

**Mémoire présenté devant l'ENSAE ParisTech
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires**

le 30 Mai 2014

Par : Laura COHEN-SALMON, Dulcy Joyce GNINGHAYE FONGANG

Titre: Modélisation du risque gel en France

Confidentialité : NON OUI (Durée : 1 an 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

*Membre présents du jury de l'Institut
des Actuaires*

signature

Entreprise :

AXA Group Risk Management

Signature :

Directeur de mémoire en entreprise :

Pierre MOUILHADE

Signature :

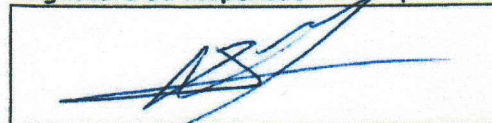
Invité :

Nom :

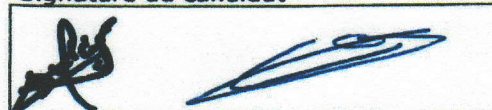
Signature :

**Autorisation de publication et de
mise en ligne sur un site de
diffusion de documents actuariels
(après expiration de l'éventuel délai de
confidentialité)**

Signature du responsable entreprise



Signature du candidat



Secrétariat

Bibliothèque :

Résumé

Mots clés : catastrophe naturelle, gel, modèle catastrophe, classification, EOF, modèle linéaire généralisé, dépassement de seuil, série temporelle, copule, courbe de destruction, AEP, OEP.

Depuis plusieurs années, les catastrophes naturelles ont tendance à engendrer des coûts de plus en plus importants et peuvent impacter ainsi directement la solvabilité des organismes d'assurance. Les assureurs ont donc besoin de quantifier les risques que présentent les catastrophes naturelles et d'estimer les pertes qu'elles pourraient entraîner. Or, la survenance imprévisible de ces phénomènes rend la tâche difficile à réaliser. Pour cela, des modèles catastrophe sont spécifiquement construits pour permettre d'accroître significativement la connaissance des assureurs sur ces types de risque.

L'objectif de ce mémoire est de construire un modèle catastrophe pour le péril gel en France. La survenance de gel étant directement liée à la température, la première partie a consisté à segmenter la France en plusieurs régions présentant des températures homogènes afin de réduire la dimension du problème et de faciliter les analyses. Ensuite, la deuxième partie a permis de modéliser le phénomène de gel dans le but de générer plusieurs scénarios d'évènements susceptibles de se produire en France pendant une année. Puis, il a fallu transformer la survenance d'évènements gel en perte économique en fonction de l'exposition du portefeuille d'AXA grâce à la construction de courbes de destruction. Enfin, dans une dernière partie, les résultats ont été mis en commun pour déterminer une distribution des pertes subies par AXA en raison du déclenchement d'évènements de gel en France.

Abstract

Key words: natural disaster, freeze, catastrophe modelling, clustering, EOF, Generalized Linear Model, Peak Over Threshold, time series, copula, vulnerability curve, AEP, OEP.

Over the past few years, the economic impact of natural catastrophes on populations has notoriously increased, jeopardizing the financial strength of insurance companies. Insurers thus need to measure the underlying risk and the losses their clients would experience from it. However, due to the unpredictability of natural disasters, specific catastrophe models must be conceived in order to increase insurers' understanding of these risks and their consequences in terms of insurance.

This paper aims at conceiving a catastrophe model for the freeze peril in France. Freeze being directly linked to temperature, the first part consisted in reducing the dimension of the study by defining temperature-homogeneous regions, enabling to later model each of them as a whole. In the second part, freeze was modelled from historical data, enabling the stochastic simulation of possible future events in France. After building custom vulnerability curves, we were able to deduce the economic losses triggered by the simulated events on AXA's portfolio. Finally, results associated to each event were put together to determine AXA's distribution of losses associated to freeze events in France.

Remerciements

Nous aimerions avant tout adresser nos remerciements à notre encadrant, Pierre Mouilhade, pour ses conseils avisés, sa patience et sa disponibilité tout au long de l'année.

Nous tenons également à remercier vivement Fabrice Balland pour sa collaboration et son aide précieuse dans l'élaboration de ce mémoire.

Enfin, nous souhaitons remercier les professeurs et intervenants de l'ENSAE pour leurs réponses à nos questions concernant le mémoire. Nous pensons notamment à Olivier Lopez et Philippe Tann.

Sommaire

Introduction	4
I. Le risque gel et sa modélisation	5
1. Les risques catastrophes naturelles	5
1.1. Définition	5
1.2. Impact des catastrophes naturelles pour un assureur	5
1.3. Le risque gel	6
2. Intérêt de la modélisation du risque catastrophe	8
2.1. Optimisation du capital réglementaire	8
2.2. Tarification de traités de réassurance optimaux	9
3. La modélisation catastrophe	11
3.1. Définition d'un modèle catastrophe	11
3.2. Le module Aléa	12
3.3. Le module Vulnérabilité	12
3.4. Le module Financier	14
3.5. Résultats d'un modèle catastrophe	16
4. La modélisation catastrophe adaptée au risque gel.....	18
4.1. Caractérisation du risque gel.....	18
4.2. Présentation des données utilisées	18
II. La classification des régions	20
1. L'Empirical Orthogonal Function (EOF)	20
1.1. Méthode EOF et quelques variantes	21
1.2. La matrice de données.....	21
1.3. Méthode de résolution.....	22
1.4. Critères de choix des dimensions.....	23
1.5. Analyse des résultats	23
2. La classification	25
2.1. Notion de distances et de similarités	26
2.2. Les méthodes de classification.....	26
3. Première approche : combinaison des méthodes EOF et CAH	29
3.1. La distance euclidienne	29
3.2. Le critère de Ward pour les distances euclidiennes.....	29
3.3. Les résultats de la CAH sur les données issues de la méthode EOF	30
4. Deuxième approche : CAH avec utilisation de la distance DTW	34
4.1. La distance DTW.....	34
4.2. Les résultats de la CAH avec utilisation de la DTW	37

III. Construction du module Aléa.....	39
1. Caractérisation de l'évènement gel.....	39
1.1. Les déterminants de la survenance du gel	40
1.2. Modélisation de la fréquence de gel	45
2. Modélisation de la température.....	56
2.1. Présentation des données.....	56
2.2. Le modèle linéaire additif simple	57
2.3. Le modèle ARMA	58
2.4. Le modèle à variance périodique.....	61
2.5. Application de la méthode aux autres régions	68
3. Modélisation de la dépendance entre régions	72
3.1. Présentation des données.....	72
3.2. La théorie des copules	73
3.3. Les copules paramétriques	74
3.4. Méthode d'estimation	77
3.5. Sélection de la copule optimale.....	79
4. Simulations de scénarios.....	85
4.1. Scénarios de température	85
4.2. Scénarios d'évènements	86
IV. Construction du module Vulnérabilité.....	88
1. Modéliser la perte assurantielle	88
2. La méthode MBBEFD	88
2.1. MBBEFD à deux paramètres.....	89
2.2. MBBEFD à un paramètre (Hyperbolic MBBEFD)	90
3. Taux de destruction médian.....	91
4. Construction des courbes de destruction.....	93
V. Résultats du modèle.....	95
1. Obtention des pertes journalières par police et par scénario	95
2. Construction des courbes AEP et OEP	96
2.1. Agrégation des pertes journalières pour la réassurance.....	96
2.2. La courbe OEP.....	96
2.3. La courbe AEP	96
Conclusion	98
Table des figures	99

Liste des tableaux.....	102
Bibliographie.....	103
Annexe A - Schéma climatique de la France métropolitaine	105
Annexe B - Fonctions d'autocorrélations et tests statistiques utilisés .	106
Annexe C - Résultats des tests de significativité.....	108
Annexe D - Dépendance de queue.....	108
Annexe E - Dépendogrammes et fonctions de Kendall	109
Annexe F - Courbes de destruction associées à chaque LoB	119

Introduction

L'entrée en vigueur prochaine de Solvabilité II¹ impose à tous les assureurs d'avoir la meilleure connaissance possible de l'ensemble des risques auxquels ils s'exposent. Dans cette directive, un traitement spécifique est notamment prévu pour les catastrophes naturelles dont les impacts peuvent s'avérer dramatiques pour les compagnies d'assurance. A titre d'exemple, l'ouragan Andrew, survenu en 1992 en Floride, a causé la faillite de onze assureurs. De telles conséquences incitent les compagnies d'assurance à anticiper leurs pertes éventuelles en cas de réalisation d'évènements extrêmes et à constituer des fonds propres spécifiques pour pouvoir y faire face. La prise en compte des risques de catastrophes naturelles suscite ainsi un intérêt croissant chez les assureurs. Pour ce faire, ces risques doivent inévitablement être modélisés.

Parmi les évènements extrêmes coûteux pour une compagnie d'assurance, on retrouve les tremblements de terre, les ouragans ou encore les éruptions volcaniques. Pour ces risques, certains assureurs possèdent déjà des modèles, appelés modèles catastrophe, qui permettent de simuler la survenance du risque et les pertes associées. Ces modèles sont complexes car ils doivent à la fois modéliser le phénomène physique mais aussi les répercussions financières en cas de réalisation du péril. Des entreprises spécialisées dans la modélisation catastrophe existent et ont pour rôle de créer des modèles adaptés à différents périls.

Lorsque l'on évoque les catastrophes naturelles, il est rare de mentionner le risque gel car sa survenance n'a généralement pas un impact aussi spectaculaire que les périls énoncés ci-dessus. Les entreprises spécialisées dans la modélisation catastrophe ne modélisent donc pas ce risque en priorité. Cela pousse les assureurs à développer leurs propres modèles afin de mieux cerner ce risque. En effet, la durée d'un évènement gel et la forte accumulation des polices touchées lors de sa survenance peuvent entraîner de lourdes pertes. Tel a été le cas d'AXA durant l'année 2012 qui fut une année particulièrement froide.

Dans ce mémoire, nous appliquerons au gel l'approche usuelle utilisée pour modéliser les catastrophes naturelles. Cela confèrera à l'assureur de multiples avantages comme la couverture de ce risque par un traité de réassurance ou la constitution de fonds propres nécessaires à la couverture de ce risque.

Pour cela, nous allons procéder en quatre étapes. Tout d'abord, nous effectuerons une classification des régions de France en fonction de leurs températures afin de réduire la quantité d'information à notre disposition. Puis, dans les deux parties suivantes, nous détaillerons la construction de deux modules d'un modèle catastrophe, en les adaptant spécifiquement au péril gel. Le premier module permettra de générer des évènements de gel en France sur une année et le deuxième module aura pour objectif de calculer l'impact de ces évènements sur le portefeuille d'AXA, traduisant les évènements physiques en pertes économiques. Enfin, dans la dernière partie, nous étudierons la distribution des pertes causées par le gel obtenue en sortie du modèle.

¹Directive européenne réglementaire du monde de l'assurance

I. Le risque gel et sa modélisation

1. Les risques catastrophes naturelles

1.1. Définition

Une catastrophe naturelle est un événement d'origine naturelle, subit et brutal, pouvant générer une accumulation d'importants dégâts matériels et humains. Une catastrophe naturelle se distingue d'autres événements par une fréquence de survenance faible et une forte intensité.

Les catastrophes naturelles ont longtemps été considérées comme des événements inévitables et imprévisibles mais les progrès de la science permettent à présent de mieux les caractériser et de prévoir, dans certains cas, leur survenance.

Ci-après, une liste des principaux événements classés dans la catégorie catastrophe naturelle :

- tremblement de terre, tsunami ;
- ouragan, typhon, cyclone ;
- tornade ;
- inondation ;
- tempête de neige ;
- éruption volcanique ;
- avalanche, glissement de terrain ;
- températures extrêmes (gel ou sécheresse).

1.2. Impact des catastrophes naturelles pour un assureur

Un assureur s'engage contractuellement à prendre en charge, contre le paiement d'une prime, le règlement de sinistres futurs. Afin de pouvoir honorer ses engagements, le risque sous-jacent au sinistre doit être évalué aussi précisément que possible.

Une compagnie d'assurance peut caractériser un risque via trois grandeurs : sa fréquence, sa sévérité et son ampleur géographique. Ces grandeurs doivent donc inévitablement être évaluées afin de pouvoir gérer au mieux le risque.

La fréquence d'occurrence Par définition, une catastrophe naturelle a une fréquence d'occurrence faible. Cela signifie qu'après plusieurs années sans sinistre, un assureur peut connaître une année marquée par une catastrophe naturelle d'intensité très forte pouvant lui être fatale. Par ailleurs, la fréquence de survenance de certaines catastrophes naturelles, comme les inondations ou la sécheresse, a tendance à s'accroître en raison du réchauffement climatique.

Dans le cas général, pour des sinistres non extrêmes, la probabilité de survenance de sinistre est évaluée par des méthodes statistiques à l'aide de l'historique des sinistres survenus dans le portefeuille de l'assureur. Cependant, ces méthodes ne sont pas adaptées aux catastrophes naturelles qui surviennent trop rarement pour avoir un historique de sinistres significatif. Pour une prise en compte adéquate des catastrophes naturelles, les méthodes statistiques doivent être prolongées au moyen de méthodes scientifiques afin de prendre en compte les caractéristiques du phénomène physique.

La localisation géographique. La localisation géographique est un facteur crucial dans la survenance de catastrophes naturelles. En effet, certaines régions « à risque » sont plus touchées

que d'autres en raison de leur localisation. Par exemple, le risque de séismes, particulièrement important au Japon, est quasiment nul en France.

La Figure I-1 représente la fréquence de survenance de catastrophes naturelles par région et souligne cette forte dépendance géographique en France.

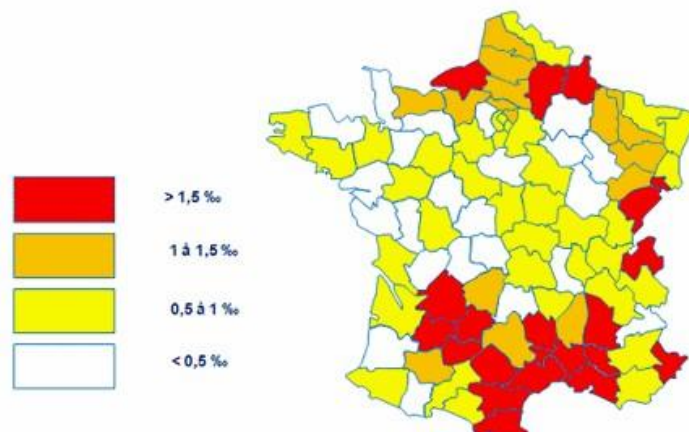


Figure I-1 Fréquence de sinistres en France entre 1998 et 2006
Source : FFSA

L'intensité de l'évènement. Lorsqu'elle survient, une catastrophe naturelle peut provoquer d'importants dommages et avoir une incidence sur la solvabilité des assureurs. Par exemple, suite au cyclone Andrew qui a ravagé la Floride en 1992, onze assureurs ont fait faillite par sous-estimation du risque. Ainsi, les assureurs sont de plus en plus incités à se doter de modèles spécifiques afin d'évaluer les risques auxquels ils sont exposés dans le but de mieux tarifer les primes d'assurance d'une part et de bien dimensionner les cessions en réassurance d'autre part.

1.3. Le risque gel

1.3.1. Définition d'un évènement gel

Par définition, le « gel » correspond à la transformation d'un liquide en état solide par des températures inférieures au point de solidification de la matière, c'est-à-dire par des températures négatives.

Néanmoins, un évènement gel n'engendre pas nécessairement de pertes (dues au gel) pour un assureur. Ce que nous chercherons à capter dans cette étude est un évènement gel au sens assurantiel et donc un évènement déclenchant des pertes.

1.3.2. Le risque gel comme catastrophe naturelle

Bien que le gel engendre des sinistres moins médiatisés que d'autres catastrophes naturelles (tremblement de terre, cyclones,...), il peut entraîner des pertes importantes pour un assureur lorsqu'une longue période de froid est observée. Il convient de noter que la sinistralité liée à ce péril s'est aggravée ces dernières années en raison d'une moindre qualité des constructions récentes sur l'isolation et des dalles moins épaisses. Le gel entraîne essentiellement des sinistres liés à :

- des ruptures de conduites d'eau qui peuvent nécessiter de lourds travaux de réparation ;
- des ruptures de structures liées à un sol gorgé d'eau qui se rétracte rapidement ;
- des pertes d'exploitation agricole dans le cas du gel de cultures.

La Figure I-2 représente les vagues de froid observées en France entre 1947 et 2012. Nous pouvons ainsi voir que la France a connu 21 vagues de froid depuis 1947. La survenance d'un tel évènement est donc faible historiquement.

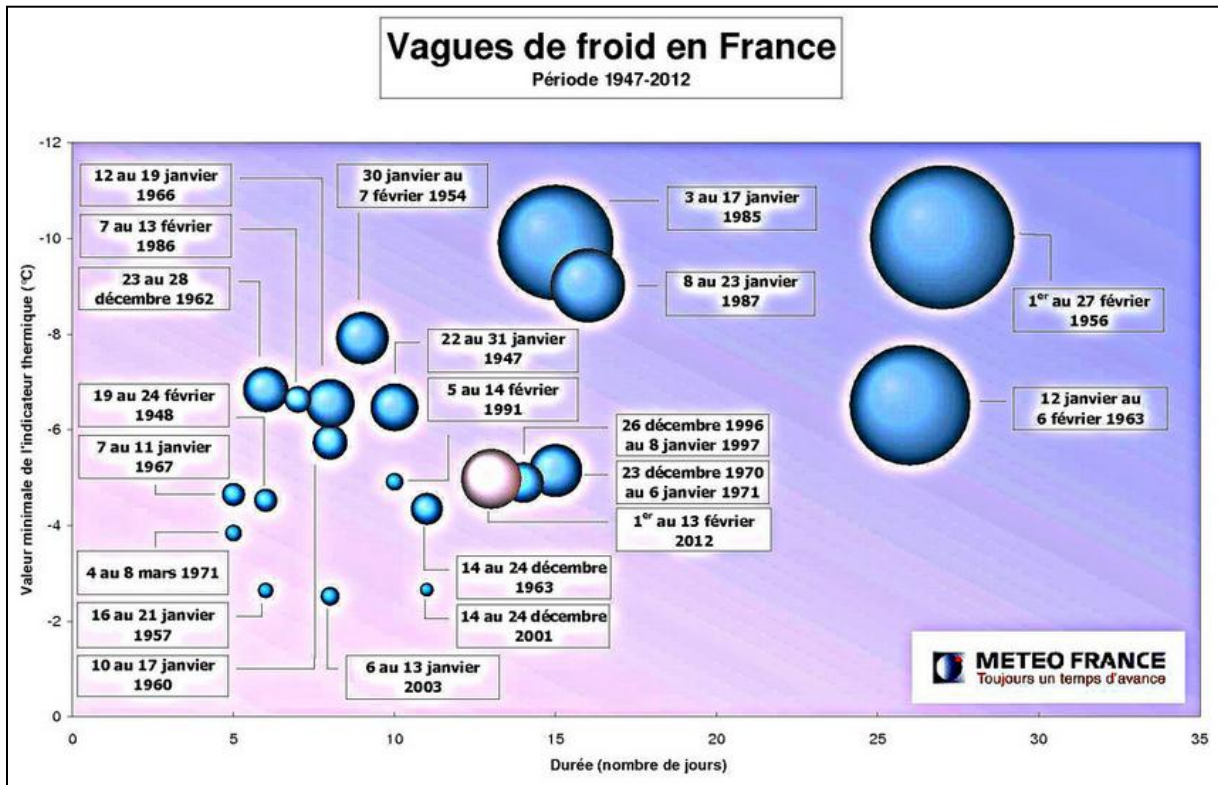


Figure I-2 Vagues de froid en France
Source : Météo France

Sur la Figure I-2, le diamètre des sphères symbolise l'intensité globale des vagues de froid, les sphères les plus grandes correspondant aux vagues de froid les plus sévères. Nous pouvons voir que les plus grandes vagues de froid se sont déroulées en 1956, 1963 et 1985, ont duré jusqu'à 30 jours et atteint la température de -32°C.

La vague de froid survenue en 2012 fait partie des six plus grosses vagues de froid observées depuis 1947. A ce jour, cette vague de froid est à l'origine du sinistre gel le plus significatif dans l'historique d'AXA France.

La Figure I-3 montre que le gel est un évènement qui génère de façon irrégulière des pertes importantes. En effet, des périodes de gel sont observées chaque hiver mais l'intensité particulièrement forte de l'évènement gel en 2012 justifie que cet évènement soit classé dans la catégorie catastrophe naturelle puisqu'il en vérifie les trois principales caractéristiques :

- montant de pertes irrégulier dépendant d'un phénomène physique ;
- évènement pouvant impacter plusieurs polices indépendantes ;
- forte dépendance géographique.

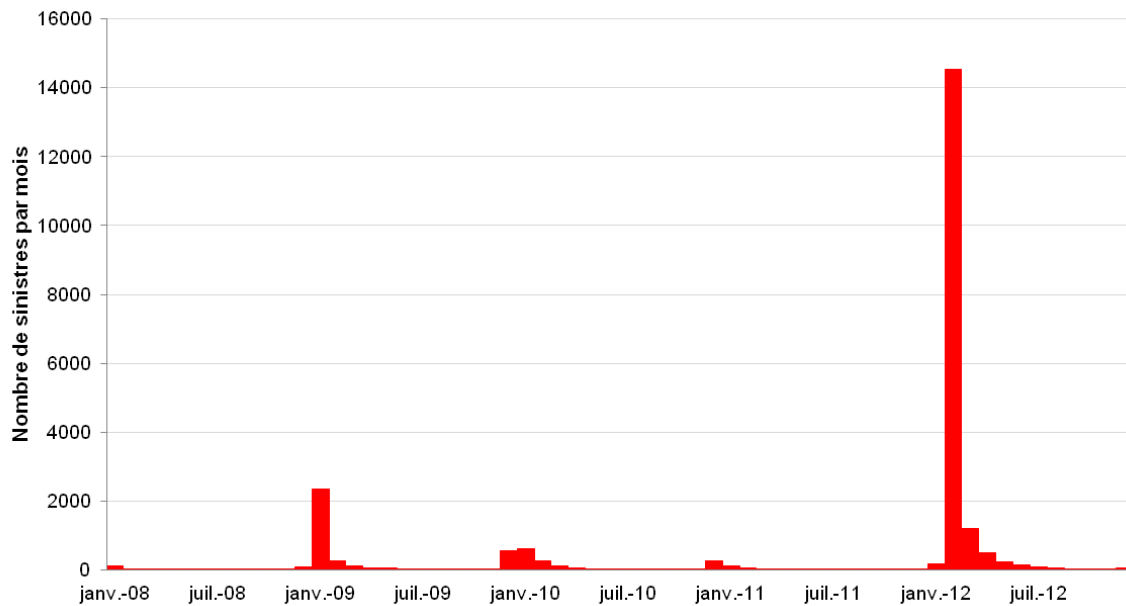


Figure I-3 Nombre de sinistres survenus entre 2008 et 2013 à cause du gel pour AXA

2. Intérêt de la modélisation du risque catastrophe

Les catastrophes naturelles constituent d'importants risques pour un assureur qui est fortement incité à les modéliser pour deux raisons principales : l'optimisation du capital réglementaire et l'optimisation de ses traités de réassurance.

2.1. Optimisation du capital réglementaire

La réforme Solvabilité II (SII) est une réforme réglementaire européenne dont l'un des objectifs est de mieux réglementer le montant de fonds propres minimal que doivent détenir des organismes d'assurances en fonction de leurs risques de bilan afin d'avoir une probabilité de défaut inférieure à 0.5% à horizon d'un an. Les risques d'un assureur se classent en quatre grandes catégories :

- risque financier ;
- risque de passif vie ;
- risque de passif non-vie ;
- risque systématique (ce dernier ne s'appliquant qu'aux 10 plus gros assureurs mondiaux).

Le risque de passif non-vie est le risque qui nous intéresse dans cette étude. Il se décompose en plusieurs composantes :

- le risque de prime lié à l'incertitude des polices souscrites ou renouvelées dans l'année à venir ;
- le risque de réserve lié à l'incertitude d'obligations existantes d'assurance ou de réassurance ;

- le risque de catastrophe lié à la survenance d'évènements d'origine naturelle ou humaine.

La directive SII prévoit le calcul des fonds propres réglementaires pour ce dernier type de risque via deux approches différentes : la formule standard et les modèles internes.

Approche 1 – Formule standard : L'assureur calcule ici l'impact sur son résultat de scénarios catastrophes fournis par le régulateur², ce qui lui permet de déduire le montant de capital qu'il doit détenir pour se couvrir contre le risque catastrophe étudié. Dans le cas où aucun scénario n'est fourni, une formule standard est appliquée sur la base des primes reçues. Ces deux méthodes peuvent être combinées, sachant que le résultat le plus pessimiste doit être retenu.

Approche 2 – Modèle interne : L'assureur peut également construire un modèle interne adapté à ses spécificités. Au sein d'AXA, plusieurs approches sont employées :

- l'utilisation de modèles tiers qui permettent de modéliser le phénomène physique, la vulnérabilité d'un portefeuille liée à ces phénomènes physiques et donc la distribution de risque du portefeuille en question ;
- le développement de modèles CAT (ou modèle catastrophe) propres à AXA ;
- La calibration de scénarios de manière plus raffinée que la formule standard.

Des trois approches, l'approche par le modèle CAT est la plus précise, car elle est la seule à refléter le profil de risque réel de l'assureur et à s'adapter à des changements de structure de portefeuille (géographique, type de risque assurés,...). Cependant, cette méthode étant spécifique à chaque assureur, la validation d'un modèle catastrophe fait l'objet d'un contrôle pointu de la part du régulateur.

Le péril gel n'étant pas considéré comme un péril majeur par Solvabilité 2³, aucun capital réglementaire n'est prévu par la directive dans la formule standard pour la couverture de ce risque. Cependant, comme nous l'avons introduit dans le paragraphe précédent, le risque gel peut engendrer des pertes importantes, voire être le péril majeur dans certains pays comme l'Irlande. Modéliser le gel permet à l'assureur de se doter d'un modèle interne et de connaître précisément la distribution des risques encourus. Il pourra ainsi en déduire le capital requis à la souscription de ce risque. Nous appliquerons dans ce mémoire la méthodologie utilisée pour construire un modèle catastrophe.

2.2. Tarification de traités de réassurance optimaux

2.2.1. Intérêt de la réassurance

Les catastrophes naturelles génèrent une charge et surtout une volatilité de bilan conséquente pour un assureur. Dans le but de se prémunir d'une trop grande volatilité, l'assureur transfère une partie de ses risques d'accumulation grâce aux mécanismes de réassurances (voir Figure I-4).

² L'Autorité de Contrôle Prudentiel et de Résolution en France

³ Seuls les périls tempête, inondation, tremblement de terre, grêle et affaissement de terrain sont pris en compte dans la formule standard.

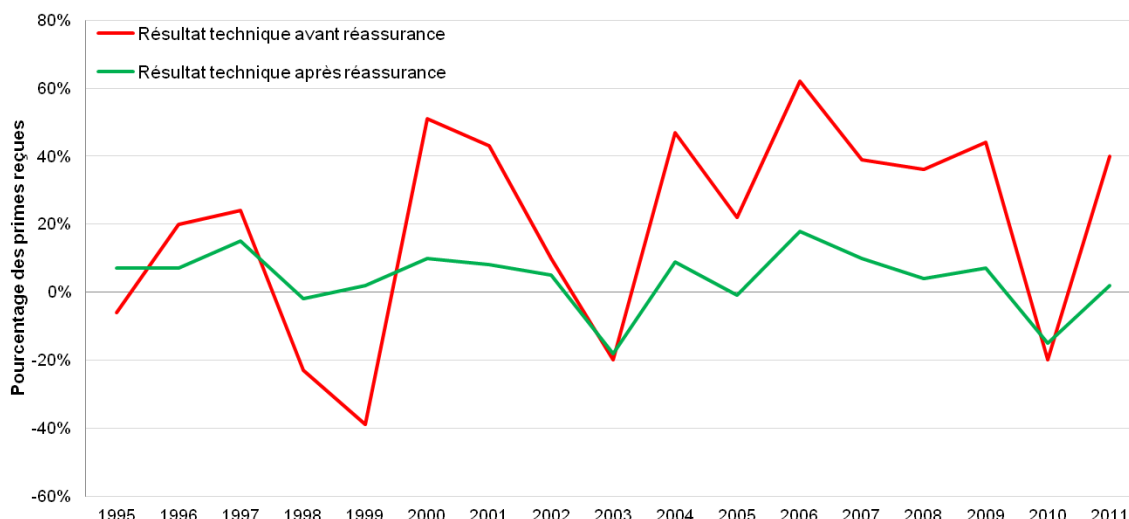


Figure I-4 Impact de la réassurance sur le ratio de sinistralité en France
Source : FFSA

2.2.2. Les traités de réassurance

Deux familles de traités de réassurance existent : la réassurance proportionnelle et la réassurance non proportionnelle.

La réassurance proportionnelle consiste à répartir les primes et les sinistres entre l'assureur et le réassureur selon un ratio défini de manière contractuelle. Ce type de traité n'est pas adapté à la couverture de catastrophes naturelles car lors de la survenance d'un événement extrême, l'assureur aura toujours à sa charge un ratio des sinistres qui pourrait lui être défavorable lorsque le montant total de sinistres est élevé.

La réassurance non proportionnelle consiste à transférer au réassureur les sinistres au-delà d'un montant spécifié (la priorité ou rétention) et jusqu'à un certain montant défini (la portée). La priorité est en général suffisamment élevée pour que le traité ne se déclenche qu'en cas d'évènement extrême. La réassurance non proportionnelle est particulièrement adaptée aux catastrophes naturelles.

La Figure I-5 illustre les mécanismes de ces deux familles de traités.

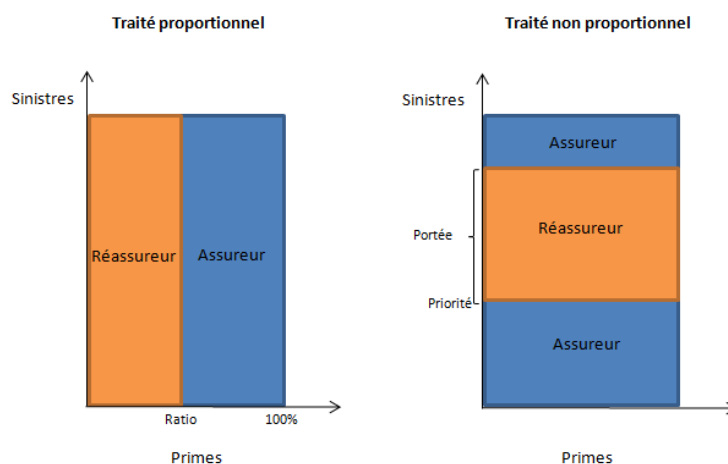


Figure I-5 Mécanismes de traités de réassurance

Le prix d'un traité de réassurance non proportionnel dépend directement des caractéristiques du contrat, à savoir la priorité et la portée. La modélisation de l'évènement gel permettra à l'assureur de connaître les risques qu'il encourt et de déterminer ainsi de façon optimale la valeur de ces deux éléments en fonction du profil de risque qu'il souhaite adopter.

3. La modélisation catastrophe

3.1. Définition d'un modèle catastrophe

Généralement, pour évaluer la distribution des pertes liées à un risque spécifique en assurance non vie, l'approche « fréquence-coûts » est utilisée. Elle se base directement sur les pertes historiques de l'assureur pour modéliser une loi de fréquence (par exemple avec une loi de Poisson) et une loi de coûts de sinistres (par exemple avec une loi de Pareto). Cependant, cette approche n'est pas adaptée aux risques catastrophes puisqu'elle repose, en général, sur l'hypothèse que la distribution des événements passés est représentative de la distribution des événements futurs. Or, les catastrophes naturelles sont par définition des événements extrêmes et rares, dépendants de phénomènes physiques, qui ne peuvent pas être modélisés par l'approche statistique traditionnelle. Ainsi, dans les régions historiquement peu impactées par des événements catastrophes, le risque induit par les périls naturels sera sous-estimé avec une approche purement statistique.

Afin d'estimer plus précisément les risques catastrophes, il est donc nécessaire d'utiliser des approches par exposition plutôt que par historique. Pour ce faire, le secteur de l'assurance a beaucoup investi ces dernières années dans le développement de modèles catastrophes. Ce sont des modèles qui combinent la représentation mathématique de la survenance des sinistres avec les caractéristiques scientifiques du risque assuré pour générer une distribution de pertes.

Les modèles catastrophes sont devenus indispensables car ils permettent aux assureurs de :

- tarifier les contrats de réassurance pour optimiser le transfert de risque à un réassureur ;
- contrôler et diversifier les risques ;
- estimer les provisions à constituer pour faire face à une perte ;
- minimiser le capital requis par Solvabilité 2 ;
- anticiper les catastrophes naturelles et prévenir les risques pour augmenter la résilience.

Un modèle catastrophe est structuré en trois modules indépendants comme le montre la Figure I-6 : le module Aléa, le module Vulnérabilité et le module Financier. Ces différents modules sont détaillés dans les sections suivantes.

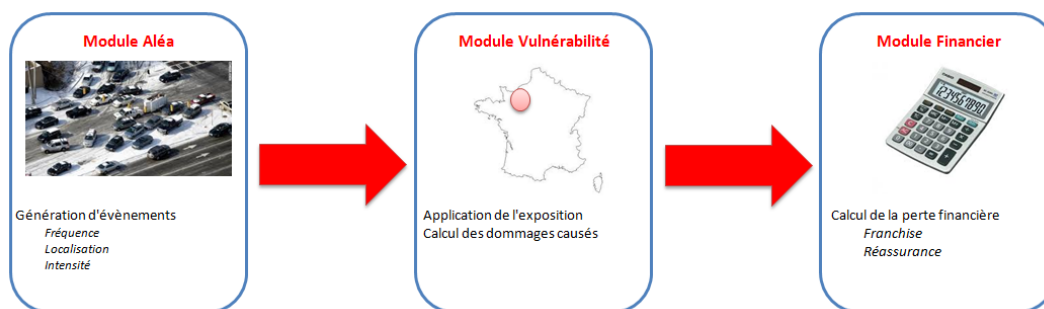


Figure I-6 Structure d'un modèle catastrophe

3.2. Le module Aléa

Le module Aléa a pour objectif de reproduire le phénomène physique sous-jacent au péril. Par exemple, il permet de simuler :

- la vitesse des vents pour les périls tempête, typhon ou cyclone ;
- la hauteur d'eau pour une inondation ;
- la propagation d'une onde pour un tremblement de terre ou une explosion.

Dans le cas du péril gel, nous chercherons à reproduire des phénomènes physiques liés à la température.

A partir d'évènements observés resimulés stochastiquement ou de modèles physiques complets, le module Aléa vise à produire un ensemble conséquent d'évènements physiquement réalistes et probabilisés permettant de représenter la multitude de « trajectoires » possibles du phénomène sous-jacent. Cependant, la constitution d'évènements réalistes est soumise à des contraintes de corrélation spatiale qui doivent être prises en compte dans la modélisation. Par exemple, le module Aléa fonctionne généralement par bassin pour représenter l'intégralité du risque tempête au niveau de l'Europe entière. Dans le cas du gel, si un évènement est observé dans une région, il y a de fortes probabilités pour que les régions voisines soient aussi affectées.

Le module Aléa est un élément central dans tout modèle catastrophe. Dans ce mémoire, nous nous focaliserons principalement sur la construction de ce module qui regroupe les principales difficultés de la modélisation catastrophe. Par sa complexité, le module Aléa nécessite de procéder à des choix de modélisation qui seront détaillés au fur et à mesure de l'étude.

3.3. Le module Vulnérabilité

Le module Aléa ayant constitué un dictionnaire conséquent d'évènements physiques, le module Vulnérabilité permet ensuite de quantifier la sinistralité de chaque évènement. Connaissant les caractéristiques physiques de la structure assurée, la perte peut être estimée. Lors de la survenance du risque, la perte est estimée police par police puis agrégée pour obtenir une perte au niveau d'un portefeuille. Cette démarche est appliquée à chaque évènement probabilisé. Pour ce faire, deux étapes doivent être réalisées :

- une modélisation du portefeuille ;
- l'application des courbes de vulnérabilité.

3.3.1. L'exposition du portefeuille

L'exposition du portefeuille représente les caractéristiques des risques souscrits par l'assureur. On observe que pour une même intensité, des dommages très divers peuvent être occasionnés suivant le type de bien assuré. Par exemple, un immeuble aura généralement une vulnérabilité moindre qu'une maison et sera donc plus résistant. Par ailleurs ces spécificités sont aussi géographiques. Par exemple, au Japon, pays particulièrement exposé aux tremblements de terre, les constructions sont plus résilientes aux tremblements de terre qu'en Turquie qui est aussi une zone à risque.

Dans la pratique, on définit cinq caractéristiques inhérentes aux biens assurés :

- **La garantie du contrat** qui peut être de trois types : Bâtiment, Contenu ou Perte d'exploitation. La garantie Bâtiment assure la structure du bâtiment comme les murs ou les fenêtres. La garantie Contenu couvre des objets comme des ordinateurs ou des télévisions. Enfin, la garantie Perte d'exploitation en assurance professionnelle couvre, par exemple, les pertes financières liées à l'arrêt d'une activité ou les frais nécessaires au relogement si ce dernier n'est plus habitable.
- **Le type de bien assuré** représente l'utilisation du bâtiment assuré. En effet, des dommages très différents peuvent être causés suivant que le bâtiment assuré soit un musée, une usine ou un immeuble résidentiel.
- **La structure du bien assuré** représente le mode de construction du bien assuré (matériaux utilisés, agencement,...), ce qui entraîne une résistance du bien plus ou moins importante aux périls naturels.
- **La hauteur du bâtiment** : le nombre d'étages d'un bâtiment est déterminant en particulier pour les inondations ou les tremblements de terre.
- **L'année de construction** car les normes de constructions évoluent et imposent des standards aux constructions neuves.

Toutes ces caractéristiques ont un impact sur les dommages causés par un évènement gel, bien que certaines caractéristiques soient plus importantes que d'autres.

3.3.2. Les courbes de vulnérabilité

Afin de quantifier l'impact d'un évènement sur une police assurée, on définit la vulnérabilité comme un ratio de pertes sur la valeur assurée pour un certain type de bien. Dans un modèle catastrophe, le module Vulnérabilité est constitué de plusieurs courbes de vulnérabilité qui permettent de prendre en compte à la fois le type de bien assuré et l'intensité de l'évènement. Ainsi, une courbe de vulnérabilité permet, pour un type de bien, d'exprimer le taux de destruction causé par le péril en fonction de son intensité.

Pour chaque combinaison des cinq caractéristiques énoncées ci-dessus, une courbe de vulnérabilité doit être construite. Pour cela, deux méthodes sont possibles :

- **La méthode statistique** : des enquêtes sont menées afin de recueillir le plus grand nombre de sinistres possible survenus suite à un évènement. Ces sinistres sont ensuite mis en relation avec l'intensité des évènements associés par une régression statistique. Cela permet ainsi de calibrer une courbe de vulnérabilité. Cette méthode nécessite un large historique des sinistres survenus ainsi que l'historique des évènements. Il faut noter que les données de sinistre fournissent rarement des informations détaillées

concernant les caractéristiques des biens assurés, ce qui complexifie la construction de courbes pour chaque combinaison des caractéristiques étudiées.

- **La méthode physique** : des tests sont effectués pour construire une courbe de vulnérabilité élémentaire en testant l'impact de l'intensité du risque étudié sur un bâtiment spécifique. En général, la physique du solide est utilisée pour étudier la résistance des matériaux et déterminer ainsi le taux de destruction en fonction de la variable utilisée. Ensuite, ces courbes sont recombinaison pour chaque combinaison des caractéristiques énoncées. Cette méthode a pour avantage de ne nécessiter aucune donnée historique et de fournir un résultat dans tous les cas. Cependant, la mise œuvre de cette méthode est lourde et peut entraîner des biais importants suivant le bâtiment utilisé pour la réalisation des tests.

La différence observée entre les méthodes peut conduire à des résultats très différents, ce qui entraîne souvent la nécessité de recalibrer les courbes obtenues. De manière générale, la méthode statistique est utilisée lorsqu'un large historique des sinistres survenus suite au péril étudié est disponible.

La Figure I-7 représente un exemple de courbe de vulnérabilité pour deux types de biens assurés. Comme attendu, le pourcentage de destruction est bien croissant avec l'intensité de l'évènement observé. De plus, ces courbes sont différentes suivant le type de biens assuré, ce qui justifie l'utilisation de plusieurs courbes de vulnérabilité dans le modèle.

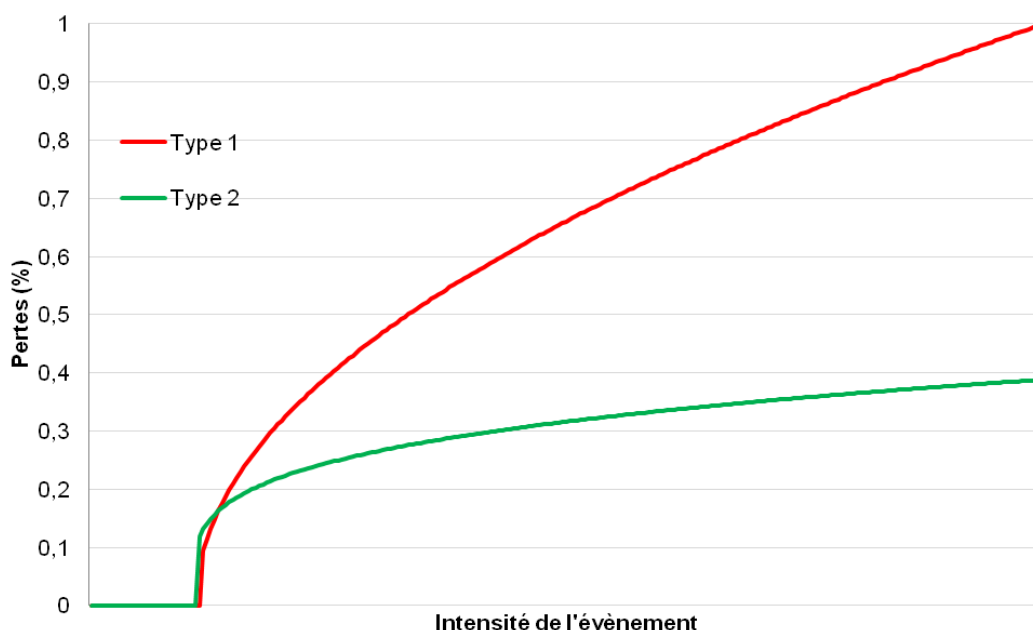


Figure I-7 Exemples de courbes de vulnérabilité

3.4. Le module Financier

Ce dernier module a pour objectif de déterminer les pertes finales de l'assureur nettes des conditions assurantielles et ainsi, de déduire la perte à la charge de l'assureur. Pour chaque scénario, le module Financier détermine la perte finale de l'assureur après application des caractéristiques associées à chaque contrat souscrit, des conditions de réassurances choisies par l'assureur et la part de la coassurance. Ces caractéristiques sont donc :

- **Les franchises et plafonds** qui sont indiqués dans chaque contrat souscrit : ils déterminent la part du sinistre laissée à la charge de l'assuré. L'assureur a donc à sa charge les sinistres compris entre la franchise et le plafond. L'assuré doit régler le montant restant.
- **La part des coassureurs** dans chaque contrat, qui représente un pourcentage des pertes générées.
- **L'application de facultatives** : un assureur peut souscrire quelques polices qui, par leur taille disproportionnée par rapport aux autres polices, pourraient à elles seules mettre à risque tout le portefeuille (par exemple les grands comptes, grands chantiers de construction, ...). L'assureur peut alors céder une partie du risque spécifiquement sur ces polices grâce à des couvertures facultatives par opposition aux traités qui s'appliquent à l'intégralité du portefeuille.
- **Les conditions de réassurance** : à cause de leurs natures extrêmes, les pertes causées par les catastrophes naturelles ne peuvent pas être supportées seulement par les assureurs dans la majeure partie des cas. Ces derniers font appel à des réassureurs pour souscrire des traités de réassurance. Il existe plusieurs type de traités, les plus communs étant les traités en XS (ou traités non proportionnels) qui permettent de transférer au réassureur les sinistres au-delà d'un montant spécifié (la priorité) et jusqu'à un certain montant défini (la portée) comme le montre la Figure I-8. Ce type de traité est subdivisé entre le traité par risque (perte subie sur une police ou causée par une police) et le traité par événement (consécutif à une accumulation de pertes sur plusieurs polices causées par un événement). Dans le cas d'un traité XS par événement, lorsqu'un événement spécifique se produit (gel, inondation,...), le traité se déclenche et le réassureur prendra à sa charge les règlements de sinistres causés par l'événement dans la limite de la portée et de la priorité spécifiées. En présence de deux traités (par risque et par événement), le traité par risque jouera en premier. Il convient de noter que le traité par événement est le mieux adapté dans la gestion du risque catastrophe.

Ces caractéristiques permettent donc de diminuer de façon significative la charge finale de l'assureur et doivent être modélisées dans un modèle catastrophe.

Le Figure I-8 résume les principales caractéristiques et montre comment elles contribuent à diminuer les pertes de l'assureur. On se place dans le cas d'un contrat présentant une franchise et un plafond, sur lequel une partie est cédée à un coassureur et qui est réassuré avec un traité de type XS avec une priorité et une portée définies.

Ainsi, les parties grises représentent la charge supportée par l'assuré via l'application de la franchise et du plafond. La partie verte correspond à la part des coassureurs impactés dans le contrat. La part bleue correspond à l'application d'un traité de réassurance avec une portée et une priorité définies. Finalement, la part que l'assureur devra payer après application de l'ensemble des caractéristiques associées au contrat est représentée en rouge.

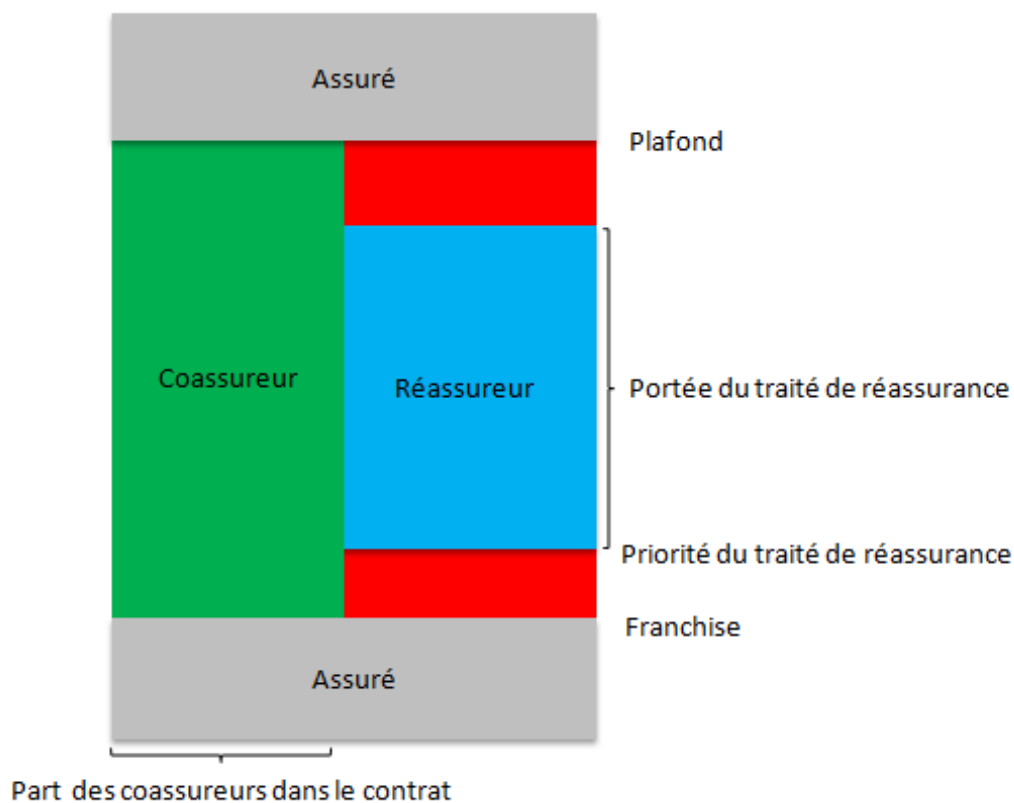


Figure I-8 Module Financier

3.5. Résultats d'un modèle catastrophe

Par définition, chaque modèle comporte des incertitudes et la modélisation catastrophe ne fait pas exception à la règle. En effet, on ne sait pas si la série d'évènements générée dans le module Aléa est réellement représentative de la réalité future. De plus, compte tenu de la multiplicité des facteurs mis en jeu dans l'évaluation des pertes (localisation, période de survenance), le montant de perte observé dans un scénario simulé ne peut être la valeur certaine d'un évènement réel. Cependant, la moyenne de l'ensemble des sinistres modélisés donnera une bonne estimation du risque si les modules Aléa et Vulnérabilité ont été correctement modélisés.

L'objectif d'un modèle catastrophe est de mesurer l'exposition du portefeuille de l'assureur au péril étudié. Pour cela, deux types de résultats peuvent être générés en sortie d'un modèle catastrophe, en fonction de l'objectif de l'assureur.

Table d'évènements Tout d'abord, un modèle catastrophe génère une table d'évènements qui fournit une analyse détaillée des pertes pour chaque évènement du dictionnaire généré par le module aléa. Cela permet d'estimer les pertes associées au dictionnaire d'évènements simulés ainsi que l'incertitude concernant la perte.

Distribution de pertes En utilisant la table d'évènement, on construit la distribution des pertes de l'assureur causées par le péril étudié. Pour cela, on introduit la notion de période de retour utilisée en modélisation catastrophe. Une période de retour correspond à l'inverse de la fréquence statistique d'un évènement parmi un ensemble d'évènements (annuels) équiprobables. Elle doit donc être interprétée comme une probabilité statistique. Elle est exprimée en années et correspond à la probabilité d'occurrence d'un scénario. Par exemple, un évènement de période de retour de 200 ans a une probabilité de 0,5 % de se produire sur une année.

Pour définir cette distribution de pertes, deux mesures sont généralement utilisées en réassurance :

- **La courbe OEP** (*Occurrence Exceedance Probability*) : associe une période de retour (en années) à la perte maximale engendrée par un évènement sur une année. L'OEP caractérise donc la probabilité annuelle qu'un évènement unique entraîne des pertes supérieures à un certain montant.
- **La courbe AEP** (*Aggregate Exceedance Probability*) : associe une période de retour (en années) à la perte totale engendrée par l'ensemble des évènements sur une année. L'AEP caractérise donc la probabilité annuelle que l'ensemble des évènements sur une année entraîne des pertes supérieures à un certain montant. Ainsi, une courbe AEP est toujours au-dessus d'une courbe OEP.

La Figure I-9 donne un exemple de courbes OEP et AEP. Nous notons que la courbe AEP est bien toujours supérieure à la courbe OEP mais l'AEP se rapproche de l'OEP pour des périodes de retour élevées. En effet, à période de retour élevée, l'OEP caractérise le montant de pertes générées par un évènement extrême. Cet évènement comprendra en fait une grande proportion des pertes associées à l'ensemble des évènements survenus pendant l'année représentées par l'AEP.

Ces courbes sont principalement utilisées pour deux raisons : la courbe AEP permet de déterminer le capital réglementaire requis sous Solvabilité 2 pour la couverture du risque spécifié et la courbe OEP permet de quantifier la sévérité du sinistre maximal pour optimiser la structuration de réassurance.

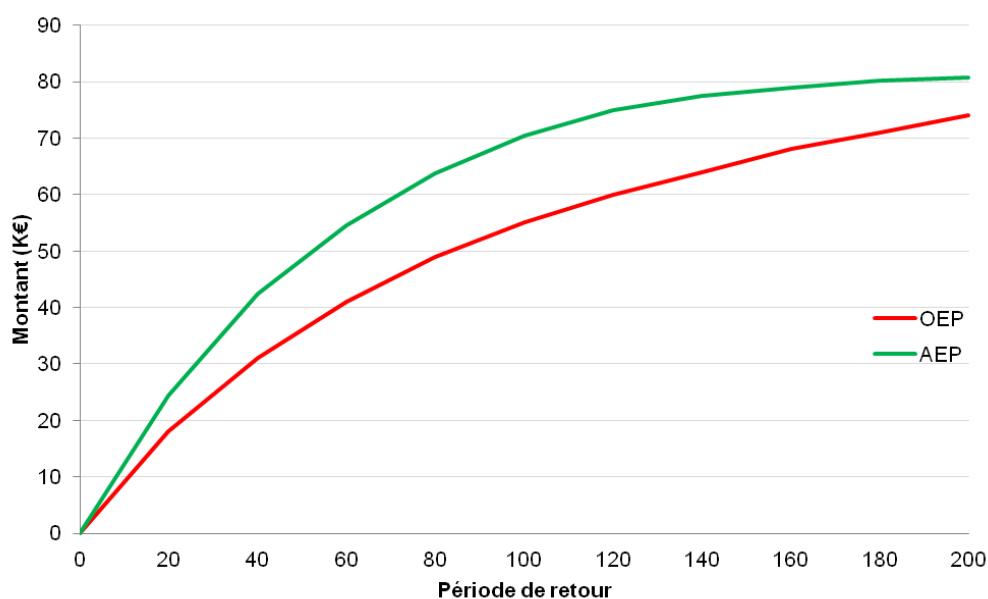


Figure I-9 Exemple de courbes OEP et AEP

Pour définir le montant de capital réglementaire, nous utilisons la courbe AEP car nous nous intéressons à la perte totale engendrée par l'ensemble des évènements survenus pendant une année. Ainsi, le capital réglementaire correspond au montant associé à la période de retour de 200 ans de la courbe AEP qui est donc de 80K€ sur la Figure I-9.

Pour structurer des traités de réassurance de catastrophes naturelles, la courbe OEP est privilégiée. En effet, c'est l'évènement générant la perte maximale annuelle (à période de retour fixée) qui peut être fatale pour un assureur. Afin de réduire au maximum ce risque, l'assureur souscrit des contrats de réassurance. Les courbes OEP permettent de déduire les caractéristiques des contrats de

réassurance (la priorité et la portée) à souscrire pour se protéger d'une perte maximale. Comme les contrats de réassurance souscrits pour les catastrophes naturelles sont généralement des traités XS par évènement, un traité qui permet de se protéger contre l'évènement générant la perte maximale permettra de se protéger aussi contre les autres évènements dont la perte associée est modélisée par la courbe AEP.

4. La modélisation catastrophe adaptée au risque gel

Le gel étant classé dans la catégorie catastrophe naturelle, nous allons décrire, dans ce mémoire, la construction d'un modèle catastrophe en tenant compte des spécificités de ce risque. Cela nous permettra de déduire les courbes OEP/AEP et donc une distribution des pertes d'AXA causées par le gel.

4.1. Caractérisation du risque gel

Afin de construire le module Aléa, il faut impérativement définir le facteur physique à l'origine d'un évènement gel. Ce facteur physique sera alors modélisé afin de générer un dictionnaire d'évènements. Cependant, comme l'apparition de gel ne génère pas forcément des sinistres, nous chercherons plutôt à modéliser les évènements de gel au sens assurantiel, c'est-à-dire les évènements qui causent des sinistres. Intuitivement, comme un évènement gel au sens physique est déclenché par des températures négatives, nous pouvons penser qu'un évènement gel au sens assurantiel est aussi directement lié aux températures. Nous montrerons dans la suite de l'étude (paragraphe III.1) que des températures extrêmement froides couplées à une chute brutale de température sont à l'origine de la formation de gel et de pertes pour un assureur.

Ainsi, la température devra être modélisée dans le module Aléa. Or, comme la caractérisation physique de la température est particulièrement complexe, nous adopterons une approche statistique pour la modéliser. Nous générerons ainsi plusieurs scénarios de température puis, nous en déduirons la probabilité de survenance d'un évènement gel. Cette approche est donc différente de la méthode traditionnelle utilisée dans les modèles catastrophe car elle ne s'appuie pas sur la modélisation physique du risque sous-jacent mais sur une modélisation statistique.

4.2. Présentation des données utilisées

4.2.1. Les données de températures

Une modélisation statistique cohérente de la température nécessite un grand nombre de données historiques. Pour cela, nous avons utilisé des données fournies par l'« *European Climate Assessment & Dataset* » (ECA&D)⁴. Cela nous a permis de disposer des données de températures minimales journalières réparties sur 250 points équidistants en France de 1950 jusqu'au 1^{er} Juillet 2013.

Pour cela, les relevés de température disponibles sur 106 stations de France (représentées sur la Figure I-10) ont été extrapolés sur une grille de pas de 50km à l'aide de modèles climatiques de propagation de la température. Ce mémoire ne détaillera pas la construction de ce modèle dont la documentation est disponible sur internet⁴.

⁴ Données disponibles sur internet à l'adresse suivante : <http://eca.knmi.nl/>

Pour vérifier la cohérence de ces données, nous avons calculé les différences entre les relevés historiques disponibles des 106 stations avec les points extrapolés les plus proches. Les résultats étant très faibles (inférieurs à 0,1°C), nous considérons les données valides.

Finalement, nous disposons de 250 séries temporelles de température (extrapolées) qui contiennent chacune 23 193 données (63 années).

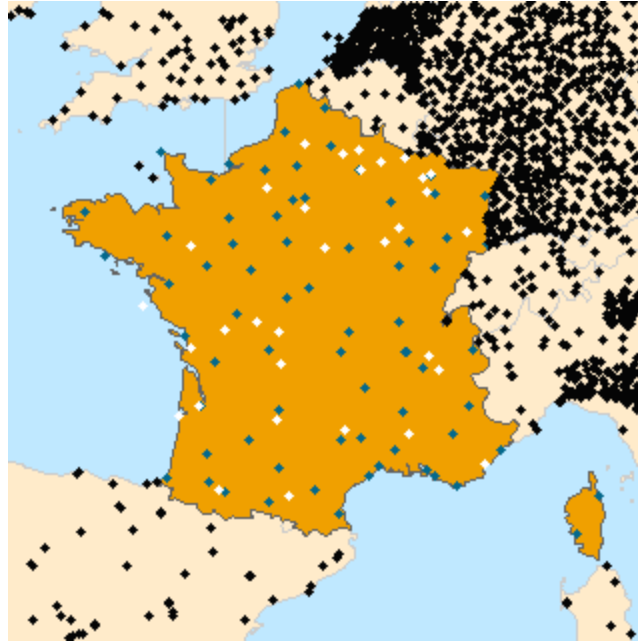


Figure I-10 Stations des relevés de température en France

4.2.2. Les données de sinistres d'AXA France

Afin de modéliser la vulnérabilité et de caractériser la survenance d'événements gel en France, nous nous sommes appuyées sur les données de sinistres d'AXA causés par le gel en France. Nous disposons donc d'un historique des sinistres survenus pendant cinq années (de 2008 à 2013) avec les variables suivantes associées à chaque sinistre :

- date d'occurrence du sinistre ;
- lieu d'occurrence ;
- montant du sinistre ;
- valeur du bien assuré ;
- branche d'activité ou *Line of Business* (LoB) : Agricole, Habitation, Immeuble ou Industrie.

II. La classification des régions

Le gel dépendant directement de la température observée, nous allons devoir simuler différents scénarios de température dans un premier temps pour en déduire la survenance d'évènements gel dans un second temps. Pour cela, nous disposons des données historiques (sur 23 193 jours) de températures journalières minimales relevées en 250 points localisés dans toute la France, ce qui représente un nombre important de données. Or, modéliser la température en chacun de ces 250 points serait fastidieux. Il est donc nécessaire de réduire cette quantité importante d'information dans l'espace et dans le temps pour obtenir des régions de France homogènes en température. Une fois les régions obtenues, nous pourrions caractériser leur température associée par la moyenne des températures des différents points géographiques qui appartiennent à ladite région. Notons par ailleurs que la classification des points géographiques en régions homogènes permettra également de tenir compte de la localisation géographique dans les modélisations qui seront effectuées par la suite.

Afin de regrouper les 250 localisations géographiques de la France, les méthodes statistiques d'analyse factorielle de données et/ou de classification sont appropriées. Dans certains cas, il est possible de combiner ces deux types de méthodes. En effet, certains auteurs recommandent d'appliquer une méthode d'analyse factorielle puis une classification sur les résultats de la première méthode. Cela a l'avantage d'affiner l'information et les résultats obtenus. Dans cette partie, nous étudierons deux approches différentes pour répondre au problème.

Dans la première approche, nous combinerons une méthode d'analyse factorielle de données avec une classification. Dans un premier temps, nous appliquerons la méthode Empirical Orthogonal Function (EOF) qui se rapproche d'une Analyse en Composantes Principales (ACP), mais pour des données spatio-temporelles. Cette méthode est donc adaptée à nos données de température prises en différentes localisations géographiques. Dans un second temps, nous effectuerons une méthode de classification sur les premières coordonnées principales obtenues avec la méthode EOF.

La seconde approche consistera à appliquer une méthode de classification ascendante hiérarchique avec l'utilisation de la distance Dynamic Time Warping (DTW). L'intérêt de la distance DTW est qu'elle permet de s'affranchir de la « rigidité » de la distance euclidienne appliquée à des séries temporelles. Néanmoins, les temps de calcul considérables nécessaires à l'implémentation de cette méthode nous ont conduites à privilégier l'approche précédente ; ce d'autant plus que les régions obtenues avec ces deux approches sont quasiment identiques à quelques points près sur les frontières des régions.

Dans la suite, nous présenterons les résultats obtenus avec ces deux approches. Nous nous attarderons particulièrement sur celle qui a été privilégiée, à savoir la classification obtenue suite à l'application de la méthode EOF.

1. L'Empirical Orthogonal Function (EOF)

La méthode EOF vise à déterminer un nouvel échantillon de variables qui permettent de capturer une partie importante de l'information observée sur les données à travers une combinaison linéaire des variables initiales. Ainsi présenté, la méthode EOF est proche de l'ACP. Nous verrons par la suite que les étapes d'implémentation ainsi que les principaux résultats sont identiques. Historiquement, l'ACP apparaît en 1933 avec les travaux de Hotelling alors que la méthode EOF apparaît en 1956 avec les travaux de Lorenz et est appliquée à des données atmosphériques. Elle est généralement utilisée par les météorologues. Cela nous a incitées à privilégier cette méthode. La

légère différence entre les méthodes EOF et ACP réside donc essentiellement dans le champ d'application.

1.1. Méthode EOF et quelques variantes

La simplicité et la facilité d'analyse associée à la méthode EOF expliquent son succès dans les domaines de la science liés à l'atmosphère. La principale critique qui peut être adressée à cette méthode est qu'elle n'est pas toujours interprétable physiquement à cause de l'orthogonalité dans l'espace et dans le temps des vecteurs propres⁵. Pour parer à cette limite, la méthode Rotated Empirical Orthogonal Function (REOF) a été développée qui consiste à trouver des facteurs de rotation qui permettent d'obtenir des structures simples et physiquement interprétables (HANNACHI, 2004).

La méthode EOF dérive sur des structures stationnaires, c'est-à-dire n'évoluant pas dans le temps. Les structures stationnaires obtenues prennent en compte la corrélation spatiale ou corrélation entre les localisations géographiques. C'est une représentation « statique » de l'espace sur une période de temps. Il en est de même pour la méthode REOF dont le « seul » avantage, dans certains cas, est l'interprétation physique. Afin de tenir compte de la corrélation temporelle, une méthode supplémentaire a été créée et est appelée Extended Empirical orthogonal Function (EEOF).

Nous ne nous intéressons qu'au caractère corrélé de nos points géographiques que nous voulons rassembler en un nombre réduit de régions. A cet effet, nous n'appliquerons que la méthode EOF et pas ses variantes

1.2. La matrice de données

Généralement, pour la méthode EOF, les données sont tridimensionnelles ou bidimensionnelles pour l'espace et unidimensionnelle dans le temps. Dans notre cas, les données sont une fonction du temps t , de la latitude θ et de la longitude ϕ et nous pouvons donc les écrire ainsi :

$$F_{ijk} = F(t_i, \theta_j, \phi_k)$$

Pour des besoins d'optimisation dans les calculs, nous nous ramenons à un tableau de données à deux dimensions où les coordonnées spatiales sont regroupées afin de ne former qu'un point géographique. La deuxième coordonnée est le temps. Cette nouvelle matrice de données est notée $X(t, s) = (x_{ij})_{i=1, \dots, n; j=1, \dots, p}$. Ainsi, comme nous avons 250 points géographiques avec 23 193 données de température, notre matrice de données transposées contient $p = 250$ colonnes et $n = 23\ 193$ lignes.

Cette matrice de données peut se réécrire ainsi :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Chaque colonne j de la matrice représente une évolution temporelle de la localisation x_j et chaque ligne i de la matrice représente une carte de la France à l'instant t_i . Les données doivent

⁵Cette assertion est vraie lorsqu'il s'agit d'étudier les modes normaux de phénomènes physiques (HANNACHI (2004), SIMMONS et al. (1983)). Par modes normaux, on entend des modes propres d'oscillation de système climatique.

ensuite être centrées à l'aide du vecteur des moyennes de température en chaque localisation $\bar{x} = (\bar{x}_1, \dots, \bar{x}_p)$ avec $\bar{x}_j = \frac{1}{n} \sum_{k=0}^n x_{kj}$. Nous posons :

$$X' = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

En pratique, lorsque $p \leq n$, la matrice de données utilisée dans le logiciel d'implémentation est la transposée de X' , notée X'^t ; ce qui est le cas pour nos données. Ainsi, notre matrice de données finales comporte automatiquement 250 lignes et 23 193 colonnes.

Dans le cadre de l'ACP, il est souvent commode de centrer et de réduire les variables, ce qui revient à utiliser une matrice diagonale des poids contenant l'inverse de la variance de chaque variable j . Dans le cas de la méthode EOF, la matrice de poids n'est employée que pour contrecarrer un artefact géométrique qui apparaît lorsque les données observées ne sont pas uniformément distribuées. Par exemple, on peut avoir des observations sur des carrés de résolution 5° latitude \times 5° longitude. Dans de tels cas, la meilleure manière d'attribuer des poids aux localisations est de considérer le cosinus de la latitude comme poids. La matrice des poids est donnée par $D_\theta = \text{diag}(\cos(\theta_1), \dots, \cos(\theta_p))$. La matrice des données avec les poids est donnée par :

$$X_\omega = X'D_\theta$$

Nos données ne sont pas sujettes à cet artefact géométrique donc la matrice de poids n'a pas été intégrée. Les observations sont donc équipondérées.

Par ailleurs, la matrice de covariance est définie par :

$$\Sigma = \frac{1}{n-1} (X')^t X'$$

1.3. Méthode de résolution

Comme dans le cas de l'ACP, l'objectif de l'EOF est de projeter les individus dans un sous-espace de projection F_k , de dimension k avec $k \leq p$, qui maximise l'inertie totale. La méthode EOF vise donc à trouver une combinaison linéaire des variables qui explique le maximum de la variance. Il s'agit de trouver un axe principal $a = (a_1, \dots, a_p)^t$ de \mathbb{R}^p telle que $X'a$ ait la variabilité maximale. Ainsi, on a :

$$\text{Max } a^T \Sigma a$$

$$\text{s. c. } a^T a = 1$$

La solution de cette équation est donnée par $\Sigma a = \lambda a$. La matrice Σ est symétrique définie positive. On obtient p vecteurs propres associés aux axes principaux. A chaque vecteur propre u_k est associé une valeur propre λ_k . Les valeurs propres sont triées suivant un ordre décroissant et le pourcentage d'inertie expliquée par le k -ième axe principal est donné par :

$$\frac{\lambda_k}{\sum_{k=1}^p \lambda_k}$$

L'espace de projection est donc le sous-espace engendré par les k vecteurs propres associés aux k premières (ou plus grandes) valeurs propres du obtenus avec le programme de maximisation ci-dessus.

1.4. Critères de choix des dimensions

Le but de la méthode EOF/ACP étant de réduire le nombre de dimensions pour la représentation des observations, il existe des critères théoriques et des critères empiriques permettant de déterminer ce nombre de dimensions.

Les critères théoriques ont pour objectif de vérifier que les valeurs propres sont significativement différentes entre elles à partir d'un certain rang. Le cas échéant, seules les premières valeurs propres sont conservées. Seulement, ce test repose sur une hypothèse forte que les individus sont distribués suivant une loi normale. Cette hypothèse forte est contraignante sur la loi utilisée par la statistique de test qui doit suivre une loi du χ^2 (SAPORTA, 2011). Dans la pratique, il est plus usuel d'utiliser les critères empiriques.

Le critère de Kaiser est le critère empirique le plus connu. Ce critère conduit à considérer tous les axes principaux dont la valeur propre est supérieur à 1 pour des données centrées réduites. Par ailleurs, SAPORTA (2011) interroge sur la « pertinence » du seuil. Est-ce significativement différent de considérer 1,1 comme seuil au lieu de 1 ? C'est ce qui l'a conduit à développer une formule pour le choix des valeurs propres :

$$\lambda > 1 + 2 \sqrt{\frac{p-1}{n-1}}$$

Un autre critère, plutôt visuel, autant valable dans le cas des variables centrées réduites que dans d'autres cas, est la méthode du coude. Il s'agit de détecter sur un diagramme de décroissance des valeurs propres l'existence d'un coude permettant de séparer les valeurs propres « utiles » de celles qui sont peu différentes et n'apportent pas d'information. C'est ce critère que nous utiliserons.

1.5. Analyse des résultats

Après application de la méthode EOF, on obtient une matrice de données contenant nos 250 points initiaux et 250 axes principaux en colonnes. Ces axes principaux sont deux à deux orthogonaux. Ils ont été obtenus par combinaison linéaire des variables temporelles initiales. Le premier axe contient plus d'information que le second, et le second plus que le troisième et ainsi de suite. A partir d'un certain moment, l'information contenue par un axe principal est très petite. Il est donc intéressant de se focaliser sur les axes principaux qui apportent le plus d'information pour l'analyse.

La Figure II-1 présente l'éboulis des valeurs propres ainsi que les pourcentages d'inerties associées aux dix premières composantes principales.

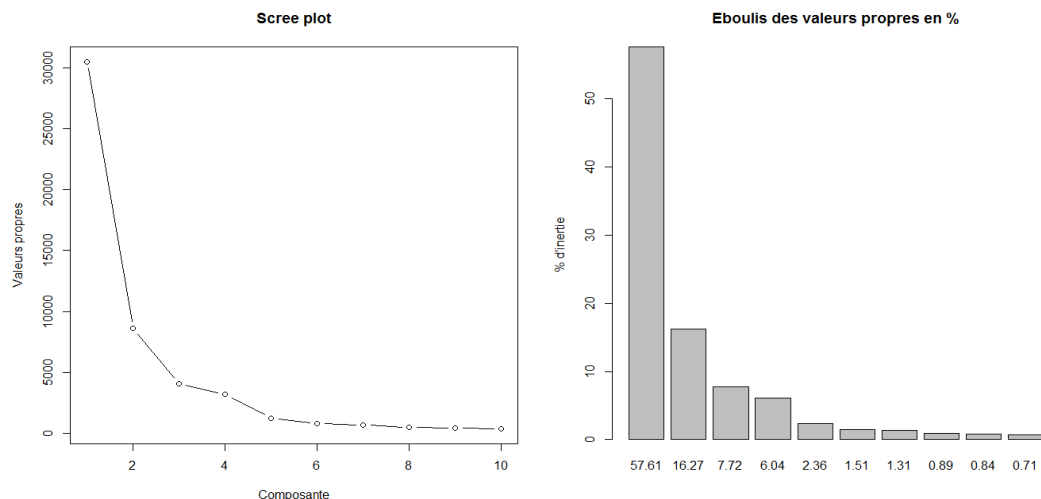


Figure II-1 Eboulis des valeurs propres

A la vue du premier graphique de la Figure II-1, la règle du coude revient à choisir les 4 premières valeurs propres. A l'aide du second graphique, on constate que l'inertie expliquée par les 4 premiers axes principaux est de 87,63 %.

Nous allons à présent décrire rapidement les grandes tendances associées à ces quatre axes principaux. Pour ce faire, nous allons représenter sur la France les points ayant une contribution significative à la formation de chaque axe principal considéré, c'est-à-dire les points les mieux représentés et ayant les coordonnées les plus élevées en valeurs absolues sur l'axe principal considéré. Par exemple, sur le premier axe principal, on considère les points dont les coordonnées sont inférieures à -275 ou supérieures à 275. L'intérêt de cette représentation est d'arriver à cerner rapidement quelle est la principale information apportée par chaque axe. Ainsi, la Figure II-2 permet de mettre en évidence différentes régions qui sont mises en opposition sur les différents axes.

Sur le premier axe factoriel, nous notons une opposition entre les massifs montagneux (points en rouge) et les régions ayant un climat méditerranéen et des hivers plutôt doux (points en jaune). Le second axe factoriel met clairement en évidence une opposition entre les points du Nord avec un climat océanique et les points du Sud ayant un climat plus méditerranéen. Sur le troisième axe factoriel, on a une opposition entre les régions avec un climat sous influence montagnarde (points en jaune) et les zones ayant plutôt un climat sous influence semi-continentale (points en rouge). Enfin, le quatrième axe factoriel fait apparaître une opposition entre les zones ayant un climat méditerranéen, notamment sur la Corse et la côte d'azur (points en rouge) et le massif central (points en jaune).

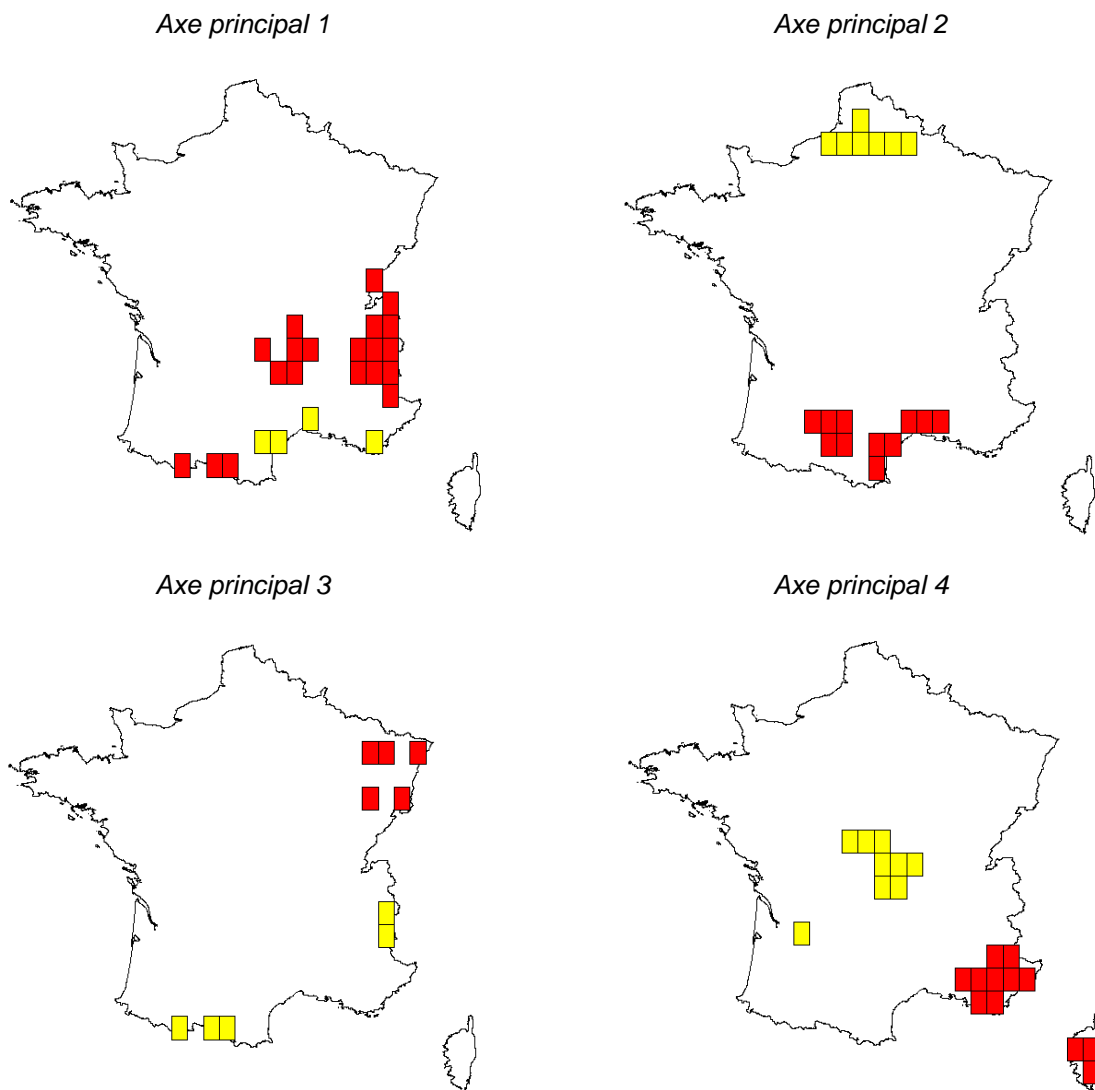


Figure II-2 Points géographiques les mieux représentés sur les 4 premiers axes principaux

Dans la suite, en implémentant la première approche combinant la méthode EOF avec une méthode de classification, nous effectuerons le regroupement des 250 points géographiques de la France en nous servant de leurs coordonnées sur les 4 premiers axes principaux obtenus avec la méthode EOF.

2. La classification

Les méthodes de classification permettent de regrouper des objets en fonction de la distance mesurée entre chacun d'eux. Deux notions importantes apparaissent ici à savoir la méthode de partitionnement et la mesure de distance (ou de dissimilarité) utilisées.

Quelle que soit la méthode de partitionnement ou la mesure de dissemblance choisie, le tableau de données en entrée de la classification est toujours un tableau $n \times n$ contenant les distances entre les n observations à classer.

2.1. Notion de distances et de similarités

En notant E l'ensemble des n objets à classer, une distance est une application de $E \times E$ dans \mathbb{R}^+ telle que :

$$\begin{cases} d(i, j) = d(j, i) \\ d(i, j) \geq 0 \\ d(i, j) = 0 \leftrightarrow i = j \\ d(i, j) \leq d(i, k) + d(k, j) \end{cases}$$

On parle de dissimilarité lorsque

$$\begin{cases} d(i, j) = d(j, i) \\ d(i, j) \geq 0 \\ d(i, i) = 0 \end{cases}$$

La similarité est une application s de $E \times E$ dans \mathbb{R}^+ telle que

$$\begin{cases} s(i, j) = s(j, i) \\ s(i, j) \geq 0 \\ s(i, i) \geq s(i, j) \end{cases}$$

Il existe de nombreux indices de similarités dans la littérature (Jaccard, Russel et Rao, Ochiaï). Les indices de dissimilarité peuvent être obtenus à partir des précédents indices en faisant la complémentation à 1. Dans le cadre de ce mémoire, nous nous sommes essentiellement intéressés à la distance euclidienne et la distance DTW en implémentant la classification.

2.2. Les méthodes de classification

Les méthodes de partitionnement ont initialement été développées pour des données statiques. Certaines ont néanmoins pu être adoptées pour des séries chronologiques. Les principales techniques⁶ utilisées pour la classification des séries temporelles sont :

- Les méthodes de partitionnement du type « nuées dynamiques » (*k-means*, *fuzzy c-means*).
- Les méthodes de classification hiérarchique (ascendante, descendante).

2.2.1. Les méthodes de partitionnement

Ces méthodes visent à regrouper rapidement un ensemble important d'objets en optimisant localement suivant un critère d'inertie. Suivant ce critère, chaque objet doit être affecté à une seule classe. L'idée générale de ces méthodes est de regrouper les n objets en k classes en maximisant l'inertie entre les classes, encore appelée inertie interclasse, et de minimiser l'inertie à l'intérieur de chaque classe. A titre d'exemple, lorsque l'on considère la distance euclidienne, l'inertie se définit comme la moyenne des carrés des distances au centre de gravité.

De manière pratique, l'algorithme *k-means* cherche à minimiser une fonction objective, laquelle est choisie comme étant la distance totale entre chaque point et le centre de la classe à laquelle il

⁶Il existe également des méthodes basées sur des modèles notamment celles utilisant le modèle bayésien ou les réseaux de neurones. Elles sont peu utilisées dans le domaine d'intérêt car elles ne sont pas toujours facilement interprétables.

appartient. Le nombre de classes doit donc être connu et/ou choisi a priori. Cet algorithme comporte deux principales étapes :

- Choisir arbitrairement un nombre k de classes initiales et calculer les coordonnées des centres correspondants ;
- Réallouer les objets dans les classes et mettre à jour des centres de classes.

L'algorithme alterne entre ces deux étapes jusqu'à ce que la valeur de la fonction objective ne puisse plus être réduite.

La minimisation de la fonction objective pour la méthode *k-means* est donnée par :

$$\text{Min } J_1(U, V) = \sum_{i=1}^c \mu_{ik} \|x_k - v_i\|^2$$

Avec :

- U l'ensemble des poids de chaque pattern k
- V l'ensemble des centres de classe
- X l'ensemble des objets

La *fuzzy means* est une généralisation de la méthode *k-means*. Il y a simplement une modification de la fonction objective : on attribue une puissance m aux poids des classes. Ainsi pour la méthode *fuzzy means*, il faut donc de minimiser :

$$\text{Min } J_m(U, V) = \sum_{i=1}^c \mu_{ik}^m \|x_k - v_i\|^2$$

Un avantage important de l'algorithme décrit est qu'il converge toujours. Néanmoins, un certain nombre d'inconvénients lui sont associés. Outre la nécessité de connaître le nombre de classes a priori, ces méthodes sont applicables uniquement pour des séries de longueur équivalente. Cela peut être contraignant dans le cas de séries temporelles. Par ailleurs, la structure finale dépend beaucoup du choix initial des centres de classes. Dans la pratique, il est nécessaire de répéter cet algorithme un nombre important fois pour obtenir un regroupement stable ou de dégager les formes fortes. Cela est très coûteux en calcul.

Cependant, cette méthode ne semble pas adaptée à notre problème puisque nous ne connaissons pas à l'avance le nombre de classes à considérer. De plus, afin de nous affranchir du problème de stabilité et de temps de calcul, nous optons pour une méthode de classification hiérarchique.

2.2.2. Les méthodes de classification hiérarchique

Il existe deux méthodes de classification hiérarchique :

- La méthode ascendante qui consiste à appliquer une stratégie de bas en haut pour l'algorithme.
- La méthode descendante a une logique inverse et consiste à appliquer une stratégie de haut en bas.

L'algorithme utilisé dans le cadre de la classification ascendante hiérarchique (la plus populaire) consiste à avoir initialement autant de classes que d'objets. L'étape suivante consiste à regrouper les objets présentant les plus fortes similarités : cela forme un nœud. A cette étape, il ne reste plus que $n - 1$ objets. Ce processus est itéré jusqu'à ce que le regroupement complet soit effectué. A chaque étape, toutes les combinaisons possibles de deux classes sont effectuées. La variance intraclasse est calculée pour le nouveau groupe formé et la combinaison présentant la variance intraclasse minimale est choisie. Le regroupement des objets dépend donc de la mesure de similarité/dissimilarité utilisée.

Les méthodes hiérarchiques consistent à regrouper les données en un arbre de classes ou dendrogramme. Un exemple de dendrogramme est fourni sur la Figure II-3:

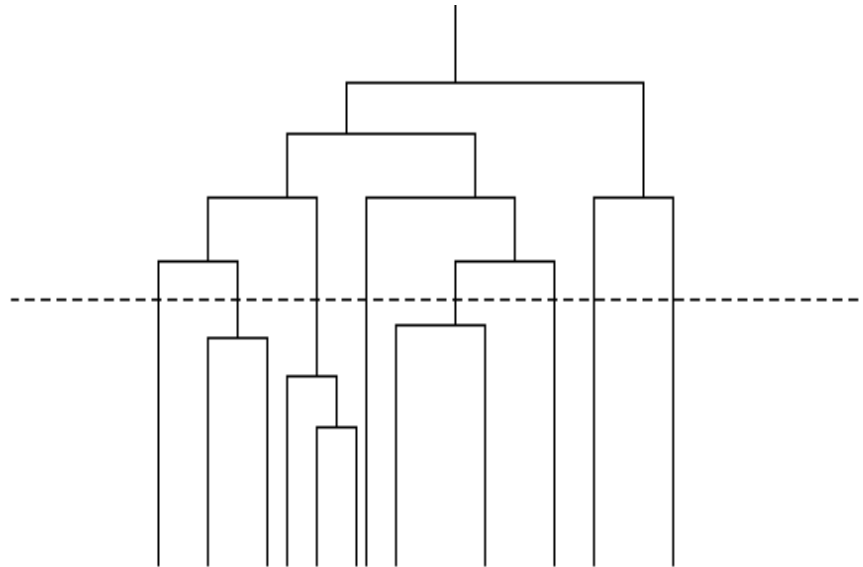


Figure II-3 Exemple d'un dendrogramme

A la racine de l'arbre (c'est-à-dire l'axe horizontal en bas du dendrogramme) sont représentés les individus à classer. Initialement, chaque individu représente une classe. Ensuite, les classes d'individus sont jointes deux à deux successivement en suivant l'algorithme de classification ascendante hiérarchique décrit précédemment, jusqu'à ce que l'on retrouve une seule classe avec tous les individus.

L'arbre est coupé horizontalement pour obtenir un nombre de classes précis. Par exemple, sur la Figure II-3, les traits en pointillés désignent la hauteur à laquelle le dendrogramme a été coupé. A ce niveau de coupure, on obtient 8 classes. La longueur d'une branche est proportionnelle à la perte d'inertie interclasse lors du regroupement des deux classes correspondantes. De manière générale, il faut donc chercher à couper l'arbre à l'endroit où l'augmentation de la longueur des branches est la plus grande.

L'avantage des méthodes de classification hiérarchique est qu'il n'est pas nécessaire de connaître *a priori* le nombre de classes. En effet, à partir du dendrogramme et du niveau de détail souhaité, l'utilisateur peut choisir son nombre de classes. L'inconvénient de cette méthode est qu'une fois le partitionnement effectué, il est impossible de le défaire.

Tout comme *k-means*, les méthodes de classification hiérarchique sont utilisées pour les séries de longueur identique. Néanmoins, les méthodes de classification hiérarchique peuvent également être utilisées sur des séries de tailles différentes à l'aide de distances appropriées telle que DTW.

3. Première approche : combinaison des méthodes EOF et CAH

À la suite de la méthode EOF, nous obtenons une matrice qui représente nos 250 localisations géographiques sur les 4 premiers axes principaux résultant de l'implémentation de la méthode EOF. Nous souhaitons à présent regrouper ces points en un nombre réduit de régions. Nous allons appliquer une Classification Ascendante Hiérarchique (CAH) à ces points en utilisant le critère de Ward.

3.1. La distance euclidienne

Nous avons choisi d'utiliser, pour cette approche, la distance euclidienne qui est une mesure de distance géométrique naturelle, très utilisée et très simple à mettre en œuvre.

Soit $X := (x_1, x_2, \dots, x_T)$ et $Y := (y_1, y_2, \dots, y_T)$ deux séries temporelles. La distance euclidienne entre X et Y est définie par :

$$d(X, Y) = \sqrt{\sum_{i=1}^T (x_i - y_i)^2}$$

Ainsi, la distance euclidienne entre deux séries temporelles est obtenue en additionnant la mesure la distance pour chaque pas de temps t entre x_t et y_t . Cependant, cette distance a pour inconvénient de ne pas prendre en compte les décalages temporels qui peuvent exister au sein des séries de mesures. Nous avons cependant décidé d'adopter cette distance pour des raisons de temps de calcul.

3.2. Le critère de Ward pour les distances euclidiennes

La méthode de Ward est la méthode de classification hiérarchique de référence pour les distances euclidiennes. Par ailleurs, cette méthode est complémentaire de l'ACP et repose sur un critère d'optimisation assez naturel (SAPORTA, 2011). Cela explique notre intérêt pour cette méthode à la suite de l'implémentation de l'EOF.

Le regroupement des objets se fait en rassemblant les objets qui font le moins varier l'inertie intraclasse ou maximise l'inertie interclasse. L'indice de dissimilarité entre deux classes correspond à la perte d'inertie interclasse résultant de leur regroupement.

Si on considère deux classes nommées A et B , g_A et g_B leurs centres de gravité respectifs, p_A et p_B leurs poids respectifs, le centre de gravité résultant de la réunion de ces deux classes, noté g_{AB} , s'obtient par la formule ci-après :

$$g_{AB} = \frac{p_A g_A + p_B g_B}{p_A + p_B}$$

L'inertie interclasse est la moyenne des carrés des distances des centres de classe au centre de gravité total. La variation d'inertie entre les classes A et B est donnée par :

$$\delta(A, B) = p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_A + p_B) d^2(g_{AB}, g)$$

Où g est le centre de gravité du nuage de points de E .

Nous obtenons alors :

$$\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$$

Ainsi, si l'on dispose d'un certain nombre d'objets à regrouper en deux classes notées A et B, la classification sera effectuée de manière à ce que $\delta(A, B)$ soit maximal.

Cette formule généralisée par Lance et Williams (SAPORTA, 2011) est donnée par :

$$\delta((A, B); C) = \frac{(p_A + p_C)\delta(A, C) + (p_B + p_C)\delta(B, C) - p_C\delta(A, B)}{p_A + p_B + p_C}$$

Il convient de noter que l'inertie totale du nuage est égale à la somme niveaux d'agrégation des différents nœuds du dendrogramme.

3.3. Les résultats de la CAH sur les données issues de la méthode EOF

Le dendrogramme issu de la CAH est présenté sur la Figure II-4.

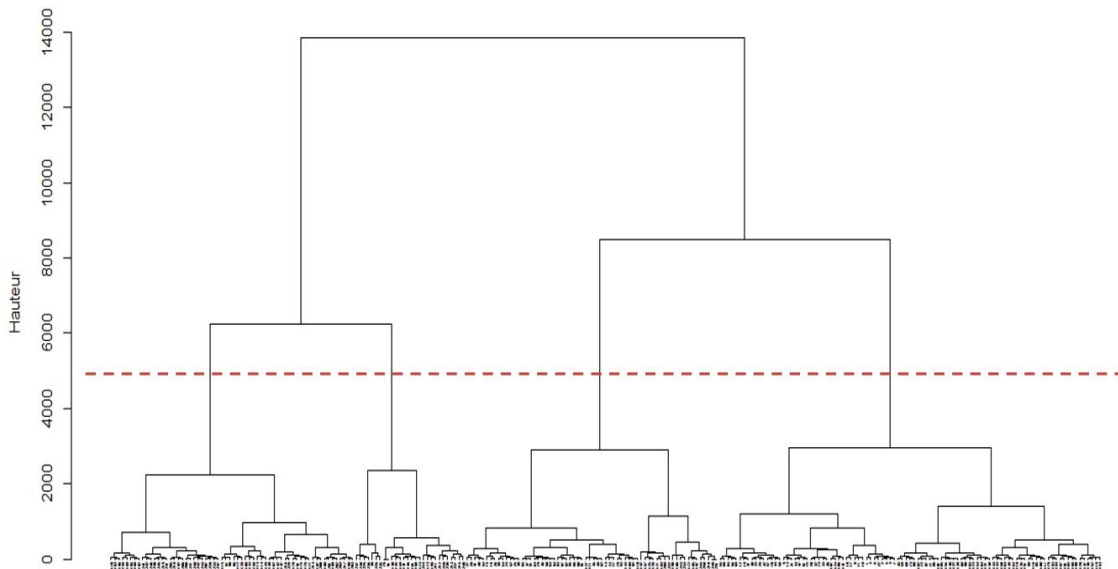


Figure II-4 Dendrogramme associé à la Classification Ascendante Hiérarchique (CAH) sur les coordonnées des 4 premiers axes principaux de l'EOF

Différentes classes peuvent être obtenues suivant le niveau où le dendrogramme est « coupé ». Afin d'obtenir des régions dans lesquelles la température est homogène, nous aimerions que les variances à l'intérieur des classes considérées soient très petites. Par exemple, il serait intéressant que l'écart type, au sein d'une région issue de la CAH, soit inférieur ou proche de 2° Celsius. Par ailleurs, nous souhaiterions avoir le plus petit nombre de classes possible.

Nous avons donc décidé de couper le dendrogramme à différentes hauteurs afin d'obtenir des nombres distincts de classes. Nous avons eu successivement de 3 à 6 classes et nous avons comparé l'évolution des variances intraclasse pour les classes considérées. Ces variances intraclasse sont présentées sur le Tableau II-1.

Classes	Variance intraclasse	Ecart Type	Nombre points par classe
3	1,990	1,411	97
	2,814	1,677	64
	5,507	2,347	90
4	1,990	1,411	97
	2,814	1,677	64
	4,707	2,170	28
	2,286	1,512	62
5	1,567	1,252	45
	2,814	1,677	64
	1,161	1,078	52
	4,707	2,170	28
	2,286	1,512	62
6	1,567	1,252	45
	1,302	1,141	44
	1,161	1,078	52
	4,707	2,170	28
	2,286	1,512	62
	2,950	1,718	20

Tableau II-1 Variances intraclasse pour différents nombres de classes

Nous constatons, sur le Tableau II-1, qu'en débutant avec trois classes, on obtient une seule classe dont l'écart type excède 2°C. En effet, avec trois classes, la variance intraclasse observée sur la dernière classe est de 5,5 °C². Cette classe est ensuite divisée en deux par une autre coupure du dendogramme pour obtenir quatre classes. De manière générale, en regardant le dendogramme, il est facile de constater que pour avoir une classe supplémentaire (avec un nombre maximum de classes fixé à six ici), une classe devra être divisée en deux. Cela se perçoit également visuellement sur les graphiques de la Figure II-5.

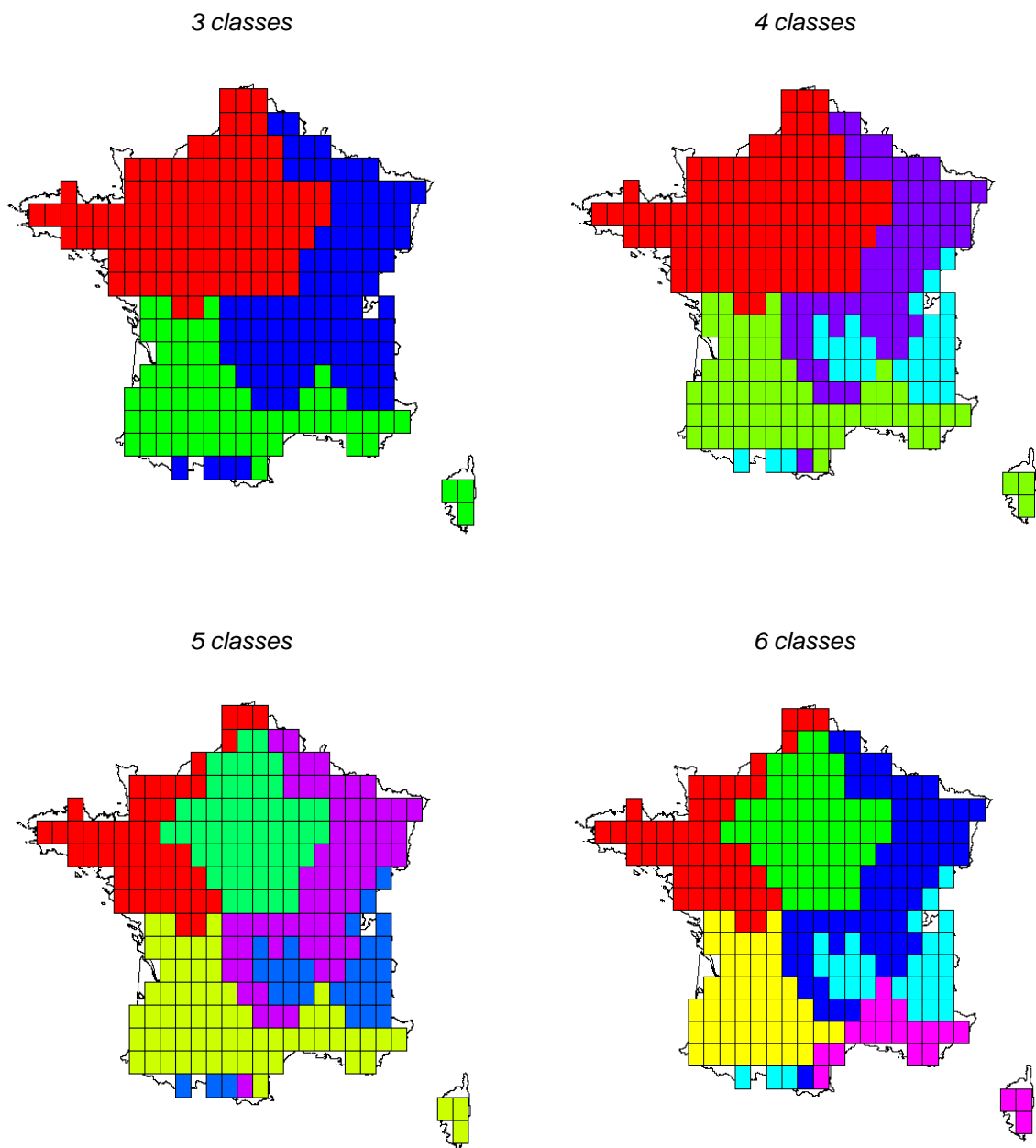


Figure II-5 Graphiques obtenus avec la CAH pour différents nombres de classes

Nous sélectionnons donc quatre classes par mesure de parcimonie et au regard de la cohérence de l'interprétation géographique. De plus, en analysant le dendrogramme de la Figure II-4, un regroupement en quatre classes semble être un choix convenable si on suit la règle explicitée plus haut sur la manière de « couper » l'arbre.

Finalement, les différentes régions sélectionnées sont représentées sur la Figure II-6.

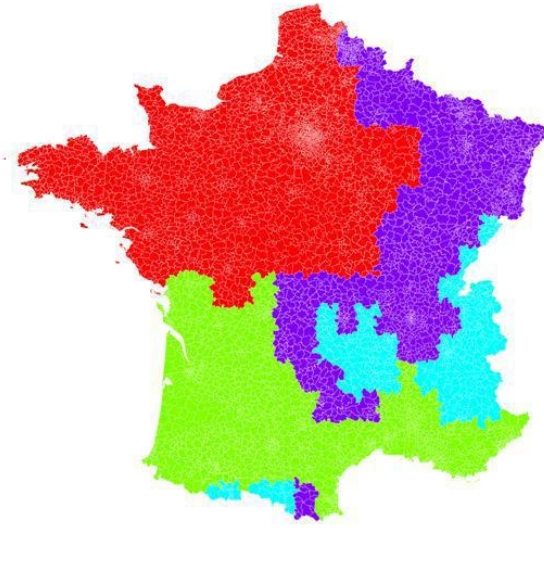


Figure II-6 Segmentation de la France en 4 régions (résultats de la CAH à la suite de l'EOF)

La première classe, en vert, regroupe une partie du Sud de la France. Elle est caractérisée par des villes ayant un climat tempéré avec des hivers plutôt doux. La zone au Sud-Ouest de la France est caractérisée par un climat océanique aquitain. On y retrouve des villes comme Bordeaux ou Limoges. Le reste des localisations géographiques dans cette 1^{ère} région est caractérisé par un climat méditerranéen. On y retrouve des villes telles que Nice, Marseille, Perpignan, Toulon et toute la Corse.

La deuxième classe, en rouge, englobe le Nord-Ouest de la France et prend en compte les zones sous influence du climat océanique. On y retrouve des villes comme Paris, Lille, Caen, Rennes.

La troisième classe, en violet, regroupe le Nord-Est de la France et une partie du Centre. C'est la zone sous influence d'un climat semi-continentale. On y retrouve des villes telles que Strasbourg, Nancy, Dijon. Elle également un point au Sud-Est de la France.

La quatrième classe, en bleu cyan, caractérise des zones sous influence montagnarde. On y retrouve des villes telles que Saint-Etienne et Grenoble.

Nous trouvons en annexe A les différentes régions de France en fonction de leur climat et nous pouvons constater que nous aboutissons à des régions semblables avec notre méthode, ce qui nous conforte dans ce choix de segmentation des régions.

Dans la suite, nous nommons les classes issues de la segmentation par les attributs ci-après :

- Région 1 : région Sud (en vert)
- Région 2 : Région Nord-Ouest (en rouge)
- Région 3 : Région Nord-Est (violet)
- Région 4 : Région Montagneuse (bleu cyan)

4. Deuxième approche : CAH avec utilisation de la distance DTW

Dans cette deuxième approche, nous souhaitons prendre en compte les déformations temporelles des séries de température via une mesure de distance spécifique : la distance DTW. Nous appliquons donc directement une Classification Ascendante Hiérarchique avec la distance DTW à nos 250 localisations géographiques sur leurs coordonnées initiales (latitude*longitude et temps). Notons qu'il aurait également été possible d'effectuer une EOF puis une CAH avec DTW. Cette approche n'a pu être implémentée en raison de temps de calcul trop importants.

4.1. La distance DTW

4.1.1. DTW classique

La distance DTW est une distance spécialement adaptée aux séries temporelles. En effet, nous observons souvent des déformations temporelles dans des séries de température dus, par exemple, au déplacement d'anticyclones. L'objectif de DTW est de trouver un alignement optimal entre deux séries temporelles prenant en compte des décalages temporels éventuels.

Soit $X := (x_1, x_2, \dots, x_N)$ et $Y := (y_1, y_2, \dots, y_M)$ deux séries temporelles.

Nous introduisons $c : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+$ une mesure de distance locale, c pourrait être par exemple la distance euclidienne. Dans ce cas, c serait définie de la façon suivante :

$$c : (x, y) \rightarrow \sqrt{(x - y)^2}$$

Soit C la matrice coût de taille $(N \times M)$ telle que $\forall (n, m) \in [1 : N] \times [1 : M]$,

$$C(n, m) = c(x_n, y_m)$$

L'objectif de DTW est de trouver un alignement entre X et Y qui minimise le coût total.

Un alignement est une suite $p = (p_1, p_2, \dots, p_L)$ avec $p_l = (n_l, m_l) \in [1 : N] \times [1 : M]$ satisfaisant les trois conditions suivantes :

- (i) $p_1 = (1, 1)$ et $p_L = (N, M)$
- (ii) $n_1 \leq n_2 \leq \dots \leq n_L$ et $m_1 \leq m_2 \leq \dots \leq m_L$
- (iii) $p_{l+1} - p_l \in \{(0, 1), (1, 0), (1, 1)\}$

Le coût d'un alignement p entre X et Y est défini de la manière suivante :

$$c_p(X, Y) := \sum_{l=1}^L c(x_{n_l}, y_{m_l})$$

Un alignement optimal entre X et Y est un alignement p^* dont le coût est minimal parmi tous les alignements possibles. Ce coût est la distance DTW entre X et Y :

$$DTW(X, Y) := c_{p^*}(X, Y) = \min\{c_p(X, Y) / p \text{ est un "alignement"}\}$$

Afin de trouver l'alignement optimal p^* , nous pourrions tester tous les alignements possibles entre X et Y , mais cela mènerait à un algorithme de complexité exponentielle en $N * M$.

Nous allons donc définir un algorithme de complexité $O(N * M)$ permettant de trouver la distance DTW.

Pour cela, nous introduisons la matrice D de taille $(N \times M)$ telle que :

$$\forall (n, m) \in [1: N] \times [1: M], D(n, m) = \text{DTW}(X_n, Y_m)$$

Avec $X_n := (x_1, x_2, \dots, x_n)$ et $Y_m := (y_1, y_2, \dots, y_m)$

Donc $D(N, M) = \text{DTW}(X, Y)$.

L'algorithme consiste à remplir récursivement l'ensemble des cases de la matrice D de la manière suivante, la distance DTW étant le point $D(N, M)$:

$$D(n, 1) = \sum_{k=1}^n c(x_k, y_1) \forall n \in [1: N]$$

$$D(1, m) = \sum_{k=1}^m c(x_1, y_k) \forall m \in [1: M]$$

$$D(n, m) = c(x_n, y_m) + \min\{D(n-1, m-1), D(n-1, m), D(n, m-1)\} \forall (n, m) \in [2: N] \times [2: M]$$

Ainsi, l'avantage de cette distance par rapport aux mesures standards est qu'elle permet de ne pas systématiquement comparer chaque point d'une série avec celui de l'autre série qui intervient au même instant t . Elle peut comparer un point d'une série avec un ou plusieurs autres points de l'autre série, ceux-ci pouvant être décalés dans le temps et prendre ainsi en compte les décalages temporels. Cependant, la complexité de l'algorithme entraîne des temps de calculs considérables.

4.1.2. Décalage temporel

La condition $p_{i+1} - p_i \in \{(0,1), (1,0), (1,1)\}$ pour définir un alignement permet de prendre en compte des décalages temporels. Il faut néanmoins utiliser cette condition avec parcimonie, en ayant bien observé les données. Par exemple pour des séries de température, prendre en compte un décalage temporel de un ou deux jours serait acceptable, mais prendre en compte un décalage temporel d'un mois n'aurait plus de sens.

En effet, cette condition permet de prendre en compte des décalages temporels de n'importe quelle taille, en théorie, tant que l'alignement présente le coût minimal. Ainsi, on pourrait associer avec cette méthode un élément de X avec 30 éléments de Y par exemple, et prendre en compte un décalage temporel d'un mois.

Afin d'éviter ce genre de problème, on peut paramétrer l'écart temporel maximum permis à la mesure pour comparer des points. Soit d le décalage temporel maximum qu'on souhaite prendre en compte. A un instant t , un point de X ne pourra être comparé qu'avec les points de Y intervenant aux instants $[t - d, t + d]$.

4.1.3. Exemple illustrant la distance DTW

Afin de fixer les idées et comprendre l'intérêt de cette distance, notamment pour des données de température, nous avons choisi de mesurer la distance entre deux séries temporelles périodiques, simples, décalées de trois jours. Nous pouvons aisément assimiler ces séries à des séries de température. Soit $x_t = t \forall t \in \{1, \dots, 103\}$:

$$\begin{cases} X_t = \sin(x_t) \\ Y_t = \sin(x_{t+3}) \end{cases} \forall t \in \{1, \dots, 100\}$$

Les Figure II-7 et Figure II-8 représentent ces deux séries. Les droites en pointillées joignant les deux séries représentent les distances calculées entre chaque point avec la distance DTW sur la Figure II-7 et avec la distance euclidienne sur la Figure II-8. A première vue, ces séries semblent identiques et il serait cohérent que leur distance soit faible. Or, nous pouvons voir que la distance DTW compare les points décalés deux à deux (et leur attribue ainsi une distance nulle) alors que la distance euclidienne compare les points sans décalage (et attribue systématiquement une distance positive à part lorsque les séries se croisent).

Ainsi, la distance obtenue avec la méthode DTW est égale à 1,9 et la distance obtenue avec la distance euclidienne est égale à 74,2. En effet, avec DTW, seuls les trois premiers et les trois derniers points n'attribuent pas une distance nulle et engendrent une distance de 1,9. Même si la série de données était plus longue, la distance avec DTW serait donc toujours égale à 1,9 alors qu'elle augmenterait avec la distance euclidienne.

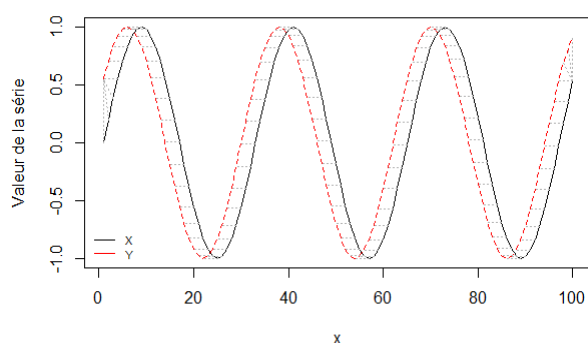


Figure II-7 Distance DTW

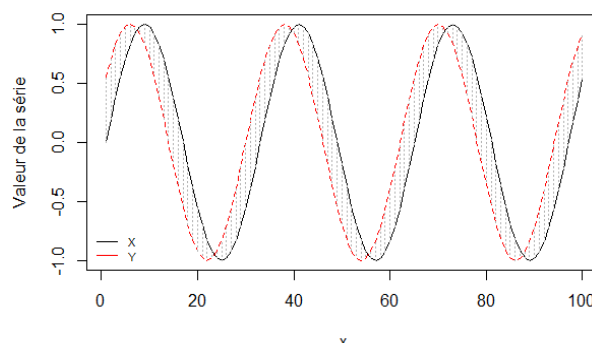


Figure II-8 Distance euclidienne

La méthode DTW est donc a priori adaptée aux séries de températures. En effet, notre exemple pourrait se retrouver dans la pratique avec les températures : on pourrait facilement imaginer deux points de France subissant des températures décalées à cause de certains phénomènes (comme la vitesse du vent par exemple). De plus, pour la classification des régions, nous souhaitons obtenir des régions dans lesquelles la température aurait un comportement similaire, ce qui est le cas dans notre exemple, et donc leur attribuer une faible distance.

La Figure II-9 montre l'alignement optimal trouvé avec DTW à gauche et l'alignement avec la distance euclidienne à droite. Pour prendre en compte le décalage de trois jours avec DTW, les trois premiers points de X sont donc comparés avec le premier point de Y et les trois derniers points de Y sont comparés avec le dernier point de X. A l'inverse, la distance euclidienne compare toujours chaque point de façon linéaire et ne prend donc pas en compte le décalage temporel.

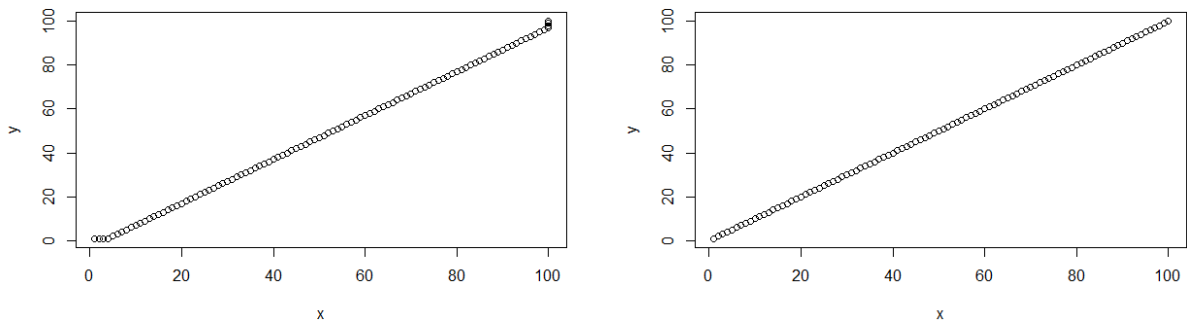


Figure II-9 Alignement avec la distance DTW (à droite) et la distance euclidienne (à gauche)

4.2. Les résultats de la CAH avec utilisation de la DTW

L'implémentation de la CAH avec la distance DTW a été effectuée en tenant compte d'un décalage temporel maximal de quatre jours, qui permet d'accepter des décalages de température dus, par exemple, à la vitesse du vent. Au-delà de quatre jours, le décalage temporel n'est plus très pertinent au regard du phénomène physique. Nous présentons sur la Figure II-10 le graphique obtenu en implémentant la CAH avec la distance DTW pour un choix de quatre classes en le comparant avec le graphique correspondant de la méthode précédente.

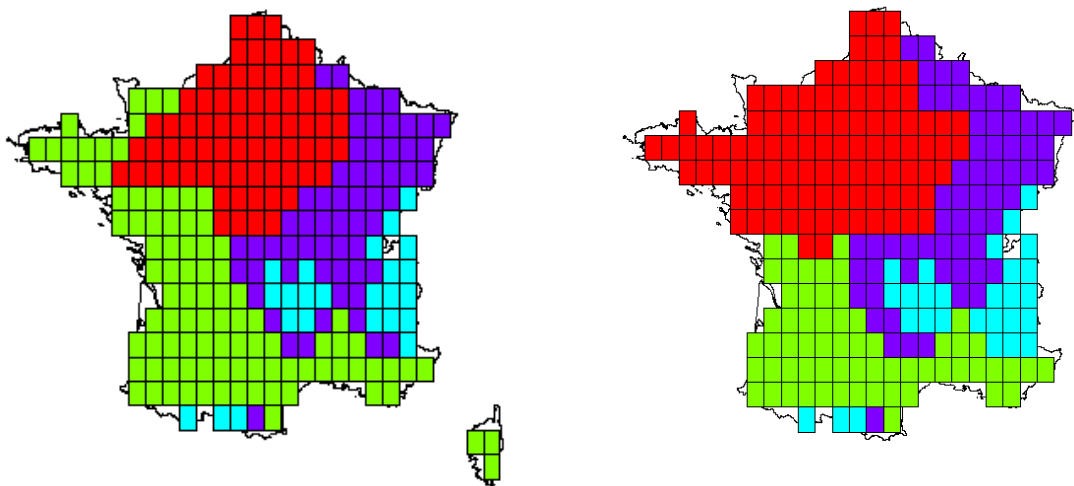


Figure II-10 Quatre régions de la France obtenues par la méthode CAH avec DTW (à gauche) et par la méthode EOF puis CAH (à droite)

En comparant les régions obtenues avec (i) la CAH sur les premières composantes principales de la méthode EOF (1^{ère} méthode) et (ii) avec la CAH utilisant la distance DTW (2^{ème} méthode), force est de constater que les régions 2 et 3, respectivement en violet et bleu cyan, sont quasiment identiques avec les deux méthodes.

La différence importante réside sur la classe en vert. Dans le cas de la CAH avec DTW, cette classe s'étend jusque sur la partie Nord-ouest de la France en incluant la Bretagne et une partie de la Manche ; ce qui contribue à réduire l'étendue de la première région en rouge.

Outre le temps calcul important pour une matrice de données telle que la notre (de taille 250x23 193) avec la distance DTW, il nous apparaît plus cohérent de considérer la classification obtenue avec la combinaison des méthodes EOF et CAH qui semble mieux correspondre au schéma

présenté en annexe illustrant les différents climats observés en France (cf. schéma climatique de la France en annexe A).

Cette partie a permis de réduire l'information à disposition sur les températures afin de se ramener à quatre régions homogènes en température en France. Nous nous servirons de ces régions dans la construction du module Aléa.

III. Construction du module Aléa

Le module Aléa vise à simuler plusieurs scénarios de gel en France. La classification effectuée dans la partie précédente, nous a permis de réduire la dimension du problème pour pouvoir simuler des scénarios de gel dans chaque région, et donc uniquement en quatre points.

Pour ce faire, dans un premier temps, nous allons caractériser ce qui déclenche un évènement de gel à l'aide de l'historique des sinistres survenus chez AXA. Nous verrons qu'un évènement gel dépend directement des températures observées. Nous proposerons donc, dans un deuxième temps, une modélisation statistique de la température tenant compte des corrélations inter-régions. Nous pourrons par la suite simuler plusieurs scénarios de température à partir desquels seront déduits les évènements de gel.

1. Caractérisation de l'évènement gel

Le gel est caractérisé physiquement par des températures négatives. Or, nous voulons caractériser ici un évènement gel au sens assurantiel, c'est-à-dire un évènement gel qui déclenche des sinistres et engendre des pertes pour un assureur. Cette caractérisation diffère de la caractérisation physique, car on observe historiquement que des températures négatives n'entraînent pas nécessairement de sinistres. Cela complexifie donc la caractérisation du phénomène.

Il est intuitif de supposer que la survenance d'un évènement gel au sens assurantiel est aussi lié à la température mais il faut se questionner sur la façon dont la température influe sur ce phénomène. Dans cette partie, nous allons déterminer si, par exemple, le gel est lié à la température du jour d'occurrence du sinistre ou est lié aux écarts de température sur une période précédant le sinistre. Ainsi, pour caractériser le gel, il est nécessaire de déterminer les variables de température influençant son apparition. Une fois connues, elles seront utilisées comme variables explicatives permettant de faire le lien entre le niveau température et les sinistres déclarés.

La caractérisation du gel passe donc par deux étapes essentielles. La première étape est dédiée à la recherche de variables de température susceptibles d'expliquer la survenance du gel. Ensuite, dans la seconde étape, les sinistres historiques seront modélisés au regard des variables précédentes.

Pour ce faire, différentes données seront exploitées, notamment :

- Les fichiers de pertes dues au gel d'AXA en France entre 01/01/2008 et 31/12/2012,
- Le fichier de polices assurées au gel d'AXA, en France
- Les données de températures au niveau des 250 localisations géographiques ou au niveau de la région selon le niveau de détail souhaité.

Dans cette section, ces données seront utilisées à différents niveaux d'agrégation. Dans la suite, pour chaque étape, les hypothèses émises, les données utilisées et les approches méthodologiques adoptées seront décrites succinctement.

1.1. Les déterminants de la survenance du gel

1.1.1. Les différents modèles possibles

Dans un premier temps, nous supposons *a priori* que le gel est causé par des températures froides ou par des chutes de température sur une période donnée. Nous proposons donc de tester les variables suivantes comme variables explicatives de la survenance de gel :

- La température minimale le jour j de la survenance du sinistre, notée t_{min} ,
- Les écarts de température minimale entre le jour de la survenance du sinistre j et les jours précédents $j - n$ (avec $n = 10, 15, 20$ jours), notés t_{min_n} ,
- Les écarts entre le maximum et le minimum de température minimale atteinte entre le jour de la déclaration du sinistre j et les jours précédents $j - n$ (avec $n = 10, 15, 20$ jours), notés $MinMaxT_n$.

Pour chaque variable citée ci-dessus, un modèle sera construit. Ces modèles permettront de déterminer si les variables considérées influent sur l'apparition de gel. Au total, sept modèles seront donc implémentés. Afin de déterminer les variables expliquant la survenance du gel, nous souhaitons avoir le détail le plus fin possible. Nous travaillerons donc sur les relevés des températures sur les 250 localisations géographiques. Par ailleurs, le gel étant susceptible de se produire uniquement en hiver lorsque les températures sont négatives (d'après la caractérisation physique), les données de pertes utilisées ici sont celles pour lesquelles la température observée le jour de survenance du sinistre est négative et pour lesquelles la déclaration s'est faite en hiver. Cela nous permettra de minimiser le biais induit par la déclaration de sinistres tardifs. Par exemple, il est évident que les sinistres déclarés en été sont survenus antérieurement en réalité. Nous avons 1 827 jours pour l'ensemble des hivers des années 2008 à 2012 pour 250 localisations géographiques, soit 456 750 observations. En ne considérant que les jours de l'hiver où la température minimale est négative, notre historique est réduit à 62 852 observations.

Nous supposons également que la survenance du gel peut être différenciée suivant la région en France. Pour prendre en compte cette différenciation, la région sera incluse dans le modèle en tant que variable qualitative.

La variable à expliquer est la survenance ou non du sinistre. La survenance peut être approximée par une variable binaire Y qui prend la valeur 1 lorsqu'au moins un sinistre est déclaré et 0 dans le cas contraire. Dans le cas où la variable dépendante est binaire, il convient d'utiliser un modèle dichotomique.

1.1.2. Les modèles dichotomiques

Les modèles dichotomiques spécifient la probabilité d'observer $y_i = 1$ conditionnellement aux réalisations des variables explicatives x_i par une fonction $F(\cdot)$. Ce qui se traduit par :

$$P(y_i = 1|x_i) = F(x_i\beta) = p_i.$$

La fonction $F(\cdot)$ caractérise à la fois la distribution de probabilité mais également l'espérance conditionnelle du modèle. D'où l'expression ci-après :

$$E(y_i|x_i) = \sum_{j=0}^1 jP(y_i = j|x_i) = F(x_i\beta) = p_i$$

Le domaine de définition de Y étant $\{0; 1\}$, il est nécessaire que la fonction F soit croissante, bornée inférieurement par 0 et supérieurement par 1. Toute fonction de répartition (de n'importe quelle loi de probabilité) est donc appropriée. Néanmoins, dans la pratique, deux modèles sont utilisés⁷ :

- Le modèle probit dont la fonction de répartition associée est celle d'une loi normale.

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt$$

- Le modèle logit repose sur une fonction logistique et les erreurs associées suivent une loi binomiale.

$$F(z) = \frac{e^z}{1 + e^z}$$

La différence essentielle entre ces modèles réside dans le choix de la loi du terme des erreurs. Dans le cadre de ce mémoire, le choix est porté sur le modèle logit car on suppose a priori que les erreurs sont distribuées suivant une loi binomiale. En effet, la loi binomiale est adaptée lorsque l'on s'intéresse à la survenance ou non d'un évènement.

Dans le cadre du modèle logit à une variable explicative, la probabilité de succès pour un individu s'écrit :

$$p_i = F(x_i\beta) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

En généralisant la formule de la probabilité de succès avec k variables explicatives, on obtient :

$$p = F(\alpha + \beta_1x_1 + \dots + \beta_kx_k) = \frac{e^{\alpha + \beta_1x_1 + \dots + \beta_kx_k}}{1 + e^{\alpha + \beta_1x_1 + \dots + \beta_kx_k}}$$

Le ratio $\frac{p}{1-p}$ représente le *odds ratio* ou « rapport de cotes ». Lorsque le *odds ratio* est égal à 2, cela signifie que la probabilité d'apparition du sinistre est 2 fois plus élevée que sa non survenance. Le $\log\left(\frac{p}{1-p}\right)$ est le logit de p . La fonction logit permet donc de linéariser le *odds ratio*.

En développant, il vient que $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1x_1 + \dots + \beta_kx_k$ où p est la probabilité de succès.

1.1.3. Estimation des paramètres

Les paramètres du modèle sont estimés par la méthode du maximum de vraisemblance. La vraisemblance est donnée par :

$$L(y_1, \dots, y_n; \beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

Dans la pratique, on a plus souvent recours à la maximisation de la log-vraisemblance donnée par la formule ci-contre :

⁷CREPON & JACQUEMET (2010) pensent que ces deux modèles sont privilégiés du fait de la concavité de la fonction de vraisemblance associée. En effet, cette condition de concavité globale assurée sur la fonction de log-vraisemblance permet de garantir que l'optimum trouvé est celui recherché.

$$\log\{L(y_1, \dots, y_n; \beta)\} = \sum_{i=1}^n y_i \log(p_i) + \sum_{i=1}^n (1 - y_i) \log(1 - p_i)$$

La déviance D est égale à la log-vraisemblance à une constante négative près. Ainsi, maximiser la log-vraisemblance est équivalent à minimiser la déviance.

$$D = -2 \left[\sum_{i=1}^n y_i \log(p_i) + \sum_{i=1}^n (1 - y_i) \log(1 - p_i) \right]$$

1.1.4. Les critères de comparaison des modèles

Lorsque l'on dispose de plusieurs modèles, il est naturel de se demander quel est le meilleur. Sachant que la vraisemblance permet de mesurer l'adéquation du modèle aux données, il est possible de comparer la log-vraisemblance, et donc la déviance, calculée en $\hat{\beta}$ des modèles afin de détecter le modèle le plus performant. Néanmoins, le critère de la vraisemblance ne permet pas toujours de choisir le meilleur modèle car la log-vraisemblance croît avec le nombre de paramètres estimés. Il convient donc d'utiliser des critères de vraisemblance pénalisés à savoir :

- Le critère d'Akaike (AIC)

$$AIC = -2\ln(L(\hat{\beta})) + 2k \text{ où } k \text{ est le nombre de paramètres du modèle}$$

- Le critère de Schwartz (BIC)

$BIC = -2\ln(L(\hat{\beta})) + \ln(n)k$ où k est le nombre de paramètres du modèle et n le nombre d'observations.

De manière générale, le « meilleur » modèle est celui qui aura l'AIC et/ou le BIC minimal. A la différence du critère l'AIC, le BIC tient compte du nombre de paramètres en plus de la taille de l'échantillon. Pour de grands échantillons, le BIC aura tendance à favoriser des modèles à moins de paramètres que l'AIC.

Il convient de noter que nous ne nous intéresserons pas au meilleur modèle mais aux meilleurs modèles. En effet, on pourra choisir un panel de modèles suivant la significativité des coefficients d'estimation et les critères de comparaison évoqués.

1.1.5. Résultats des régressions logistiques

Dans cette partie, nous allons estimer sept modèles différents, chacun visant à prédire la probabilité d'occurrence du gel. S'agissant de la spécification, chaque modèle contient une variable à expliquer binaire et deux variables explicatives. La première variable explicative est quantitative et est une variable de température. La seconde est qualitative et contient les numéros de régions obtenues avec la segmentation de la partie II.

Nous considérons que les modèles performants sont ceux dont toutes les variables explicatives sont significatives. L'intérêt premier étant la variable de température susceptible d'expliquer l'apparition du gel, nous pouvons tolérer que les estimateurs des modalités de la région ne soient pas significatifs.

Le Tableau III-1 présente les résultats des différents modèles. Le seuil de significativité des coefficients est fixé à 5 %. On constate que pour tous les modèles, la variable de température est significative sauf pour le second modèle (avec t_{min10}). Ce second modèle sera donc exclu de nos choix possibles. Sachant que six variables de température sur sept ont leurs estimateurs significatifs,

on conclut que de manière générale, la survenance du gel est bien due à la température journalière ou à des écarts de température sur une période donnée.

Parmi les résultats des régressions présentées, force est de constater que le coefficient de la région 2 n'est pas significatif pour les modèles 3 à 6. Par contre, deux modèles ont des résultats satisfaisants. Toutes les estimations des variables explicatives de ces deux modèles sont significatives. Il s'agit du premier modèle dont la variable de température est tmin et le septième modèle dont la variable de température est MinMaxT₂₀.

Modèles	Variables	Paramètres	Erreur	t-Student	p-valeur	Significativité
Modèle 1	constante	-3,240	0,034	-95,556	<2e-16	✓
	tmin	-0,320	0,004	-72,597	<2e-16	✓
	region2	0,170	0,044	3,824	0,000	✓
	region3	-1,673	0,051	-32,973	<2e-16	✓
	region4	-0,592	0,038	-15,463	<2e-16	✓
Modèle 2	constante	-2,052	0,042	-49,254	<2e-16	✓
	tmin10	-0,005	0,004	-1,285	0,199	✗
	region2	0,074	0,042	1,752	0,080	✗
	region3	-0,823	0,046	-18,066	<2e-16	✓
	region4	-0,252	0,035	-7,173	0,000	✓
Modèle 3	constante	-3,018	0,049	-62,170	<2e-16	✓
	tmin15	0,091	0,004	23,573	<2e-16	✓
	region2	0,043	0,042	1,020	0,308	✗
	region3	-0,542	0,046	-11,845	<2e-16	✓
	region4	-0,200	0,035	-5,667	0,000	✓
Modèle 4	constante	-3,649	0,052	-70,814	<2e-16	✓
	tmin20	0,136	0,004	37,332	<2e-16	✓
	region2	0,066	0,042	1,549	0,121	✗
	region3	-0,372	0,046	-8,078	0,000	✓
	region4	-0,163	0,036	-4,578	0,000	✓
Modèle 5	constante	-3,713	0,055	-67,896	<2e-16	✓
	MinMaxT10	0,157	0,004	35,536	<2e-16	✓
	region2	0,003	0,042	0,078	0,938	✗
	region3	-0,700	0,045	-15,592	<2e-16	✓
	region4	-0,325	0,036	-9,097	<2e-16	✓
Modèle 6	constante	-5,662	0,068	-83,272	<2e-16	✓
	MinMaxT15	0,282	0,005	61,001	<2e-16	✓
	region2	0,064	0,044	1,462	0,144	✗
	region3	-0,609	0,046	-13,195	<2e-16	✓
	region4	-0,425	0,037	-11,491	<2e-16	✓
Modèle 7	constante	-6,795	0,077	-88,428	<2e-16	✓
	MinMaxT20	0,331	0,005	70,432	<2e-16	✓
	region2	0,167	0,045	3,741	0,000	✓
	region3	-0,516	0,047	-10,956	<2e-16	✓
	region4	-0,461	0,038	-12,160	<2e-16	✓

Tableau III-1 Résultats des régressions logistiques pour la probabilité d'occurrence du gel

Le Tableau III-2 montre également que le modèle 7, directement suivi par le modèle 1, sont les meilleurs modèles. Sachant que le nombre de paramètres à estimer est identique dans chacun de ces modèles, il est légitime ici de limiter la comparaison entre les modèles à la déviance ou à la log-vraisemblance. En utilisant les critères AIC et/ou BIC, on aboutit sans surprise aux mêmes conclusions.

Modèles	Variable 1	Déviance	Degré de liberté	Log vraisemblance	AIC	BIC
Modèle1	Tmin	31 328	5	-15 664	31 338	31 383
Modèle2	tmin10	36 976	5	-18 488	36 986	37 031
Modèle3	tmin15	36 430	5	-18 215	36 440	36 485
Modèle4	tmin20	35 591	5	-17 796	35 601	35 647
Modèle5	MinMaxT10	35 736	5	-17 868	35 746	35 791
Modèle6	MinMaxT15	32 980	5	-16 490	32 990	33 035
Modèle7	MinMaxT20	31 183	5	-15 592	31 193	31 238

Tableau III-2 Comparaison des modèles de régression logistique

La régression logistique ayant pour but de définir la probabilité d'occurrence du gel en fonction des données de température, il est intéressant de savoir si cette probabilité d'occurrence est plus importante dans une région plutôt qu'une autre. Pour cela, nous allons utiliser le modèle 7 qui est le meilleur dans l'ensemble des régressions logistiques effectuées et qui contient des estimateurs significatifs pour toutes les régions. La variable de température d'intérêt pour ce modèle est $MinMaxT_{20}$, c'est-à-dire l'écart entre le maximum et le minimum des températures minimales sur une durée de 20 jours précédant la déclaration du sinistre. Il s'agit d'exhiber le lien entre le niveau de $MinMaxT_{20}$ et la probabilité d'occurrence du gel suivant les régions.

La Figure III-1 présente la probabilité d'occurrence du gel suivant les écarts de température (entre le minimum et le maximum) sur 20 jours. On constate que la probabilité de gel croît lorsque cet écart sur 20 jours est important. Lorsque cet écart est strictement inférieur à 10°C , la probabilité de gel est faible ($<0,1$) et est peu différenciée sur les régions. Lorsque cet écart dépasse 10°C , la probabilité de survenance du gel est plus forte dans la région 2 (région Nord-Ouest) suivie de la région 1 (région Sud) et ensuite dans les régions 4 (région Nord-Est) et 3 (région montagneuse). Pour ces deux dernières régions, les probabilités sont assez proches. Plus l'écart est important, plus les probabilités sur les régions sont différentes. Pour un écart supérieur à 22°C , la probabilité de survenance du gel est supérieure à $0,5$.

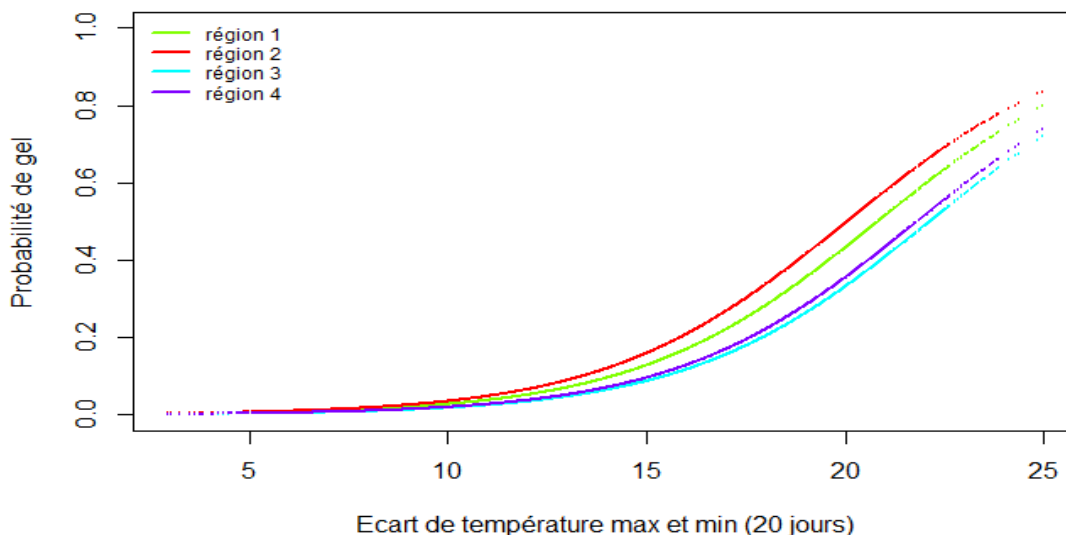


Figure III-1 Probabilité d'occurrence du gel par région (avec $MinMaxT_{20}$)

A première vue, il peut sembler surprenant que la probabilité de survenance de gel soit très importante dans la région Sud. Cela tient du fait que nous modélisons l'occurrence du gel au sens assurantiel. Cette région étant moins habituée à subir de sévères chutes de températures, on peut postuler que les structures des bâtiments sont moins robustes au péril gel.

1.2. Modélisation de la fréquence de gel

La sous-section précédente a permis de sélectionner deux variables de température qui expliquent la survenance du gel : les écarts de température sur 20 jours et la température minimale journalière. Nous utiliserons ces deux variables pour la modélisation de la fréquence journalière des sinistres donnée par :

$$\text{fréquence} = \frac{\text{Nombre de sinistres journaliers}}{\text{Exposition}}$$

La fréquence dépend à la fois du nombre de sinistres journaliers et de l'exposition. L'exposition représente le nombre de polices assurées par AXA. L'exposition étant connue, la fréquence s'obtient aisément après la modélisation du nombre de sinistres journaliers. Dans la suite, nous nous intéressons donc à la modélisation de ce nombre de sinistres.

Pour la modélisation de variables de comptage telles que le nombre de sinistres, nous présenterons l'approche classique des régressions de poisson, géométrique⁸ ou binomiale négative. Cependant, le gel étant un événement dont l'occurrence est faible et très localisée temporellement, cela amènera à évoquer les modèles avec excès de zéro. Bien que ces modèles estiment bien les nombres de sinistres d'un point de vue global, ils tiennent mal compte des extrêmes. Cela nous conduira finalement à retenir les modèles avec dépassement de seuil. Il convient de noter que les différents modèles ont été testés pour chaque combinaison des deux variables de température sélectionnées dans la section précédente. Ainsi, pour chaque étape d'implémentation et pour chaque régression utilisée, nous avons trois modèles :

- un modèle avec la température minimale journalière et les modalités des régions ;
- un modèle avec des écarts de température sur 20 jours et les modalités des régions ;
- un modèle avec les deux variables de température et les modalités de régions.

Le dernier s'est avéré être le meilleur suite au backtesting des simulations au regard des historiques. Nous présenterons uniquement ce dernier modèle.

Avant de présenter les résultats issus du modèle à deux variables de température, il est important de préciser les données utilisées.

1.2.1. Des données par région

Dans cette partie, nous nous concentrons sur la modélisation des fréquences des sinistres causés par le gel au niveau de la région. En effet, comme les simulations d'évènements dans le module Aléa sont effectuées pour chaque région de la classification (cf. partie II), il est nécessaire d'avoir des données agrégées à la région.

Nous regroupons donc les polices d'AXA, classées initialement par code INSEE⁹, par région en utilisant des outils de géocodage. Pour chaque région, nous agrégeons le nombre de polices couvertes par AXA pour obtenir l'exposition par région. Nous effectuons la même procédure sur l'historique des sinistres survenus.

Par nécessité de confidentialité, l'exposition par région ne sera pas présentée. Par contre, il est à noter que la région 1 (Sud) a la plus grosse part d'exposition. Elle est suivie par la région 2 (Nord-

⁸ Le modèle de régression géométrique est un cas particulier de modèle de régression binomiale négative, lorsque le paramètre de forme est égal à l'unité. Nous nous limitons au cas général. Nous ne mentionnerons donc pas le modèle de régression géométrique.

⁹ INSEE (Institut National de la Statistique et des Etudes Economiques)

Ouest), la région 4 (Nord-Est) et enfin, la région 3 (région montagneuse). Pour le modèle, l'exposition par région est supposée constante chaque année. Cette hypothèse est justifiée par le fait que la part d'exposition par région est quasiment constante d'une année à l'autre, que le nombre de polices assurées augmente ou non.

De même, nous agrégeons les variables de température déterminées précédemment au niveau des régions. Nous obtenons alors les variables de température explicatives suivantes :

- La température minimale journalière moyenne sur la région considérée (que nous appelons t_{minMoy}) à la place de t_{min}
- Les écarts moyens observés sur la région considérée entre le maximum et le minimum des températures sur une durée de 20 jours (que nous appelons $MinMaxT20Moy$) à la place de $MinMaxT20$

Comme dans la sous-section III.1.1, nous ne considérons que les sinistres ayant eu lieu durant les mois d'hiver des années 2008 à 2012. Notre tableau de données contient également des restrictions sur la température minimale du jour. Précédemment, l'information était plus fine et la table de données était plus restrictive car réduit aux jours de température minimale négative. En se ramenant au niveau de la région, nous considérons une moyenne de température minimale (t_{minMoy}) mais nous posons une condition plus souple à savoir $t_{minMoy} < 3^{\circ}C$. En effet, bien que les classes obtenues dans la partie II soient homogènes, il y subsiste un écart type non nul. Cet écart type excède $2^{\circ}C$ dans la classe 3 mais reste en dessous de $3^{\circ}C$ pour toutes les régions. Ainsi, même si la température moyenne au sein d'une région est de $1^{\circ}C$ par exemple, certains points de la région auront sûrement des températures négatives et pourront potentiellement être impactés par le gel. Cela justifie la condition posée sur t_{minMoy} . Finalement, le tableau de données contient 1 646 observations.

1.2.2. Modèles classiques de comptage

Dans la section III.1.1, la survenance d'un sinistre a été modélisée par une variable binaire. Dans cette partie, nous voulons modéliser le nombre de sinistres. Pour cela, des modèles de comptage sont appropriés.

Les modèles de comptage classiques font partie de la famille des modèles linéaires généralisés. Les modèles linéaires généralisés décrivent la dépendance d'un vecteur $y_i, i = 1, \dots, n$ sur un ensemble de régresseurs x_i . La distribution conditionnelle de $y_i | x_i$ est une famille linéaire exponentielle dont la fonction de densité est donnée par

$$f(y; \lambda, \phi) = c(y, \phi) \exp \left(\frac{y\lambda - b(\lambda)}{\phi} \right)$$

Où λ est le paramètre canonique et ϕ le paramètre de dispersion.

La dépendance de l'espérance conditionnelle $E(y_i | x_i) = \mu_i$ est donnée par la fonction de lien canonique connue $g(\mu_i) = x_i' \beta$. La vraisemblance est la densité de la loi jointe de Y_1, \dots, Y_n :

$$L(y; \lambda, \phi) = \prod_{i=1}^n f(y_i; \lambda_i, \phi) = \left(\prod_{i=1}^n c(y_i, \phi) \right) \exp \left(\frac{\sum_{i=1}^n y_i \lambda_i - b(\lambda_i)}{\phi} \right)$$

Dans la suite, les différents modèles de comptage seront présentés en prenant en compte la spécificité de nos données. La normalisation des données est effectuée à l'aide de l'exposition qui est l'offset. Dans notre cas, l'offset est donc égal à l'exposition sur la région.

Régression poisson

La régression de poisson est le modèle de comptage le plus simple. Sa fonction de densité s'exprime ainsi :

$$f(y; \mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

Le lien canonique est donnée par $g(\mu) = \log(\mu)$, ce qui signifie une relation log-linéaire entre l'espérance de Y et le prédicteur linéaire. Par ailleurs, le modèle de régression de Poisson est basé sur des hypothèses contraignantes : $E(Y) = V(Y) = \mu$ et le coefficient de dispersion $\phi = 1$. L'estimation se fait par maximisation de la fonction de vraisemblance.

Les résultats du modèle de régression sont présentés dans le Tableau III-3.

	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>Constante</i>	-18,412	0,054	-343,020	<2e-16	✓
<i>MinMaxT20Moy</i>	0,459	0,004	131,053	<2e-16	✓
<i>tminMoy</i>	-0,056	0,002	-28,512	<2e-16	✓
<i>region2</i>	0,982	0,019	50,588	<2e-16	✓
<i>region3</i>	0,325	0,027	12,165	<2e-16	✓
<i>region4</i>	0,099	0,021	4,793	1,65E-06	✓

Tableau III-3 Résultats de l'estimation avec la régression de poisson

Ce tableau montre que tous les coefficients sont significatifs au seuil de 5 %. Néanmoins, compte tenu de la dispersion importante sur les nombres de sinistres, il est important de savoir si les observations sont surdispersées.

Test de dispersion : poisson ou binomiale négative ?

Lorsqu'il y a surdispersion, c'est la régression binomiale négative qui est privilégiée. On parle de surdispersion lorsque $\phi > 1$. On peut donc tester l'hypothèse de dispersion suivante :

$$\begin{cases} H_0: \phi = 1 \\ H_1: \phi > 1 \end{cases}$$

D'après les travaux de CAMERON & TRIVEDI (2005), sous l'hypothèse nulle, on a $V(Y) = \mu$. En cas de surdispersion, $V(Y) = \mu + \alpha * \varphi(\mu)$ avec $\alpha > 0$ et $\varphi(\cdot)$ une fonction. En effet, CAMERON & TRIVEDI (2005) considèrent deux formes pour la fonction $\varphi(\cdot)$:

- $\varphi(\mu) = \mu$ dans le cas où on a une spécification linéaire de la forme de la variance,
- $\varphi(\mu) = \mu^2$ dans le cas où on a une spécification quadratique de la forme de la variance.

Le test de dispersion peut se réécrire ainsi :

$$\begin{cases} H_0: \alpha = 0 \\ H_1: \alpha > 0 \end{cases}$$

La statistique de test est $Z = \frac{\sqrt{n}(\hat{\alpha} - \alpha)}{\sigma}$. Les résultats obtenus à partir de nos données sont présentées dans le Tableau III-4.

alpha	Z	p-value
46,20	4,11	1,97E-05

Tableau III-4 Résultats du test de dispersion

D'après ce tableau, α est significativement différent de 0 au seuil de 5 %. On conclut que nos observations sont surdispersées. Il est donc plus intéressant d'appliquer un modèle linéaire généralisé avec des nombres de sinistres distribués suivant une binomiale négative.

Régression binomiale négative

La binomiale négative est une composition de loi gamma et de loi de poisson. La densité d'une loi binomiale négative ajoute un paramètre de forme θ au paramètre de la loi de poisson afin de tenir compte de la dispersion. La fonction de densité est donnée ci-après :

$$f(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta) * y!} * \frac{\mu^y \theta^\theta}{(\mu + \theta)^{y+\theta}}$$

Comme précédemment, l'estimation est effectuée par maximisation de la vraisemblance. Les résultats de l'estimation sont donnés dans le Tableau III-5.

	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>constante</i>	-17,443	0,234	-74,548	<2e-16	✓
<i>MinMaxT20Moy</i>	0,389	0,018	22,061	<2e-16	✓
<i>tminMoy</i>	-0,065	0,015	-4,182	0,000	✓
<i>region2</i>	0,458	0,143	3,200	0,001	✓
<i>region3</i>	0,414	0,137	3,022	0,003	✓
<i>region4</i>	0,361	0,133	2,718	6,57E-03	✓

Tableau III-5 Résultats de l'estimation avec la régression binomiale négative

Tous les coefficients sont significatifs au seuil de 5 %. Par ailleurs, $\hat{\theta}$ est significativement différent de zéro au seuil de 5 %. La valeur de $\hat{\theta}$ est 0,3414.

En comparant les effectifs estimés et les effectifs pour chaque nombre de sinistres, l'ajustement avec la distribution binomiale négative est meilleur qu'avec la loi de poisson. Néanmoins, on constate que l'estimation pour les petits nombres de sinistres (notamment les sinistres nuls) est sous-estimée avec la régression de poisson et mal estimée avec la binomiale négative. Le problème de l'ajustement du nombre de sinistres avec une loi binomiale négative est que nos données sont très dispersées : l'écart type avoisine 50 sinistres sachant que le minimum du nombre observé de sinistres est nul et son maximum est d'environ 700 sinistres. La binomiale négative n'est pas très performante compte tenu des variations importantes d'amplitude.

En outre, il faut relever que les données présentent un nombre important de zéros. En effet, sur les 1646 observations de la table de données, 716 ont un nombre de sinistres nul (ces observations correspondent aux jours où il n'y a pas eu de déclaration de sinistres). Dans le cas de données possédant de nombreux zéros, les modèles classiques de comptage ne sont pas toujours adaptés. Nous nous tournons donc vers les modèles de comptage avec excès de zéros.

1.2.3. Modèles de comptage avec excès de zéros

Afin de tenir compte du nombre important de valeurs nulles (les jours où aucun sinistre n'est déclaré), de nouveaux modèles ont été développés : les modèles *zero inflated* et *hurdle*. Ces modèles permettent également d'utiliser un offset en vue de la normalisation des données. Comme précédemment, l'offset est l'exposition par région.

Modèle zero inflated

Un modèle *zero inflated* est la combinaison d'un Dirac en 0 $I_{\{0\}}(y)$ et d'une distribution de comptage $f_{\text{compte}}(y; x, \beta)$. Dans ce modèle, il y a deux sources d'apparition des zéros. Les zéros peuvent apparaître à cause de l'existence d'un Dirac ou être générés par le modèle de comptage. Cela se traduit par deux phénomènes : soit l'évènement (ici, le gel) générant le sinistre ne s'est pas produit, soit l'évènement s'est produit mais n'a généré aucun sinistre.

Pour calibrer la distribution de comptage, notée $f_{\text{compte}}(\cdot)$, nous partons d'un modèle classique de comptage (régression de poisson, régression géométrique ou régression binomiale négative). La probabilité d'apparition du zéro ou la fonction de densité en zéro notée $f_{\text{zero}}(0; z; \gamma)$ est modélisée en faisant appel à un modèle binaire de type probit ou logit. Il apparaît bien que la probabilité d'observer une valeur de comptage nulle est augmentée de la probabilité d'obtenir un zéro dans le modèle en zéro.

La fonction de densité dans le cas du modèle *zero inflated* est spécifiée ainsi qu'il suit :

$$f_{\text{zeroinflated}}(y; x, z, \beta, \gamma) = f_{\text{zero}}(0; z, \gamma)I_{\{0\}}(y) + (1 - f_{\text{zero}}(0; z, \gamma))f_{\text{comptage}}(y; x, \beta)$$

Dans l'équation précédente, β et γ sont les paramètres à estimer. On notera que x et z sont des observations. Avec le modèle *zero inflated*, il est possible d'utiliser des données différentes pour les fonctions de densité f_{zero} et f_{comptage} .

Dans le cadre du mémoire, les mêmes observations seront utilisées pour spécifier le modèle en zéro et le modèle de comptage, donc $x = y$. Par ailleurs, une régression logistique sera utilisée pour approcher le modèle en zéro. Sachant que le test de dispersion implémenté précédemment a montré que les données étaient sur-dispersées, la distribution de comptage sera approchée par une régression binomiale négative.

L'estimation des coefficients se fait par maximisation de la vraisemblance. Les coefficients estimés sont présentés dans le Tableau III-6. Dans le modèle de comptage, les coefficients sont tous significatifs au seuil de 10 %. A ce même seuil, il convient de noter que la variable de dispersion est significativement différente de zéro. En effet, $\hat{\theta} = 0,3592$.

Par contre, pour le modèle en zéro, seule la variable des écarts de température $\text{MinMaxT}_{20}\text{Moy}$ a un estimateur significativement différent de zéro. Pour le modèle en zéro, seule la variable $\text{MinMaxT}_{20}\text{Moy}$ permet d'expliquer l'apparition de zéro, les autres coefficients estimés étant tous nuls.

Modèle de comptage					
	Paramètres	Erreur	t-Student	p-valeur	Significativité
(Intercept)	-17,270	0,224	-77,101	<2e-16	✓
MinMaxT20Moy	0,375	0,017	22,449	<2e-16	✓
tminMoy	-0,064	0,015	-4,391	0,000	✓
region2	0,676	0,156	4,348	0,000	✓
region3	0,414	0,137	3,022	0,003	✓
region4	0,366	0,130	2,813	4,90E-03	✓
Modèle en zéro					
	Paramètres	Erreur	t-Student	p-valeur	Significativité
(Intercept)	-9,089	320,973	-0,028	0,9774	✗
MinMaxT20Moy	-0,726	0,303	-2,398	0,0165	✓
tminMoy	0,273	0,209	1,303	0,193	✗
region2	15,300	320,962	0,048	0,962	✗
region3	-0,801	744,095	-0,001	0,999	✗
region4	-1.0378	1006.9618	-0.001	0.9992	✗

Tableau III-6 Résultats de l'estimation avec la régression *zero inflated*

Dans la suite, nous présentons les résultats des estimations avec le modèle *hurdle*.

Modèle hurdle

Un modèle *hurdle* est un modèle à deux composantes : un modèle binaire censuré à droite (en $y = 1$ dans notre cas) et un modèle de comptage tronqué à gauche (en $y = 1$ dans notre cas). Le premier modèle prédit la probabilité d'avoir un nombre de sinistres nul. Le second modélise la probabilité d'avoir un nombre de sinistres non nul.

Contrairement au modèle *zero inflated*, le modèle *hurdle* ne suppose pas deux sources de zéros mais que les valeurs nulles et les valeurs non nulles du nombre de sinistres sont générées par deux lois différentes. Dans l'optique du *hurdle*, la seconde composante du modèle n'est estimée que si la première composante n'est pas satisfaisante (par exemple, lorsque la probabilité estimée de n'avoir aucun sinistre est très faible).

Le modèle *hurdle* se spécifie ainsi :

$$\begin{cases} P(y = 0) = f_1(0) \\ P(y = j) = \frac{1 - f_1(0)}{1 - f_2(0)} f_2(j), \quad j > 0 \end{cases}$$

Avec f_1 la loi de densité du modèle binaire dit modèle *hurdle* de zéros et f_2 la loi de densité du second modèle tronqué pour les valeurs non nulles du nombre de sinistres.

Les paramètres du modèle final sont une combinaison des paramètres de chaque loi de densité et sont obtenus via l'estimation des paramètres des deux modèles sous-jacents. Pour chaque modèle sous-jacent, les paramètres sont estimés à l'aide de la maximisation de log-vraisemblance. Les paramètres du modèle global s'estiment donc à l'aide du maximum de vraisemblance.

La première composante du modèle *hurdle* est généralement modélisée par une loi binomiale (avec une fonction de lien canonique logit). La seconde composante du modèle *hurdle* est implémentée à l'aide des modèles de comptage classique (poisson, géométrique, binomiale négative). Comme dans le cas du modèle *zero inflated*, la distribution binomiale négative sera utilisée pour approximer le modèle de comptage.

Avec un seuil de 10 %, les coefficients estimés avec la régression *hurdle* sont tous significatifs comme le montre le Tableau III-7. Les résultats sur le modèle en zéro sont meilleurs que ceux trouvés avec la régression *zero inflated* effectuée précédemment (cf. Tableau III-6).

Modèle de comptage					
	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>Constante</i>	-24,750	14,643	-1,690	0,09099	✓
<i>MinMaxT20Moy</i>	0,435	0,027	16,028	<2e-16	✓
<i>tminMoy</i>	-0,048	0,023	-2,056	0,040	✓
<i>region2</i>	0,802	0,234	3,424	0,001	✓
<i>region3</i>	0,485	0,219	2,218	0,027	✓
<i>region4</i>	0,445	0,201	2,218	2,66E-02	✓
Modèle en zéro					
	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>Constante</i>	-1,654	0,283	-5,853	4,81E-09	✓
<i>MinMaxT20Moy</i>	0,226	0,022	10,145	<2e-16	✓
<i>tminMoy</i>	-0,086	0,020	-4,295	0,000	✓
<i>region2</i>	-0,776	0,179	-4,344	0,000	✓
<i>region3</i>	-1,570	0,171	-9,183	<2e-16	✓
<i>region4</i>	-0,667	0,170	-3,929	8,52E-05	✓

Tableau III-7 Résultats de l'estimation avec le modèle *hurdle*

1.2.4. Comparaison des modèles précédents

Les modèles estimés précédemment (Poisson, Binomiale négative ou « NB », *Hurdle* avec la binomiale négative ou « Hurdle-NB », *zero inflated* avec la loi binomiale négative ou « ZINB ») ont un nombre différent de paramètres. Le critère de comparaison utilisé est donc le critère d'Akaike.

Le Tableau III-8 montre que le critère AIC du modèle de poisson est très important, soit 38 286 alors que pour les autres modèles, ce critère reste inférieur à 8 000. Globalement, les valeurs du critère AIC sont les moins élevées pour les modèles *hurdle* et *zero inflated* (environ 7 244 et 7 370 respectivement). Il convient de noter que les critères AIC des régressions binomiale négative et *zero inflated* binomiale négative sont très proches. Cela peut être dû au fait que la composante en zéro du modèle *zero inflated* est peu significative. En effet, un seul coefficient est significatif dans le modèle en zéro de la régression *zero inflated* (cf. Tableau III-6). Finalement, le meilleur modèle est la régression *hurdle* avec la binomiale négative comme proxy de la distribution de comptage.

Modèles	Degrés de liberté	Log vraisemblance	AIC
<i>Poisson</i>	6	-19 137	38 286
<i>NB</i>	7	-3 683	7 379
<i>Hurdle-NB</i>	13	-3 609	7 244
<i>ZINB</i>	13	-3 672	7 370

Tableau III-8 Comparaison des différents modèles de comptage

L'intérêt principal de l'implémentation des modèles *hurdle* et *zero inflated* réside dans le comptage de zéros. Le Tableau III-9 détaille le nombre observé de zéros et les nombres estimés de zéros pour les différents modèles implémentés. S'agissant des modèles de comptage classique, la régression de poisson sous-estime beaucoup le nombre de zéros tandis que la régression binomiale négative surestime mais n'est pas loin du nombre observé. Sans surprise, les nombres estimés de

zéros par la NB et la ZINB sont proches. S'agissant des modèles avec excès de zéros, nous constatons que la régression *hurdle* estime exactement le nombre de zéros observés tandis que la régression ZINB surestime ce nombre de zéros. Cela est normal car pour la ZINB, il y a deux sources de zéros alors que pour Hurdle-NB, les zéros ont leur propre loi génératrice. En réalité, le nombre de zéros estimés sera toujours égal au nombre de zéros observés par construction du modèle *hurdle*.

Observés	Poisson	NB	Hurdle-NB	ZINB
716	375	771	716	777

Tableau III-9 Effectifs des nombres de sinistres nuls (observés et estimés)

Après un backtesting du modèle simulant le nombre de sinistres au travers d'une régression *hurdle*, il ressort que ce modèle sous-estime le nombre de sinistres sur l'ensemble des années. En effet, le total estimé de sinistres sur l'ensemble des années est égal à 81% du total historique de sinistres sur les années 2008 à 2012. En faisant le détail de cette estimation par année, on se rend compte que l'année 2012¹⁰, qui à elle seule compte près de 76% du nombre total de sinistres, est très largement sous-estimée dans le modèle, l'écart de prédiction étant de -72%. En effet, l'année 2012 a été très froide et compte les nombres les plus importants de sinistres. Le modèle *hurdle* a du mal à prendre en compte ces extrêmes. Il a en fait lissé les observations : il surestime très fortement les années où il y a eu peu de sinistres et sous-estime celles où il y a eu de nombreux sinistres. Or, nous souhaitons justement capter les événements extrêmes dans un modèle catastrophe. Même si en moyenne l'estimation semble pertinente, elle ne convient pas pour l'étude.

Les températures extrêmement froides et le nombre de sinistres importants observés en 2012 nous incitent à penser que les sinistres se déclenchent massivement dès qu'un seuil de température est atteint. Face à cette problématique et à la variabilité importante du nombre de sinistres, nous avons décidé d'implémenter différents modèles suivant un seuil de température afin de capter les nombres de sinistres élevés lorsque les températures sont extrêmement basses.

1.2.5. Modèles avec dépassement de seuil

L'objectif de cette partie est de déterminer un seuil à partir duquel la sinistralité observée est extrême. Pour ce faire, nous utiliserons la notion de *mean excess function*, empruntée à la théorie des valeurs extrêmes. Nous ne nous attarderons pas sur la théorie des valeurs extrêmes. Nous exposerons uniquement les généralités de cette méthode nécessaires à la compréhension de la *mean excess function*. Une fois le seuil déterminé, nous chercherons à estimer le meilleur modèle correspondant au déclenchement des sinistres au-delà du seuil. Nous ferons de même pour les valeurs de température en dessous du seuil. Nous obtiendrons finalement deux modèles différents suivant le seuil de température observé.

Dans la suite, nous ferons un rappel succinct des principaux résultats de la théorie des valeurs extrêmes qui seront utilisées pour aboutir à la détermination du seuil. Puis, les modèles retenus seront présentés.

La détermination du seuil

Dans la théorie des valeurs extrêmes, deux approches principales sont utilisées pour déterminer les extrêmes :

- L'approche des blocs maxima qui consiste à s'intéresser au maximum dans un échantillon X_1, \dots, X_n . Dans cette approche, le maximum est choisi périodiquement (par exemple annuellement). C'est une approche peu exploitable en assurance du fait de la rareté des sinistres extrêmes.

¹⁰ Année ayant générée des pertes extrêmes pour les sinistres de gel.

- L'approche Peak Over Threshold (POT) qui vise à répondre à la question de savoir : « Etant donné un sinistre extrême, à quel point ce sinistre est-il extrême ? ». Cette approche tient compte des dépassements de seuil.

Dans POT, au lieu de considérer le maximum M_n d'un échantillon X_1, \dots, X_n , on s'intéresse aux dépassements du seuil u_n notés N_n , avec $N_n = \sum_{i=1}^n 1_{(X_i > u_n)}$. Ainsi, on s'intéresse aux observations $(X_i - u_n)^+$ qui sont strictement positives. Ces observations sont caractérisées par une loi de Pareto généralisée (GPD).

La distribution de Pareto généralisée $GPD(\beta, \xi)$ se définit ainsi :

$$G_{\beta, \xi}(x) = \begin{cases} 1 - [1 + \xi(x/\beta)]_+^{-1/\xi} & \text{si } \xi \neq 0 \\ 1 - e^{-x/\beta} & \text{si } \xi = 0 \end{cases}$$

Où :

$$x \geq 0 \text{ si } \xi \geq 0$$

$$0 \leq x \leq -\frac{\beta}{\xi} \text{ si } \xi < 0$$

La Figure III-2 illustre les observations étudiées dans le cadre de dépassements d'un seuil u .

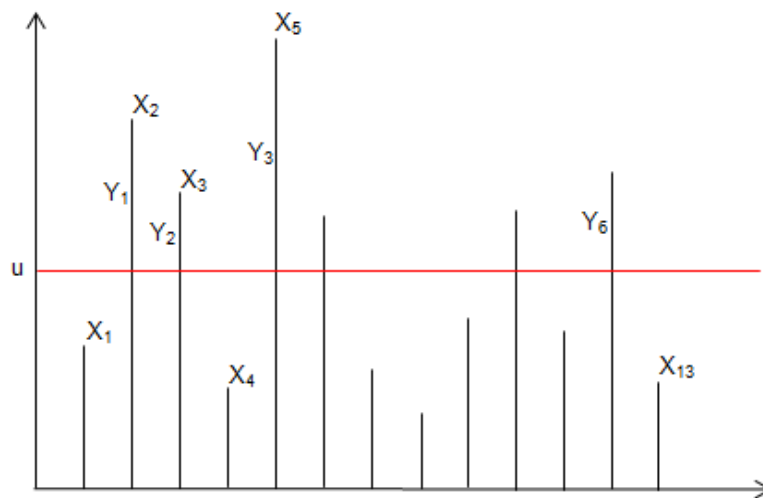


Figure III-2 Illustration de la méthodologie POT

Un outil graphique important pour étudier les extrêmes est la fonction de dépassement moyen (en anglais, *mean excess function*) ou durée de vie résiduelle. Elle se définit théoriquement par :

$$e(u) = E(X - u | X > u)$$

Lorsque X suit une $GPD(\beta, \xi)$, sa *mean excess function* est linéaire en u :

$$e(u) = E(X - u | X > u) = \frac{\beta + \xi u}{1 - \xi}$$

Empiriquement, la *mean excess function* se calcule ainsi :

$$e_n(u) = \frac{\sum_{i \in \Delta_n(u)} (X_i - u)}{\text{card } \Delta_n(u)}$$

Avec $\Delta_n(u) = \{i: i = 1, \dots, n, X_i > u\}$

La Figure III-3 illustre le tracé de la *mean excess function* pour la variable *tminMoy*. Pour faciliter la lecture graphique de la *mean excess function*, nous avons tracé *tminMoy*. En théorie, le seuil de température extrême est celui à partir duquel les températures peuvent être modélisées par une loi de Pareto généralisé. En pratique, le seuil choisi doit correspondre à un point de cassure dans l'allure de la *mean excess function*. Nous fixons donc notre seuil à $-5,5^\circ$.

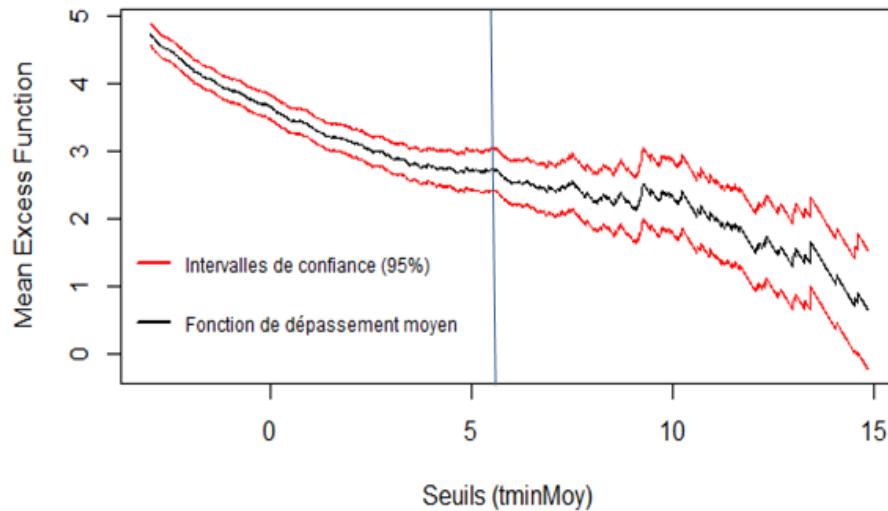


Figure III-3 Mean Excess fonction de *tminMoy*

Modèles suivant le seuil

Le seuil pour la moyenne régionale des températures minimales est fixé à $-5,5^\circ\text{Celsius}$. Le seuil a permis de partager nos données en deux sous-populations :

- les observations pour lesquelles $-5,5^\circ < tminMoy < 3^\circ$, soit 1 352 observations ;
- les observations pour lesquelles $tminMoy \leq -5,5^\circ$, soit 294 observations.

Pour chaque sous-population, on cherche le modèle qui ajuste le mieux le nombre de sinistres.

Nous effectuons d'abord la même démarche que précédemment sur le premier sous-échantillon (quand $-5,5^\circ < tminMoy < 3^\circ$) ; du fait de la surdispersion des nombres de sinistres et un excès de zéros dans les données. Nous ne présentons que les résultats du modèle avec les meilleurs résultats après backtesting. Il s'agit du modèle *hurdle* avec une composante de comptage approchée avec la loi binomiale négative. Le Tableau III-10 présente les résultats de l'estimation avec ce modèle *hurdle*.

Modèle de comptage					
	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>Constante</i>	-30,318	94,963	-0,319	0,74953	✘
<i>MinMaxT20Moy</i>	0,443	0,029	15,236	<2e-16	✔
<i>tminMoy</i>	0,040	0,038	1,058	0,290	✘
<i>region2</i>	0,744	0,240	3,096	0,002	✔
<i>region3</i>	0,695	0,234	2,968	0,003	✔
<i>region4</i>	0,415	0,209	1,981	4,76E-02	✔
Modèle en zéro					
	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>Constante</i>	-1,770	0,299	-5,922	3,17E-09	✔
<i>MinMaxT20Moy</i>	0,234	0,024	9,732	<2e-16	✔
<i>tminMoy</i>	-0,065	0,027	-2,369	0,018	✔
<i>region2</i>	-0,772	0,180	-4,291	0,000	✔
<i>region3</i>	-1,498	0,175	-8,558	<2e-16	✔
<i>region4</i>	-0,639	0,173	-3,702	2,14E-04	✔

Tableau III-10 Résultats du modèle *hurdle* ($tminMoy > -5,5^\circ$)

Ce tableau montre que tous les coefficients sont significativement différents de zéro au seuil de 10 %, exception faite de la constante et de la variable *tminMoy* dans la composante de comptage. Dans la composante de comptage, la variable de température qui influence le nombre de sinistres est la moyenne des écarts de température maximum et minimum sur une durée de 20 jours. Par ailleurs, le paramètre de dispersion $\hat{\theta}$ est nul.

Pour le second sous échantillon (quand $tminMoy \leq -5,5^\circ$), les modèles de comptage classiques sont les mieux adaptés car le problème d'excès de zéros dans la base de données a disparu. Par contre, les données sont toujours surdispersées. C'est donc la régression binomiale négative qui est la plus appropriée. Les résultats sont présentés dans le Tableau III-11. Au seuil de 10 %, les estimateurs des variables de températures sont tous significatifs mais les régions 3 et 4 ont des coefficients nuls. Par ailleurs, $\hat{\theta}$ est significativement différent de 0 au seuil de 10 %.

	Paramètres	Erreur	t-Student	p-valeur	Significativité
<i>Constante</i>	-15,887	0,639	-24,876	<2e-16	✔
<i>MinMaxT20Moy</i>	0,225	0,042	5,335	9,56E-08	✔
<i>tminMoy</i>	-0,251	0,050	-5,028	0,000	✔
<i>region2</i>	1,459	0,496	2,939	0,003	✔
<i>region3</i>	-0,560	0,355	-1,580	0,114	✘
<i>region4</i>	-0,031	0,352	-0,089	0,929	✘

Tableau III-11 Résultats de la régression binomiale négative ($tminMoy \leq -5,5^\circ$)

En cas de sinistralité extrême, le modèle avec dépassement de seuil (régression binomiale négative sur $tminMoy \leq -5,5^\circ$) fournit une meilleure estimation que le modèle *hurdle* sur toutes les observations (de la sous-section III.1.2.3). En effet, avec le modèle *hurdle* de la sous-section III.2.3, l'écart de prédiction sur l'année 2012 est de -72 %. En tenant compte du seuil, le modèle prédit près de 97% du nombre de sinistres de 2012, soit un écart de prédiction de -3%.

Pour conclure, au terme de cette section, nous retenons deux modèles pour caractériser la fréquence de gel. Il s'agit d'une régression *hurdle* lorsque $tminMoy > -5,5^\circ$ et de la régression binomiale négative lorsque $tminMoy \leq -5,5^\circ$.

2. Modélisation de la température

L'étude précédente a montré que la survenance d'un évènement gel dépend directement des températures observées et plus particulièrement de l'amplitude des décalages de températures. Il est donc nécessaire d'effectuer une modélisation de la température en France pour pouvoir en déduire l'apparition d'évènements gel.

Initialement, les données à disposition sont des relevés de températures minimales journalières depuis 1950 en 250 points équidistants en France. Il aurait cependant été fastidieux de modéliser la température en chacun de ces 250 points. Une approche raisonnable consiste à considérer les températures moyennes dans chaque région obtenue avec la classification de la section II. En effet, la classification a été réalisée dans le but d'obtenir des températures homogènes et très peu dispersées au sein de chaque région.

Finalement, nous avons donc quatre séries de températures journalières correspondant aux températures moyennes observées dans les quatre régions de la classification depuis 1950. Dans cette partie, chacune de ces séries sera modélisée afin de pouvoir simuler plusieurs scénarios de température pour l'année suivante. Nous pourrons en déduire, ensuite, la survenance d'évènements gel dans la suite.

Plusieurs approches seront étudiées ici, pour finalement retenir un modèle linéaire à variance périodique qui s'est révélé particulièrement adapté à nos données de température¹¹.

2.1. Présentation des données

Par souci de clarté, nous ferons toujours référence à la série temporelle associée à la région 1 (la région Sud) dans cette partie. Le raisonnement et la démarche suivie seront appliqués aux trois autres régions.

La série à modéliser correspond à la moyenne des températures minimales journalières observées dans la région 1 (la région Sud), du 1^{er} janvier 1950 au 2 juillet 2013, soit 23 193 données. Pour des besoins de cohérence dans la définition de la périodicité de la série, nous avons choisi de supprimer les 29 février afin d'avoir des années de taille équivalente (365 jours). Finalement, supprimer les 29 février revient à ôter 16 valeurs à l'ensemble des données, ce qui est négligeable en comparaison de la taille initiale de la série.

La Figure III-4 montre que la série étudiée présente une forte saisonnalité, ce qui est prévisible dans le cas de données de températures. Afin d'avoir une meilleure visibilité et une analyse plus précise, les graphiques présentés dans cette partie ne montrent pas toujours l'intégralité des données (depuis 1950).

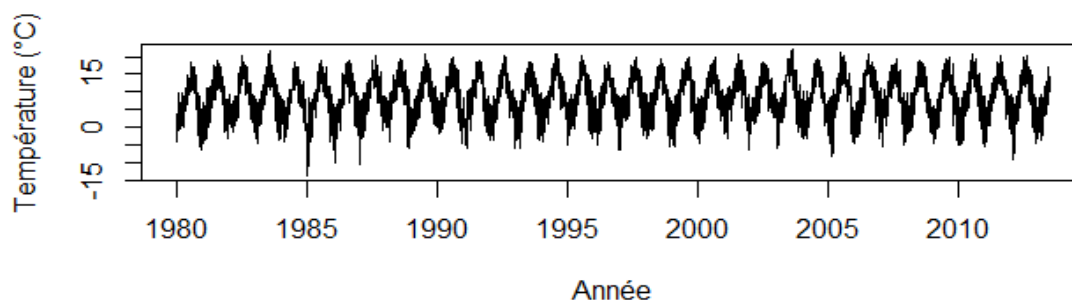


Figure III-4 Températures moyennes journalières observées dans la région 1 depuis 1980

¹¹Cette section s'inspire des travaux réalisés par ROUSTANT (2003).

2.2. Le modèle linéaire additif simple

Nous allons dans un premier temps utiliser l'approche Box & Jenkins (1976) pour décomposer notre série de données en une réalisation de plusieurs phénomènes simultanés. Nous décomposons donc notre série initiale de température (X_t) de la manière suivante :

$$X_t = m_t + s_t + Y_t$$

- Le premier phénomène est la **tendance** (m_t) qui représente l'évolution temporelle de la série. Il est vu aussi comme le comportement « basse fréquence » et permet de modéliser l'évolution de la série sur le long terme.
- Le deuxième phénomène est la **saisonnalité** (s_t) qui est une composante périodique de période T . La saisonnalité de période T vérifie $\forall t \in \mathbb{Z}, s_{t+T} = s_t$ et $\sum_{t=1}^T s_t = 0$. La saisonnalité permet, quant à elle, de modéliser le comportement sur le court terme de la série.
- Enfin, le troisième phénomène, appelé **série résiduelle** (Y_t), est la composante qui explique les écarts du modèle obtenu à partir de la tendance et de la saisonnalité. Elle peut se voir comme la réalisation d'un processus aléatoire stationnaire.

Cette décomposition est obtenue à l'aide du logiciel R. Pour cela, la tendance m_t est d'abord estimée par une méthode de régression locale¹². Cette méthode repose sur un algorithme qui permet de déterminer localement un modèle de régression non paramétrique adapté aux données locales. Puis, la saisonnalité s_t et la série résiduelle Y_t sont déduites de la série sans tendance $X_t - m_t$. Les résultats sont présentés dans la Figure III-5.

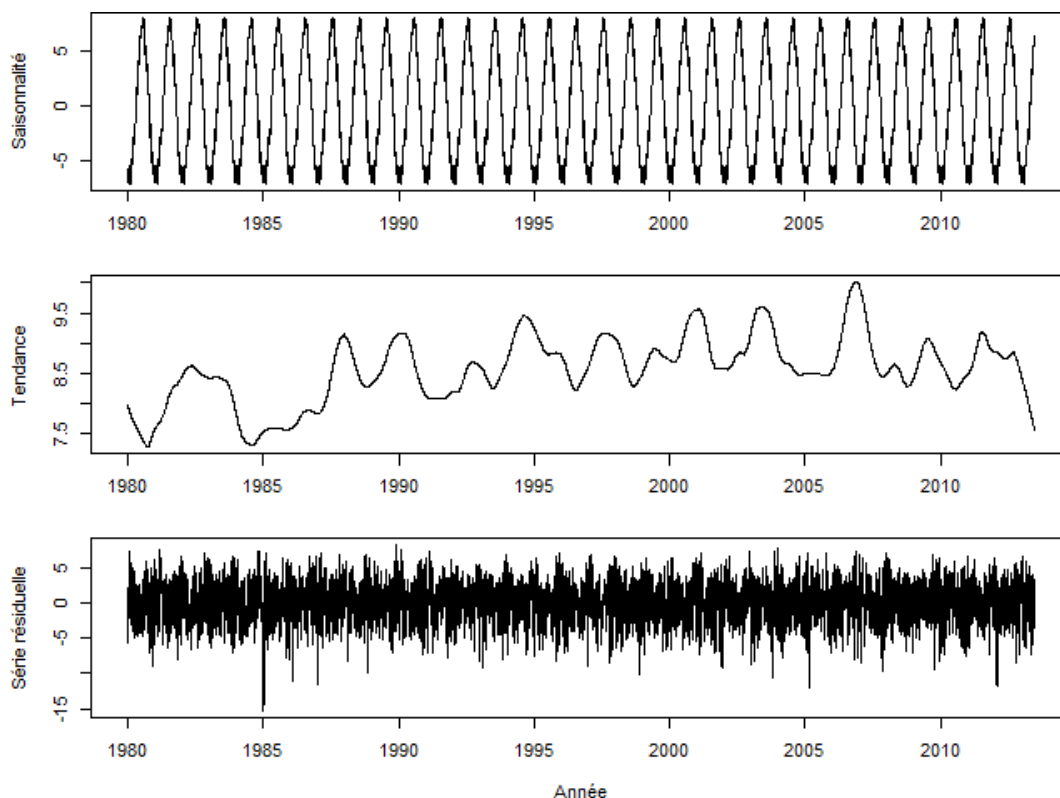


Figure III-5 Décomposition de la série

¹² Régression « LOESS » : Local regrESSion »

La Figure III-5 montre que la saisonnalité de la série étudiée est périodique annuellement, ce qui est cohérent pour des données de température. De plus, une tendance légèrement haussière est observée (2^{ème} graphique de la Figure III-5). Cela peut être dû au réchauffement climatique constaté depuis quelques années dans le monde. Cette tendance haussière sera confirmée dans la suite de l'étude. Enfin, la série résiduelle (3^{ème} graphique de la Figure III-5) ne semble présenter à première vue ni tendance, ni saisonnalité.

Pour pouvoir développer des prédictions intéressantes, il faut que la série résiduelle présente certaines propriétés structurelles : soit de la "rigidité" qui permet en partie d'extrapoler, soit une forme d'invariance statistique, qu'on appelle stationnarité, qui permet d'apprendre le présent et prévoir le futur à partir du passé. De manière générale, pour qu'un modèle soit correctement spécifié, il faut que la série résiduelle soit un bruit blanc, ce qui permet de prendre en compte les propriétés structurelles souhaitées. La série résiduelle est un bruit blanc (faible) si ses termes ont une moyenne nulle, une variance constante et la covariance de ses termes pris deux à deux est nulle. La série résiduelle est un bruit blanc fort si ses termes sont soit indépendants et identiquement distribués (iid) avec une moyenne nulle et de variance constante. On parle de bruit blanc gaussien lorsque les résidus sont iid et suivent une loi normale centrée réduite.

Pour vérifier si la série résiduelle est un bruit blanc, le test de Ljung Box ou test de Portmanteau est couramment utilisé. Le test de Ljung Box permet de vérifier l'indépendance sérielle entre les résidus. L'hypothèse nulle est que les coefficients de corrélation entre les résidus sont nuls. La p-value de ce test étant inférieure à 10^{-16} pour Y_t (Tableau III-12), cela nous amène à rejeter l'hypothèse nulle. La série résiduelle n'est donc pas un bruit blanc et le modèle linéaire simple est également rejeté. Le fonctionnement de ce test est fourni dans l'annexe B.

	Résultat
Test de non corrélation des résidus : p-value <i>Test de Ljung-Box</i>	$< 10^{-16}$ x

Tableau III-12 Test de Ljung Box

2.3. Le modèle ARMA

2.3.1. Définition

La série résiduelle n'étant pas un bruit blanc, nous allons tenter d'ajuster un modèle ARMA à Y_t . Les modèles ARMA (*Auto Regressive Moving Average*) ont été développés par Box et Jenkins en 1970 et représentent une classe très populaire de processus dans l'étude de séries temporelles. Ces processus ont, par exemple, l'avantage de pouvoir s'appliquer à de nombreux domaines, comme la météorologie, et sont facilement interprétables.

Un modèle ARMA(p,q) se décompose en deux processus : un processus AR (*Auto Regressive*) d'ordre p et un processus MA (*Moving Average*) d'ordre q. Soit (X_t) un processus ARMA(p,q), il vérifie l'équation suivante :

$$X_t = \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Le processus ε_t est un bruit blanc et correspond au terme d'erreur. Le premier terme correspond à un processus AR(p) qui permet d'exprimer le processus en fonction de ses p valeurs passées. Enfin, le deuxième terme correspond au processus MA(q) qui permet d'exprimer le processus en fonction des q termes d'erreurs passés.

2.3.2. Estimation d'un modèle ARMA

Dans un premier temps, nous devons sélectionner les ordres p et q qui correspondent respectivement au terme autorégressif et au terme de moyenne mobile.

Les diagrammes des autocorrélations (ACF) et des autocorrélations partielles (PACF)¹³ sont représentés respectivement par les Figure III-6 et Figure III-7. Ils nous amènent à sélectionner un processus autorégressif. En effet, quand le diagramme des autocorrélations (ACF) décroît exponentiellement vers 0 et que le diagramme des autocorrélations partielles est nul à partir du $(p+1)^{\text{ème}}$ retard, alors un processus AR(p) doit être sélectionné. Ici, les conditions sont donc réunies pour sélectionner un processus autorégressif. De plus, les autocorrélations partielles (PACF) n'étant plus significatives à partir de l'ordre 3, nous pouvons supposer que Y_t suit un modèle AR(3), c'est-à-dire un processus auto-régressif d'ordre 3.

Intuitivement, un processus auto-régressif semble adapté à la modélisation de données de températures puisque cela signifie que la température un jour donné dépend de la température observée durant les jours précédents. Nous allons cependant vérifier si cette hypothèse est bien acceptable par quelques tests.

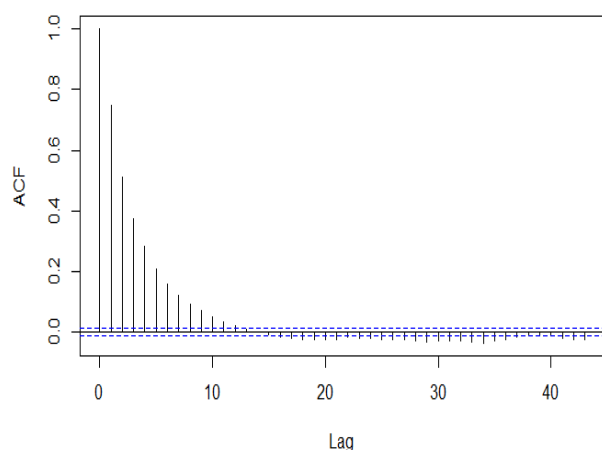


Figure III-6 Diagramme des autocorrélations de la série résiduelle

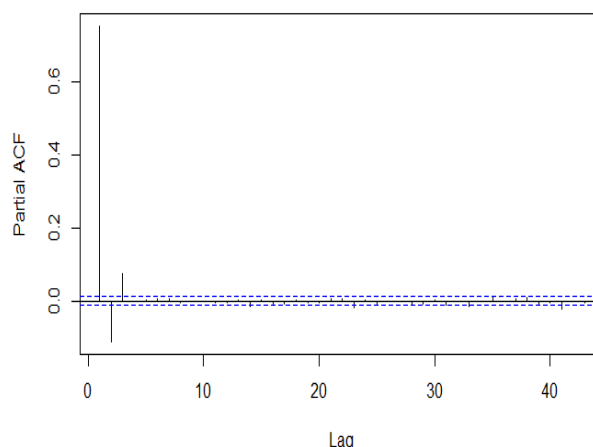


Figure III-7 Diagramme des autocorrélations partielles de la série résiduelle

Nous estimons donc les coefficients φ_1 , φ_2 et φ_3 en maximisant la vraisemblance afin que Y_t soit un processus auto-régressif d'ordre 3 :

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + \varepsilon_t$$

Le terme ε_t est communément appelé résidu.

Pour accepter ce modèle, plusieurs vérifications doivent impérativement être réalisées. Il faut tout d'abord que ε_t soit un bruit blanc gaussien :

- Les résidus doivent être stationnaires, c'est-à-dire de variance constante au cours du temps. Pour vérifier cela, nous utiliserons un critère graphique (le box plot).
- Les résidus doivent être indépendants, ce qui sera testé par le test de non corrélation de Ljung-Box.
- Les résidus doivent être gaussiens. Nous avons recours pour cela au test de normalité Kolmogorov-Smirnov.

¹³ Voir définitions de ACF et PACF en Annexe B - Fonctions d'autocorrélations et tests statistiques utilisés.

Enfin, les paramètres estimés doivent tous être significatifs. Cette propriété est vérifiée avec le test de Student.

Les fonctionnements de l'ensemble de ces tests sont détaillés dans l'annexe B et les résultats sont présentés dans le Tableau III-13:

	Résultats
Significativité des coefficients <i>Test de Student</i>	✓
Test de non corrélation des résidus : p-value <i>Test de Ljung-Box</i>	97,7% ✓
Test de normalité des résidus : p-value <i>Test de Kolmogorov-Smirnov</i>	5% ✓

Tableau III-13 Tests d'adéquation avec un modèle AR(3)

Afin de s'assurer de la stationnarité des résidus (ε_t), nous avons tracé, sur la Figure III-8, un box plot mettant en évidence cette propriété. Le box plot résume quelques caractéristiques de la série étudiée (médiane, quartiles, minimum, maximum et déciles) et permet ainsi de vérifier rapidement la stationnarité de la série. Pour cela, les valeurs prises par ε_t ont été regroupées par mois et les caractéristiques correspondantes sont mises en évidence sur le graphique. Cela permet d'observer la dispersion des valeurs de ce processus par mois. Il apparaît très clairement que les valeurs de ε_t sont plus dispersées pendant les mois d'hiver que pendant les mois d'été. Cela prouve donc que ε_t n'est pas un bruit blanc et le modèle ARMA est rejeté.

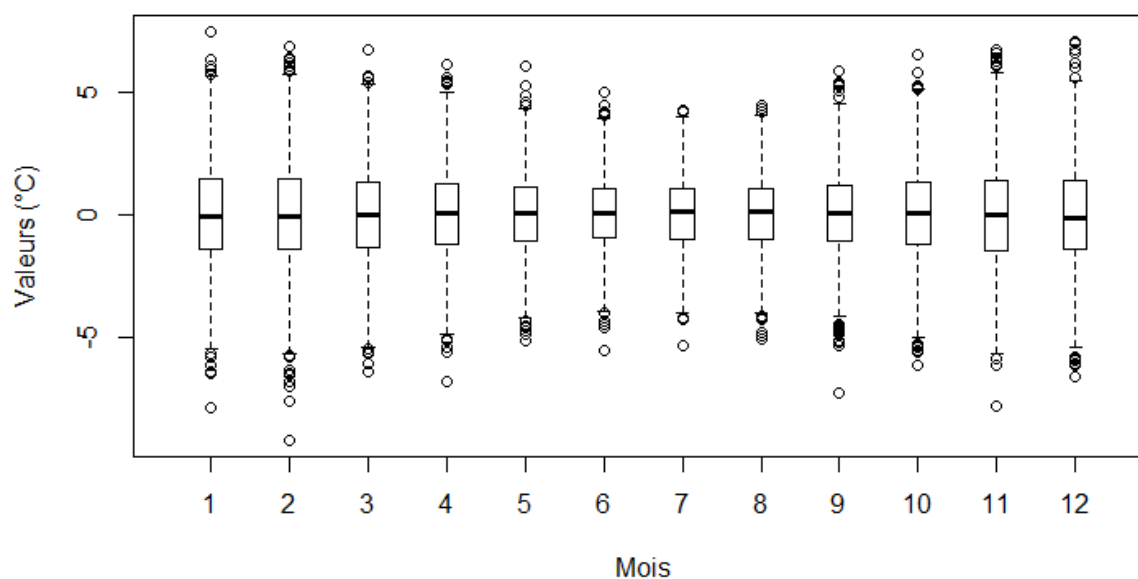


Figure III-8 Dispersion des résidus par mois

Il faudra donc prendre en compte cette dispersion dans la suite de l'étude pour obtenir un modèle convenable. De plus, la dispersion semble saisonnière car elle est visiblement plus importante pendant les mois d'hiver. Nous allons tester dans la section suivante un modèle linéaire à variance périodique pour tenir compte de la saisonnalité des résidus.

2.4. Le modèle à variance périodique

2.4.1. Introduction

La partie précédente a mis en évidence la non stationnarité des résidus dans le modèle ARMA. La série des températures étant périodique de périodicité annuelle, nous pouvons supposer que la variance des résidus est périodique de la même façon. Cette hypothèse sera vérifiée par la suite.

Nous proposons donc le modèle suivant :

$$X_t = m_t + s_t + \rho_t Y_t$$

Avec :

- X_t : la série initiale de températures
- s_t : la saisonnalité
- m_t : la tendance
- Y_t : un processus ARMA à estimer avec $\text{Var}(Y_t) = 1$. Dans la section précédente, $\text{Var}(Y_t)$ n'était pas constant.
- ρ_t : l'écart type de la série, supposé périodique de périodicité annuelle

Afin de faciliter la prédiction des séries dans la partie suivante, le modèle doit être entièrement paramétrique. Chacun des éléments ci-dessus sera donc modélisé dans les sections suivantes.

2.4.2. Tendance

Tout d'abord, la tendance de la série est modélisée par une droite. Pour cela, une régression linéaire est effectuée afin de trouver les coefficients a et b tels que :

$$m_t = at + b$$

Le résultat de la régression est illustré sur la Figure III-9 et les coefficients obtenus sont : $a = 6,668 \times 10^{-5}$ et $b = 7,372$.

Nous pouvons cependant nous interroger quant à la significativité de ces coefficients, particulièrement pour le coefficient a qui semble tellement faible qu'il pourrait éventuellement être nul en théorie. Pour vérifier cela, nous avons effectué des tests dont les résultats sont présentés en annexe C. Les résultats se sont révélés positifs, ce qui signifie que la tendance est bien haussière et confirme donc l'observation faite au paragraphe précédent.

La droite théorique permet juste de capter la tendance de la courbe originale, mais pas ses variations. Les écarts observés entre la droite théorique (en rouge) et la courbe originale (en noir) seront modélisés dans la série résiduelle associée au processus ARMA.

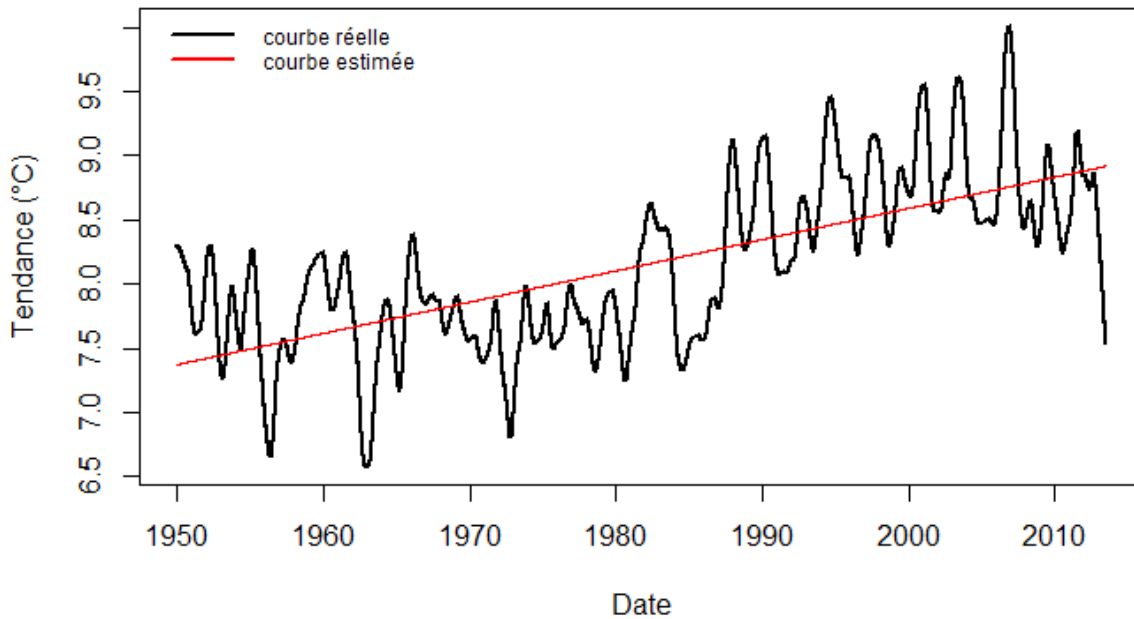


Figure III-9 Modélisation de la tendance

2.4.3. Saisonnalité

Nous savons que la saisonnalité s_t de la série est de périodicité annuelle. Ainsi, $s_t = s_{t+365} \forall t$. Nous cherchons donc à modéliser ici s_t pour $t \in [1, 365]$.

Néanmoins, nous pouvons chercher si une périodicité plus fine que la périodicité annuelle existe. Deux approches ont donc été testées :

- La périodicité annuelle en effectuant une régression linéaire sur $(\cos(\omega t), \sin(\omega t))$. Cette modélisation est représentée par la courbe bleue sur la Figure III-10.
- La périodicité semi-annuelle en effectuant une régression linéaire sur $(\cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t))$. Cette modélisation est représentée par la courbe rouge sur la Figure III-10.

$$\text{Avec } \omega = \frac{2\pi}{365}$$

Les coefficients estimés de la seconde régression étant tous significatifs (résultats du test de Student en annexe C), il apparait donc clairement qu'une périodicité semi-annuelle est plus représentative des données initiales.

Il existe donc des coefficients a_1, b_1, a_2 et b_2 tels que :

$$s_t = a_1 \cos(\omega t) + b_1 \sin(\omega t) + a_2 \cos(2\omega t) + b_2 \sin(2\omega t)$$

Finalement, les valeurs obtenues sont :

$$a_1 = -6,18, b_1 = -3,02, a_2 = 0,019 \text{ et } b_2 = 0,58.$$

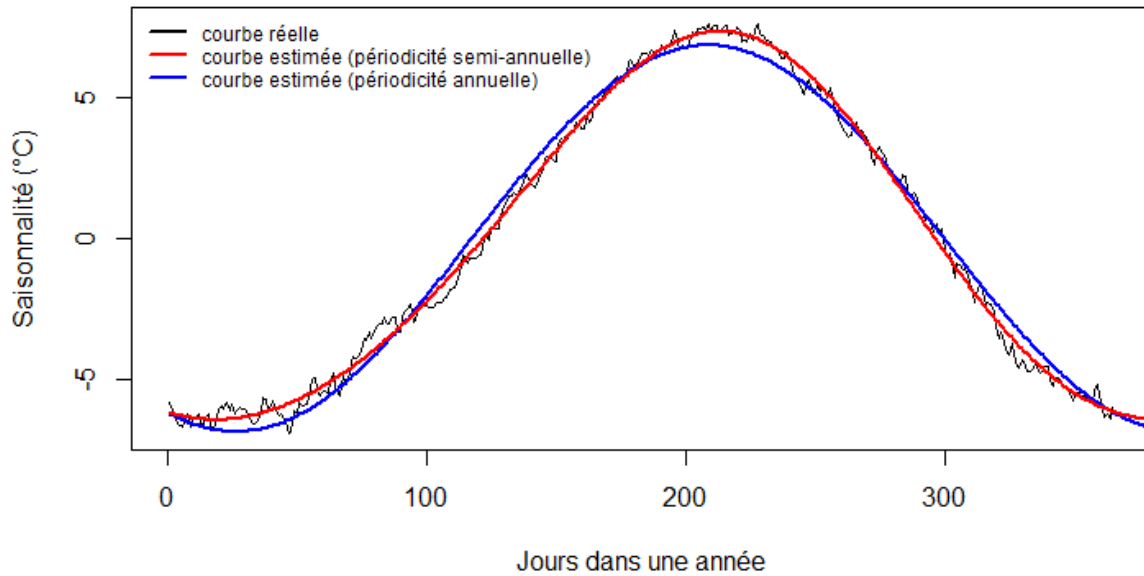


Figure III-10 Modélisation de la saisonnalité

2.4.4. Estimation de l'écart type

L'objectif de cette section est d'estimer l'écart type ρ_t par une fonction paramétrique telle que :

$$X_t = m_t + s_t + \rho_t Y_t \text{ et } \text{Var}(Y_t) = 1$$

Pour cela, nous définissons la série $(Z_t)_{1 \leq t \leq n}$ telle que : $Z_t = X_t - m_t - s_t = \rho_t Y_t$ avec m_t et s_t les fonctions paramétriques estimées dans les sections précédentes.

$$\text{Ainsi, } \rho_t = \sqrt{\text{Var}(Z_t)}.$$

Le box plot présenté à la Figure III-8 montre que la variance de la température est plus importante pendant les mois d'hiver. Nous supposons donc que ρ_t a une périodicité annuelle, et peut s'écrire de la façon suivante :

$$\rho_t = \alpha + \beta \cos(\omega t) + \gamma \sin(\omega t) \text{ avec } \omega = \frac{2\pi}{365}$$

Afin de calculer l'écart type de la série $(Z_t)_{1 \leq t \leq n}$, nous calculons un écart-type mobile centré en t . Cet écart type est dit mobile parce qu'il est recalculé de façon continue, en utilisant à chaque calcul un sous-ensemble d'éléments dans lequel un nouvel élément remplace le plus ancien. Pour cela, un paramètre de lissage h est déterminé et permet de définir le sous-ensemble sur lequel est calculé l'écart-type.

Soit $Z_{t,h}$ la série extraite de $(Z_t)_{1 \leq t \leq n}$ telle que $Z_{t,h} = (Z_t)_{t-h \leq t \leq t+h}$. Ainsi, ρ_t est défini comme l'écart type de $Z_{t,h}$ pour tout t appartenant à l'intervalle $[h, n - h]$:

$$\rho_t = \frac{1}{2h + 1} \sqrt{\sum_{k=-h}^h (Z_{t+k} - \mu_t)^2}$$

Avec μ_t la moyenne mobile définie de la même façon par :

$$\mu_t = \frac{1}{2h + 1} \sum_{k=-h}^h Z_{t+k}$$

Le paramètre h est bien un paramètre de lissage car plus h augmente, plus le nombre de données sur lequel est calculé l'écart type augmente et plus la série $(\rho_t)_{h \leq t \leq n-h}$ est lisse. La difficulté majeure est de trouver un compromis entre le lissage et la précision des données. Il faut que le lissage soit suffisant pour pouvoir déterminer un modèle de régression cohérent. Cependant, il ne faut pas que le paramètre soit trop élevé pour pouvoir capter aussi les valeurs singulières.

Afin de coller au maximum avec les données initiales et contrebalancer l'effet du lissage, nous avons choisi d'introduire une fonction de pondération dans le calcul de l'écart-type mobile afin d'attribuer un poids plus important aux valeurs proches de t . Cette fonction de pondération est obtenue en appliquant une pondération triangulaire et est illustrée sur la Figure III-11 : un poids de 1 est donné à la valeur de l'élément en t , et les poids sont de moins en moins importants à mesure que l'on s'écarte de t .

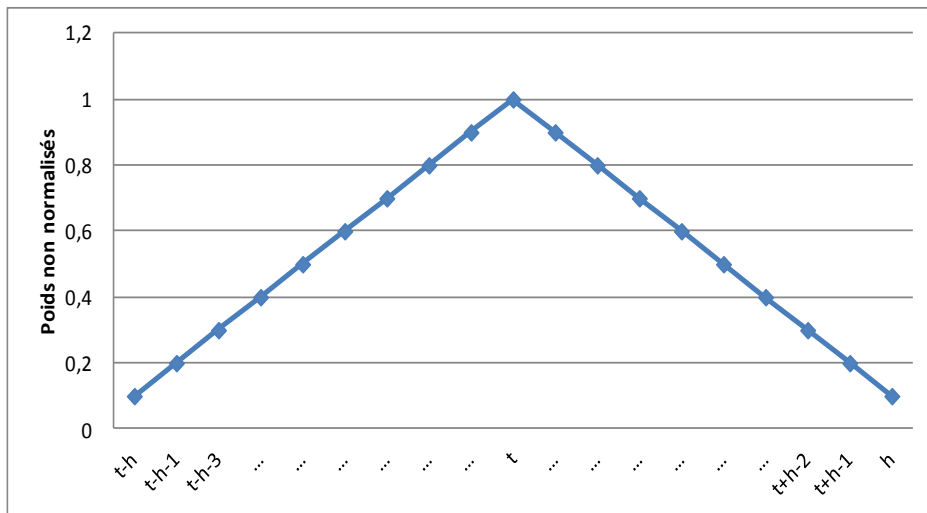


Figure III-11 Fonction de pondération triangulaire

La Figure III-12 permet de représenter ρ_t pour différents paramètres de lissage afin de sélectionner un paramètre optimal pour nos données. Nous pouvons observer que la différence entre des paramètres de lissage à 25 et à 50 est très importante alors que la différence entre un paramètre de lissage à 50 et à 75 est moins évidente. Nous pouvons ainsi considérer qu'un paramètre de lissage de 50 constitue un bon compromis pour nos données dans la mesure où cela permet de lisser suffisamment tout en conservant au mieux les variations de ρ_t .

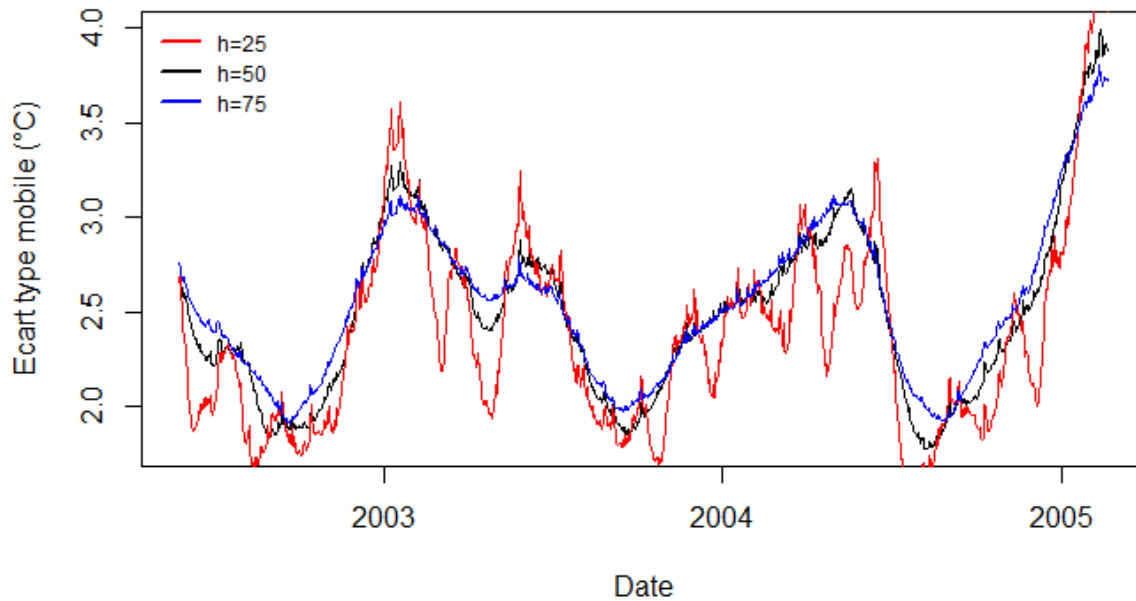


Figure III-12 Choix du paramètre de lissage

Une régression linéaire est ensuite effectuée sur $(\cos(\omega t), \sin(\omega t))$ pour trouver les coefficients α , β et γ qui se révèlent être tous significatifs par l'application d'un test de Student. La Figure III-13 montre que l'hypothèse faite au début de la section concernant la saisonnalité de la variance est bien vérifiée avec des données de température car une périodicité annuelle est clairement identifiée ici.

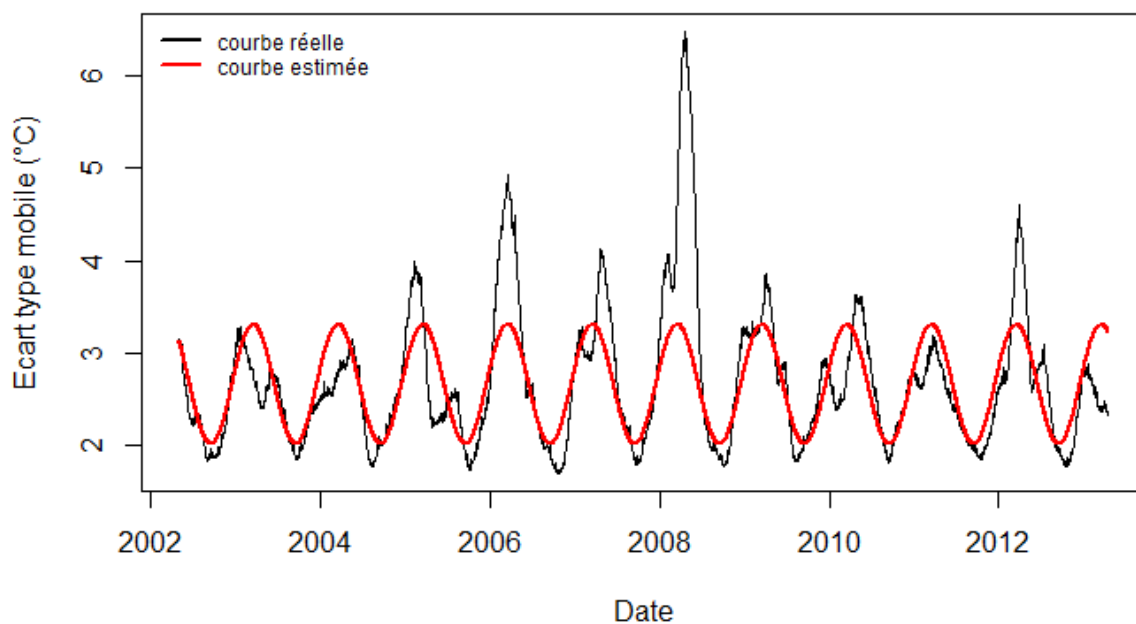


Figure III-13 Ecart type mobile – h=50

Les valeurs de ρ_t sont bien toujours positives, ce qui est une condition indispensable car ρ_t représente un écart-type. Les valeurs des coefficients obtenus sont les suivantes :

$$\alpha = 2,67, \beta = 0,64 \text{ et } \gamma = 0,07$$

Les résultats du test de Student pour vérifier la significativité de ces coefficients sont fournis en annexe C et sont positifs.

Nous pouvons nous demander encore une fois si les écarts entre la courbe réelle (en noir) et la courbe estimée (en rouge) ne fausseraient pas notre modèle : les pics observés sur la courbe réelle (par exemple en 2008) pourraient nous amener à penser que les pics extrêmes de température seraient atténués avec la courbe lissée dans notre modèle. Or, le modèle qu'on souhaite réaliser cherche justement à capter les pics de température afin de prévoir les pertes extrêmes dues aux périodes de grand froid. Ces pics seront en réalité pris en compte dans le modèle ARMA défini dans la section suivante.

2.4.5. Ajustement d'un modèle ARMA

Le dernier élément du modèle à paramétrer est Y_t tel que :

$$Y_t = \frac{X_t - m_t - s_t}{\rho_t}$$

Avec m_t , ρ_t et s_t les fonctions paramétriques définies dans les sections suivantes qui représentent respectivement la tendance, la variance mobile et la saisonnalité de la série initiale. Ainsi, le processus Y_t permettra de prendre en compte les écarts entre la fonction réelle observée (X_t) et les fonctions paramétriques estimées. Nous capterons ainsi implicitement les pics observés par rapport aux courbes réelles dans les sections précédentes, et donc de capter aussi la survenance de températures extrêmes.

Pour modéliser Y_t , nous allons ajuster un modèle ARMA et tester ensuite la validité de ce modèle. Comme dans le paragraphe précédent, nous sélectionnons les ordres p et q du modèle ARMA en traçant les diagrammes des autocorrélations (ACF) et des autocorrélations partielles (PACF) représentés sur les Figure III-14 et Figure III-15. Les diagrammes étant similaires à la section précédente, cela nous amène encore à sélectionner un processus autorégressif d'ordre 3.

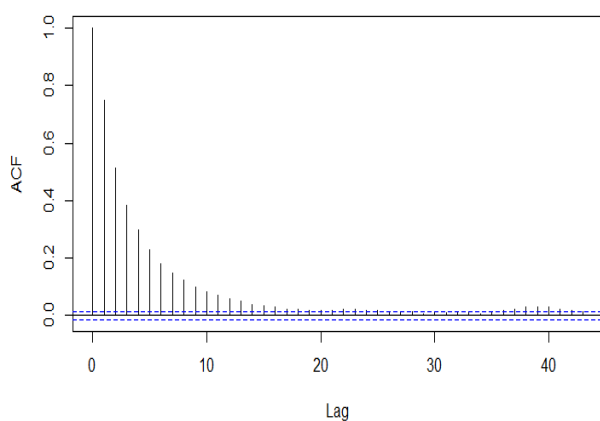


Figure III-14 Diagramme des autocorrélations

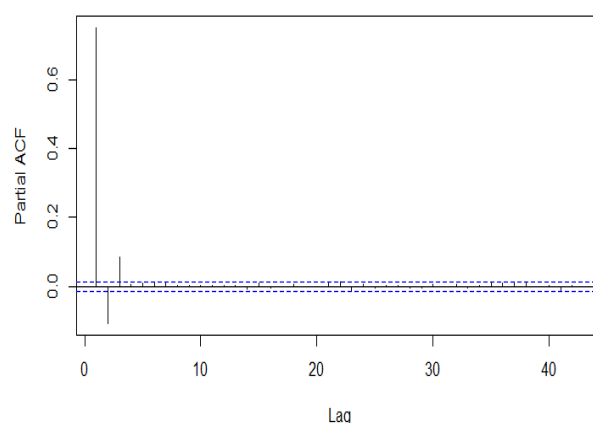


Figure III-15 Diagramme des autocorrélations partielles

Les diagrammes semblent identiques à ceux obtenus dans le paragraphe précédent car nous avons seulement extrait la saisonnalité de la variance des résidus dans ce nouveau modèle. Il existe donc des coefficients φ_1 , φ_2 et φ_3 tels que :

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + \varepsilon_t$$

Avec $\varphi_1 = 0,84$, $\varphi_2 = -0,17$, $\varphi_3 = 0,08$ et ε_t la série résiduelle. Les estimations de ces coefficients sont réalisées comme précédemment par maximisation de la vraisemblance. Les coefficients obtenus sont significatifs (résultats obtenus à l'aide du test de Student et détaillé en annexe C).

2.4.6. Validation du modèle

Pour valider le modèle, nous souhaiterions que ε_t soit un bruit blanc gaussien, ce qui implique la stationnarité du processus. Dans le modèle précédent, l'hypothèse de stationnarité n'était pas respectée. C'est précisément cela qui nous a conduit à tester le modèle à variance périodique.

Contrairement au box plot précédent (Figure III-8), la dispersion des résidus représentée sur la Figure III-16 semble constante pour tous les mois de l'année, ce qui valide l'hypothèse de stationnarité des résidus.

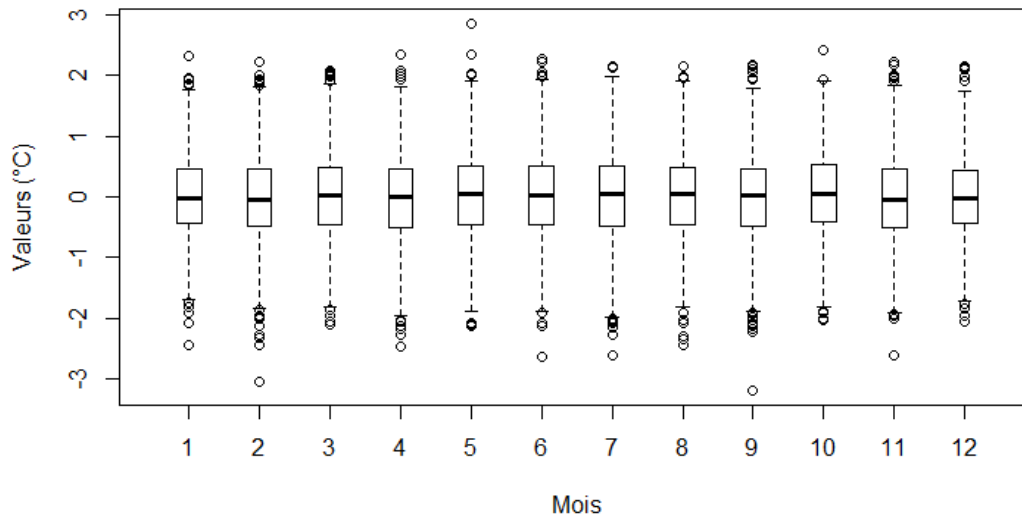


Figure III-16 Dispersion des résidus par mois

De plus, le test de Ljung-Box permet d'affirmer que la série résiduelle (ε_t) est bien un bruit blanc (Tableau III-14).

Enfin, pour vérifier le caractère gaussien de la série résiduelle, nous traçons un QQ-plot (Figure III-17). Ce type de graphique est souvent utilisé pour vérifier la cohérence d'une modélisation statistique. Pour cela, les quantiles théoriques suivis par une loi normale sont représentés par l'axe des abscisses et les quantiles relatifs à ε_t sont représentés par l'axe des ordonnées. Ainsi, plus les points tracés sont proches de la première bissectrice, plus l'estimation réalisée en affirmant que ε_t est gaussien est correcte. Sur la Figure III-17, les points sont très proches de la bissectrice, ce qui justifie bien l'hypothèse de normalité des résidus.

Par ailleurs, le test de Kolmogorov-Smirnov conduit aussi à accepter l'hypothèse de normalité des résidus (Tableau III-14).

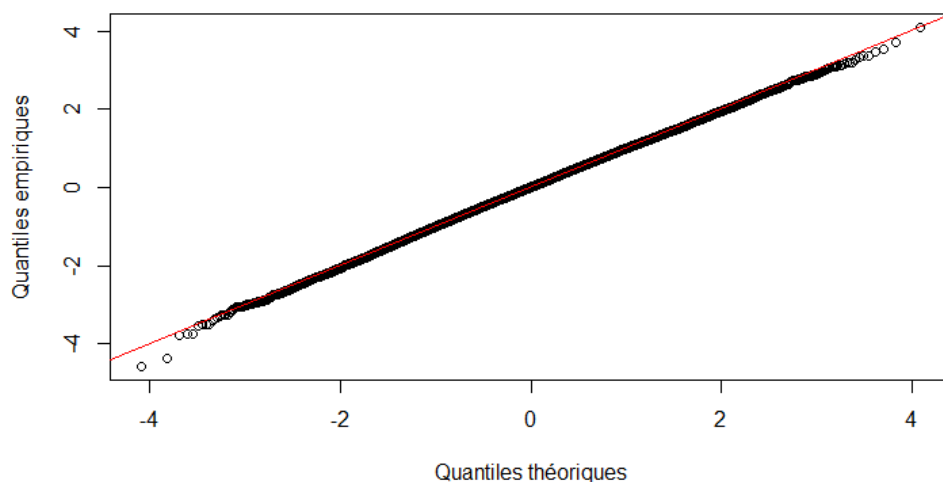


Figure III-17 QQ plot comparant la distribution de ε_t et la distribution d'une loi gaussienne

Enfin, nous testons comme précédemment la significativité des coefficients estimés du processus auto-régressif.

L'ensemble des résultats des tests effectués sont donnés dans le Tableau III-14 :

	Résultats
Significativité des coefficients <i>Test de Student</i>	✓
Test de non corrélation des résidus : p-value <i>Test de Ljung-Box</i>	98,2% ✓
Test de normalité des résidus : p-value <i>Test de Kolmogorov-Smirnov</i>	14% ✓

Tableau III-14 Tests d'adéquation avec un modèle AR(3)

Finalement, l'ensemble des hypothèses sont vérifiées et nous pouvons donc accepter ce modèle à variance périodique pour la modélisation de la série moyenne de températures pour la région 1 (région Sud). Le paragraphe suivant permettra d'adapter ce modèle aux trois autres régions.

2.5. Application de la méthode aux autres régions

Dans cette partie, le modèle à variance périodique est adapté à chacune des régions obtenues avec la segmentation réalisée dans la partie II. Nous allons comparer, pour chaque région, chacun des éléments définissant le modèle, puis nous vérifierons la validité du modèle.

2.5.1. La tendance

La tendance est l'élément principal qui va caractériser chaque région. En effet, les températures en France ne sont pas identiques, les régions au sud étant généralement plus chaudes que les régions se trouvant au nord.

Sur la Figure III-18, nous pouvons clairement voir que la tendance est différente pour chaque région. Comme nous pouvions nous y attendre, la région montagneuse (en bleue) a la tendance la plus faible, alors que la région du sud (en vert) a la tendance la plus élevée.

Cependant, la pente associée à chaque région est positive, ce qui confirme bien la tendance légèrement haussière des températures avec le temps. En effet, le réchauffement climatique touche bien de la même façon l'ensemble des régions dans le monde.

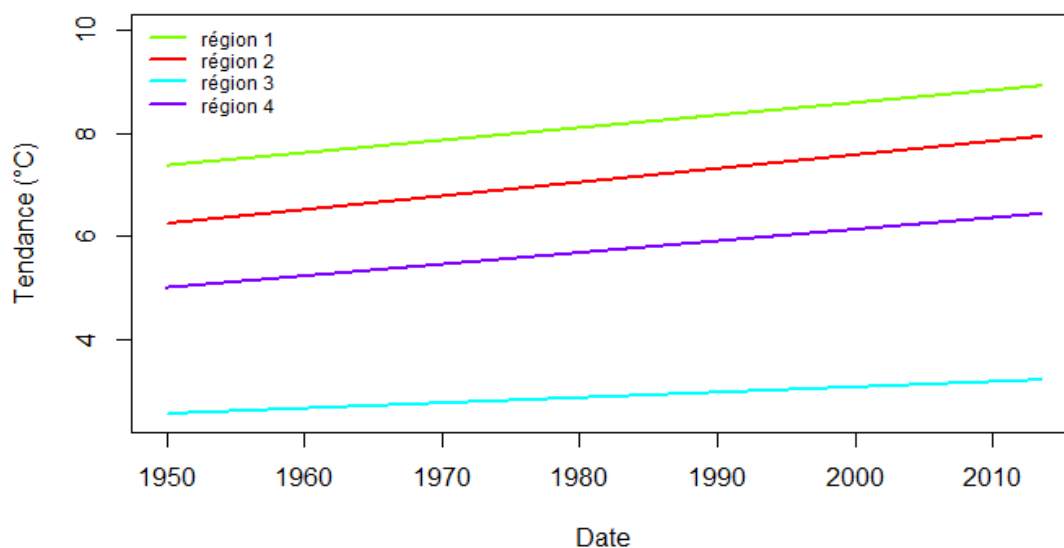


Figure III-18 Comparaison des tendances dans chaque région

2.5.2. La saisonnalité et l'écart-type mobile

La saisonnalité et l'écart-type mobile semblent similaires dans chaque région au regard des Figure III-19 et Figure III-20 puisque les courbes sont très proches, et se confondent même parfois. L'effet saisonnier caractéristique de la température impacte toutes les régions bien que nous pouvons constater de légères différences entre les régions.

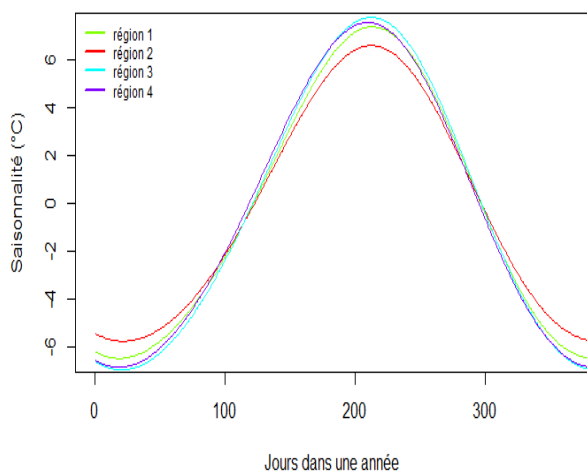


Figure III-19 Comparaison de la saisonnalité

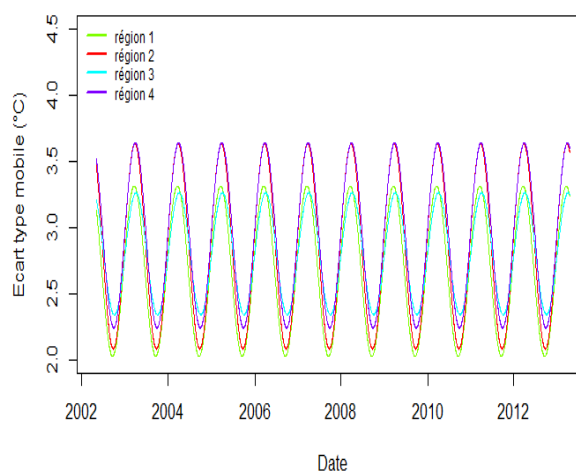


Figure III-20 Comparaison de l'écart-type mobile

2.5.3. Ajustement d'un modèle ARMA

De la même façon que dans la partie précédente, nous traçons les diagrammes des autocorrélations (ACF) et des autocorrélations partielles (PACF) pour sélectionner un modèle ARMA cohérent. Les Figure III-21 et Figure III-22 nous indiquent, comme précédemment, qu'un processus auto-régressif d'ordre 3 est le processus le mieux adapté à nos données.

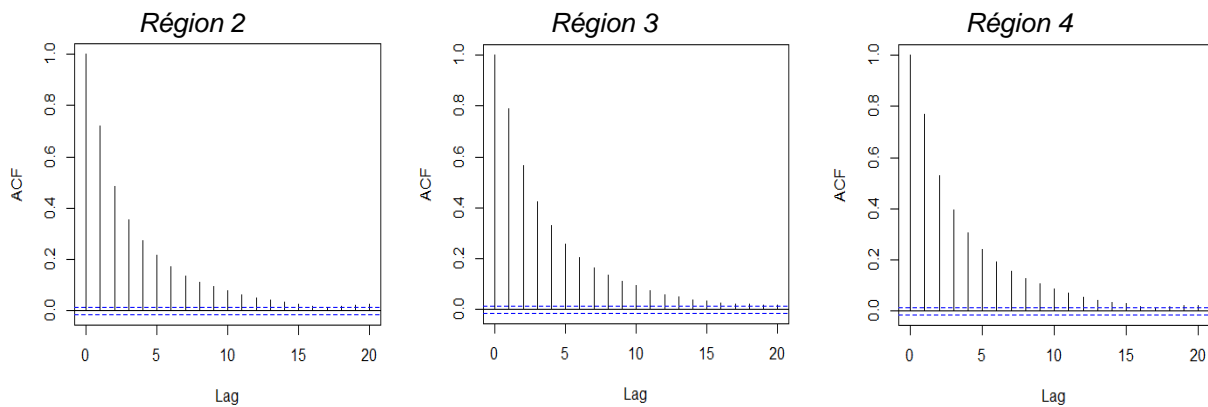


Figure III-21 Diagramme des autocorrélations pour chaque région

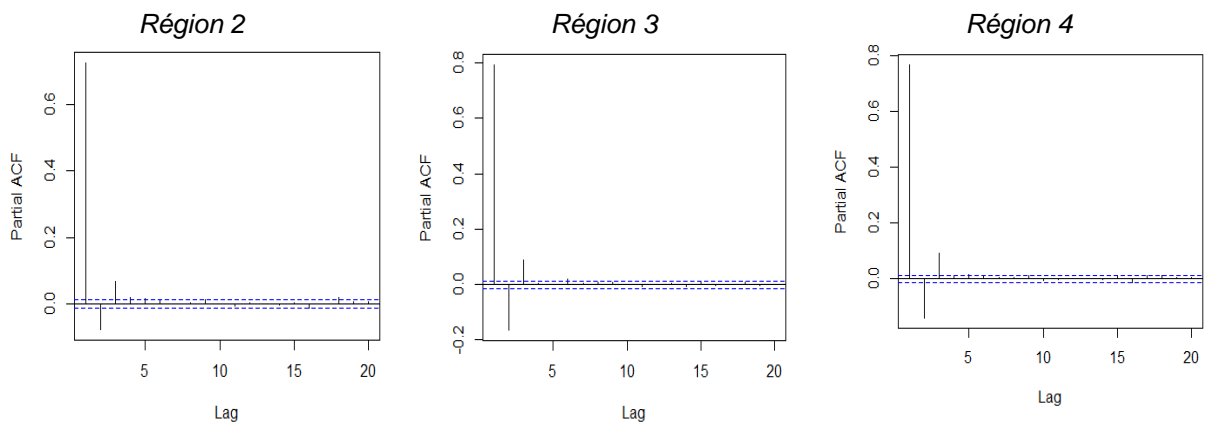


Figure III-22 Diagramme des autocorrélations partielles pour chaque région

Pour pouvoir accepter les modèles sélectionnés, il faut à présent s'assurer que certaines propriétés sont vérifiées.

2.5.4. Validation

Comme dans la partie précédente, afin de valider le modèle, il faut s'assurer que, pour chaque région, les résidus sont des bruits blancs gaussiens.

Dans un premier temps, les p-value du test de Ljung-Box nous assurent que les résidus associés à chaque région sont bien des bruits blancs (Tableau III-15).

Nous vérifions par ailleurs que les résidus associés à chaque région suivent bien une loi gaussienne grâce aux QQ-plot représentés sur la Figure III-23. En effet, les points sont proches de la bissectrice sur chacun des trois graphiques ce qui confirme le caractère gaussien des résidus. Cependant, l'hypothèse pourrait être remise en question pour la région 3 qui présente des points un peu éloignés de la bissectrice aux extrémités mais les résultats du test de Kolmogorov-Smirnov (Tableau III-15) nous permettent tout de même d'accepter l'hypothèse de normalité.

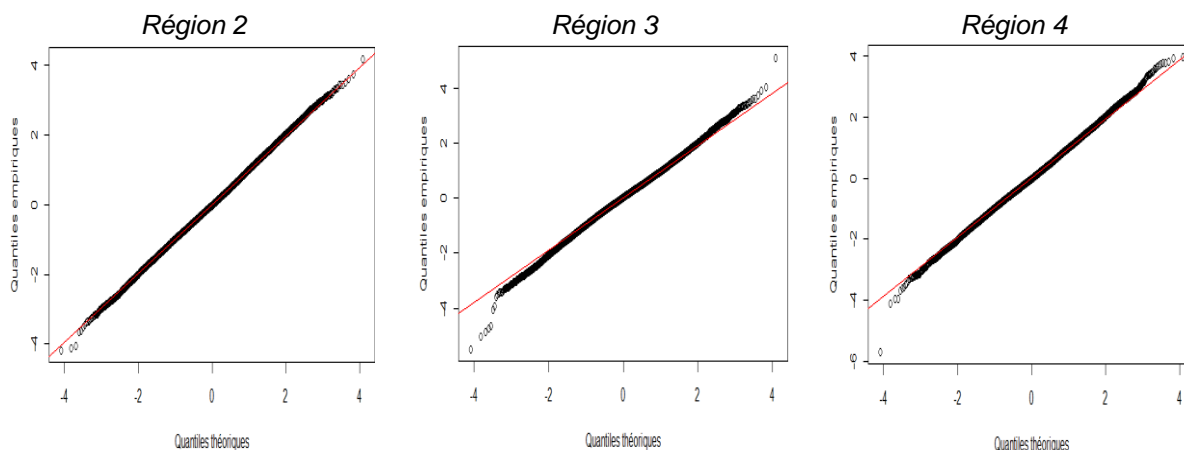


Figure III-23 QQ Plot permettant de vérifier le caractère gaussien des résidus de chaque région

Enfin, l'hypothèse de stationnarité des résidus est vérifiée comme précédemment à l'aide de box plots. La Figure III-24 montre que la dispersion des résidus semble globalement constante, ce qui permet d'accepter l'hypothèse de stationnarité des résidus :

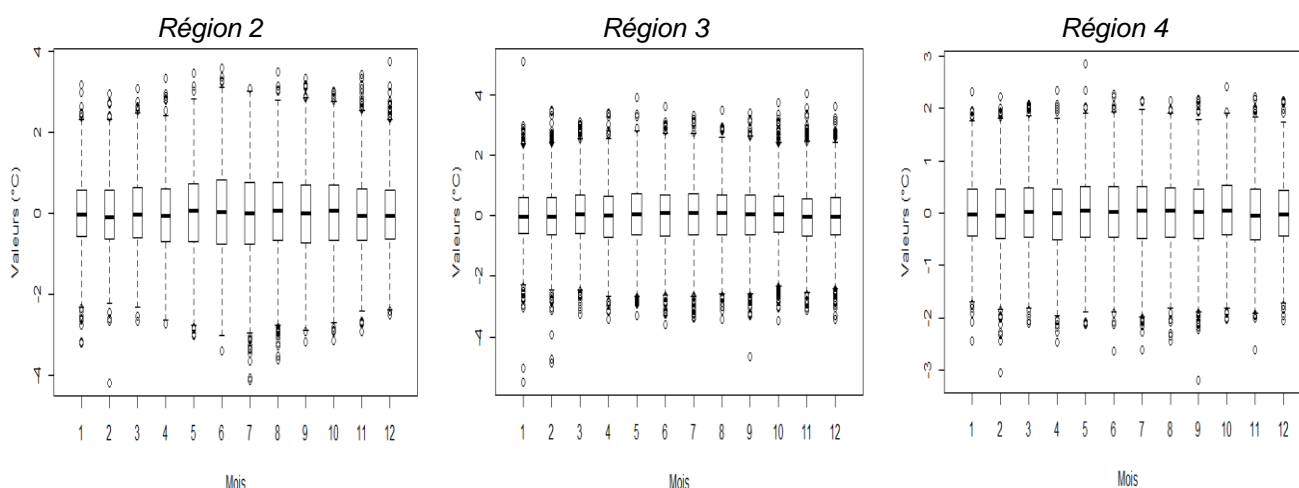


Figure III-24 Dispersion des résidus par mois

Les résultats des tests pour les trois régions sont répertoriés dans le Tableau III-15 :

	Région 2	Région 3	Région 4
Significativité des coefficients <i>Test de Student</i>	✓	✓	✓
Test de non corrélation des résidus : p-value <i>Test de Ljung-Box</i>	96,3% ✓	99,6% ✓	98,3% ✓
Test de normalité des résidus : p-value <i>Test de Kolmogorov-Smirnov</i>	15,7% ✓	8,1% ✓	19,2% ✓

Tableau III-15 Tests d'adéquation pour les régions 2, 3 et 4

Finalement, le modèle à variance périodique est bien transposable à chaque région. Ce modèle sera donc utilisé par la suite pour générer des scénarios de températures pour l'année suivante.

Néanmoins, l'utilisation de séries temporelles pour capter les températures extrêmes pourrait être contestée. Ce choix peut cependant être justifié par plusieurs raisons. En effet, la longueur des séries de températures utilisées permet de prendre en compte à la fois le réchauffement climatique et l'ensemble les vagues de froid survenues depuis 1950. Cela constitue des informations importantes pour générer ces valeurs extrêmes de températures. En outre, la modélisation des séries temporelles a bien intégré la cyclicité des températures et la dispersion saisonnière. Enfin, dans la section III.1,

nous avons observé que les scénarios extrêmes étaient principalement liés au franchissement d'un certain seuil de température ($-5,5^{\circ}\text{C}$), ce que les séries modélisées sont à même de restituer.

3. Modélisation de la dépendance entre régions

La partie précédente nous a permis de déterminer une modélisation statistique de la température dans quatre régions de France. La prochaine étape sera donc de simuler différents scénarios de température à partir de ces quatre modèles. Cependant, les températures dans les quatre régions ne peuvent pas être simulées indépendamment. En effet, la température est un phénomène homogène. Si un pic de température est observé dans une région de France, il est très probable que les autres régions soient aussi affectées par ce pic. Par conséquent, nous devons également modéliser la dépendance entre les différentes régions pour pouvoir effectuer des prédictions cohérentes.

La modélisation statistique des régions a fait apparaître que le seul terme non déterministe du modèle est la série résiduelle de l'ARMA. Nous allons donc, dans cette partie, modéliser les quatre séries de résidus associées aux séries de température en prenant en compte la dépendance entre les régions. Pour cela, nous aurons recours à la théorie des copules qui permet de caractériser la dépendance entre plusieurs variables aléatoires.

3.1. Présentation des données

Nous avons à notre disposition quatre séries de résidus suivant des lois normales centrées réduites. Cette propriété a été vérifiée dans la partie précédente.

A priori, nous pouvons nous attendre à ce qu'il y ait une forte corrélation entre les températures des différentes régions. Nous allons donc vérifier cela en traçant les nuages de points des résidus deux à deux sur la Figure III-25 :

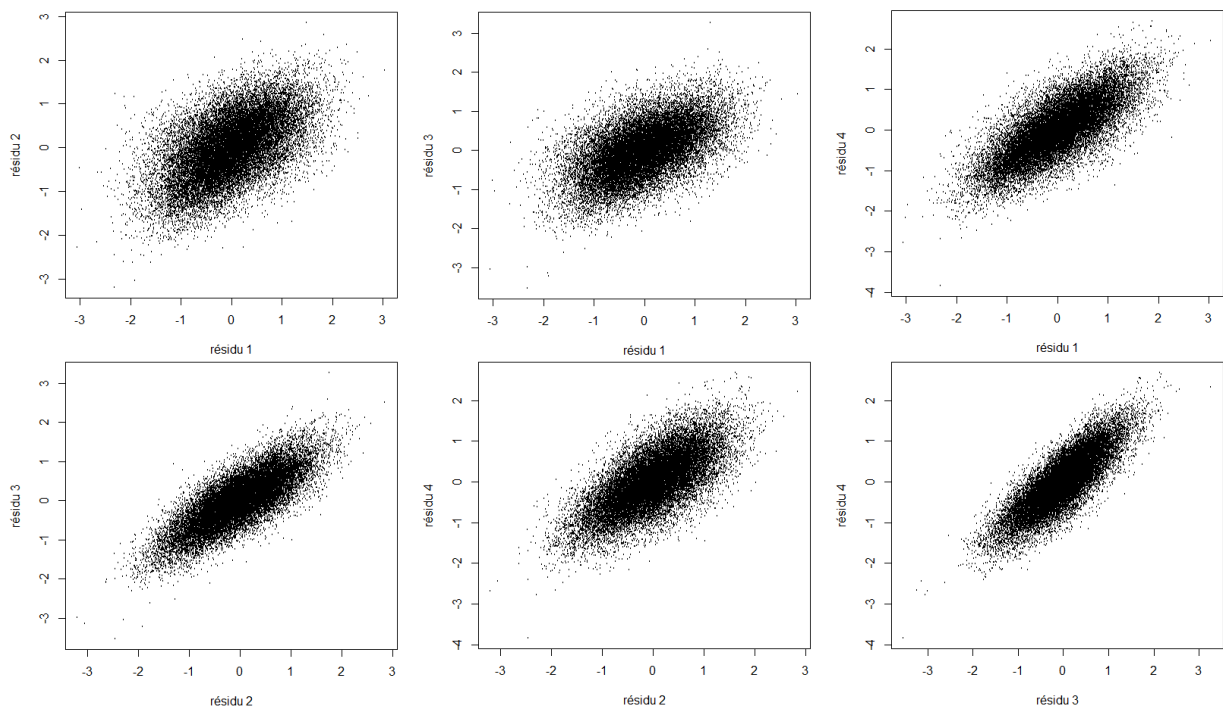


Figure III-25 Nuages de points des résidus empiriques

Nous pouvons remarquer que les corrélations entre les résidus semblent fortes, particulièrement entre les résidus des régions 3 (région Nord-Est) et 4 (région Montagneuse).

Afin de mesurer cette corrélation, nous utiliserons la corrélation de Spearman empirique représentée par le Tableau III-16. Le coefficient de corrélation de Spearman permet de mesurer la corrélation entre deux variables qui n'ont pas forcément de relation linéaire entre elles. Ainsi, le coefficient de corrélation de Spearman (appelé aussi communément Rho de Spearman) est défini comme le coefficient de corrélation linéaire entre les rangs des deux variables.

Rho de Spearman	Région 1	Région 2	Région 3	Région 4
Région 1	100%	55%	54%	74%
Région 2	55%	100%	78%	68%
Région 3	54%	78%	100%	82%
Région 4	74%	68%	82%	100%

Tableau III-16 Matrice de corrélation de Spearman empirique

Comme nous pouvions nous y attendre, les corrélations entre les résidus sont élevées, supérieure à 54%, les résidus les plus corrélés étant les résidus associés aux régions 3 et 4, ce qui était clairement identifié sur les graphiques précédents.

Enfin, nous observons le comportement des résidus empiriques sur les vingt-cinq derniers jours de données à notre disposition sur la Figure III-26. Cela permet de valider graphiquement la dépendance entre les résidus, particulièrement entre les résidus des régions 3 (région Nord-Est) et 4 (région Montagneuse) représentés en bleu et violet.

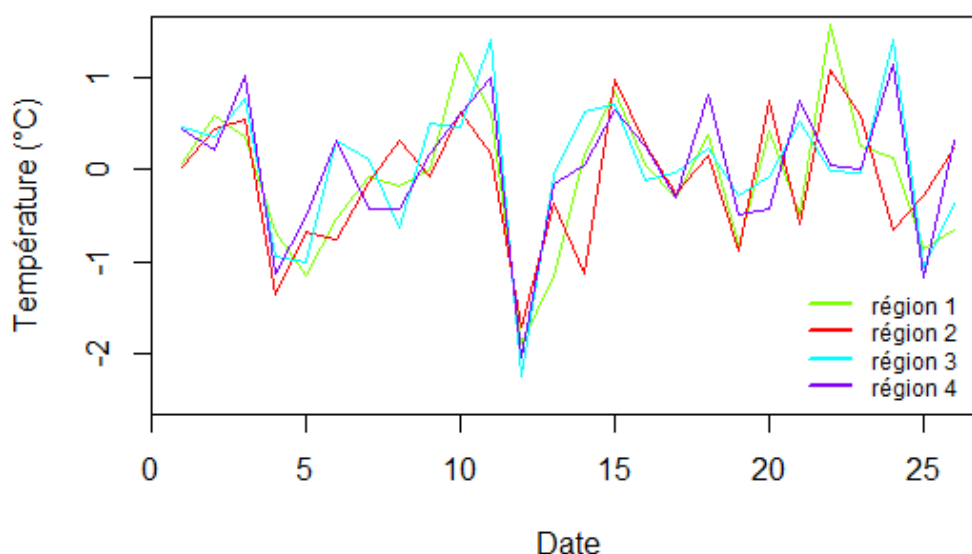


Figure III-26 Résidus observés les 25 derniers jours

3.2. La théorie des copules

Pour mesurer la dépendance, la corrélation linéaire de Pearson est fréquemment utilisée. Cette mesure est performante lorsque la relation de dépendance entre deux variables est linéaire, ce qui n'est souvent pas le cas dans la pratique. C'est pourquoi, nous avons recours à d'autres indicateurs se fondant sur les discordances et concordances observées dans un échantillon comme le tau de Kendall ou le rho de Spearman. Ce sont de bons indicateurs qui permettent de mesurer globalement la dépendance entre deux variables aléatoires à l'aide d'un nombre compris entre -1 et 1. Cependant, ces indicateurs contiennent l'ensemble de l'information de la dépendance entre deux variables dans

un seul nombre. Ils ne permettent donc pas de prendre en compte la complexité sous-jacente à la structure de dépendance empirique.

Pour remédier à cela, Abe Sklar a introduit la théorie des copules en 1959. Les copules fournissent une structure de dépendance plus précise et plus générale à des vecteurs de données. Plus précisément, elles permettent de « coupler » des lois marginales afin d'obtenir une loi multivariée qui correspond simplement à des fonctions de répartition en dimension quelconque.

Définition La copule bivariable C fonction de $[0,1]^2 \rightarrow [0,1]$ est définie par les caractéristiques suivantes :

(i) $C(u, 0) = C(0, u) = 0 \forall u \in [0,1]$

(ii) $C(u, 1) = C(1, u) = u \forall u \in [0,1]$: les marges des distributions marginales sont des marges uniformes.

(iii) C est 2-croissante : $C(v_1, v_2) - C(v_1, u_2) - C(u, v_2) + C(u, u_2) \geq 0$
 $\forall [u_1, u_2] \times [v_1, v_2] \in [0,1]^2$

Cette définition se généralise aisément en dimension supérieure à 2.

L'ensemble des résultats de la théorie des copules repose sur le théorème de Sklar qui précise le lien existant entre les fonctions de répartition marginales et la fonction de distribution multivariée :

Théorème Soit F la fonction de répartition multivariée du vecteur $X = (X_1, \dots, X_d)$ de lois marginales F_1, \dots, F_d . Alors il existe une copule C telle que :

$$\forall (x_1, \dots, x_d) \in \mathbb{R}^d, F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d))$$

Si, en plus, les lois marginales F_1, \dots, F_d sont continues, alors C est unique.

Ce théorème est fondamental car il permet de modéliser la fonction de répartition multivariée d'un vecteur de variables aléatoires et de modéliser ainsi la dépendance en deux étapes : modéliser d'abord les lois marginales, puis définir la structure de dépendance pour « lier » ces lois marginales. Grâce à la fonction de répartition multivariée, nous pourrions simuler différents scénarios de résidus tout en respectant les dépendances.

Pour cela, nous devons estimer les lois marginales F_1, \dots, F_d d'une part et la copule C d'autre part. Les lois marginales sont connues dans notre problème puisque nous avons montré, dans la partie précédente, que les résidus suivaient des lois normales centrées réduites. Il nous reste donc à définir la copule C qui permettra de modéliser la dépendance entre les résidus.

Cependant, déterminer la copule C n'est pas trivial. Dans la pratique, des copules paramétriques sont utilisées afin de simplifier le problème. En effet, les copules paramétriques permettent de décrire de nombreuses formes de dépendance et sont suffisantes dans de nombreux cas. Elles ont par ailleurs l'avantage d'être entièrement définies par un nombre fini de paramètres. Il suffit donc de calibrer les paramètres associés à la copule paramétrique choisie pour la définir. La difficulté majeure réside ainsi dans le choix de la copule paramétrique.

3.3. Les copules paramétriques

L'objectif ici est de déterminer la copule paramétrique la mieux adaptée à nos données et reflétant au mieux les relations de dépendance entre résidus. Il existe deux grandes familles de copules paramétriques : les copules elliptiques et les copules archimédiennes. La recherche de la copule optimale est réalisée sur un ensemble de quatre copules paramétriques sélectionnées *a priori* et présentées dans cette section, deux copules appartenant à la famille des copules elliptiques et deux copules appartenant à la famille de copules archimédiennes.

3.3.1. Copule elliptiques

Les copules elliptiques ont la particularité d'être relativement simples d'utilisation. Elles sont associées à une distribution multivariée symétrique, ce qui semble correspondre aux données de résidus (Figure III-25). Plus particulièrement, elles sont utilisées lorsque la distribution du vecteur considéré suit une distribution elliptique.

Définition Le vecteur aléatoire $X = (X_1, \dots, X_d)$ suit une distribution elliptique si et seulement s'il peut s'écrire sous forme affine :

$$X = AY + b$$

Avec:

- Y vecteur de dimension d suivant une loi sphérique ($OY =_{loi} Y$ pour toute matrice orthogonale O)
- A matrice racine carrée telle que $A'A = \Sigma$
- b vecteur réel de dimension d

Parmi les copules elliptiques, les copules gaussiennes ainsi que les copules de Student sont les plus répandues.

La copule gaussienne La copule gaussienne (ou normale) est définie par :

$$\forall (u_1, \dots, u_d) \in [0,1]^d, C(u_1, \dots, u_d) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$$

Avec Φ_{Σ} la fonction de répartition multivariée normale standard de matrice de corrélation Σ et Φ la fonction de répartition marginale de $N(0,1)$.

Cette copule n'a pas de dépendance de queue (sauf pour une corrélation parfaite de 1)¹⁴ et ne permet donc pas de corrélérer des valeurs extrêmes. Ainsi, cette copule n'est pas adaptée lorsque l'on souhaite modéliser une dépendance non linéaire ou entre événements extrêmes. Nous pouvons donc questionner son utilisation dans notre étude vu que nous désirons également modéliser des événements rares et extrêmes (le gel).

La copule de Student La copule de Student est définie par :

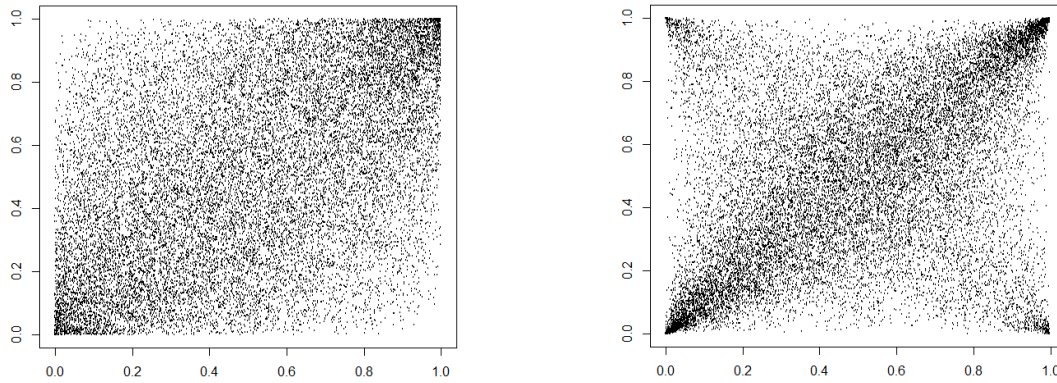
$$\forall (u_1, \dots, u_d) \in [0,1]^d, C(u_1, \dots, u_d) = T_{\Sigma, \nu}(t_{\nu}^{-1}(u_1), \dots, t_{\nu}^{-1}(u_d))$$

Avec $T_{\Sigma, \nu}$ la fonction de répartition d'une loi de Student multivariée de matrice de corrélation Σ et de ν degrés de liberté et t_{ν} la fonction de répartition univariée de ν degrés de liberté.

La copule de Student a l'avantage de mieux modéliser les dépendances de queue par rapport à la copule gaussienne et semble donc mieux adaptée à notre problème *a priori*. En effet, le paramètre ν permet de faire varier le niveau de dépendance des queues de distribution.

La Figure III-27 représente 20 000 simulations d'un couple de variables aléatoires modélisé par une copule gaussienne et une copule de Student.

¹⁴ Détails en annexe D.



Copule gaussienne de paramètre de corrélation 0,5 *Copule de Student de paramètre de corrélation 0,5 et de degré de liberté 1*

Figure III-27 Comparaison des copules gaussiennes et de Student de dimension 2

Nous pouvons donc observer sur la Figure III-27 la différence entre ces deux copules, la copule de Student permettant d'accentuer la dépendance des points extrêmes.

Les copules gaussiennes et de Student sont définies d'une part par la matrice de corrélation du vecteur de données Σ et par le degré de liberté ν pour la copule de Student. Nous avons à notre disposition quatre séries de données. Cela conduit donc à l'estimation de six paramètres pour la copule gaussienne et de sept paramètres pour la copule de Student¹⁵.

3.3.2. Copule archimédiennes

Les copules archimédiennes représentent une classe importante de copules. A la différence des copules elliptiques, les copules archimédiennes ont l'avantage de décrire des structures de dépendance très diverses prenant en compte notamment les dépendances asymétriques. Elles sont particulièrement utilisées pour décrire des événements extrêmes, ce qui pourrait convenir à notre étude.

De plus, contrairement aux copules gaussiennes et aux copules de Student, la fonction de répartition multivariée associée à une copule archimédienne n'est pas déduite du théorème de Sklar. Elles présentent des formes analytiques fermées et sont donc faciles à simuler :

Définition Une copule C est strictement archimédienne si et seulement si il existe une fonction φ continue, strictement décroissante, convexe, allant de $]0,1[$ à $[0, +\infty[$ avec $\varphi(0) = +\infty$ et $\varphi(1) = 0$, tel que :

$$\forall (u_1, \dots, u_d) \in [0,1]^d, C(u_1, \dots, u_d) = \varphi^{-1}\left(\sum_{i=1}^d \varphi(u_i)\right)$$

Dans ce cas, φ est un générateur strict de C .

Les copules archimédiennes sont donc entièrement définies par la fonction génératrice φ . Nous allons nous focaliser ici sur deux copules paramétriques de la famille des copules archimédiennes : la copule de Gumbel et la copule de Clayton. La Figure III-28 représente 20 000 simulations d'un couple de variables aléatoires modélisé par une copule de Gumbel et une copule de Clayton.

¹⁵ Pour d marginales, l'estimation de la matrice de corrélation nécessite $\frac{d(d-1)}{2}$ paramètres. La copule gaussienne nécessite donc $\frac{d(d-1)}{2}$ paramètres et la copule de Student en nécessite $\frac{d(d-1)}{2} + 1$, le paramètre supplémentaire correspondant au degré de liberté.

La copule de Gumbel La copule de Gumbel est définie par sa fonction génératrice φ :

$$\varphi(u) = (-\ln(u))^a \text{ avec } a \geq 1$$

Comme nous pouvons le constater sur la Figure III-28, la copule de Gumbel permet de modéliser des dépendances extrêmes à droite. La copule de Gumbel est entièrement définie par le paramètre a qui accentue la dépendance quand le paramètre augmente.

La copule de Clayton La copule de Clayton est définie par sa fonction génératrice φ :

$$\varphi(u) = \frac{u^{-a} - 1}{a} \text{ avec } a > 0$$

A la différence de la copule de Gumbel, la copule de Clayton permet de modéliser les dépendances extrêmes à gauche, comme nous pouvons le constater sur la Figure III-28. La copule de Clayton est aussi entièrement définie par le paramètre a qui accentue la dépendance quand il augmente.

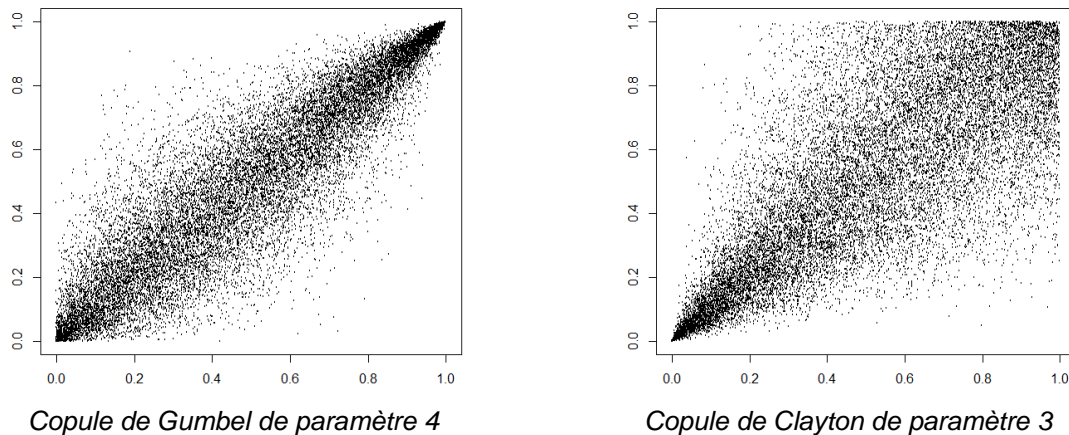


Figure III-28 Comparaison des copules de Gumbel et de Clayton de dimension 2

Ainsi, les copules archimédiennes présentent de nombreux avantages. Elles permettent d'une part de modéliser les dépendances extrêmes, ce qui est un critère important dans la modélisation de catastrophes naturelles. Elles permettent d'autre part d'avoir recours à l'estimation d'un seul paramètre, à la différence des copules gaussiennes et de Student qui en nécessitent six et sept. Cependant, un seul paramètre sera-t-il suffisant pour représenter la complexité de la dépendance des résidus ?

3.4. Méthode d'estimation

Qu'elles appartiennent à la famille elliptique ou à la famille archimédienne, les copules paramétriques sont toutes caractérisées par un nombre fini de paramètres. L'estimation des paramètres est donc une étape décisive dans la sélection de la copule optimale. Pour ce faire, nous avons choisi d'adopter la méthode CML (Canonical Maximum Likelihood) ou maximisation de la pseudo-vraisemblance. Cette méthode permet d'estimer les paramètres d'une copule sans estimer les fonctions de répartition des marginales, qui sont bien connues dans notre cas (des lois normales centrées réduites).

Soit α le vecteur de paramètres associé à la copule paramétrique que l'on souhaite calibrer. Nous estimons α par maximisation de la vraisemblance :

$$\hat{\alpha} = \arg \max_{\alpha} \sum_{i=1}^n \ln c(F_1(x_1^i), \dots, F_d(x_d^i))$$

Avec (x_1^i, \dots, x_d^i) pour $i \in \{1, \dots, n\}$, n observations du vecteur aléatoire qu'on souhaite modéliser et c la densité de la copule de paramètre α .

Nous avons sélectionné la méthode CML par rapport aux autres méthodes disponibles car ces dernières permettent d'estimer, en plus des paramètres de la copule, les lois marginales. Comme nous avons connaissance des lois marginales, leur estimation pourrait entraîner un biais dans l'estimation des paramètres. De plus, la méthode CML permet de réduire le nombre de paramètres à estimer ce qui donne des temps de calcul moins importants et des résultats plus précis.

Ainsi, pour chaque copule paramétrique sélectionnée, nous trouvons les paramètres associés estimés avec la méthode CML. Tout d'abord, les paramètres correspondants aux matrices de corrélation des copules gaussienne et de Student sont regroupés respectivement dans les Tableau III-17 et Tableau III-18. Dans chaque cellule nous pouvons lire deux chiffres. Le premier représente le coefficient de corrélation estimé pour les deux séries indiquées en ligne et en colonne. Le second, entre-parenthèses, indique l'écart-type estimé pour ce coefficient.

Matrice de corrélation Copule gaussienne	<i>Région 1</i>	<i>Région 2</i>	<i>Région 3</i>	<i>Région 4</i>
<i>Région 1</i>	100% (0%)	57% (6%)	54% (9%)	74% (5%)
<i>Région 2</i>	57% (6%)	100% (0%)	78% (17%)	71% (13%)
<i>Région 3</i>	54% (9%)	78% (17%)	100% (0%)	82% (11%)
<i>Région 4</i>	74% (5%)	71% (13%)	82% (11%)	100% (0%)

Tableau III-17 Estimation des paramètres de corrélation de la copule gaussienne avec la méthode CML

Matrice de corrélation Copule de Student	<i>Région 1</i>	<i>Région 2</i>	<i>Région 3</i>	<i>Région 4</i>
<i>Région 1</i>	100% (0%)	57% (7%)	55% (9%)	75% (6%)
<i>Région 2</i>	57% (7%)	100% (0%)	80% (20%)	71% (11%)
<i>Région 3</i>	55% (9%)	80% (20%)	100% (0%)	83% (9%)
<i>Région 4</i>	75% (6%)	71% (11%)	83% (9%)	100% (0%)

Tableau III-18 Estimation des paramètres de corrélation de la copule de Student avec la méthode CML

Nous remarquons que les paramètres de corrélation estimés pour les copules gaussiennes et de Student sont très proches. Par ailleurs, nous notons que la corrélation estimée entre les régions 3 (région Nord-Est) et 4 (région montagneuse) est bien la plus importante, ce qui confirme l'observation faite en début de section.

Ensuite, le paramètre estimant le nombre de degrés de liberté de la copule de Student est de 4,49, avec un écart-type estimé à 9%, ce qui est plutôt caractéristique d'une queue de distribution peu épaisse.

Enfin, les paramètres de dépendances α estimés pour les copules de Gumbel et de Clayton valent respectivement 2,79 et 1,18 avec des écarts-types de 5% et 8%.

3.5. Sélection de la copule optimale

3.5.1. Comparaison des résultats

Grâce à la méthode d'estimation présentée précédemment, nous avons trouvé pour chacune des quatre copules paramétriques sélectionnées, les paramètres correspondants maximisant la vraisemblance. L'objectif à présent est de comparer les résultats obtenus avec l'utilisation de ces quatre copules avec les données de résidus empiriques. Nous sélectionnerons ainsi la copule paramétrique qui modélise au mieux la structure de dépendance des résidus.

Pour ce faire, nous utiliserons deux types de critère : des critères graphiques et des critères quantitatifs qui permettront de comparer les résultats obtenus avec l'utilisation des différentes copules paramétriques étudiées.

Critères graphiques

La fonction de Kendall La fonction de Kendall est un outil permettant une comparaison directe entre la copule empirique et la copule théorique. Elle nous permettra de sélectionner rapidement les copules paramétriques ayant un comportement semblable aux données empiriques.

Soit (U, V) un couple de variables aléatoires de distribution jointe C et $Z = C(U, V)$. La fonction de Kendall est définie comme la fonction de répartition du couple le long de la première bissectrice :

$$K(t) = \mathbb{P}(C(U, V) \leq t) = \mathbb{P}(Z \leq t)$$

Nous supposons que nous disposons de n observations des variables U et V : soit (U_i, V_i) les observations obtenues pour $i \in \{1, \dots, n\}$. Nous pouvons estimer K de la façon suivante :

$$\hat{K}(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(Z_i \leq t)$$

Nous définissons Z de la façon suivante :

$$Z_i = \frac{1}{n-1} \sum_{j \neq i} \mathbf{1}(U_j \leq U_i, V_j \leq V_i)$$

Nous voulons à présent comparer les fonctions de Kendall empiriques avec les fonctions de Kendall théoriques pour chaque copule paramétrique étudiée. Nous traçons alors sur un même graphique la fonction de Kendall empirique (obtenue avec les données des résidus à notre disposition) en noir et la fonction de Kendall théorique (obtenue avec les données simulées par une copule paramétrique) en rouge.

La Figure III-29 représente, pour chaque copule paramétrique testée, la fonction de Kendall associée en rouge et la fonction de Kendall associée aux observations empiriques en noir. Nous nous limitons ici à l'étude de la dépendance des résidus correspondant aux régions 1 (région Sud) et 2 (région Nord-Ouest). Les fonctions de Kendall correspondants aux cinq autres couples de résidus ont donné des résultats similaires et sont disponibles en annexe E. Ainsi, plus les deux courbes sont proches, plus la copule paramétrique est adaptée.

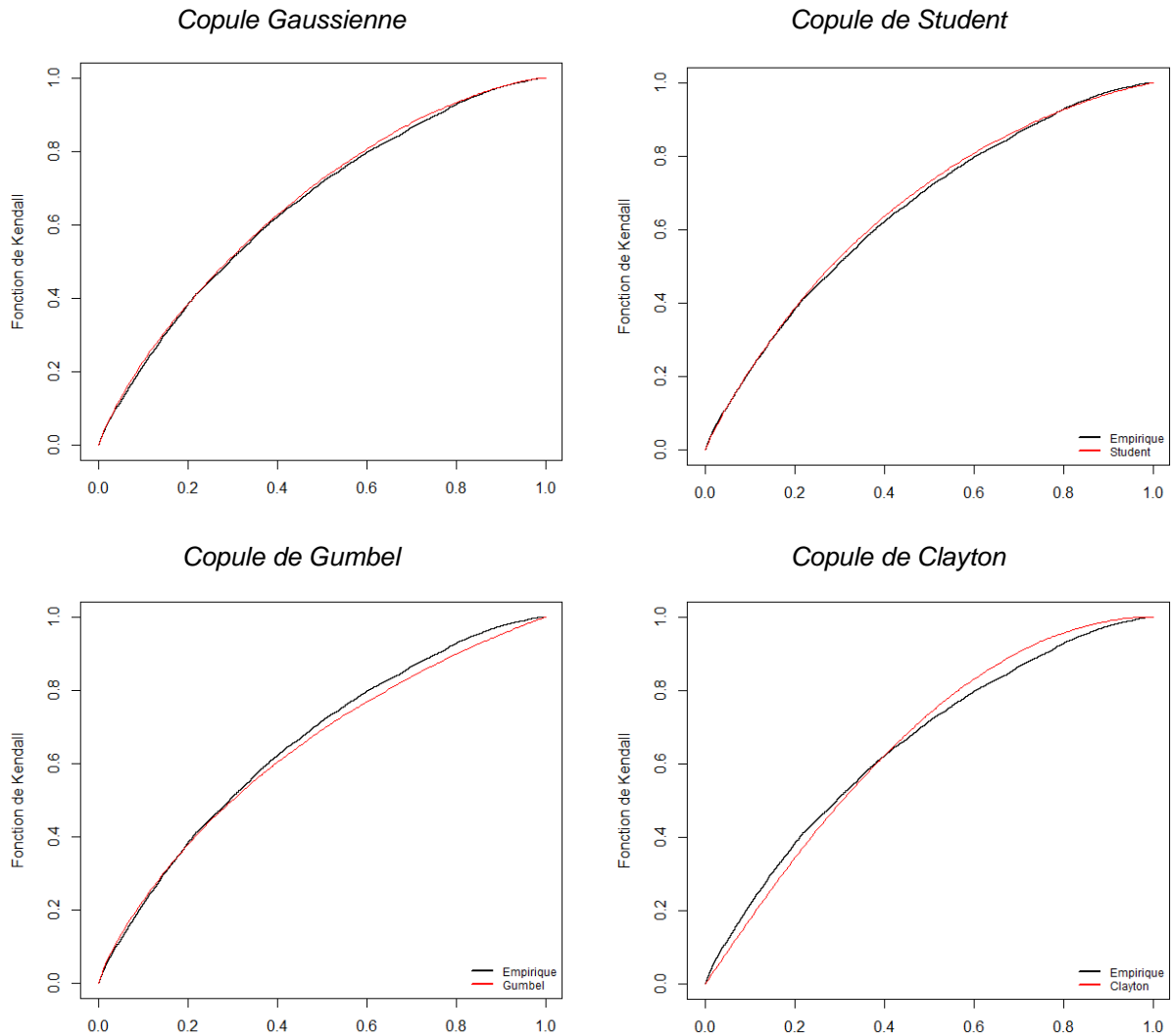


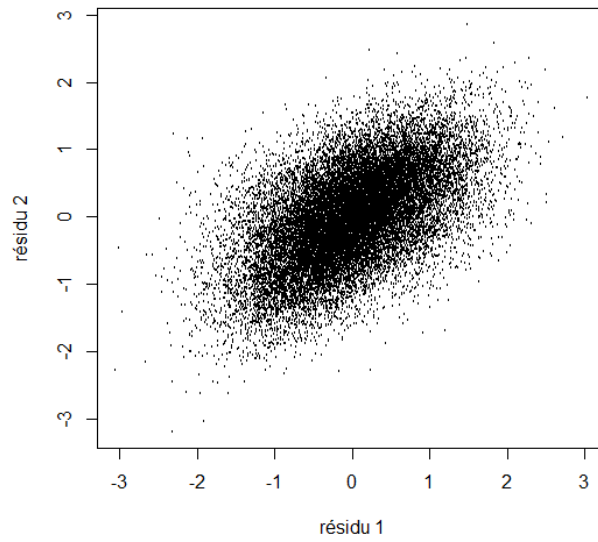
Figure III-29 Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 1 et 2

Nous pouvons ainsi clairement identifier ici que les copules elliptiques (gaussienne et Student) représentent le mieux les dépendances empiriques des résidus, la courbe rouge étant la plus proche de la courbe noire empirique pour ces copules.

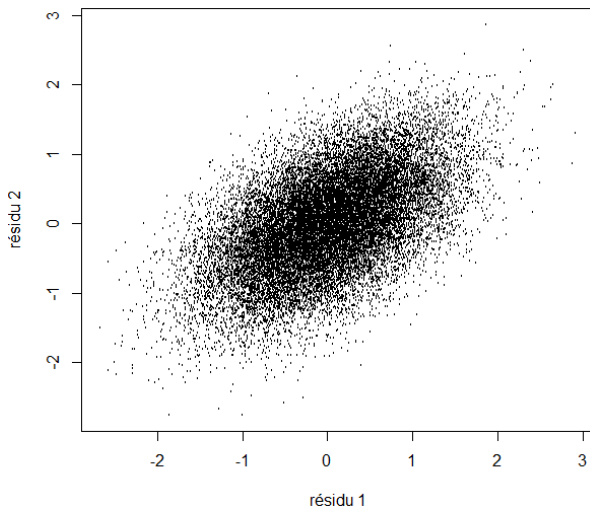
Le dépendogramme. Le dépendogramme représente la structure de dépendance sous la forme d'un nuage de points des simulations d'une copule théorique. Nous comparons le dépendogramme de la copule empirique à celui des quatre copules paramétriques sélectionnées. Le dépendogramme permet d'observer aussi le caractère plus ou moins simultané des réalisations issues de l'échantillon. Dans les queues plus précisément, il sera utile d'analyser si la simultanéité est forte et donc s'il est nécessaire de calibrer sur notre échantillon une copule avec une dépendance de queue.

Nous avons représenté sur la Figure III-30 les dépendogrammes associés à la copule empirique et aux copules paramétriques pour les résidus correspondants à la région 1 (région Sud) et à la région 2 (région Nord-Ouest) pour 20 000 simulations. Cela nous permettra de sélectionner la copule paramétrique qui semble correspondre le mieux à la copule empirique. Les dépendogrammes associés aux cinq autres couples sont disponibles en annexe E et donnent des résultats similaires.

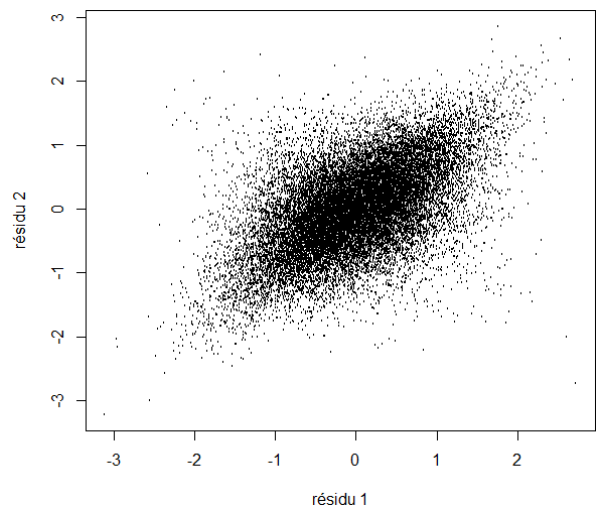
Copule empirique



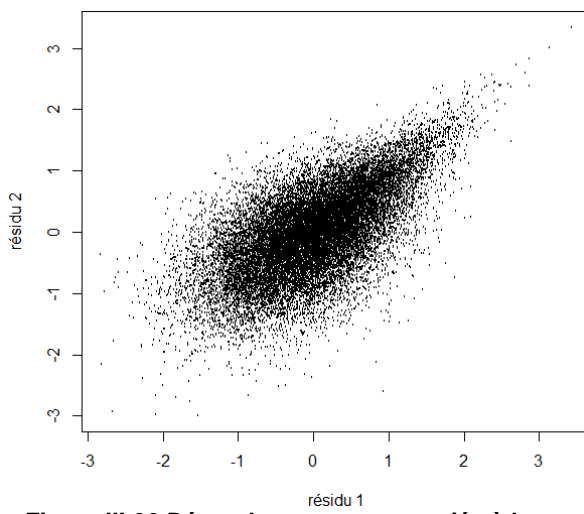
Copule Gaussienne



Copule de Student



Copule de Gumbel



Copule de Clayton

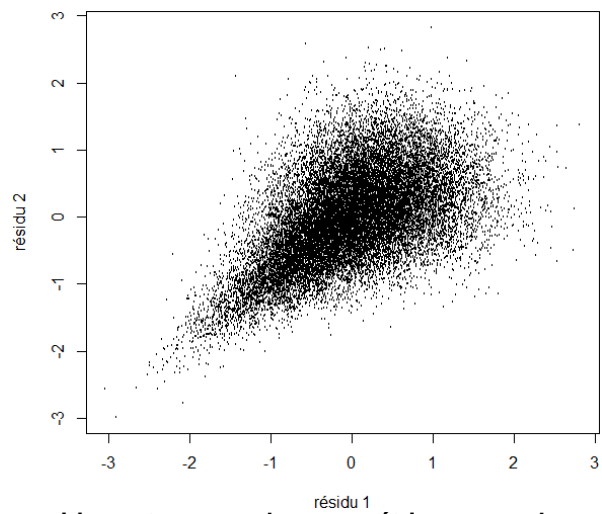


Figure III-30 Dépendogrammes associés à la copule empirique et aux copules paramétriques pour les résidus 1 et 2

Tout d'abord, il est clair que les copules de Gumbel et de Clayton ne sont pas satisfaisantes car elles présentent une asymétrie. Or, le nuage de points associé à la copule empirique semble symétrique. Les copules gaussiennes et de Student se rapprochent du nuage de points de la copule empirique et semblent convenir. Nous pouvons noter cependant que la copule de Student permet de mieux modéliser la dépendance des points extrêmes et semble donc mieux adaptée aux données empiriques.

Les critères graphiques nous incitent donc à sélectionner plutôt une copule de la famille elliptique et plus particulièrement la copule de Student. Cependant, les critères graphiques ne tiennent pas compte du nombre de paramètres à estimer pour définir le modèle. Les copules elliptiques nécessitent six et sept paramètres alors que les copules archimédiennes n'en nécessitent qu'un seul. Il n'est donc pas surprenant que les copules elliptiques parviennent à mieux reproduire la structure de dépendance grâce à leur surparamétrisation par rapport aux copules archimédiennes. Les critères quantitatifs permettront dans la suite de prendre en compte le nombre de paramètres dans la définition du modèle et de pénaliser les modèles nécessitant beaucoup de paramètres. Cependant, le caractère asymétrique associé aux copules archimédiennes nous incite à sélectionner une copule elliptique.

Critères quantitatifs

Pour vérifier et comparer l'adéquation de chacune des copules paramétriques aux données de résidus, nous choisissons plusieurs critères.

Tout d'abord, nous calculons et nous comparons la vraisemblance associée à chaque copule paramétrique. Puis, les critères BIC et AIC nous permettent de pénaliser le nombre de paramètres associés à chacun des modèles. Le Tableau III-19 regroupe les résultats obtenus pour chaque copule testée :

Copule	Log Vraisemblance	Nombre de paramètres	BIC	AIC
<i>Gaussienne</i>	35306	6	- 70 551	- 70 600
<i>Student</i>	36621	7	- 73 171	- 73 228
<i>Gumbel</i>	24919	1	- 49 827	- 49 836
<i>Clayton</i>	23047	1	- 46 083	- 46 092

Tableau III-19 Critères quantitatifs

De manière générale, le « meilleur » modèle sera celui qui aura un AIC et un BIC minimal. Il apparaît clairement sur le tableau que la copule de Student semble la mieux adaptée. Bien qu'elle ait le plus grand nombre de paramètres, la vraisemblance du modèle est suffisamment grande pour qu'elle ne soit pas pénalisée par son nombre élevé de paramètres.

Au regard des résultats des critères graphiques et quantitatifs, nous retenons la copule de Student pour la suite de l'étude.

3.5.2. Simulations obtenues avec la copule de Student

Nous vérifions à présent l'adéquation de la copule de Student pour caractériser la structure de dépendance des résidus.

Qualité d'adéquation Nous avons superposé les nuages de points associés aux données empiriques et aux données simulées avec la copule de Student pour vérifier l'adéquation du modèle sur la Figure III-31. Les points noirs représentent les données empiriques (23 177 points) et les points rouges représentent les données simulées avec la copule paramétrique de Student (50 000 points simulés).

Les résultats obtenus sont satisfaisants car les points théoriques et empiriques semblent bien coïncider. De plus, les valeurs extrêmes des points empiriques semblent ne pas dépasser les valeurs extrêmes des points simulés qui sont plus nombreux. Ils sont donc bien pris en compte dans les valeurs simulées. Les périodes de froid extrême pourront ainsi être captées avec cette modélisation.

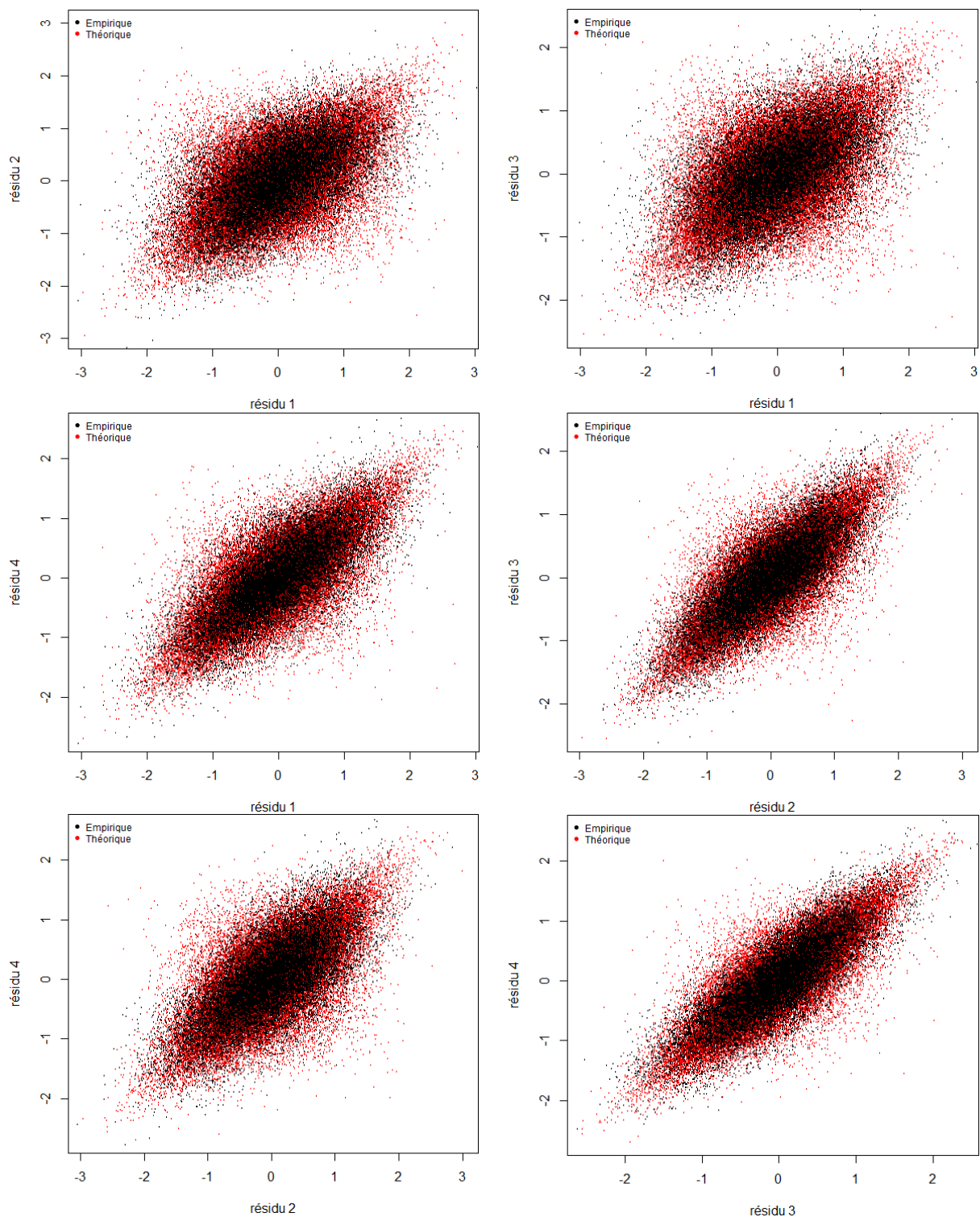


Figure III-31 Superposition des nuages de points théoriques et simulés avec la copule de Student

Matrice de corrélation simulée Après avoir simulé 50 000 résidus avec la copule de Student, nous souhaitons comparer la matrice de corrélation de Spearman obtenue avec les données simulées (Tableau III-20) et les données empiriques. Nous remarquons, en comparant à la matrice de corrélation empirique (Tableau III-16), que la matrice de corrélation associée à la copule paramétrique est légèrement sous-estimée par rapport à la matrice de corrélation empirique. Cependant, les deux matrices sont proches, l'erreur entre la matrice théorique et empirique étant de 0,4%. Les résultats sont donc acceptables.

Rho de Spearman	Région 1	Région 2	Région 3	Région 4
Région 1	100%	51%	50%	70%
Région 2	51%	100%	75%	64%
Région 3	50%	75%	100%	80%
Région 4	70%	64%	80%	100%

Tableau III-20 Matrice de corrélation de Spearman simulée avec la copule de Student

Simulation de la série résiduelle sur vingt-cinq jours La Figure III-32 représente la série résiduelle simulée avec la copule paramétrique de Student sur vingt-cinq jours. Nous pouvons observer aisément que les quatre séries sont corrélées. De plus, la corrélation observée semble cohérente avec celle observée sur les données empiriques (Figure III-26) et conforte notre choix.

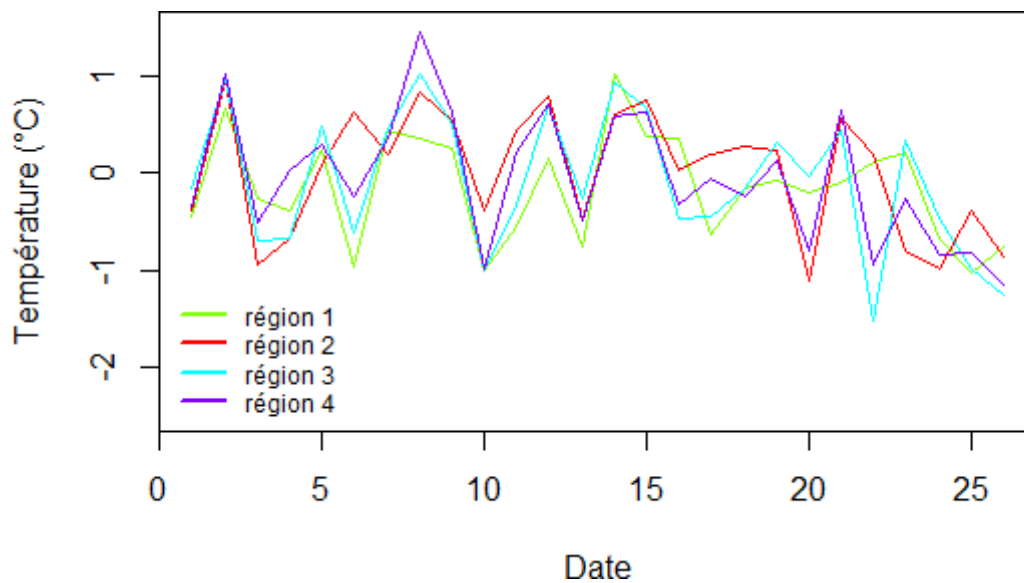


Figure III-32 Résidus simulés avec la copule de Student pendant 25 jours

A présent, nous avons à notre disposition les outils suffisants pour construire des scénarios de température pour l'année qui suit. Ce sera donc l'objectif de la prochaine section.

4. Simulations de scénarios

L'objectif de cette section est de regrouper l'ensemble des résultats précédents pour construire des scénarios de gel sur une année en France. Pour cela, nous simulons dans un premier temps plusieurs scénarios de température sur un an pour pouvoir en déduire ensuite la survenance de périodes de gel.

4.1. Scénarios de température

Tout d'abord, nous nous servons de la modélisation des températures dans chaque région définie au paragraphe III.2 pour prédire différents scénarios de température. Par ailleurs, nous utilisons la théorie des copules détaillée au paragraphe III.3 pour simuler conjointement les résidus associés à chaque région. Nous simulons alors 10 000 scénarios de température pour l'année suivante. Nos données de température s'arrêtant au 1^{er} Juillet 2013, nous construisons 10 000 scénarios de température du 2 Juillet 2013 au 2 Juillet 2014 pour chacune des quatre régions. Notons que les températures simulées correspondent à la moyenne des températures minimales au sein de chaque région.

La Figure III-33 représente les températures prédites dans chaque région pour le premier scénario. La courbe noire représente les températures observées disponibles depuis Août 2011 et la courbe rouge un scénario de températures pour l'année suivante.

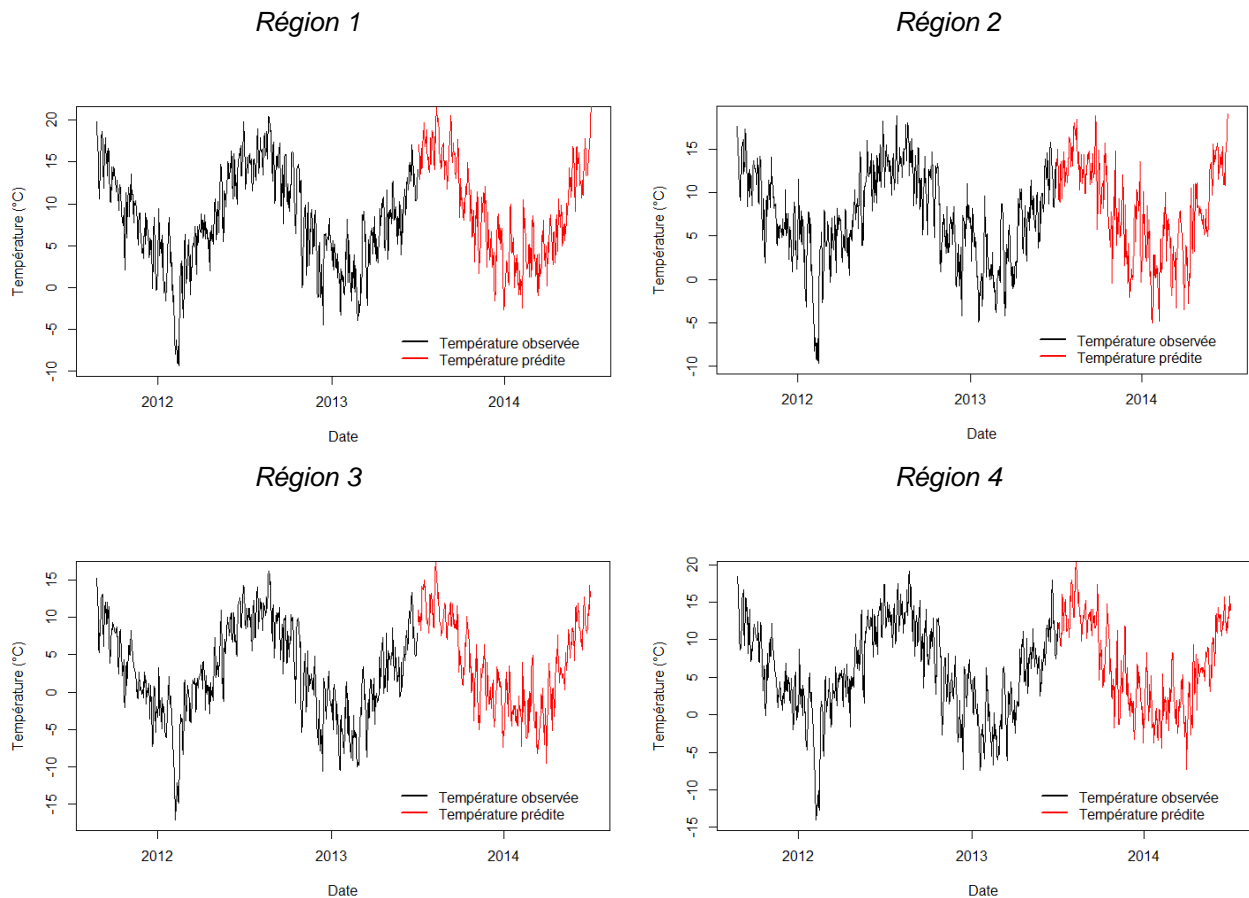


Figure III-33 Exemple d'un scénario de température par région de Juillet 2013 à Juillet 2014

La Figure III-34 représente les températures simulées dans les quatre régions. Ce graphique fait bien apparaître une corrélation entre les régions qui a été obtenue grâce à l'utilisation de copules.

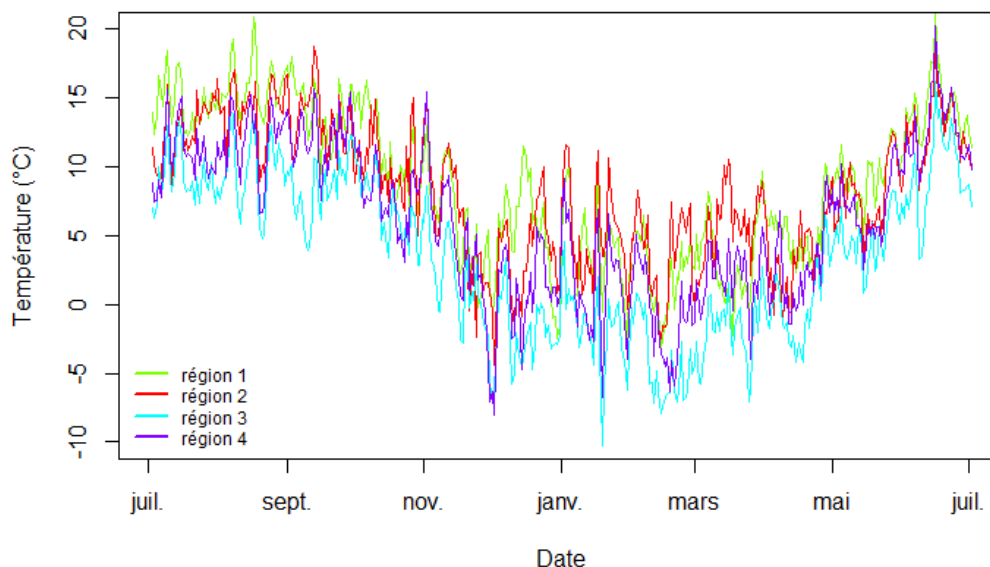


Figure III-34 Exemple d'un scénario de température dans les quatre régions

Les températures minimale et maximale sur l'ensemble des scénarios sont respectivement de -30°C et de 34°C , ce qui semble à première vue être des valeurs acceptables au regard de l'historique. Ces deux valeurs sont plus extrêmes que les valeurs historiques (respectivement à -19°C et à 23°C), ce qui semble cohérent puisque 10 000 scénarios ont été simulés et que nous disposions d'un historique de 63 années. Il convient de noter que la température minimale observée en France depuis 1947 était de -32°C .

4.2. Scénarios d'évènements

Nous nous servons des températures simulées ainsi que des modèles retenus pour prédire la fréquence de l'évènement gel (section III.1.2.5) afin de générer 10 000 scénarios de sinistres journaliers pour chaque région pour la période annuelle définie précédemment. Une fois les sinistres simulés, les fréquences s'obtiennent en divisant le nombre de sinistres simulé par l'exposition régionale correspondante.

Pour obtenir les sinistres prédits, on dispose déjà de la température moyenne minimale dans chaque région. A partir de cette information, nous déduisons les écarts de température (minimum et maximum) sur 20 jours. Nous avons également fait l'hypothèse que l'exposition par région est constante d'une année à l'autre. On dispose de l'exposition par région sur l'année 2012. Les informations par région sont donc toutes connues.

Ensuite, il s'agit de prédire les nombres de sinistres dans chaque région en tenant compte du seuil de $-5,5^{\circ}\text{C}$, de la température le jour du sinistre et de la saison. En effet, la prédiction doit être effectuée pendant l'hiver uniquement et lorsque la température est inférieure à 3°C . Par ailleurs, lorsque cette température est strictement supérieure à $-5,5^{\circ}\text{C}$, les prédictions de sinistres utilisent la régression *hurdle* présentée dans la sous-section III.1.2.5. Sinon, la prédiction utilise la régression binomiale négative présentée dans cette même sous-section III.1.2.5. Dans le reste des cas, en été ou lorsque la température est supérieure à 3°C , on prédit un nombre nul de sinistres. Une fois que les nombres de sinistres sont obtenus sur les 10 000 scénarios pour chaque région et sur les 365 jours, les fréquences se déduisent aisément.

Finalement, nous obtenons 10 000 scénarios donnant la probabilité de sinistralité journalière (ou probabilité de survenance de l'évènement gel pour une police donnée) au sein de chaque région pendant une période d'un an.

IV. Construction du module Vulnérabilité

1. Modéliser la perte assurantielle

Le module Aléa construit précédemment a permis de générer 10 000 scénarios journaliers donnant la probabilité de survenance de l'évènement gel dans chacune des quatre régions. Nous voulons à présent connaître la perte associée aux différents biens assurés lorsqu'un évènement gel survient.

Le module Vulnérabilité doit permettre de quantifier le sinistre produit par un évènement gel sur un bien donné. Pour cela, ce module fournit des courbes de destruction (ou courbes de vulnérabilité) qui modélise la fonction de répartition du taux de destruction¹⁶ observé suite à un évènement gel. Ces courbes diffèrent des courbes présentées dans la partie I qui associaient un taux de destruction différent selon l'intensité de l'évènement considéré. Pour le risque gel, nous considérons que les montants de sinistres causés par le gel ne dépendent pas de l'intensité de l'évènement. En effet, quand il survient, un évènement gel cause généralement des sinistres très similaires, peu importe leur seuil de déclenchement. Nous faisons donc l'approximation ici que le montant de perte lors de la survenance d'un sinistre gel dépend uniquement du type et de la valeur du bien assurée et non de l'intensité de l'évènement gel qui est prise en compte dans la probabilité de sinistralité dans le module Aléa.

Par leur nature (habitation, immeuble, locaux professionnels), les biens assurés par AXA ne présentent pas les mêmes pertes lorsqu'un évènement gel survient. Pour cette raison, nous calibrerons des courbes de vulnérabilité adaptées aux spécificités propres à chaque type de biens.

Finalement, le module Vulnérabilité permettra d'appliquer un évènement gel aux spécificités du portefeuille de biens, et de traduire une probabilité d'évènement en un montant de perte pour l'assureur.

2. La méthode MBBEFD

Pour modéliser la vulnérabilité des biens au gel, nous allons calibrer des courbes de la famille paramétrique MBBEFD (Maxwell-Boltzman, Bose-Einstein, Fermi-Dirac). Cette méthode a été popularisée par un article de S. BERNEGGER¹⁷ et est devenue un standard dans le secteur de l'assurance pour plusieurs raisons :

- Les courbes MBBEFD permettent de modéliser la probabilité de destruction d'un bien en prenant en compte des discontinuités. Par exemple, certains biens peuvent subir des destructions soit mineures soit totales et la fonction de répartition du taux de destruction présente donc des discontinuités.
- Elles peuvent prendre des formes très diverses : convexes, concaves, points d'inflexion, ...

Deux méthodes permettent de définir une courbe MBBEFD. Tout d'abord, il y a la méthode générale qui nécessite deux paramètres. Il existe aussi un cas particulier de la méthode générale,

¹⁶ Montant du sinistre divisé par la valeur du bien assuré

¹⁷ "Swiss RE exposure curves and the MBBEFD distribution class", 1997

appelée Hyperbolic MBBEFD, qui nécessite un seul paramètre. Dans la pratique, la seconde méthode est souvent préférée pour sa simplicité.

2.1. MBBEFD à deux paramètres

Nous décrivons dans un premier temps la méthode générale qui nécessite de calibrer deux paramètres a et b . L'avantage de cette méthode est la possibilité de modéliser de nombreuses formes de la fonction de répartition du taux de destruction : elle peut être concave, convexe ou convexe puis concave avec un point d'inflexion, en fonction des valeurs des paramètres a et b .

Soit F la fonction de répartition associée au taux de destruction (ou courbe de destruction). Elle est définie sur l'intervalle $[0,1]$, elle est continue sur $[0,1[$ et elle a une masse au point 1, c'est-à-dire $\mathbb{P}(DR = 1) = p > 0$ avec DR la variable aléatoire associée au taux de destruction.

On définit alors la courbe d'exposition G par :

$$G(x) = \frac{\mathbb{E}[DR|DR < x]}{\mathbb{E}[DR]} = \frac{\int_0^x (1 - F(y)) dy}{\int_0^1 (1 - F(y)) dy}$$

Avec $x =$ taux de destruction $= \frac{\text{Montant de sinistre}}{\text{Somme assurée}}$

La fonction F s'exprime donc aisément en fonction de G :

$$F(x) = \begin{cases} 1 & \text{si } x = 1 \\ 1 - \frac{G'(x)}{G'(0)} & \text{si } 0 \leq x < 1 \end{cases}$$

La famille MBBEFD, définie par S. BERNEGGER, impose une expression particulière à la fonction G en fonction des deux paramètres a et b :

$$G(x) = \frac{\ln(a + b^x) - \ln(a + 1)}{\ln(a + b) - \ln(a + 1)}$$

Nous pouvons alors réécrire la fonction F de la façon suivante :

$$F(x) = \begin{cases} 1 & \text{si } x \geq 1 \\ 1 - \frac{(a + 1) \cdot b^x}{a + b^x} & \text{si } 0 < x < 1 \end{cases}$$

Nous avons donc F croissante, $F(x) = 0$ si $x < 0$ et $F(x) = 1$ si $x \geq 1$. Cela définit bien une fonction de répartition. Par ailleurs, F est discontinue en $x = 1$. Notons p la probabilité associée à une destruction totale, c'est-à-dire :

$$p = \mathbb{P}(DR = 1) = 1 - F(1) = \frac{(a + 1)b}{a + b}$$

La fonction densité associée à F est alors définie par :

$$f(x) = \frac{-a(a + 1)b^x \ln(b)}{(a + b^x)^2} \cdot \mathbf{1}_{0 \leq x < 1} + p \cdot \mathbf{1}_{x=1}$$

Sa moyenne est donc égale à :

$$\mathbb{E}(\text{DR}) = (a + 1) \frac{\ln\left(\frac{a+b}{a+1}\right)}{\ln(b)}$$

2.2. MBBEFD à un paramètre (Hyperbolic MBBEFD)

La méthode MBBEFD à un paramètre permet de calibrer plus facilement les courbes de destruction d'une part et d'obtenir des résultats plus robustes d'autre part. Nous avons donc choisi d'adopter cette méthode dans l'étude.

La méthode MBBEFD à un paramètre est un cas particulier de la méthode à deux paramètres en prenant les paramètres a et b tels que :

$$\begin{cases} a = (1 - p) \left(\frac{p}{p - b} - 1 \right) \\ b \rightarrow 1 \end{cases}$$

Soit m le taux de destruction médian. La fonction de répartition Hyperbolic MBBEFD(m), dans ce cas particulier, s'exprime simplement en fonction de ce paramètre :

$$F(x) = P(\text{DR} < x) = \left[1 - \frac{1}{1 + \frac{x}{m}} \right] \cdot \mathbf{1}_{0 < x < 1} + \mathbf{1}_{x \geq 1}$$

Nous avons donc : $F(x) = 0$ si $x < 0$ et $F(x) = 1$ si $x \geq 1$ ce qui définit bien une fonction de répartition.

Puisque F est discontinue en $x = 1$, la probabilité de destruction totale p est donnée par :

$$p = P(\text{DR} = 1) = 1 - F(1) = \frac{m}{1 + m}$$

La fonction de densité associée à F est alors définie par :

$$f(x) = \frac{1}{m \cdot \left(1 + \frac{x}{m}\right)^2} \cdot \mathbf{1}_{0 \leq x < 1} + p \cdot \mathbf{1}_{x=1}$$

Sa moyenne est donc égale à :

$$\mathbb{E}(\text{DR}) = m \cdot \ln\left(1 + \frac{1}{m}\right)$$

Enfin, son quantile d'ordre α est défini par :

$$F^{-1}(\alpha) = m \frac{\alpha}{1 - \alpha} \mathbf{1}_{0 \leq \alpha \leq 1-p} + \mathbf{1}_{\alpha > 1-p}$$

Ainsi, avec la connaissance du taux de destruction médian m , nous pouvons construire la courbe de destruction F correspondante.

3. Taux de destruction médian

Le paragraphe précédent a permis de montrer qu'une courbe de destruction de type MBBEFD hyperbolique est entièrement définie par le taux de destruction médian que nous allons donc modéliser ici.

Comme introduit, nous supposons que le montant des pertes liées à un sinistre gel ne dépend que du type et de la valeur du bien assurée. Pour le vérifier, nous nous baserons sur l'historique des pertes gel du portefeuille France d'AXA entre 2008 et 2012 et nous montrerons qu'il existe un modèle linéaire entre les sommes assurées et les pertes historiques.

Pour débiter, il est nécessaire de partitionner le portefeuille en branches d'activité (*Line of Business* en anglais ou LoB) afin de distinguer les types de biens couverts :

- Habitation ;
- Agricole ;
- Immeuble ;
- Industrie.

Pour chaque LoB, nous appliquons la démarche suivante :

1. Regrouper les pertes en couches de sommes assurées croissantes.
2. Calculer pour chacune des pertes le taux de destruction qu'elle représente, comparativement à la valeur assurée du bien.
3. Calculer pour chaque couche k la somme assurée¹⁸ médiane SI et le taux de destruction médian m : $(SI(k), m(k))$.
4. Tester l'existence d'un lien log-linéaire linéaire sur l'ensemble des couples $(\log(SI(k)), \log(m(k)))$.

La Figure IV-1 illustre la régression linéaire effectuée entre $\log(SI)$ et $\log(m)$ pour les quatre branches d'activité considérées. Lorsque cela est possible, nous regroupons les branches d'activité qui ont les mêmes caractéristiques de régression. Dans notre cas, nous obtenons deux regroupements : les assurances de particuliers (LoB habitations) d'une part et les assurances professionnelles d'autre part qui regroupent les LoB agricole, immeuble et industrie.

¹⁸ La somme assurée correspond à la valeur d'un bien assuré

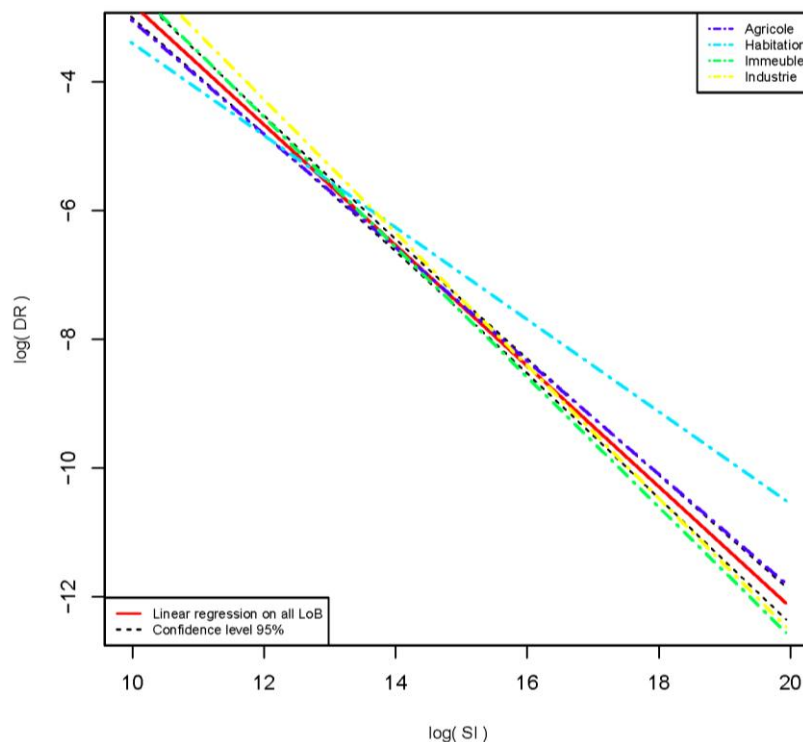


Figure IV-1 Régression Log-linéaire sur les points médians

Nous obtenons alors deux modèles de régression (un modèle pour la LoB habitation et un modèle pour les autres LoB) :

$$\log(m) = a \times \log(SI) + b$$

$$m = \exp(a \times \log(SI) + b)$$

Avec (a, b) les paramètres associés à la régression donnés dans le Tableau IV-1 :

	a	b
<i>Modèle 1 (LoB Habitation)</i>	-0,715	3,75
<i>Modèle 2 (LoB Professionnelles) Agricole – Immeuble - Industrie</i>	-0,947	6,68

Tableau IV-1 Paramètres de la régression

Nous illustrons sur la Figure IV-2 les quatre régressions linéaires associées aux quatre branches d'activité en observant notamment l'écart avec le regroupement des trois branches d'activité professionnelles (représenté par la droite rouge). Le modèle log-linéaire semble ainsi bien adapté à nos données avec des coefficients $R^{2^{19}}$ satisfaisants (égaux à 86% pour la LoB Habitation et à 99% pour les autres LoB).

Finalement, grâce à ces deux modèles, nous pouvons pour une branche d'activité et une somme assurée médiane données, associer un taux de destruction médian.

¹⁹ Dans le cas d'une régression linéaire simple, le coefficient de détermination (R^2) est un indicateur de la qualité d'adéquation du modèle aux données. Ce coefficient représente la part de la variance expliquée par le modèle dans la variance totale. Il est compris entre 0 et 1. Plus le R^2 est élevé, plus le modèle est bien ajusté aux données.

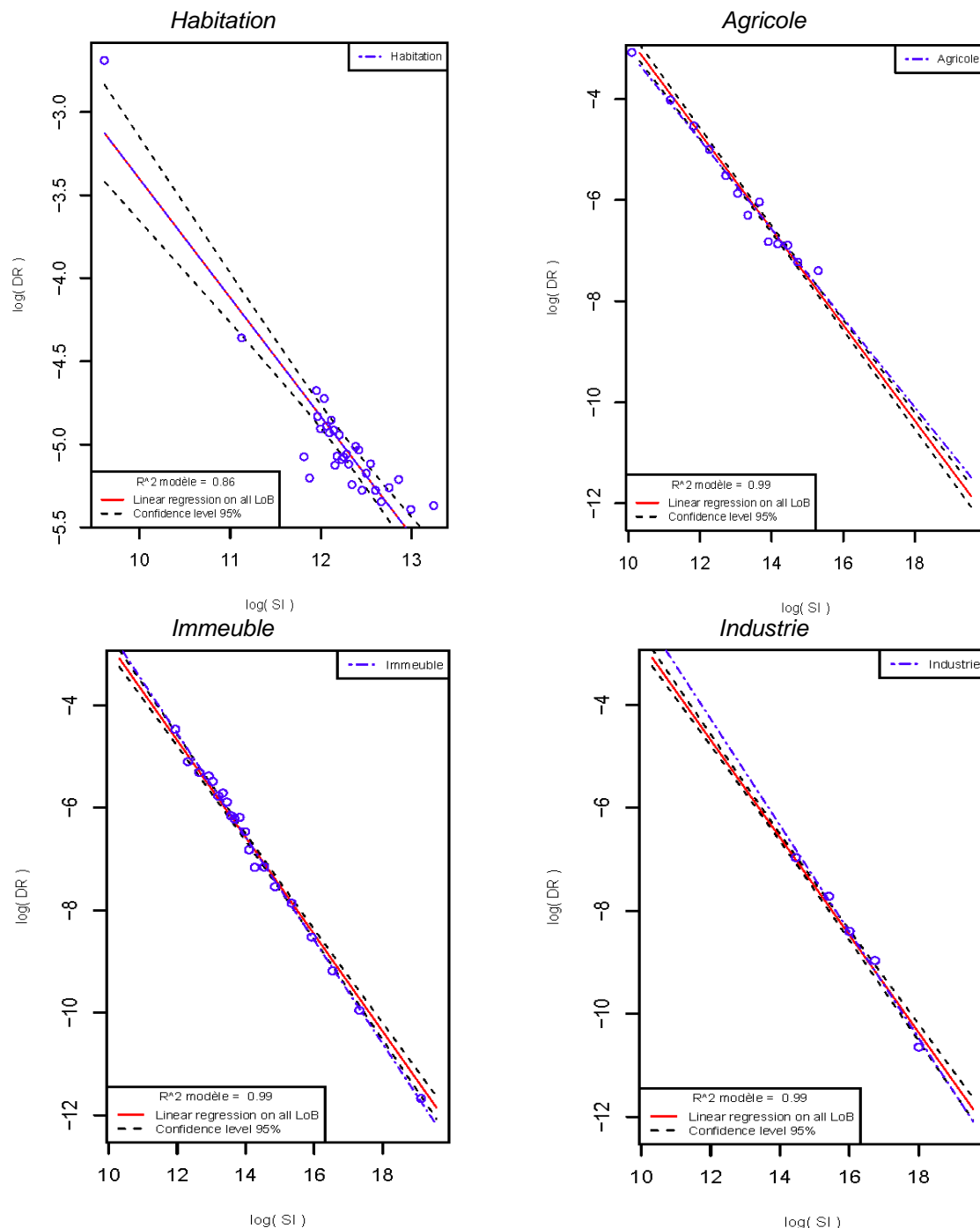


Figure IV-2 Régressions log-linéaires sur chaque ligne de business

4. Construction des courbes de destruction

Grâce au modèle de régression proposé, nous pouvons à présent construire la courbe de destruction d'un bien en fonction de sa somme assurée et de sa branche d'activité. A titre d'exemple, nous traçons les courbes de destruction F associées à cinq valeurs de somme assurée pour chaque branche d'activité. La Figure IV-3 donne les résultats obtenus pour la branche d'activité Habitation. Les courbes associées aux autres branches d'activité sont fournies en annexe F.

Pour la branche d'activité Habitation, nous avons sélectionné cinq valeurs de sommes assurées différentes données dans la légende sur la Figure IV-3. Ainsi, grâce au modèle de régression défini

dans le paragraphe précédent, nous trouvons cinq valeurs différentes pour le paramètre m en fonction des SI (Tableau IV-2), ce qui donne donc cinq courbes de destruction associées à chaque tranche.

Tranches	Somme assurée (SI)	Paramètre m (en %)
Tranche 1	15 000	4,4
Tranche 2	37 000	2,3
Tranche 3	92 000	1,2
Tranche 4	228 000	0,6
Tranche 5	566 000	0,3

Tableau IV-2 Paramètre de l'Hyperbolic MBBEFD

Par ailleurs, nous pouvons voir sur la Figure IV-3 que plus la somme assurée (SI) augmente, plus la courbe de destruction associée se rapproche d'un Dirac en zéro. En effet, plus le bien est cher, moins il sera impacté par un évènement gel proportionnellement à sa valeur totale.

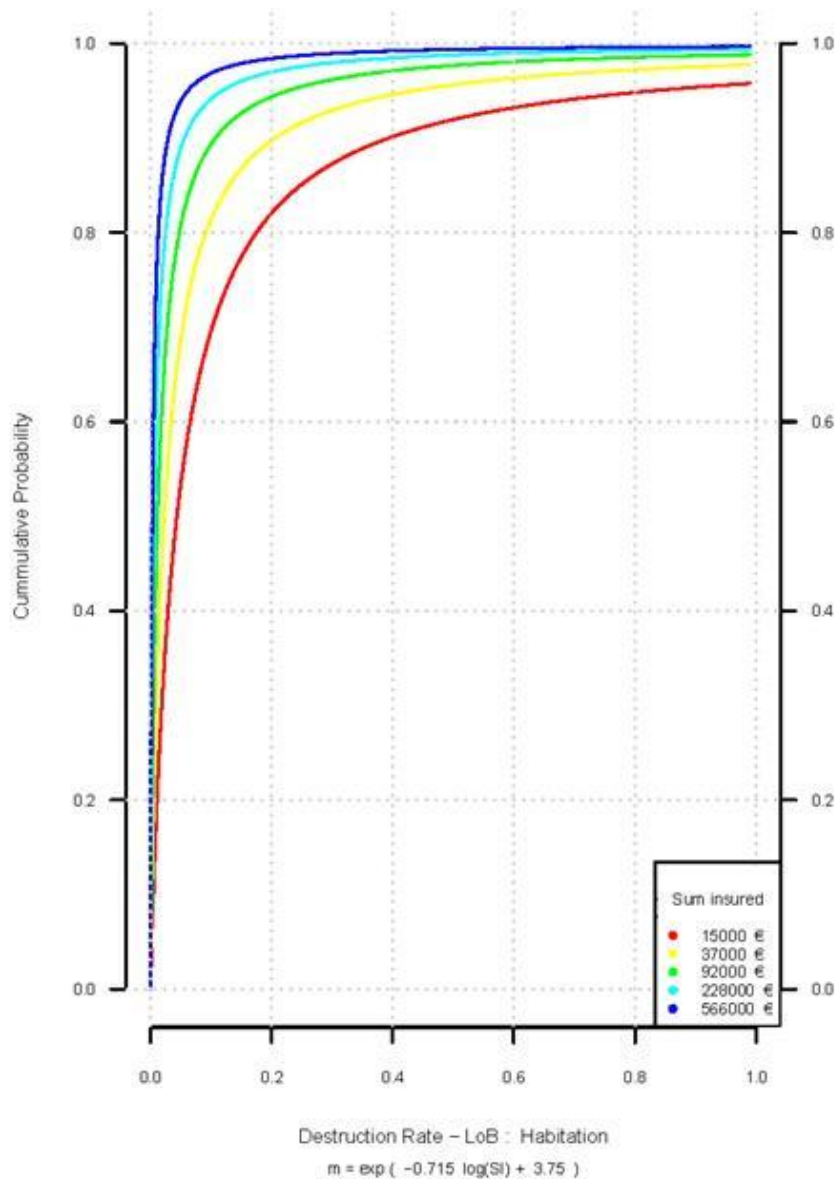


Figure IV-3 Courbes de destruction pour la LoB Habitation

V. Résultats du modèle

Les principaux résultats d'un modèle catastrophe sont synthétisés par les courbes AEP²⁰ et OEP²¹. Comme nous l'avons vu dans la première partie, ces courbes permettent de définir respectivement le capital réglementaire requis dans le cadre de Solvabilité II et la couverture de réassurance optimale à souscrire.

Les courbes AEP et OEP de notre modèle sont obtenues en couplant les modules Aléa, Vulnérabilité et Financier. Pour une question de confidentialité, le module Financier permettant de prendre en compte les caractéristiques des contrats souscrits et les conditions de réassurance n'a pas été modélisé. Nous considérons donc ici la perte brute avant l'application des conditions financières.

1. Obtention des pertes journalières par police et par scénario

Les événements gel, générés de façon quotidienne pour chaque région avec le module Aléa, doivent être appliqués au portefeuille de polices pour traduire la probabilité d'occurrence du risque en une perte financière. A cette fin, les polices du portefeuille d'AXA France ont été regroupées selon leur appartenance géographique à l'une des 4 régions de notre segmentation. Pour chaque jour d'un scénario régional annuel, nous commençons par évaluer la perte causée par le gel sur les polices du portefeuille régional.

Chaque police p possède les caractéristiques associées suivantes :

- la somme assurée représentant la valeur totale du bien assuré : p_{SI} ;
- la branche d'activité de la police représentant le type de bien assuré : p_{LOB} ;
- la région de la classification dans laquelle elle se trouve : $p_{région}$.

Le taux de destruction DR d'un bien touché par un événement gel a été modélisé dans la partie précédente. Au sein de chaque famille de biens assurés, nous avons montré qu'il existe un lien log-linéaire entre la somme assurée et le taux de destruction historique constaté. Nous avons ainsi calibré une courbe MBBEFD hyperbolique adaptée à chaque police du portefeuille. Pour une police p donnée, le taux de destruction DR ne dépend donc que de p_{SI} et p_{LOB} (paramètres nécessaires à la calibration de la courbe MBBEFD).

Au sein d'une région donnée, considérons une police de somme assurée SI et de taux de destruction associé DR. Notons \mathbb{P} la probabilité de survenance de l'évènement gel pour un jour donné d'un scénario. L'espérance de perte associée à cette police le jour considéré est :

$$\text{Perte}(SI, DR, \mathbb{P}) = SI \times DR \times \mathbb{P}$$

La probabilité \mathbb{P} de l'évènement gel a été générée dans le module Aléa pour 10 000 scénarios de 365 jours pour chaque région de la classification. Pour chaque région, la probabilité associée à la police p du scénario i et du jour j dans l'année simulée est notée $\mathbb{P}(i, j, p_{région})$.

²⁰ *Aggregate Exceedance Probability*

²¹ *Occurrence Exceedance Probability*

Nous pouvons écrire l'espérance des pertes par police, par jour et par scénario de la façon suivante :

$$\text{Perte}_{i,j,p} = p_{SI} \times DR(p_{SI}, p_{LoB}) \times \mathbb{P}(i, j, p_{\text{region}})$$

Où i est le numéro du scénario, j le jour au sein de l'année simulée et p la police considérée.

La perte du portefeuille AXA France pour un scénario et un jour donnés est donc :

$$\sum_{p \text{ police}} p_{SI} \times DR(p_{SI}, p_{LoB}) \times \mathbb{P}(i, j, p_{\text{region}})$$

2. Construction des courbes AEP et OEP

2.1. Agrégation des pertes journalières pour la réassurance

La courbe OEP représente la distribution de la perte maximale causée par un seul évènement pendant une année et la courbe AEP représente la distribution de la perte causée par l'ensemble des évènements survenus pendant une année. Pour construire ces courbes, il faut donc définir ce qu'est un évènement gel.

Dans le cadre du risque de gel, les dispositions contractuelles de réassurance sont établies sur des périodes de 21 jours consécutifs. Ces périodes sont déterminées de façon à maximiser le montant des pertes qu'elles regroupent, pertes qui peuvent ensuite être transférées par le biais de traités de réassurance. Elles définissent donc un évènement gel pour la construction des courbes AEP et OEP.

2.2. La courbe OEP

La courbe OEP présente les périodes de retour associées à l'évènement de perte maximale sur 21 jours associé à chaque année. Nous disposons ainsi de 10 000 pertes, engendrées par la période de 21 jours de chaque scénario ayant le plus sévèrement impacté le portefeuille. Après avoir ordonné ces pertes par ordre décroissant, nous pouvons tracer la courbe OEP en associant à chaque perte, une période de retour. Soit N le nombre total de scénarios, y le classement d'une perte (dans l'ordre décroissant), alors la période de retour t associée à la perte considérée est :

$$t = \frac{N}{y}$$

Par exemple, au sein de nos 10 000 relevés de pertes annuelles maximales sur 21 jours, la 50^{ème} perte la plus sévère de notre simulation aura une période de retour de $10000 / 50 = 200$ ans.

2.3. La courbe AEP

La courbe AEP représente les périodes de retour associées à la distribution des pertes annuelles. Pour obtenir la perte annuelle d'un scénario, nous sommes les pertes de l'ensemble des évènements de 21 jours du scénario considéré. Nous disposons ainsi de 10 000 pertes annuelles issues des 10 000 scénarios que nous avons simulés.

Pour tracer la courbe AEP, nous procédons ensuite de la même façon que pour construire la courbe OEP.

La Figure V-1 représente les courbes AEP et OEP prises sur des périodes de retour de 1 à 250 ans. Pour des raisons de confidentialité, les montants de pertes ne sont pas communiqués dans ce mémoire. Cependant, la valeur x associée à l'axe des ordonnées représente la perte brute totale d'AXA causée par le gel pendant l'année 2012. Nous pouvons donc déterminer que la période de retour associée à cet évènement sur la courbe AEP est de huit ans. Or, dans la partie I.2 (Figure I-2), nous avons vu que l'année 2012 correspondait à la sixième plus grosse vague de froid survenue durant les soixante dernières années, et donc correspondait à un évènement de période de retour de 10 ans. Ainsi, une période de retour de huit ans estimée avec le modèle semble cohérente et prudente d'un point de vue risk management.

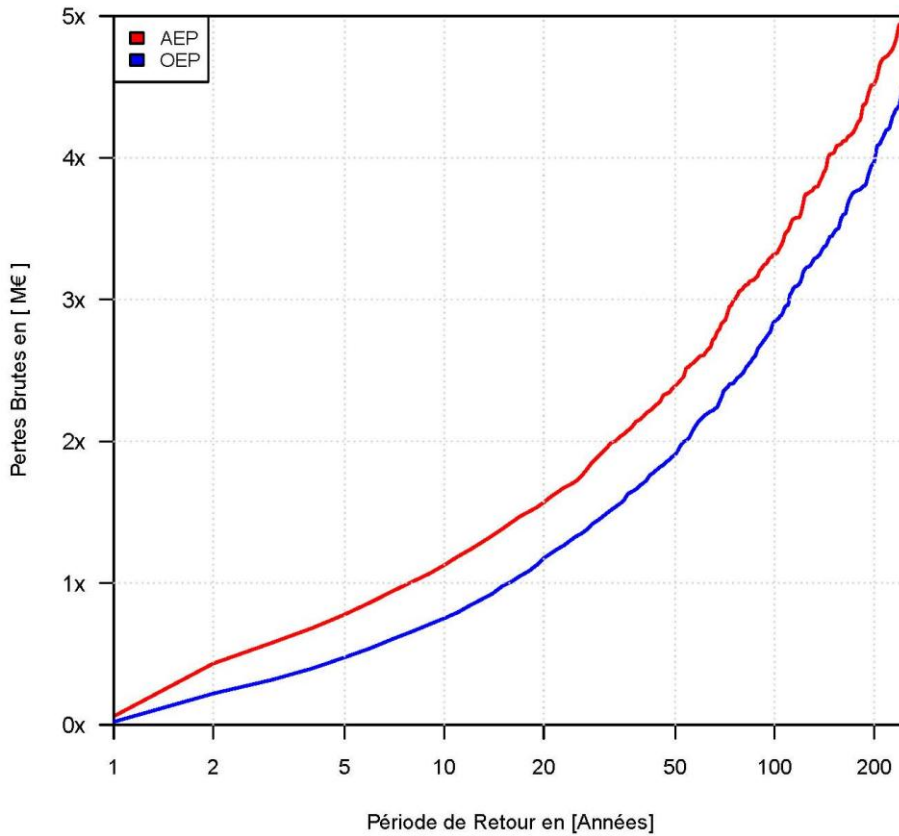


Figure V-1 Courbes AEP et OEP

Conclusion

L'objectif de ce mémoire était de construire un modèle catastrophe modélisant le risque de gel en assurance non vie. Un modèle catastrophe combine trois modules, chaque module apportant une information précise. Ce mémoire est focalisé sur la construction des deux premiers : le module Aléa et le module Vulnérabilité. Nous obtenons en résultat final une distribution des pertes brutes issue de l'application de la modélisation du risque gel sur le portefeuille exposé. Le module financier, dernière étape de la modélisation, est réalisé par l'équipe interne d'AXA pour une question de confidentialité des clauses. La prise en compte des conditions financières avec ce dernier module permettra d'affiner le résultat final via l'obtention d'une distribution des pertes nettes.

Le module Aléa est l'élément central d'un modèle catastrophe. Ce module, dédié à la caractérisation physique du phénomène naturel étudié, aboutit à la génération d'un catalogue d'évènements reproduisant le péril. Pour construire ce module, nous nous sommes appuyées sur un ensemble de méthodes statistiques afin de pouvoir simuler aléatoirement des évènements de gel. L'évènement gel a ainsi pu être caractérisé du point de vue de l'assureur en intégrant l'information sur les sinistres historiques d'une part, mais aussi des éléments sur le phénomène physique tels que les températures et la localisation géographique d'autre part. A partir de cette caractérisation du gel et des températures simulées, nous avons proposé un modèle de prédiction des probabilités journalières de survenance du gel en chaque région de France.

A la différence des modèles catastrophes classiques, où les évènements sont simulés ponctuellement et donc de façon non consécutive, l'approche adoptée dans ce mémoire permet d'apporter une information journalière sur la survenance de sinistres liés au gel. Cette information présente un atout majeur pour un assureur dans le cadre de l'optimisation de ses traités de réassurance et de sa gestion du risque de gel.

La démarche que nous avons adoptée a le mérite d'être souple et adaptable. Le modèle a été pensé de façon à pouvoir intégrer de nouveaux pays tout en tenant compte de leur corrélation. Actuellement, l'étude est étendue à la Suisse et sera poursuivie pour l'ensemble des pays de l'Union Européenne. En effet, les résultats sur la France sont crédibles car cohérents avec les pertes historiques d'AXA ; ce qui suscite un intérêt important pour l'application du modèle aux autres pays du bassin Europe.

Enfin, le modèle construit pourrait être adapté à un autre risque climatique important : le risque de sécheresse. En effet, le phénomène physique à l'origine de la sécheresse étant également la température, la construction d'un modèle catastrophe pour la sécheresse pourrait s'inspirer de la méthode décrite dans ce mémoire. Il faudrait alors considérer cette fois les températures maximales et tenir compte de la typologie des sols.

Table des figures

Figure I-1 Fréquence de sinistres en France entre 1998 et 2006 Source : FFSA	6
Figure I-2 Vagues de froid en France Source : Météo France.....	7
Figure I-3 Nombre de sinistres survenus entre 2008 et 2013 à cause du gel pour AXA.....	8
Figure I-4 Impact de la réassurance sur le ratio de sinistralité en France Source : FFSA.....	10
Figure I-5 Mécanismes de traités de réassurance	10
Figure I-6 Structure d'un modèle catastrophe	12
Figure I-7 Exemples de courbes de vulnérabilité	14
Figure I-8 Module Financier.....	16
Figure I-9 Exemple de courbes OEP et AEP	17
Figure I-10 Stations des relevés de température en France.....	19
Figure II-1 Eboulis des valeurs propres.....	24
Figure II-2 Points géographiques les mieux représentés sur les 4 premiers axes principaux.....	25
Figure II-3 Exemple d'un dendogramme	28
Figure II-4 Dendogramme associé à la Classification Ascendante Hiérarchique (CAH) sur les coordonnées des 4 premiers axes principaux de l'EOF.....	30
Figure II-5 Graphiques obtenus avec la CAH pour différents nombres de classes	32
Figure II-6 Segmentation de la France en 4 régions (résultats de la CAH à la suite de l'EOF).....	33
Figure II-7 Distance DTW.....	36
Figure II-8 Distance euclidienne	36
Figure II-9 Alignement avec la distance DTW (à droite) et la distance euclidienne (à gauche)	37
Figure II-10 Quatre régions de la France obtenues par la méthode CAH avec DTW (à gauche) et par la méthode EOF puis CAH (à droite).....	37
Figure III-1 Probabilité d'occurrence du gel par région (avec MinMaxT20)	44
Figure III-2 Illustration de la méthodologie POT	53
Figure III-3 Mean Excess function de tminMoy	54
Figure III-4 Températures moyennes journalières observées dans la région 1 depuis 1980	56
Figure III-5 Décomposition de la série.....	57

Figure III-6 Diagramme des autocorrélations de la série résiduelle.....	59
Figure III-7 Diagramme des autocorrélations partielles de la série résiduelle	59
Figure III-8 Dispersion des résidus par mois.....	60
Figure III-9 Modélisation de la tendance	62
Figure III-10 Modélisation de la saisonnalité	63
Figure III-11 Fonction de pondération triangulaire	64
Figure III-12 Choix du paramètre de lissage	65
Figure III-13 Ecart type mobile – h=50	65
Figure III-14 Diagramme des autocorrélations	66
Figure III-15 Diagramme des autocorrélations partielles	66
Figure III-16 Dispersion des résidus par mois.....	67
Figure III-17 QQ plot comparant la distribution de ϵt et la distribution d'une loi gaussienne.....	68
Figure III-18 Comparaison des tendances dans chaque région	69
Figure III-19 Comparaison de la saisonnalité.....	69
Figure III-20 Comparaison de l'écart-type mobile	69
Figure III-21 Diagramme des autocorrélations pour chaque région.....	70
Figure III-22 Diagramme des autocorrélations partielles pour chaque région	70
Figure III-23 QQ Plot permettant de vérifier le caractère gaussien des résidus de chaque région	71
Figure III-24 Dispersion des résidus par mois.....	71
Figure III-25 Nuages de points des résidus empiriques.....	72
Figure III-26 Résidus observés les 25 derniers jours.....	73
Figure III-27 Comparaison des copules gaussiennes et de Student de dimension 2	76
Figure III-28 Comparaison des copules de Gumbel et de Clayton de dimension 2.....	77
Figure III-29 Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 1 et 280	
Figure III-30 Dépendogrammes associés à la copule empirique et aux copules paramétriques pour les résidus 1 et 2	81
Figure III-31 Superposition des nuages de points théoriques et simulés avec la copule de Student...	83
Figure III-32 Résidus simulés avec la copule de Student pendant 25 jours	84

Figure III-33 Exemple d'un scénario de température par région de Juillet 2013 à Juillet 2014	85
Figure III-34 Exemple d'un scénario de température dans les quatre régions	86
Figure IV-1 Régression Log-linéaire sur les points médians.....	92
Figure IV-2 Régressions log-linéaires sur chaque ligne de business	93
Figure IV-3 Courbes de destruction pour la LoB Habitation.....	94
Figure V-1 Courbes AEP et OEP	97

Liste des tableaux

Tableau II-1 Variances intraclases pour différents nombres de classes	31
Tableau III-1 Résultats des régressions logistiques pour la probabilité d'occurrence du gel	43
Tableau III-2 Comparaison des modèles de régression logistique	44
Tableau III-3 Résultats de l'estimation avec la régression de poisson	47
Tableau III-4 Résultats du test de dispersion	48
Tableau III-5 Résultats de l'estimation avec la régression binomiale négative.....	48
Tableau III-6 Résultats de l'estimation avec la régression <i>zero inflated</i>	50
Tableau III-7 Résultats de l'estimation avec le modèle <i>hurdle</i>	51
Tableau III-8 Comparaison des différents modèles de comptage.....	51
Tableau III-9 Effectifs des nombres de sinistres nuls (observés et estimés)	52
Tableau III-10 Résultats du modèle <i>hurdle</i> ($t_{\min} \text{Moy} > -5,5^\circ$)	55
Tableau III-11 Résultats de la régression binomiale négative ($t_{\min} \text{Moy} \leq -5,5^\circ$)	55
Tableau III-12 Test de Ljung Box	58
Tableau III-13 Tests d'adéquation avec un modèle AR(3).....	60
Tableau III-14 Tests d'adéquation avec un modèle AR(3).....	68
Tableau III-15 Tests d'adéquation pour les régions 2, 3 et 4	71
Tableau III-16 Matrice de corrélation de Spearman empirique	73
Tableau III-17 Estimation des paramètres de corrélation de la copule gaussienne avec la méthode CML	78
Tableau III-18 Estimation des paramètres de corrélation de la copule de Student avec la méthode CML	78
Tableau III-19 Critères quantitatifs	82
Tableau III-20 Matrice de corrélation de Spearman simulée avec la copule de Student.....	84
Tableau IV-1 Paramètres de la régression.....	92
Tableau IV-2 Paramètre de l'Hyperbolic MBBEFD	94

Bibliographie

BERNEGGER S., 1997, *The Swiss RE exposure curves and the MBBEFD distribution class*

BJÖNRSSON H. & VENEGAS S.A., 1997, *A manual for EOF and SVD – Analyses of Climatic Data*, McGill University

CAMERON, A.C. & TRIVEDI, P.K., 1998, *Regression analysis of count data*, Cambridge, Cambridge University Press

CAMERON, A.C. & TRIVEDI, P.K., 2005, *Microeconometrics: Methods and Applications*, Cambridge, Cambridge University Press.

CASIEZ G., *Dynamic Time Warping : déformation temporelle dynamique*, Université de Lille

CHARPENTIER A., DUTANG C., *L'Actuariat avec R*

CHAVEZ-DEMOULIN V. & DAVISON A.C., 2012, *Modelling time series extremes*, Volume 10, Number 1, P 109–133

CREPON B. & JACQUEMET N., 2010, *Econométrie : méthodes et applications*, De Boeck

DENUIT M. & CHARPENTIER A., 2005, *Mathématiques de l'assurance non-vie*, Tome II, Economica

EUROPEAN CLIMATE ASSESSMENT & DATASET, <http://eca.knmi.nl/>

FERMANIAN, J.D., 2014, *Théorie des copules*, support de cours ENSAE

FFSA, 2012, *Rapport annuel*

FLYNN M. & FRANCIS L. A., 2009, *More flexible GLMs zero-inflated models and hybrid models*, Casualty Actuarial Society E-forum

GORGE G., 2013, *Insurance Risk Management and Reinsurance*

HANNACHI A., 2004, *A Primer for EOF Analysis of Climate Data*, Department of Meteorology, University of Reading U.K

HUGUES G., RAO S. & RAO T., 2006, *Statistical analysis and time-series models for minimum/maximum temperatures in the Antarctic Peninsula*

KHAROUBI-RAKOTOMALALA C. 2008, *"Les fonction copules en finance"*, Université de la Sorbonne

LEDOLTER J., 2013, *Datamining and business analytics with R*, Wiley

LOPEZ O., 2014, *Econométrie de l'assurance*, support de cours ENSAE

ROBERT Y. C., *Théorie des Valeurs Extrêmes*, support de cours ENSAE

RONCALLI T., 2002, *Gestion des risques multiples ou copules et aspects multidimensionnels du risque*, support de cours ENSAI

RONCALLI T., FRACHOT A., 2009, *La Gestion des Risques Financiers*, Economica

ROUSTANT O., 2003, *Produits dérivés climatiques : aspects économétriques et financiers*

SAKOE H., CHIBA S., 1978, *Dynamic programming algorithm optimization for spoken word recognition*

SAPORTA G., 2011, *Probabilités, Analyse de données et Statistique*, Technip

SKLAR A., 1959, *Fonctions de répartition à n dimensions et leurs marges*, Publications de l'Institut de Statistique de l'Université de Paris

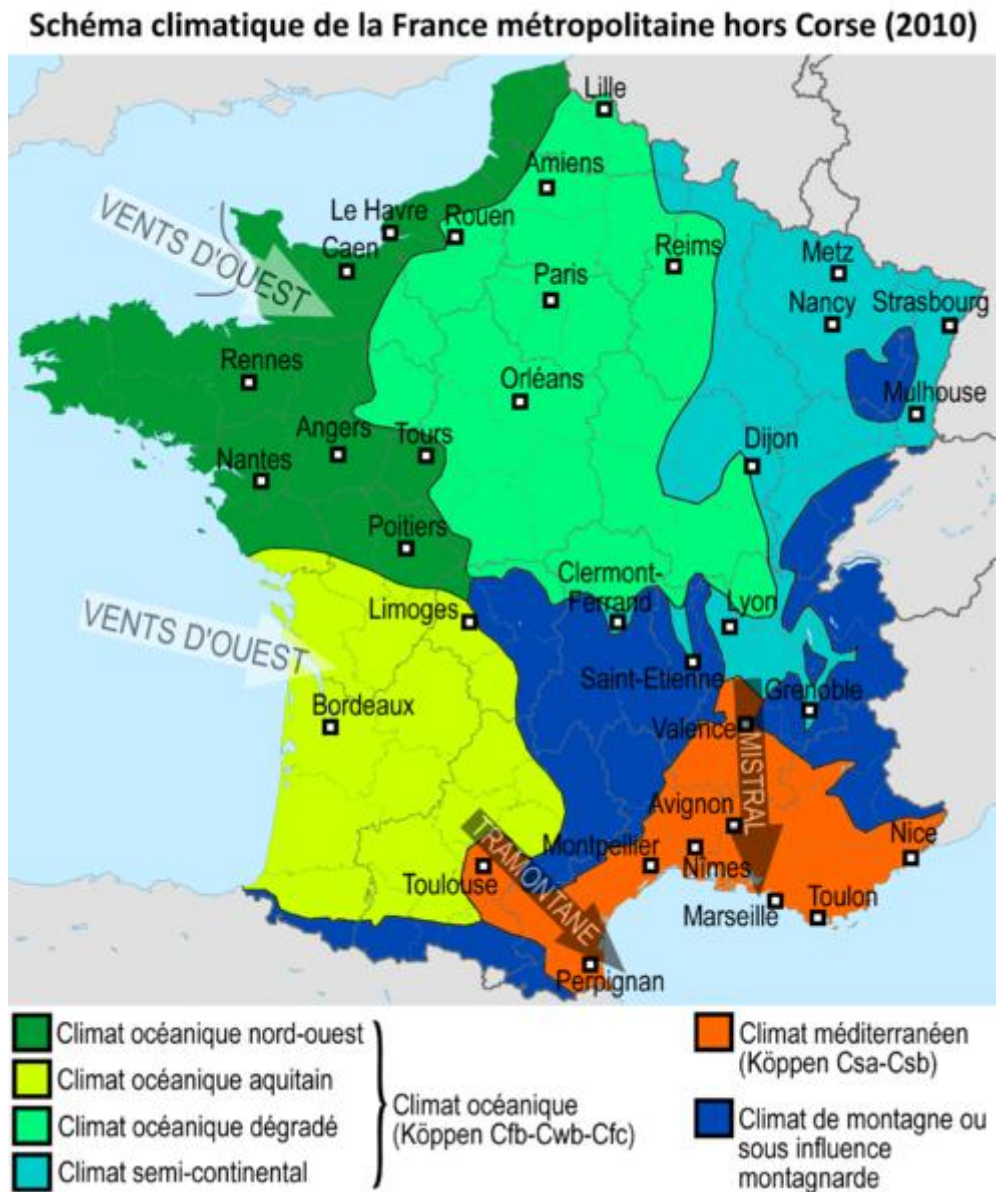
STRAUSS D., 2009, *Empirical Orthogonal Function Analysis (Principal Component Analysis)*, ICTP

SWISS RE, 2003, *Catastrophes naturelles et réassurance*

SWISS RE, 2003, *Introduction à la réassurance*

ZEILEIS A., KLEIBER C. & JACKMAN S., 2008, *Regression models for count data in R*, Journal of statistical software, Volume 27, Issue 8

Annexe A - Schéma climatique de la France métropolitaine



Cette carte est disponible via le lien ci-après :

http://fr.wikipedia.org/wiki/Climat_de_la_France

Annexe B - Fonctions d'autocorrélations et tests statistiques utilisés

1) Fonction d'autocorrélations (ACF) et fonction d'autocorrélations partielles (PACF)

La fonction d'autocorrélation d'un processus stationnaire $(X_t)_{t \in \mathbb{Z}}$, la fonction à valeurs dans \mathbb{Z} définie par :

$$\forall h \in \mathbb{Z}, \rho(h) = \frac{\text{cov}(X_t, X_{t+h})}{\text{Var}(X_t)} = \frac{\gamma(h)}{\gamma(0)} = \text{Corr}(X_t, X_{t+h})$$

La fonction d'autocorrélation partielle permet de définir le lien entre X_t et X_{t+h} , une fois que l'on exclut l'influence des variables entre X_t et X_{t+h} . La fonction d'autocorrélation partielle d'ordre h est définie par :

$$\begin{aligned} r(h) &= \text{Corr}(X_t - \text{EL}(X_t | X_{t-1}, \dots, X_{t-h+1}), X_{t-h} - \text{EL}(X_{t-h} | X_{t-1}, \dots, X_{t-h+1})) \\ &= \frac{\text{Cov}(X_t - \text{EL}(X_t | X_{t-1}, \dots, X_{t-h+1}), X_{t-h} - \text{EL}(X_{t-h} | X_{t-1}, \dots, X_{t-h+1}))}{\sqrt{\text{Var}(X_t - \text{EL}(X_t | X_{t-1}, \dots, X_{t-h+1})) * \text{Var}(X_{t-h} - \text{EL}(X_{t-h} | X_{t-1}, \dots, X_{t-h+1}))}} \end{aligned}$$

Avec EL l'opérateur pour l'espérance linéaire. $\text{EL}(X_t | X_{t-1}, \dots, X_{t-h+1})$ est la projection orthogonale de X_t sur le sous-espace engendré par $X_{t-1}, \dots, X_{t-h+1}$.

2) Test de Ljung Box

Le test de Ljung Box permet d'évaluer la corrélation entre les résidus et est très utilisé dans l'étude de séries temporelles pour vérifier que les résidus sont des bruits blancs.

Nous souhaitons tester :

H_0 : les résidus sont indépendamment distribués contre

H_a : les résidus ne sont pas indépendamment distribués

La statistique du test est définie par :

$$Q = n(n+2) \sum_{k=1}^h \frac{\widehat{\rho}_k}{n-k}$$

Avec :

n : la taille de l'échantillon

$\widehat{\rho}_k$: l'autocorrélation de l'échantillon au lag k

h : le nombre de lags testés

Alors, pour un niveau de confiance α , la zone de rejet de l'hypothèse nulle est :

$$Q > \chi_{1-\alpha, h}^2$$

Avec $\chi_{1-\alpha, h}^2$ le quantile α d'une distribution du chi-deux à h degrés de liberté.

3) Test de Kolmogorov Smirnov

Le test de Kolmogorov Smirnov est utilisé pour déterminer si un échantillon suit bien une loi donnée (la loi normale centrée réduite dans notre étude).

Soit X la variable étudiée et $(x_i)_{1 < i < n}$ les observations de X . Nous souhaitons tester :

H_0 : la distribution de X suit une loi normale centrée réduite contre

H_a : la distribution de X ne suit pas une loi normale centrée réduite

Soit $F_{réelle}$ la fonction de répartition empirique et $F_{théo}$ la fonction de répartition de la loi normale centrée réduite. On calcule :

$$K = \sup_i |F_{réelle}(x_i) - F_{théo}(x_i)|$$

Sous H_0 , K suivra une distribution selon la fonction de Kolmogorov Smirnov. Soit D_α la valeur critique de la loi Kolmogorov Smirnov pour un seuil α . Alors, la zone de rejet de l'hypothèse nulle est :

$$K > D_\alpha$$

4) Test de Student

Le test de Student permet de tester la significativité d'un paramètre et de déterminer s'il est significativement non nul.

Soit β le paramètre étudié et $\hat{\beta}$ son estimation. Nous souhaitons tester :

H_0 : $\beta = 0$ contre

H_a : $\beta \neq 0$

La statistique du test est définie par :

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\text{Var}(\hat{\beta})}}$$

T suit alors une loi de Student. Donc, pour un niveau de confiance α , la zone de rejet de l'hypothèse nulle est :

$$|T| > t_{\alpha, h}$$

Avec $t_{\alpha, h}$ le quantile d'ordre α d'une distribution de Student à h degrés de libertés et h dépendant du nombre de paramètre du modèle.

Annexe C - Résultats des tests de significativité

Le tableau ci-dessous permet de présenter l'ensemble des tests de significativité (avec le test de Student) effectués dans la partie III.2 :

	Valeur	Ecart-type	p-value
a	6,668e-05	5,286e-07	<2e-16
b	7,372	7,073e-03	<2e-16
a_1	-6,18	0,003525	<2e-16
b_1	-3,02	0,003525	<2e-16
a_2	0,019	0,003525	<2e-16
b_2	0,58	0,003525	<2e-16
α	2,67	0,002987	<2e-16
β	0,64	0,004230	<2e-16
γ	0,07	0,004219	<2e-16

Annexe D - Dépendance de queue

Soit (U_1, U_2) un couple de variables aléatoires. La dépendance de queue est alors définie par :

$\lambda_u = \lim_{u \rightarrow 1^-} \mathbb{P}[U_1 > u | U_2 > u]$: la dépendance de queue à droite

$\lambda_l = \lim_{u \rightarrow 0^+} \mathbb{P}[U_1 < u | U_2 < u]$: la dépendance de queue à gauche

La dépendance de queue d'une copule gaussienne bivariée de paramètre ρ est²² :

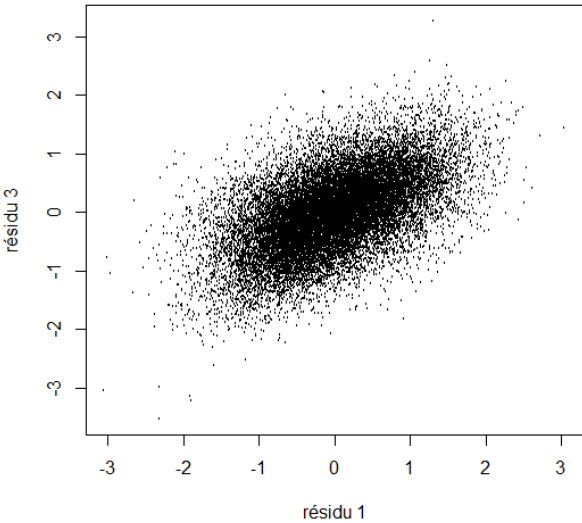
$$\begin{cases} \lambda_u = \lambda_l = 1 & \text{si } \rho = 1 \\ \lambda_u = \lambda_l = 0 & \text{si } \rho < 1 \end{cases}$$

La copule gaussienne ne permet donc pas de corrélérer des dépendances extrêmes.

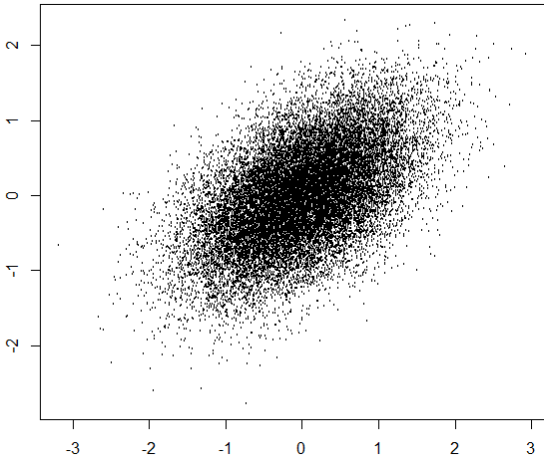
²² Pour plus de détails, voir RONCALLI T. (2002).

Annexe E - Dépendogrammes et fonctions de Kendall

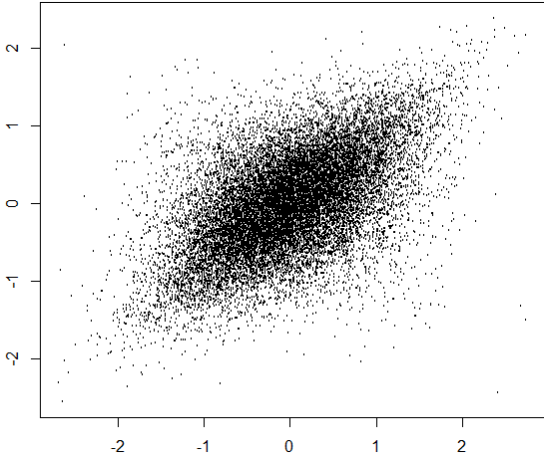
Copule empirique



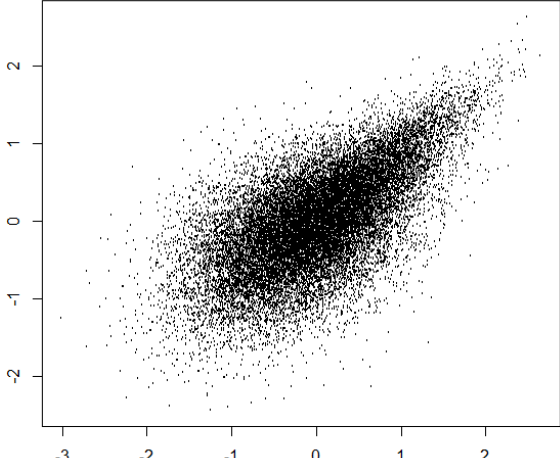
Copule Gaussienne



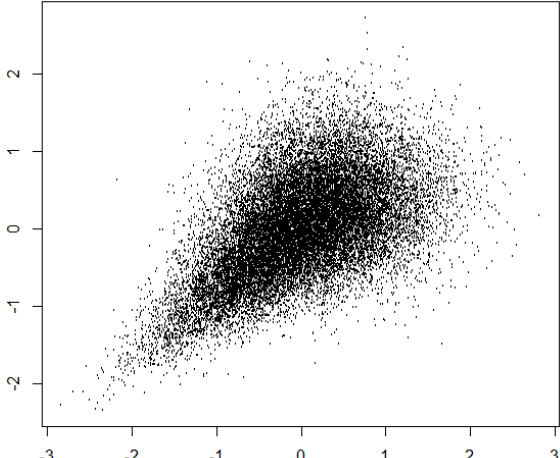
Copule de Student



Copule de Gumbel

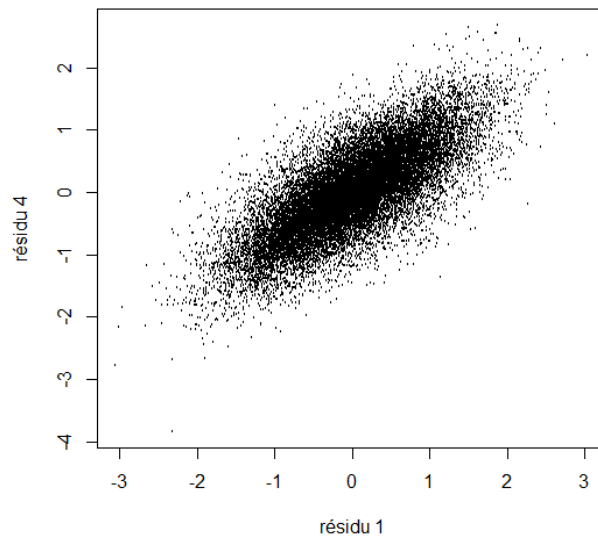


Copule de Clayton

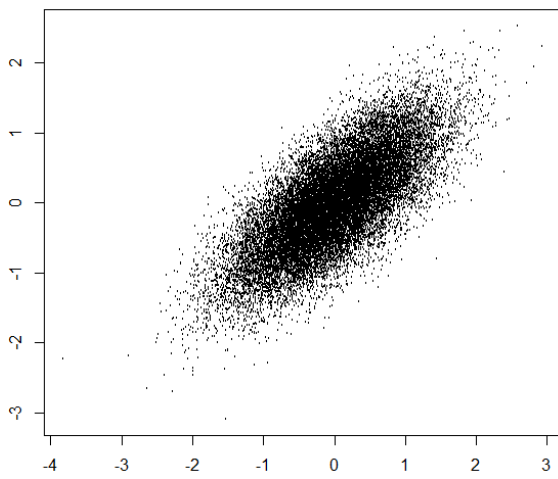


Nuages de points associés à la copule empirique et aux copules paramétriques pour les résidus 1 et 3

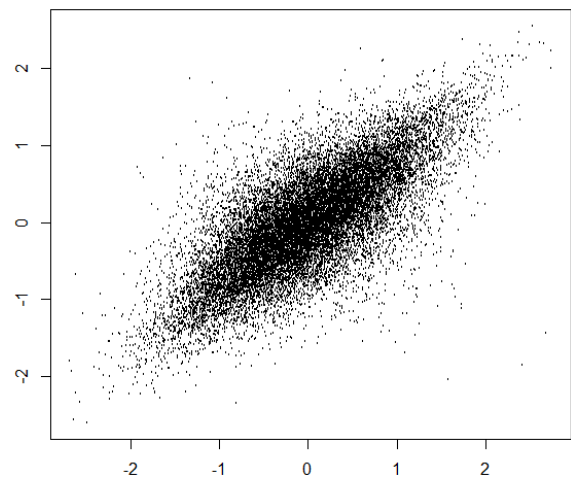
Copule empirique



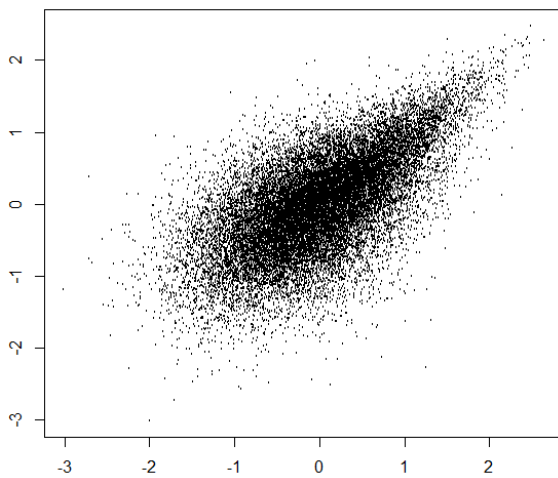
Copule Gaussienne



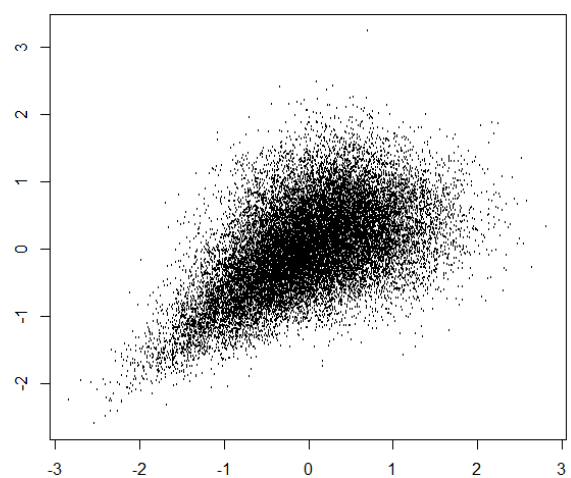
Copule de Student



Copule de Gumbel

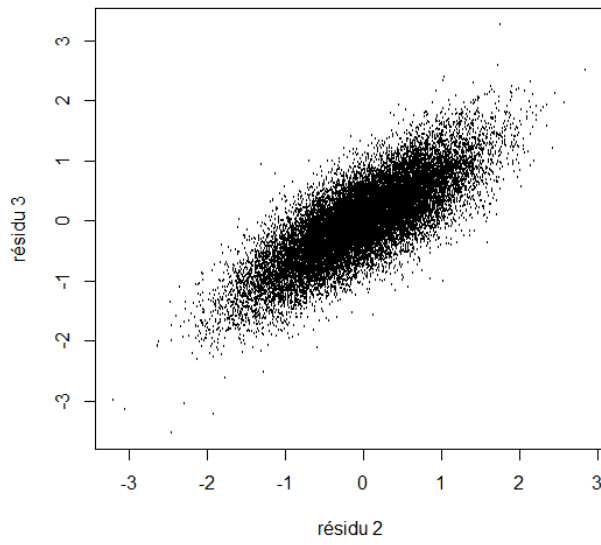


Copule de Clayton

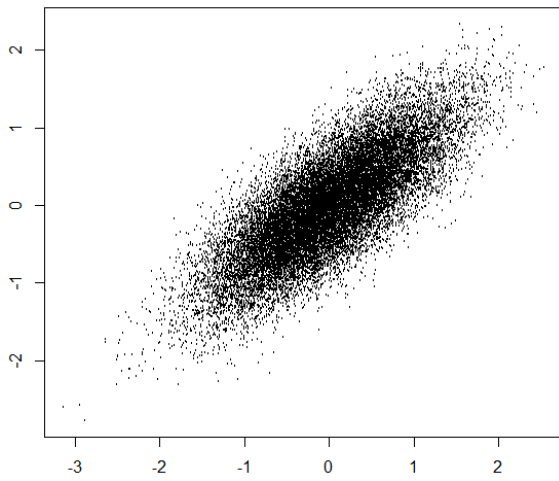


Nuages de points associés à la copule empirique et aux copules paramétriques pour les résidus 1 et 4

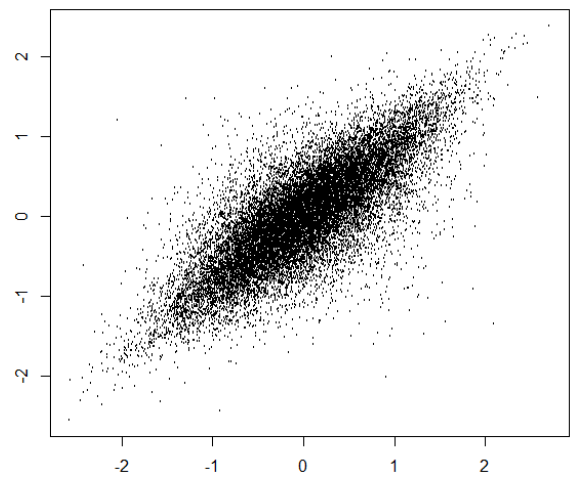
Copule empirique



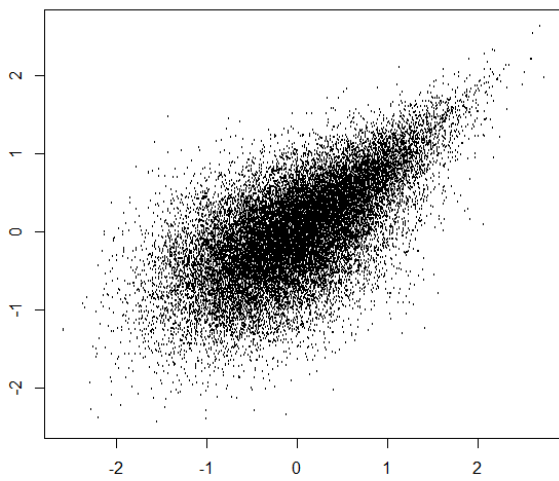
Copule Gaussienne



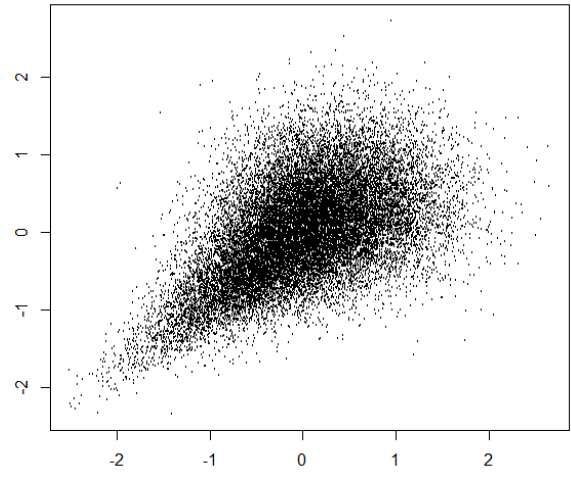
Copule de Student



Copule de Gumbel

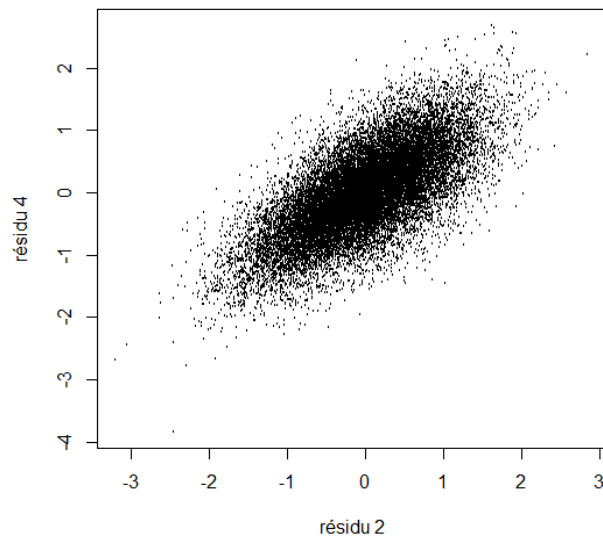


Copule de Clayton

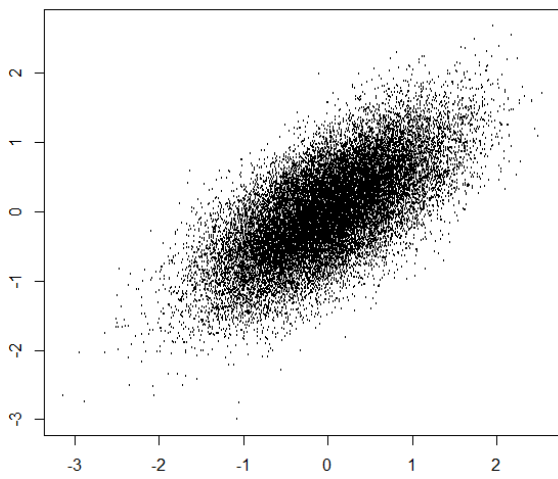


Nuages de points associés à la copule empirique et aux copules paramétriques pour les résidus 2 et 3

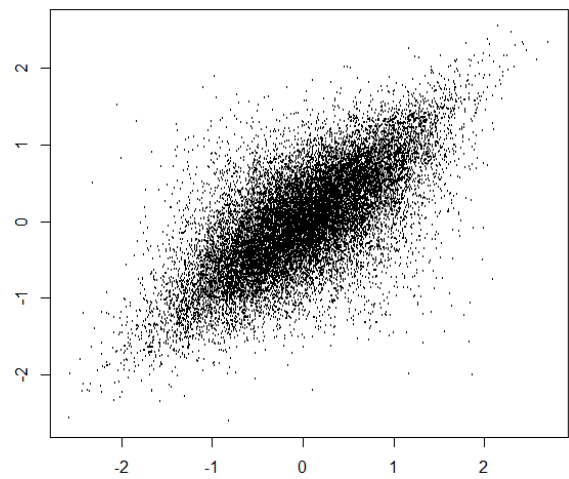
Copule empirique



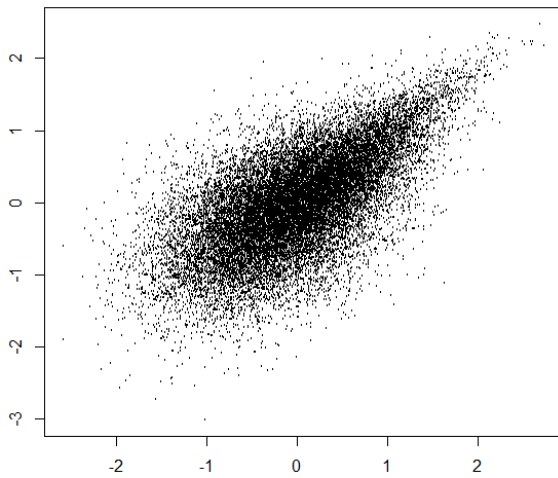
Copule Gaussienne



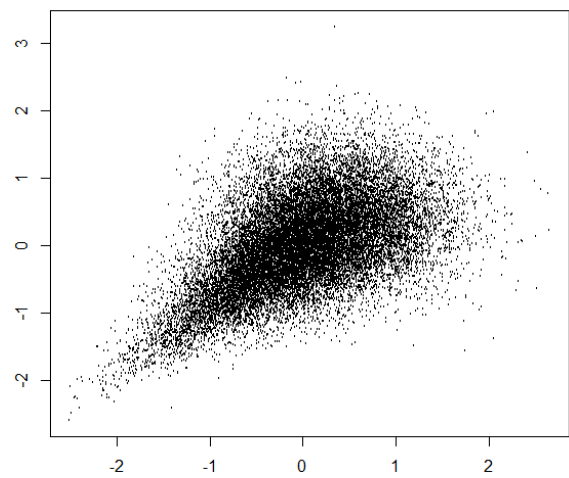
Copule de Student



Copule de Gumbel

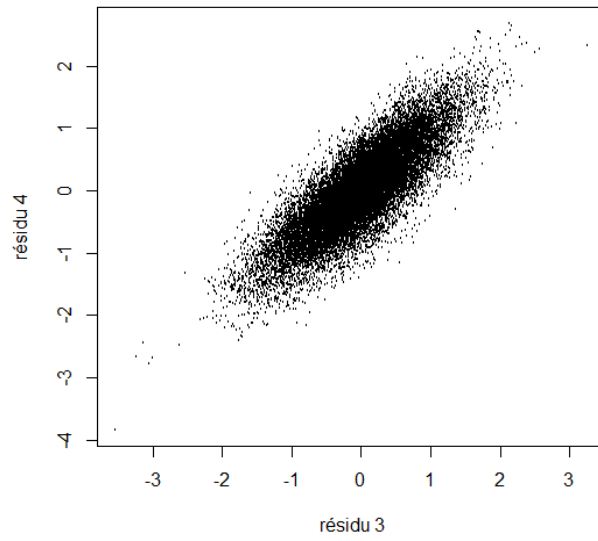


Copule de Clayton

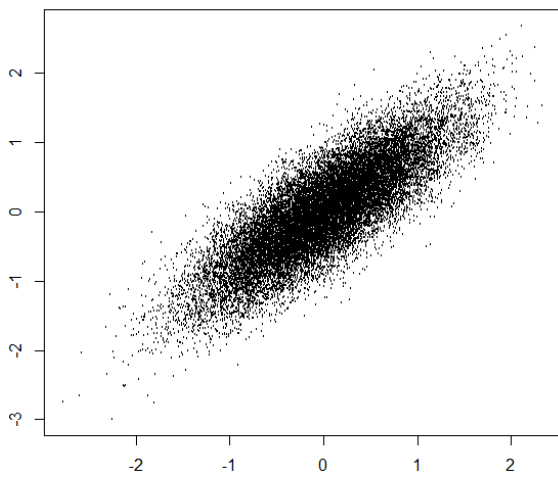


Nuages de points associés à la copule empirique et aux copules paramétriques pour les résidus 2 et 4

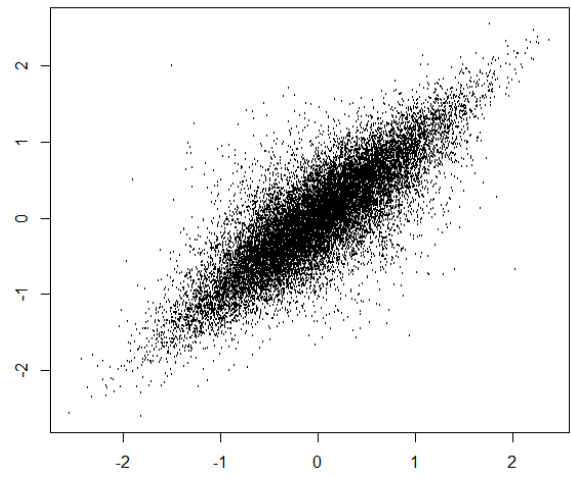
Copule empirique



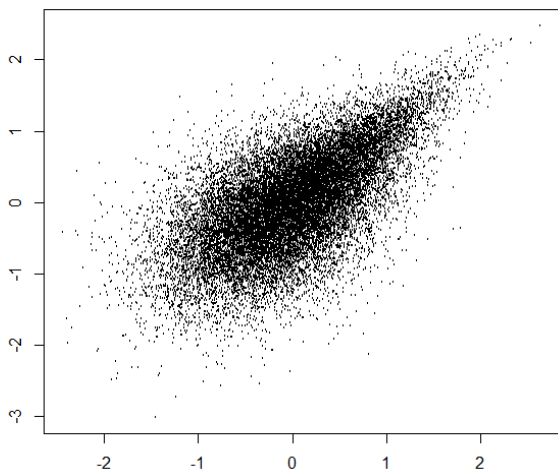
Copule Gaussienne



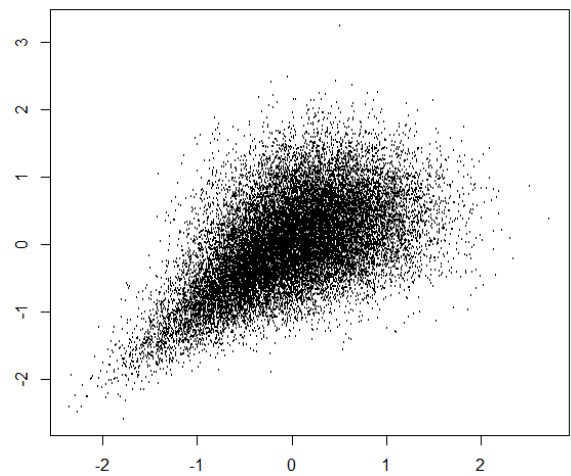
Copule de Student



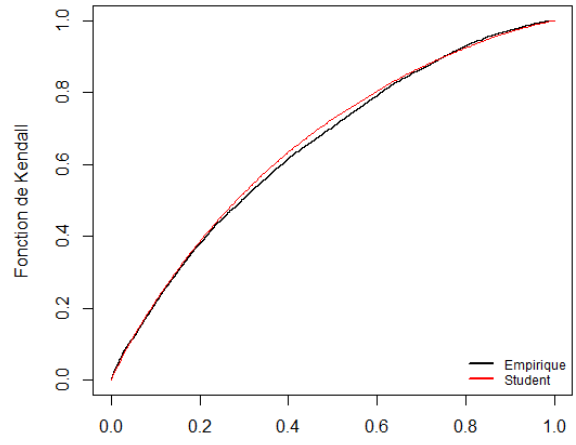
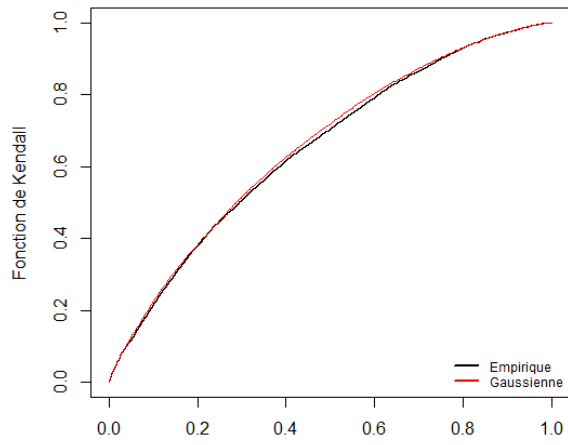
Copule de Gumbel



Copule de Clayton

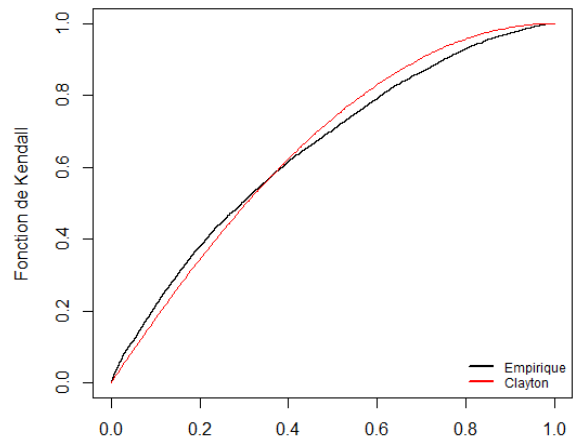
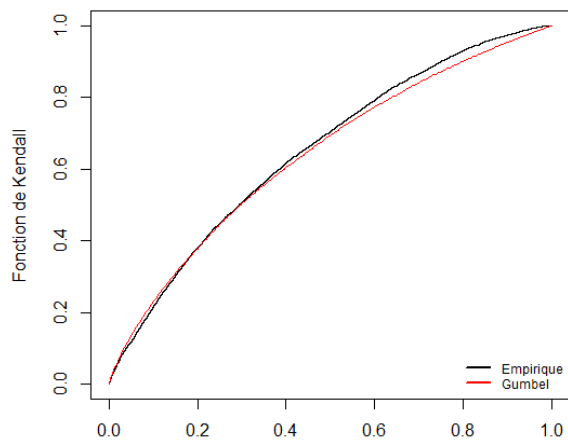


Nuages de points associés à la copule empirique et aux copules paramétriques pour les résidus 3 et 4



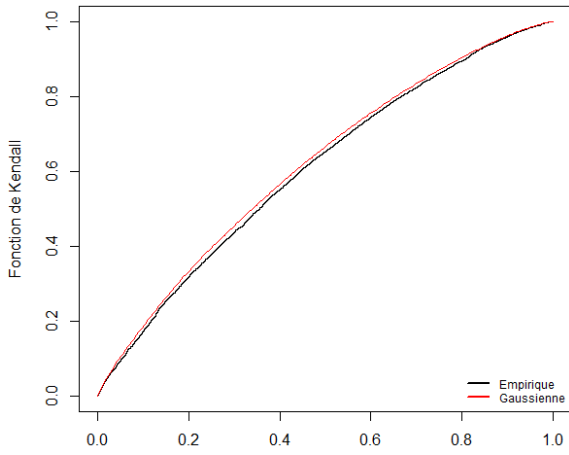
Copule de Gumbel

Copule de Clayton

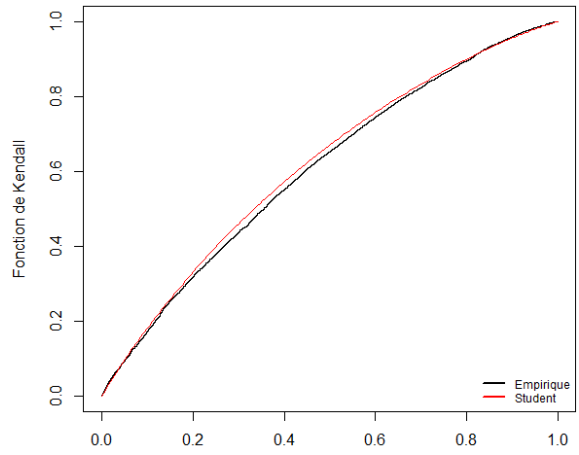


Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 1 et 2

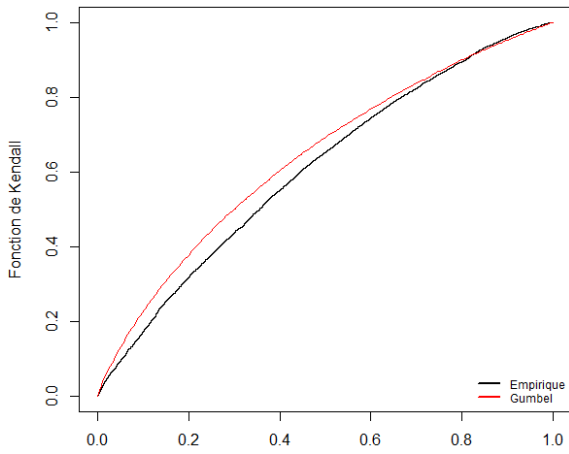
Copule Gaussienne



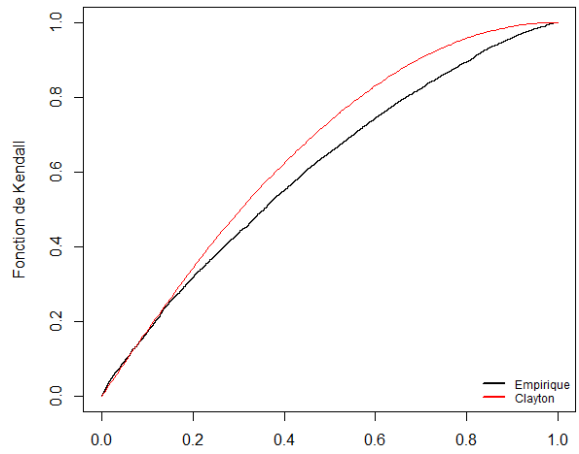
Copule de Student



Copule de Gumbel

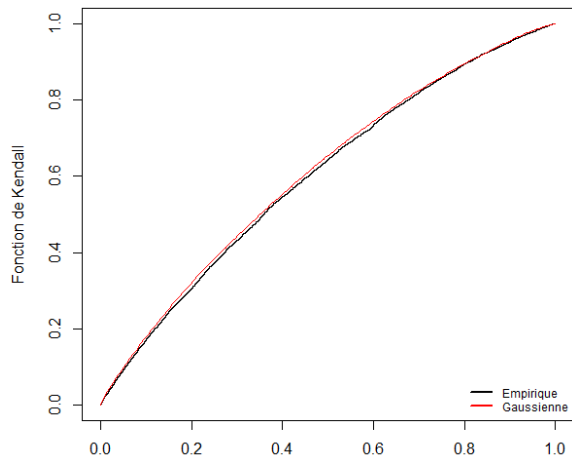


Copule de Clayton

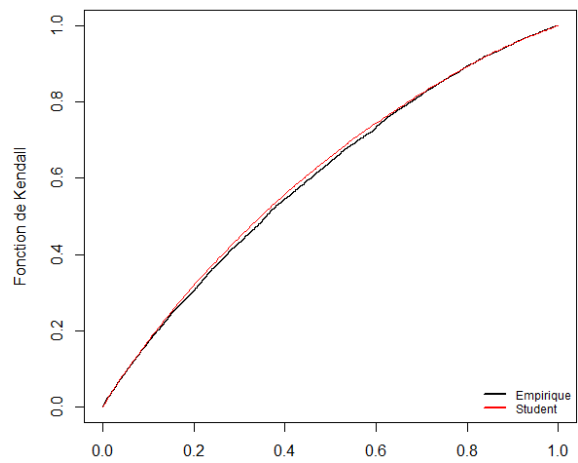


Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 1 et 4

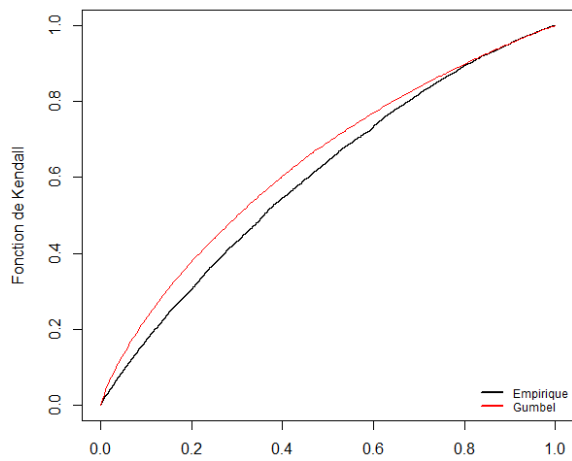
Copule Gaussienne



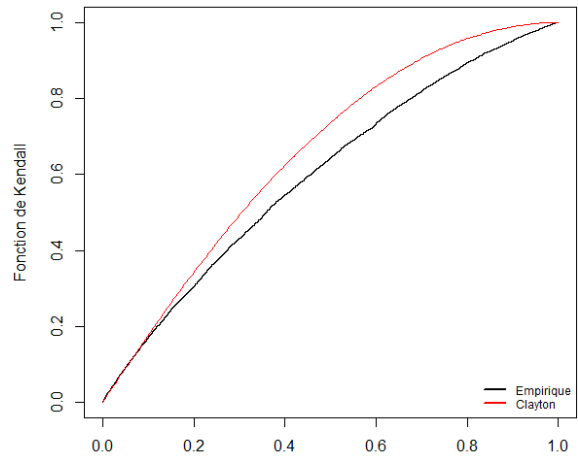
Copule de Student



Copule de Gumbel

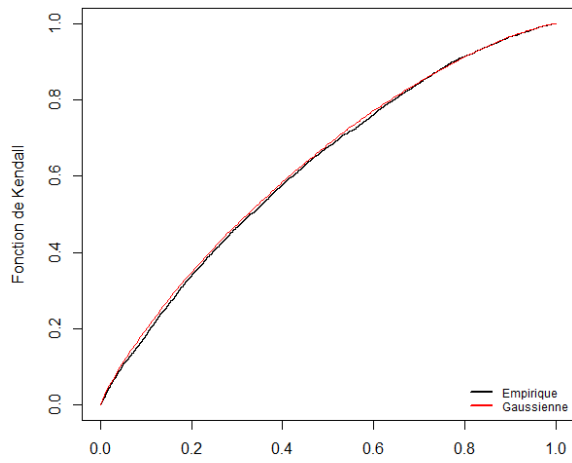


Copule de Clayton

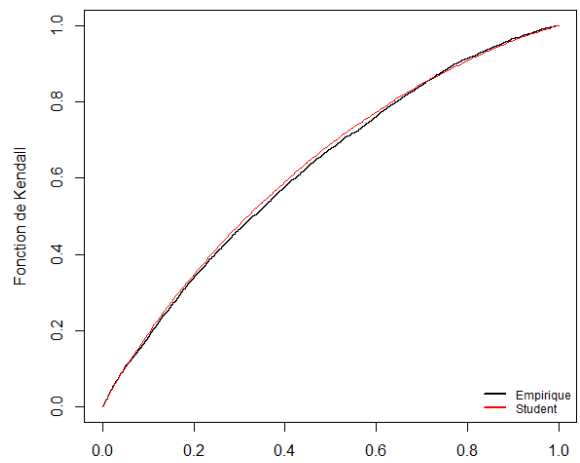


Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 2 et 3

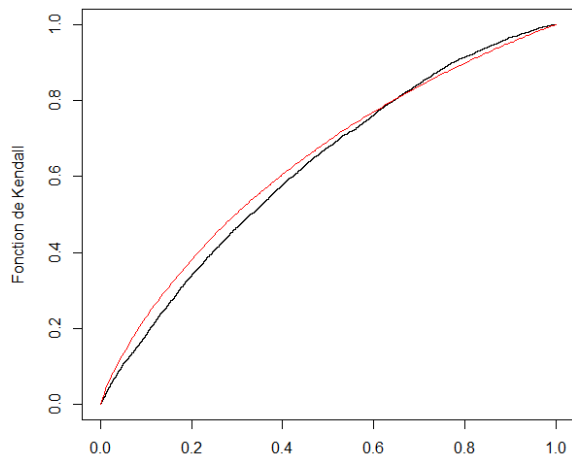
Copule Gaussienne



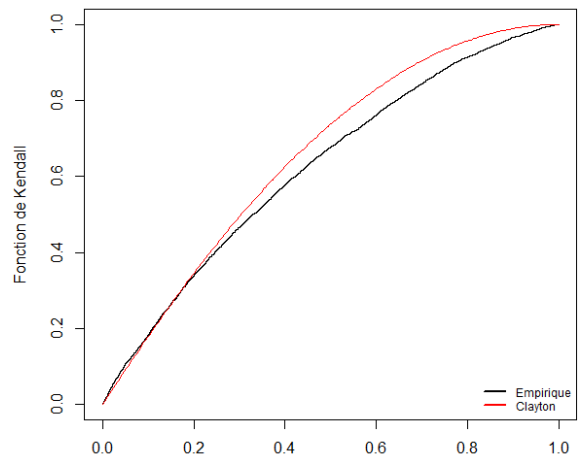
Copule de Student



Copule de Gumbel

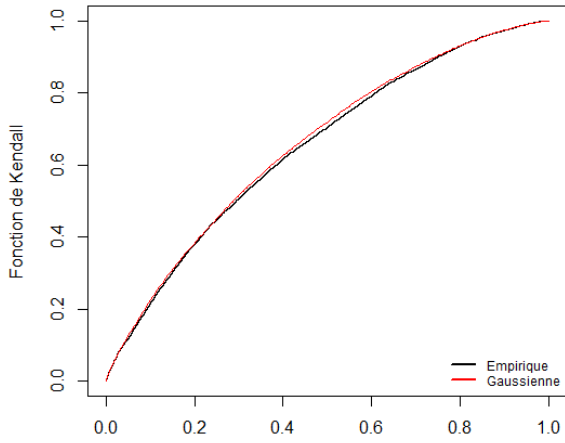


Copule de Clayton

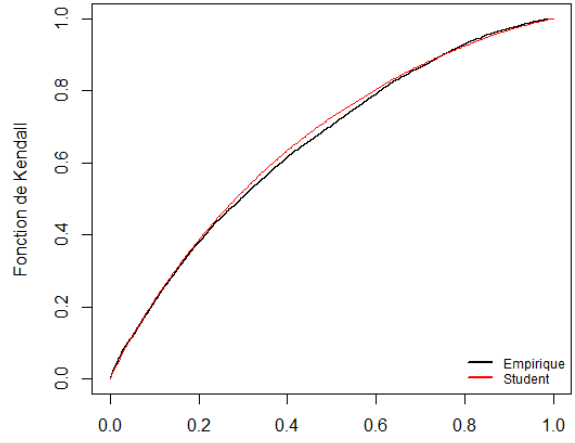


Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 2 et 4

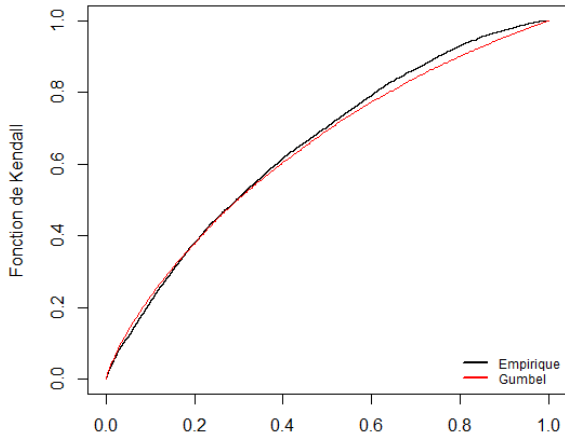
Copule Gaussienne



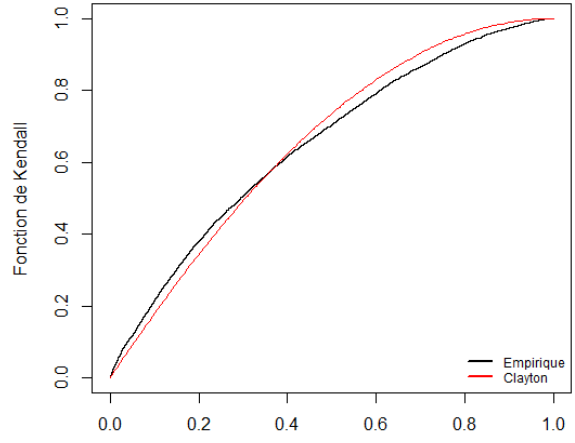
Copule de Student



Copule de Gumbel

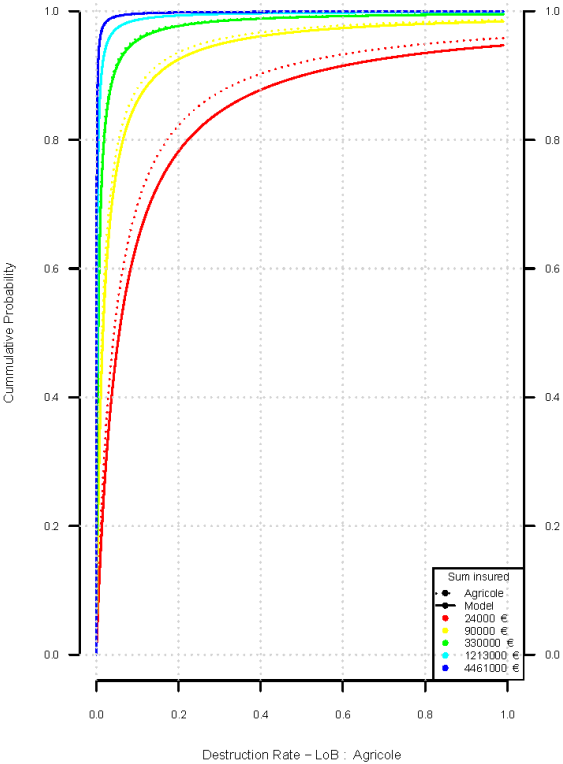


Copule de Clayton

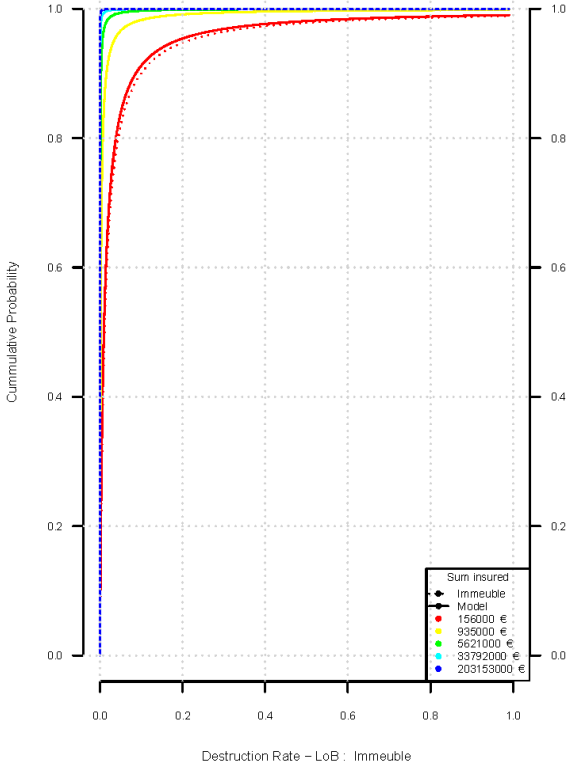


Comparaison des fonctions de Kendall empiriques et théoriques pour les résidus 3 et 4

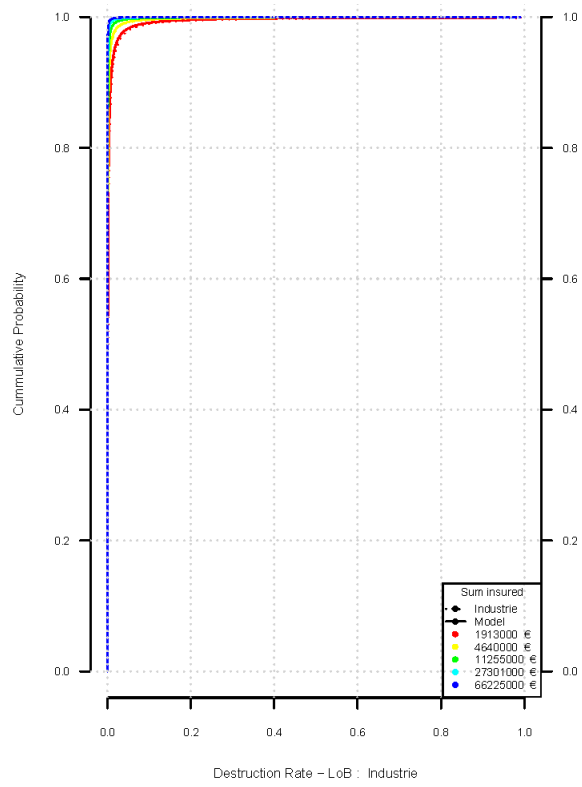
Annexe F - Courbes de destruction associées à chaque LoB



Courbe de destruction Agricole



Courbe de destruction Immeuble



Courbe de destruction Industrie