_{i+1}:c_l}, Z_{d_1:d_m}) \mathbb{P}(Z_{d_1:d_m})} \times \\ & \frac{\prod_{i=1}^{l-1} \mathbb{P}(Z_{c_i} \mid Z_{c_{i+1}:c_l}, X, Y, Z_{d_1:d_m}) \mathbb{P}(X, Y, Z_{d_1:d_m})}{\prod_{i=1}^{l-1} \mathbb{P}(Z_{c_i} \mid Z_{c_{i+1}:c_l}, Y, Z_{d_1:d_m}) \mathbb{P}(Y, Z_{d_1:d_m})} \end{aligned} \tag{2.6.9}$$

Qui est une chaîne de rapports de vraisemblance logarithmiques, qui peut également être traitée comme un test d'information mutuelle.

2.6.4 Score d'un réseau

2.6.4.1 DBNs

Le score d'un réseau bayésien dynamique (DBN) est souvent calculé en utilisant un score décomposable, qui est une somme des scores locaux pour chaque nœud X_i en fonction de ses parents Π_{X_i} [42] :

$$\text{Score}(G) = \sum_{i=1}^{N} \text{Score}(X_i \mid \Pi_{X_i}) \tag{2.6.10}$$

Dans la littérature deux méthodes de scoring sont couramment citées [26] [42] :

- **Critère d'Information Bayésien (BIC) :**

$$\text{BIC}(G) = \sum_{i=1}^{N} \log \text{P}(X_i \mid \Pi_{X_i}) - \frac{|\Theta_{X_i}|}{2} \log n \tag{2.6.11}$$

  où $|\Theta_{X_i}|$ représente le nombre de paramètres pour X_i, et n est la taille de l'échantillon.

- **Équivalent Bayésien de Dirichlet (BDe) [42] :**

$$\text{BDe}(G) = \sum_{i=1}^{N} \log \left[\int \text{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i}) \text{P}(\Theta_{X_i} \mid \Pi_{X_i}) d\Theta_{X_i} \right] \tag{2.6.12}$$

  Ce score intègre les paramètres Θ_{X_i} en supposant une distribution a priori uniforme sur les structures du réseau et les paramètres.

2.6.4.2 Vraisemblances pénalisées

Les vraisemblances pénalisées sont très populaires tout autant pour les CLGBNs, que les DBNs et les GBNs. Pour les CLGBNs, la log-vraissemblance et le nombre de paramètres associés à une distribution locale (Δ_{X_i} étant les parents discrets, Γ_{X_i} les parents continus) sont [42] :

$$\text{LL}(X_i, \Pi_{X_i}) = \prod_{m=1}^{n} N(x_m; \mu_{X_i,\delta_m} + \gamma_m \beta_{X_i,\delta_m}, \sigma^2_{X_i,\delta_m}), \quad |\Theta_{X_i}| = |\Delta_{X_i}| \times (|\Gamma_{X_i}| + 1)| \tag{2.6.13}$$

Pour les DBNs :

$$\text{LL}(X_i, \Pi_{X_i}) = \prod_{m=1}^{n} \text{P}(X_i = x_m \mid \Pi_{X_i} = \pi_m), \quad |\Theta_{X_i}| = R \times |\Pi_{X_i}| \tag{2.6.14}$$

Pour les GBNs :

$$\text{LL}(X_i, \Pi_{X_i}) = \prod_{m=1}^{n} N(x_m; \mu_{X_i} + \pi_m \beta_{X_i}, \sigma^2_{X_i}), \quad |\Theta_{X_i}| = |\Pi_{X_i}| + 1 \tag{2.6.15}$$

2.6.5 Discrétisation

La discrétisation est un procédé utilisé pour transformer une variable continue $X = \{x_1, x_2, \ldots, x_n\}$ en une variable discrète. Est abordé ici trois méthodes : la discrétisation par intervalles égaux, par quantiles, et la méthode de Hartemink très populaire dans les BNs.

2.6.5.1 Discrétisation par intervalles égaux

Cette méthode divise la plage de valeurs de X en k intervalles de largeur égale. La largeur de chaque intervalle est donnée par [52] :

$$\delta = \frac{x_{\max} - x_{\min}}{k} \tag{2.6.16}$$

et construit les bornes des intervalles, ou seuils, à $x_{\min} + i\delta$ où $i = 1, \ldots, k-1$.

2.6.5.2 Discrétisation par quantiles

Cette méthode divise les données en k groupes contenant chacun un nombre égal d'observations. Les quantiles q_j sont définis par :

$$q_j = F^{-1}\left(\frac{j}{k}\right), \quad j = 1, 2, \ldots, k-1 \tag{2.6.17}$$

où F^{-1} est la fonction inverse de la distribution cumulative. Les intervalles Q_j sont alors définis comme suit :

$$Q_j = [q_{j-1}, q_j), \quad j = 1, 2, \ldots, k \tag{2.6.18}$$

avec $q_0 = \min(X)$ et $q_k = \max(X)$. Chaque valeur $x_i \in X$ est assignée au quantile Q_j correspondant.

2.6.5.3 Méthode de Hartemink

La méthode Hartemink a pour objectif de discrétiser en minimisant la perte d'information grâce à une méthode de maximisation de la corrélation entre les variables discrétisées et une variable cible Y. Elle commence par évaluer l'entropie conditionnelle $I(X;Y)$ pour différentes partitions de X. La discrétisation optimale est celle qui minimise cette entropie [51] :

$$I(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{2.6.19}$$

2.6.6 Validation croisée

La validation croisée est l'une des techniques les plus simples et les plus populaires pour évaluer la performance des modèles statistiques et sélectionner les valeurs optimales des hyperparamètres. Cette méthode consiste à diviser aléatoirement les données en k sous-ensembles égaux, appelés folds. Le modèle est ensuite entraîné sur $k-1$ de ces sous-ensembles, tandis que le sous-ensemble restant est utilisé pour tester le modèle. Ce processus est répété k fois, chaque sous-ensemble jouant successivement le rôle du jeu de test. Les performances obtenues à chaque itération sont ensuite moyennées pour fournir une estimation globale de la performance du modèle. La figure 2.6.1 illustre ce processus.

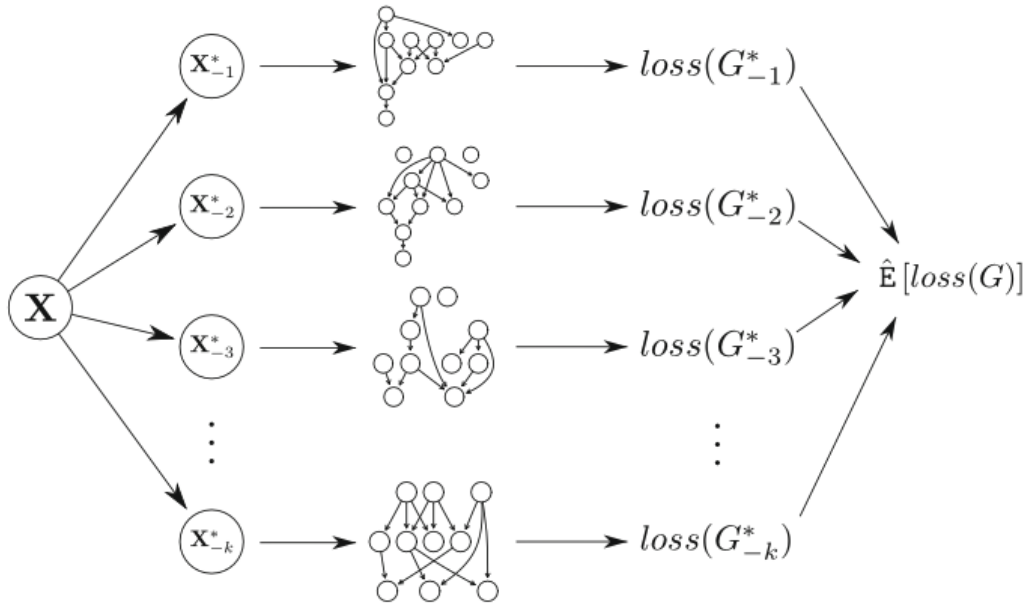


FIGURE 2.6.1 – Estimation par validation croisée k-fold d'une fonction de perte pour un algorithme d'apprentissage de réseau bayésien [26].

Deux fonctions de perte sont principalement utilisées pour évaluer la performance des modèles :

— **Le maximum de vraisemblance** : Il permet d'évaluer la qualité globale de l'ajustement du réseau en mesurant la probabilité que le modèle ait généré les données observées.

— **La corrélation prédictive** : Elle mesure la force de la relation entre les prédictions du modèle et les observations réelles d'une variable cible, offrant ainsi une évaluation de la précision des prédictions.

2.6.7 Extreme Value Theory (EVT)

2.6.7.1 Les deux grandes méthodes de l'EVT

L'EVT est une branche des statistiques qui se concentre sur la modélisation des comportements extrêmes, c'est-à-dire les queues de distribution des données. Elle est particulièrement utile pour estimer les risques associés à des événements rares et extrêmes, comme les catastrophes naturelles [23].

L'EVT propose principalement deux approches pour analyser les valeurs extrêmes [23] :

— **Block Maxima Approach** : Cette méthode consiste à diviser les données en blocs de même taille et à ne conserver que le maximum de chaque bloc. Bien que simple à mettre en œuvre, cette méthode peut entraîner une perte d'information en négligeant des valeurs non maximales importantes.

— **Peaks-Over-Threshold (POT)** : Dans cette approche, on s'intéresse à tous les points qui dépassent un certain seuil. Elle permet de mieux utiliser l'information contenue dans les valeurs extrêmes en modélisant directement les excès au-dessus du seuil sélectionné.

2.6.7.2 Peaks-Over-Threshold (POT)

Soit X une variable aléatoire représentant les pertes, et μ le seuil choisi. La variable des excès est définie comme $Y = X - \mu$ avec $Y > 0$. La fonction de répartition des excès au-delà de μ, notée $F_\mu(y)$, est donnée par :

$$F_\mu(y) = P(Y \leq y \mid Y > 0) = \frac{F(\mu + y) - F(\mu)}{1 - F(\mu)}$$

2.6.7.3 Distribution Pareto Généralisée (GPD)

La GPD est la distribution utilisée pour modéliser les excès au-delà du seuil μ. Elle est définie par la fonction de répartition suivante [48] :

$$G_\xi(x) = \begin{cases} 1 - \left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{x-\mu}{\sigma}\right), & \text{si } \xi = 0 \end{cases}$$

où ξ est le paramètre de forme et σ le paramètre d'échelle.

2.6.7.4 Estimateur de Hill

L'estimateur de Hill est un outil utilisé pour sélectionner le seuil en identifiant la stabilité du paramètre de forme ξ. Pour une variable aléatoire $X_1, \ldots, X_n$ ordonnée, l'estimateur de Hill est défini comme [49] :

$$\xi_{k,n}^{\text{Hill}} = \frac{1}{k} \sum_{i=1}^{k} \log\left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right)$$

où k est le nombre de statistiques d'ordre utilisées dans le calcul. Le Hill's plot est un outil graphique pour choisir k en observant la stabilité de $\xi_{k,n}^{\text{Hill}}$ en fonction de k.

2.6.7.5 Value at Risk

La Value at Risk permet d'estimer la perte maximale que l'on s'attend à ne pas dépasser avec une probabilité de p sur une période donnée. En supposant que les excès suivent une GPD, la VaR à p est calculée comme [48] :

$$\widehat{\text{VaR}}_p = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(\frac{n}{N_u} p \right)^{-\hat{\xi}} - 1 \right] \tag{2.6.20}$$

— u : Le seuil à partir duquel les dépassements sont modélisés par la distribution généralisée de Pareto (GPD).

— $\hat{\sigma}$: L'estimation du paramètre d'échelle de la GPD.

— $\hat{\xi}$: L'estimation du paramètre de forme de la GPD.

— n : Le nombre total d'observations dans l'échantillon.

— N_u : Le nombre d'observations au-dessus du seuil u.

2.6.7.6 Expected Shortfall

La formule de l'Expected Shortfall est définie comme la moyenne des pertes qui dépassent la VaR au niveau de confiance p. En supposant que les excès suivent une GPD, l'Expected Shortfall est calculé comme suit [48] :

$$\widehat{ES}_p = \frac{\widehat{\mathrm{VaR}}_p}{1-\hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi}u}{1-\hat{\xi}} \tag{2.6.21}$$

Chapitre 3

Application

3.1 Sources et présentation des données

La base de données utilisée dans cette étude provient de plusieurs sources accessibles en ligne, couvrant divers aspects pertinents pour cette recherche. Les principales sources incluent *The International Disaster Database (EM-DAT)*, *International Energy Agency (IEA)*, *Goddard Institute for Space Studies (GISS)*, et *United Nations University Institute for Environment and Human Security (UNU-EHS)*.

— *The International Disaster Database*

EM-DAT fournit des informations sur plus de 26 000 catastrophes mondiales depuis 1900, bien que les données antérieures à l'an 2000 puissent manquer de précision. Cette base de données, compilée à partir de multiples sources, est distribuée par le Centre de Recherche sur l'Épidémiologie des Désastres (*CRED*) de l'Université catholique de Louvain et est disponible en accès libre pour tout usage non-commercial [27]. De cette base de données ont été extraites des informations historiques sur les catastrophes climatiques survenues en Europe (type et lieu) ainsi que le montant des dégâts ajustés. [1]

— *International Energy Agency*

L'*IEA*, basée à Paris, est une organisation qui fournit des analyses et des données sur le secteur énergétique mondial [30]. Les rapports de l'IEA ont fourni des données sur les émissions de CO_2 (exprimées en gigatonnes (Gt)) provenant de la production d'énergie et des processus industriels. [2]

— *Goddard Institute for Space Studies*

Le *GISS*, basé aux États-Unis, est un centre de recherche spécialisé dans l'étude de l'atmosphère terrestre. L'institut met à disposition des données variées sur le réchauffement climatique [29]. Cette étude utilise les données relatives au changement de la température de la surface du globe par rapport à la moyenne à long terme de 1951 à 1980. [3]

1. La base de données *EM-DAT* peut être consultée à l'adresse suivante : `https://doc.emdat.be/docs/introduction/`.

2. Les données de l'*IEA* peuvent être consultées à l'adresse suivante : `https://www.iea.org/data-and-statistics/charts/total-increase-in-energy-related-co2-emissions-1900-2023`.

3. Les données sur les températures globales peuvent être consultées à l'adresse suivante : `https://climate.nasa.gov/vital-signs/global-temperature/?intent=121`.

— *United Nations University Institute for Environment and Human Security (UNU-EHS)*

Le *WorldRiskIndex* propose plusieurs indices pour évaluer le risque de catastrophe d'un pays causé par des événements naturels extrêmes et les effets négatifs du changement climatique. Développé par l'*United Nations University Institute for Environment and Human Security (UNU-EHS)* [28], parmi les indices que l'institut propose, ont été extraits des informations sur l'exposition et la vulnérabilité pour tous les pays présents dans l'historique européen d'EM-DAT. [4]

Une fois rassemblée et compilée, la base de données (aperçu en annexe 5) se compose de 605 observations contenant les éléments suivants :

- **Historique climatique** : Informations sur les émissions de CO_2 et l'augmentation moyenne des températures globales par rapport à la moyenne à long terme de 1951 à 1980.
- **Historique des périls et des dégâts associés** : Inclut les types de périls tels que la grêle, les inondations, les tempêtes, et les mouvements de terrain, conformément au sous-module *Cat Nat (non-life)* de *Solvency II*. Les données comprennent l'année de survenance, le montant des dégâts ajustés pour chaque catastrophe, ainsi que le pays affecté (indiqué par son code ISO).
- **Données sur la vulnérabilité et l'exposition** : Informations issues des indices sur la vulnérabilité et l'exposition pour chaque pays présent dans l'historique des catastrophes climatiques.

3.1.1 Préparation des données

3.1.1.1 Horizon de temps

Conformément aux consignes de l'*EM-DAT*, qui avertit que les données collectées avant l'an 2000 sont sujettes à des imprécisions [31], les données antérieures à l'année 2000 ont été exclues de l'analyse afin d'éviter l'utilisation de données potentiellement erronées et de garantir une meilleure fiabilité des résultats.

3.1.1.2 Variables discrètes

Les différents types d'inondation (fluviale, côtière, etc.) ont été regroupés en une seule catégorie générale "Flood (General)", ce qui aboutit à quatre catégories principales de catastrophes naturelles : {Ground movement, Flood (General), Hail, Storm (General)}. Les codes ISO des pays ont également été utilisés pour identifier 37 pays différents. Certains pays, comme l'Italie et la France, sont beaucoup plus représentés, tandis que d'autres, tels que le Luxembourg ou la Suède, n'apparaissent qu'une seule fois.

3.1.1.3 Variables continues et discrétisation

Pour réduire la complexité de notre modèle, éviter les singularités, et permettre de relâcher les hypothèses de linéarité, il est intéressant de discrétiser certaines variables continues. Cependant, les variables *Temperature*, $GtCO_2$, et *Damage* ne seront pas discrétisées. Ce choix

4. Les données du *WorldRiskIndex* peuvent être consultées à l'adresse suivante : `https://data.humdata.org/dataset/worldriskindex?`.

est délibéré et vise à conserver une plus grande flexibilité dans les inférences que le modèle pourra réaliser.

Très utilisée dans les BN pour préserver les dépendances entre les variables, la méthode de Hartemink, présentée en section 2.6.5, sera employée pour discrétiser *Vulnerability* et *Exposure*, en les catégorisant en cinq niveaux : {Très faible, Faible, Moyen, Élevé, Très élevé}.

3.1.1.4 Valeurs manquantes

Il existe trois types de données manquantes [47] :

— **Données manquantes complètement aléatoires (MCAR)** : Les données sont absentes indépendamment des valeurs, qu'elles soient observées ou manquantes, formant ainsi un sous-ensemble aléatoire.

— **Données manquantes aléatoirement (MAR)** : La probabilité de données manquantes est liée aux données observées, mais pas aux données manquantes elles-mêmes.

— **Données manquantes non aléatoires (MNAR)** : L'absence de données dépend des valeurs manquantes elles-mêmes, créant ainsi un biais significatif.

Les données manquantes de type MNAR sont considérées comme non-ignorables [42], car il est nécessaire de modéliser le mécanisme qui a conduit à leur absence pour les traiter adéquatement. En revanche, les données manquantes de type MCAR et MAR sont souvent considérées comme ignorables.

Dans le contexte des BNs, chaque variable X_i suit une distribution locale $X_i \sim P(X_i \mid \Pi_{X_i})$ si les données sont complètes. Cependant, si X_i contient des données manquantes [42] :

— Dans le cas MCAR :

$$X_i \sim \begin{cases} P(X_i \mid \Pi_{X_i}) & \text{pour les données observées } X_i^{(O)} \\ P(X_i \mid \Pi_{X_i}) & \text{pour les données manquantes } X_i^{(M)} \end{cases}$$

— Dans le cas MAR, la même distribution s'applique puisque le manque de données dépend des Π_{X_i}.

— En revanche, dans le cas MNAR :

$$X_i \sim \begin{cases} P(X_i^{(O)} \mid \Pi_{X_i}, M) & \text{pour les données observées } X_i^{(O)} \\ P(X_i^{(M)} \mid \Pi_{X_i}, M) & \text{pour les données manquantes } X_i^{(M)} \end{cases}$$

où M représente le mécanisme des données manquantes. M est non-ignorable car il est impossible d'estimer correctement la distribution locale de X_i sans comprendre le mécanisme qui cause ces données manquantes.

Malheureusement, dans la base de données utilisée, certaines informations liées aux dégâts sont absentes. À ce sujet, l'EM-DAT avertit [27] :

> "Les dommages économiques résultant des catastrophes sont largement sous-déclarés. Dans la base de données EM-DAT, les chiffres ne sont disponibles que pour les catastrophes à fort impact dans des pays bénéficiant d'une couverture d'assurance et de réassurance."

Cela correspond à un cas de MNAR, et il est crucial de tenir compte de ce biais. Ces situations sont particulièrement complexes à gérer, car les raisons pour lesquelles les variables sont absentes dépendent des variables elles-mêmes ou d'autres variables non observées. Dans ce travail, deux approches sont envisagées et comparées pour traiter ces données manquantes : l'imputation à l'aide de l'algorithme *Expectation-Maximisation* ou l'omission des données manquantes en appliquant directement la maximisation de la vraisemblance.

3.1.1.5 Hypothèses

Après avoir effectué ces ajustements, notre modèle suit un *CLGBNs*, modélisant une combinaison de variables discrètes et continues. Dans ce contexte, les hypothèses suivantes doivent être considérées :

1. La distribution de chaque variable continue, conditionnée sur ses parents, doit suivre une distribution gaussienne.
2. Les dépendances entre les variables sont linéaires.

Une étude des résidus de la variable *Damage* sera réalisée a posteriori pour vérifier ces hypothèses. Il est cependant déjà possible de jeter un œil à cette variable en examinant sa distribution :

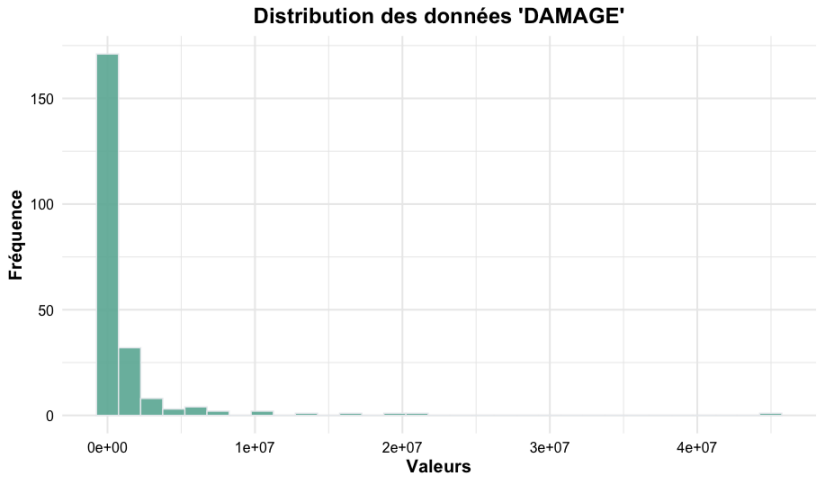


FIGURE 3.1.1 – Distribution de la variable "DAMAGE" présentée sous forme d'histogramme

L'histogramme montre une distribution asymétrique avec une forte concentration de valeurs faibles et une longue queue à droite. Cette distribution n'est pas représentative d'une gaussienne, qui est symétrique, mais elle est typique d'une distribution selon la loi log-normale ou la loi de Pareto. Un artifice présenté en section 2.6.1 est d'appliquer une transformation logarithmique à nos données, puis de vérifier si elles suivent une loi normale à l'aide du test de Shapiro-Wilk.
Les résultats du test, présentés dans le tableau 3.1.1, ainsi que l'analyse des résidus dans le graphique 3.1.2, permettent raisonnablement de penser qu'ils sont normalement distribués, bien que de légères déviations soient observables dans les quantiles extrêmes.

Le coefficient de Pearson entre nos deux variables linéaires, Damage et Temperature, est de 0,1272302, ce qui indique une relation linéaire très faible.

	Statistique W	Valeur-p
Shapiro-Wilk	0.99231	0.2833

TABLE 3.1.1 – Résultats du test de normalité de Shapiro-Wilk

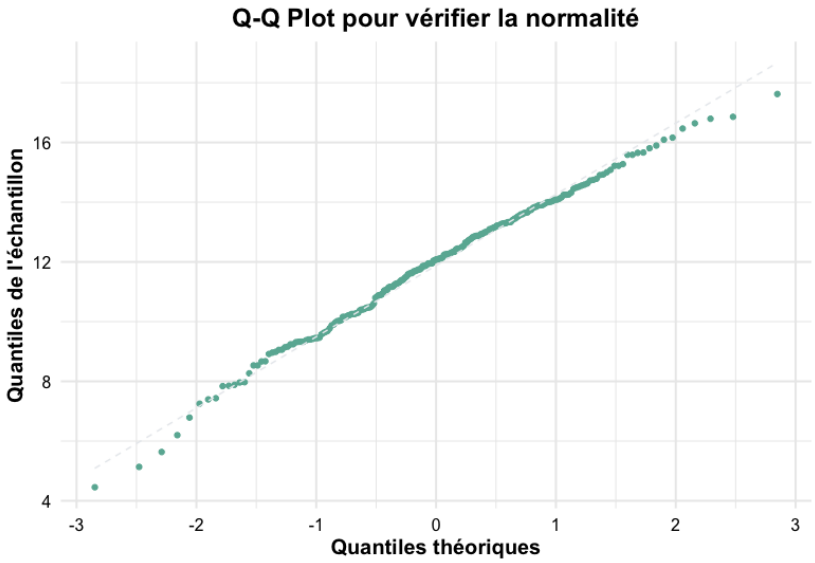


FIGURE 3.1.2 – QQ plot de *DAMAGE* après transformation logarithmique

3.2 Apprentissage de la structure

3.2.1 Par la littérature et l'expertise

De manière générale, la littérature décrit précisément les relations causales entre les différents effets météorologiques et climatiques. Dans notre cadre de travail, la science et les principes logiques offrent un fondement solide pour modéliser ces interactions, permettant de construire un réseau cohérent 3.2.1 sans recourir à des algorithmes, mais en s'appuyant sur les réflexions suivantes :

— **Exposure** → **Damage** : Plus une région ou une population est exposée à un événement climatique, plus les dommages potentiels sont élevés.

— **Vulnerability** → **Damage** : La vulnérabilité d'une population ou d'une infrastructure détermine également l'ampleur des dégâts.

— **GtCO2** → **Temperature** : Les émissions de CO_2 influencent directement la température globale, un lien bien établi maintes fois dans la littérature.

— **Temperature** → **Damage** : L'augmentation des températures peut accroître la fréquence et l'intensité des événements climatiques extrêmes.

— **Disaster_Subtype** → **Damage** : Le type de catastrophe (inondation, tempête, etc.) affecte directement les dommages, car chaque type d'événement a des caractéristiques et des impacts spécifiques.

— **ISO** → **Exposure, ISO** → **Vulnerability** : Le code ISO des pays est utilisé pour

capturer les caractéristiques spécifiques à chaque pays en termes d'exposition et de vulnérabilité.

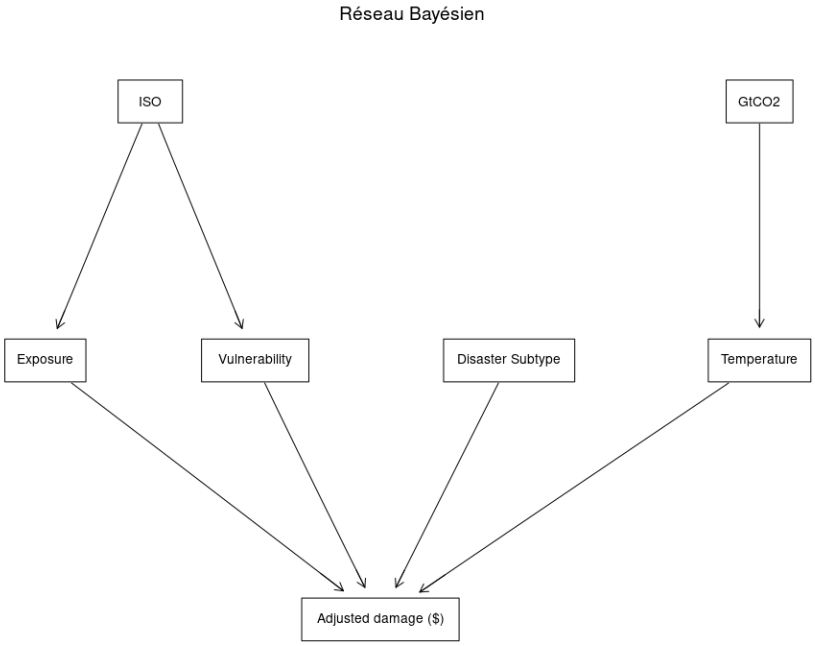


FIGURE 3.2.1 – Réseau Bayésien construit à partir de la littérature et de l'expertise

3.2.2 Par maximisation du score

Vu la petite taille du réseau, il est tout à fait envisageable de *brute-force*[5] le BN doté du meilleur score (voir annexe 5) avec l'algorithme Hill-Climbing 3.

Bien que cette version du réseau montre quelques similarités avec le précédent en capturant des relations causales établies (comme $GtCO_2$ et Temperature), elle diffère sur certains points avec une certaine irrationalité apparente, par exemple la variable *Vulnérabilité* qui influence le type de désastre qui survient.

3.2.3 Par contrainte

L'algorithme ICA 1 peut également se montrer computationnellement accessible dans ce genre de scénario et nous offre sa version du BN. (voir annexe 5)
Cependant l'ICA n'a trouvé qu'une seule relation de dépendance entre les pays, leur vulnérabilité et leur exposition respectives, et suggère que *Disaster_Subtype*, *Adjusted Damage ($)*, $GtCO_2$, et *Temperature* sont indépendants du reste de la structure.

5. Consiste à tester chaque combinaison possible, une par une.

3.2.4 Choix

Pour garantir que le réseau bayésien repose sur des relations causales fondées sur des connaissances validées scientifiquement, le modèle construit à partir de la littérature sera considéré pour la suite.

3.3 Paramétrisation et validation des modèles

Deux méthodes d'ajustement sont considérées pour le CLGBN : la maximisation de la log-vraisemblance, qui ignorera les lignes de données contenant des valeurs manquantes, et l'*Expectation Maximization*, qui cherchera d'abord à imputer des valeurs aux données manquantes. Des tests d'évaluation ont été effectués à l'aide d'une validation croisée. Deux fonctions de perte différentes ont également été utilisées : la corrélation prédictive a posteriori, qui permet de juger du degré de prédiction d'une variable cible (ici la variable "Adjusted Damage ($)", la variable d'intérêt), et la log-vraisemblance négative afin d'évaluer l'ajustement global du BN.
Une évaluation comparative de ces deux méthodes, à l'aide de ces deux fonctions de perte, a été réalisée sur plusieurs structures alternatives du réseau :

- Un modèle complet
- Un modèle sans *ISO*
- Un modèle sans *exposure*
- Un modèle sans *vulnérability*
- Un modèle sans *exposure* ni *vulnérability* ni *ISO*
- Un modèle sans *Disaster_Subtype*
- Un modèle uniquement avec la température

Pour chacune de ces structures, différents niveaux de catégorisation ont été appliqués à *vulnerability* et *exposure* :

- Une catégorisation inchangée à 5 niveaux
- Une catégorisation simplifiée à 2 niveaux
- Pas de catégorisation

Un total de 84 combinaisons a été testé pour couvrir toutes les possibilités. Les résultats sont consultables dans les figures 3.3.1 et 3.3.2 ci-dessous.

Dans la majorité des scénarios utilisant la fonction de perte de la log-vraisemblance négative, la méthode a rencontré des difficultés dues à des matrices singulières, empêchant l'algorithme de converger correctement. Cette singularité est souvent le signe que le modèle est trop complexe ou mal spécifié pour les données disponibles.

De manière générale, il apparaît que le modèle préfère des structures plus simples et moins larges. Les scénarios où des variables ont été retirées ou où la complexité du modèle a été réduite affichent de meilleurs scores. Ces résultats suggèrent que le modèle de base comportant toutes les variables est surdimensionné.

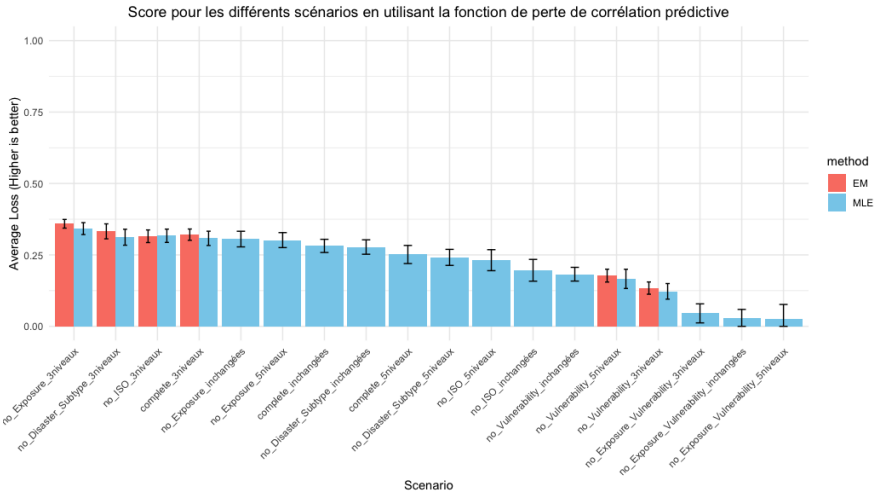


FIGURE 3.3.1 – Scores pour les différents scénarios utilisant la fonction de perte de corrélation prédictive

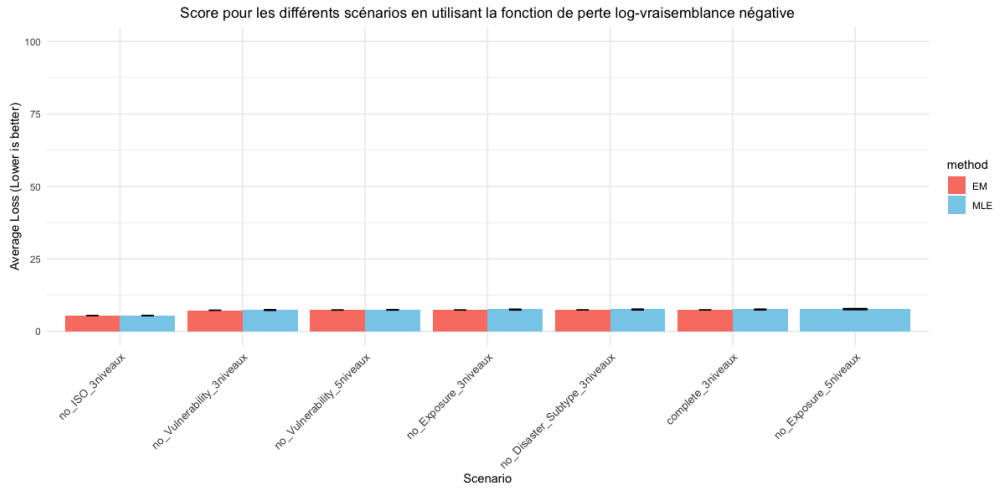


FIGURE 3.3.2 – Scores pour les différents scénarios utilisant la fonction de perte de log-vraisemblance négative

Parmi les candidats pour le modèle le mieux ajusté, le retrait de la variable catégorielle *ISO*, qui distingue les pays, a permis d'obtenir le meilleur ajustement global, comme en témoigne le score de log-vraisemblance négative. Ce résultat n'est pas surprenant, car la présence de cette variable complexifie grandement le modèle avec ses nombreux niveaux de catégorisation.

Toutefois, en considérant le score de la corrélation prédictive, l'impact de la variable *ISO* semble négligeable. Les scénarios sans *ISO* obtiennent des scores similaires à ceux de leurs homologues où la variable est incluse. Cela s'explique par le fait que *ISO* n'appartient pas à la *Markov Blanket* de la variable *Damage*, la rendant conditionnellement indépendante de *ISO* lorsque *Vulnerability* ou *Exposure* sont observées.

En revanche, le retrait de la variable *Vulnerability* a un impact négatif significatif sur les performances prédictives du modèle pour la variable cible *Damage*. Cette dégradation des

performances souligne l'importance de *Vulnerability* et indique qu'elle joue un rôle crucial dans la compréhension des facteurs influençant les dommages dans les scénarios étudiés.

Cependant, il est important de noter que, de manière générale, même le meilleur modèle prédictif reste peu fiable, avec une précision correcte dans seulement 35% des cas. Cette faiblesse pourrait être attribuée à plusieurs facteurs. D'une part, la variable *Damage* s'étend sur une large gamme de valeurs, ce qui complexifie la tâche du modèle, surtout dans un contexte où la base de données est relativement petite. Il est possible que le modèle soit confronté à une situation de type *small n, large p*, où le nombre de variables (p) est important par rapport au nombre d'observations (n), rendant difficile la capture de toutes les subtilités de *Damage*. D'autre part, une mauvaise spécification des autres variables pourrait également contribuer à cette performance prédictive.

Une amélioration notable pour renforcer la performance prédictive du modèle serait de disposer d'une base de données plus fournie et plus fiable. [6] La limitation actuelle du modèle semble en grande partie liée à la taille restreinte et à la qualité des données disponibles, le rendant pas assez bien équipé pour capturer les relations complexes qui mènent au montant de dégâts.

Une autre piste d'amélioration consisterait à intégrer des variables catégorielles supplémentaires qui expliqueraient mieux la relation entre le changement climatique et les dégâts liés aux catastrophes. Par exemple, des variables comme le niveau de précipitations pourraient jouer un rôle crucial dans la modélisation des impacts liés aux inondations. Pour les besoins de cette démonstration, ce type d'information n'était pas disponible pour l'ensemble de la période couverte par notre base de données, ni pour chaque localisation spécifique associée à chaque péril. L'ajout de ces variables pourrait significativement affiner la précision des prédictions en capturant des dynamiques supplémentaires.

Compte tenu de ses performances relatives par rapport aux autres modèles, le scénario retenu pour la suite est le modèle sans *Exposure*, mais avec deux niveaux de catégorisation pour la variable *Vulnerability*.

3.4 Propriétés du modèle

3.4.1 CPDs et CTPs

3.4.1.1 Distribution de probabilité conditionnelle : Temperature | GtCO2

La table 3.4.1 montre que le nœud *Temperature* est modélisé par une distribution normale conditionnelle, avec pour moyenne une fonction linéaire positive du nœud parent $GtCO_2$.

	Intercept	GtCO2
Coefficient	-0.47105035	0.03732219

TABLE 3.4.1 – Coefficients de la régression pour `Temperature | GtCO2`

6. Voir les biais de la base de données EM-DAT : `https://doc.emdat.be/docs/known-issues-and-limitations/`

3.4.1.2 Table de probabilité marginale : Disaster_Subtype

La table 3.4.2 présente la distribution de probabilité marginale de la variable catégorielle *Disaster_Subtype*, et donne des indications sur la fréquence des différents types de sinistres présents dans la base de données.

	Flood (General)	Ground movement	Hail	Storm (General)
Probability	0.78842975	0.08099174	0.01322314	0.11735537

TABLE 3.4.2 – Probabilités marginales pour Disaster_Subtype

3.4.1.3 Table de probabilités conditionnelles : Vulnerability | ISO

La table 3.4.3 montre la table de probabilités conditionnelles de *Vulnerability* avec son parent *ISO*. Beaucoup de pays gardent le même profil de vulnérabilité, mais d'autres ont évolué/régressé au fil des années.

	ALB	AUT	BEL	BGR	BIH	BLR	CHE	CZE	DEU	ESP
Élevée	0.5625	1.0000	0.7500	0.0385	0.0952	1.0000	1.0000	1.0000	1.0000	1.0000
Faible	0.4375	0.0000	0.2500	0.9615	0.9048	0.0000	0.0000	0.0000	0.0000	0.0000
	FIN	FRA	GBR	GRC	HRV	HUN	IRL	ITA	LTU	LUX
Élevée	1.0000	0.6296	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Faible	0.0000	0.3704	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LVA	MDA	MKD	MNE	NLD	NOR	POL	PRT	ROU	RUS
Élevée	1.0000	0.0000	0.1818	0.2500	1.0000	1.0000	1.0000	1.0000	0.2667	0.0000
Faible	0.0000	1.0000	0.8182	0.7500	0.0000	0.0000	0.0000	0.0000	0.7333	1.0000
	SCG	SRB	SVK	SVN	SWE	UKR				
Élevée	1.0000	0.0476	1.0000	1.0000	1.0000	0.2500				
Faible	0.0000	0.9524	0.0000	0.0000	0.0000	0.7500				

TABLE 3.4.3 – Table de probabilités conditionnelles Vulnerability | ISO

3.4.1.4 Table de probabilité marginale : ISO

La table 3.4.4 présente la distribution de probabilité marginale de la variable catégorielle *ISO*, et donne des indications sur la fréquence des différents pays dans la base de données.

	ALB	AUT	BEL	BGR	BIH	BLR	CHE	CZE	DEU	ESP	FIN	FRA	GBR
Probability	0.0264	0.0182	0.0198	0.0430	0.0347	0.0033	0.0116	0.0231	0.0364	0.0463	0.0017	0.0893	0.0545
	GRC	HRV	HUN	IRL	ITA	LTU	LUX	LVA	MDA	MKD	MNE	NLD	NOR
Probability	0.0678	0.0281	0.0248	0.0116	0.0893	0.0033	0.0017	0.0033	0.0066	0.0182	0.0066	0.0033	0.0033
	POL	PRT	ROU	RUS	SCG	SRB	SVK	SVN	SWE	UKR			
Probability	0.0215	0.0149	0.0744	0.1041	0.0116	0.0347	0.0215	0.0132	0.0017	0.0264			

TABLE 3.4.4 – Probabilités marginales pour ISO

3.4.1.5 Distribution de probabilité conditionnelle : Damage | Temperature, Disaster Subtype, Vulnerability

La table 3.4.5 présente la distribution de probabilité conditionnelle de la variable *Damage* selon ses trois nœuds parents : *Disaster_Subtype*, *Temperature* et *Vulnerability*. Les conclusions suivantes peuvent être tirées :

— L'intercept est généralement plus faible pour une vulnérabilité faible que pour une vulnérabilité élevée.

— Une augmentation de la température a un impact positif sur les dégâts, à l'exception des situations de grêle.

— Dans les configurations à faible vulnérabilité, l'impact de la température sur les dégâts est plus prononcé que dans les configurations à vulnérabilité élevée.

— Les évènements de grêle ainsi que les tempêtes en vulnérabilité faible montrent des résultats hors-normes.

Configuration	`Flood (General)`		`Ground movement`		`Hail`		`Storm (General)`	
	`Élevée`	`Faible`	`Élevée`	`Faible`	`Élevée`	`Faible`	`Élevée`	`Faible`
Intercept	12.41	8.82	11.92	7.16	20.80	13.46	10.57	0.86
Temperature	0.21	2.70	0.92	5.31	-7.68	0.00	1.31	12.05

TABLE 3.4.5 – Coefficients de la régression pour `Adjusted_damage`

3.4.2 Analyse des résidus de *Damage*

L'analyse des résidus permet d'évaluer l'ajustement du modèle, en vérifiant si les hypothèses sous-jacentes, comme la normalité des résidus, sont respectées. Pour les 8 configurations présentées dans la table 3.4.6, les Q-Q plots des résidus ont été générés pour observer leur distribution.

Configuration	`Vulnerability`	`Disaster_Subtype`
0	Élevée	Flood (General)
1	Faible	Flood (General)
2	Élevée	Ground movement
3	Faible	Ground movement
4	Élevée	Hail
5	Faible	Hail
6	Élevée	Storm (General)
7	Faible	Storm (General)

TABLE 3.4.6 – Configurations parentales des variables

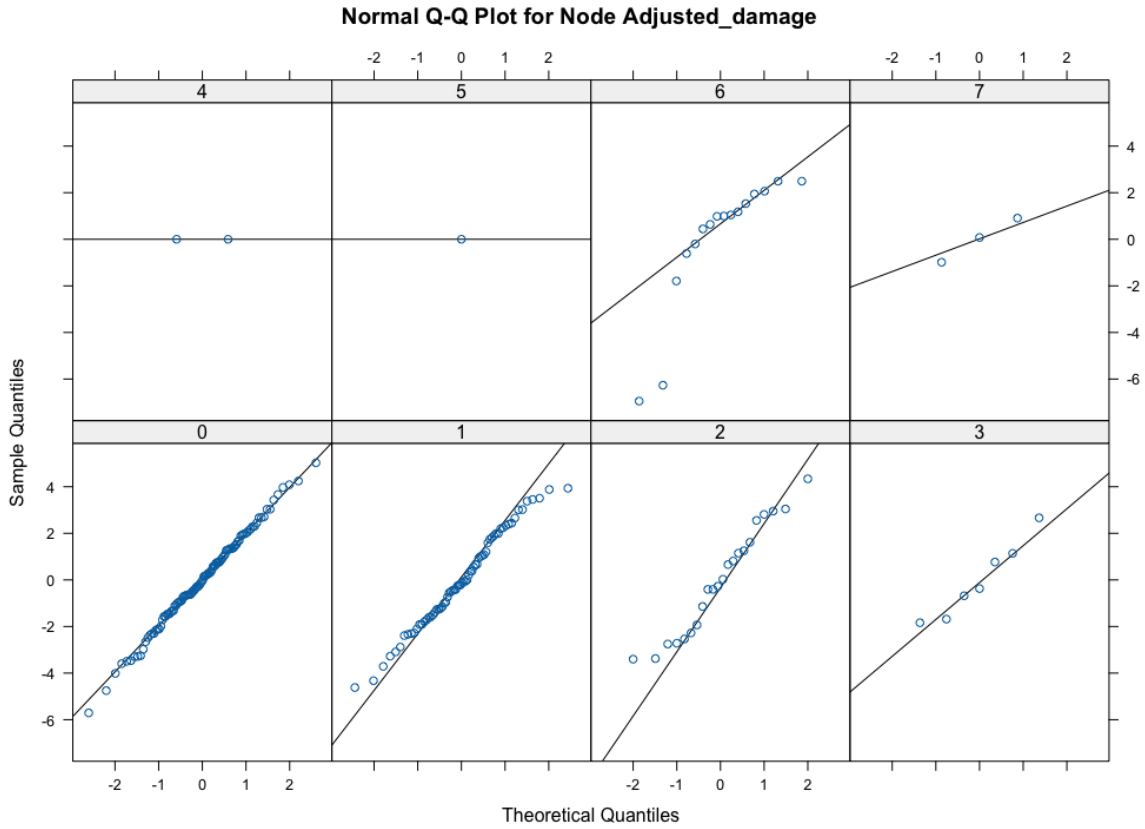


FIGURE 3.4.1 – Q-Q plots des résidus pour les différentes configurations

L'analyse des Q-Q plots révèle plusieurs points intéressants :

- Le péril de grêle comporte très peu d'observations, ce qui se traduit par une quasi-absence de variance des résidus, rendant difficile toute interprétation fiable pour ces configurations.
- Les configurations 3,6 et 7, souffrent également d'un nombre limité d'observations.
- Pour les autres configurations, les résidus suivent bien la droite théorique, avec seulement quelques déviations légères dans les quantiles extrêmes, suggérant que le modèle est généralement bien ajusté pour une distribution gaussienne, malgré des petites imperfections.

3.5 Prédiction

Une fois le modèle sélectionné et opérationnel, il devient possible de l'appliquer pour estimer les risques financiers auxquels une assurance pourrait être exposée lors de futurs événements.

Dans ce cadre, l'une des questions clefs pourrait concerner la sévérité attendue d'un prochain évènement type inondation dans notre position actuelle vis à vis du réchauffement climatique (1°C par rapport à la moyenne à long terme de 1951 à 1980). Il s'agit de faire du beleif updating, le BN met à jour la distribution de probabilité de la variable continue en fonction des nouvelles *evidences* reçues, qui, dans ce cas, correspondent à l'information *Disaster Subtype = inondation* et à *Temperature = 1.*

Ensuite, à partir du *Likelihood Weighting* **6**, des milliers de scénarios sont générés, et des moments de la distribution peuvent être calculés afin de construire la densité de probabilité de ces simulations. Une fois remis à leur échelle logarithmique, le résultat suivant est obtenu :

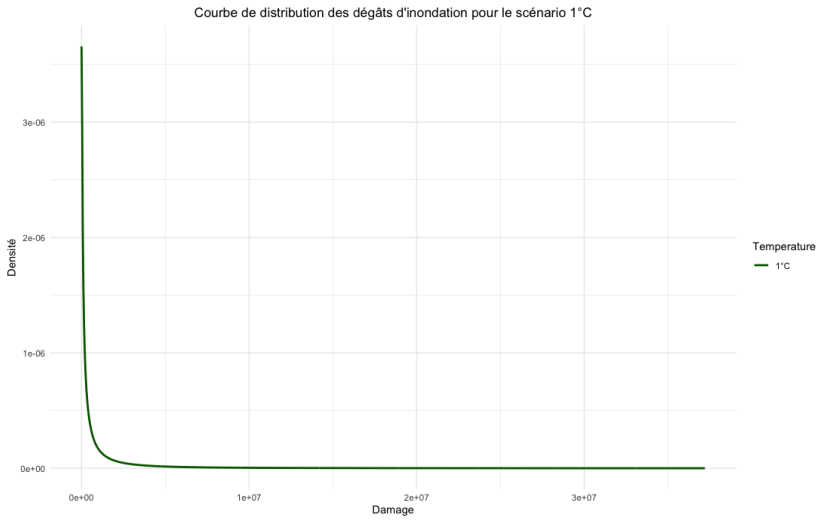


FIGURE 3.5.1 – Densité de probabilité des dégâts attendus pour un événement de type inondation à une température de 1°C au-dessus de la moyenne de 1951-1980

La courbe de distribution obtenue permet d'extraire des statistiques cruciales telles que la moyenne des dégâts, leur variance, ou de calculer une *Value at Risk* pour déterminer un seuil de perte qui ne serait dépassé qu'avec une faible probabilité.
Cependant, il est important de noter que sans un modèle de fréquence qui estime la probabilité d'occurrence de ces événements, ces statistiques ne permettront pas de tirer des conclusions précises ou d'apporter une aide complète à la décision.

3.6 Scénario futur

L'intérêt majeur de l'analyse causale dans la problématique abordée par cette étude réside dans sa capacité à permettre des inférences causales et des analyses contre-factuelles, souvent formulées sous la forme de requêtes "what-if". Contrairement aux modèles traditionnels basés uniquement sur l'historique des données passées, qui se limitent à capturer des corrélations observées, l'approche causale permet de simuler l'impact de changements hypothétiques dans

des variables clefs. Par exemple, un modèle traditionnel ne pourrait pas répondre à une question telle que : "Que se passerait-il si les émissions de CO_2 chutaient drastiquement en raison de l'introduction d'une nouvelle loi carbone ?".

Les BN, en tant que modèle causal, surmonte cette limitation en offrant la capacité de simuler de tels scénarios. Les scénarios proposés par le *Network for Greening the Financial System* (NGFS) offrent plusieurs trajectoires possibles en fonction de ces variables clefs. En appliquant ces scénarios dans un BN, il devient possible d'anticiper et de se préparer à une variété de situations hypothétiques.

Une des trajectoires prévues par le NGFS indique que le réchauffement de la planète pourrait atteindre 1,5 °C dans les années 2030. En utilisant notre modèle, cette hypothèse peut être intégrée comme une nouvelle *evidence*. Cette intégration permet d'observer son impact direct sur la distribution des dommages, représentée graphiquement par un déplacement de la courbe vers la droite :

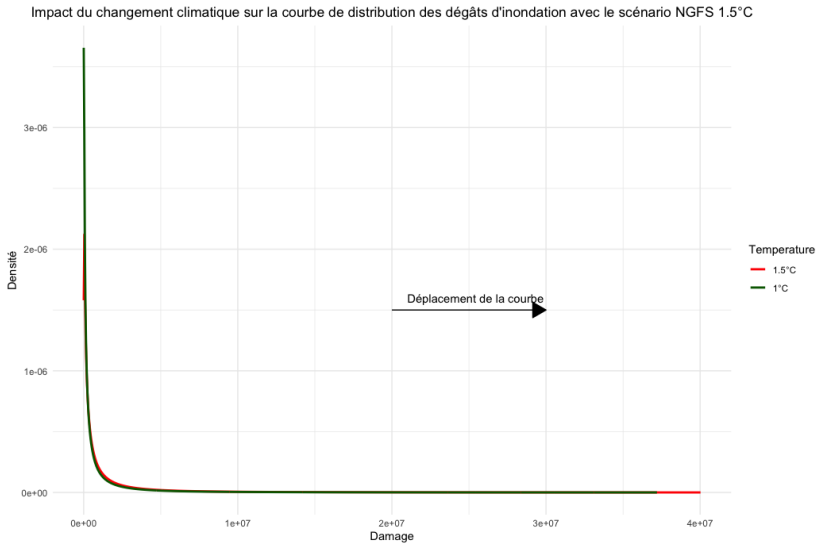


FIGURE 3.6.1 – Impact du réchauffement à 1,5 °C sur la distribution des dommages attendus

Ce déplacement de la courbe vers la droite suggère que, sous ce nouveau scénario, les dommages augmenteraient en moyenne, ce qui pourrait signaler une hausse des coûts pour les compagnies d'assurance. Avec un réseau plus étendu et plus riche, il devient possible d'explorer une gamme plus large de scénarios contre-factuels, et d'y modifier plusieurs variables à la fois. Chaque observation influencera la suivante, jusqu'à l'estimiation de la sévérité des évènements. En observant ces variables en amont, il est aussi possible d'évaluer comment les changements se propagent à travers le réseau et de trouver des leviers d'actions potentiels pour mitiger ce risque.

3.7 Association avec un modèle de fréquence

3.7.1 Introduction

Un modèle de sévérité, bien qu'essentiel pour évaluer les pertes individuelles, ne suffit pas à fournir une analyse complète du risque. Pour enrichir cette analyse, il est crucial de l'associer à un modèle de fréquence, permettant ainsi de modéliser les pertes agrégées sur une année, ainsi que d'estimer des mesures connexes telles que la Value at Risk à l'horizon d'un an et l'Expected Shortfall associé.

3.7.2 Exploration préliminaire

Le graphique 3.7.1 illustre la distribution annuelle du nombre d'événements de type "Flood (General)" entre les années 2000 et 2023.

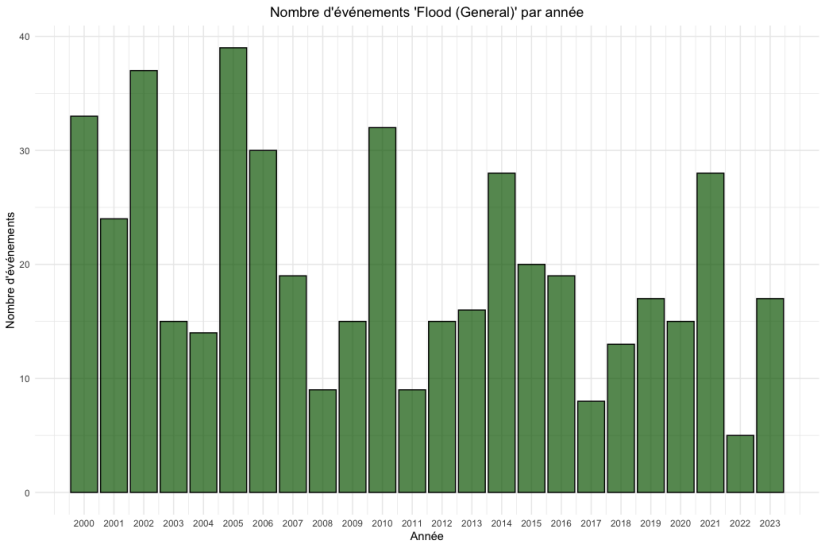


FIGURE 3.7.1 – Distribution annuelle du nombre d'événements "Flood (General)" entre 2000 et 2023

Cette distribution montre une variabilité importante du nombre d'événements d'une année à l'autre, avec une moyenne annuelle de $19,875$ événements de type flood et une variance de $88,636$.

3.7.3 Ajustement et validation d'un modèle

Pour modéliser la fréquence des événements annuels "Flood (General)", deux modèles de distribution ont été considérés : le modèle de Poisson et le modèle binomial négatif. Ces distributions ont été ajustées aux données à l'aide de la méthode du maximum de vraisemblance, et les résultats de ces ajustements sont résumés dans la table 3.7.1.
Le modèle binomial négatif fournit un meilleur ajustement que le modèle de Poisson, comme en témoignent ses valeurs plus faibles de l'AIC et du BIC, ainsi qu'une log-vraisemblance plus élevée. Ce résultat est conforme aux attentes, étant donné la surdispersion marquée dans les données observées.

Modèle	Paramètres Estimés	Log-vraisemblance	AIC / BIC
Poisson	$\lambda = 19,875$	-107,8572	217,7143 / 218,8924
Binomial Négatif	$size = 5,9004$, $\mu = 19,8743$	-86,1252	176,2505 / 178,6066

TABLE 3.7.1 – Score pour les modèles Poisson et Binomial Négatif

3.7.4 Simulations des pertes agrégées

Le modèle binomial négatif, ajusté précédemment, a été utilisé pour simuler le nombre d'occurrences des événements d'inondation sur une période d'une année, répété sur 1000 simulations différentes.

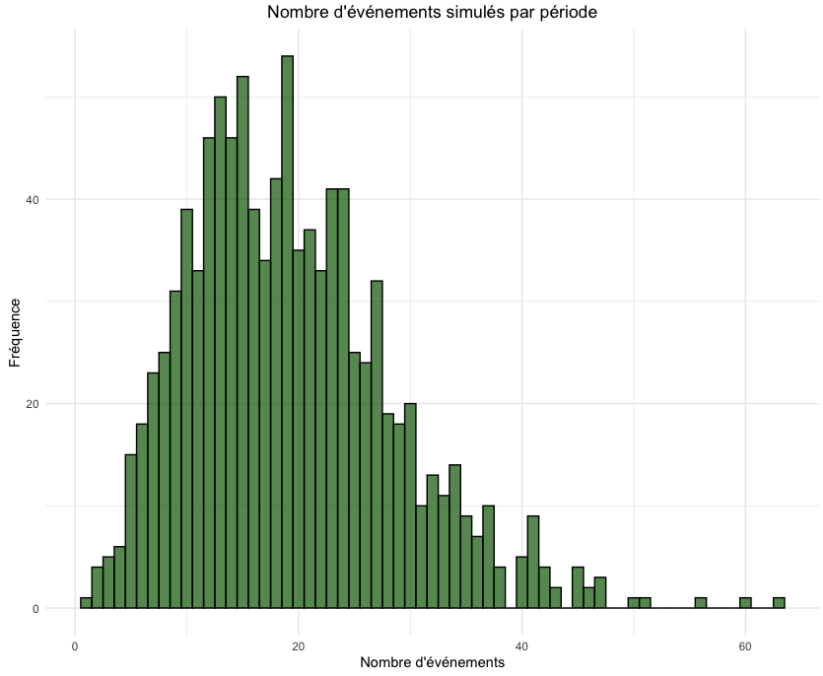


FIGURE 3.7.2 – Distribution simulée du nombre annuel d'inondations sur 1000 périodes

La variabilité qu'on y retrouve concorde bien avec les caractéristiques précédemment observées.

Ensuite, pour chaque période simulée est associée un nombre d'événements donné, et pour chacun de ces événement, des sévérités de pertes ont été tirées de notre distribution log-normale prédit par le BN dans le cas où $Temperature = 1$ et $Disaster\_Subtype = inondation$. Ce qui permet de calculer les pertes agrégées annuelles pour 1000 simulations différentes. La figure 3.7.3 met en évidence la nature asymétrique des pertes, avec une longueue queue à droite qui indique la possibilité de pertes rares mais dévastatrices.

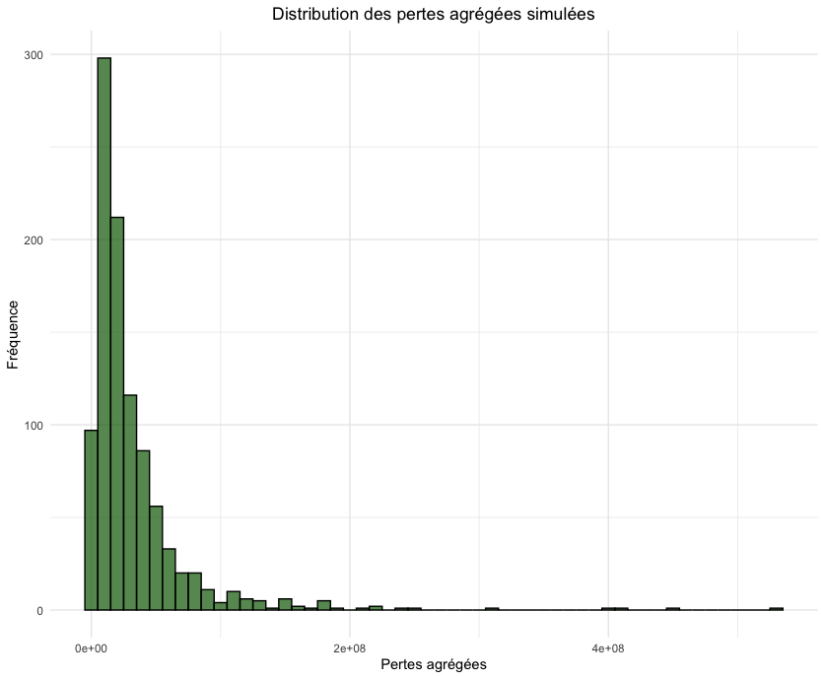


FIGURE 3.7.3 – Distribution des pertes agrégées annuelles simulées pour des événements d'inondation

3.8 Application de la théorie des valeurs extrêmes

Après avoir simulé les pertes agrégées, l'étape suivante consiste à appliquer la théorie des valeurs extrêmes pour se concentrer sur les événements rares mais sévères. Pour modéliser cette longue queue de distribution, la GPD est très utile en nous décrivant la distribution des excès au-dessus d'un seuil élevé. La première étape consiste à trouver judicieusement ce seuil, trop bas et il incluera un biais, trop haut et le nombre d'observations ne sera pas suffisant.

3.8.0.1 Sélection du seuil

Une des principales utilisations du Hill plot est de trouver la zone où le paramètre α se stabilise. La figure 3.8.1 montre une légère zone de stabilisation aux alentours de la statistique d'ordre 250, avant que le paramètre ne commence à diminuer. Ce point servira de seuil pour ajuster un modèle GPD.

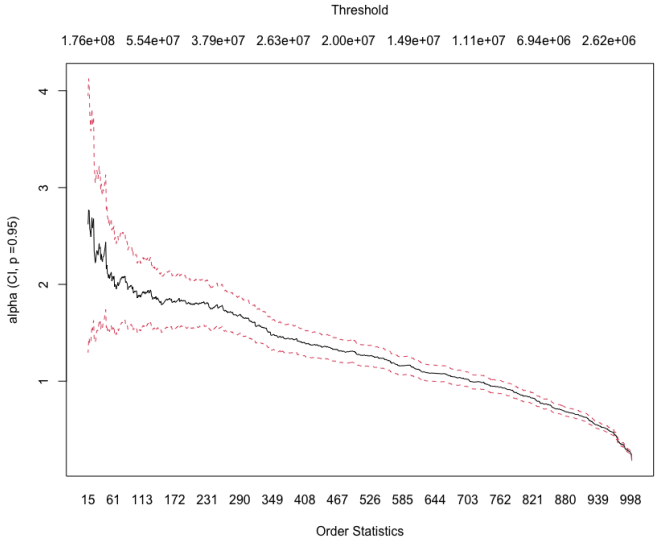
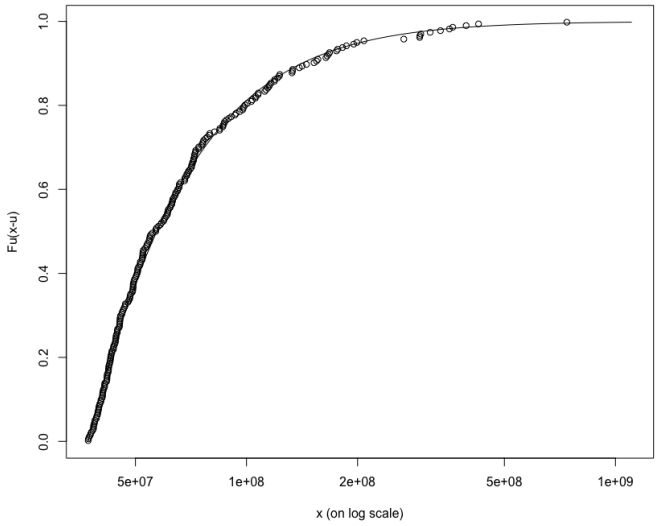


FIGURE 3.8.1 – Hill plot pour la sélection du seuil

3.8.0.2 Validation du modèle GPD

La courbe de la figure 3.8.0.2 montre que les excès au-dessus du seuil se répartissent conformément aux attentes théoriques. Les points empiriques suivent de près la courbe théorique, ce qui suggère un bon ajustement du modèle GPD qui possède les caractéristiques suivantes :

— **Paramètre de forme** (ξ) : $\xi = 0{,}499$
— **Paramètre d'échelle** (β) : $\beta = 24\ 078\ 639$
— **Seuil** : Seuil = 37 210 617 (correspondant à la statistique d'ordre 250)



Distribution des excès à partir du seuil

Après avoir validé l'ajustement du modèle GPD, nous passons à l'estimation de deux mesures cruciales de gestion des risques : la Value at Risk et l'Expected Shortfall à un niveau de confiance de 99,5% pour l'horizon d'une année.

3.8.0.3 Calcul de la Value at Risk et de l'Expected Shortfall à 99,5%

En utilisant les paramètres du modèle GPD, à savoir le paramètre de forme (ξ), le paramètre d'échelle (β), et le seuil choisi, il est désormais possible d'avoir une représentation de notre Value at Risk et de l'Expected Shortfall pour un horizon d'une année.

$$\text{VaR}_{99.5\%} = 668\ 353\ 284\$$$

$$\text{ES}_{99.5\%} = 1\ 345\ 698\ 815\$$$

Ces chiffres signifient que, il y a 99,5% de chances que les pertes ne dépassent pas 668,4 millions de dollars sur une année, soit 1 évènement tous les 200 ans. Cependant, si cet évènement arrive, alors la perte moyenne attendue serait d'environ 1,35 milliard de dollars.

3.8.0.4 Conclusion

Grâce à l'approche causale établie, il devient possible de calibrer plus précisément la charge au risque CAT, en intégrant non seulement les corrélations historiques observées, mais aussi les relations causales susceptibles de modifier les tendances futures.

Ces métriques jouent un rôle crucial dans le secteur de l'assurance, car elles déterminent le capital économique nécessaire pour maintenir la probabilité de ruine à 0,5% sur un horizon d'un an. En particulier, les analyses "what-if" permettent de simuler l'impact de scénarios hypothétiques, tels qu'une augmentation de la température moyenne, sur la courbe de sévérité des pertes. Cette approche holistique améliore la capacité d'anticipation du secteur assurantiel et garantit que les décisions en matière de capital économique restent robustes et bien informées, même dans un contexte de changements climatiques.

De plus, en personnalisant le modèle selon le profil spécifique d'un assureur, notamment en intégrant des nœuds de vulnérabilité et d'exposition propres à son portefeuille, il devient possible d'identifier des leviers d'action pertinents pour la prise de décision.

Chapitre 4

Limitations et développements futurs

Le modèle développé dans ce mémoire offre une première approche solide pour la modélisation des risques climatiques, en particulier à travers l'utilisation des BN pour capturer les relations causales entre les variables climatiques et les pertes potentielles. Cependant, plusieurs limitations et défis techniques sont apparus au cours de ce travail, ouvrant la voie à des améliorations et développements futurs.

Une des principales limitations du modèle actuel réside dans la granularité des données. Les indices d'exposition et de vulnérabilité utilisés dans le modèle sont uniformes à l'échelle nationale, sans distinction géographique en fonction de l'emplacement spécifique des événements climatiques. Cela peut introduire des biais dans la modélisation, car un événement climatique dans une région particulièrement vulnérable pourrait être sous-estimé, tandis qu'un autre dans une région moins exposée pourrait être surestimé.

De plus, la modélisation de la sévérité des événements est limitée par le manque de composantes climatiques précises. Par exemple, des variables telles que les précipitations, les températures extrêmes, ou les caractéristiques géographiques locales ne sont pas directement intégrées dans le modèle de sévérité, ce qui limite la capacité du modèle à expliquer les variations des dommages d'un événement à l'autre. Aussi, de nombreuses données sont également manquantes ou incomplètes, ce qui a nécessité des hypothèses simplificatrices qui peuvent affecter la robustesse des résultats.

Une autre limitation notable réside dans la complexité des phénomènes météorologiques, souvent influencés par des processus chaotiques. Particulièrement, la modélisation de la fréquence des événements climatiques pose un défi particulier. Capturer la fréquence de tels événements dans un BN standard nécessite l'inclusion de nombreuses variables pour représenter les divers facteurs qui les influencent, ce qui peut entraîner une structure de réseau plus complexe et exigeante en données climatiques.

Une perspective de développement sont les Spatio-Temporal Bayesian Networks qui offrent une solution particulièrement adaptée. Contrairement aux BN traditionnels, les STBN permettent de modéliser les dépendances spatio-temporelles entre les variables, capturant ainsi les variations géographiques et temporelles des événements climatiques. Ils relaxent également l'hypothèse de dépendance linéaire, une contrainte lourde qu'impose les BN standards.

Chapitre 5

Conclusion

L'objectif de ce mémoire était de développer une approche innovante pour la gestion des risques climatiques en utilisant une approche causale mieux adaptée à la problématique. Dans un contexte où les modèles traditionnels de gestion des risques reposent généralement sur des analyses statistiques des données historiques, ils montrent leurs limites face à l'incertitude croissante liées aux changements que notre monde vit. L'élaboration d'un modèle préliminaire de catastrophe basé sur les réseaux bayésiens a permis de capturer efficacement ces dynamiques.

Cependant, ce travail a été confronté à plusieurs contraintes techniques. L'un des principaux obstacles résidait dans la gestion de la granularité et du volume des données, ainsi que dans la modélisation de la fréquence des événements climatiques extrêmes en tenant compte de leur complexité causale. Ces limitations ont restreint la robustesse du modèle, illustrant les problèmes inhérents à la modélisation dans un contexte de données imparfaites et de phénomènes complexes.

Pour enrichir l'analyse fournie par notre modèle préliminaire, il a été complété avec un modèle de fréquence classique. Cette approche combinée a permis de simuler des mesures de risques plus détaillées et utilisables, telles que la Value at Risk et l'Expected Shortfall utilisés comme base pour la détermination du capital cible dans *Solvency II*.

Des pistes d'amélioration ont également été identifiées. En collaborant étroitement avec des fournisseurs de données, il serait possible d'enrichir les variables causales de notre modèle en intégrant des informations plus détaillées et spécifiques à chaque type d'aléa climatique, tout en tenant compte de l'exposition et de la vulnérabilité propres à un portefeuille ou à une zone géographique. Ainsi, en segmentant les différents aléas et en s'appuyant sur un modèle plus large et enrichi de données, nous pourrions capturer l'étendue des dégâts de manière plus fine, rendant le modèle plus capable d'expliquer les variations des dommages d'un événement à l'autre. Avec un tel modèle, il serait également possible d'identifier des leviers d'action permettant de mitiger le risque, en adaptant les stratégies aux spécificités locales ou aux caractéristiques de chaque portefeuille.

Bibliographie

[1] 4 AVRIL 2014. - Loi relative aux assurances. s. d. Consulté le 13 avril 2024. https://www.ejustice.just.fgov.be/cgi_loi/loi_a1.pl?language=fr&la=F&cn=2014040423&table_name=loi&&caller=list&F&fromtab=loi&tri=dd+AS+RANK&rech=1&numero=1&sql=(text+contains+(%27%27))#Art.123

[2] Crowe, Michelle Adcock, Radhika Bains, Thomas. 2024. «Quantifying Climate Risk in Insurance - KPMG Global». KPMG, 18 mars 2024. https://kpmg.com/xx/en/home/insights/2022/10/quantifying-climate-risk-in-insurance.html

[3] The European Actuary. 2023. «DEVELOPING STRATEGIES FOR NATURAL CATASTROPHE». Quarterly Magazine of the Actuarial Association of Europe, 9 septembre 2023.

[4] ecm114. 2020. « Climate Change and the Insurance Industry : Managing Risk in a Risky Time ». Georgetown Journal of International Affairs (blog). 9 juin 2020. https://gjia.georgetown.edu/2020/06/09/climate-change-and-the-insurance-industry-managing-risk-in-a-risky-time/

[5] Golnaraghi, Maryam. s.d. «Climate Change Risk Assessment for the Insurance Industry».

[6] Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., et al. (2021). Weather and climate extreme events in a changing climate. In V. Masson-Delmotte, P. Zhai, A. Pirani, et al. (Eds.), Climate Change 2021 : The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.

[7] Rakotomalala, R. (s.d.). Tests de normalité : Techniques empiriques et tests statistiques (Version 2.0). Université Lumière Lyon 2.

[8] Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., et al. (2021). Weather and climate extreme events in a changing climate. In V. Masson-Delmotte, P. Zhai, A. Pirani, et al. (Eds.), Climate change 2021 : The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (pp. 1513–1766). Cambridge University Press. https ://doi.org/10.1017/9781009157896.013

[9] Majeed, Abdul, et Ibtisam Rauf. «Graph Theory : A Comprehensive Survey about Graph Theory Applications in Computer Science and Social Networks». Inventions 5, n 1 (mars 2020) : 10. https ://doi.org/10.3390/inventions5010010.

[10] Catanzaro, Daniele. «Coding Project - Projet de Programmation - MINFO1302»

[11] Pearl, J., Glymour, M., Jewell, N. P. (2016). Causal inference in statistics : A primer. Wiley.

[12] Ocean acidification | National Oceanic and Atmospheric Administration. Consulté 30 juillet 2024, à l'adresse https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-acidification

[13] Lee, R. (2023). Climate Risk Assessment and Scenario Analysis.

[14] Kaikkonen, L., Parviainen, T., Rahikainen, M., Uusitalo, L., & Lehikoinen, A. (2021). Bayesian Networks in Environmental Risk Assessment : A Review. Integrated Environmental Assessment and Management, 17(1), 62-78. https://doi.org/10.1002/ieam.4332

[15] Gomes Marques, I., Vieites-Blanco, C., Rodríguez-González, P. M., Segurado, P., Marques, M., Barrento, M. J., Fernandes, M. R., Cupertino, A., Almeida, H., Biurrun, I., Corcobado, T., Costa E Silva, F., Díez, J. J., Dufour, S., Faria, C., Ferreira, M. T., Ferreira, V., Jansson, R., Machado, H., ... Jung, T. (2024). The ADnet Bayesian belief network for alder decline : Integrating empirical data and expert knowledge. Science of The Total Environment, 947, 173619. https://doi.org/10.1016/j.scitotenv.2024.173619

[16] Chen, S. H., & Pollino, C. A. (2012). Good practice in Bayesian network modelling. Environmental Modelling & Software, 37, 134-145. https://doi.org/10.1016/j.envsoft.2012.03.012

[17] Cain, J., 2001. Planning Improvements in Natural Resources Management. In : Guidelines for Using Bayesian Networks to Support the Planning and Management of Development Programmes in the Water Sector and Beyond. Centre for Ecology and Hydrology, Wallingford, UK.

[18] Marcot, B. G., Steventon, J. D., Sutherland, G. D., & McCann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. Canadian Journal of Forest Research, 36(12), 3063–3074. https ://doi.org/10.1139/x06-135

[19] Spirtes, P., Glymour, C., Scheines, R., & Heckerman, D. (2001). *Causation, Prediction, and Search*. MIT Press.

[20] Margaritis, D. (2003). Learning Bayesian Network Model Structure from Data. *Ph.D. thesis*, School of Computer Science, Carnegie-Mellon University.

[21] Tsamardinos, I., Aliferis, C. F., Statnikov, A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 673–678.

[22] Yaramakala, S., & Margaritis, D. (2005). Speculative Markov blanket discovery for optimal feature selection. *Proceedings of the Fifth IEEE International Conference on Data Mining*, 809-812.

[23] Thierry Roncalli. Handbook of financial risk management. CRC press, 2020.

[24] Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2003). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1), 31-78.

[25] Mean Field Variational Bayes for Elaborate Distributions - Scientific Figure on ResearchGate. Available from : https://www.researchgate.net/figure/Markov-blankets-for-each-of-the-six-parameters-hidden-nodes-in-the-example-Bayesian_fig1_254212653

[26] Nagarajan, R., Scutari, M., Lèbre, S. (2013). Bayesian Networks in R : With Applications in Systems Biology. Springer New York. https://doi.org/10.1007/978-1-4614-6446-4

[27] EM-DAT and the CRED. (2023, avril 17). https ://doc.emdat.be/docs/about/emdat-and-the-cred/

[28] WorldRiskIndex—Humanitarian Data Exchange. (s. d.). Consulté 9 mai 2024, à l'adresse https ://data.humdata.org/dataset/worldriskindex ?

[29] Change, N. G. C. (s. d.). Global Surface Temperature | NASA Global Climate Change. Climate Change : Vital Signs of the Planet. Consulté 9 mai 2024, à l'adresse https ://climate.nasa.gov/vital-signs/global-temperature ?intent=121

[30] https ://www.iea.org/about

[31] Understanding Particular Data Quality Concerns in EM-DAT. (2023, avril 13). https ://doc.emdat.be/docs/known-issues-and-limitations/specific-biases/

[32] ck. (s. d.). WeltRisikoBericht. WeltRisikoBericht. Consulté 9 mai 2024, à l'adresse https ://weltrisikobericht.de/en/

[33] Mitchell-Wallace, K. (Éd.). (2017). Natural catastrophe risk management and modelling : A practitioner's guide. John Wiley and Sons, Inc.

[34] Ghosh, S., Resnick, S. (2010). A discussion on mean excess plots. Stochastic Processes and their Applications, 120(8), 1492-1517. https ://doi.org/10.1016/j.spa.2010.04.002

[35] Giovanni Gualtieri. Reliability of era5 reanalysis data for wind resource assessment : A comparison against tall towers. Energies, 14(14) :4169, 2021. Received : 16 June 2021 / Revised : 7 July 2021 / Accepted : 8 July 2021 / Published : 10 July 2021.

[36] Chevallier, Matthieu, et al. "The Rise of Data-Driven Weather Forecasting : A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context." Bulletin of the American Meteorological Society, vol. 105, no. 6, 2024, pp. E864–E883. doi :10.1175/BAMS-D-23-0162.1.

[37] "Comparing Data-Driven and Mechanistic Models for Predicting Phenology in Deciduous Broadleaf Forests." arXiv preprint arXiv :2401.03960, 2023.

[38] Jakeman, Anthony J., et al. "Ten iterative steps in development and evaluation of environmental models." Environmental Modelling Software, vol. 21, no. 5, 2006, pp. 602-614.

[39] Pollio, G., et al. "Good practice in Bayesian network modelling." Environmental Modelling Software, 2012.

[40] Lee, Robert, et al. "Climate Risk Assessment and Scenario Analysis : A Case Study on Climate Change Impact on Home Prices Utilizing Bayesian Network Qualitative to Quantitative Analysis." Society of Actuaries, 2023. Sponsor : Catastrophe and Climate Strategic Research Program Steering Committee.

[41] Pearl, Judea. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[42] Scutari, M. (2017). Understanding Bayesian Networks—With Examples in R.

[43] Zheng, Y., Zhang, W., Lopez, D. M. B., Ahmad, R. (2021). Scien-tometric analysis and systematic review of multi-material additivemanufacturing of polymers. Polymers. https ://doi.org/10.3390/polym13121957

[44] Kaikkonen, L., Parviainen, T., Rahikainen, M., Uusitalo, L., Lehikoinen, A. (2021). Bayesian Networks in Environmental Risk Assessment : A Review. Integrated Environmental Assessment and Management, 17(1), 62-78. https ://doi.org/10.1002/ieam.4332

[45] Ramazi, P. (2023). Bayesian and Causal Bayesian Networks. https ://openlibrary-repo.ecampusontario.ca/jspui/handle/123456789/1790

[46] Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge University Press.

[47] Imbert, A., Vialaneix, N. (2018). Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : Une revue des approches existantes. 159(2).

[48] Singh, A. K., Allen, D. E., Powell, R. J. (s. d.). Value at Risk Estimation using Extreme Value Theory.

[49] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 3(5), 1163-1174. https ://doi.org/10.1214/aos/1176343247

[50] Université Claude Bernard Lyon 1. (s.d.). Cours en ligne : Probabilités et statistiques. Université Claude Bernard Lyon 1. https ://mathsv.univ-lyon1.fr/app/cours/ ?theme=probachap=2

[51] Dimitrova, E. S., Licona, M. P. V., McGee, J., Laubenbacher, R. (2010). Discretization of Time Series Data. Journal of Computational Biology, 17(6), 853-868. https ://-doi.org/10.1089/cmb.2008.0023

[52] Dougherty, J., Kohavi, R., Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features. In Machine Learning Proceedings 1995 (p. 194-202). Elsevier. https ://doi.org/10.1016/B978-1-55860-377-6.50032-3

Des modèles LLM ont aussi été utilisés au cours de la rédaction de ce mémoire à des fins de correction orthographique, de traduction des formules mathématiques en langage LaTeX, ainsi que pour l'aide au débogage du code R.

Annexes

Aperçu de la base de données

	ISO	Disaster Subtype	Damage ($)	GtCO2	Temperature	Exposure	Vulnerability
1	CZE	Flood (General)	141557.00	25.50	0.39	0.10	14.22
2	ROU	Flood (General)	885.00	25.50	0.39	0.69	14.29
3	RUS	Flood (General)	9330.00	25.50	0.39	29.28	32.79
4	ROU	Flood (General)	176946.00	25.50	0.39	0.69	14.29
5	HUN	Flood (General)	97321.00	25.50	0.39	0.91	0.11
6	SCG	Flood (General)		25.50	0.39	0.17	0.00
7	RUS	Flood (General)		25.50	0.39	29.28	32.79
8	GBR	Flood (General)		25.50	0.39	7.21	2.59
9	RUS	Flood (General)		25.50	0.39	29.28	32.79
10	GRC	Ground movement		25.50	0.39	10.65	8.47
11	AUT	Storm (General)	35389.00	25.50	0.39	0.17	9.02
12	GBR	Flood (General)	32116.00	25.50	0.39	7.21	2.59
13	RUS	Flood (General)	12386.00	25.50	0.39	29.28	32.79
14	FRA	Flood (General)		25.50	0.39	2.72	13.21
15	FRA	Flood (General)		25.50	0.39	2.72	13.21
16	ESP	Flood (General)		25.50	0.39	7.83	11.72
17	RUS	Flood (General)		25.50	0.39	29.28	32.79
18	UKR	Storm (General)		25.50	0.39	0.48	14.55
19	RUS	Flood (General)	53084.00	25.50	0.39	29.28	32.79
20	RUS	Ground movement	1628.00	25.50	0.39	29.28	32.79
21	RUS	Flood (General)	2870.00	25.50	0.39	29.28	32.79
22	ITA	Flood (General)		25.50	0.39	8.85	12.78
23	UKR	Storm (General)		25.50	0.39	0.48	14.55
24	RUS	Flood (General)		25.50	0.39	29.28	32.79
25	ITA	Flood (General)		25.50	0.39	8.85	12.78
26	GBR	Flood (General)	10439837.00	25.50	0.39	7.21	2.59
27	ITA	Flood (General)	14155712.00	25.50	0.39	8.85	12.78
28	IRL	Storm (General)	176946.00	25.50	0.39	3.40	1.48
29	NOR	Flood (General)		25.50	0.39	1.05	9.89
30	ESP	Flood (General)	132710.00	25.50	0.39	7.83	11.72

Réseau obtenu avec l'algorithme Hill-Climbing

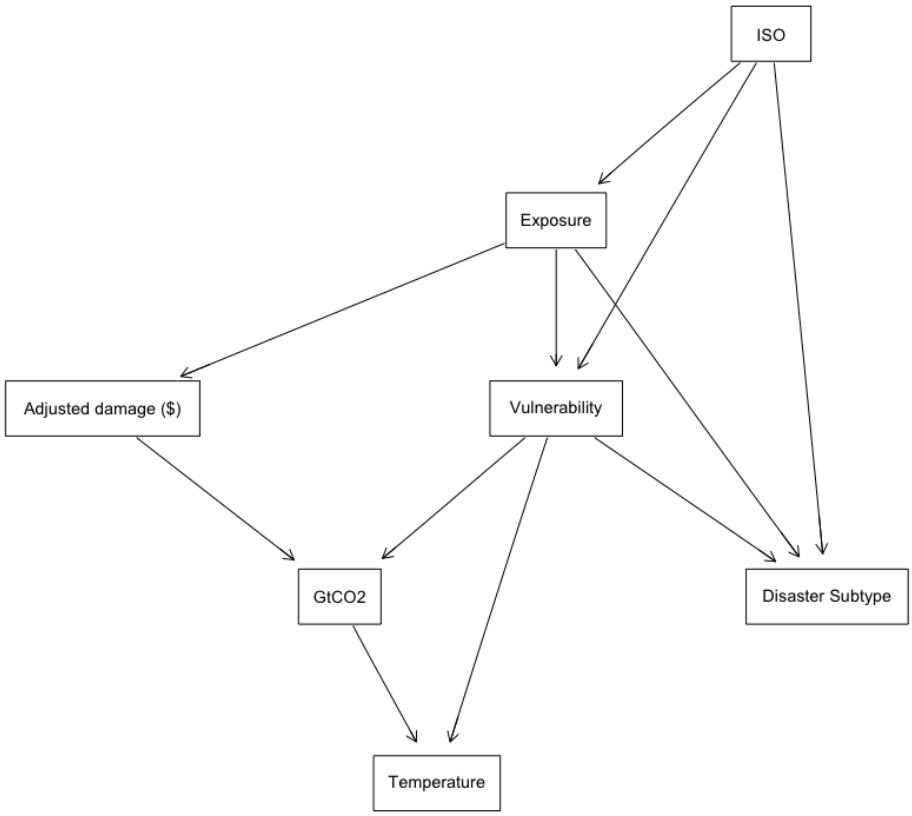


FIGURE .0.1 – DAG à 7 nœuds généré par l'algorithme Hill-Climbing

Réseau obtenu avec l'algorithme ICA

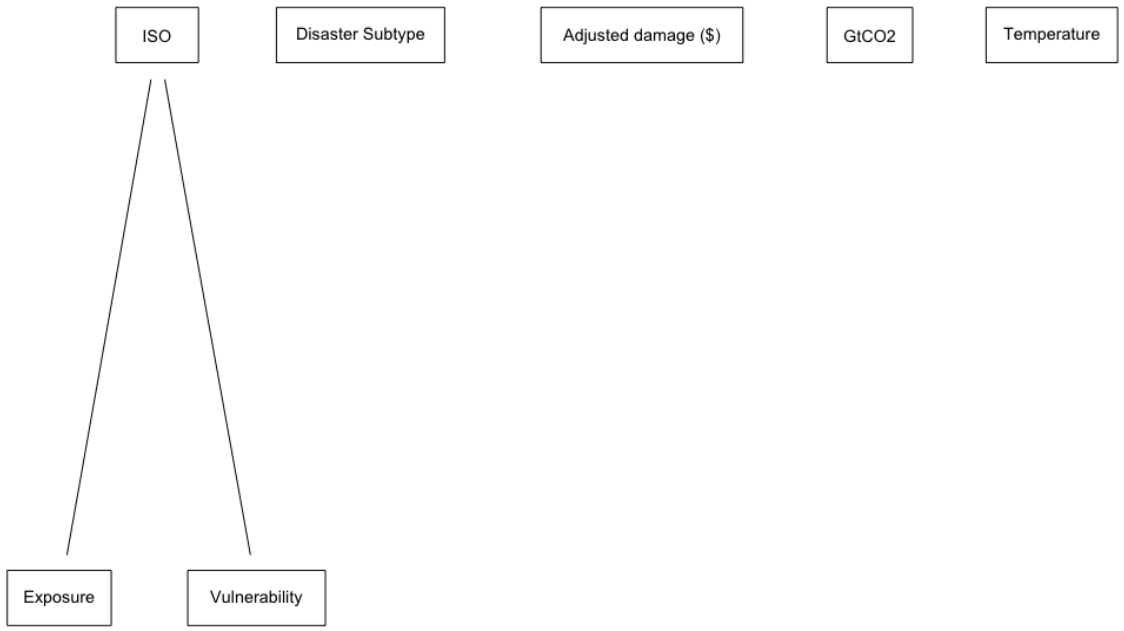


FIGURE .0.2 – DAG à 7 nœuds généré généré par l'algorithme ICA

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**