Master Actuariat de Dauphine

Mémoire présenté devant l'Université Paris Dauphine pour l'obtention du diplôme du Master Actuariat et l'admission à l'Institut des Actuaires

le _____

Par : Estèphe ARNAUD

Titre: Modélisation du risque sécheresse en France

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

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

Membre présent du jury de l'Institut des Actuaires :

Signature :

Entreprise :

Nom : AXA Group Risk Management

Signature :

Directeur de mémoire en entreprise :

Nom : Hugo d'ANTIN

Signature :

Membres présents du jury du Master Actuariat de Dauphine :

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 :

Secrétariat :

Bibliothèque :

Signature du candidat :

Résumé

Les risques de catastrophes naturelles, dont la fréquence d'occurrence est faible mais dont les coûts de sinistralité peuvent être très élevés, ne cessent d'augmenter depuis plusieurs décennies. Cela menace directement la solvabilité des compagnies d'assurance qui doivent faire face à des coûts de plus en plus importants. Dès 2016, la réglementation européenne Solvabilité II imposera aux assureurs de détenir suffisamment de fonds propres pour couvrir des événements extrêmes pouvant affecter le respect de leur engagement. Il est alors nécessaire de connaître le plus finement possible les risques encourus.

Le risque sécheresse a de plus en plus de poids parmi les risques de catastrophes naturelles. Pourtant, il n'a pas été encore étudié autant que les autres. Ce mémoire vise à modéliser le risque sécheresse en France. Pour cela, nous allons dans un premier temps chercher des variables et construire des indicateurs capables d'expliquer les périodes de sécheresse. Nous allons ensuite les modéliser afin de générer un grand nombre de scénarios météorologiques réalistes et probabilisés. Dans un second temps, nous allons utiliser l'historique de sinistralité d'AXA lié à la sécheresse afin de quantifier le lien entre ces variables explicatives et la fréquence de sinistralité. La modélisation des pertes financières sera réalisée indépendamment de la fréquence de sinistralité. Les scénarios météorologiques seront alors traduits en scénarios de pertes financières liées à la sécheresse. Nous pourrons alors finalement représenter la distribution des pertes financières d'AXA causées par la sécheresse.

Mots clés : catastrophe naturelle, sécheresse, classification, série temporelle, théorie des copules, modèle linéaire généralisé, apprentissage automatique, forêt aléatoire, courbe de destruction, MBBEFD, *AEP*, *OEP*

Abstract

Natural catastrophe risk – which a risk with low frequency and high severity – has kept increasing over the last decades. This directly threatens the solvency of insurance companies that need to face increasingly important costs. From 2016, the European Solvency II regulations have introduced requirements for insurers to have enough capital to cover extreme events that may affect the compliance of their commitment. Thus, it is necessary to understand the risks as precisely as possible.

The drought risk is becoming more and more important among natural catastrophe risks. Yet, it has not been studied as much as the others. This thesis aims to model the drought risk in France. In that respect, we will first look for variables and construct indicators capable of explaining droughts. We will then model them in order to generate a large number of realistic and probabilized meteorological scenarios. Secondly, we will use AXA's historical background of claims in relation to drought in order to quantify the relationship between these variables and the frequency of claims. The modeling financial losses will be carried out regardless of the frequency of claims. The meteorological scenarios will then be translated into financial loss scenarios related to drought. We will, eventually, be able to represent the distribution of AXA's financial losses caused by drought.

Keywords: natural catastrophe, drought, classification, time series, copula theory, generalized linear model, machine learning, random forest, destruction curve, MBBEFD, AEP, OEP

Synthèse

L'objectif de ce mémoire est de modéliser le risque sécheresse en France afin d'apporter les outils nécessaires d'aide à la décision pour l'optimisation de la gestion de portefeuille au titre de la sécheresse. Cette nécessité intervient dans le cadre de la réglementation européenne Solvabilité II, qui entrera en vigueur dès le 1er janvier 2016, et qui contraint les assureurs de disposer de suffisamment de fonds propres pour couvrir une perte bicentenaire, c'est-à-dire une perte qui arrive en moyenne une fois tous les 200 ans.

Le poids du risque de sécheresse ne cesse de croître au sein des risques de catastrophes naturelles en raison du changement climatique et du développement urbain dans des zones vulnérables. Peu de modèles de sécheresse ont déjà été développés.

La modélisation du risque sécheresse est divisée en trois étapes :

- Définition d'un ensemble de variables explicatives de la sécheresse, et modélisation de ces variables. Cela permet de générer un catalogue de scénarios réalistes et probabilisés d'évolution mensuelle des variables explicatives de la sécheresse.
- Modélisation de la fréquence de sinistralité à l'aide d'une forêt aléatoire (technique d'apprentissage automatique) et modélisation des coûts de sinistralité à l'aide de la méthode MBBEFD (technique développée chez Swiss Re).
  Ainsi, les scénarios d'évolution mensuelle des variables explicatives de la sécheresse sont traduits en scénarios d'évolution mensuelle des pertes financières générées par la sécheresse. Les pertes financières enregistrées dans l'historique de sinistralité liée à la sécheresse sont nettes des conditions contractuelles.
- Synthèse des résultats : représentations de la distribution des pertes financières causées par la sécheresse.

Définition des variables explicatives de la sécheresse

La sécheresse est essentiellement expliquée par les évolutions de comportement des précipitations mensuelles (déficit pluviométrique empêchant le bon remplissage des nappes phréatiques) et des températures maximales journalières (une période prolongée de températures élevées accentue l'assèchement des sols par le phénomène d'évapotranspiration).

C'est sur la modélisation de ces deux variables que se base celle de l'ensemble des variables explicatives de la sécheresse. Les connaissances scientifiques sur ce sujet et le croisement de données avec l'historique de sinistralité d'AXA liée à la sécheresse nous ont amené à caractériser la sécheresse par sept variables :

- Localisation géographique : on utilise les codes *CRESTA (Catastrophe Risk Evaluation and Standardizing Target Accumulations)* qui, pour la France, correspondent aux départements.
- La précipitation mensuelle.
- La température maximale enregistrée dans le mois.
- Le nombre maximal de jours depuis 60 jours où la température journalière est supérieure à 30°C
- L'indicateur *SPI (Standardized Precipitation Index)*, fondé sur la probabilité de précipitations estimée par une loi Gamma, qui quantifie l'écart des précipitations d'une période, déficit ou surplus, par rapport aux précipitations moyennes historiques de la période.

- L'indicateur *SPEI (Standardized Precipitation Evapotranspiration Index)*, construit de la même manière que le *SPI*, mais fondé sur la probabilité de précipitations nettes d'évapotranspiration estimée par une loi log-logistique.
- Le niveau moyen d' « aléa retrait-gonflement des argiles » : certains minéraux argileux présents dans les sols peuvent varier de volume en fonction de la teneur en eau des terrains, ce qui change la structure des sols et peut entraîner des dégâts matériels aux bâtiments (fissures, décollement). L' « aléa » mesure alors, pour une zone géographique donnée, à quel point les minéraux argileux peuvent varier de volume : plus il est élevé, plus il y a de chances d'avoir des dégâts en période de sécheresse.

Modélisation des variables explicatives de la sécheresse

Les précipitations ont été modélisées par *CRESTA* et par mois. En effet, pour un *CRESTA* donné, nous pouvons supposer que le processus des précipitations cumulées est stationnaire par année, mais pas par mois. Une loi Gamma suffit alors à représenter la distribution des précipitations cumulées par mois et par *CRESTA*. Nous disposons de 1152 modèles de précipitations mensuelles[1]. Les résultats sont très satisfaisants.

Les températures maximales journalières ont été modélisées pour cinq régions en France, homogènes en termes de températures[2]. Pour chaque région, la série temporelle des températures est décrite par trois composantes :

- Une tendance, modélisée par régression linéaire.
- Une saisonnalité, modélisée par régression sinusoïdale.
- Une série résiduelle qui, après avoir été « réduite » par l'écart-type mobile de cette série, est modélisée par un processus ARMA.

Après avoir développé un modèle de température pour chacune des cinq régions, nous avons modélisé la dépendance entre les régions, à l'aide de la théorie des copules, afin d'obtenir une cohérence entre la température d'une région et celle d'une autre pour un jour donné.

Pour tester la pertinence des résultats, nous avons effectué un *backtesting*, permettant de comparer la répartition des trajectoires obtenues avec les données réelles sur une partie restreinte des historiques. Les résultats obtenus sont satisfaisants.

Il est alors possible de simuler un grand nombre de trajectoires donnant une évolution des précipitations mensuelles et des températures maximales journalières pour l'année à venir.

Les variables explicatives de la sécheresse qui sont aléatoires et qui doivent être modélisées découlent des précipitations mensuelles et des températures maximales journalières. Les modélisations de ces dernières permettent alors de générer un catalogue de 10 000 scénarios d'évolution mensuelle des variables explicatives de la sécheresse pour l'année à venir.

[1] Il y a 96 *CRESTA* en France, et 12 mois dans une année. Il y a donc $96 \times 12 = 1152$ modèles.

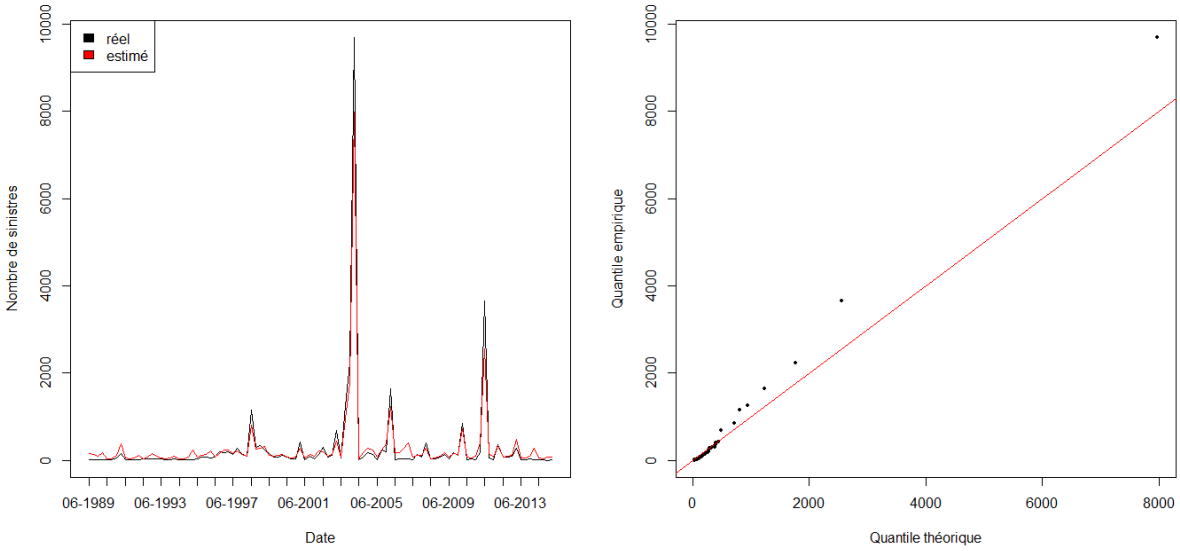
[2] Les relevés de températures sont disponibles pour 251 points équirépartis en France. Au lieu de développer un modèle de température en chacun de ces points, nous allons les regrouper en plusieurs régions de manière optimale afin d'avoir un nombre restreint de modèles en perdant le minimum d'informations.

Modélisation de la fréquence de sinistralité

La fréquence de sinistralité est le rapport entre le nombre de sinistres et le nombre de contrats présents dans le *CRESTA* considéré. Il suffit alors de modéliser le nombre de sinistres.

Afin de quantifier le lien entre les variables explicatives de la sécheresse et le nombre de sinistres qui en découle, nous proposons une technique récente d'apprentissage automatique : les forêts aléatoires. Cet algorithme effectue un apprentissage sur de multiples arbres binaires de décision entraînés sur des sous-ensembles de données légèrement différents. Pour chaque arbre de décision, on utilise une partie restreinte des données et une partie restreinte des variables explicatives. On sépare ensuite les données de manière optimale selon les valeurs prises par les variables explicatives.

Les résultats obtenus sont synthétisés par le graphique suivant :



Modélisation du nombre de sinistres

Le graphique de gauche représente l'évolution du nombre mensuel réel de sinistres (en noir) et l'évolution du nombre mensuel estimé de sinistres (en rouge).

Le graphique de droite représente le *Q-Q plot* associé, qui compare la répartition des quantiles empiriques avec les quantiles théoriques générés par le modèle.

Les distributions réelles et estimées du nombre mensuel de sinistres sont semblables : l'interprétation graphique est satisfaisante. Cependant, le critère quantitatif utilisé (le « pseudo-R^2 » du modèle), mesurant le rapport entre la variance des résidus et la variance totale pour chaque arbre, nous indique que le modèle n'est pas totalement satisfaisant. Les modèles plus traditionnels, tels que le modèle linéaire généralisé, n'étaient pas capables de reproduire les pics de sinistralité, contrairement au modèle construit avec une forêt aléatoire. Nous décidons alors de conserver ce modèle.

Modélisation des coûts de sinistralité

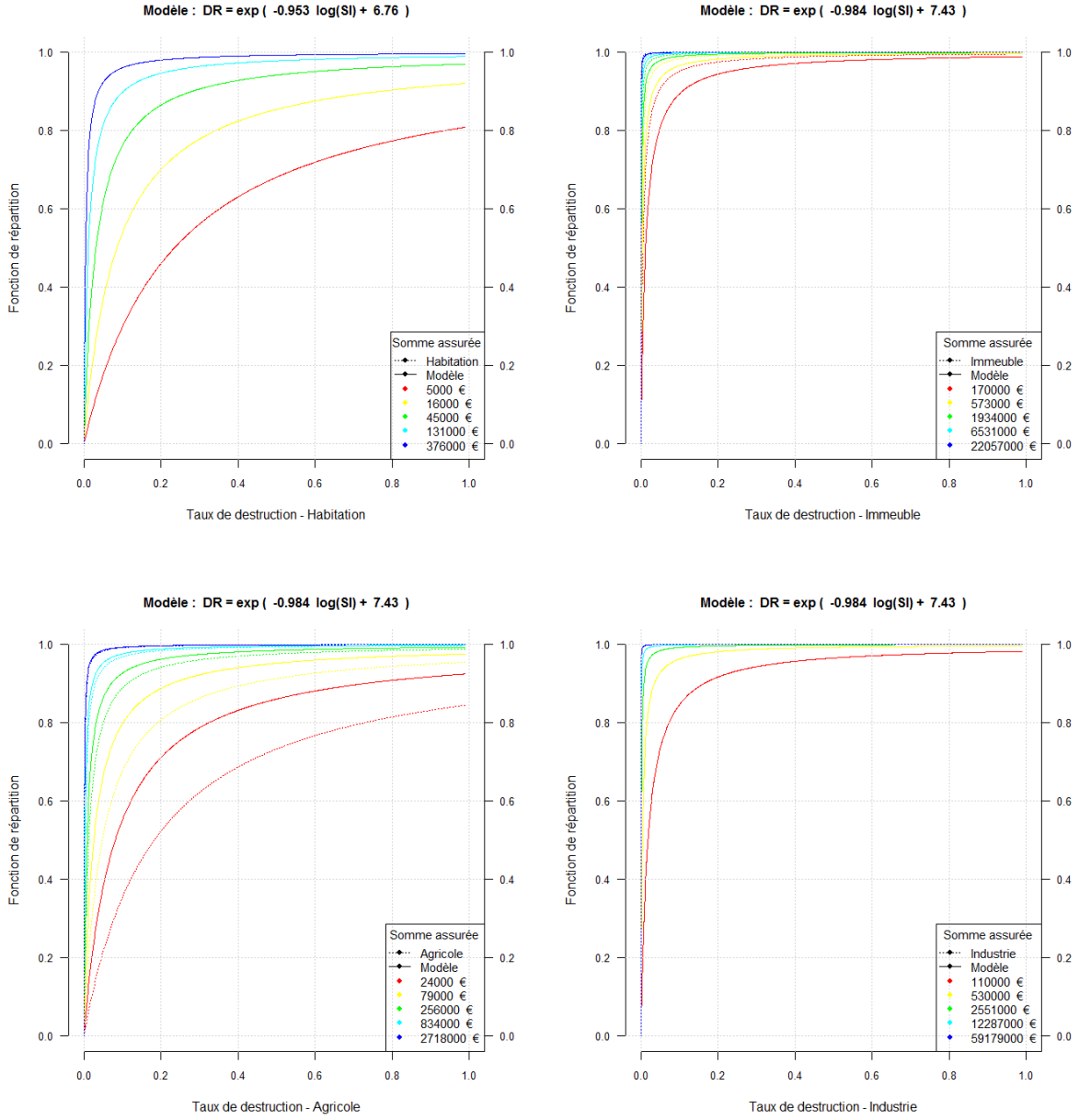
Le coût de sinistralité d'un objet assuré en période de sécheresse peut être considéré comme indépendant des variables explicatives de la sécheresse : le coût dépend essentiellement de la catégorie de l'objet assuré (MRH, Agricole, Immeuble, Industrie) et de la somme assurée.

Pour chaque catégorie d'objet assuré, et par couche de somme assurée, on observe un lien log-linéaire décroissant entre les taux de destruction[3] médians et les sommes assurées.

La méthode MBBEFD, développée récemment chez Swiss Re, permet de définir une fonction de répartition caractérisée par la médiane de la distribution.

Ainsi, pour une catégorie d'objet assuré et une somme assurée données, l'estimation du taux de destruction médian paramètre la fonction de répartition des taux de destruction.

Les graphiques suivants donnent les résultats obtenus pour les quatre catégories citées ci-dessus.



Les résultats sont cohérents : si nous considérons un dégât d'un montant fixé, alors le taux de destruction est décroissant avec la somme assurée.

[3] $\text{taux de destruction} = \frac{\text{montant du sinistre}}{\text{somme assurée}}$

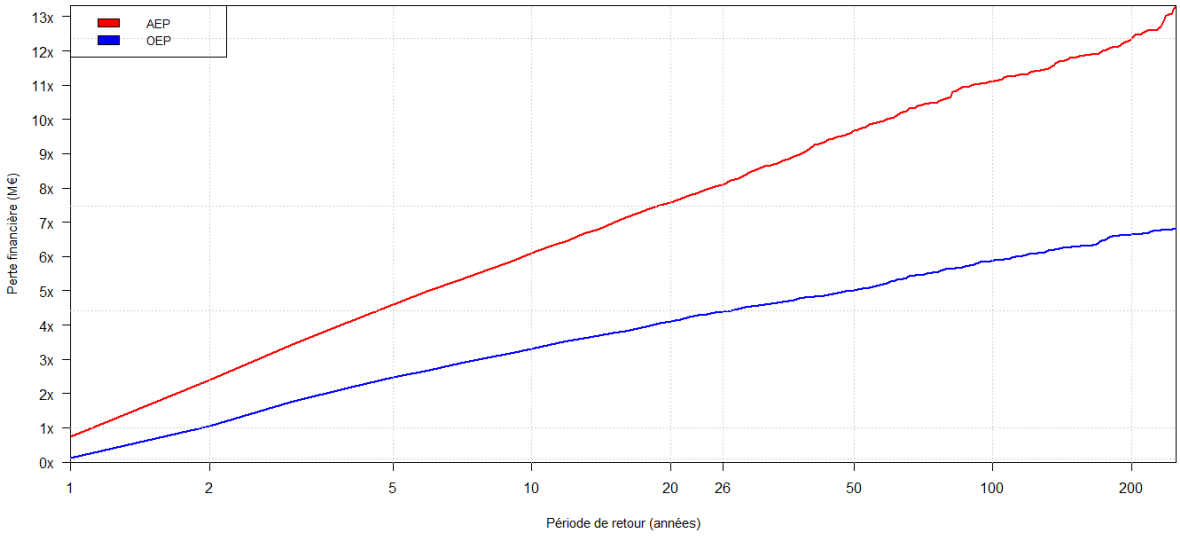
Ainsi, les scénarios d'évolution mensuelle des variables explicatives de la sécheresse peuvent être traduits en nombre de sinistres, puis en pertes financières.

Synthèse des résultats

Nous disposons d'un catalogue de 10 000 scénarios réalistes et probabilisés d'évolution mensuelle des pertes financières causées par la sécheresse. Nous définissons un événement sécheresse comme étant la perte financière accumulée sur un mois. Afin de synthétiser les résultats, nous construisons deux types de courbes :

- La courbe *AEP* (*Annual Exceedance Probability*) : associe une période de retour[4] au coût total des événements sur une année. Cette courbe permet de déterminer le capital réglementaire requis sous la réglementation européenne Solvabilité II, correspondant au montant associé à la période de retour de 200 ans.
- La courbe *OEP* (*Occurence Exceedance Probability*) : associe une période de retour au coût maximal d'un événement sur une année. Cette courbe aide donc à optimiser la structuration des traités de réassurance, en quantifiant la distribution du coût maximal annuel d'un événement (pour une période de retour donnée).

Le graphique suivant représente les résultats obtenus. Les résultats étant confidentiels, nous ne donnerons que des ordres de grandeur.



Courbes *AEP* et *OEP*

La valeur x associée à l'axe des ordonnées correspond à la perte nette moyenne d'AXA causée par la sécheresse durant une année. Le pic de sinistralité enregistré en 2003 est presque huit fois plus élevé que la sinistralité annuelle moyenne. La période de retour associée est estimée à 20 ans, alors que notre historique de sinistralité s'étale sur 26 ans. La perte mensuelle maximale de 2003 est bien associée à une période de retour de 26 ans. Pour de faibles périodes de retour, le modèle a tendance à surestimer les pertes. Il est important de se rappeler que la fréquence de sinistralité causée par la sécheresse va augmenter ces prochaines décennies.

La perte bicentenaire estimée est presque deux fois plus élevée que la perte enregistrée en 2003.

[4] Période de retour : temps statistique entre deux occurrences de même intensité

Limites du modèle

Le lien entre les variables explicatives et la fréquence de sinistralité est difficile à quantifier : l'historique de sinistralité dont nous disposons n'est pas nécessairement fidèle au véritable historique des événements sécheresse (d'un point de vue purement physique et non assurantiel). Beaucoup de sinistres sont enregistrés en début ou fin de mois. Il est alors difficile d'obtenir une base de données des sinistres (*reporting* en anglais) de qualité suffisante pour représenter de manière optimale l'historique de la sécheresse. L'agrégation mensuelle des données allège malgré tout ce problème.

La définition d'un événement sécheresse reste une question délicate. En effet, il est difficile de déterminer quand commence et finit une période de sécheresse, uniquement à partir de l'historique de sinistralité. Nous avons alors défini un événement comme étant la perte financière accumulée mensuellement mais cela peut être affiné. Cependant, la connaissance de la distribution de la perte annuelle est indépendante de la définition d'un événement sécheresse. Cela permet d'obtenir une vision complète du risque sécheresse en France et de calculer le capital réglementaire imposé par la réforme Solvabilité II qui s'appliquera dès 2016.

Mots clés : catastrophe naturelle, sécheresse, classification, série temporelle, théorie des copules, modèle linéaire généralisé, apprentissage automatique, forêt aléatoire, courbe de destruction, MBBEFD, *AEP*, *OEP*

Synthesis

The objective of this paper is to model the drought risk in France and to bring the necessary tools of decision to optimize the portfolio management during drought. This need comes as part of the European Solvency II regulations, that will be applied in 1 January 2016. Insurers will have to hold sufficient capital to cover a bicentennial loss, which is a loss that happens once every 200 years.

The weight of the drought risk continues to grow among natural catastrophe due to climate change and urban development in vulnerable areas. Only a few drought models have already been developed.

Modeling drought risk is divided into three steps:

- Definition of a set of explanatory variables for the drought, and modeling of these variables. This will generate a catalog of realistic and probabilized meteorological scenarios.
- Modeling of the claims' frequency with a random forest approach (machine learning tool) and modeling of the claims' cost with the MBBEFD method (developed at Swiss Re).
  Thus, the meteorological scenarios can be interpreted as financial losses scenarios that are caused by drought. Financial losses in the history of loss due to drought are net of contractual conditions.
- Summary of results: representations of the distribution of financial losses caused by drought.

Definition of the drought's explanatory variables

Drought is mainly explained by the behavior of the monthly rainfalls (rainfall deficit preventing the proper filling of ground water) and daily maximum temperatures (a prolonged period of high temperatures increases the dewatering of the soils with an evapotranspiration phenomenon).

It is on the modeling of these two variables that of all the explanatory variables of the drought are based. Scientific knowledge on this subject and the data crossing with AXA claims history related to drought have led us to characterize drought with seven variables:

- Geographical location: we use the CRESTA (Catastrophe Risk Evaluation and Standardizing Target Accumulations) codes, which, for France, correspond to the departments.
- The monthly precipitation.
- The maximum temperature recorded in the month.
- The maximum number of days on a 60 days basis where the daily temperature is above 30 °C.
- The SPI (Standardized Precipitation Index) indicator, based on the probability of precipitation estimated by a gamma distribution, which quantifies the difference in precipitation of a period, deficit or surplus, compared with historical average rainfall for the period.
- The SPEI (Standardized Precipitation Evapotranspiration Index) indicator, constructed in the same manner as the SPI, but based on the probability of precipitation net of evapotranspiration estimated by a log-logistic distribution.
- The average level of "shrink–swell capacity": certain clay minerals in soils can vary in volume according to the land water content, which changes the structure of the soil and may cause damage to buildings (cracks, uprising). The "shrink–swell capacity" measures, for a given geographical area, how clay minerals may vary in volume: the higher it is, the more likely it is to have damage during drought.

Modeling drought explanatory variables

The precipitations were modeled by CRESTA and by month. Indeed, for a given CRESTA, we can assume that the process of cumulative rainfall is stationary per year, but not per month. A gamma distribution is then enough to represent the distribution of cumulative rainfall per month and per CRESTA. We have 1152 models of monthly precipitation[5]. The results are very satisfactory.

The maximum daily temperatures were modeled for five regions in France, homogeneous in terms of temperatures. For each region, the series of temperatures is described by three components:

- A trend, modeled by linear regression.
- Seasonality, modeled by sinusoidal regression.
- A residual series, which after having been "reduced" by the moving standard deviation of this series, is modeled by an ARMA process.

After developing a temperature model for each of the five regions, we modeled the dependence between regions[6], using the copula theory, in order to have consistent results regarding the different regions' temperature for a given day.

To test the relevance of the results, we performed a backtest, to compare the distribution of the trajectories obtained with a partial set of the real data of AXA's historical database. The results obtained are satisfactory.

It is then possible to simulate a large number of trajectories giving monthly rainfall and daily maximum temperatures for the coming year.

The drought's explanatory variables, that are random and that must be modeled, result of the monthly rainfall and daily maximum temperatures. Their modeling allow us to generate a catalog of 10 000 scenarios of the drought's explanatory variables for coming year.

Modeling of frequency claims

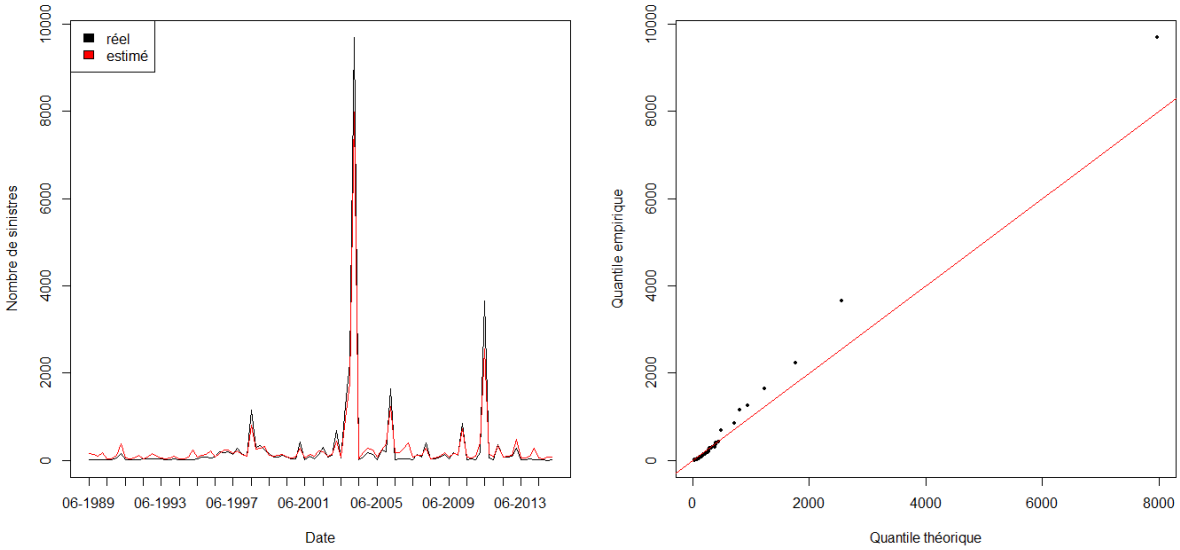
The claims' frequency is the ratio of the number of claims and the number of contracts considered in the present CRESTA. It is then sufficient to model the number of claims.

To quantify the relationship between the drought's explanatory variables and the related number of claims, we propose a new technique of machine learning: the random forests. This algorithm performs a learning on multiple binary decision trees trained on slightly different data subsets. For each decision tree, it uses a part of the data and a part of the explanatory variables. It separates optimally the data according to the values taken by the explanatory variables.

The results obtained are summarized in the following graph:

[5] There are 96 CRESTA in France and 12 months in a year. So there are $96 \times 12 = 1152$ models.

[6] The temperature readings are available for 251 points equally distributed in France. Instead of developing a temperature model each of these points, we will group them into several optimal regions in order to have a limited number of models losing the minimum information.



Modeling the number of claims

The left graph shows the evolution of the real monthly number of claims (in black) and evolution of the monthly estimated number of claims (in red).

The right graph represents the associated Q-Q plot that compares the empirical distribution with the theoretical quantile generated by the model.

The actual monthly distributions and estimated number of claims are similar: the graphical interpretation is satisfactory. However, the quantitative criterion (the "pseudo-R^2" model), measuring the ratio of the variance of the residuals and the total variance for each tree indicates that the model is not completely satisfactory. The more traditional models, such as generalized linear model, were not able to reproduce the peaks of loss, unlike the model built with a random forest. We decide then to maintain this model.

Modeling of loss

The cost of loss of an insured object during drought may be considered independent explanatory variables of drought: cost mainly depends on the category of the insured object (Home, Agricultural, Building, Industry) and the insured values.

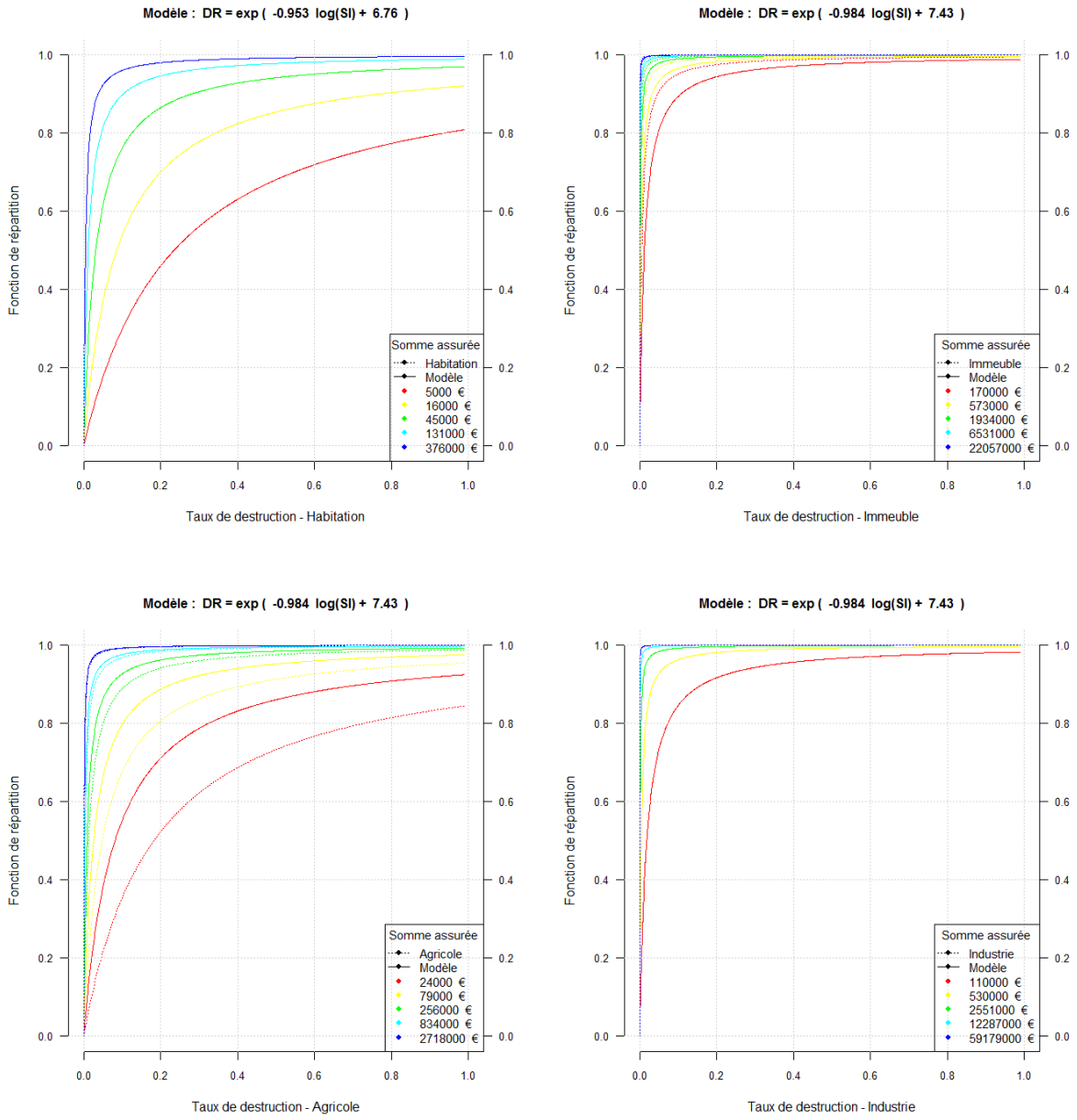
For each category of insured object, and layer of insured values, there is a log-linear and decreasing relationship between the median destruction rate[7] and the insured values.

MBBEFD method has recently developed at Swiss Re, to define a distribution function characterized by the median of the distribution.

Thus, for a category of insured object and insured values, the median estimate of destruction rate parameter the distribution function of destruction rate.

The following graphs show the results obtained for the four categories mentioned above.

[7] $\text{destruction rate} = \frac{\text{amount of loss}}{\text{insured values}}$



Modeling cost of loss

The results are consistent: if we consider a damage of a fixed amount, then destruction rate is decreasing with the insured values.

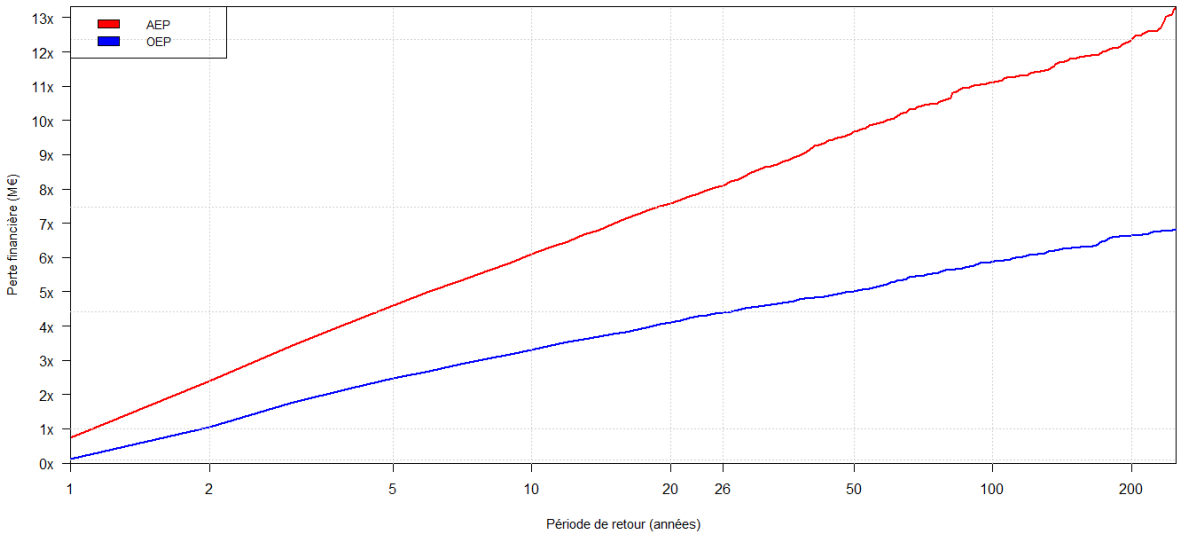
Thus, the scenarios of monthly evolution of drought explanatory variables can be translated into number of claims and in financial losses.

Summary of results

We have a catalog of 10 000 realistic scenarios and probabilistic monthly evolution of financial losses caused by drought. We define an event as drought accumulated financial loss over a month. To summarize the results, we build two types of curves:

- The AEP curve (Annual Exceedance Probability): combines a return period[8] to the total cost of events in one year. This curve is used to determine the regulatory capital required under European regulations Solvency II, corresponding to the amount associated with the return period of 200 years.
- The OEP curve (Occurence Exceedance Probability): combines a return period to the maximum cost of an event over a year. Therefore, this curve helps to optimize the structuring of reinsurance treaties, quantifying the distribution of the maximum annual cost of an event (for a given return period).

The following graph shows the results obtained. The results are confidential, we only give orders.



Curve AEP and OEP

The value associated with the x-axis corresponds to the average net loss of AXA caused by drought for a year. The peak of loss recorded in 2003 is almost eight times higher than the average annual loss. The associated return period is estimated at 20 years, while our historical loss experience spans 26 years. The maximum monthly loss of 2003 is associated with a return period of 26 years. For low return periods, the model tends to overestimate losses. It is important to remember that the frequency of loss caused by drought will increase in the coming decades.

The estimated loss bicentennial is almost twice higher than the loss recorded in 2003.

Limitations of the model

The link between the explanatory variables and the frequency of claims is difficult to quantify: the history of loss experience we have is not necessarily true to the real history of drought events (from a purely physical point of view and not insurance). Many claims are recorded at the beginning or end of the month. It is then difficult to obtain a reporting of sufficient quality to represent an optimal manner the history of drought. The monthly data aggregation alleviates this problem anyway.

[8] Return period: statistical time between occurrences of the same intensity

The definition of a drought event remains a sensitive issue. Indeed, it is difficult to determine when to begin and end a drought, only from claims history. We then defined an event as the monthly accumulated financial loss but it can be refined. However, knowledge of the distribution of the annual loss is independent of the definition of a drought event. This provides a comprehensive view of drought risk in France and it allows to calculate the regulatory capital required by Solvency II will apply in 2016.

Keywords: natural disasters, drought, classification, time series, copula theory, generalized linear model, machine learning, random forest, destruction curve, MBBEFD, AEP, OEP

Remerciements

Je remercie l'équipe CAT du GIE AXA pour m'avoir permis de réaliser mon stage de fin d'étude dont la problématique m'intéresse particulièrement. Leur accueil et le cadre de travail m'ont permis de profiter pleinement de cette expérience professionnelle.

Je tiens spécialement à remercier mon maître de stage, Hugo d'ANTIN, pour ses explications, ses suggestions, sa gentillesse et sa clairvoyance. Il a su me guider dans les moments de blocage et m'a aidé dans l'élaboration de mon mémoire d'actuariat.

Enfin, je souhaite exprimer mes meilleurs sentiments aux autres stagiaires et tiens à transmettre toute mon amitié à Omar JERRARI avec qui j'ai pu échanger de longues discussions tout au long du stage.

Sommaire

Résumé i
Synthèse ii
Remerciements xiv
Sommaire xv

Introduction 1

I. Présentation du risque sécheresse 3
A. La sécheresse, un risque météorologique et agricole 3
1. La sécheresse comme type de catastrophe naturelle 3
2. Origines et conséquences de la sécheresse 3
3. Cas historiques de sécheresse 4
B. La sécheresse, un risque assurantiel 6
1. Principes de l'assurance et de la modélisation des risques 6
2. Le régime CAT NAT en France 7
3. Intérêts de la modélisation des catastrophes naturelles 8
C. Synthèse des données sur la sécheresse 12
1. Les événements majeurs de sécheresse survenus en France 12
2. Les chiffres de la CCR 12
3. Présentation des données utilisées pour la modélisation 15
D. Etapes de modélisation du risque sécheresse 18
1. Module Aléa 18
2. Module Vulnérabilité 19
3. Module Financier 19
4. Résultats du modèle 20

II. Module Aléa : simulations des variables expliquant la sécheresse 23
A. Les variables explicatives de l'événement sécheresse 23
1. Les précipitations mensuelles 23
2. L'indice *SPI* (*Standardized Precipitation Index)* 25

3. Les températures maximales journalières 27
4. L'évapotranspiration 29
5. L'indice *SPEI* (*Standardized Precipitation Evapotranspiration Index)* 30
6. L'aléa retrait-gonflement des argiles 32
B. Modélisation des précipitations mensuelles 34
1. Stationnarité du processus des précipitations mensuelles 34
2. Estimation paramétrique de la loi des précipitations mensuelles 36
C. Segmentation de la France en régions homogènes en termes de température 40
1. Les principes de la classification 40
2. L'analyse par composante principale 41
3. La classification hiérarchique ascendante 43
D. Modélisation des températures maximales journalières 47
1. Décomposition des séries temporelles des températures 47
2. Modélisation des résidus (Yt) 53
3. Validation du modèle 55
E. Modélisation des dépendances entre régions 56
1. La théorie des copules 56
2. Construction des copules paramétrées 58
3. Critère pour choisir la meilleure copule 58
4. Application 59
F. Simulations 63
1. Simulations des précipitations et des températures 63
2. Simulations des variables explicatives 64

III. Module Vulnérabilité et Financier : simulations des pertes financières causées par la sécheresse 65
A. Modélisation de la fréquence de sinistralité avec les modèles linéaires généralisés .. 66
1. Le modèle linéaire généralisé 66
2. Modélisation sans dépassement de seuil 69
3. Modélisation avec indicateur à seuil 72
B. Modélisation de la fréquence de sinistralité avec une forêt aléatoire 76
1. L'apprentissage automatique et les forêts aléatoires 76
2. Application 79
C. Modélisation des coûts de sinistralité par les courbes de vulnérabilité 80
1. La méthode MBBEFD 81

2. Liens entre les taux de destruction médians et les sommes assurées 83
3. Construction des courbes de vulnérabilité 85

IV. Résultats du modèle 87
A. Ajustement pertes estivales / pertes annuelles 87
B. Construction des courbes *AEP* et *OEP* 89
1. Définition d'un événement sécheresse du point de vue assurantiel 89
2. Les courbes *AEP* et *OEP* 89
C. Approche fréquence/coût 91
1. Hypothèses 91
2. Résultats 92

Conclusion 94

Annexe A : Evapotranspiration – Coefficients de correction 95
Annexe B : Dépendances entre régions – Comparaisons entre les copules empiriques et les copules elliptiques 97
Annexe C : Variables explicatives – Graphiques 102
Annexe D : Tests statistiques 107
Annexe E : Ajustement de modèle – Coefficients 110

Table des figures 114
Liste des tableaux 116
Bibliographie 117

Introduction

Le changement climatique annoncé par la communauté scientifique et le développement urbain dans des zones vulnérables permettent d'anticiper une augmentation des risques de catastrophes naturelles d'ici ces prochaines années. Le monde de l'assurance est particulièrement affecté par cette problématique. Nous pouvons citer l'ouragan Katrina de 2005 aux Etats-Unis qui a coûté aux assureurs 80 milliards de dollars, le tsunami de 2011 au Japon qui a coûté près de 40 milliards de dollars, ou même la tempête Xynthia de 2010 en France qui a coûté 2,5 milliards d'euros.

Il est alors fondamental de connaître et de quantifier les risques afin de garantir la solvabilité des assureurs. La réglementation européenne Solvabilité II, qui entrera en vigueur dès le 1er janvier 2016, contraint les assureurs de disposer de fonds propres en quantité suffisante pour couvrir une perte bicentenaire, c'est-à-dire une perte qui arrive en moyenne une fois tous les 200 ans. Connaître le risque d'un point de vue quantitatif permet alors de calculer ce niveau de fonds propres exigé. En outre, cela permet d'optimiser le transfert de risques par le biais de la réassurance.

Afin de modéliser les risques de catastrophes naturelles, plusieurs approches existent. Une des plus développée est une approche stochastique qui consiste à simuler l'événement physique pour ensuite le traduire en perte financière. La modélisation du phénomène physique est essentielle car tout en découle. Plusieurs sociétés spécialisées en catastrophes naturelles vendent des licences d'utilisation de modèles aux assureurs et réassureurs. Parmi les risques concernés, on peut citer les tremblements de terre, les tornades, les inondations, les tempêtes de neige, les tornades, ou les glissements de terrain. Cependant, le risque sécheresse n'a pas encore été à l'étude par ces sociétés. Les assureurs doivent donc développer leur propre modèle pour ce risque.

La sécheresse est un risque en pleine expansion. Sa modélisation va devenir de plus en plus nécessaire pour les acteurs de l'assurance. En effet, l'association MRN[9] estime que le tiers des indemnités catastrophes naturelles versées depuis 1982 concernent des sinistres sécheresse, contre 12 % vu fin 1993. Certaines périodes de sécheresse ont généré de très nombreux dégâts : on peut notamment citer la sécheresse de 2003 qui a coûté aux assureurs près de 1,3 milliards d'euros.

Afin de modéliser le risque sécheresse, nous allons appliquer l'approche stochastique de modélisation des catastrophes naturelles en trois étapes.

Dans une première partie, nous allons définir un ensemble de variables explicatives de la sécheresse. Pour cela, nous utiliserons l'historique de sinistralité d'AXA lié à la sécheresse et nous le croiserons avec des variables et des indicateurs afin de ne garder que les plus pertinents. Nous modéliserons ensuite ces variables afin de les simuler et ainsi établir un catalogue de scénarios réalistes et probabilisés d'évolution mensuelle des variables explicatives de la sécheresse.

Dans une seconde partie, nous allons modéliser indépendamment la fréquence et les coûts de sinistralité. Nous allons quantifier les liens existant entre la sinistralité liée à la sécheresse et les variables explicatives de la sécheresse. La fréquence de sinistralité sera d'abord modélisée par une approche traditionnellement utilisée en assurance : le modèle linéaire généralisé. Nous testerons ensuite une approche récente développée en apprentissage automatique : les forêts aléatoires. Les coûts de sinistralité seront modélisés grâce à une méthode récemment développée chez Swiss Re : la méthode MBBEFD. Les simulations d'évolution des variables explicatives de la sécheresse seront alors traduites en pertes financières. Nous

[9] MRN : Mission Risques Naturels

disposerons finalement d'un catalogue de scénarios réalistes et probabilisés d'évolution mensuelle des pertes financières au titre de la sécheresse.

Dans une dernière partie, nous synthétiserons l'ensemble des simulations effectuées par une distribution des pertes financières causées par la sécheresse. Cela permettra d'avoir une vision complète du risque de sécheresse et ainsi apporter les informations nécessaires au calcul du capital réglementaire et à l'optimisation de réassurance.

I. Présentation du risque sécheresse

A. La sécheresse, un risque météorologique et agricole

1. La sécheresse comme type de catastrophe naturelle

Une catastrophe naturelle est un événement brutal, d'origine naturelle, qui se produit avec une fréquence faible et une intensité élevée. Elle peut engendrer des destructions matérielles et humaines massives.

La sécheresse est un phénomène dévastateur, qui se produit avec une fréquence faible et une intensité élevée. Elle peut donc être considérée comme une catastrophe naturelle. Mais contrairement aux autres catastrophes, sa phase de destruction est lente : elle peut durer plusieurs mois et avoir des effets ravageurs sur les écosystèmes et les productions agricoles.

2. Origines et conséquences de la sécheresse

Deux principales sources alimentent le risque de sécheresse : la faible quantité de précipitations (déficit pluviométrique) et les températures élevées sur une période prolongée.

A partir de l'hiver, les nappes phréatiques (masse d'eau contenue dans les fissures du sous-sol) se remplissent d'eau de pluie. A partir du printemps, cette eau peut s'évaporer au niveau du sol ou par la transpiration des plantes qui l'absorbent. Cela s'appelle l'évapotranspiration.

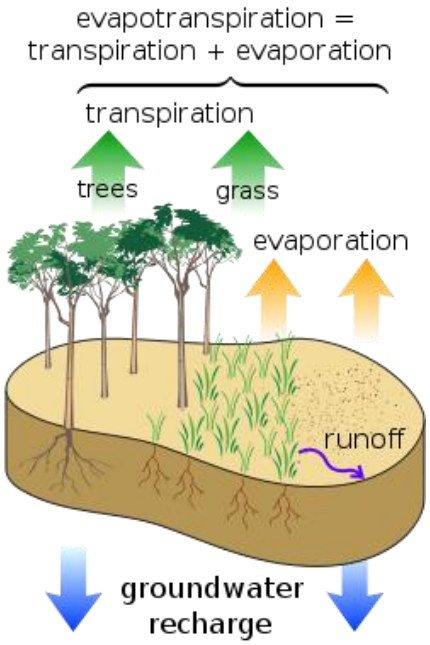


Figure 1 - Représentation schématique du bilan évapotranspiration – Source : Wikipédia

La sécheresse peut alors se produire lorsque la quantité de précipitations est suffisamment inférieure aux normales et depuis suffisamment longtemps empêchant le bon remplissage des nappes phréatiques. Cela peut être combiné avec des températures supérieures aux normales de saison et depuis suffisamment longtemps, accélérant le phénomène d'évapotranspiration et provoquant un assèchement des sols.

On peut distinguer deux types de sécheresse :

- La sécheresse « météorologique » : caractérisée par un déficit de précipitations sur une période prolongée.
- La sécheresse « agricole » : caractérisée par un déficit de la présence d'eau dans les sols, dépendant donc des précipitations et de l'évapotranspiration des sols et des plantes.

Les conséquences de la sécheresse peuvent être :

- La modification, les perturbations, voire la destruction des écosystèmes
- La diminution des rendements des cultures
- Des dommages matériels aux bâtiments (fissures, décollement) souvent causés par le phénomène de retrait-gonflement des argiles que nous décrirons dans la suite.

Les deux photos suivantes illustrent les conséquences de la sécheresse souvent constatées (fissures et pertes agricoles) :

Figure 2 - Conséquences de la sécheresse

3. Cas historiques de sécheresse

En France, les sécheresses les plus connues sont celles de 1976, 2003, et 2011 qui ont beaucoup dégradé les écosystèmes et les cultures, et ont eu par ailleurs un impact important sur la mortalité. Ces sécheresses sont associées à des périodes où les précipitations et les températures ont des comportements anormaux par rapport aux normales.

Le graphique suivant donne les écarts enregistrés des températures et des précipitations par rapport aux normales de saison, entre 1959 et 2015.

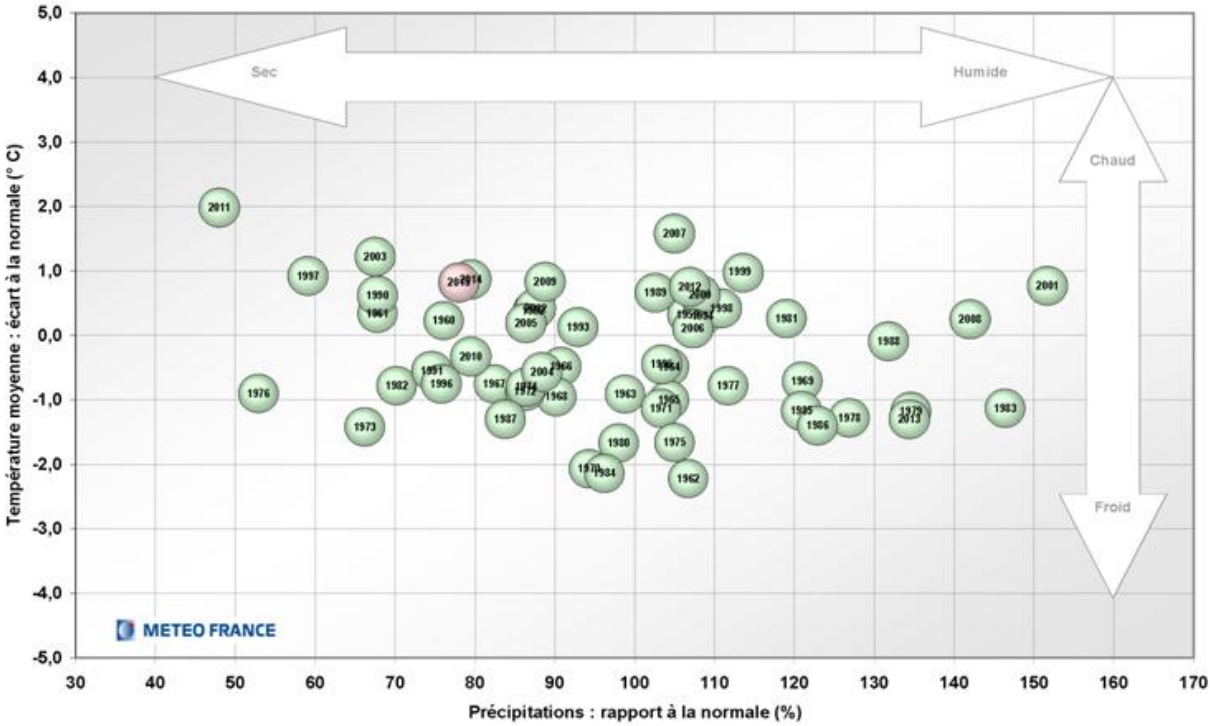


Figure 3 - Températures et précipitations au printemps de 1959 à 2015 – Source : Météo France

Ce graphique illustre bien le fait que l'apparition et l'intensité d'une sécheresse est d'abord liée à un manque exceptionnel de précipitations, mais aussi à une période exceptionnellement chaude (accentuant le phénomène d'évapotranspiration).

B. La sécheresse, un risque assurantiel

1. Principes de l'assurance et de la modélisation des risques

L'ensemble des destructions engendrées par ces catastrophes impacte directement le portefeuille d'un assureur.

En effet, une assurance est un service qui fournit une prestation, contre le paiement d'une prime, lors de la survenance d'un événement incertain et aléatoire (appelé "risque"). L'assureur doit donc être capable de mesurer et représenter le plus fidèlement possible la répartition de ce risque, afin d'honorer ses engagements.

Pour cela, il doit d'abord connaître son exposition. Cela se mesure par le type de contrat (MRH, Immeuble, Agricole, …), les caractéristiques du bien assuré (structure du bâtiment, …), la somme assurée (la valeur du bien assuré), ainsi que la localisation du risque assuré (par code INSEE[10], code *CRESTA*[11], ou code postal). L'assureur doit être capable de hiérarchiser les risques entre des zones préalablement définis : le sud de la France est plus exposé au risque de sécheresse que la Bretagne.

Ensuite, il devra modéliser la fréquence de survenance du risque concerné. Dans notre cas, le risque de catastrophe naturelle a, par définition, une fréquence de survenance faible. A première vue, cela peut sembler impossible de capter un risque qui n'arrive que très rarement, mais les progrès scientifiques (sur la caractérisation du phénomène physique) associés aux développements de nouveaux modèles statistiques donnent des résultats prometteurs.

Le graphique suivant donne la fréquence des catastrophes naturelles de 1988 à 2006 :

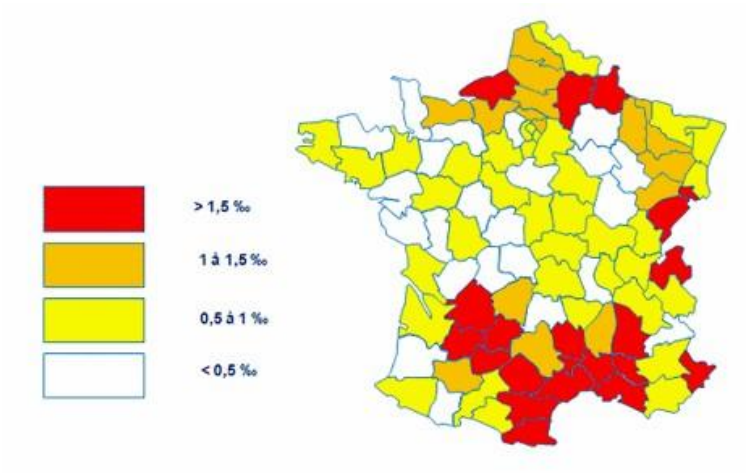


Figure 4 - Fréquence des catastrophes naturelles sur la période 1988-2006 – Source : FFSA

Enfin, sachant qu'il y a un événement ayant généré des sinistres, l'assureur doit être capable d'en estimer l'intensité et les coûts de sinistralité générés. Il doit alors modéliser l'intensité du risque concerné et les coûts de chaque contrat affecté. Dans notre cas, le risque de catastrophe naturelle a, par définition, une intensité élevée : lorsqu'une catastrophe se produit, elle peut être fatale à l'assureur si celui-ci avait sous-estimé l'intensité, même s'il avait correctement modélisé sa fréquence d'occurrence. A titre d'exemple, nous pouvons citer le cyclone Andrew de 1992 en Floride, qui a provoqué la faillite de onze assureurs.

[10] Code INSEE : code officiel géographique à cinq chiffres répertoriant les 36 794 communes françaises
[11] Code *CRESTA* : *Catastrophe Risk Evaluation and Standardizing Target Accumulations*. En France, les zones *CRESTA* correspondent aux départements.

Après avoir hiérarchisé les risques par zone d'exposition, modélisé la fréquence de survenance et l'intensité du risque concerné, et modélisé les coûts de sinistralité par contrat, l'assureur est capable de modéliser le coût global du risque concerné.

Le graphique suivant donne le coût global des catastrophes naturelles de 2002 à 2007 :

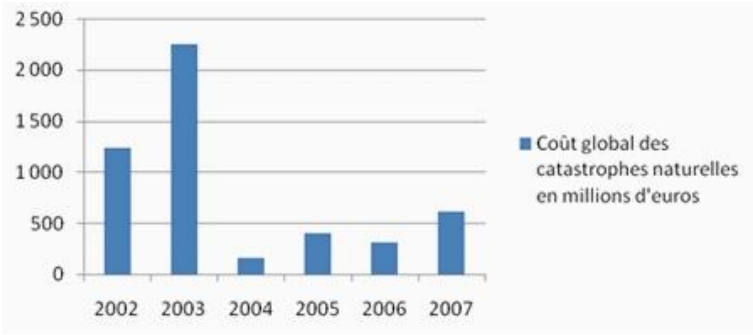


Figure 5 - Coût global des catastrophes naturelles en millions d'euros – Source : FFSA

Par une connaissance pointue de l'ensemble des risques assurés, il aura tous les outils en mains pour optimiser la gestion de son portefeuille.

Plusieurs modèles cherchant à estimer la sinistralité potentielle de catastrophes naturelles ont déjà été créés. Il en existe deux grands types :

- Les modèles produits par les assureurs et réassureurs
- Les modèles produits par des sociétés spécialisées qui vendent des licences d'utilisation aux assureurs et réassureurs

Les trois principales sociétés spécialisées développant ces modèles sont :

- *Risk Management Solutions (RMS)*, qui développe le logiciel *RiskLink*
- *EQECAT*, qui développe le logiciel *RQE*
- *Applied Insurance Research (AIR)*, qui développe le logiciel *Touchstone*

Ces logiciels sont souvent considérés comme des « boîtes noires » : les assureurs manquent parfois de documentation leur empêchant d'avoir le recul suffisant pour analyser finement les risques. Ils souhaitent alors développer leurs propres modèles en complément. Depuis 2009, plusieurs acteurs européens du marché de l'assurance et de la réassurance (Allianz, Axa, Generali, Groupama, Munich Re, Partner Re, Swiss Re) se sont regroupés, à travers une initiative baptisée « PERILS », pour constituer une base de données à partir des expositions aux risques des assureurs et ainsi affiner les modèles déjà existants. Les modèles proposés s'appliquent à la tempête, l'inondation, et le tremblement de terre, mais pas encore à la sécheresse. Ils s'appliquent à des grands pays, en raison de leur plus grande exposition.

2. Le régime CAT NAT en France

Aux termes de la loi initiée en juillet 1982, sont considérés comme effets des catastrophes naturelles « des dommages matériels directs non assurables ayant eu pour cause déterminante l'intensité anormale d'un agent naturel lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises » (Article L. 125-1 alinéa 3 du Code des assurances). Les événements le plus

souvent constatés sont les inondations, les coulées de boue, la sécheresse et, dans une moindre mesure, l'action mécanique des vagues, les glissements et affaissements de terrain, les avalanches, raz de marée et les tremblements de terre.

En assurant ses biens contre l'incendie, les dégâts des eaux, le vol…[12], l'assuré est automatiquement couvert contre les dégâts dus aux catastrophes naturelles. La garantie catastrophes naturelles prévoit la prise en charge des dommages matériels causés aux biens assurés.

Cette garantie joue **seulement** si un arrêté interministériel paru au Journal officiel constate l'état de catastrophe naturelle. Il permet l'indemnisation des dommages directement causés aux biens assurés.

La France fait partie des rares pays qui garantissent à chacun de ses citoyens une indemnisation correcte en cas de sinistre causé par une catastrophe naturelle.

Son régime de catastrophes naturelles (CAT NAT) est financé par :

- Des primes : tout assuré souscrivant un contrat de dommage aux biens paie une prime d'assurance forfaitaire. Quelques soient le type de risque et l'exposition aux risques naturels, un taux de surprime est fixé par l'Etat et est exprimé en pourcentage de la prime dommage payée par l'assuré : une surprime de 12 % de la prime pour un contrat multirisque habitation, et une surprime de 6 % de la prime pour un contrat d'assurance d'un véhicule.
- Des franchises : le niveau de franchise, non indexé, a été fixé par arrêté. Il varie de 380 € à 3050 € en fonction des biens couverts à usage des particuliers ou des professionnels.

La CCR[13], entreprise de réassurance française créée en 1946 et détenue en totalité par l'Etat français, propose aux assureurs de partager la moitié des coûts de sinistralité jusqu'à 200 % des primes. Au-delà de cette limite, elle prend en charge la totalité des coûts de sinistralité. Cela signifie que les assureurs ne peuvent pas perdre plus que les primes acquises. Ils peuvent donc couvrir leurs assurés contre les catastrophes naturelles sans connaître parfaitement les risques encourues.

Cependant, une mauvaise connaissance du risque n'optimise pas la gestion du portefeuille et peut générer d'importantes pertes financières.

3. Intérêts de la modélisation des catastrophes naturelles

Comme dit précédemment, les catastrophes naturelles peuvent avoir un impact désastreux sur le portefeuille des assureurs. Ces derniers doivent alors nécessairement modéliser ces risques pour :

- Optimiser le transfert de risque par la cession d'une partie de ses primes (perçues) à un réassureur
- Calculer le niveau de fonds propres qu'il doit détenir sous les nouvelles réglementations comptables de Solvabilité II pour couvrir des pics de sinistralité

[12] Assurance dommages aux biens
[13] CCR : Caisse Centrale de Réassurance

a) Optimisation de réassurance

La réassurance est l'assurance des sociétés d'assurance. Pour s'assurer, l'assureur peut céder une partie des primes qu'il perçoit. On l'appelle « cédante ». La réassurance permet aux cédantes de faire face à des pics de sinistralités exceptionnels, diminuant ainsi leur probabilité de faillite.

En plus d'augmenter la sureté financière de la cédante en intervenant les années où il existe de nombreux sinistres, la réassurance lui permet aussi de souscrire plus d'affaires, et de lui lisser ses bilans et ses résultats financiers d'une année à l'autre.

Le graphique suivant illustre l'impact de la réassurance sur le résultat technique :

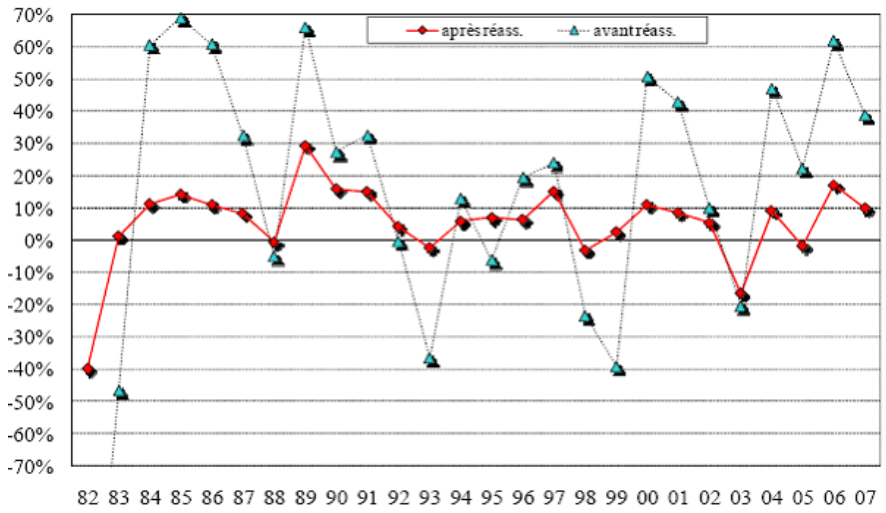


Figure 6 - Impact de la réassurance sur le résultat technique – Source : FFSA

Le résultat technique est exprimé en pourcentage des primes avant et après réassurance. Nous pouvons observer l'effet lissant de la réassurance.

On distingue deux types de réassurance :

- la réassurance proportionnelle : participation proportionnelle du réassureur aux gains (primes) et pertes (sinistres) de la cédante
- la réassurance non proportionnelle : le réassureur n'intervient qu'à partir d'un certain seuil de sinistre ou de perte de la cédante

Les traités de réassurance proportionnels sont :

- Traités quote-part : la cédante partage avec son réassureur un pourcentage fixé des primes et des sinistres pour toutes les polices
- Traités XP (excédent de plein) : la cédante partage avec son réassureur un pourcentage fixé des primes et des sinistres pour chaque police

Les traités de réassurance non-proportionnels sont :

- Traités XS (excédent de sinistres): défini par une priorité (ou franchise) et une portée (ou plafond). Le réassureur prend en charge la partie de tout sinistre qui excède la priorité du traité et dans la limite de la portée du traité (différence entre le plafond et la franchise).
- Traités *Stop-Loss* (excédent de perte annuelle) : intervention du réassureur lorsque la charge annuelle globale de sinistres (sur une branche donnée) dépasse un seuil déterminé

Les catastrophes naturelles sont essentiellement prises en charge par des traités de réassurance non-proportionnels. Ainsi, une connaissance pointue du risque concerné permettra de fixer de manière optimale la priorité et la portée de chaque traité.

b) Calcul du capital réglementaire

Solvabilité II est une réforme réglementaire européenne, qui a pour objectif principal la meilleure adaptation des fonds propres exigés des compagnies d'assurance et de réassurance face aux risques auxquels elles sont exposées dans leur activité. Elle s'inspire de la réforme Bâle II qui a eu lieu dans le secteur bancaire. Initialement prévue en 2010, la mise en place effective est prévue le 01/01/2016, du fait de la complexité du projet.

La solvabilité d'une compagnie d'assurance est sa capacité à répondre de ses engagements envers les bénéficiaires des contrats. Cette réforme permet de protéger la solvabilité de l'entreprise. Pour cela, elle lui impose de détenir un capital, appelé *SCR*[14], majorant la probabilité de faillite sur un horizon d'un an à 0,5 %[15].

Pour calculer leur *SCR*, les compagnies d'assurance ont le choix d'utiliser la formule standard ou de développer un modèle interne (partiel ou global).

Dans la formule standard, le *SCR* est la somme de trois éléments :

- Le *BSCR* (*Basic Solvency Capital Requirement*), composé de plusieurs modules de risques. Chaque module permet de couvrir un type de risque : risque de souscription vie, risque de souscription non-vie, risque de souscription santé, risque de marché, et risque de contrepartie.
  Chacun de ces modules est ensuite divisé en sous-module de risques.
- Le *SCR* opérationnel, correspondant aux risques qui ne sont pas pris en compte dans le *BSCR* : fraudes, fautes, erreurs de calcul, …
- Un ajustement, pour la capacité d'absorption des pertes par les provisions et impôts différés

Le schéma ci-dessous explique de manière plus visuelle la composition du *SCR* par la formule standard.

[14] *SCR : Solvability Capital Requirement*

[15] Cela correspond à une faillite tous les 200 ans en moyenne ($0{,}5\ \% = \frac{1}{200}$).

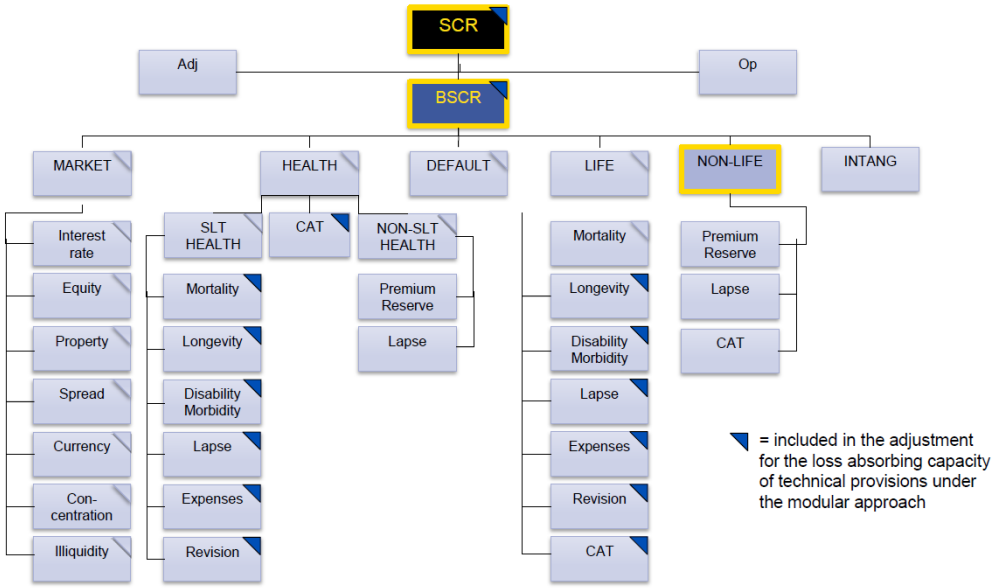


Figure 7 - Calcul du *SCR* par une combinaison de modules – Source : *EIOPA*

Nos données regroupent l'ensemble des dommages matériels causés par la sécheresse. Le module non-vie est donc à retenir dans notre étude. Il se divise en trois sous-modules de risques, dont celui du risque catastrophe.

L'*EIOPA*[16] a proposé l'élaboration de scénarios standardisés pour l'estimation de la charge des risques de catastrophe. Les catastrophes naturelles prises en compte sont les tempêtes, inondations, tremblements de terre, et grêle. Mais la sécheresse n'a pas été à l'étude.

Pour des risques non pris en compte dans la formule standard, ou pour apporter une mesure plus précise à des risques spécifiques, l'assureur peut aussi mettre en place un modèle interne.

Pour analyser les catastrophes naturelles au sein de son modèle interne, et ainsi déterminer son SCR consacré aux catastrophes, AXA a recours :

- A ses propres modèles CAT
- Aux modèles développés par les sociétés spécialisés citées en I.B.1
- A une calibration de scénarios de manière plus adaptée que la formule standard

La première de ces trois approches s'adapte particulièrement bien au profil de risque de l'assureur et aux changements de structure de portefeuille. Néanmoins, la validation de ce modèle doit être contrôlée minutieusement par l'ACPR[17], autorité administrative indépendante adossée à la Banque de France chargée de la surveillance et du contrôle des assurances françaises.

[16] *EIOPA : European Insurance and Occupational Pensions Authority*
[17] ACPR : Autorité de contrôle prudentiel et de résolution

C. Synthèse des données sur la sécheresse

1. Les événements majeurs de sécheresse survenus en France

De 1989 à 2014, on peut distinguer cinq principales périodes de sécheresse intense qui ont affecté le marché de l'assurance et de la réassurance :

- 1990 : déficit hydrologique après une période sèche (manque de précipitations pour reconstituer le capital hydrique) entre l'été 1989 et le printemps 1990. Le coût total des dommages pour les assureurs est estimé à 0,4 Md €.
- 1997 : déficit exceptionnel de précipitations. Le coût total des dommages est estimé à 0,4 Md €.
- 2003 : printemps exceptionnellement chaud et sec où les températures atteignent à certains endroits déjà 30 °C fin avril. La canicule estivale qui a suivi a été extrêmement forte. Cette sécheresse se démarque des deux précédentes par l'intensité des températures, plus que par l'intensité de la déshydratation. Le coût total des dommages est estimé à 1,3 Md €.
- 2005 : mois de juin particulièrement chaud touchant particulièrement l'Ouest de la France. Le coût total des dommages est estimé à 0,4 Md €.
- 2011 : déficit hydrologique intense enregistré en 2011 après une période sèche entre l'automne 2010 et le printemps 2011. Le coût total des dommages est estimé à 0,6 Md €.

2. Les chiffres de la CCR

a) Reconnaissance CAT NAT

La carte suivante présente la répartition du nombre d'arrêtés ministériels ayant déclaré l'état de catastrophe naturelle :

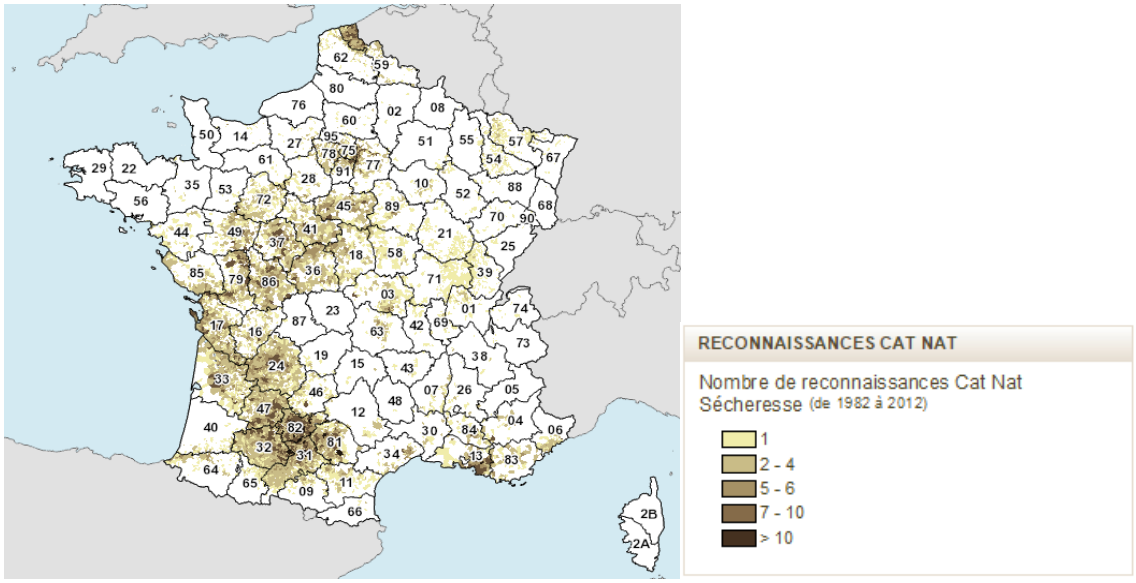


Figure 8 - Reconnaissance CAT NAT – Source : CCR

Sur le plan météorologique et agricole, nous pouvons observer que l'Ouest de la France (sous forme d'arc que nous appellerons « arc de l'Ouest ») est particulièrement sensible au risque de sécheresse, ainsi que la pointe Nord et le département du Var.

b) Fréquence moyenne de sinistres

La carte suivante présente la répartition de la fréquence moyenne de sinistres liés à la sécheresse sur le marché de l'assurance :

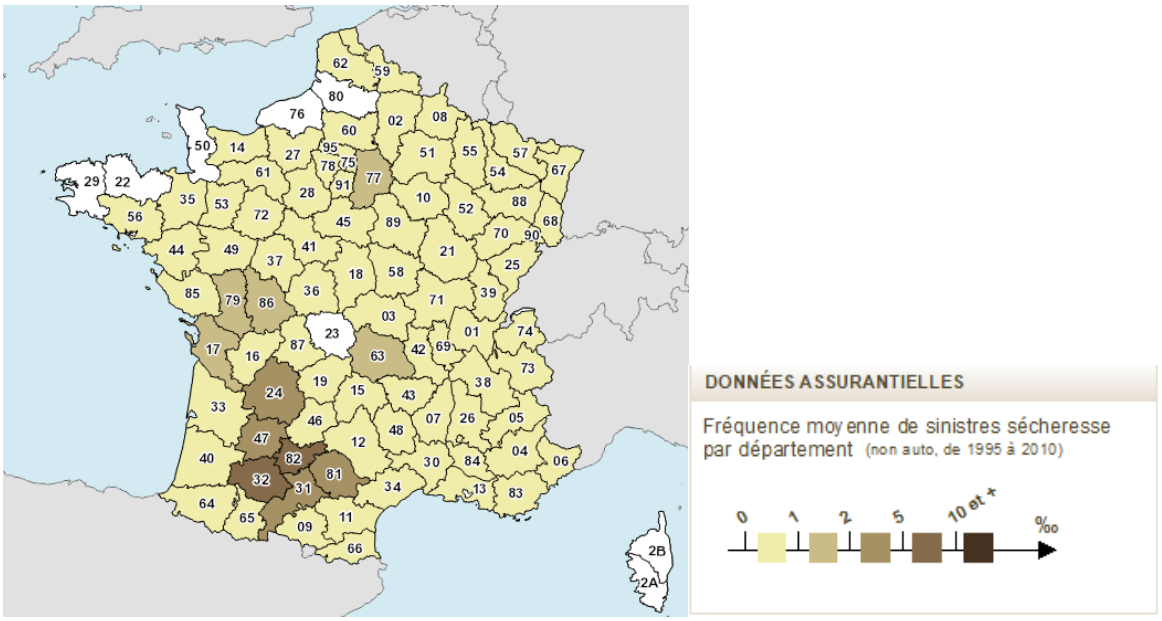


Figure 9 - Fréquence moyenne des sinistres – Source : CCR

Sans surprise, la répartition ressemble à celle du nombre d'arrêtés sécheresse. Cependant, « l'arc de l'Ouest » est moins dessiné, mais les zones les plus exposées restent le Tarn-et-Garonne, le Lot-et-Garonne, la Haute-Garonne, et le Gers.

c) Coûts par départements

La carte suivante présente la répartition des coûts totaux liés à la sécheresse :

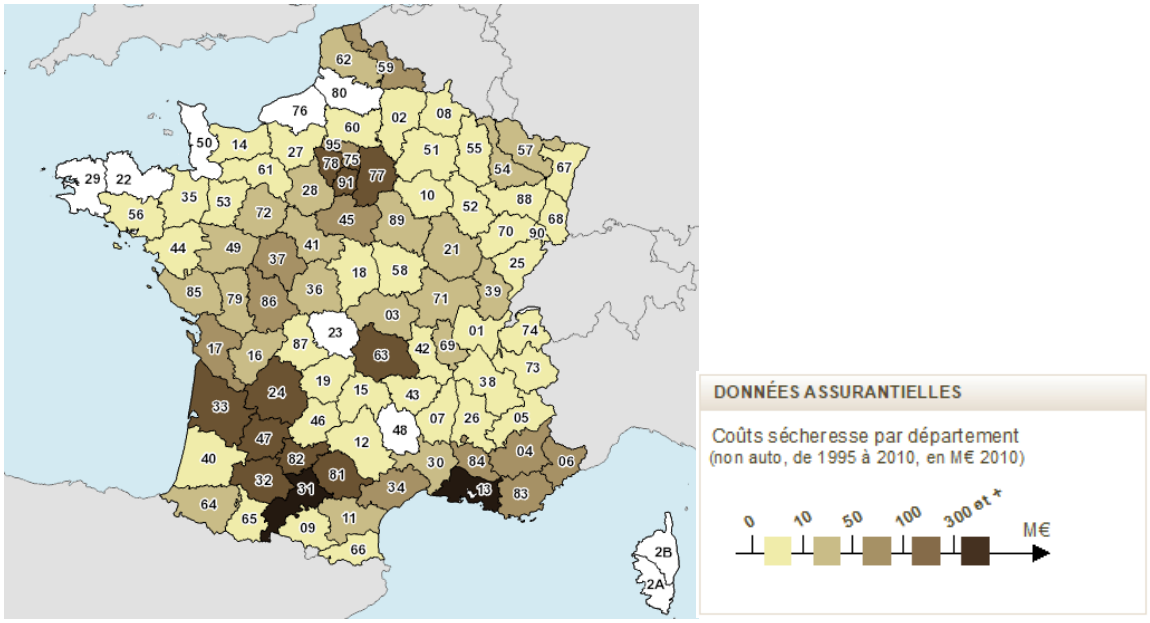


Figure 10 - Coûts par départements – Source : CCR

Après une catastrophe naturelle, on constate souvent que pour une même intensité, l'ampleur des dommages occasionnés peut être très variable.

Nous pouvons supposer que dans les zones où la fréquence de survenance d'événements sécheresse n'est pas élevée, les constructions sont moins préparées à ce type de péril, et donc sont plus vulnérables. La répartition des coûts liés à la sécheresse est alors moins marquée que celle du nombre d'arrêtés sécheresse (l'arc observé auparavant est moins visible, et la répartition est plus homogène).

d) Coût moyen des sinistres

La carte suivante présente la répartition des coûts moyens par sinistre :

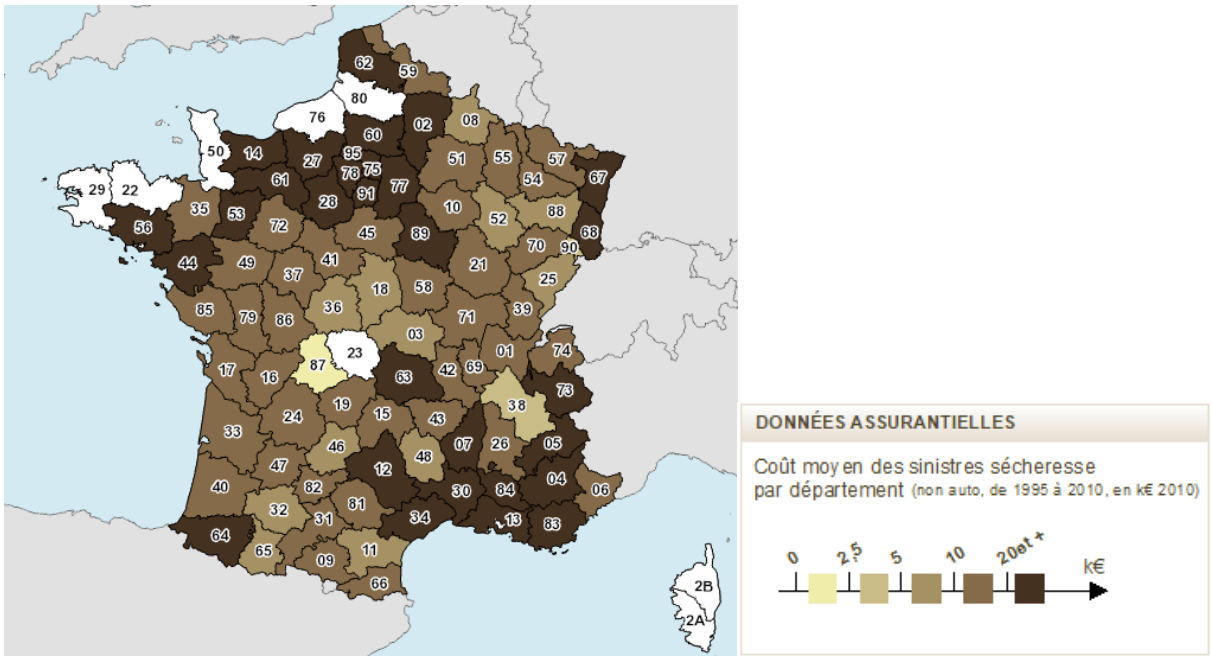


Figure 11 - Coût moyen des sinistres – Source : CCR

Dans la même logique que le point précédent, la vulnérabilité des constructions présentes dans des zones qui ne sont pas très exposées au risque de sécheresse affecte considérablement le coût moyen de chaque sinistre.

La différence est ici encore plus flagrante, car on étudie les coûts moyen et non pas les coûts totaux : une petite structure (ex : une maison) sera plus affectée qu'une grosse (ex : un immeuble).

3. Présentation des données utilisées pour la modélisation

a) La sinistralité historique d'AXA

Pour nous permettre de modéliser l'impact de la sécheresse sur le portefeuille d'AXA, nous avons à disposition son historique de sinistres liés à la sécheresse.

Chaque sinistre est renseigné par :

- Date de survenance : entre 1989 et 2014
- Lieu de survenance : recensé par un code INSEE
- Segment : MRH[18], Immeuble, Agricole, Risques industriels, multirisques professionnels
- Charge totale du sinistre (en €), vue à avril 2015
- Valeur de l'objet assuré (en €), vue à avril 2015

Il est fondamental que le modèle obtenu soit construit à partir de « bonnes » données, c'est-à-dire que la base de données des sinistres (*reporting* en anglais) doit **refléter** correctement l'historique des événements sécheresse.

La sécheresse étant un risque peu connue et pouvant s'étendre sur plusieurs jours contrairement aux autres catastrophes naturelles, l'historique à notre disposition présente des défauts :

- Beaucoup de sinistres sont enregistrés en fin d'exercice le 31/12
- Pour une période de forte sinistralité donnée, les sinistres sont souvent enregistrés seulement en début ou fin de mois.

Afin de modéliser la sécheresse sur une période allant de juin à septembre, nous devons être capables de contourner ces défauts.

Pour contourner le premier point, nous effectuerons une régression entre les pertes enregistrées sur l'année entière et les pertes enregistrées sur la période étudiée.

Pour contourner le deuxième point, nous agrègerons nos données de manière mensuelle, afin d'avoir plus de visibilité dans le croisement de nos données. La méthode d'agrégation sera décrite dans la suite.

b) Les précipitations et les températures maximales journalières

Un manque de précipitations et des températures élevées sur une période prolongée expliquent essentiellement une période de sécheresse. Nous allons donc étudier l'historique de ces variables pour les mettre en parallèle avec l'historique des sinistres liés à la sécheresse.

Pour cela, nous avons utilisé la base de données de *l'European Climate Assessment & Dataset* (*ECA&D*).

Elle contient les valeurs historiques des précipitations et des températures maximales journalières :

- Entre le 01/01/1950 et le 31/12/2014, correspondant à 23 741 jours ;
- En des points de l'Europe équidistants de 0,5°, ce qui correspond en France à environ 50 km.

[18] MRH : multirisques habitation

Après s'être restreint à la France, nous obtenons les relevés de précipitations et de températures en 251 points. Les données ont été estimées par extrapolation à partir des relevés de 106 stations météorologiques en France. Pour simplifier, nous supposerons que ces points représentent des stations météorologiques.

Pour les précipitations et pour les températures, nous disposons alors de 251 séries temporelles. Chaque série contient 23 741 données représentant l'évolution des précipitations ou des températures maximales journalières enregistrées en un point donné.

c) L'aléa retrait-gonflement des argiles

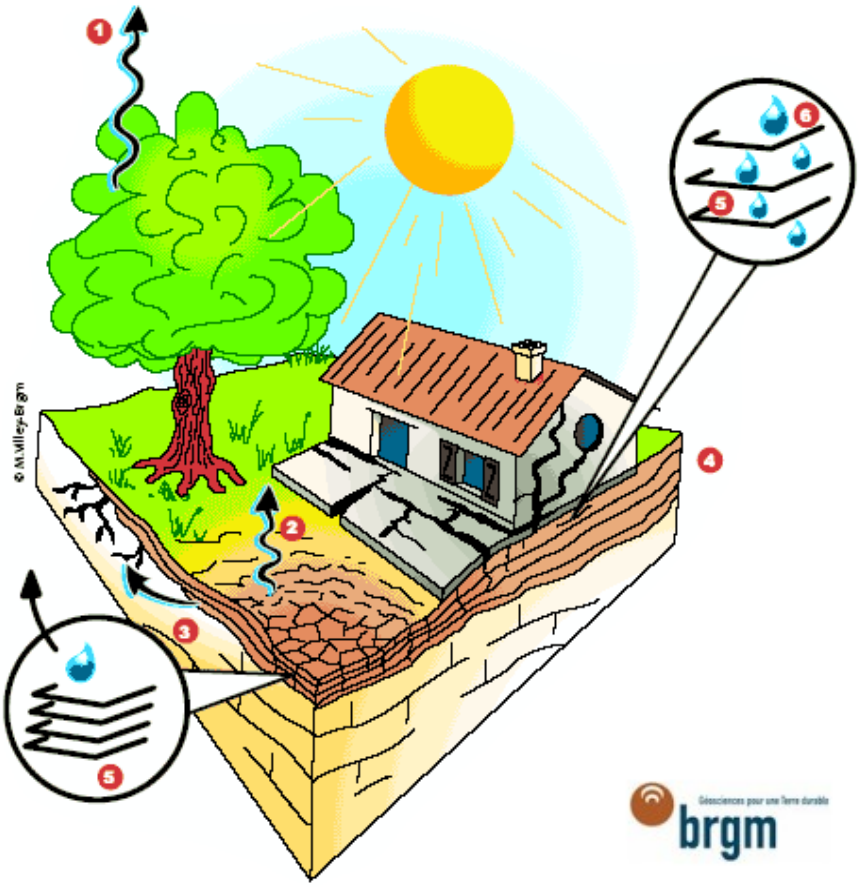
Le phénomène de retrait-gonflement des argiles, intimement lié aux périodes de sécheresse, génère chaque année de nombreux dégâts aux bâtiments, affectant ainsi le portefeuille d'AXA.

En effet, la consistance des matériaux argileux se modifie en fonction de leur teneur en eau et peut engendrer des variations de volume : on parle alors de retrait-gonflement des argiles.
En climat tempéré, ce qui est le cas en France, les argiles sont proches de leur limite de gonflement en eau. Cela implique qu'elles sont éloignées de leur limite de retrait.
Comme vu précédemment, la sécheresse (agricole) est caractérisée par un déficit de la présence d'eau dans les sols. Les plus forts mouvements de retrait d'argiles se produisent donc pendant les périodes de sécheresse.

Il y a alors un tassement du terrain argileux, et une ouverture de fissures, provoquant des dégâts importants aux bâtiments (fissuration, décollement, …)

Les minéraux argileux présentent une structure en feuillets, à la surface desquels les molécules d'eau peuvent s'absorber (gonflement des argiles), ou se dissiper (retrait des argiles). Mais certaines familles de minéraux argileux possèdent des liaisons particulièrement lâches entre feuillets. Cela augmente la capacité de retrait des argiles, impliquant alors de fortes variations de volume du matériau argileux.

Le schéma suivant illustre ce phénomène de rétractation de l'argile qui engendre des dégâts aux bâtiments.



Légende du dessin :
(1) Evapotranspiration
(2) Evaporation
(3) Absorption par les racines
(4) Couches argileuses
(5) Feuillets argileux
(6) Eau interstitielle

Figure 12 - Phénomène de retrait-gonflement des argiles – Source : BRGM

Il est alors intéressant de détecter les zones où ce phénomène de retrait-gonflement des argiles est important. Le site internet Georisques, géré par le Ministère de l'Écologie, du Développement durable et de l'Énergie, nous a permis de récolter des données, qui attribuent une probabilité (appelé « aléa ») que ce phénomène survienne de manière importante sur un secteur géographique donné. Nous avons ensuite agrégé les données en donnant la moyenne d'aléa par zone INSEE.
Il sera alors intéressant de comparer la répartition de cet « aléa » avec celle du nombre de sinistres liés à la sécheresse.

D. Etapes de modélisation du risque sécheresse

Un modèle est une représentation simplifiée de la réalité. A partir de toutes les données dont on dispose, on va extraire un maximum d'informations permettant de représenter au mieux la répartition de la sinistralité liée à la sécheresse.

Pour cela, nous allons diviser notre travail en trois modules répondant à différentes problématiques :

- Module Aléa (ou physique) : l'objectif est de construire un catalogue d'événements contenant une multitude de scénarios météorologiques réalistes et probabilisés d'évolution mensuelle des variables explicatives de la sécheresse.
- Module Vulnérabilité (ou de sinistralité) : l'objectif est de modéliser la fréquence de survenance de l'événement sécheresse et les dommages subis par les objets assurés en fonction de l'intensité de l'événement, et des caractéristiques de l'objet assuré (localisation, nature et valeur).
- Module Financier : l'objectif est de modéliser les pertes finales pour l'assureur, nettes des conditions contractuelles.

Afin de modéliser la sinistralité due à la sécheresse et son impact sur le portefeuille d'AXA, nous ferons la liaison entre ces différents modules : nous aurons à disposition des milliers de scénarios simulés grâce au module Aléa et, pour chaque scénario, nous mesurerons l'impact sur les modules de Vulnérabilité et Financier.

1. Module Aléa

L'exposition au phénomène de sécheresse, d'un point de vue purement physique, dépend de plusieurs paramètres :

- Répartition géographique
- Fréquence de survenance
- Intensité des événements

Afin de modéliser ces paramètres, nous allons mettre en relation les données dont nous disposons : l'historique des précipitations, l'historique des températures, l'historique des événements sécheresse, ainsi que les connaissances scientifiques en lien avec ces événements. Plus la période d'étude des historiques est longue, plus nos données seront riches en information, et plus notre modèle captera de façon suffisante la diversité des caractéristiques de la sécheresse.

Nous allons définir un ensemble de variables qui caractérise la sécheresse : les variables explicatives. Pour chaque zone géographique préalablement définie, on va chercher à simuler plusieurs milliers de scénarios réalistes d'évolution mensuelle des variables explicatives de la sécheresse afin d'obtenir le plus de trajectoires possibles et pour prendre en compte des comportements extrêmes.

Afin d'étudier la performance de ces simulations, nous pourrons effectuer des « *backtests* » : nous baserons notre modélisation sur une partie restreinte des historiques, pour comparer la répartition des trajectoires obtenues avec les données réelles.

Le module Aléa nous permet alors d'extrapoler des événements aussi probables que les événements historiques observés, mais surtout nous fournit une base de données de scénarios météorologiques réalistes et probabilisés générant des événements sécheresse.

2. Module Vulnérabilité

Pour une zone géographique donnée, le module Aléa nous a ainsi fournit une base de données contenant des milliers de scénarios physiques réalistes et probabilisés d'évolution mensuelle des variables explicatives de la sécheresse.

Dans ce mémoire, la fréquence d'occurrence et l'intensité de la sécheresse seront synthétisées en une seule variable : le nombre de sinistres[19]. Pour chaque scénario, le module Vulnérabilité fait le lien entre les variables explicatives de la sécheresse et le nombre de sinistres qui en découle. Cela permet ensuite d'estimer la sinistralité des objets assurés en fonction de l'intensité de l'événement, et des caractéristiques de l'objet assuré (nature et valeur).

Pour une même intensité, l'ampleur des dommages peut être très diverse. En effet, les dégâts sur les bâtiments peuvent varier selon le type de construction. La sinistralité dépend donc de l'intensité de l'événement, mais aussi des spécificités de l'objet assuré.

Malheureusement, les données de sinistralité nous informent rarement sur les caractéristiques des biens assurés. Etant donné la multitude de type d'objets et par manque de données détaillées, les objets assurés seront regroupés par catégorie. Nous reprendrons les catégories citées précédemment en regroupant les contrats agricoles avec les contrats multirisques professionnels : MRH, Immeuble, Agricole et Risques industriels. Pour chaque catégorie, on étudiera la répartition du taux de destruction[20] en fonction de l'intensité d'un événement. Finalement, nous obtiendrons pour chaque catégorie d'objet assuré une courbe de vulnérabilité appliquée au risque de sécheresse.

Le graphique ci-dessous représente une courbe de vulnérabilité pour un type d'objet assuré.

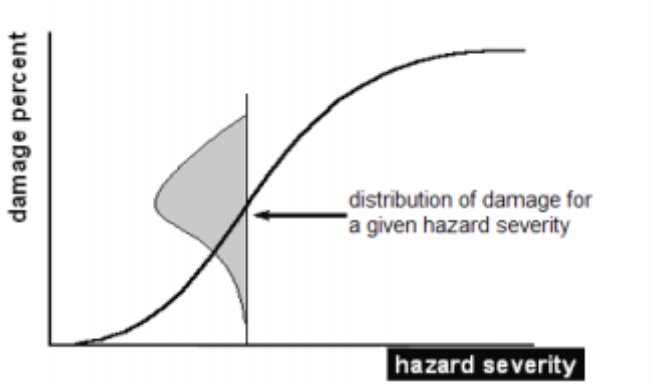


Figure 13 - Exemple d'une courbe de vulnérabilité pour un type d'objet assuré – Source : EQUECAT

Le taux moyen de destruction est bien croissant en fonction de l'intensité. Chaque type d'objet assuré donne une courbe de vulnérabilité particulière.

3. Module Financier

Par zone géographique, le module Aléa permet de générer des scénarios météorologiques réalistes et probabilisés de variables expliquant la sécheresse. Pour chaque scénario, le module Vulnérabilité permet de traduire l'intensité des variables expliquant la sécheresse en nombre de sinistres, puis en taux de sinistralité

[19] En effet, un nombre strictement positif indique qu'il y a au moins un sinistre, et sa valeur mesure l'intensité.
[20] Taux de destruction : montant du sinistre par rapport à la valeur totale de l'objet concerné.

pour chaque type d'objets assurés. Le module Financier détermine alors la perte financière de l'assureur nette des caractéristiques de chaque contrat associé aux biens assurés.

Les différentes caractéristiques sont :

- La franchise et le plafond : l'assureur doit payer la part du montant des sinistres compris entre la franchise et le plafond. Le reste est à la charge de l'assuré.
- La part des coassureurs : la coassurance est le partage d'un même risque entre plusieurs sociétés d'assurance, chacune étant garante de la seule partie qu'elle a acceptée de prendre en charge. Ainsi, chaque assureur devra payer le coût du ou des sinistres en fonction du pourcentage correspondant à son niveau d'engagement dans la couverture du risque.
- La part des réassureurs : dans la plupart des cas, lorsqu'une catastrophe naturelle survient, l'assureur ne peut pas régler l'intégralité du sinistre. La réassurance permet de se couvrir contre ces pics de sinistralité. Comme expliqué précédemment, il existe plusieurs types de traités de réassurance. Le plus souvent, les traités non-proportionnels sont souscrits. Le réassureur prend en charge la partie de tout sinistre qui excède la priorité du traité et dans la limite de la portée du traité. Cela permet de maximiser la perte de l'assureur.

Ces caractéristiques permettent de diminuer de manière significative la perte financière produit par le module Vulnérabilité.

4. Résultats du modèle

La combinaison des modules Aléa, Vulnérabilité et Financier permet de répondre au problème initial : modéliser la sinistralité probable de l'année à venir (*expected annual loss* en anglais), afin d'optimiser les conditions de réassurance et de calculer le capital réglementaire.

Les trois modules construits précédemment influencent tous le résultat de la modélisation de la sinistralité : le résultat final ne peut pas être plus performant que le module le moins performant.

Les résultats des modules sont souvent synthétisés sous forme de courbe représentant la distribution des pertes. Cette distribution est exprimée en fonction d'une période retour, qui correspond au temps statistique entre deux occurrences de même intensité. Par exemple, un événement dont la période de retour est de 100 ans se produira en moyenne une fois tous les cent ans. Ainsi, cet événement aura une chance sur cent de se produire sur une année.

Pour représenter cette distribution des pertes, on distingue deux types de courbe :

- La courbe *AEP* (*Annual Exceedance Probability* en anglais) : associe une période de retour au coût total des événements sur une année.
  Cette courbe permet donc de déterminer le capital réglementaire requis sous Solvabilité II, correspondant au montant associé à la période de retour de 200 ans.
- La courbe *OEP* (*Occurence Exceedance Probability* en anglais) : associe une période de retour au coût maximal d'un événement sur une année.
  Cette courbe aide donc à optimiser la structuration des traités de réassurance, en quantifiant la distribution du coût maximal annuel d'un événement (pour une période de retour donnée).

Le graphique suivant donne un exemple de courbe AEP et OEP :

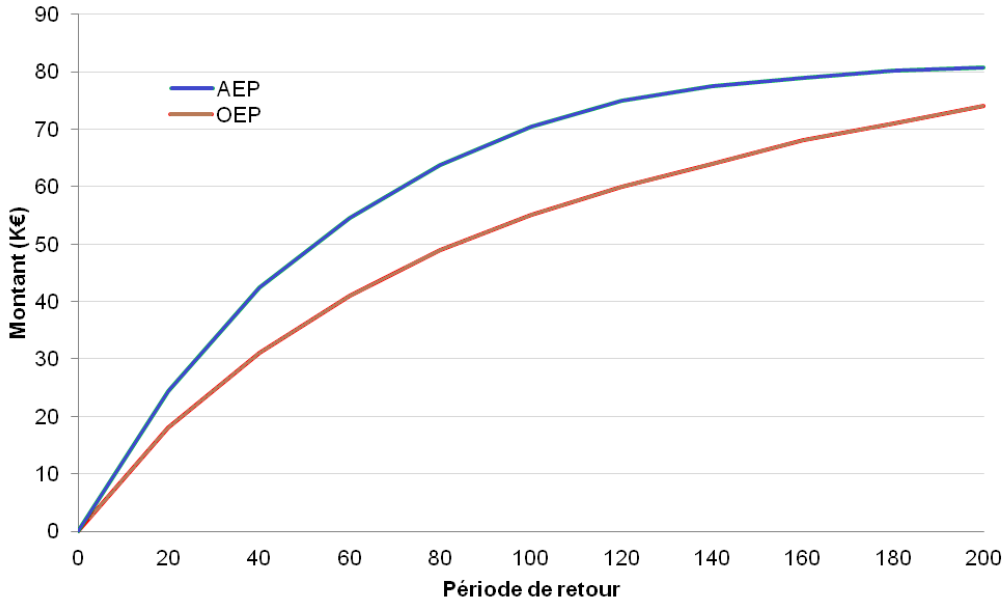


Figure 14 - Courbes *AEP* et *OEP*

II. Module Aléa : simulations des variables expliquant la sécheresse

Nous savons que la sécheresse est essentiellement liée aux précipitations et aux températures maximales.

Dans un premier temps, nous allons chercher un ensemble de variables explicatives de la sécheresse construit à partir des données de précipitations et de températures.

Nous modéliserons ensuite les précipitations et les températures maximales afin de pouvoir simuler 10 000 scénarios d'évolution mensuelle de ces variables.

Enfin, nous pourrons en déduire 10 000 simulations de l'ensemble des variables explicatives de la sécheresse.

A. Les variables explicatives de l'événement sécheresse

Dans cette partie, nous allons chercher à définir un ensemble de variables explicatives de la sécheresse. Pour ce faire, nous allons croiser le nombre de sinistres mensuels de chaque *CRESTA* avec les valeurs prises par un certain nombre de variables.

Les graphiques en annexe C tendent à montrer qu'il est plus pertinent de croiser la sinistralité avec les valeurs des variables explicatives des mois précédents, et non avec le mois courant.

1. Les précipitations mensuelles

Le manque de précipitations sur une période prolongée joue un rôle majeur dans l'apparition de la sécheresse.

Le graphique suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec les précipitations cumulées observées le mois dernier ($SumPrecip.Mois1$).

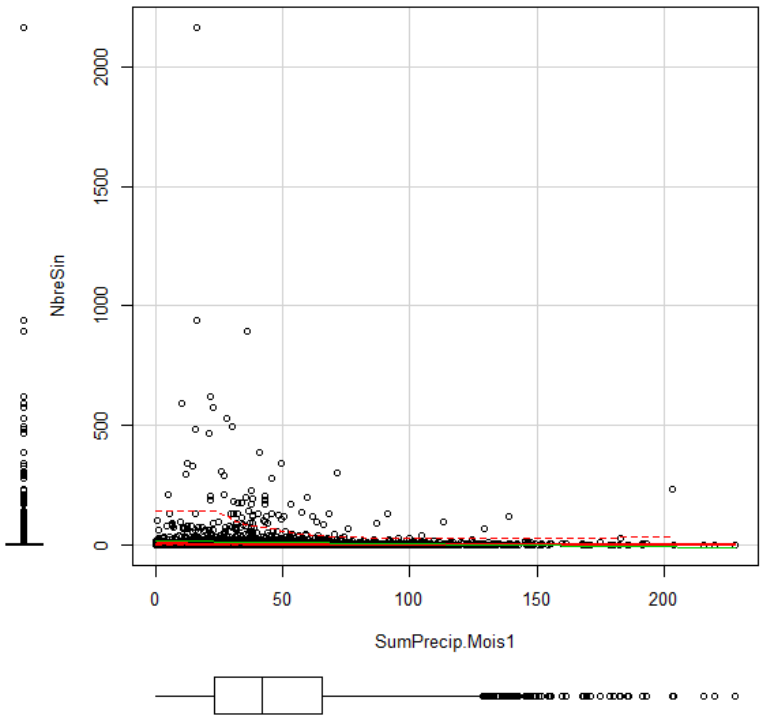


Figure 15 - Sinistralité et précipitations mensuelles

Sur chaque axe est insérée une boîte à moustache afin de représenter la répartition de la variable donnée. Une boîte à moustaches est un outil graphique qui résume les principales caractéristiques de la répartition d'une série statistique : médiane, quartiles, déciles, minimum et maximum.

De plus, différentes régressions ont été effectuées :

- Une régression linéaire, afin de détecter une tendance générale (en vert)
- Des régressions non-paramétriques lissées, dont la méthode sera décrite plus tard, pour détecter des tendances locales (en rouge)

Nous remarquons que les gros sinistres (supérieurs à 100) sont liés à des valeurs faibles de $SumPrecip.Mois1$.

De plus, nous observons que lorsque $SumPrecip.Mois1$ est plus faible que d'habitude (pour des valeurs inférieures à la médiane, c'est-à-dire autour de 45 mm), une tendance à la hausse se dégage nettement et beaucoup plus de sinistres sont observés.

Ensuite, il est intéressant de mettre en lien la sinistralité avec ce qu'il a pu se passer les deux derniers mois.

Le graphique 3D suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec les précipitations cumulées observées les deux derniers mois. Pour plus de visibilité, seuls les gros sinistres (supérieurs à 100) sont représentés.

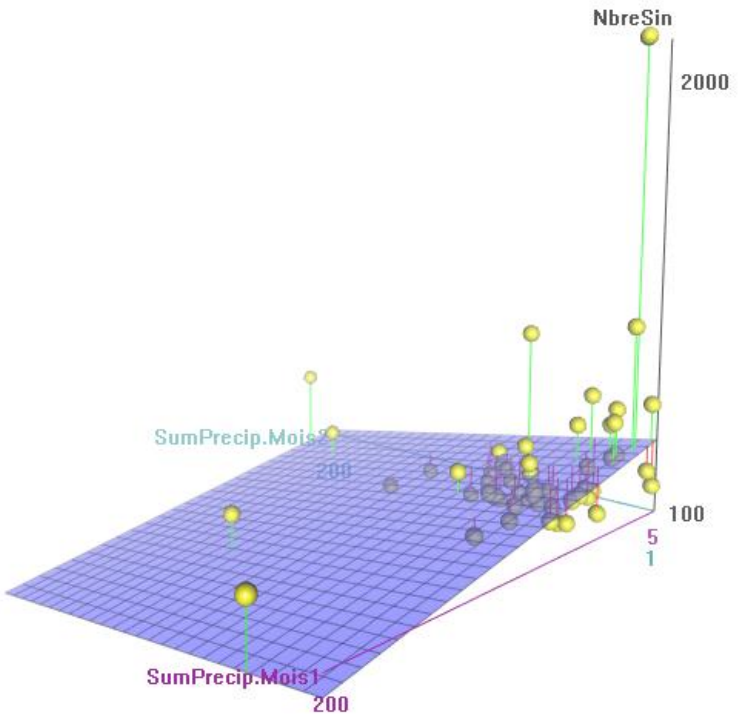


Figure 16 - Sinistralité et précipitations mensuelles (2)

Nous observons que la majorité des gros sinistres sont liés à des périodes prolongées (deux mois successifs) de déficit pluviométriques. Cela confirme que l'apparition de la sécheresse est en partie expliquée par un manque prolongé de précipitations.

Nous allons donc modéliser les précipitations mensuelles dans chacune des zones *CRESTA* de France. Nous nous apercevrons qu'une loi Gamma peut suffire à simuler les précipitations mensuelles partout en France.

2. L'indice *SPI* (*Standardized Precipitation Index)*

Plusieurs indicateurs de sécheresse ont été développés depuis les années 80. Parmi eux, l'un des plus utilisés est l'indice *SPI* (*Standardized Precipitation Index)*. Il a été développé par McKee *et al.* en 1993.

L'indice *SPI* est fondé sur la probabilité de précipitations (cumulées sur une période donnée) estimée par une loi Gamma. La probabilité des précipitations observées est transformée en un indice normalisé. Cela permet de quantifier l'écart des précipitations d'une période, déficit ou surplus, par rapport aux précipitations moyennes historiques de la période.

$$SPI = -\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) \quad pour\ 0 < H(x) < 0{,}5$$

$$SPI = +\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3}\right) \quad pour\ 0{,}5 < H(x) < 1$$

- x : précipitations cumulées sur une période donnée (nous étudierons les périodes mensuelles)
- $H : x \rightarrow q + (1-q)G(x)$

où G est la fonction de répartition d'une loi Gamma paramétrée par maximum de vraisemblance sur les données de précipitations mensuelles, et q la probabilité d'avoir $x = 0$ (estimée par le rapport historique du nombre de mois sans précipitations et le nombre de mois étudié)

- $t = \sqrt{\log\left(\frac{1}{(H(x))^2}\right)} \quad pour\ 0 < H(x) < 0{,}5$
- $t = \sqrt{\log\left(\frac{1}{(1-H(x))^2}\right)} \quad pour\ 0{,}5 < H(x) < 1$
- $c_0 = 2{,}515517 \qquad c_1 = 0{,}802853 \qquad c_2 = 0{,}010328$
- $d_1 = 1{,}43278 \qquad d_2 = 0{,}189269 \qquad d_3 = 0{,}001308$

Pour rappel, la fonction de répartition d'une loi Gamma (servant à modéliser les précipitations) est :

$$G(x;\alpha,\beta) = \int_0^x t^{\alpha-1}\frac{\beta^\alpha e^{-\beta t}}{\Gamma(\alpha)}dt \qquad avec\ \Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt$$

En théorie, le sol est asséché lorsque le *SPI* est négatif, et est humide lorsque le *SPI* est positif. Plus on s'éloigne de 0, plus l'intensité est élevée.

Le graphique suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec la valeur de l'indice *SPI* du mois dernier ($SPI.Mois1$).

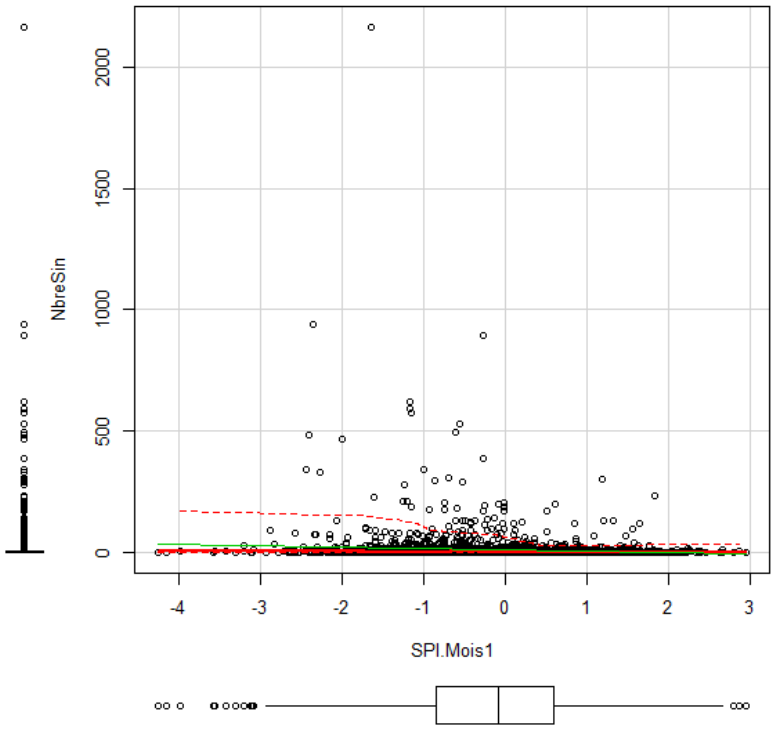


Figure 17 - Sinistralité et *SPI*

Une tendance haussière est observée lorsque le *SPI* est négatif. Cela confirme l'intérêt de cet indicateur dans l'apparition de sécheresse.

Les indices de sécheresse basés uniquement sur les précipitations supposent deux hypothèses :

- La variabilité des précipitations est beaucoup plus élevée que celle des autres variables, telles que la température et l'évapotranspiration.
- Les autres variables sont stationnaires (i.e. aucune tendance temporelle).

Cela suppose que l'importance de ces autres variables est négligeable, et que les périodes de sécheresse sont caractérisées entièrement par la variabilité temporelle des précipitations.

Cependant, certains auteurs ont remis en question le fait de négliger systématiquement l'évolution de la température sur les conditions de sécheresse. Des études empiriques ont montré que des températures élevées sur une période prolongée affectent nettement la gravité des sécheresses.

Le rôle important des températures sur la gravité de la sécheresse était évident en 2003 : des températures très élevées ont augmenté considérablement le phénomène d'évapotranspiration accentuant la gravité de la sécheresse.

Il est donc nécessaire de s'intéresser à la modélisation des températures maximales quotidiennes et à celle de l'évapotranspiration.

3. Les températures maximales journalières

La sécheresse est intimement liée à des périodes où les températures sont restées supérieures à un seuil pendant suffisamment longtemps.

Le graphique suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec la température maximale enregistrée au mois précédent ($Tmax.Mois1$).

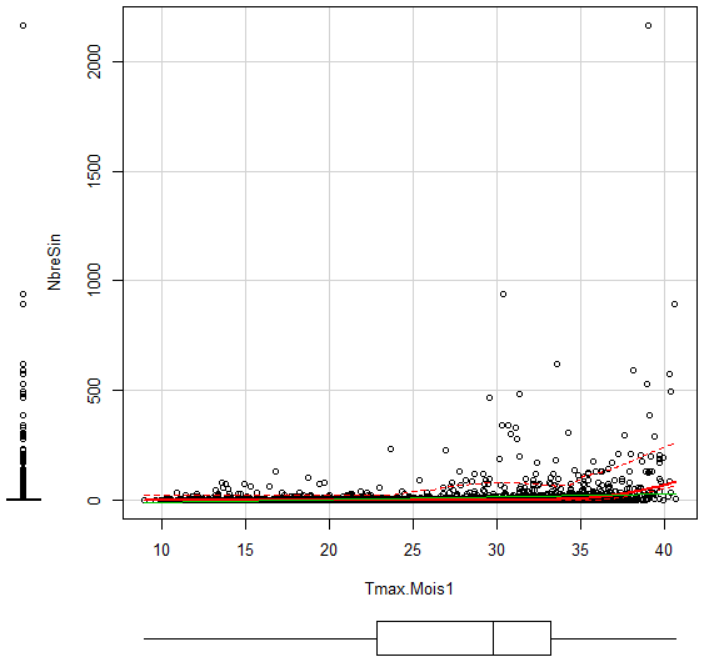


Figure 18 - Sinistralité et température maximale mensuelle

Nous remarquons que les gros sinistres (supérieurs à 100) sont liés à des valeurs élevées de $Tmax.Mois1$.

De plus, nous observons que lorsque $Tmax.Mois1$ est beaucoup plus élevé que d'habitude (pour des valeurs supérieures au dernier quartile, c'est-à-dire autour de 35°C), une tendance à la hausse se dégage nettement et beaucoup plus de sinistres sont observés.

Ensuite, comme dans le cas des précipitations, il est intéressant de mettre en lien la sinistralité avec ce qu'il a pu se passer les deux derniers mois.

Le graphique 3D suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec les températures maximales observées les deux derniers mois. Pour plus de visibilité, seuls les gros sinistres (supérieurs à 100) sont représentés.

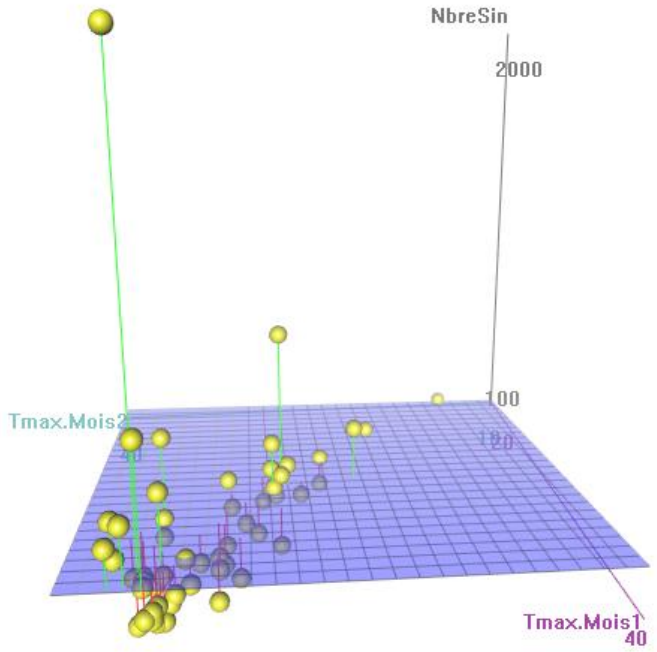


Figure 19 - Sinistralité et température maximale mensuelle (2)

Nous observons que la majorité des gros sinistres sont liés à des périodes prolongées (deux mois successifs) de températures élevées.

Une forte corrélation est donc observée entre le nombre de sinistres et les températures élevées sur une période prolongée. Il semble alors nécessaire de modéliser les températures.

Les principales étapes de la modélisation des températures maximales quotidiennes sont :

- La segmentation de la France en régions homogènes en termes de température. L'objectif étant de simuler plusieurs milliers de scénarios d'évolution de la température pour une zone géographique donnée, il est plus intéressant de regrouper les zones présentant des similitudes dans l'évolution de la température plutôt que d'effectuer ces simulations sur les 251 stations météorologiques.
- L'ajustement d'un certain type de processus pour représenter les séries temporelles des températures pour chaque région définie à la première étape. Cela permet de détecter le plus finement possible l'ensemble des dynamiques stochastiques contenues dans les données.
- La modélisation de la dépendance entre régions. En effet, si une vague de chaleur apparaît dans une région donnée, les autres régions seront aussi, dans une certaine mesure, exposées à ces conditions

exceptionnelles. Il est donc nécessaire de prendre en compte cette dépendance pour avoir des scénarios réalistes.

- La validation du modèle.

Cela permettra de générer une liste de 10 000 scénarios d'évolution de la température maximale journalière, contenant des configurations jamais connues auparavant (plusieurs jours où la température a pu atteindre 45°C).

4. L'évapotranspiration

Des températures élevées sur une période prolongée diminuent la quantité d'eau présente dans les sols par le phénomène d'évapotranspiration, accentuant l'assèchement des sols.

Plusieurs formules peuvent être utilisées afin d'estimer l'évapotranspiration mensuelle, seulement à partir des données de températures et latitudes.

Celle développée par le géographe et climatologue C.W. Thornthwaite a l'avantage d'être simple et robuste pour différentes latitudes données :

$$E_m = 16\left(\frac{10T_m}{I_m}\right)^{a_m} F_{m,\lambda}$$

- E_m : évapotranspiration du mois m (en mm)
- T_m : moyenne des températures au mois m (en °C)
- I_m : somme des 12 indices thermiques mensuels

$$I_m = \sum_{m=1}^{12}\left(\frac{T_m}{5}\right)^{1,514}$$

- $a_m = 6{,}75.10^{-7}.{I_m}^3 - 7{,}71.10^{-5}.{I_m}^2 + 1{,}79.10^{-2}.I_m + 0{,}49$
- $F_{m,\lambda}$: coefficient de correction dépendant de la latitude λ et du mois m concernés, et est donné par une table (voir annexe A)

L'ensemble des scénarios d'évolution de la température maximale journalière pour l'année à venir permettra d'en déduire une estimation de l'évapotranspiration.

Le graphique suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec l'évapotranspiration estimée au mois précédent ($Evapo.Mois1$).

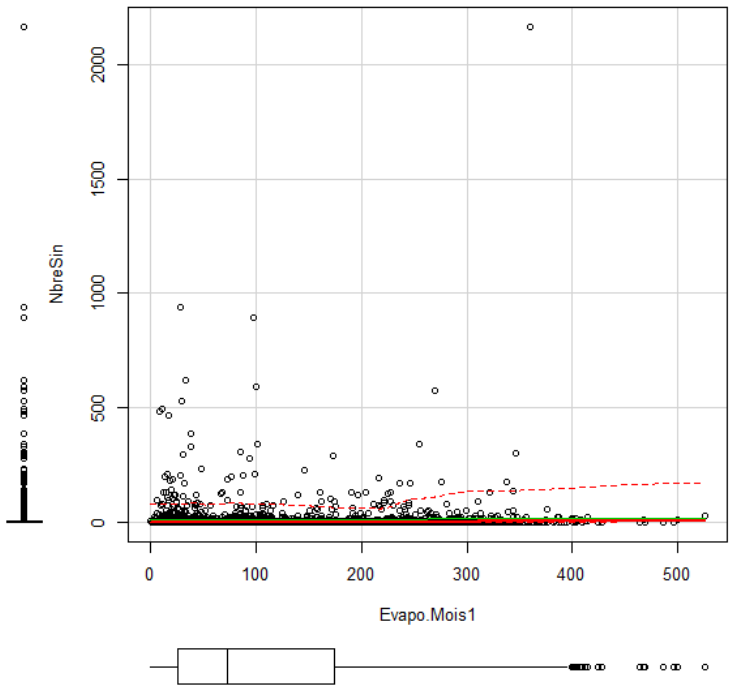


Figure 20 - Sinistralité et évapotranspiration

Comme dans le cas des températures, on remarque que lorsque les valeurs d'évapotranspiration sont très élevées (supérieures à 200 mm), certains gros sinistres sont détectés.

Cependant, beaucoup de gros sinistres sont liés à de faibles valeurs d'évapotranspiration, correspondant à des périodes présentant uniquement un déficit pluviométrique. Nous n'utiliserons pas directement l'évapotranspiration comme variable explicative, mais elle servira à calculer l'indice de sécheresse couramment utilisé : l'indice *SPEI*.

5. L'indice *SPEI* (*Standardized Precipitation Evapotranspiration Index)*

L'indice *SPI* se calcule uniquement à partir des données de précipitations. Il explique la sécheresse en fonction de la quantité d'eau qui tombe dans les sols. Il ne peut pas identifier le rôle de l'augmentation de la température dans les futures conditions de sécheresse.

Or, nous avons à disposition des données de températures permettant d'estimer l'évapotranspiration, c'est-à-dire la quantité d'eau qui s'évapore des sols. A partir de ces données, nous pouvons donc estimer les précipitations nettes représentant la quantité d'eau restante dans les sols.

$$\text{précipitations nettes} = \text{précipitations} - \text{évapotranspiration}$$

L'indice *SPEI* (*Standardized Precipitation Evapotranspiration Index*) a été développé pour tenir compte des effets possibles des températures extrêmes sur l'accentuation de l'évapotranspiration et donc sur l'assèchement des sols.

Il est construit de la même manière que le *SPI*. Les seules différences se trouvent dans les données utilisées et dans le choix de la loi pour modéliser ces données :

- x : précipitations nettes sur une période donnée (nous étudierons les périodes mensuelles)
- $H : x \rightarrow q + (1-q)G(x)$

  où G est la fonction de répartition d'une loi log-logistique paramétrée par maximum de vraisemblance sur les données de précipitations mensuelles, et q la probabilité d'avoir $x = 0$ (estimée par le rapport historique du nombre de mois sans précipitations et le nombre de mois étudié)

Pour rappel, la fonction de répartition d'une loi log-logistique (servant à modéliser les précipitations nettes) est :

$$G(x;\alpha,\beta) = \frac{1}{1+(x/\alpha)^{-\beta}} = \frac{(x/\alpha)^{-\beta}}{1+(x/\alpha)^{-\beta}} = \frac{x^{\beta}}{\alpha^{\beta}+x^{\beta}} \qquad o\grave{u}\ x>0, \alpha>0, \beta>0$$

Comme pour le *SPI*, un *SPEI* négatif indique une période sèche, et un *SPEI* positif indique une période humide. Plus on s'éloigne de 0, plus l'intensité est élevée.

Le graphique suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec la valeur de l'indice *SPEI* du mois précédent ($SPEI.Mois1$).

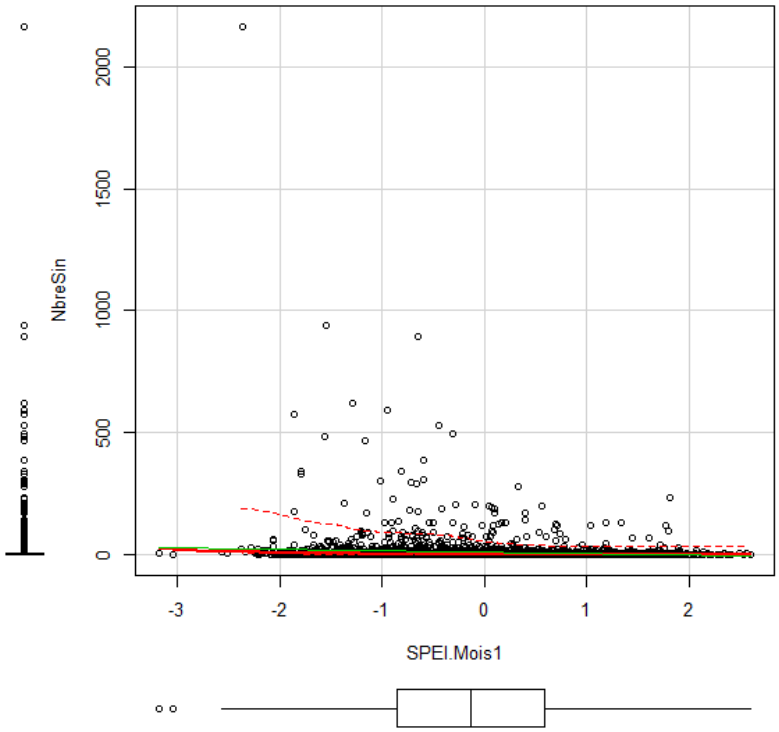


Figure 21 - Sinistralité et *SPEI*

Cet indicateur semble prend en compte plus de configurations que le *SPI*. La tendance haussière, lorsque le *SPEI* devient négatif, est davantage affirmée à force qu'on s'éloigne de 0.

C'est donc un indicateur précieux de sinistralité liée à la sécheresse.

6. L'aléa retrait-gonflement des argiles

Comme vu en première partie, il est intéressant de comparer la répartition du phénomène de retrait-gonflement des argiles (mesuré par l' « aléa ») avec celle du nombre de reconnaissances sécheresse et avec celle du nombre de sinistres liés à la sécheresse dans le portefeuille d'AXA.

Les cartes suivantes permettent de comparer ces répartitions :

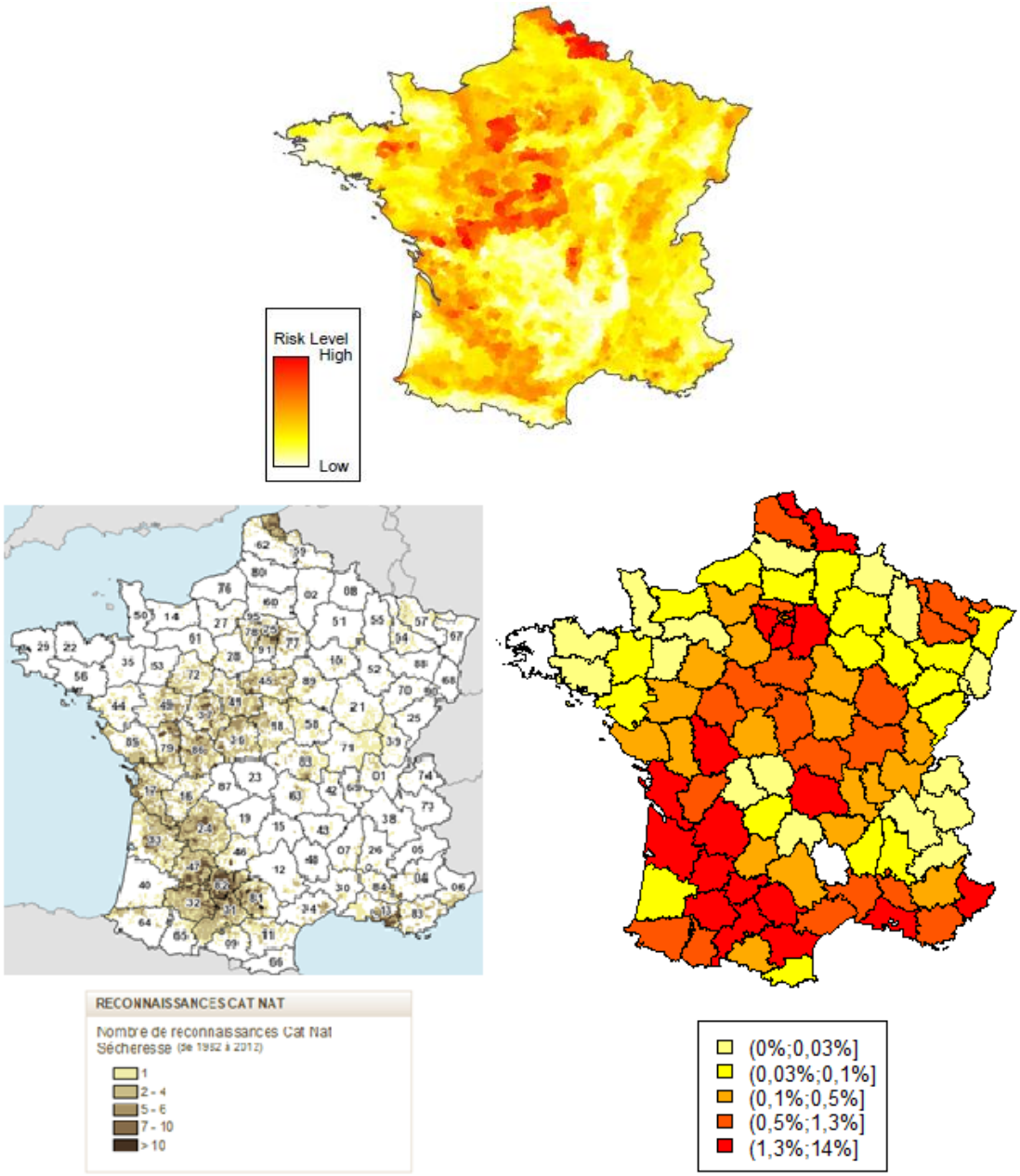


Figure 22 - Comparaison de la répartition de l'aléa retrait-gonflement des argiles avec la répartition du nombre de sinistres liés à la sécheresse

Sur la carte en bas à droite, nous avons représenté la répartition de la sinistralité d'AXA, en donnant le pourcentage de la totalité des sinistres présents dans un *CRESTA* donné.

Nous observons que les répartitions sont liées, en particulier dans le Pays de la Loire, le Centre et le Poitou Charente. Une forte présence du phénomène de retrait-gonflement des argiles est aussi présente à la

pointe Nord de la France. Cependant, le Sud-Ouest de la France n'est pas autant exposé à ce phénomène qu'au risque de sécheresse, mais nous pouvons tout de même reconnaître « l'arc de l'Ouest ».

Cet aléa permet de donner un poids plus conséquent aux zones géographiques présentant une forte exposition au phénomène de retrait-gonflement des argiles, du fait de sa corrélation avec le risque de sécheresse.

B. Modélisation des précipitations mensuelles

Dans la partie précédente, nous avons décrit certains indicateurs qui se basaient sur les précipitations mensuelles.

Afin de pouvoir générer des scénarios d'évolution mensuelle des variables explicatives de la sécheresse pour l'année à venir, nous allons devoir modéliser la série temporelle (P_t) des précipitations mensuelles.

Pour cela, nous allons partager l'étude en 2 parties :

- D'abord, on étudie l'hypothèse selon laquelle la précipitation cumulée (P_t) d'un mois de l'année est stationnaire. Si nous validons cette hypothèse, cela implique que les propriétés statistiques que nous aurons observé sont constantes (annuellement) et indépendantes du temps.
- Nous chercherons alors à estimer les paramètres d'une certaine distribution pour qu'elle s'ajuste de manière optimale avec la distribution des précipitations cumulées de chaque mois.

1. Stationnarité du processus des précipitations mensuelles

Nous allons étudier l'hypothèse selon laquelle la précipitation cumulée (P_t) d'un mois de l'année est stationnaire.

Pour cela, nous avons à disposition les données de précipitations journalières depuis le 01/01/1950. Nous avons ensuite calculé les précipitations cumulées mensuelles pour chaque *CRESTA*, comme étant la somme des précipitations journalières de chaque mois. Il s'agit donc des réalisations d'un processus aléatoire discret, ou série temporelle, noté (P_t).

Nous cherchons à caractériser les propriétés essentielles de ce processus. Le problème est largement simplifié s'il est stationnaire.

Un processus discret (Z_t) est stationnaire, au sens fort, si pour toute fonction mesurable f et pour tout entier t et k :

$$f(Z_1, Z_2, \dots, Z_t) \text{a la même loi que } f(Z_{1+k}, Z_{2+k}, \dots, Z_{t+k})$$

Cela signifie que toutes les propriétés statistiques caractérisant le processus se conservent et sont indépendantes du temps. Autrement dit, le processus se comportera de la même manière que l'on situe à l'instant t ou à l'instant $t + k$.

Une étude graphique permettra de visualiser les comportements mensuels des précipitations, et ainsi appréciera si l'hypothèse est vraisemblable. Pour chaque année, nous allons utiliser une boîte à moustaches donnant, de manière concise, la répartition des précipitations cumulées de chaque mois.

Le graphique suivant présente les boîtes à moustaches des précipitations mensuelles par année, tous *CRESTA* et mois confondus. Pour chaque boîte à moustaches (année fixée), chaque donnée utilisée est la précipitation mensuelle pour un certain *CRESTA* et un certain mois pour l'année considérée.

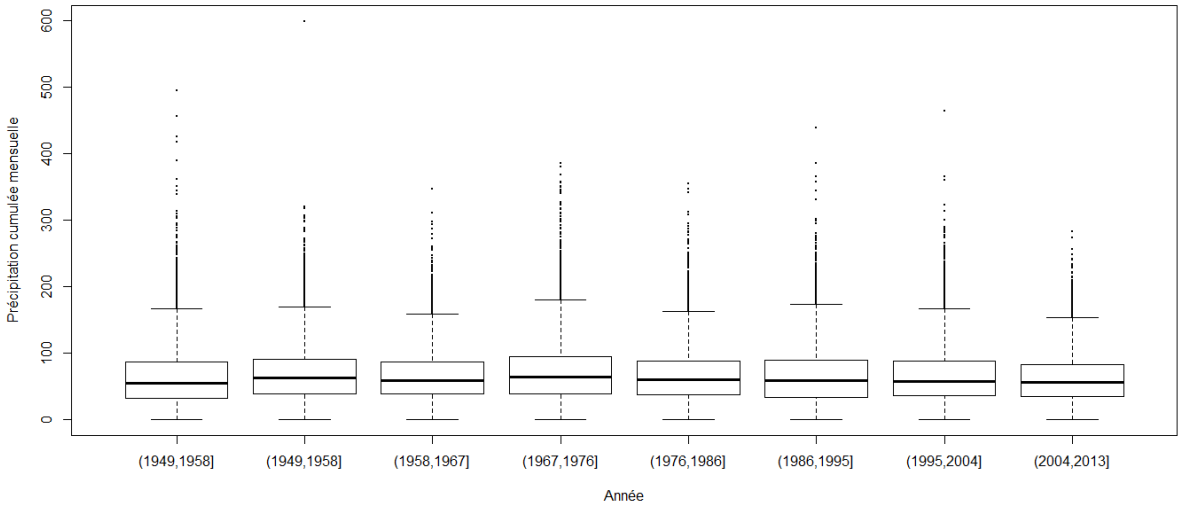


Figure 23 - Boîtes à moustaches des précipitations mensuelles par année, tous *CRESTA* et mois confondus

Nous remarquons que la répartition des précipitations mensuelles est globalement constante d'année en année. Une estimation paramétrique pour une certaine loi pourra alors suffire à modéliser les précipitations mensuelles.

Cependant, nous ne savons toujours pas si nous pouvons simuler la même loi de précipitation mensuelle pour tous les mois de l'année.

Le graphique suivant présente les boîtes à moustaches des précipitations par mois, tous *CRESTA* et années confondus. Pour chaque boîte à moustaches (mois fixé), chaque donnée utilisée est la précipitation mensuelle pour un certain *CRESTA* et une certaine année pour le mois considéré.

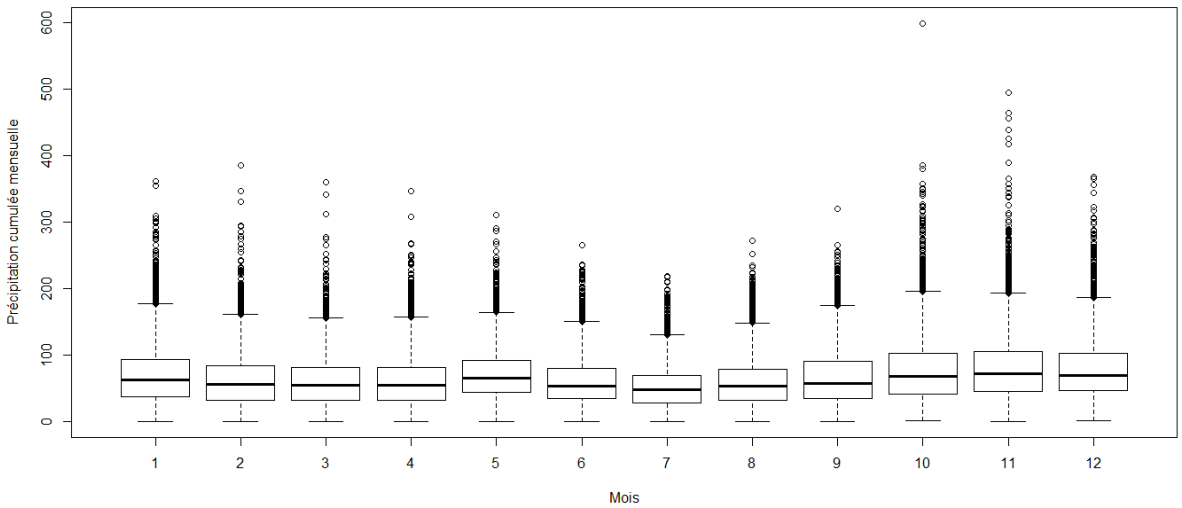


Figure 24 - Boîtes à moustaches des précipitations cumulées mensuelles, tous *CRESTA* et années confondus

Nous remarquons que la répartition des précipitations mensuelles n'est pas constante au cours d'une année. En effet, les principaux quartiles sont globalement stationnaires, mais les derniers déciles ne le sont pas. La variance des précipitations mensuelles est plus élevée en automne que pour le reste de l'année.

Il faudra donc modéliser les précipitations mensuelles pour chaque mois de l'année.

2. Estimation paramétrique de la loi des précipitations mensuelles

Dans la section précédente, nous avons validé l'hypothèse selon laquelle les précipitations mensuelles sont stationnaires d'une année à l'autre au sens fort. Nous allons alors supposer que la précipitation cumulée, pour un mois et un *CRESTA* donnés, est stationnaire.

Pour chaque *CRESTA* et chaque mois de l'année, nous allons finalement paramétrer une loi particulière pour modéliser la précipitation cumulée.

Nous allons procéder en plusieurs étapes :

- On choisit une loi particulière.
- On estime les paramètres de cette loi qui permettent d'ajuster de manière optimale sa distribution avec la distribution empirique des précipitations mensuelles. On utilisera la méthode du maximum de vraisemblance.
- On apprécie la qualité du modèle grâce à la méthode du *Q-Q plot* décrite ci-après.

a) Hypothèses

Soit $(P_t)_{t\in\mathbb{N}}$ une suite de variables aléatoires donnant la précipitation cumulée mensuelle. On les suppose indépendantes et identiquement distribuées, et de même loi que P.

Les données utilisées sont des réalisations de P.

On pose :

$$\begin{cases} \bar{P}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} P_i \\ s_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} (P_i - \bar{P}_n)^2 \end{cases}$$

Nous allons étudier si la loi de P est une des lois suivantes :

- La loi Normale $\mathcal{N}(\mu,\ \sigma^2)$, dont la densité s'écrit :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \mu, \sigma \in \mathbb{R}$$

  Pour connaître les paramètres optimaux, la méthode du maximum de vraisemblance[21] les estime de la manière suivante :

$$\begin{cases} \hat{\mu} = \bar{P}_n \\ \widehat{\sigma^2} = s_n^2 \end{cases}$$

- La loi Exponentielle $\mathcal{E}(\lambda)$, dont la densité s'écrit :

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda \in \mathbb{R}_+^*$$

  Pour connaître le paramètre optimal, la méthode du maximum de vraisemblance l'estime de la manière suivante :

[21] Si l'on suppose que X suit une loi $\mathcal{L}$, on va chercher les paramètres de $\mathcal{L}$ qui maximise la vraisemblance de la distribution observée de X, c'est-à-dire ceux qui maximisent la probabilité d'avoir la distribution observée.

$$\hat{\lambda} = \frac{1}{\bar{X}_n}$$

- La loi Gamma $\gamma(\alpha, \beta)$, dont la densité s'écrit :

$$f(x) = t^{\alpha-1}\frac{\beta^{\alpha}e^{-\beta x}}{\Gamma(\alpha)}\,\mathbb{I}_{x\geq 0}\,, \quad \mu, \sigma \in \mathbb{R}_+^* \text{ et } \Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt$$

Pour connaître les paramètres optimaux, la méthode du maximum de vraisemblance les estime de la manière suivante :

$$\begin{cases} \hat{a} = \dfrac{\bar{X}_n{}^2}{s_n^2} \\ \hat{\beta} = \dfrac{\bar{X}_n}{s_n^2} \end{cases}$$

Nous n'avons pas cherché à paramétrer d'autres lois car la loi Gamma s'avère être très performante.

b) Résultats

Nous allons maintenant étudier la qualité d'ajustement de chacune des lois énoncées ci-dessus et choisir la meilleure. Un outil graphique, très simple à implémenter et à interpréter, est largement utilisé pour comparer des distributions : le *Q-Q plot*.

Le *Q-Q plot* est un outil graphique puissant qui permet d'évaluer la pertinence de l'ajustement d'une distribution donnée avec celle des données empiriques. On compare la position d'un certain nombre de quantiles empiriques (observés) avec celle de quantiles théoriques (déduits du modèle).
Si les distributions sont semblables, chaque quantile empirique d'un certain niveau sera correctement estimé par le quantile théorique de même niveau (ils seront quasiment égaux).

En pratique, pour des paramètres fixés, on simule une certaine loi pour ensuite trier les réalisations dans l'ordre croissant. Puis nous comparons cette suite croissante de valeurs avec celle des réalisations observées empiriquement triées dans l'ordre croissant. Si les distributions sont semblables, l'ensemble des points forme une bissectrice sur le *Q-Q plot* et les distributions associées peuvent être considérées comme semblables. Nous regarderons donc la disposition des points $(\text{Quantile théorique}, \text{Quantile empirique})$ par rapport à la droite $y = x$.

Les graphiques suivants présentent les *Q-Q plot* obtenus pour chacune des lois concernées avec les paramètres estimés par maximum de vraisemblance. Les données utilisées sont les précipitations mensuelles, tous *CRESTA* et mois confondus.

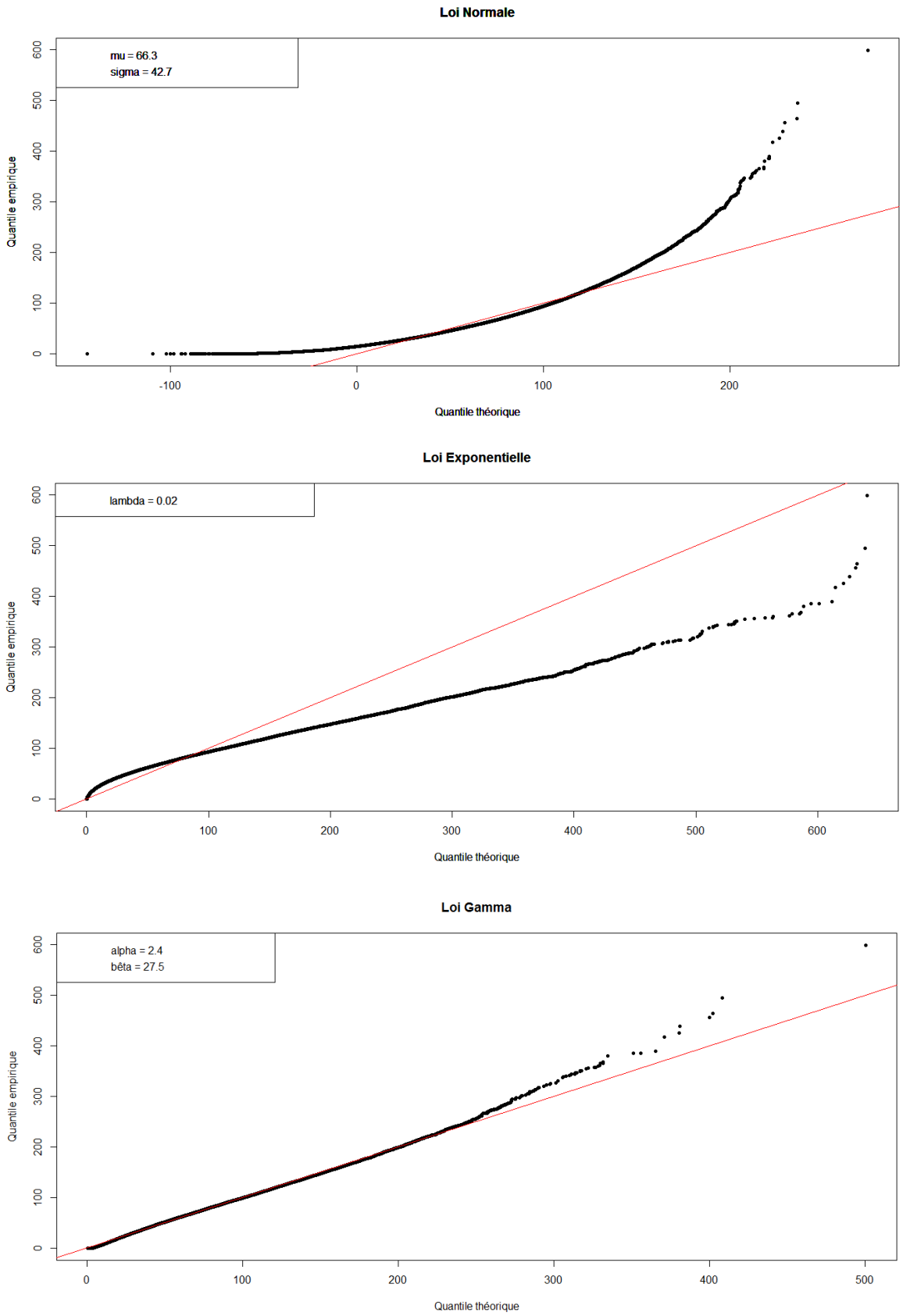


Figure 25 - Tous *CRESTA* et mois confondus : *Q-Q plot* des précipitations mensuelles avec une loi Normale, Exponentielle et Gamma

La droite rouge est celle qui a pour équation $y = x$. Nous observons que la loi Normale sous-estime les précipitations mensuelles et que la loi Exponentielle les surestime. Elles ne sont pas adéquates.

La loi Gamma semble très adéquate pour modéliser les précipitations mensuelles[22] : les quantiles empiriques observés correspondent globalement aux quantiles de la loi paramétrée. Cependant, le modèle sous-estime les extrêmes. En effet, il est indifférent au mois et au *CRESTA* concerné. Or, nous avons vu sur la figure 24 que la variance des précipitations mensuelles n'était pas stationnaire. Une unique loi ne peut donc représenter la répartition des précipitations mensuelles tous *CRESTA* et mois confondus. Le résultat fournit par le modèle Gamma est donc particulièrement satisfaisant.

Pour affiner la précision de notre modèle, nous allons répéter l'opération pour chaque *CRESTA* et pour chaque mois, en ajustant une loi Gamma.

Nous donnons un exemple de *Q-Q plot* obtenu :

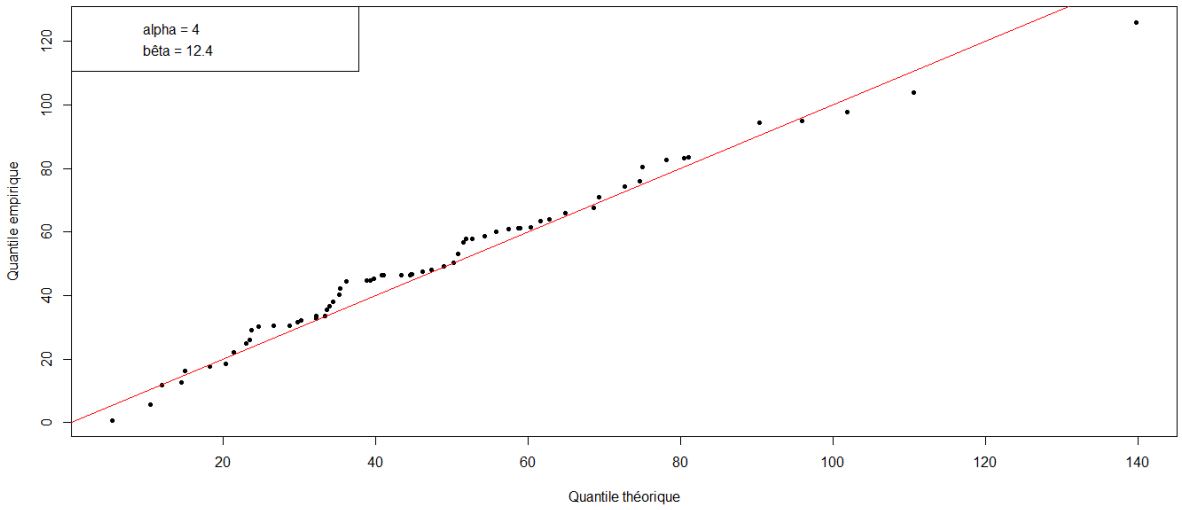


Figure 26 - *CRESTA* 92 en juin : *Q-Q plot* des précipitations mensuelles avec une loi Gamma

Les résultats étant tous satisfaisants, nous modéliserons les précipitations mensuelles par une loi Gamma paramétrée pour chaque *CRESTA* et pour chaque mois. Il y aura donc 1152[23] modèles de précipitations mensuelles.

[22] Cela est conforme avec l'hypothèse que les précipitations mensuelles utilisées dans la construction de l'indicateur *SPI* suivent une loi Gamma.
[23] $1152 = 96 \times 12$

C. Segmentation de la France en régions homogènes en termes de température

La partie précédente permet de simuler 10 000 scénarios mensuels d'évolution des précipitations cumulées. Il reste à modéliser les autres variables : température, évapotranspiration, indicateurs *SPI* et *SPEI*. Il suffit alors de modéliser les températures maximales journalières. Nous allons nous inspirer d'un modèle récemment développé au GIE AXA qui visait à modéliser le risque grêle à partir des températures minimales journalières.

La modélisation des températures maximales journalières est divisée en trois étapes :

- Segmentation de la France en plusieurs régions distinctes, où chacune d'elles est homogène en termes de température.
- Modélisation de la série temporelle des températures maximales journalières pour chaque région.
- Modélisation de la dépendance entre les régions en termes de température.

1. Les principes de la classification

Dans la première section, nous avons décrit l'importance de prendre en compte l'évolution des températures dans la modélisation de la sécheresse. Afin de générer un catalogue de scénarios probabilisés de sinistralité pour l'année à venir, nous devons être capables de simuler les variables explicatives de la sécheresse (précipitations, températures, …) partout en France.

Il serait fastidieux de développer un modèle pour chaque station météorologique, et ensuite d'étudier les dépendances entre chacune d'elles. Il est plus intéressant de regrouper les stations dans des zones relativement homogènes en température, pour baser notre modélisation sur seulement quelques points au lieu des 251 initiaux. Pour chaque zone, les températures associées seront la moyenne des températures des stations se trouvant dans la zone.

Dans cette optique, deux étapes seront nécessaires :

- L'analyse par composante principale (ACP) : l'objectif est de simplifier les données pour éviter des redondances. On cherche à projeter nos points sur un espace de dimension réduite[24] en perdant le minimum d'informations. La méthode de résolution optimale est décrite ci-après. Des graphiques nous aideront à interpréter les résultats de l'ACP pour choisir un espace de dimension satisfaisante sur lequel projeter les 251 points initiaux.
- La classification ascendante hiérarchique (CAH) : l'objectif est de regrouper les points présentant des comportements similaires. Dans notre cas, nous voulons regrouper les stations météorologiques dont les relevés de températures sont similaires par rapport aux autres. La CAH est un algorithme de classification qui crée à chaque étape une partition obtenue en agrégeant deux à deux les éléments[25] les plus proches, constituant ainsi des classes d'éléments. Une notion de distance est donc à définir. Dans notre cas, les coordonnées de nos points ne sont pas liées à une localisation géographique, mais à des températures. Cet algorithme hiérarchise donc les partitions et finit par agréger tous les éléments en une seule classe.

[24] Dans notre cas, chacun des 251 points possède 23 741 coordonnées, ce qui est énorme.
[25] Un élément peut être une station météorologique ou un ensemble de stations déjà regroupées (ou « classe »).

2. L'analyse par composante principale

a) Hypothèses

Les données de températures disponibles forment une matrice de la forme :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad avec \begin{cases} n = 251 \\ k = 23\,741 \end{cases}$$

x_{nk} est la température maximale enregistrée dans la n-ième station météorologique et le k-ème jour après le 01/01/1950.

La n-ème ligne de X représente l'évolution des températures maximales quotidiennes pour la n-ème station météorologique.

La k-ième colonne de X représente le relevé des températures maximales enregistrées dans chacune des stations pour le k-ième jour.

Les données doivent ensuite être centrées : pour chaque colonne, le vecteur des températures est centré par la moyenne.

$$Y = \begin{pmatrix} x_{11} - \overline{x}_{.1} & \cdots & x_{1k-} \overline{x}_{.k} \\ \vdots & \ddots & \vdots \\ x_{n1} - \overline{x}_{.1} & \cdots & x_{nk} - \overline{x}_{.k} \end{pmatrix}$$

$$o\grave{u} \quad \forall j \in \{1, \dots, k\}, \ \overline{x}_{.j} = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$$

La matrice de variance-covariance entre les stations, de dimension $k \times k$, est définie de la manière suivante :

$$\Sigma = \frac{1}{n-1} Y'Y$$

b) Méthode de résolution

Initialement, il y a un nuage de $n = 251$ points situés dans un espace à $k = 23\,741$ dimensions.

L'objectif est de projeter ces points dans un espace $E_{k'}$ à k' dimensions, avec $k' < n < k$, tout en conservant un maximum d'informations. On va chercher un vecteur normalisé $a = (a_1, a_2, \dots, a_k)' \in \mathbb{R}^k$ tel que Ya soit le plus variable possible, pour pouvoir expliquer la variance Σ le mieux possible.

Le problème d'optimisation est le suivant :

$$\begin{cases} Max \ \ a'\Sigma a \\ s.c. \ \ a'a = 1 \end{cases}$$

La solution de cette équation vérifie $\Sigma a = \lambda a$. ($a'\Sigma a = \lambda$)

Σ est une matrice variance-covariance et est donc définie positive. Il y a alors exactement k valeurs propres (notées $\lambda_l \in \mathbb{R}, 1 \leq l \leq k$) , associées aux vecteurs propres (notés $u_l \in \mathbb{R}^k, 1 \leq l \leq k$). Les valeurs propres λ_l sont ensuite classées dans l'ordre décroissant.

Les points initiaux sont projetés sur les axes dirigés par les u_l. Les premiers axes sont ceux perdant le moins d'informations, car ils sont associés aux plus grandes valeurs propres qui mesurent « l'inertie » de la projection.

La quantité d'information contenue dans l'axe Δu_l peut être évaluée par le taux d'inertie expliqué :

$$\frac{\lambda_l}{\sum_{l=1}^{p} \lambda_l}$$

λ_l est l'inertie de la projection du nuage des points initiaux sur l'espace E_k.

$\sum_{l=1}^{p} \lambda_l$ est l'inertie totale.

c) Résultats

Après application de la méthode énoncée précédemment, on obtient une matrice ayant 251 lignes et 251 colonnes. Chaque colonne représente un axe sur lequel sont projetés les 251 points représentant l'emplacement des stations météorologiques : pour un axe donné, la n-ième coordonnée est la coordonnée de la n-ième station météo projetée sur cet axe. Pour chaque axe, on calcule l'inertie des valeurs propres représentant la quantité d'information conservée. Le premier axe contient plus d'informations que le deuxième, et le deuxième plus que le troisième, et ainsi de suite.

Le graphique suivant représente le taux d'inertie expliqué pour chacun des axes obtenus.

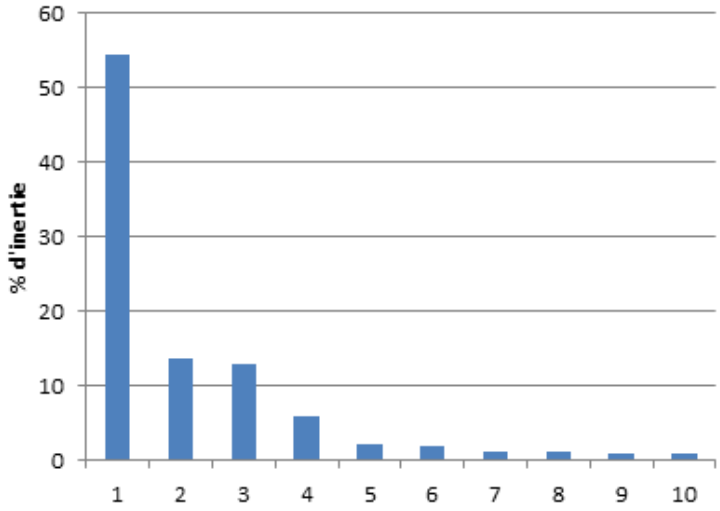


Figure 27 - Taux d'inertie expliqué pour chacun des axes obtenus

A partir du cinquième axe, le taux d'inertie expliqué est inférieur à 3 % : l'information est donc essentiellement contenue dans les quatre premiers axes. Cela n'est pas surprenant, car il y en a autant que de saisons dans une année. La température se comporte relativement de la même manière tous les ans, et donc les comportements moyens par saison peuvent caractériser les stations météorologiques : chaque point est identifié par quatre coordonnées.

On fixe les quatre premiers axes sur lesquels on a projeté les 251 points représentant l'emplacement des stations météorologiques.

Les données étant simplifiées, nous pouvons passer à la classification.

3. La classification hiérarchique ascendante

a) Hypothèses

Pour faciliter la modélisation des températures et diminuer les redondances d'informations, on cherche à regrouper dans une même classe les stations météorologiques présentant des similarités dans les relevés de températures, afin de se restreindre à un nombre limité de classes.

La classification permet de regrouper des points en plusieurs classes en fonction de la distance qui sépare chacun d'eux. Une fois les premières classes constituées, on recommence l'opération en regroupant les éléments (point ou classe) les plus proches.

Deux types de distances sont donc à définir : une distance entre les points et une distance entre les classes. Dans notre cas, chaque point représente une station météorologique et ses coordonnées sont les relevés de températures.

Les résultats de l'ACP nous ont poussés à projeter les $n = 251$ points initiaux dans un espace à $k' = 4$ dimensions.

On considère alors un ensemble fini $E \subset \mathbb{R}^{k'}$ contenant n points représentant les relevés de températures des stations météorologiques.

Choix d'une distance entre points

Une distance est une application $d : E \times E \to \mathbb{R}_+$ vérifiant :

$$\forall (x, y) \in E \times E,\ \ d(x, y) = 0 \Leftrightarrow x = y$$
$$\forall (x, y) \in E \times E,\ \ d(x, y) = d(y, x)$$
$$\forall (x, y, z) \in E \times E \times E,\ \ d(x, y) \leq d(x, z) + d(z, y)$$

Les distances les plus usuelles sont :

- La distance euclidienne : $d(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$
- La distance de Manhattan : $d(x, y) = \sum_{i=1}^{d} |x_i - y_i|$
- La distance de Sebestyen : $d(x, y) = \sqrt{\sum_{i=1}^{d} w_i (x_i - y_i)^2}$ $\quad avec\ w_i > 0$
- La distance de Tchebychev : $d(x, y) = \max_i |x_i - y_i|$

Nous utiliserons la distance euclidienne pour mesurer la proximité entre les points.

Choix d'une distance entre classes

Après avoir regroupé au moins deux points dans une classe, il faut être capable de donner une distance entre cette classe et le reste des éléments.

Cette distance ne remplit pas nécessairement les propriétés ci-dessus, et ne peut pas être véritablement qualifiée de distance. Cependant, cette « pseudo-distance » quantifie bien la proximité entre classes.

Soient A et B deux classes. Soit d une distance entre points.

Les méthodes les plus usuelles sont :

- La méthode du saut minimal : $\bar{d}(A,B) = \inf_{x\in A, y\in B} d(x,y)$
- La méthode du saut maximal : $\bar{d}(A,B) = \sup_{x\in A, y\in B} d(x,y)$
- La méthode du saut moyen : $\bar{d}(A,B) = \frac{1}{Card(A)Card(B)} \sum_{x\in A, y\in B} d(x,y)$
- La méthode de Ward : $\bar{d}(A,B) = \frac{p_A p_B}{p_A+p_B} d(g_A,g_B)^2 \quad avec \begin{cases} p_A = \frac{Card(A)}{n},\ p_B = \frac{Card(B)}{n} \\ g_A = \frac{1}{Card(A)} \sum_{x\in A} x,\ g_B = \frac{1}{Card(B)} \sum_{x\in B} x \end{cases}$

  p_A et p_B sont respectivement les poids des classes A et B.
  g_A et g_B sont respectivement les centres de gravité des classes A et B.

Nous utiliserons la méthode de Ward, qui est la plus utilisée, pour mesurer la proximité entre les classes.

b) Méthode de résolution

Après avoir choisi la distance d entre points et la distance $\bar{d}$ entre classes, on applique l'algorithme suivant :

---

1. **Début**
2. Calculer les distances entre tous les éléments (points ou classe) pris deux à deux
3. Agréger les deux éléments (points ou classe) les plus proches pour former une nouvelle classe
4. **Si** il reste une unique classe
5. Sortir
6. **Sinon**
7. Refaire la première étape
8. **Fin**

Ainsi, la CAH permet de regrouper des classes d'éléments en remontant jusqu'à une unique classe. A chaque étape, on obtient une partition de E en différentes classes.

On peut alors définir deux types de variance :

- La variance intra-classe, qui est la moyenne des variances au sein de chaque classe. C'est un indicateur qui mesure l'homogénéité de chacune des classes.
- La variance interclasse, qui est la variance des centres de gravité au sein de chaque classe. C'est un indicateur qui mesure à quel point les classes sont distinctes.

Pour choisir le nombre de classes que nous allons garder, on cherchera simultanément à minimiser la variance intra-classe et à maximiser la variance interclasse.

c) Résultats

Après application de la méthode énoncée précédemment, on obtient un dendrogramme qui permet de visualiser la classification effectuée. Un dendrogramme se présente sous forme d'arbre binaire et représente les agrégations successives jusqu'à réunion de tous les points en une seule classe. De plus, la hauteur d'une branche est proportionnelle à la variance interclasse entre les deux éléments regroupés.

Le dendrogramme obtenu à partir de nos données et avec la distance euclidienne et la méthode de Ward est le suivant :

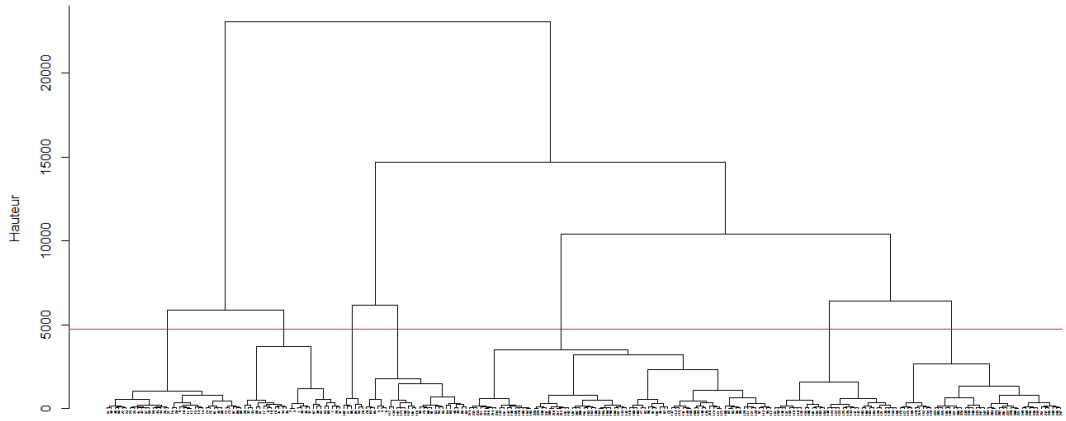


Figure 28 - Dendrogramme issu de la CAH

Pour obtenir une partition de E, on tire un trait horizontal (ici en rouge) qui coupe l'arbre en plusieurs classes (ici en 7 classes).

L'arbre est coupé à l'endroit où les sous arbres ont de faibles hauteurs (c'est-à-dire lorsque les classes sont encore très similaires). Nous décidons de partitionner E en 4 à 8 classes.

Les cartes suivantes montrent les résultats de la classification pour un nombre fixé de classes :

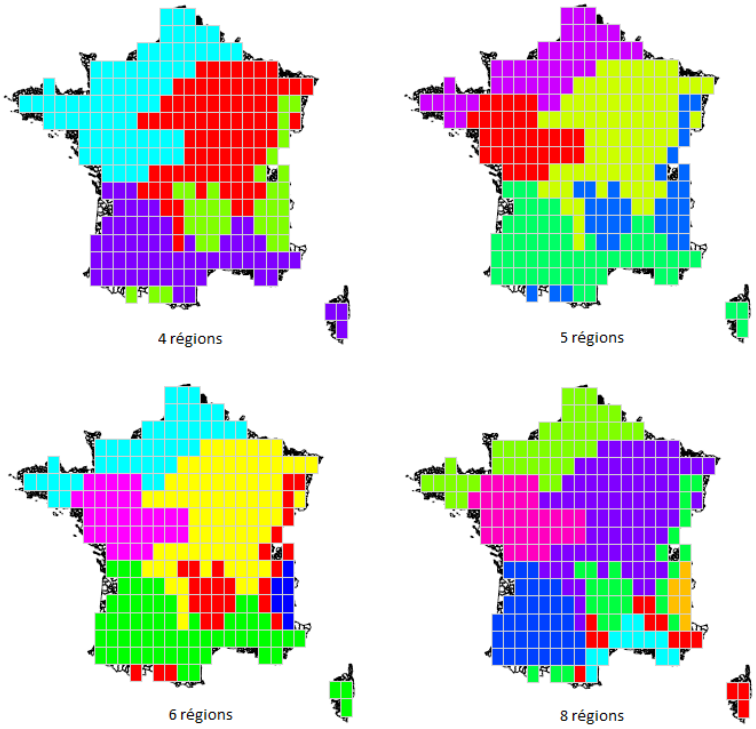


Figure 29 - Résultats obtenus avec la CAH pour différents nombres de classes

Sans surprise, différentes classes ou régions ressortent : les régions montagneuses, les régions du Sud, du Nord-Est et du Nord-Ouest.

Afin de se fixer un nombre de classes, il reste à étudier les variances intra-classes :

Nombre de classes	Variance intra-classe
4	4,3
5	3,5
6	2,6
7	2,6
8	2,3

Tableau 1 - Variances intra-classe

Pour que la classification des stations météorologiques soit cohérente, c'est-à-dire pour obtenir des régions dans lesquelles les températures sont homogènes, nous nous fixons un seuil de variance intra-classe de 4. Cela signifie que l'écart-type des températures est inférieur ou égal à 2°C. Cependant, la variance interclasse doit être maximale pour maximiser l'effet de la segmentation et obtenir des régions les moins similaires possibles.

Cela nous amène donc à segmenter la France en 5 régions homogènes en termes de températures maximales journalières :

- Région 1 : le sud de la France (en vert)
- Région 2 : les zones montagneuses (en bleu)
- Région 3 : l'est de la France (en vert clair)
- Région 4 : les pays de la Loire (en rouge)
- Région 5 : le nord de la France (en violet)

D. Modélisation des températures maximales journalières

La modélisation des températures maximales journalières étant nécessaire pour modéliser la sécheresse, la France a été segmenté en 5 régions étant chacune homogène (variance intra-classe faible) et distinctes entre elles (variance interclasse forte).

Nous avons donc à disposition 5 séries temporelles de températures : pour chaque région, une unique série temporelle est construite comme étant la moyenne des séries temporelles de températures des stations météorologiques présentes dans la région.

D'abord, chaque série temporelle est décomposée afin d'avoir une partie tendancielle et saisonnière propres aux comportements des températures. La tendance et la saisonnalité sont déterministes. La dernière composante de la série, ou série résiduelle, est aléatoire et doit être modélisée.

Ensuite, la série résiduelle est décomposée de manière à ce que la partie aléatoire soit la plus stationnaire possible.

Enfin, nous allons modéliser la dépendance entre les régions afin d'avoir une cohérence dans les futures simulations.

1. Décomposition des séries temporelles des températures

La région Sud est celle qui est la plus exposée au risque de sécheresse. Pour la décomposition des séries temporelles des températures, nous allons nous focaliser sur cette région.

Le graphique suivant donne l'évolution des températures maximales quotidiennes moyennes au sein de la région Sud :

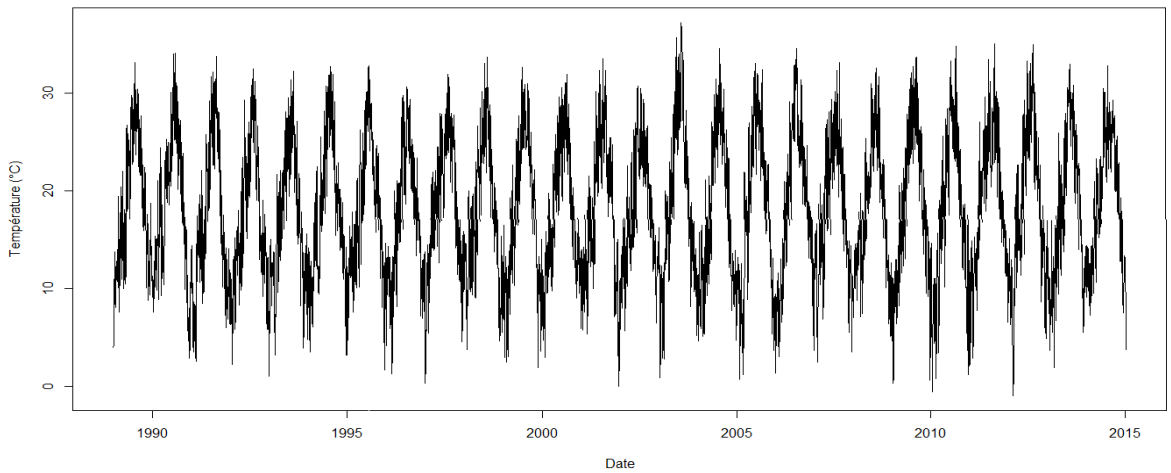


Figure 30 - Evolution des températures maximales dans la région Sud

Ce graphique illustre bien le caractère saisonnier des températures. La saisonnalité devra être prise en compte dans notre modèle, et permettra de modéliser un comportement de court terme.

De plus, le réchauffement climatique annoncé par la communauté scientifique indique qu'une tendance haussière des températures existe sur le long terme. Une tendance devra être prise en compte dans notre modèle, et permettra de modéliser un comportement de long terme.

Nous aimerions donc décomposer la série temporelle des températures en une tendance, une saisonnalité et une série résiduelle. Ainsi, la série temporelle (X_t) des températures s'écrit :

$$X_t = m_t + s_t + Z_t$$

- (X_t) : températures maximales journalières (aléatoire)
- (m_t) : tendance (déterministe)
- (s_t) : saisonnalité (déterministe)
- (Z_t) : série résiduelle (aléatoire)

a) Tendance (m_t)

La tendance des températures s'obtient en appliquant une méthode régression locale appelée *LOESS*. Il s'agit d'une méthode qui effectue une régression non-paramétrique sur des sous-ensembles locaux de données. Cette méthode présente l'avantage de ne pas définir une unique fonction globale qui ajusterait un modèle à l'ensemble des données de l'échantillon, puisque la méthode consiste à calculer autant de fonctions locales qu'il y a de segments de données.

Plus précisément, il s'agit d'une régression polynomiale avec pondération locale. Pour un sous-ensemble local de données, on cherche à effectuer une régression par un polynôme de faible degré pour éviter un sur-ajustement des données. Les coefficients du polynôme sont calculés à l'aide de la méthode des moindres carrés pondérés. La pondération sert à donner plus de poids aux points les plus proches. La fonction de pondération utilisée est une fonction cubique pondérée : $w(x) = (1 - |x|^3)\mathbb{I}_{|x|<1}$

Ensuite, nous modéliserons la tendance de long terme par une droite, construite par régression linéaire sur les données obtenues par régression locale.

Le graphique suivant illustre le résultat de cette méthode appliquée à notre série temporelle des températures :

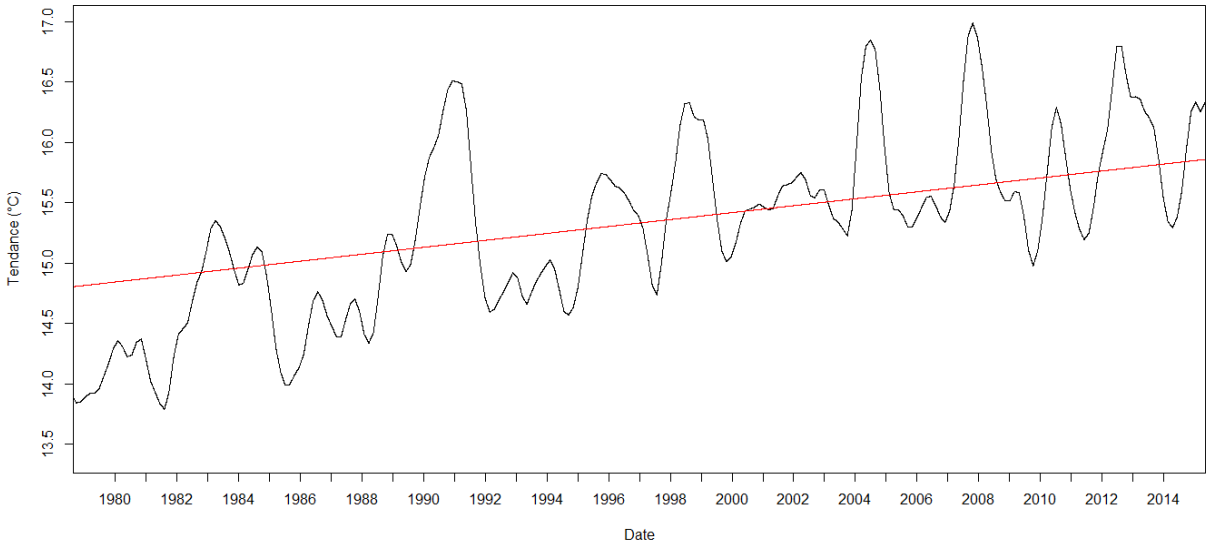


Figure 31 - Evolution de la tendance

Nous obtenons que :

$$m_t = 7{,}1.\,10^{-5} \times t + 1{,}6$$

Comme attendu, une tendance légèrement haussière est observée sur le long terme. Cela est lié au réchauffement climatique.

b) Saisonnalité (s_t)

La saisonnalité (s_t) s'obtient en étudiant la série des températures sans tendance $(X_t - m_t)$.

Elle doit vérifier, pour une périodicité T donnée : $\begin{cases} \forall t \in \mathbb{Z},\ s_t = s_{t+T} \\ \sum_{t=1}^{T} s_t = 0 \end{cases}$

Pour modéliser la saisonnalité, on va effectuer deux types de régression linéaire :

- Régression linéaire sur $t \longmapsto \big(cos(\omega t), \sin(\omega t)\big)$: périodicité annuelle
- Régression linéaire sur $t \longmapsto \big(cos(\omega t), \sin(\omega t), cos(2\omega t), \sin(2\omega t)\big)$: périodicité semi-annuelle

Avec $\omega = \frac{2\Pi}{365}$

Le graphique suivant présente les résultats :

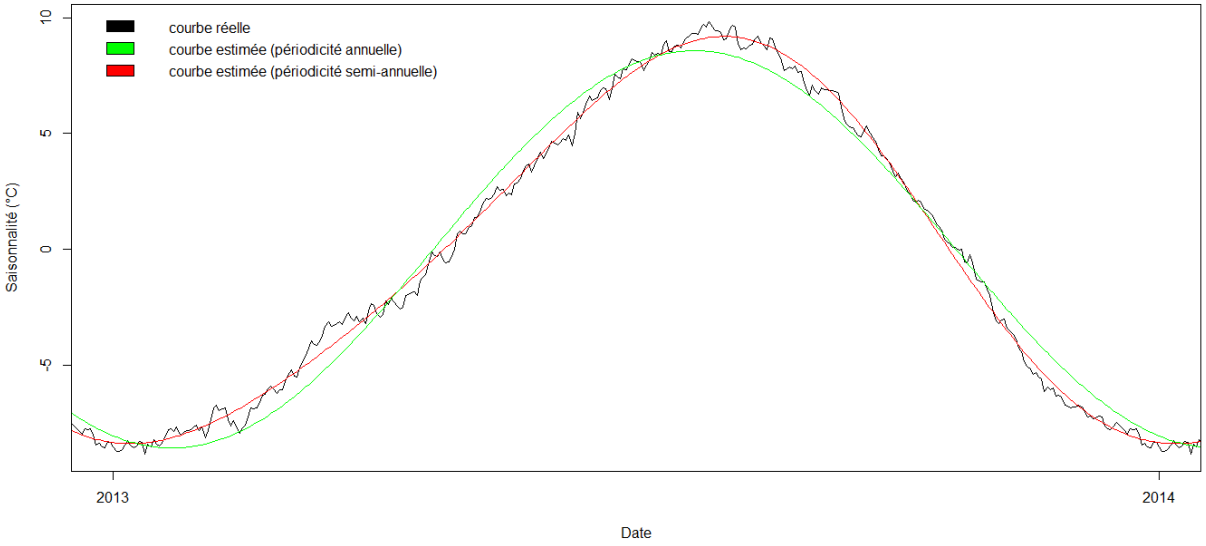


Figure 32 - Evolution de la saisonnalité

La régression effectuée avec une périodicité semi-annuelle semble mieux capter la saisonnalité des températures. On conserve donc la périodicité semi-annuelle. La régression linéaire associée nous permet de modéliser la saisonnalité de la manière suivante :

$$s_t = -8{,}2 \times cos(\omega t) - 3{,}1 \times \sin(\omega t) - 0{,}3 \times cos(2\omega t) + 1{,}1 \times \sin(2\omega t)$$

c) Série résiduelle (Z_t)

Après avoir modélisé la tendance et la saisonnalité de manière déterministe, nous allons devoir développer un modèle aléatoire capable de capter le maximum des dynamiques stochastiques présentes dans l'évolution des températures.

Il reste donc à étudier la série résiduelle (Z_t) défini par : $Z_t = X_t - m_t - s_t$

Le graphique suivant montre son évolution :

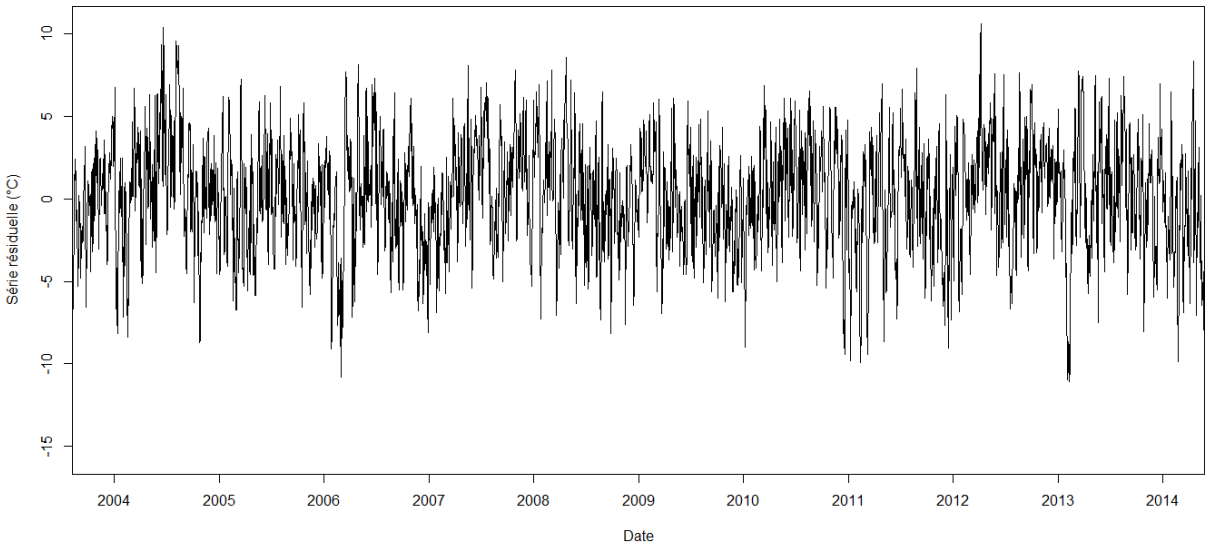


Figure 33 - Evolution de la série résiduelle

Nous cherchons donc à caractériser les propriétés essentielles des températures. Comme dans le cas des précipitations, le problème est largement simplifié si le processus est stationnaire.

Un processus discret (Z_t) est stationnaire, au sens faible, s'il remplit les propriétés suivantes :

- $\mathbb{E}(Z_i) = \mu \quad \forall i$
- $\mathbb{V}(Z_i) = \sigma^2 < +\infty \quad \forall i$
- $Cov(Z_i, Z_{i-k}) = \rho_k \quad \forall i, \forall k \geq 1$

Cela signifie qu'à chaque instant, l'espérance, la variance et la covariance, pour un écart temporel fixé, sont constantes.

La série résiduelle (Z_t) est centrée en 0 du fait que l'on a supprimé la tendance et la saisonnalité. Nous allons donc tester si (Z_t) est un bruit blanc.

Le processus (Z_t) est un bruit blanc si :

- $\mathbb{E}(Z_i) = 0 \quad \forall i$
- $\mathbb{V}(Z_i) = \sigma^2 < +\infty \quad \forall i$
- $Cov(Z_i, Z_{i-k}) = 0 \quad \forall i, \forall k \geq 1.$

Un bruit blanc est donc un processus stationnaire.

De plus, (Z_t) est un bruit blanc est gaussien si :

- (Z_t) est un bruit blanc
- $Z_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i$

Cela implique que les Z_i sont tous indépendants entre eux[26].

Pour vérifier ces propriétés, nous allons nous baser sur une étude graphique.

Le graphique suivant donne les boites à moustaches associées à (Z_t) pour tous les mois de l'année :

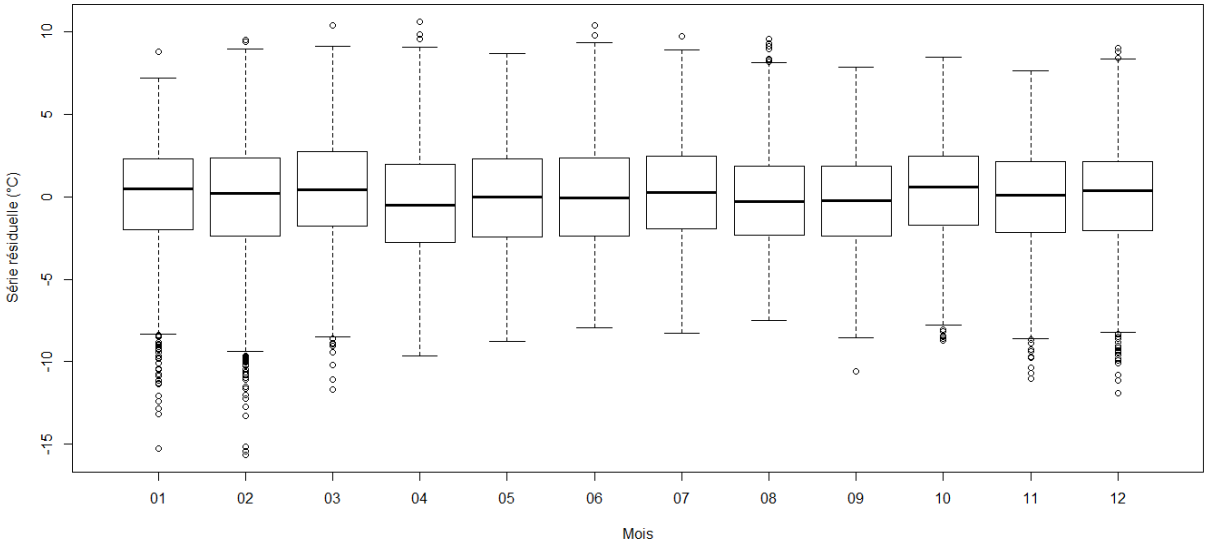


Figure 34 - Boîtes à moustaches mensuelles de (Z_t)

L'hypothèse que la variance est constante au cours du temps ne semble pas être vérifiée.

La variance n'est pas constante au cours du temps, mais semble périodique : plus on s'approche des mois d'hiver, plus les variances associées sont élevées.

Nous allons donc chercher à séparer Z_t en deux parties :

- Une partie liée à sa variance, qui semble périodique. On la modélisera avec une régression sur des fonctions périodiques.
- Une partie résiduelle. On la modélisera par un certain type de processus.

On pose :

$$Z_t = \rho_t . Y_t \quad avec \begin{cases} \rho_t = \sqrt{\mathbb{V}(Z_t)} \\ \mathbb{V}(Y_t) = 1 \end{cases}$$

Pour connaître ρ_t, on va calculer l'écart-type mobile de Z_t autour de t.
On pose :

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

h représente un paramètre de lissage : plus h est grand, plus la courbe de (ρ_t) est lisse.

[26] En effet, si X_1 et X_2 suivent des lois normales et $\mathrm{Cov}(X_1, X_2) = 0$, alors X_1 est indépendante de X_2.

Le graphique suivant présente les évolutions de ρ_t pour un paramètre h donné.

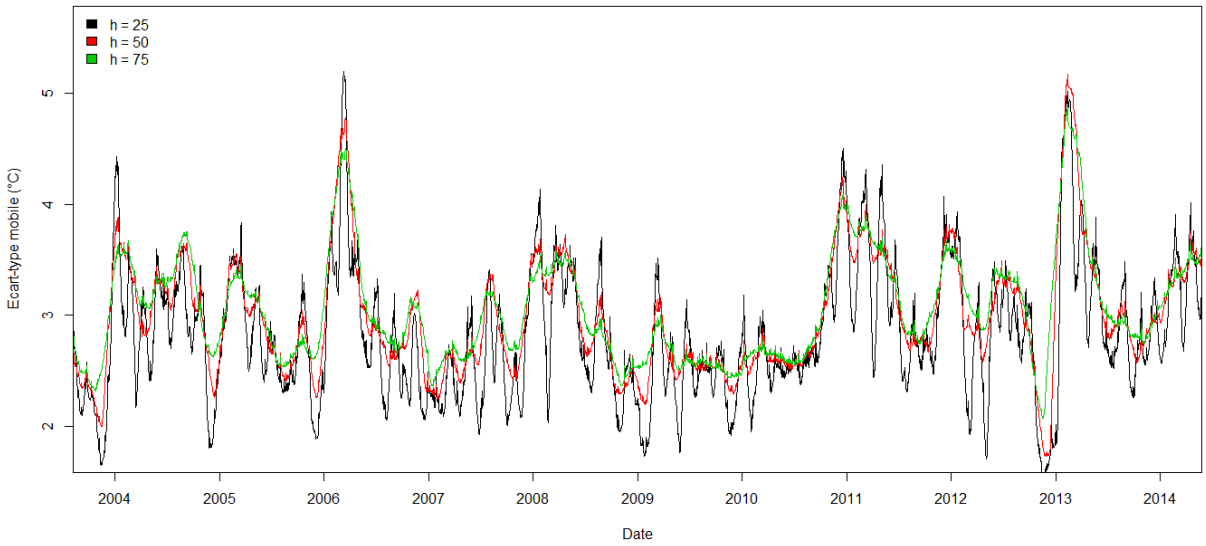


Figure 35 - Ecart-type de Z_t pour différents paramètres de lissage h

Nous gardons $h = 50$.

Les valeurs de $\rho_t$étant à présent connues, nous pouvons le modéliser par une régression linéaire sur $t \longmapsto \big(1, cos(\omega t), \sin(\omega t)\big)$.

Nous obtenons que :

$$\rho_t = 0{,}1 \times cos(\omega t) + 0{,}2 \times \sin(\omega t) + 3{,}1$$

Le graphique suivant montre les résultats obtenus :

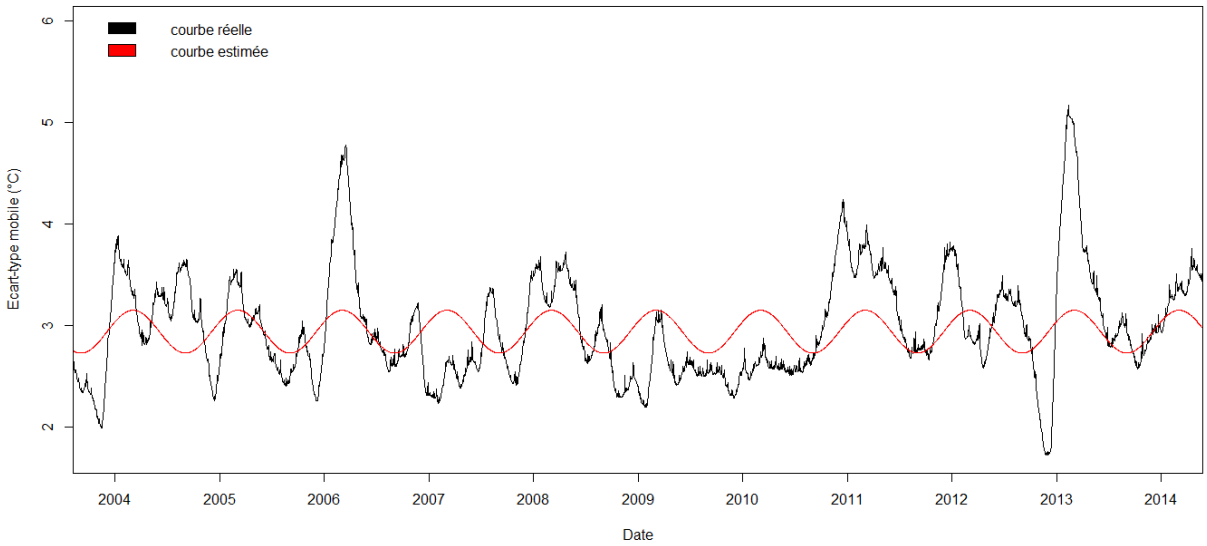


Figure 36 - Estimation de l'écart-type de Z_t

Les résultats sont plutôt satisfaisants. Une régression sinusoïdale permet donc de modéliser la série temporelle des écart-types de (Z_t). Il reste à modéliser la série (Y_t).

2. Modélisation des résidus (Y_t)

Dans la section précédente, la série temporelle des températures (X_t) est décomposée de la manière suivante :

- Une partie déterministe, qui regroupe la tendance (m_t), la saisonnalité (s_t), et les écart-types mobiles (ρ_t) de la série résiduelle $(X_t - m_t - s_t)$
- Une partie aléatoire, qui est la série des résidus réduits définis par :

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

Pour modéliser (X_t), il reste à modéliser (Y_t). Pour cela, on va ajuster les paramètres d'un certain type de processus : les processus $\mathrm{ARMA(p, q)}$.

$(Y_t)_{t\in\mathbb{N}}$ est un processus $\mathrm{ARMA(p, q)}$ s'il existe des coefficients réels $(\varphi_i)_{(1\leq i\leq p)}$ et $(\theta_j)_{(1\leq j\leq q)}$ tels que

$$Y_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Le package `forecast` du logiciel R possède une fonction qui cherche des entiers p^* et q^* afin d'optimiser l'ajustement du processus $\mathrm{ARMA}(p^*, q^*)$ sur nos données. Cependant, des valeurs maximales de p^* et q^* doivent être données.

Le processus $\mathrm{ARMA(p, q)}$, avec $p, q \leq 7$, qui approxime de manière optimal le processus $(Y_t)_{t\in\mathbb{N}}$ des résidus est un processus $\mathrm{ARMA(3,0)}$ avec :

$$Y_t = 0{,}9Y_{t-1} - 0{,}3Y_{t-2} + 0{,}1Y_{t-3} + \varepsilon_t$$

Pour valider le modèle, il serait souhaitable que $(\varepsilon_t)_{t\in\mathbb{N}}$ soit un bruit blanc gaussien pour que le processus soit stationnaire et que nous fassions de bonnes simulations.

Il faut donc vérifier les propriétés suivantes :

- $\forall i \neq j,\ \varepsilon_i$ indépendant de ε_j
- $\forall t,\ \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$

Nous allons en particulier vérifier que les résidus (ε_t) sont indépendants et identiquement distribués selon une loi normale standard[27].

Le graphique suivant représente la boîte à moustaches des résidus ε_t.

[27] En effet, $\mathbb{V}(\mathrm{Y_t}) = 1 \Rightarrow\ \mathbb{V}(\varepsilon_t) = 1$

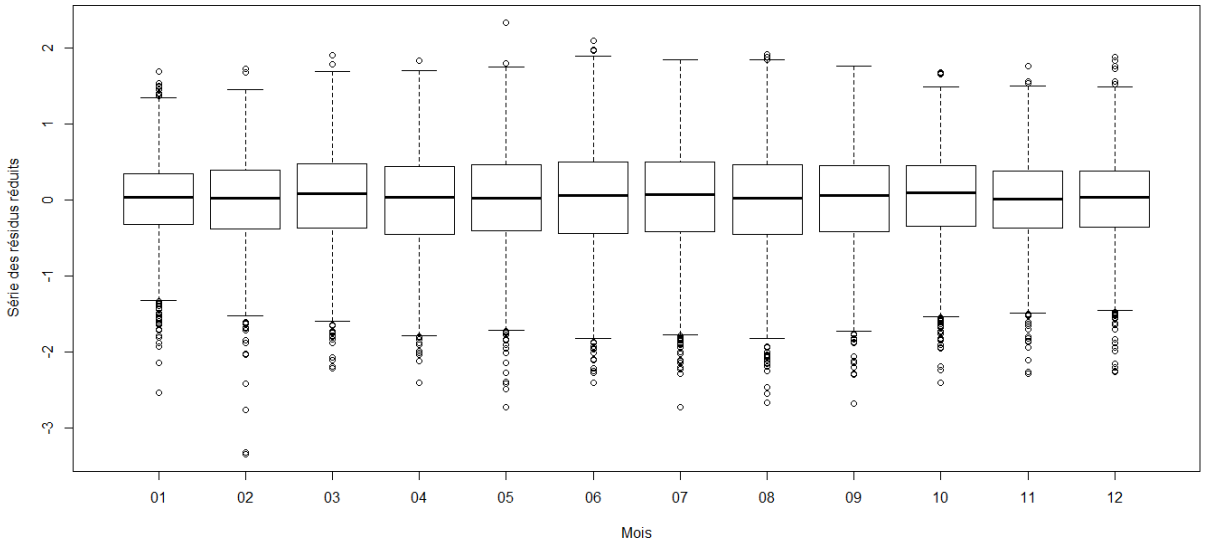


Figure 37 - Boîte à moustaches des résidus ε_t

On observe une régularité dans la dispersion de ces résidus. Cela tend à affirmer ces résidus sont indépendants et identiquement distribués, et que le processus (Y_t) est stationnaire. Cela est conforté par le test KPSS qui valide cette hypothèse. De plus, le test de Ljung-Box valide l'hypothèse que les résidus sont indépendants et que $(\varepsilon_t)_{t\in\mathbb{N}}$ est bien un bruit blanc. Ces tests sont décrits en annexe D.

Ensuite, nous avons effectué un *Q-Q plot* avec la distribution d'une normale standard pour évaluer la pertinence de l'ajustement du modèle.

Le graphique suivant est le *Q-Q plot* de la distribution des résidus (ε_t) avec celle d'une normale standard.

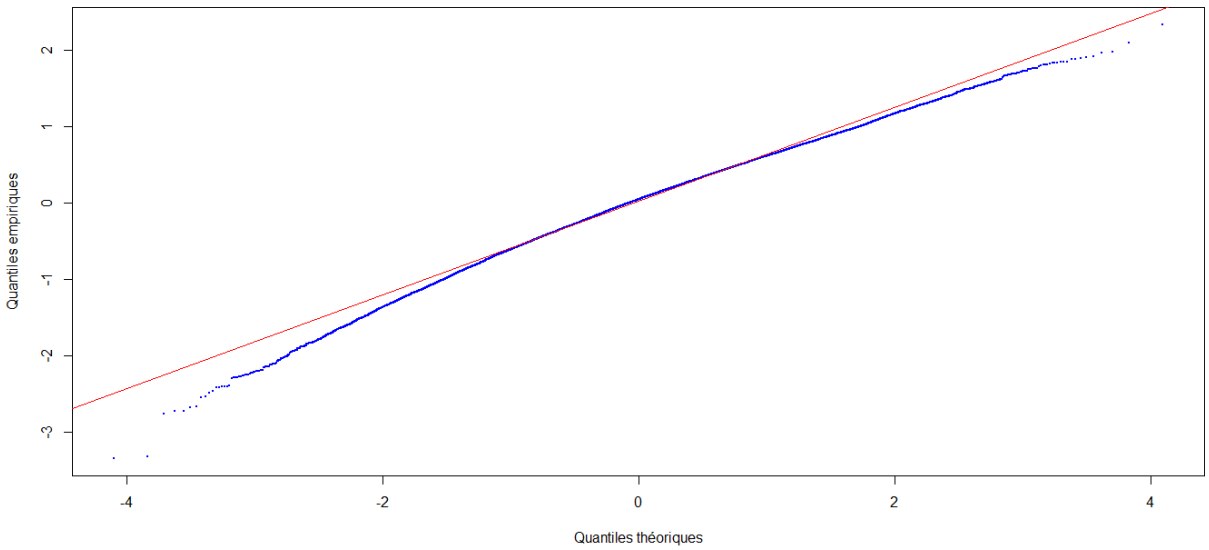


Figure 38 - *Q-Q plot* de la distribution des résidus (ε_t) avec celle d'une normale standard

On observe que les quantiles des résidus semblent en moyenne correspondre aux quantiles de la loi normale standard. Cela tend à valider l'hypothèse que les résidus ε_t suivent une loi normale standard.

En conclusion, le modèle ARMA obtenu semble adapté pour représenter l'évolution des températures maximales quotidiennes en un point donné.

3. Validation du modèle

Afin de valider le modèle, nous allons procéder de la manière suivante :

- On construit le modèle à partir des données jusqu'à fin 2013.
- On simule les évolutions de température maximale journalière pour chaque région.
- On compare le comportement global de nos simulations avec les vraies données de 2014.

Le graphique suivant montre la tendance des simulations (en rouge), et donne la région où se concentrent 98 % des simulations (délimitée par les courbes bleues) :

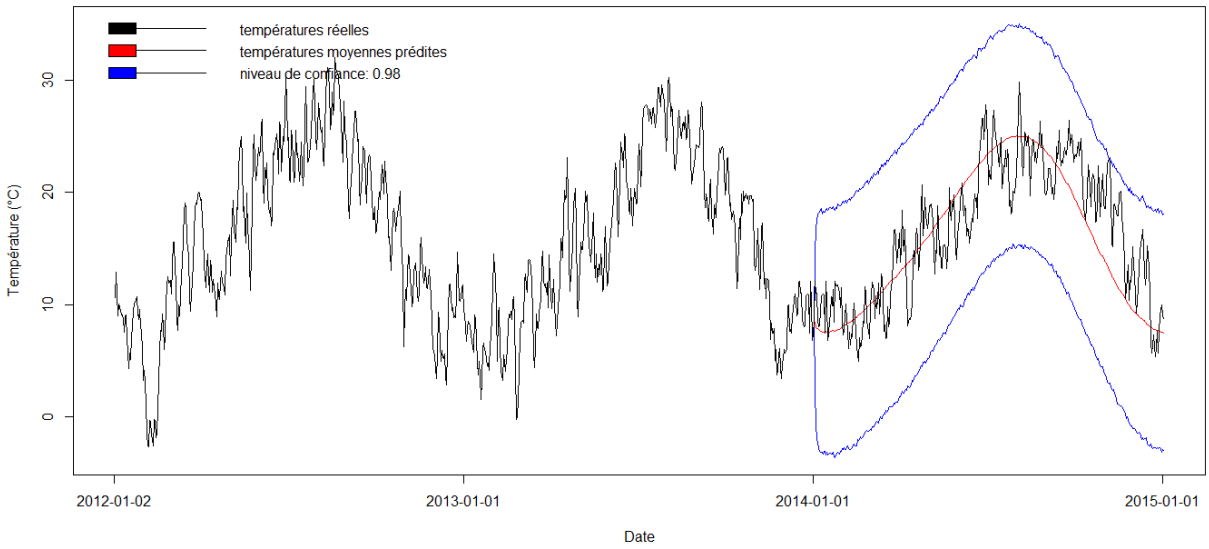


Figure 39 - *Backtesting* des températures dans la région Sud

On observe que la tendance des simulations est conforme à celle des températures réelles, et les pics de températures ont été captés par nos simulations.

De plus, nous avons étudié le comportement moyen de nos simulations au-dessus d'un seuil donné. Nous avons pris les seuils allant du quantile des températures de niveau 2 % à celui de niveau 98 %. Pour chaque seuil et pour chaque simulation, on calcule le nombre de fois que ce seuil a été dépassé dans l'année. On calcule ensuite la moyenne du nombre de dépassement par simulation, et on la compare par rapport au vrai nombre de dépassement observé.

Ensuite, pour chaque région, on pondère chaque erreur d'estimation de manière linéaire et décroissante avec le seuil : plus le seuil est faible, plus on prend en compte l'erreur de dépassement. Cela permet aussi de moins prendre en compte les erreurs liées aux seuils élevés, du fait qu'il y ait peu de jours où la température dépasse ces seuils.

Enfin, on pondère les erreurs globales de chaque région par la superficie de ces régions. On obtient finalement un taux d'erreur global de 5,9 %.

Cela nous permet de valider définitivement le modèle obtenu pour approximer le processus des températures.

E. Modélisation des dépendances entre régions

Dans un premier temps, la France est segmenté en cinq régions où chacune d'elles est homogène en termes de températures.

Les températures maximales journalières sont ensuite modélisées pour chacune de ces régions. Il est alors possible de simuler une évolution journalière de la température maximale pour chaque région.

Cependant, pour un jour donné, les cinq simulations obtenues ne doivent pas présenter trop d'incohérence : si un comportement extrême est observé dans une certaine région, il y a de forte chance d'observer ce même genre de comportement dans les autres régions.

Il est donc nécessaire de modéliser la dépendance entre régions en termes de températures. La théorie des copules est alors utilisée.

1. La théorie des copules

Soit $X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ un vecteur aléatoire. Soit F sa fonction de répartition.

On suppose que les X_i ont une densité. On note F_i la fonction de répartition de X_i.

La copule associée à X est défini de la manière suivante :

$$\tilde{X} = \begin{pmatrix} \tilde{X}_1 \\ \vdots \\ \tilde{X}_k \end{pmatrix} = \begin{pmatrix} F_1(X_1) \\ \vdots \\ F_k(X_k) \end{pmatrix}$$

La loi de la copule est alors donnée par :

$$\begin{aligned} C(t_1, \dots, t_k) &= \mathbb{P}\left(\tilde{X}_1 \leq t_1, \dots, \tilde{X}_k \leq t_k\right) \qquad \text{avec } t_i \in [0,1] \\ &= \mathbb{P}(F_1(X_1) \leq t_1, \dots, F_k(X_k) \leq t_k) \end{aligned}$$

où $\forall i \in \{1, \dots, k\}, \quad t_i \in [0,1]$

On peut montrer que :

$$\forall i \in \{1, \dots, k\}, \quad F_i(X_i) \sim Unif([0,1])$$

On connaît donc la loi de chaque composante $F_i(X_i)$, mais pas celle du vecteur aléatoire $\tilde{X}$.

La fonction C donne alors la structure de dépendance entre les variables X_i. Elle permet de donner la loi de X en fonction des lois des X_i, de la manière suivante (théorème de Sklar) :

$$F(x_1, \dots, x_k) = C(F_1(x_1), \dots, F_k(x_k))$$

Deux types de copules sont généralement utilisés : les copules elliptiques et les copules archimédiennes.

a) Les copules elliptiques

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

$\tilde{X}$ est une copule gaussienne de matrice de corrélation Σ si et seulement si sa loi s'écrit :

$$\forall(t_1, \dots, t_k) \in [0,1]^k, \qquad C_\Sigma(t_1, \dots, t_k) = \phi_\Sigma\left(\phi^{-1}(t_1), \dots, \phi^{-1}(t_k)\right)$$

où ϕ est la fonction de répartition d'une loi normale $\mathcal{N}(0,1)$

ϕ_Σ est la fonction de répartition d'une loi normale multivariée $\mathcal{N}_k(0_k, \Sigma)$

$\tilde{X}$ est une copule de Student de matrice de corrélation Σ si et seulement si sa loi s'écrit :

$$\forall(t_1, \dots, t_k) \in [0,1]^k, \qquad C_{\tilde{X}}(t_1, \dots, t_k) = T_{\Sigma,\mathrm{d}}\left(T_d^{-1}(t_1), \dots, T_d^{-1}(t_k)\right)$$

où T_d est la fonction de répartition d'une loi de Student à d degrés de liberté.

$T_{\Sigma,\mathrm{d}}$ est la fonction de répartition d'une loi de Student multivariée à d degrés de liberté.

b) Les copules archimédiennes

$\tilde{X}$ est une copule archimédienne si et seulement si sa loi s'écrit :

$$\forall(t_1, \dots, t_k) \in [0,1]^k, \qquad C_{\tilde{X}}(t_1, \dots, t_k) = \varphi^{-1}\left(\sum_{i=1}^{k} \varphi(t_i)\right)$$

où $\varphi :]0,1[\to \mathbb{R}_+$ est une fonction continue, convexe et strictement décroissante, tel que :

$$\begin{cases} \varphi(t) \underset{t\to 0}{\longrightarrow} +\infty \\ \varphi(t) \underset{t\to 1}{\longrightarrow} 0 \end{cases}$$

Nous nous intéresserons aux copules de Franck et de Gumbel.

<u>*Copule de Franck*</u> :

$$\varphi(t) = -\log\left(\frac{e^{at} - 1}{e^a - 1}\right), \quad a > 0$$

Cela implique que :

$$C(t_1, \dots, t_k) = -\frac{1}{a}\log\left(\frac{1 - \prod_{i=1}^{k}(e^{-at_i} - 1)}{(e^{-a} - 1)^{a-1}}\right)$$

<u>*Copule de Gumbel*</u> :

$$\varphi(t) = (-\log(t))^a, \quad a > 1$$

Cela implique que :

$$C(t_1, \dots, t_k) = e^{-\left(\sum_{i=1}^{k} -\log(t_i)^a\right)^{1/a}}$$

2. Construction des copules paramétrées

Notre objectif est d'être capable de simuler les températures au sein de chaque région en prenant en compte les dépendances.

La seule partie aléatoire des séries de températures est la partie résiduelle réduite[28]. Il faut donc étudier :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix}$$

où Y_i représente les résidus de températures de la i-ème région.

Nous pouvons donner la copule empirique, construite à partir des fonctions de répartition empiriques de chaque composante :

$$\tilde{Y} = \begin{pmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_k \end{pmatrix} = \begin{pmatrix} \hat{F}_1(Y_1) \\ \vdots \\ \hat{F}_k(Y_k) \end{pmatrix}$$

où $\hat{F}_i$ est la fonction de répartition empirique des résidus de la i-ème région.

Nous voulons ajuster une copule particulière (elliptique ou archimédienne) de sorte à ce qu'elle corresponde au mieux à la copule empirique. Pour cela, on calcule la matrice de corrélation du vecteur aléatoire $\tilde{Y}$, puis on choisit les paramètres de la copule à ajuster de sorte à conserver cette corrélation. Par exemple, dans le cas de la copule gaussienne, le seul paramètre à connaître est une matrice de corrélation. Le paramètre utilisé est donc la matrice de corrélation empirique.

3. Critère pour choisir la meilleure copule

Deux critères permettent de conserver la copule la plus représentative de la dispersion des données étudiées : l'interprétation graphique et la fonction de Kendall.

a) Interprétation graphique

Il existe deux manières de visualiser la répartition d'une copule à deux dimensions :

- On peut directement représenter toutes les réalisations de la copule.
- On peut représenter les quantiles associés.

Dans la suite, nous utiliserons la deuxième manière pour une vision plus concrète.

La copule retenue est celle dont la représentation se superpose le mieux avec celle de la copule empirique.

[28] La série (Y_t) : on l'appelle dorénavant « résidus »

b) Fonction de Kendall

Soit $\tilde{X} = \begin{pmatrix} \tilde{X}_1 \\ \vdots \\ \tilde{X}_k \end{pmatrix}$ une copule.

La fonction de Kendall K est définie de la manière suivante :

$$K(t) = \mathbb{P}\big(C\big(\tilde{X}_1, \dots, \tilde{X}_k\big) \leq t\big), \quad t \in [0,1]$$

Supposons que l'on dispose de n réalisations de la copule $\tilde{X}$. Soit $\tilde{x}_i^{(u)}$ la u-ième réalisation de la i-ème composante.

Pour construire la fonction de Kendall associée, on procède de la manière suivante :

- Pour chaque réalisation, on estime C par :

$$\hat{C}\left(\tilde{x}_1^{(j)}, \dots, \tilde{x}_k^{(j)}\right) = \frac{1}{n}\sum_{u=1}^{n} \mathbb{I}_{\tilde{x}_1^{(u)} \leq \tilde{x}_1^{(j)}, \dots, \tilde{x}_k^{(u)} \leq \tilde{x}_k^{(j)}}$$

- Puis on estime la fonction de Kendall par :

$$\widehat{K}(t) = \mathbb{P}\big(\hat{C}\big(\tilde{X}_1, \dots, \tilde{X}_k\big) \leq t\big) = \frac{1}{n}\sum_{u=1}^{n} \mathbb{I}_{\hat{c}\left(\tilde{x}_1^{(u)}, \dots, \tilde{x}_k^{(u)}\right) \leq t}$$

Nous allons alors comparer la fonction de Kendall empirique (celle construite avec les données initiales) avec la fonction de Kendall théorique (celle construite avec les données simulées par la copule paramétrée obtenue).

La copule retenue est celle dont la fonction de Kendall associée se rapproche le plus de la fonction de Kendall empirique.

4. Application

Le graphique suivant montre la dispersion des résidus d'une région par rapport à une autre : il s'agit de la représentation graphique des quantiles associés aux copules empiriques.

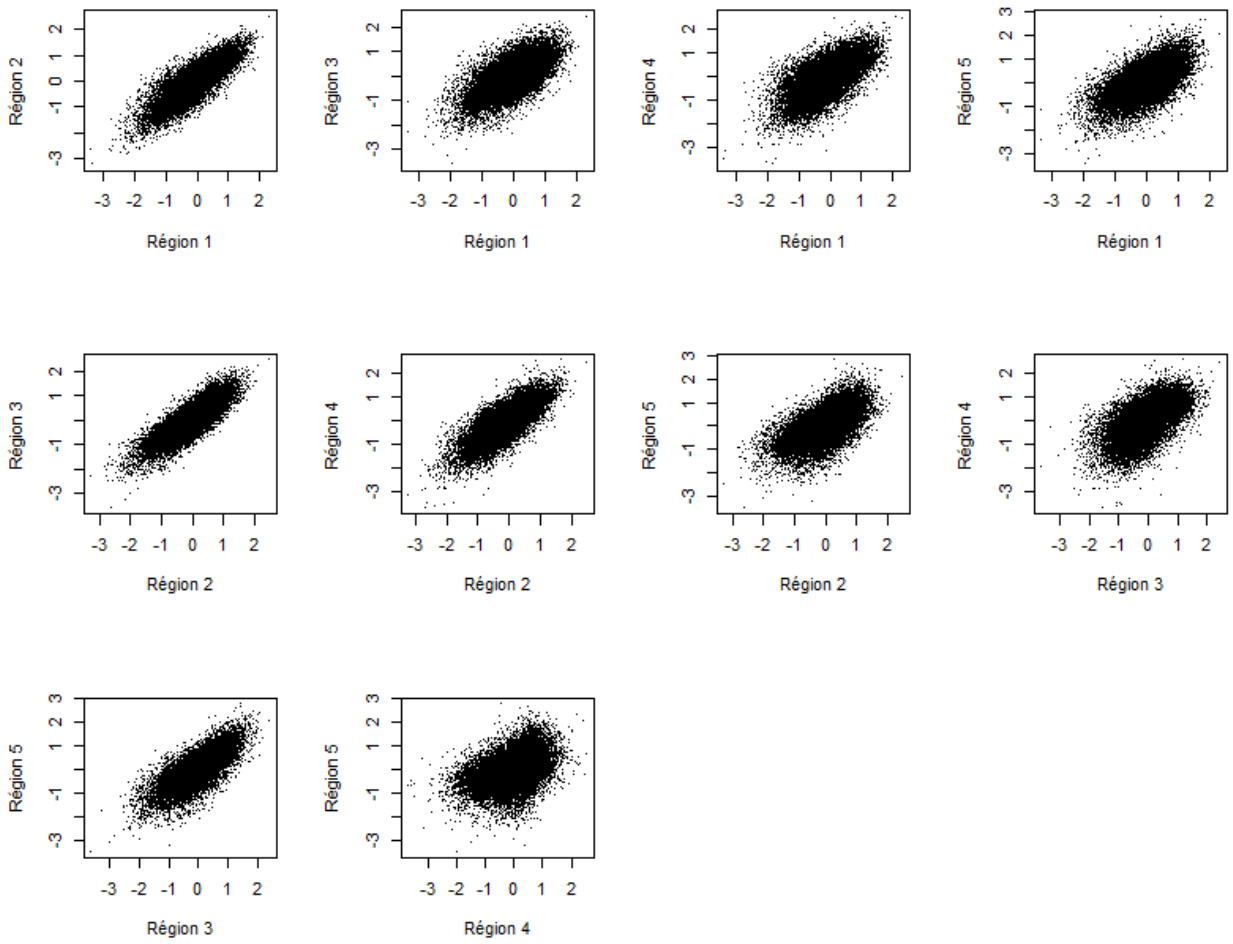


Figure 40 - Copules empiriques entre régions

La copule empirique semble avoir une distribution multivariée symétrique. Les copules elliptiques s'adapteront bien à ce type de dépendance.

Le tableau ci-après donne les corrélations empiriques des résidus entre régions.

	Région 1	Région 2	Région 3	Région 4	Région 5
Région 1	1,000	0,836	0,664	0,668	0,642
Région 2		1,000	0,846	0,798	0,632
Région 3			1,000	0,608	0,774
Région 4				1,000	0,443
Région 5					1,000

Tableau 2 - Corrélations empiriques des résidus entre régions

Pour pouvoir simuler des bruits blancs avec la même structure de dépendance représentée ci-dessus, il faut comparer les copules empiriques avec des copules paramétrées par la matrice de corrélations empiriques : nous étudions les copules gaussiennes et de Student.

Le graphique suivant compare les réalisations avec la copule empirique et les réalisations avec les copules Gaussiennes et de Student, entre le sud de la France (Région 1) et les zones montagneuses (Région 2).

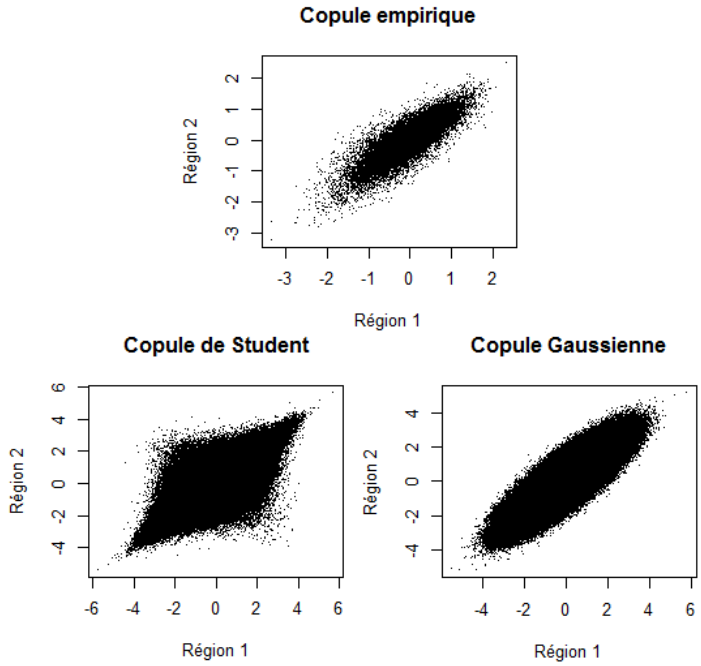


Figure 41 - Comparaison de la copule empirique avec les copules gaussiennes et de Student paramétrées

La copule empirique semble mieux se superposer avec la copule Gaussienne que la copule de Student.

Pour valider ce résultat, on va étudier les fonctions de Kendall théoriques associées à chaque copule candidate, par rapport à la fonction de Kendall empirique.

Le graphique suivant présente les résultats pour différentes copules (elliptiques et archimédiennes) :

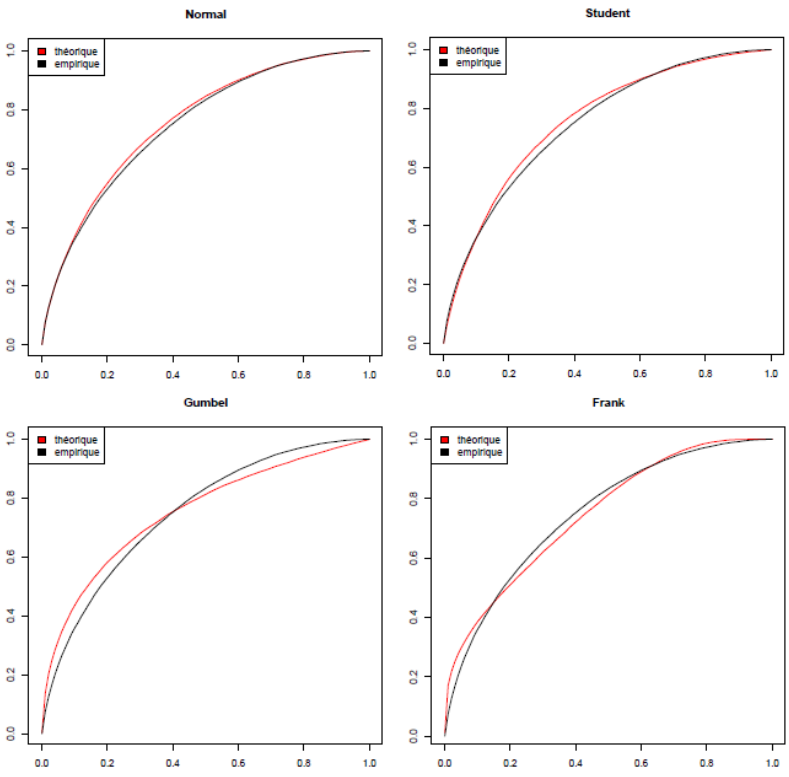


Figure 42 - Fonctions de Kendall empiriques et théoriques pour différentes copules

Les résultats sont similaires pour les autres régions (voir annexe B).

Pour l'ensemble des régions, la copule gaussienne semble donc être la plus adaptée. Des bruits blancs seront simulés avec cette structure de dépendance pour représenter les résidus Y_t. Ainsi, avec la tendance, la saisonnalité, et les résidus, nous pourrons effectuer autant de simulations que nous le souhaitons : nous allons générer 10 000 scénarios d'évolutions de la température maximale journalière.

F. Simulations

La sinistralité causée par la sécheresse est décrite par des variables explicatives construites à partir des précipitations mensuelles et des températures maximales journalières.

Un modèle pour les précipitations mensuelles et un modèle pour les températures maximales journalières ont ensuite été développés.

Cela permet d'effectuer de simuler 10 000 scénarios d'évolution mensuelle des précipitations et d'évolution journalière des températures maximales.

Ces simulations de précipitations et de températures permettent d'en déduire des simulations de l'ensemble des variables explicatives de la sécheresse.

1. Simulations des précipitations et des températures

a) Précipitations mensuelles

Le modèle des précipitations mensuelles précédemment construit permet de simuler une précipitation cumulée pour chaque mois de l'année et pour chaque *CRESTA*.

Une loi Gamma est alors utilisée : les paramètres dépendent du mois et du *CRESTA* concerné.

Ainsi, nous obtenons 10 000 scénarios d'évolution mensuelle des précipitations pour chaque *CRESTA*.

b) Températures maximales journalières

Le modèle des températures maximales journalières précédemment construit permet de simuler une température maximale pour un jour donné et un *CRESTA* donné.

La France a été initialement segmentée en 5 régions. Des simulations sont effectuées pour chacune de ces régions. Ainsi, les simulations de températures dans un *CRESTA* correspondront aux simulations de la région à laquelle il appartient.

Pour un jour donné, nous procédons de la manière suivante :

- On simule simultanément 5 bruits blancs gaussiens dont les dépendances sont structurées par la copule gaussienne paramétrée par la matrice de corrélation empirique.
- Pour chaque région, on applique le modèle ARMA correspondant pour obtenir une valeur de température à partir des valeurs prises précédemment (températures et bruits).

Le graphique suivant présente un exemple de scénario :

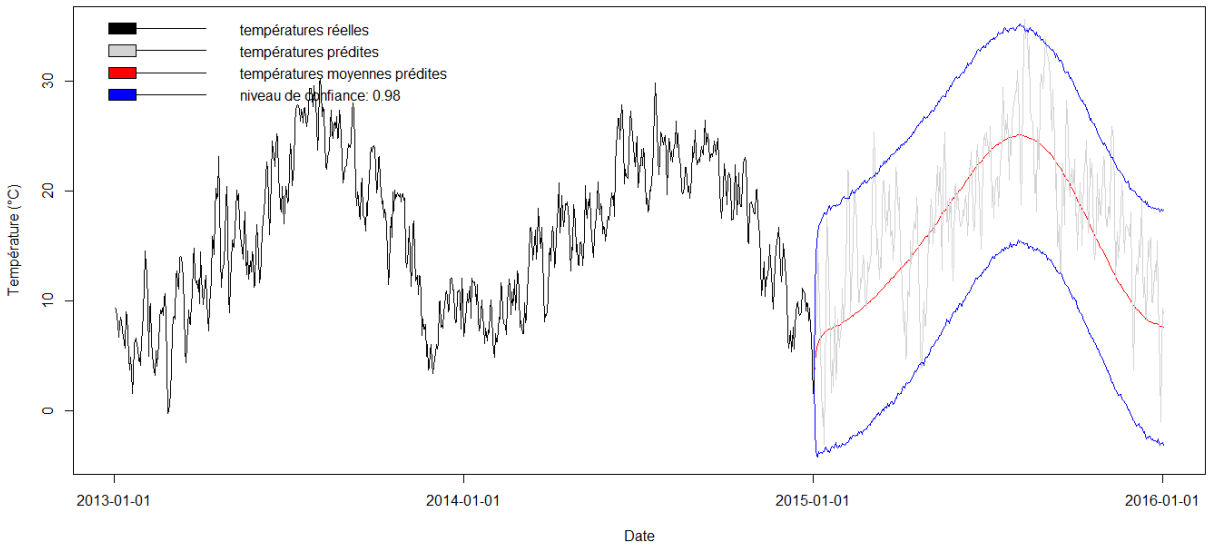


Figure 43 - Exemple de scénario de température

2. Simulations des variables explicatives

Les variables explicatives de la sécheresse sont toutes construites à partir des précipitations et des températures. Pour chaque *CRESTA,* les scénarios d'évolution mensuelle des précipitations et de températures peuvent donc être directement traduits en scénario d'évolution mensuelle des variables explicatives de la sécheresse, comme le montre le schéma suivant :

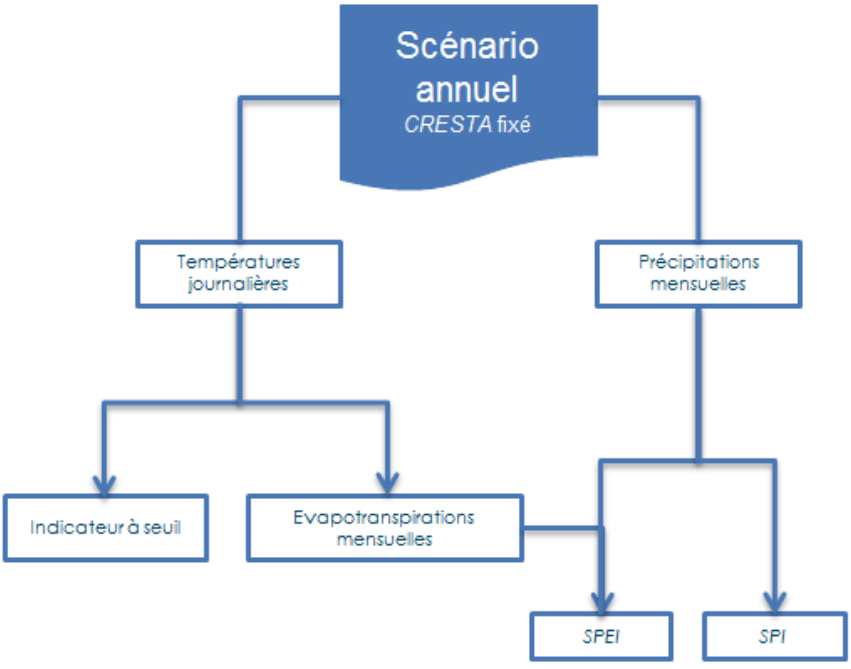


Figure 44 - Un scénario annuel des variables explicatives de la sécheresse

Cette opération est répétée 10 000 fois. Ainsi, le module Aléa a généré un catalogue de 10 000 scénarios météorologiques pour chaque *CRESTA* en France.

III. Module Vulnérabilité et Financier : simulations des pertes financières causées par la sécheresse

Le module précédent a permis d'obtenir 10 000 simulations de l'évolution mensuelle des variables explicatives de la sécheresse.

Il faut traduire ces scénarios d'aléa physique en scénarios de pertes assurantiels. On va donc chercher à quantifier le lien entre ces variables explicatives et la fréquence de sinistralité[29] liée à la sécheresse, afin de pouvoir établir des prédictions de sinistralité pour l'année à venir.

Pour cela, nous proposons deux types de modèle : un modèle linéaire généralisé (*GLM*[30] en anglais) qui est traditionnellement utilisé en assurance, et un modèle de régression plus récent construit à partir d'une « forêt aléatoire » qui effectue un apprentissage sur de multiples arbres binaires de décision entraînés sur des sous-ensembles de données légèrement différents.

Enfin, la perte financière générée par les sinistres vont être modélisés en les supposant indépendants de l'aléa physique (et donc de la fréquence de sinistralité). En effet, le coût moyen d'un sinistre pour un événement sécheresse ne semble pas corrélé à l'intensité de cet événement. Le graphique suivant représente, pour un mois donné, le coût moyen d'un sinistre et le nombre associé de sinistres (représentant l'intensité).

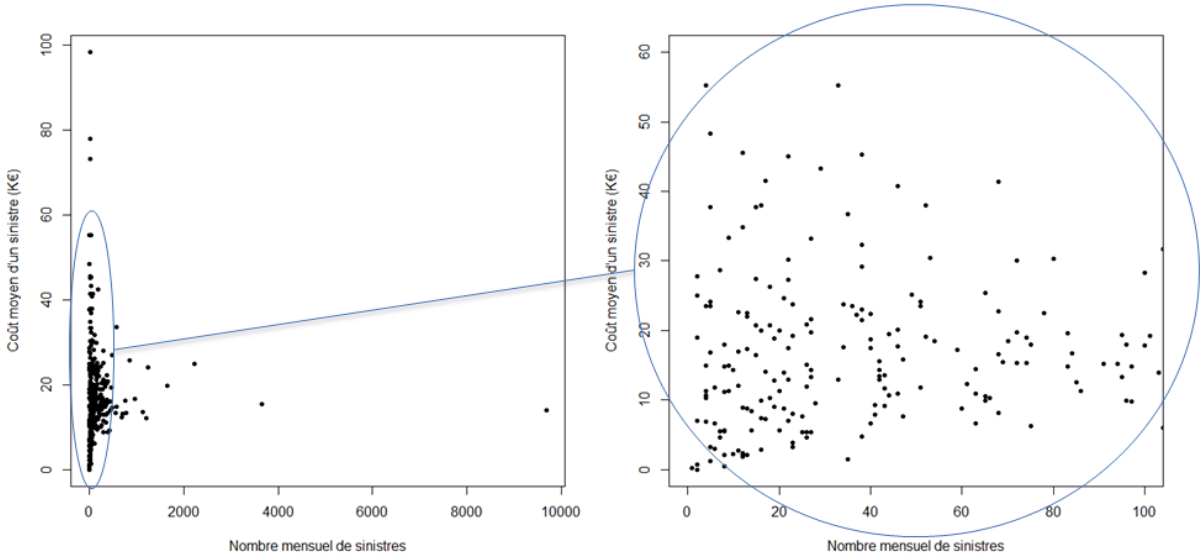


Figure 45 - Indépendance fréquence/coût

Aucune structure de dépendance ne se dégage. La modélisation des coûts de sinistralité peut s'effectuer de manière indépendante de celle de la fréquence de sinistralité.

Pour cela, nous allons utiliser une méthode récente d'analyse du risque en réassurance : la méthode MBBEFD.

[29] fréquence de sinistralité $= \frac{\text{Nombre mensuel de polices affectées par la sécheresse}}{\text{Nombre total de polices}}$, pour un $CRESTA$ donné

[30] *GLM : Generalized Linear Model*

A. Modélisation de la fréquence de sinistralité avec les modèles linéaires généralisés

1. Le modèle linéaire généralisé

Le modèle linéaire généralisé (ou *GLM* en anglais) est un modèle qui étudie la liaison entre une variable Y, dite « réponse », et un ensemble de variables $X_1, X_2, \dots, X_k$, dites « explicatives ». On cherche à expliquer les montants de sinistres par un ensemble de variables explicatives. Dans notre problématique, les variables explicatives seront la typologie des sols de la zone concernée, la précipitation cumulée mensuelle, etc. La variable réponse sera le nombre mensuel de sinistres enregistré dans la zone concernée.

Ce modèle est très répandu en assurance pour connaître les risques et aider à la tarification. Il vise à déterminer la loi de probabilité de la variable réponse sachant qu'on connait la valeur des variables explicatives. Il s'agit donc d'estimer la distribution de $Y \mid (X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$.

On va supposer connaître son type de loi, mais dont le paramètre dépend directement des variables explicatives. Cependant, sa loi doit appartenir à une certaine classe de lois : la famille exponentielle. Cette dernière regroupe les principales lois usuelles telles que la loi Normale, la loi Gamma, la loi de Poisson ou la loi Binomiale négative. Chacune des lois de la famille exponentielle possède un paramètre, dit « naturel ». Le modèle linéaire généralisé met en lien ce paramètre naturel de la loi avec une combinaison linéaire des variables explicatives. On va donc chercher à optimiser les coefficients de cette combinaison linéaire de sorte que la vraisemblance de $Y \mid (X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$ soit maximale avec le paramètre calculé à partir de cette combinaison linéaire.

Il faut donc définir 4 éléments qui composent le modèle : la variable réponse, les variables explicatives, la fonction de lien entre les variables explicatives et le paramètre naturel de la loi conditionnelle de la variable réponse, et les critères quantifiant la qualité du modèle.

a) La variable réponse

La variable réponse Y est la partie aléatoire du modèle. On souhaite être capable de la prédire. On dispose pour cela de n réalisations de Y. On suppose que sa loi appartient à la famille exponentielle.

Une variable aléatoire Y appartient à la famille exponentielle si sa densité (ou mesure de probabilité dans le cas discret) peut s'écrire sous la forme :

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

où $a, b\ et\ c$ sont des fonctions et θ, ϕ les paramètres.

On peut montrer que :

$$\mathbb{E}(Y) = b'(\theta)$$
$$\mathbb{V}(Y) = b''(\theta)\,\phi$$

θ est appelé paramètre naturel, et ϕ est appelé paramètre de dispersion.

Parmi les lois qui appartiennent à la famille exponentielle, on peut citer :

- La loi normale $\mathcal{N}(\mu, \sigma^2)$ avec :
$$\theta = \mu, \phi = \sigma^2, a(\phi) = \phi, b(\theta) = \frac{\theta^2}{2}, c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\Pi\phi)\right) \ \forall y \in \mathbb{R}.$$
- La loi de Poisson $\mathcal{P}(\lambda)$ avec :
$$\theta = \log(\lambda), \phi = 1, a(\phi) = 1, b(\theta) = \exp(\theta) = \lambda, c(y, \phi) = -\log(y) \ \forall y \in \mathbb{N}$$

b) Les variables explicatives

Il est fondamental de choisir des variables capables d'expliquer, du moins en partie, le phénomène observé pour pouvoir établir de bonnes prévisions. En effet, même si un modèle semble avoir été capable d'ajuster correctement les valeurs de la variable réponse avec celles de variables peu explicatives, de nouvelles valeurs de ces variables n'arriveront pas nécessairement à expliquer le phénomène et donc à faire de bonnes prédictions. Dans la même logique, il ne faut pas utiliser trop de variables pour que le modèle ne soit pas « sur-ajusté ». Cependant, il faut qu'il y en ait suffisamment pour que le modèle soit convenable.

c) La fonction de lien

Pour des réalisations de $(X_1, \dots, X_k)$, la meilleure prévision de la variable réponse est
$$\mathbb{E}(Y \mid X_1 = x_1, X_2 = x_2, \dots, X_k = x_k).$$

On souhaite donc établir un lien entre les réalisations des variables explicatives et le comportement de la variable réponse conditionnée par ces réalisations. Une fonction de lien inversible g représente ce lien de la manière suivante :
$$g(\mathbb{E}(Y \mid (X_1, \dots, X_k)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$
où $(\beta_0, \beta_1, \dots, \beta_k) \in \mathbb{R}^k$.

Comme g inversible, cela implique que :
$$\theta = b'\left(g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)\right)$$

Différentes fonctions de lien peuvent être utilisées :

- La fonction canonique : $g(x) = b'^{-1}(x)$. Cela implique que
$\theta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- La fonction identité : $g(x) = x$. Cela implique que
$$\mathbb{E}(Y \mid (X_1, \dots, X_k)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$
- La fonction log : $g(x) = \log(x)$. Cela implique que
$$\mathbb{E}(Y \mid (X_1, \dots, X_k)) = \exp(\beta_0) \times \exp(\beta_1 X_1) \times \dots \times \exp(\beta_k X_k)$$
- La fonction $g(x) = \log\left(\frac{x}{x+1/k}\right)$ où k représente le paramètre de la distribution binomiale négative.

Nous souhaitons donc optimiser les coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ de sorte à maximiser la vraisemblance du modèle par l'intermédiaire de la fonction de lien g choisie.

Nous avons à disposition un vecteur de n observations $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, et une matrice de n lignes correspondant aux valeurs prises par nos k variables explicatives $X_{obs} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_{1.} \\ \vdots \\ x_{n.} \end{pmatrix}$.

Un théorème nous énonce que la valeur du vecteur des coefficients qui maximise la vraisemblance peut s'approcher de la manière suivante :

$$\beta^{k+1} = (X_{obs}{}' W^k X_{obs})^{-1} X_{obs}{}' W^k z^k$$

où $\begin{cases} W^k = diag\left(\frac{(g^{-1})'(x_i'\beta^k)}{b''(\theta_i)}\right)_{i=1\dots n} \\ z^k = X_{obs}\beta^k + diag\left(\frac{1}{(g^{-1})'(x_i'\beta^k)}\right)_{i=1\dots n}(y-\mu) \end{cases}$ avec $\mu = \left(b'\left(g^{-1}(x_{i.}'\beta^k)\right)\right)_{i=1\dots n}$

Ainsi, la solution de ce problème $\hat{\beta}$ permet de prédire la variable réponse Y à partir de nouvelles valeurs des variables explicatives $(x_1, x_2, \dots, x_k)$:

$$\hat{y} = g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right)$$

d) Les critères quantifiant la qualité du modèle

Afin de mesurer la qualité du modèle, nous pouvons calculer ses « déviances ». La déviance Q d'un modèle se définit par :

$$Q(y_i, \hat{y}_i) = \sum_{i=1}^{n} 2(l(y_i) - l(\hat{y}_i))$$

Il est alors intéressant de calculer le ratio entre la déviance du modèle et la déviance originale, c'est-à-dire formellement :

$$\dot{R}^2 = 1 - \frac{Q(y_i, \hat{y}_i)}{Q\left(y_i, \overline{y}_i\right)}$$

où $\overline{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $\hat{y}_i$ l'estimation faite par le modèle avec le vecteur des variables explicatives associé.

On cherche à maximiser ce rapport. En pratique, lorsque ce rapport dépasse de 70 %, le modèle est considéré comme satisfaisant.

D'autres critères permettent de hiérarchiser les qualités de plusieurs modèles, comme le critère *AIC*[31]. Pour chaque modèle, on calcule le critère *AIC* :

$$AIC = 2k - 2\log(L)$$

où k est le nombre de variables explicatives, et L est le maximum de la fonction de vraisemblance du modèle.

Lorsque l'on rajoute une variable dans un modèle, la vraisemblance du modèle augmente, mais cela a pu ajouter du « bruit » que le modèle aura utilisé pour ajuster encore mieux (« sur-ajuster ») la variable réponse avec les variables d'entrée. On souhaite donc avoir simultanément un faible nombre de variables

[31] *AIC : Akaike Information Criterion*

explicatives (k le plus petit possible), et une vraisemblance du modèle la plus élevée possible (L le plus grand possible). On souhaite donc minimiser le critère *AIC*.

Pour un ensemble de modèles candidats, le modèle choisi est celui qui aura la plus faible valeur *d'AIC*.

2. Modélisation sans dépassement de seuil

a) Hypothèses

Nous allons ajuster un modèle linéaire généralisé à nos données. Il faut donc définir la variable réponse, les variables explicatives et la fonction de lien.

Chaque ligne de notre tableau de données correspond à un **mois**, une **année**, et une **localisation géographique**[32]. Nous avons alors à notre disposition les variables suivantes :

- Le nombre total de sinistres enregistrés : $NbreSin$
- Le niveau moyen d' « aléa » (mesurant l'intensité du phénomène retrait-gonflement des argiles) pondéré par l'aire du *CRESTA* concerné : $nALEA$
- La température maximale mensuelle : $Tmax$
- La classe de température auquel appartient le *CRESTA* concerné : $ClassTemp$
- Les précipitations cumulées du mois concerné : $SumPrecip$
- Les valeurs des indicateurs *SPI* et *SPEI* calculés à partir de la précipitation cumulée mensuelle et de l'évapotranspiration mensuel déduite des températures du mois concerné : SPI et $SPEI$.

Pour des valeurs fixées de ces variables explicatives, le modèle donnera toujours le même nombre de sinistres que l'on soit dans un *CRESTA* contenant peu ou beaucoup de contrats. Puisque le nombre de contrats varie selon le *CRESTA* concerné, il faut relativiser les valeurs prises par les variables explicatives en fonction de l'exposition[33]. Pour cela, nous pouvons introduire une variable *offset* dans le modèle, appelée $Expo$, qui normalisera nos variables explicatives en fonction de l'exposition.

Classe de sinistralité

Les 96 *CRESTA* ne seront pas considérés directement comme variables d'entrée dans le modèle. En effet, l'objectif est d'avoir simultanément un minimum de variables d'entrée et un maximum de significativité. Nous allons alors agréger les *CRESTA* en plusieurs classes, appelées classes de sinistralité ou $ClassSin$, pour simplifier le modèle et avoir une plus grande vue d'ensemble sur le lien entre les variables explicatives de la sécheresse et la sinistralité.

Nous appliquerons la méthode de classification CAH, décrite précédemment, sur les données de fréquence de sinistralité enregistrées par AXA pour chaque *CRESTA*. On utilise la méthode de Ward avec la distance euclidienne.

Le dendrogramme obtenu est le suivant :

[32] Localisée par *CRESTA*

[33] Exposition : nombre de contrats d'assurance dans le *CRESTA* concerné

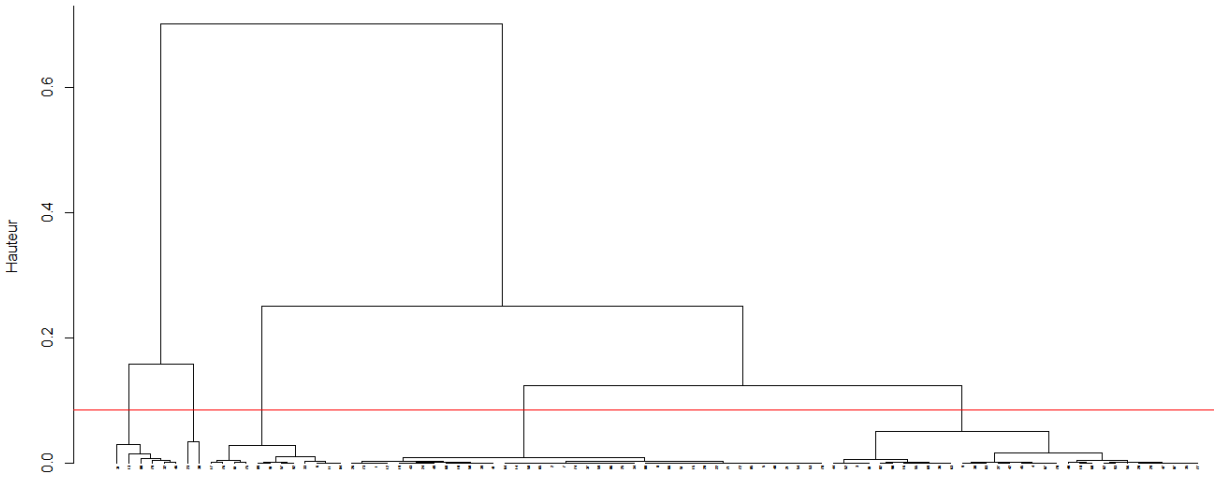


Figure 46 - Dendrogramme pour la fréquence de sinistralité

En-dessous du trait rouge, les hauteurs des sous-arbres deviennent beaucoup plus faibles : fixer 5 classes de sinistralité semble alors être un choix judicieux.

La carte suivante montre les résultats de la classification :

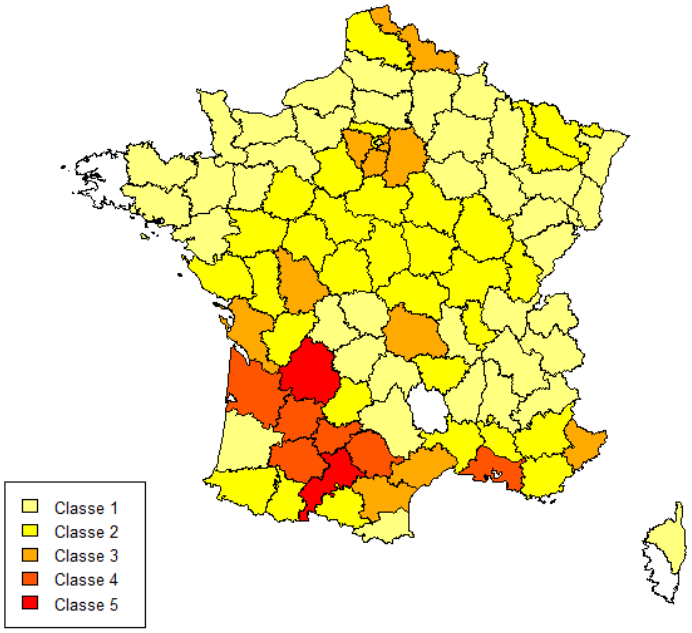


Figure 47 - Répartition des classes de sinistralité

Croisement des données avec les mois précédents

Le phénomène de sécheresse s'étendant dans le temps et les défauts de la base de données des sinistres n'étant pas négligeables, il est plus pertinent de croiser les sinistres d'un mois fixé avec les données des mois qui le précèdent. Cela est visible sur les graphiques présents en annexe C. Les valeurs retenues des variables explicatives seront celles observées les deux mois précédents.

Synthèse

Le modèle linéaire généralisé que nous allons ajuster à nos données, a donc les caractéristiques suivantes :

- Données : agrégées par mois, année, et *CRESTA*
- Variable réponse : $NbreSin$
- Variables explicatives : $ClassTemp, ClassSin, nALEA, Tmax, SumPrecip, SPI, SPEI$
  On observe les réalisations (de ces variables) des deux mois précédents.
- Offset : $Expo$
- Fonction de lien : $\log()$, correspondant au modèle de Poisson. Le modèle Binomiale Négative a été aussi testé, mais les résultats (présents en annexe E) ne sont pas plus performants.

Ainsi, on ajuste un modèle linéaire généralisé de manière optimale grâce à la fonction de lien $\log()$ et avec un certain paramètre β.

On peut alors estimer le nombre de sinistres pour le mois n et dans le *CRESTA* E, de la manière suivante :

$$\begin{aligned} \widehat{NbreSin}_n &= \mathbb{E}(NbreSin_n \mid X_n) \\ &= \log^{-1}(\beta.\ X_n) \\ &= e^{\beta.\ X_n} \end{aligned}$$

b) Résultats

Le graphique suivant présente le résultat de l'ajustement d'un modèle linéaire généralisé avec les hypothèses énoncées ci-dessus :

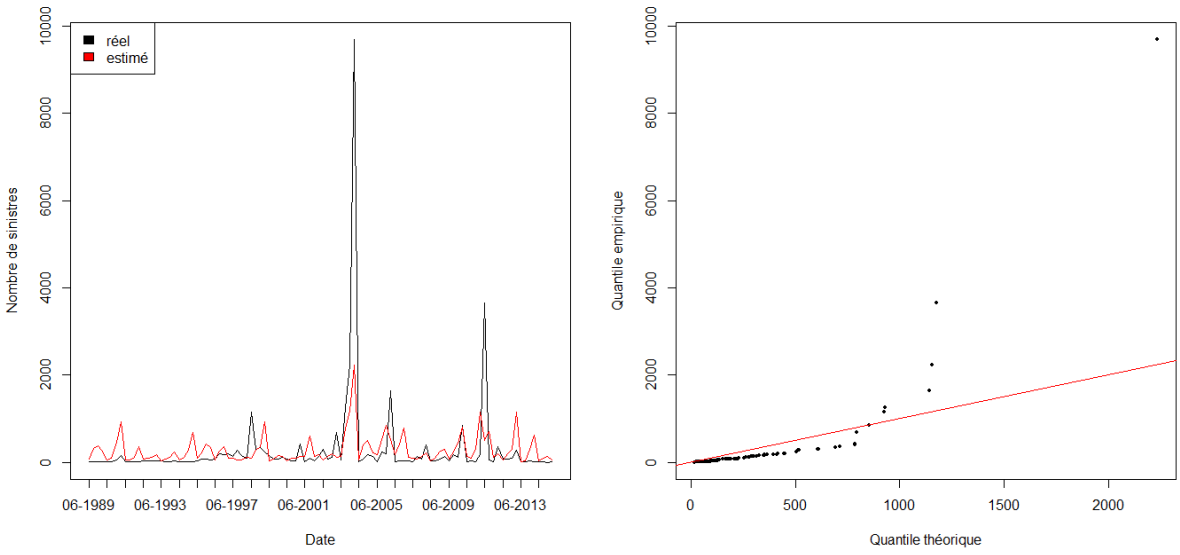


Figure 48 - Modèle de fréquence de sinistralité – Poisson

Le taux de déviance est de 51 %, ce qui est faible. Le critère *AIC* vaut 106 369. Les coefficients d'ajustement sont donnés en annexe E.

On observe que le modèle surestime la plupart des « petits sinistres », et sous-estime de loin les pics de sinistralité. Cela signifie que notre modèle ne prévoit quasiment jamais aucun sinistre, et n'est pas capable de s'adapter à des conditions extrêmes en prévoyant des pics de sinistralité.

Afin de pouvoir détecter un comportement particulier de l'ensemble des variables explicatives, et ainsi capter des conditions extrêmes, il peut être judicieux de développer des indicateurs à seuil. Ainsi, une valeur non nulle de cet indicateur indique qu'on a dépassé le seuil (défini par cet indicateur) et qu'on peut se trouver dans des conditions particulières. C'est ce que l'on va développer juste après.

3. Modélisation avec indicateur à seuil

a) Hypothèses

Dans la section précédente, nous avons vu que l'ajustement du modèle n'était pas satisfaisant : nous n'arrivions pas à capter (expliquer) les pics de sinistralité comme ceux de 2003 ou 2011.

En plus des variables explicatives du modèle précédent, nous allons développer un indicateur dépendant d'un seuil de température. Cela permettra d'amplifier l'intensité de sinistralité et donc de mieux capter les pics de sinistralité. A titre d'exemple, les relevés de températures de l'été 2003 en Haute-Garonne (fortement exposée à la sécheresse) montraient que les températures ont pu être supérieures à 30°C trente-sept fois en deux mois. Il peut être judicieux de prendre en compte ces comportements au-delà d'un seuil.

Soit s_0 un seuil de température fixée.
Soit $Tmax(t)$ la variable aléatoire donnant, à la date t, la température maximale enregistrée
Soit $Temp_{s_0}(t)$ la variable aléatoire donnant, à la date t, le nombre de jours où la température maximale journalière a été supérieure au seuil s_0 sur les 60 derniers jours.

On a alors :

$$Temp_{s_0}(t) = \sum_{0 \leq i \leq 60} \mathbb{I}_{Tmax(t) \geq s_0}$$

Nous allons alors chercher les seuils s_0 permettant à notre modèle de capter le mieux possible les pics de sinistralité.

Le graphique suivant met en lien, pour un *CRESTA* et un mois donnés, le nombre mensuel de sinistres enregistrés par AXA ($NbreSin$) avec les valeurs observées le mois dernier de l'indicateur à seuil de 30°C ($Temp_{30}$). Les indicateurs obtenus avec d'autres seuils sont présentés en annexe C.

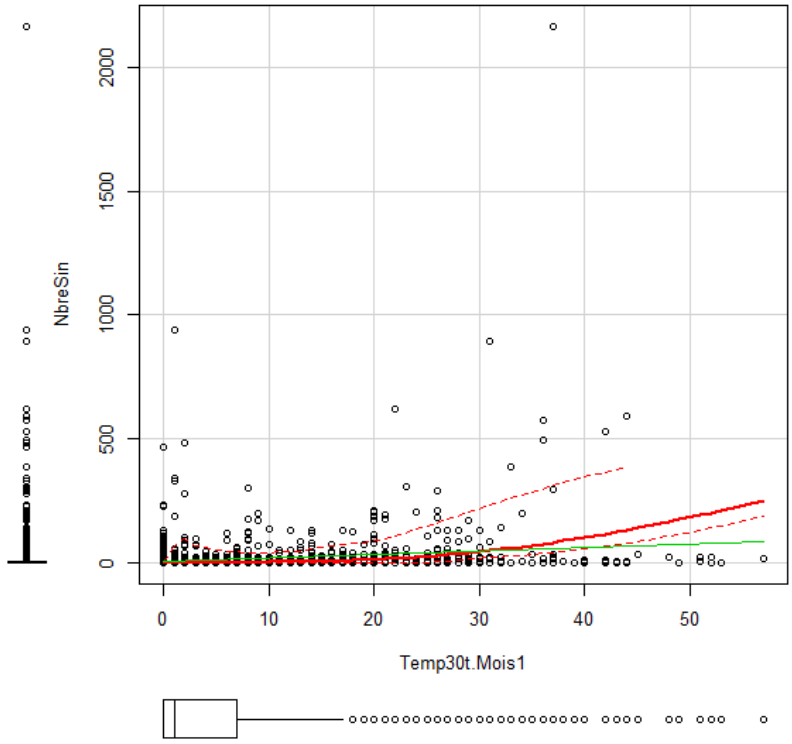


Figure 49 - Sinistralité et indicateur à seuil

Nous observons une nette corrélation entre la sinistralité et le comportement des indicateurs à seuil liés à la température. Nous allons donc reconstruire le modèle de la même manière que précédemment mais en rajoutant l'indicateur à seuil $Temp_{30}$[34].

Le modèle linéaire généralisé que nous allons ajuster à nos données, a donc les caractéristiques suivantes :

- Données : agrégées par mois, année, et *CRESTA*
- Variable réponse : $NbreSin$
- Variables explicatives : $ClassTemp, ClassSin, nALEA, Tmax, SumPrecip, SPI, SPEI, Temp_{30}$
  On observe les réalisations (de ces variables) des deux mois précédents.
- Offset : $Expo$
- Fonction de lien : $\log()$

b) Résultats

Le graphique suivant présente le résultat de l'ajustement d'un modèle linéaire généralisé avec les hypothèses énoncées ci-dessus :

[34] Comme $Tmax$, nous garderons la valeur maximale prise pour chacun de ces indicateurs pour chaque mois et chaque *CRESTA*.

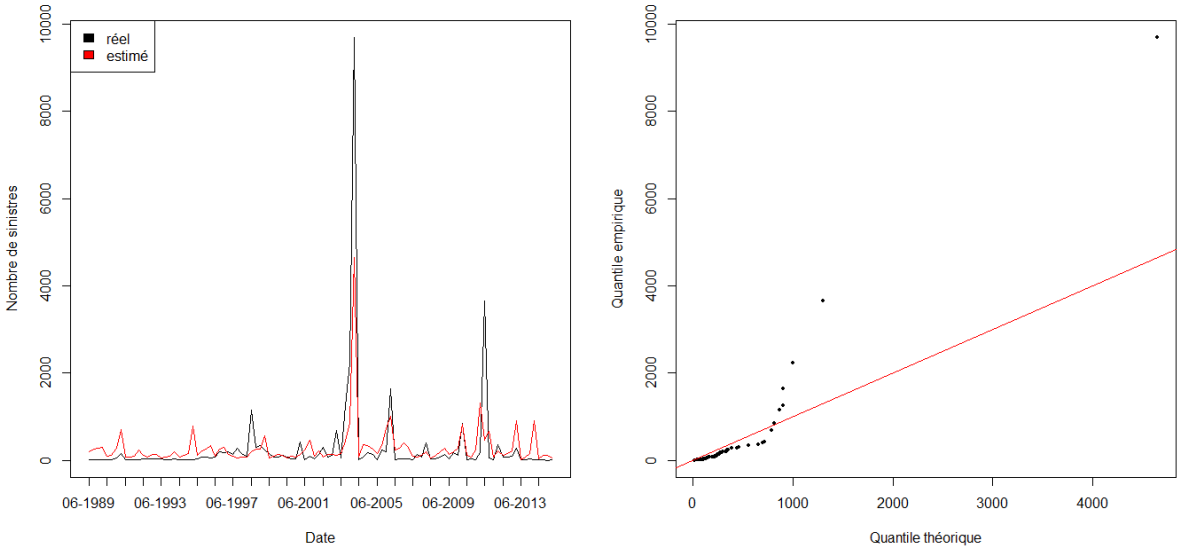


Figure 50 - Modèle de fréquence de sinistralité – Poisson avec indicateur à seuil

Le taux de déviance est de 55 %, ce qui est à peine plus. Le critère *AIC* vaut 98 890, ce qui est mieux que précédemment. Les coefficients d'ajustement sont donnés en annexe E.

Le modèle est à un peu plus performant que le précédent, mais les défauts d'ajustement restent essentiellement les mêmes : notre modèle a toujours du mal à capter les pics de sinistralité.

Notre modèle est additif : le « + » peut être traduit par « ou », et non un « et ». Cela signifie que ce modèle ne prend pas en compte le fait qu'on soit par exemple dans une classe de température particulière **et** avec des valeurs de températures particulières (pour la classe de température associée).

Nous souhaiterions alors pouvoir élargir le nombre de configurations particulières possibles en croisant les variables explicatives.

Pour cela, on va simplement remplacer les « + » par des « x » qui peuvent être traduits par « et ». Cela va permettre de recréer le modèle additif, mais où chaque variable explicative est une multiplication de certaines variables explicatives du précédent modèle.

On reprend donc la même variable réponse, les mêmes variables explicatives (dont l'indicateur à seuil), et la même fonction de lien. Nous croiserons les variables explicatives selon plusieurs catégories :

- Variables donnant un risque *a priori* : $ClassSin, nALEA$
- Variables physiques : $ClassTemp, Tmax, SumPrecip$
- Variables indicateurs : $Temp_{30}, SPEI, SPI$

Ainsi, pour chaque catégorie, on étudie toutes les combinaisons possibles de multiplications entre les variables. Toutes les variables explicatives étant présentes dans le modèle, le modèle croisé ne peut pas renvoyer des résultats pires que précédemment.

Le graphique suivant présente le résultat de l'ajustement d'un modèle linéaire généralisé avec les hypothèses énoncées ci-dessus :

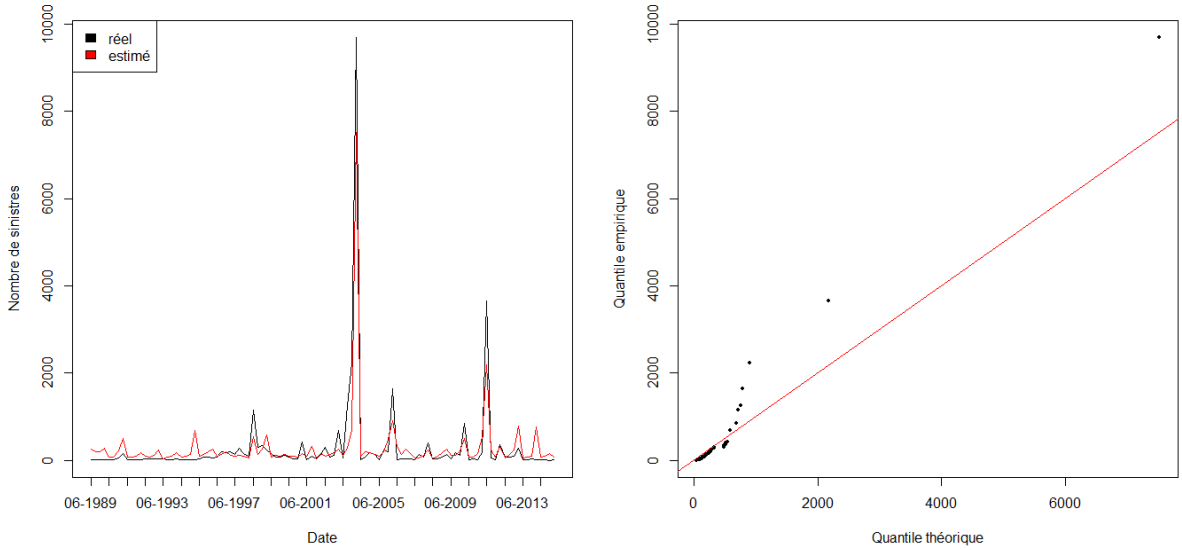


Figure 51 - Modèle de fréquence de sinistralité – Poisson croisé

Le taux de déviance est de 68 %, ce qui est nettement mieux. Le critère *AIC* vaut 70 825, ce qui est mieux que précédemment.

Le surplus de performance par rapport à ce qui précède n'est pas totalement satisfaisant : les pics de sinistralité sont encore trop sous-estimés, le modèle prévoit trop souvent des « petits » sinistres, et beaucoup de variables sont présentes dans le modèle augmentant le risque de sur-ajustement.

Les modèles linéaires généralisés, traditionnellement utilisés, ne semblent donc pas adéquats pour modéliser la fréquence de sinistralité[35].

Nous devons alors chercher un autre type de modèle capable de mieux s'adapter à la diversité des configurations. La prochaine partie présentera une méthode développée récemment en analyse de données : les forêts aléatoires.

[35] Les résultats sont aussi décevants avec des fonctions de lien comme l'identité ou celle du modèle binomiale négative.

B. Modélisation de la fréquence de sinistralité avec une forêt aléatoire

La quantité de données accessibles et la puissance de calcul des ordinateurs explosent depuis peu de temps. Cela permet de développer de nouvelles approches statistiques approfondissant considérablement les méthodes d'analyse de données.

Ces nouvelles méthodes d'analyse sont d'un immense intérêt pour le monde de l'assurance : elles peuvent aider à comprendre et à quantifier les risques de manière beaucoup plus performante qu'auparavant.

Dans cette partie, nous ferons une présentation succincte de l'apprentissage automatique qui regroupe une partie de ces méthodes. Nous appliquerons ensuite une de ces méthodes sur nos données.

1. L'apprentissage automatique et les forêts aléatoires

a) L'apprentissage automatique

L'apprentissage automatique (*machine learning* en anglais) est un domaine de l'intelligence artificielle qui vise à implémenter des méthodes capables d'apprendre des concepts non explicitement programmés et en un temps raisonnable. Il cherche à adapter un système de calculs à des sous-ensembles de données pour automatiser l'analyse statistique sur l'ensemble de données tout entier.

Formellement, le cadre théorique de l'apprentissage automatique est le suivant :

Soit $\mathcal{L}_n = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ un échantillon d'apprentissage, c'est-à-dire une suite de vecteurs aléatoires indépendants et identiquement distribués, de même loi qu'un vecteur aléatoire (X, Y). Le vecteur (X, Y) est indépendant de $\mathcal{L}_n$ et sa loi est inconnue. L'entier naturel n désigne le nombre d'observations de l'échantillon d'apprentissage.

On cherche donc à « apprendre » la loi inconnue de (X, Y) grâce à l'échantillon d'apprentissage $\mathcal{L}_n$ dont on dispose. Dans notre problématique, X est le vecteur aléatoire regroupant les variables explicatives, et Y est la variable réponse donnant le nombre mensuel de sinistres. Comme dans le cas des modèles linéaires généralisés, on cherche à matérialiser la relation entre les variables explicatives et la variable réponse.

Nous allons nous restreindre à certaines méthodes d'apprentissage automatique dont l'objectif est d'effectuer des régressions. Ces méthodes s'adaptent mieux à la diversité des données que les méthodes plus classiques, et peuvent donner des résultats beaucoup plus performants. En effet, aucune hypothèse n'est émise sur la distribution de la variable réponse, contrairement à beaucoup de modèles comme le modèle linéaire généralisé.

On souhaite séparer les données de manière optimale selon les valeurs prises par les variables explicatives. Pour cela, on peut utiliser un arbre binaire de décision.

L'algorithme suivant décrit la construction d'un arbre :

---

1. **Début**
2. $E = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ les réalisations d'un échantillon d'apprentissage
3. x_i : vecteur donnant les valeurs des k variables explicatives pour la i-ème observation
4. F_{max} : nombre maximal de feuilles
5. m : nombre minimal de données que doivent avoir chaque feuille de l'arbre

6. **Tant que** Le nombre de feuilles de l'arbre n'a pas atteint F_{max}, et que chaque feuille contienne au moins m données
7. **Faire**
8. Créer un nouveau nœud :
9. Tirer aléatoirement k' variables parmi les k variables explicatives [36]
10. Retenir la variable et le seuil associé qui minimisent la mesure d'erreur (décrite ci-dessous) au niveau du nœud
11. Créer deux branches : une avec les données dont les valeurs de la variable retenue sont inférieurs au seuil, et une autre avec les données dont les valeurs sont supérieures.
12. **Fin**
13. **Fin**
14. **Fin**

A chaque nœud de l'arbre, pour pouvoir apprécier l'importance d'une variable retenue dans l'explication de la variable réponse, et pour choisir le seuil optimal une fois la variable retenue, on définit une application « erreur » qui quantifie la dispersion de la variable réponse dans les deux sous-arbres crées. A titre d'exemple, on peut utiliser l'erreur des moindres carrés entre les sous-arbres, définie de la manière suivante :

$$Err = \sum_{y \in gauche} (y - \overline{y}_G)^2 + \sum_{y \in droit} (y - \overline{y}_D)^2$$

où

$\overline{y}_G$ est la moyenne des observations y associées à la variable retenue dont les valeurs sont inférieures au seuil
$\overline{y}_D$ est la moyenne des observations y associées à la variable retenue dont les valeurs sont supérieures au seuil

La section suivante présente une méthode qui produit plusieurs arbres de décision construits indépendamment les uns par rapport aux autres, et les combine afin de pouvoir capter le maximum de configurations possibles et ainsi établir de fines prédictions.

[36] On a bien sûr $k' \leq k$.
La plupart du temps, on prend par défaut $k' = k/3$ pour les régressions.

b) Les forêts aléatoires

Une méthode d'apprentissage pour la régression et la classification a été développée par L. Breimann en 2001 : la méthode des forêts aléatoires (*random forest* en anglais).

Une forêt aléatoire est un ensemble d'arbres de décision où chaque arbre est construit à partir d'un échantillon de données pris au hasard et où chaque nœud de l'arbre est construit à partir de variables explicatives prises au hasard.

Chaque arbre donne une prédiction pour des valeurs données des variables explicatives. La prédiction finale est la moyenne des prédictions sur l'ensemble des arbres.

L'avantage de cette méthode est que l'implémentation est simple à mettre en œuvre et que les résultats peuvent être très performants. Cependant, le temps de calcul peut être important.

Les arbres sont indépendants entre eux afin d'avoir le plus d'hétérogénéité dans les configurations possibles.

L'algorithme suivant décrit la construction d'une forêt aléatoire :

---

1. **Début**
2. $E = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ les réalisations d'un échantillon d'apprentissage
3. x_i : vecteur donnant les valeurs des k variables explicatives pour la i-ème observation
4. B : nombre d'arbres dans la forêt
5. **Pour** b allant de 1 à B
6. Tirer aléatoirement un échantillon E_b parmi E avec remise
7. Construire l'arbre basé sur l'échantillon E_b grâce à l'algorithme de construction d'arbre énoncé précédemment
8. **Fin**
9. **Fin**

Afin de mesurer la qualité du modèle, nous pouvons calculer la moyenne des « pseudo-R^2 » par arbre. Pour chaque arbre de la forêt, l'étude est basée sur une partie restreinte de l'échantillon initiale. La partie restante, appelée « *out-of-bag data* », va servir pour tester des prédictions. Dans la même logique que le R^2[37] des régressions linéaires, on étudie alors le rapport entre la variance des résidus et la variance du modèle. Plus le pseudo-R^2 est proche de 1, plus le modèle est satisfaisant. La qualité du modèle peut donc être appréciée par la moyenne des pseudo-R^2 calculé pour chaque arbre. En pratique, le modèle est considéré comme satisfaisant pour des valeurs des pseudo-R^2 supérieures à 70 %.

[37] $R^2 = 1 - \frac{\text{SCR}}{\text{SCT}}$ où SCR est la somme des carrés des résidus de la régression, et SCT est la somme des carrés totaux. Il évalue la qualité du modèle : plus il se rapproche de 1, meilleur est le modèle.

2. Application

a) Hypothèses

Nous allons reprendre les variables utilisées dans le modèle linéaire généralisé avec dépassement de seuil.

Le modèle utilise une forêt aléatoire se basant sur les hypothèses suivantes :

- Données : agrégées par mois, année, et *CRESTA*
- Variable réponse : $NbreSin$
- Variables explicatives : $ClassTemp, ClassSin, nALEA, Tmax, SumPrecip, SPI, SPEI, Temp_{30}$
  Pour chaque variable, on observe la réalisation des deux mois précédents.
- Offset : $Expo$
- Nombre de variables explicatives utilisées dans chaque arbre : 4
- Nombre maximal d'arbres dans la forêt : 500

b) Résultats

Le graphique suivant présente le résultat de l'ajustement d'une forêt aléatoire avec les hypothèses énoncées ci-dessus :

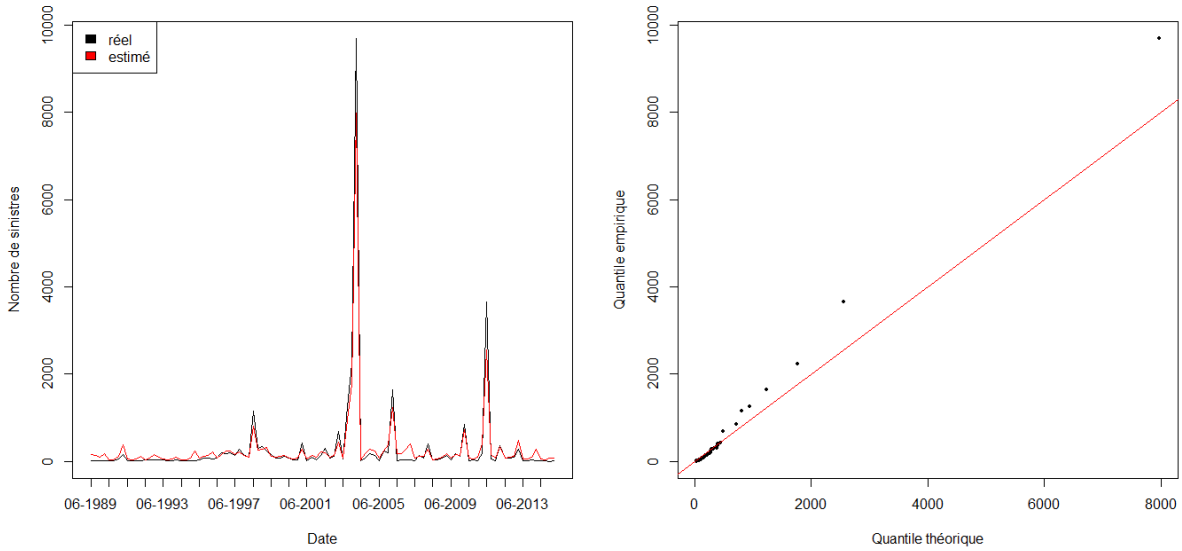


Figure 52 - Modèle de fréquence de sinistralité – Forêt aléatoire

Le moyenne des pseudo-R^2 est égal à 30 %, ce qui est faible. Cependant, les résultats sont beaucoup plus satisfaisants que ce qui précède : les pics de sinistralité sont relativement bien estimés, et le modèle est capable de prédire très peu de sinistres quand il le faut.

Nous allons finalement garder ce modèle pour expliquer les liens entre la fréquence de sinistralité due à la sécheresse et les variables explicatives.

Il nous reste donc à modéliser les coûts de de sinistralité.

C. Modélisation des coûts de sinistralité par les courbes de vulnérabilité

Nous venons de décrire le lien entre la fréquence de sinistralité liée à la sécheresse et les variables explicatives de la sécheresse. Or, nous possédons un catalogue de scénarios probabilisés d'évolution mensuelle de ces variables explicatives. Nous pouvons donc traduire ces simulations de variables physiques, en simulations de fréquence de sinistralité pour chaque zone concernée. Il reste à modéliser les coûts engendrés par ces sinistres. On définit alors le taux de destruction de la manière suivante :

$$\text{taux de destruction} = \frac{\text{montant du sinistre}}{\text{somme assurée}}$$

Les pertes financières engendrées par ces sinistres dépendent des spécificités des objets assurés. Pour une même intensité de sinistralité, l'ampleur des dommages peut être très diverse selon la nature de l'objet concerné. Par exemple, le taux de destruction médian d'une maison est plus important que celui d'un immeuble. En effet, si une sécheresse provoque une fissure dans un immeuble, cela aura beaucoup moins d'impact que si cette fissure affectait une petite maison, car le montant du sinistre reste le même mais la valeur assurée de l'immeuble est beaucoup plus élevée que celle de la maison.

Nous allons donc regrouper les objets assurés en plusieurs catégories. Pour chacune de ces catégories, nous allons étudier la répartition des taux de destruction enregistrés dans l'historique, grâce à des courbes appelées « courbes de vulnérabilité ».

Sachant qu'il y a un sinistre, la connaissance de la catégorie de l'objet assuré et de sa somme assurée pourra donner une estimation de la perte financière générée par ce sinistre[38]. Nous supposerons alors que dans le cas de la sécheresse, le montant d'un sinistre ne dépendra pas de l'intensité de l'événement (c'est-à-dire le nombre de sinistres) mais uniquement de la catégorie d'objets assurés et de la somme assurée.

Pour une catégorie d'objets assurés et une somme assurée données, nous allons chercher à caractériser la fonction de répartition F des taux de destruction, par un paramètre m. Nous appliquerons une méthode récente et développée chez Swiss Re : la méthode MBBEFD[39]. On pourra alors trouver un quantile des taux de destruction de niveau α^* tel que $q_{\alpha^*} := F^{-1}(\alpha^*) = m$.

Nous effectuerons ensuite une régression pour expliquer le lien entre les sommes assurées et les quantiles de taux de destruction q_{α^*}.

Finalement, une catégorie d'objets assurés et une somme assurée permettra d'estimer le quantile des taux de destructions de niveau α^* (grâce à la régression), pour ensuite paramétrer la fonction de répartition des taux de destruction (grâce à la méthode MBBEFD).

[38] $\text{montant du sinistre estimé} = \text{somme assurée de l'objet concerné} \times \text{taux de destruction estimé}$

[39] MBBEFD : Maxwell-Boltzman, Bose-Einstein, Fermi-Dirac

1. La méthode MBBEFD

Dans cette partie, nous souhaiterions paramétrer une fonction estimant la fonction de répartition des taux de destruction.

Pour cela, différentes approches sont possibles. Nous utiliserons une méthode qui a été développée en 1997 chez Swiss Re par S. Bernegger : la méthode MBBEFD.

Deux versions de cette méthode peuvent être utilisées :

- La méthode MBBEFD générale, qui prend en compte deux paramètres.
- La méthode MBBEFD hyperbolique, qui prend en compte un unique paramètre.

En pratique, la plupart des praticiens utilise la méthode MBBEFD hyperbolique. Elle a l'avantage d'être caractérisée directement par la médiane.

La méthode MBBEFD est fondée sur l'étude des courbes d'exposition. Les courbes d'exposition sont régulièrement utilisées en assurance pour déterminer des proportions de primes à céder en réassurance.

Nous allons faire une présentation formelle d'une courbe d'exposition, et justifier l'intérêt de sa construction.

On se fixe un contrat d'assurance. Soit $M \in \mathbb{R}_+^*$ la valeur assurée du bien concerné.
Soit $X \in \mathbb{R}_+$ la variable aléatoire représentant le montant d'un sinistre.
Soit $DR \in [0,1]$ la variable aléatoire représentant le taux de destruction. On a donc : $DR = \frac{X}{M}$.
Soit F la fonction de répartition de DR.
On suppose que $p = \mathbb{P}(DR = 1) > 0$.

La prime pure doit se calculer en s'alignant sur la répartition du risque concerné. Elle vaut donc $\mathbb{E}(X)$. Si nous voulions assurer un risque que si le montant prend des valeurs particulières, c'est-à-dire si $X \in [a, b]$ avec a et b des réels positifs, nous le tariferions par $\mathbb{E}(X\mathbb{I}_{X\in[a,b]})$.

Supposons l'existence d'un traité de réassurance qui protège le contrat avec une priorité Fet une portée illimitée. L'assureur est tenté d'étudier le rapport :

$$\frac{\text{part de la prime pure représentative des risques couverts par l'assureur}}{\text{prime totale}} = \frac{\mathbb{E}(X\mathbb{I}_{X\leq F})}{\mathbb{E}(X)} = \frac{\mathbb{E}\left(\frac{X}{M}\mathbb{I}_{\frac{X}{M}\leq\frac{F}{M}}\right)}{\mathbb{E}\left(\frac{X}{M}\right)} = \frac{\mathbb{E}(DR\mathbb{I}_{DR\leq x})}{\mathbb{E}(DR)} \text{ avec } x = \frac{X}{M}$$

On construit alors une courbe d'exposition représentant une fonction G qui va permettre de donner la part de la prime à garder pour refléter la part des risques à la charge de l'assureur.
On définit G de la manière suivante :

- $G: [0,1] \longrightarrow [0,1]$
- G continue
- $G: x \longmapsto \frac{\mathbb{E}(DR\mathbb{I}_{DR\leq x})}{\mathbb{E}(DR)}$

Nous cherchons à connaître F par l'intermédiaire de G. On remarque que :

$$G(x) = \frac{\mathbb{E}(DR\mathbb{I}_{DR\leq x})}{\mathbb{E}(DR)} = \frac{\int_{t=0}^{x}\big(1-F(t)\big)dt}{\int_{t=0}^{1}\big(1-F(t)\big)dt}$$

Cela implique que :

$$F(x) = \mathbb{I}_{x=1} + \left(1 - \frac{G'(x)}{G'(0)}\right)\mathbb{I}_{0\leq x<1}$$

Si l'on donne une bonne estimation de G, on donnera une bonne estimation de F, et donc nous aurons une bonne représentation de la répartition des taux de destruction.

La méthode MBBEFD générale propose une classe $\mathcal{C}$ de fonctions, appelée classe MBBEFD, qui permet d'estimer G. La classe de fonction utilisée est la suivante :

$$\mathcal{C} = \left\{ G_{a,b} : x \in [0,1] \mapsto \frac{\log(a+b^x) - \log(a+1)}{\log(a+b) - \log(a+1)} \;\middle|\; (a,b) \in {D_{a,b}}^2 \right\}$$

où $D_{a,b} = \{(a,b) \in \mathbb{R}^2 \mid a+1>0,\ a(1-b)>0,\ b>0\}$

Les formes des courbes représentatives des fonctions $G_{a,b}$ sont la plupart du temps suffisamment variées pour approcher la courbe empiriques de G.

On cherche à trouver un couple (a^*, b^*) qui permet d'ajuster de manière optimale la courbe empirique de G avec la courbe de G_{a^*,b^*}.

Cela implique que :

$$\begin{cases} \hat{F}(x) = \mathbb{I}_{x=1} + \left(1 - \dfrac{(a^*+1){b^*}^x}{a^* + {b^*}^x}\right)\mathbb{I}_{0\leq x<1} \\ p = \mathbb{P}(DR=1) = 1 - \hat{F}(1) = \dfrac{(a^*+1)b^*}{a^*+b^*} \end{cases}$$

Comme dit précédemment, les praticiens utilisent la majorité du temps une version simplifiée de la méthode MBBEFD générale qui prend compte un unique paramètre.

On cherche à approcher G par une fonction G_m caractérisée par un unique paramètre m. G_m va permettre d'estimer la fonction de distribution des taux de destruction F qui sera donc caractérisée par un unique paramètre. On supposera $0 < m < 1$.

La méthode MBBEFD hyperbolique consiste à reprendre la classe de fonctions $\mathcal{C}$ en jouant sur les paramètres a et b de sorte à converger vers une classe de fonctions G_m à un seul paramètre.

La fonction de répartition des taux de destruction estimée $\hat{F}$, calculée à partir de G_m, s'écrit alors :

$$\begin{cases} \hat{F}(x) = \mathbb{I}_{x=1} + \left(1 - \dfrac{1}{1+\dfrac{x}{m}}\right)\mathbb{I}_{0\leq x<1} \\ p = \mathbb{P}(DR=1) = 1 - \hat{F}(1) = \dfrac{m}{1+m} \end{cases}$$

Nous obtenons ainsi que :

$$\begin{cases} \mathbb{E}(DR) = m.\log\left(1+\dfrac{1}{m}\right) \\ F^{-1}(\alpha) = \mathbb{I}_{\alpha>1-p} + m.\dfrac{\alpha}{1-\alpha}\mathbb{I}_{0\leq\alpha\leq 1-p} \end{cases}$$

Nous constatons que le paramètre m de F est la médiane de la distribution associée : $F^{-1}\left(\frac{1}{2}\right) = m$ [40].

Ainsi, si nous sommes capables d'estimer le taux de destruction médian pour une catégorie de contrats et une somme assurée données, alors ce taux médian jouera le rôle de paramètre dans la fonction de répartition des taux de destructions.

2. Liens entre les taux de destruction médians et les sommes assurées

Nous avons à disposition l'historique de sinistralité liée à la sécheresse qui nous indique, pour chaque sinistre : le type de contrat assurant l'objet concerné, la somme assurée et le montant du sinistre.

Les objets assurés doivent être catégorisés par type. Nous définissons alors quatre catégories de contrats assurant les objets :

- MRH
- Immeuble
- Agricole
- Risques industriels

Afin de modéliser les taux de destruction liés aux sommes assurées pour chacune de ces catégories, nous allons procéder en plusieurs étapes.

D'abord, à partir du sous-historique ne contenant que les contrats d'une certaine catégorie, nous allons calculer le taux de destruction de chaque sinistre.

Ensuite, nous allons calculer les taux de destructions médians par couche de sommes assurées.

Enfin, nous allons effectuer une régression pour modéliser la relation entre les taux de destruction médians et les sommes assurées.

Soit C une catégorie fixée de contrats.

Soit N le nombre de couche de sommes assurées pour la catégorie C.

Soient $SI_k^C \in \mathbb{R}_+$ et $DR_k^C \in [0,1]$ respectivement la somme assurée médiane et le taux de destruction médian de la catégorie C et de la k-ième couche $(1 \leq k \leq N)$.

En réassurance, il n'est pas rare d'observer un lien log-linéaire entre les taux de destruction et les sommes assurées. Pour vérifier cela, nous allons effectuer pour chaque catégorie C une régression linéaire sur $(\log(SI_k^C), \log(DR_k^C))_{1\leq k\leq N}$.

Les graphiques suivants nous présentent les résultats :

[40] En effet, $0 < m < 1 \implies 1 - p = 1 - \frac{\mathrm{m}}{1+\mathrm{m}} > \frac{1}{2}$

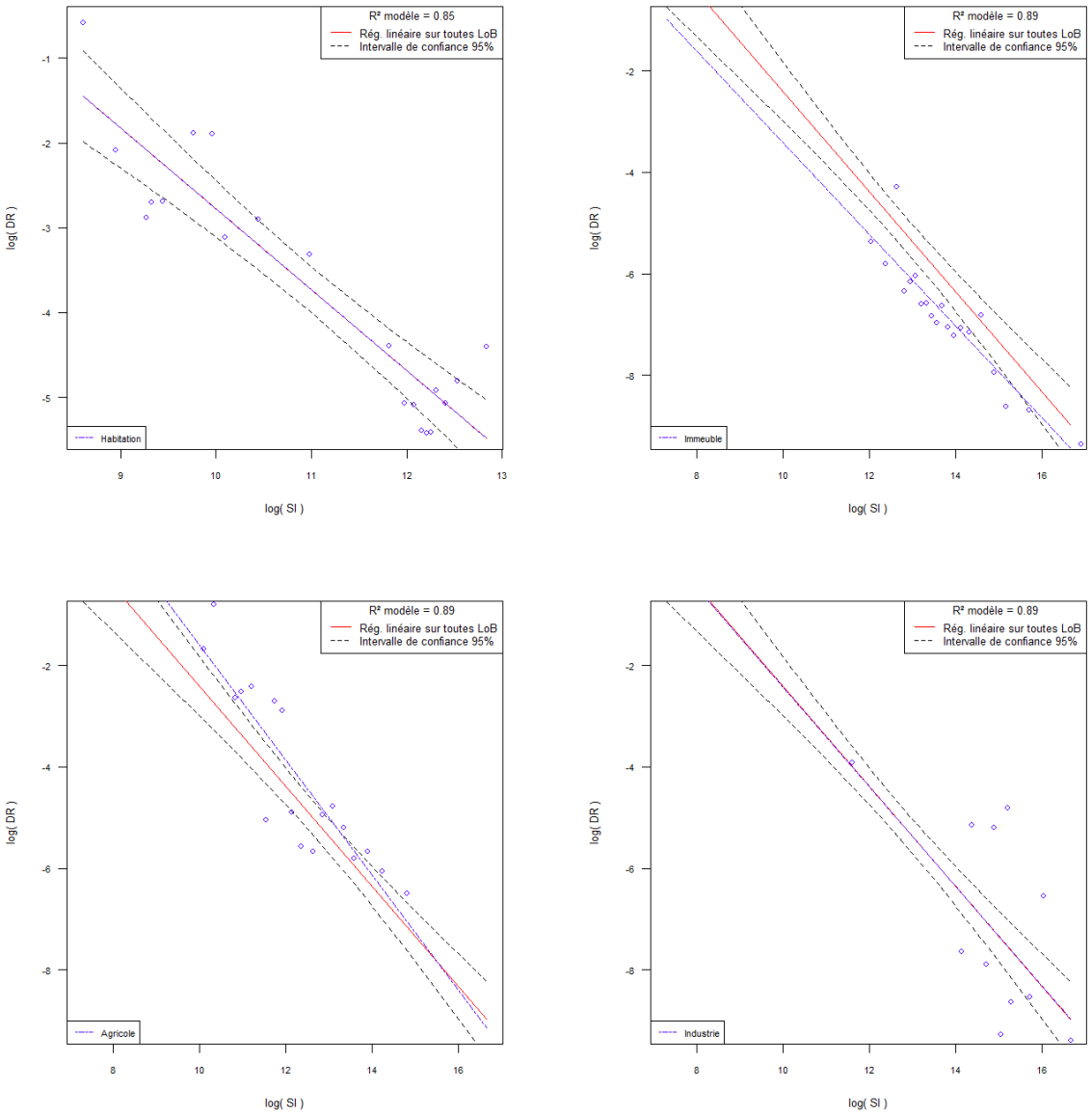


Figure 53 - Relation log-linéaire entre les taux de destruction et les sommes assurées

D'abord, on observe que globalement, la relation entre les taux de destruction et les sommes assurées est log-linéaire décroissante.

Ensuite, la régression effectuée sur les taux de destruction de la catégorie Risques Industriels n'est pas totalement satisfaisante. Il faut souligner le fait que les sinistres de cette catégorie représentent seulement moins de 1 % de la totalité des sinistres : chaque couche contient très peu de données et donc les points sont plus éparpillés.

Enfin, nous remarquons qu'en regroupant les catégories Immeuble, Agricole et Risques industriels, on obtient une régression pertinente. En effet, le R^2 obtenu est de 89 %, ce qui est satisfaisant.

Deux grandes classes de catégories de contrats peuvent être définies:

- Particulier, qui correspond aux contrats MRH ;
- Professionnel, qui correspond aux contrats Immeuble, Agricole et Risques industriels.

Nous avons alors obtenu un modèle de régression pour les contrats de particulier et un modèle de régression pour les contrats professionnels. Chaque modèle s'écrit de la manière suivante :

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

où a et b sont donnés dans le tableau suivant pour chaque type de contrats :

Particulier	Coefficient	Erreur	t valeur	Pr(>|t|)	
b	6,76	1,02	6,60	3,38e-06	***
a	-0,95	0,093	-10,29	5,70e-09	***

Professionnel	Coefficient	Erreur	t valeur	Pr(>|t|)	
b	7,43	1,04	7,13	1,22e-06	***
a	-0,98	0,080	-12,23	3,71e-10	***

Tableau 3 - Paramètres de régression – Taux de destruction et somme assurée

Les valeurs prises par la statistique de Student t[41] indiquent que les coefficients sont tous significatifs. Nous sommes maintenant capables d'estimer le taux de destruction médian pour un contrat et une somme assurée donnés.

Avec la méthode MBBEFD décrite précédemment, nous allons pouvoir estimer la fonction de répartition des taux de destruction et ainsi construire les courbes de vulnérabilité.

3. Construction des courbes de vulnérabilité

Nous avons supposé que le montant d'un sinistre (ou son taux de destruction) ne dépendait que de la catégorie de contrats concerné et de la somme assurée.

Pour une catégorie C de contrats et une couche k de sommes assurées données, nous sommes maintenant capables d'estimer le taux de destruction médian $\widehat{DR_k^C}$.

Ensuite, ce taux de destruction médian paramètre une fonction $\widehat{F_k^C}$ estimant la fonction de répartition F_k^C des taux de destruction des contrats de la catégorie C et de la couche k, de la manière suivante :

$$\begin{cases} \widehat{F_k^C}(x) = \mathbb{I}_{x=1} + \left(1 - \dfrac{1}{1 + \dfrac{x}{\widehat{DR_k^C}}}\right)\mathbb{I}_{0 \le x < 1} \\ p = \mathbb{P}(DR = 1) = 1 - \hat{F}(1) = \dfrac{\widehat{DR_k^C}}{1 + \widehat{DR_k^C}} \end{cases}$$

Cette fonction est représentée par une courbe, appelée « courbe de vulnérabilité ». Il y a donc une courbe de vulnérabilité particulière pour chaque catégorie de contrats et pour chaque couche de sommes assurées.

Les graphiques suivants présentent les résultats de la méthode MBBEFD par catégorie de contrats et de somme assurée.

[41] La statistique et le test de Student sont décrits en annexe D.

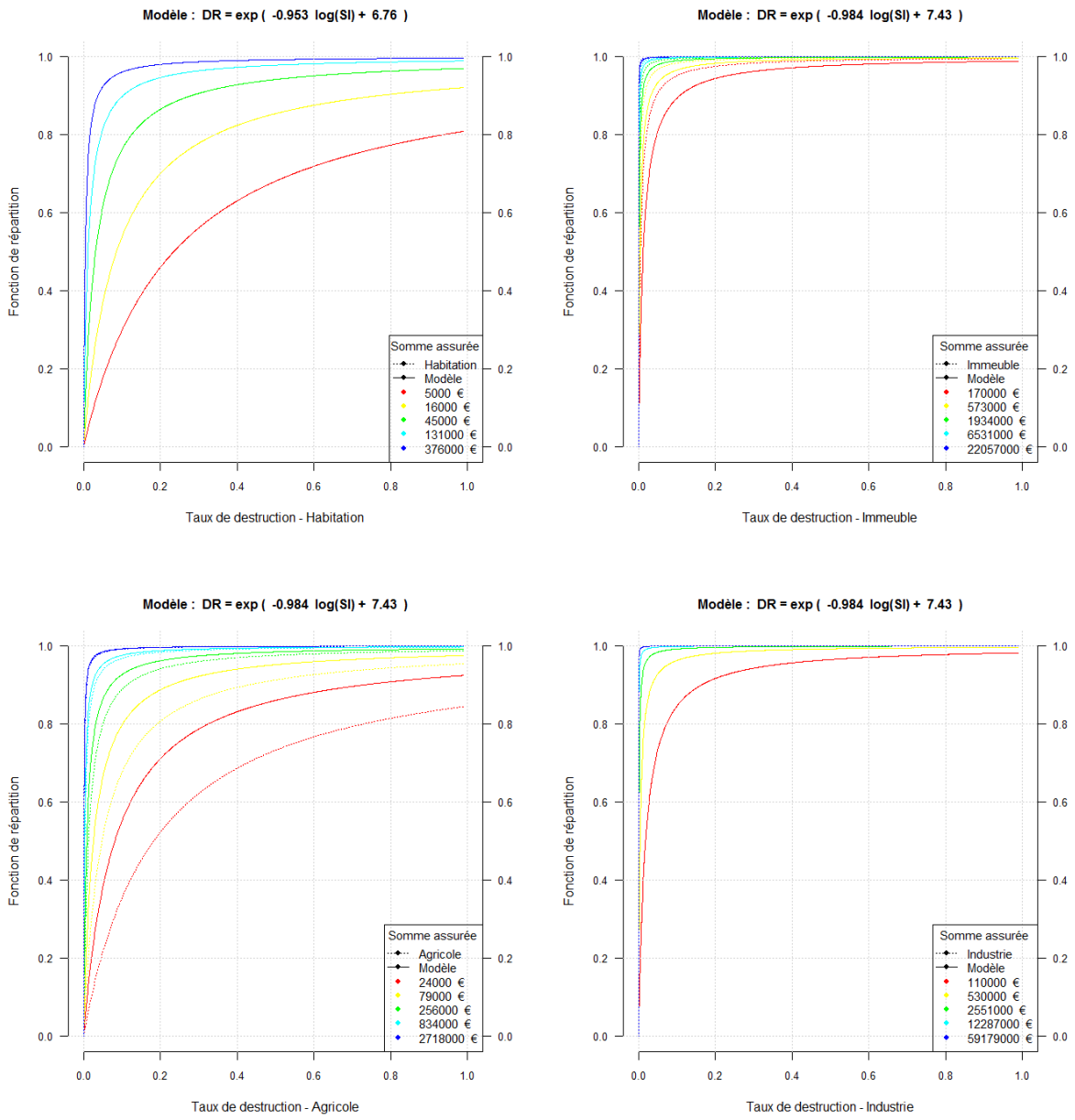


Figure 54 - Courbes de vulnérabilité

Pour une catégorie de contrats donnée, les courbes de vulnérabilité associées à des couches de sommes assurées élevées tendent à concentrer les taux de destruction en 0. Fixons un dommage particulier : une fissure sur un mur. Si c'est le mur d'une petite maison, le taux de destruction associé à ce sinistre est plus élevé que si c'est le mur d'une grande maison. En effet, le montant du sinistre reste le même mais la valeur d'une petite maison est plus faible que celle d'une grande maison.

IV. Résultats du modèle

Le module Aléa nous a permis de trouver deux variables physiques dont les comportements caractérisent le phénomène de sécheresse : les précipitations cumulées mensuelles et les températures maximales journalières.

Nous en avons déduit des indicateurs, qui constituent alors un ensemble de variables explicatives de la sécheresse.

Nous avons ensuite modélisé les précipitations et les températures afin de générer 10 000 scénarios d'évolution mensuelle des variables explicatives.

Le module Vulnérabilité nous a permis de quantifier le lien entre la fréquence de sinistralité liée à la sécheresse et ses variables explicatives. Cela nous a donc permis de traduire les 10 000 scénarios d'évolution mensuelle des variables explicatives en 10 000 scénarios d'évolution mensuelle de la fréquence de sinistralité.

Nous avons ensuite modélisé les coûts des sinistres par catégorie de contrats et de somme assurée. Cela va nous permettre de générer 10 000 scénarios d'évolution mensuelle des coûts de sinistralité.

Le module Financier n'est pas modélisé car les pertes enregistrées dans l'historique de sinistralité liée à la sécheresse sont nettes des conditions contractuelles.

La modélisation de la sécheresse est basée sur la période estivale : de juin à septembre. Pour connaître les pertes annuelles, nous allons effectuer une régression afin d'estimer la perte annuelle à partir de la perte estivale.

Il nous restera finalement à définir un événement sécheresse, au sens assurantiel, pour en déduire une distribution des pertes financières accumulées par événement. Pour synthétiser les résultats, nous construirons alors deux types de courbe :

- La courbe *AEP*, qui déterminera le capital réglementaire requis sous Solvabilité II ;
- La courbe *OEP*, qui aidera à optimiser la structuration des traités de réassurance.

A. Ajustement pertes estivales / pertes annuelles

Les pertes causées par la sécheresse s'étalent entre la fin du printemps et le début de l'automne. La modélisation de la sécheresse est alors basée entre début juin et fin septembre. Comme dit précédemment (en I.C.3.a), les données utilisées pour représenter l'historique des événements sécheresse présente des défauts non négligeables. Entre autres, beaucoup de sinistres sont enregistrés en fin d'exercice le 31/12. Pour parer à ce problème, nous allons effectuer une régression linéaire sur les pertes estivales afin d'estimer les pertes annuelles.

Le modèle s'écrit alors :

$$(Perte.annuelle) = a \times (Perte.estivale) + b$$

Les coefficients obtenus par la régression sont présentés par le tableau suivant.

	Coefficient	Erreur	t valeur	Pr(>|t|)	
b	10,43e+06	2,48e+06	4,20	0,00032	***
a	1,034	5,27e-02	19,62	2,78e-16	***

Tableau 4 - Paramètres de régression – Perte estivale et Perte annuelle

Les valeurs prises par la statistique de Student (t valeur) indiquent que les coefficients sont tous significatifs. Le R^2 est de 94 %, ce qui est satisfaisant.

Le graphique suivant présente le *Q-Q plot* obtenu. On compare la répartition des pertes annuelles réelles avec celle des pertes annuelles estimées par les pertes estivales.

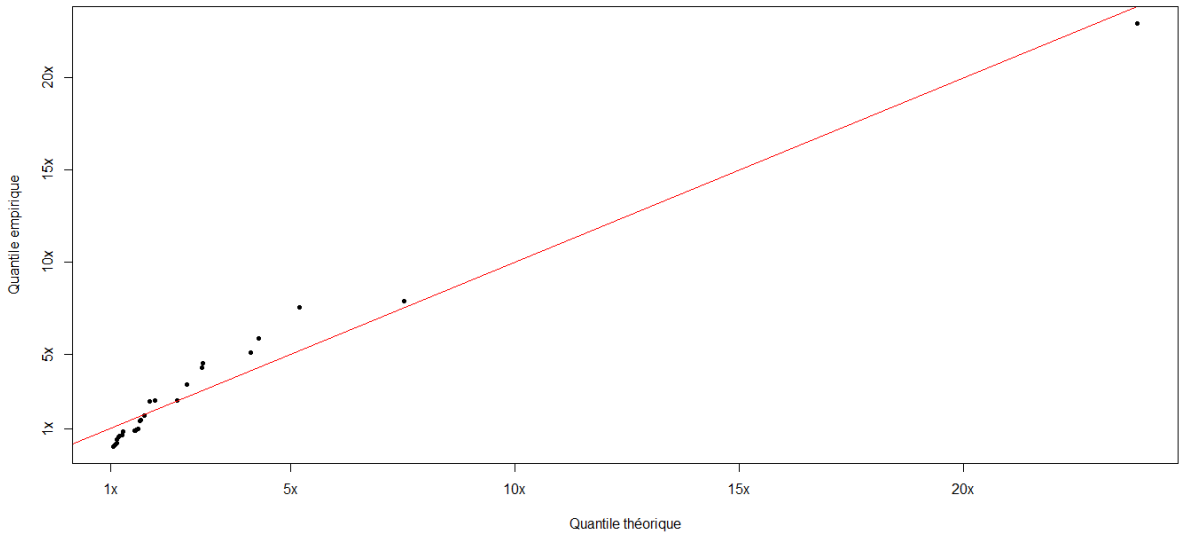


Figure 55 - Ajustement perte estivale/perte annuelle

Les résultats sont satisfaisants. Nous sommes maintenant capables de traduire les pertes estivales en pertes annuelles. Il reste à construire les courbes *AEP* et *OEP.*

B. Construction des courbes *AEP* et *OEP*

1. Définition d'un événement sécheresse du point de vue assurantiel

Il est difficile de définir ce qu'est un événement sécheresse. En effet, la garantie catastrophes naturelles ne s'applique que si un arrêté interministériel constate l'état de catastrophe naturelle. Or, contrairement aux autres catastrophes naturelles, il est difficile de situer précisément le début et la fin d'une période de sécheresse. La dimension politique rentre alors en jeu et peut fausser toute tentative de caractériser un événement sécheresse.

Nos estimations de sinistralité étant mensuelles, nous allons définir un événement comme étant simplement la perte accumulée mensuelle. Cette définition peut être remise en question. En effet, une période de sécheresse peut s'étaler sur plusieurs mois et l'historique de sinistralité ne permet pas d'en délimiter le début et la fin.

2. Les courbes *AEP* et *OEP*

Les modules Aléa, Vulnérabilité et Financier ont permis d'obtenir des simulations reflétant la distribution des pertes mensuelles et donc des événements. Ces simulations sont souvent synthétisées sous forme de courbe représentant la distribution des pertes.

Les pertes considérées correspondent à la somme des pertes cumulées pour un unique événement ou un ensemble d'événements.

Comme vu dans la première partie, modéliser les catastrophes naturelles présente deux intérêts :

- Le premier est d'optimiser les dispositions contractuelles de réassurance. On va donc s'intéresser aux événements ayant généré le plus de pertes. Le but est de représenter la distribution de la perte maximale causée par un unique événement au cours d'une année. C'est le rôle de la courbe *OEP*.
- Le deuxième est de calculer le capital requis sous Solvabilité II, afin de couvrir l'ensemble des risques à 99,5 %. On va donc s'intéresser aux pertes cumulées sur l'ensemble de tous les événements, c'est-à-dire aux pertes annuelles. Le but est de représenter la distribution des pertes annuelles. C'est le rôle de la courbe *AEP*.

Les distributions recherchées sont exprimées en fonction d'une période de retour. Une période de retour correspond au temps statistique entre deux événements de même intensité (générant des pertes semblables). Par exemple, un événement (ou un ensemble d'événements) dont la période de retour est de 200 ans se produira en moyenne une fois tous les 200 ans. Ainsi, cet événement aura une chance sur 200 de se produire sur une année.

La courbe *OEP* associe une période de retour à la perte maximale des événements sur une année.

La courbe *AEP* associe une période de retour à la perte totale des événements sur une année.

Pour couvrir les risques à 99,5 %, nous devons donc détenir le montant associé à la période de retour de 200 ans sur la courbe *AEP* obtenue[42].

Pour construire la courbe *AEP*, nous allons procéder en plusieurs étapes :

- Pour chaque scénario annuel (ou simulation), on retient la perte totale sur l'année.
- On trie les 10 000 pertes annuelles simulées par ordre décroissant.
- La période de retour associée à la k-ième perte vaut : $\frac{10000}{k}$

  Soit v_k la k-ième valeur de la suite décroissante des pertes annuelles. La perte annuelle a été supérieure ou égale à v_k, k fois en 10 000 années simulées. Ce qui revient à dire que les pertes annuelles sont supérieures ou égales à v_k une fois toutes les $\frac{10000}{k}$ années en moyenne.

Dans le cas de la courbe *AEP*, c'est la 50-ième pire[43] perte annuelle simulée qui correspondra au capital requis sous Solvabilité II.

La construction de la courbe *OEP* est similaire à celle l'*AEP*, sauf qu'on remplace la perte annuelle par la perte maximale mensuel (nous avions défini un événement comme étant la perte mensuelle).

Les parties suivantes présentent les courbes *AEP* et *OEP* pour deux modèles :

- Un modèle de fréquence de sinistralité croisé avec un modèle de coût. C'est le modèle développé dans la partie III.
- Un modèle estimant directement les pertes mensuelles en fonction des variables explicatives de la sécheresse. Nous utiliserons les forêts aléatoires.

[42] En effet, la probabilité de ne pas pouvoir couvrir à 99,5 % vaut : $(1 - 99{,}5\ \%) = 0{,}5\ \% = \frac{1}{200}$. Cela signifie qu'en moyenne, si on détient moins que le montant estimé par la courbe AEP, il y aura plus de 0,5 % de chances de ne pas pouvoir couvrir l'ensemble des risques.

[43] En effet, $\frac{10\,000}{200} = 50$

C. Approche fréquence/coût

1. Hypothèses

Soit E un *CRESTA* fixé et $p \in E$ un contrat appartenant à E.

Soit $(NbreSin_n)_{6\leq n\leq 9}$ la suite des variables aléatoires donnant le nombre total de sinistres au mois n dans E.

Soit $Expo$ le nombre de contrats présents dans E.

Soit $(X_n)_{6\leq n\leq 9}$ la suite des **vecteurs** aléatoires contenant les variables explicatives (dont les indicateurs à seuil).

Soit $(Perte_n)_{6\leq n\leq 9}$ la suite des variables aléatoires donnant la somme des pertes de chaque contrat présent dans E.

Soit $SI(p)$ la somme assurée d'un contrat p.

Soit $DR_{k(p)}^{C(p)}$le taux de destruction d'un objet assuré par un contrat p de la catégorie $C(p)$ dont la somme assurée fait partie de la couche $k(p)$.

Nous avons estimé $NbreSin$ grâce à une forêt aléatoire : pour un vecteur donné de réalisations des variables explicatives, chaque arbre de la forêt prédit $NbreSin$ et la prédiction finale est la moyenne des prédictions sur l'ensemble de la forêt.

Nous supposons que les contrats au sein d'un même *CRESTA* ont tous la même probabilité d'être touché par la sécheresse. Ainsi, pour chaque contrat, nous supposons que la probabilité d'être touché par la sécheresse le mois n vaut $\frac{NbreSin_n}{Expo}$. Cette fréquence de sinistralité mensuelle est naturellement estimée par $\frac{\widehat{NbreSin_n}}{Expo}$.

On a ensuite estimé la médiane de $DR_{k(p)}^{C(p)}$ sur tous les contrats de la catégorie $C(p)$ et de la couche $k(p)$, ce qui a permis de paramétrer une certaine loi permettant de simuler les taux de destruction.

$$\forall p \in E, \exists (a,b) \in \mathbb{R}_-^* \times \mathbb{R}, \qquad \widehat{Mediane(DR_{k(p)}^{C(p)})} = e^{a.\log(SI(p))+b}$$

Finalement, nous estimons la somme des pertes de chaque contrat présent dans E par :

$$\begin{aligned}\widehat{Perte_n} &= \mathbb{E}(Perte_n \mid X_n) \\ &= \mathbb{E}\left(\sum_{p\in E}\left(\frac{NbreSin_n}{Expo} \times SI(p) \times DR_{k(p)}^{C(p)}\right) \mid X_n\right)\end{aligned}$$

$$= \sum_{p \in E} \mathbb{E}\left(\frac{NbreSin_n}{Expo} \times SI(p) \times DR_{k(p)}^{C(p)} \mid X_n\right)$$
$$= \sum_{p \in E} SI(p) \times \widehat{DR_{k(p)}^{C(p)}} \times \mathbb{E}\left(\frac{NbreSin_n}{Expo} \mid X_n\right)$$ [44]
$$= \frac{\widehat{NbreSin_n}}{Expo} \times \sum_{p \in E} \left(SI(p) \times \widehat{DR_{k(p)}^{C(p)}}\right)$$

Nous pouvons donc simuler 10 000 scénarios d'évolution mensuelle des pertes par *CRESTA*. Pour chaque mois, nous sommons ensuite les pertes mensuelles sur l'ensemble des *CRESTA*. Nous obtenons finalement 10 000 scénarios d'évolution mensuelle des pertes sur l'ensemble de la France entre juin et septembre. Il suffit alors d'ajuster avec les coefficients donnés en IV.A pour en déduire les pertes annuelles.

Nous pouvons dès à présent appliquer la méthode de construction des courbes *AEP* et *OEP* sur nos données.

2. Résultats

Le graphique suivant présente les courbes *AEP* et *OEP* obtenues avec l'approche fréquence/coût. Pour des raisons de confidentialité, les résultats ne sont donnés que par ordre de grandeur.

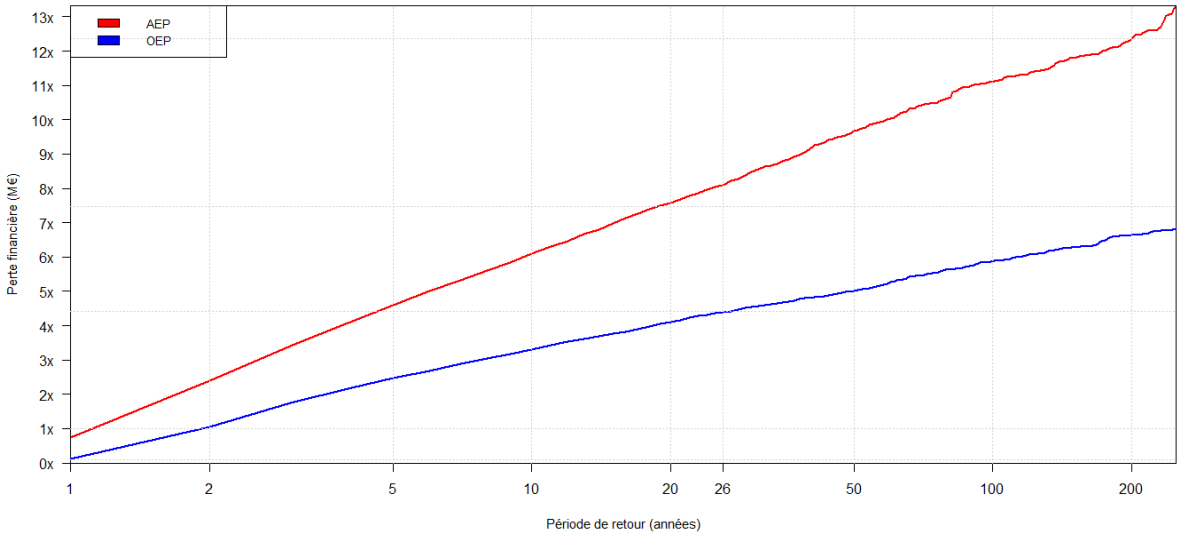


Figure 56 - Courbes *AEP* et *OEP* avec une approche fréquence/coût

Nous remarquons que la courbe *OEP* s'éloigne de la courbe *AEP* pour des périodes de retour élevées. Les périodes de retour élevées sont associées à des années connaissant une forte sécheresse. Or, la sécheresse a la particularité de s'étendre dans le temps, contrairement aux autres catastrophes naturelles. On peut alors supposer que les années très sèches ont généré des sinistres répartis sur plusieurs mois. Pour la construction de l'*OEP,* nous avions défini un événement comme étant la perte mensuelle. Si une sécheresse s'étend sur

[44] Le taux de destruction (aléatoire) ne dépend que de la catégorie du contrat concerné et de la somme assurée, et est donc indépendant des variables explicatives de la sécheresse.

plusieurs mois, le mois connaissant le plus de sinistres aura un poids moindre dans la perte annuelle (représentée par l'*AEP)*. Il n'est donc pas étonnant que la courbe *OEP* s'éloigne de celle de l'*AEP.*

La valeur x associée à l'axe des ordonnées correspond à la perte nette moyenne d'AXA causée par la sécheresse durant une année. Le pic de sinistralité enregistré en 2003 est presque huit fois plus élevé que la sinistralité annuelle moyenne. La période de retour associée est estimée à 20 ans, alors que notre historique de sinistralité s'étale sur 26 ans. La perte mensuelle maximale de 2003 est bien associée à une période de retour de 26 ans. Pour de faibles périodes de retour, le modèle a tendance à surestimer les pertes. Il est important de se rappeler que la fréquence de sinistralité causée par la sécheresse va augmenter ces prochaines décennies.

La perte bicentenaire estimée est presque deux fois plus élevée que la perte enregistrée en 2003.

Conclusion

La modélisation du risque sécheresse en France a été divisée en trois modules indépendants.

Le module Aléa nous a permis de générer un catalogue de 10 000 scénarios réalistes et probabilisés des évolutions mensuelles de l'ensemble des variables explicatives de la sécheresse, que nous avons sélectionné en croisant l'historique de sinistralité d'AXA lié à la sécheresse avec un ensemble de variables et d'indicateurs. Les variables explicatives sont toutes construites à partir des précipitations mensuelles et des températures maximales journalières.

Le module Vulnérabilité nous a permis de quantifier le lien existant entre la fréquence de sinistralité causée par la sécheresse et les variables explicatives. Nous avons modélisé les coûts de sinistralité de manière indépendante en supposant qu'ils dépendaient uniquement des catégories de contrats et de sommes assurées, et non des variables explicatives de la sécheresse. Cela nous a permis de traduire les simulations de variables explicatives en simulations de pertes financières.

Le module Financier n'a pas été développé dans ce mémoire car les pertes financières enregistrées dans l'historique de sinistralité liée à la sécheresse sont nettes de franchise, de coassurances, ... Le module Vulnérabilité nous permet donc d'obtenir directement 10 000 simulations d'évolution mensuelle des pertes financières nettes des conditions contractuelles.

Nous avons finalement synthétisé les simulations par deux courbes donnant une vision de la distribution des pertes causées par la sécheresse : une courbe donnant la distribution de la perte financière accumulée sur l'année et une courbe donnant la distribution de la perte financière maximale sur un événement sécheresse.

La modélisation des variables explicatives de la sécheresse a été un succès. Le lien entre ces variables explicatives et la fréquence de sinistralité est un problème délicat. En effet, l'historique de sinistralité dont nous disposons n'est pas nécessairement fidèle au véritable historique des événements sécheresse (d'un point de vue purement physique et non assurantiel). Il est difficile d'obtenir une base de données de sinistres de qualité suffisante pour représenter de manière optimale l'historique de la sécheresse. L'agrégation mensuelle des données allège malgré tout ce problème.

Une autre limite du modèle peut être soulevée : la définition d'un événement sécheresse, ce qui permettrait de connaître la distribution de l'événement annuel générant un maximum de pertes et pouvant être couvert par un traité XS dont la priorité et la portée pourraient être optimisées. En effet, il est difficile de déterminer quand commence et finit une période de sécheresse, uniquement à partir de l'historique de sinistralité. Nous avons alors défini un événement comme étant la perte financière accumulée mensuellement entre juin et septembre, mais cela peut être affiné. Néanmoins, la connaissance de la distribution de la perte annuelle est indépendante de la définition d'un événement sécheresse. Cela permet d'obtenir une vision complète du risque sécheresse en France et de calculer le capital réglementaire imposé par la réforme Solvabilité II qui s'appliquera dès 2016.

Annexe A : Evapotranspiration – Coefficients de correction

Le tableau suivant donne les coefficients de correction en fonction de la latitude et du mois concerné dans le calcul de l'évapotranspiration potentielle de la formule de Thornthwaite.

Latitude Nord	J	F	M	A	M	J	J	A	S	O	N	D
0	1,04	0,94	1,04	1,01	1,04	1,01	1,04	1,04	1,01	1,04	1,01	1,04
5	1,02	0,93	1,03	1,02	1,06	1,03	1,06	1,05	1,01	1,03	0,99	1,02
10	1	0,91	1,03	1,03	1,08	1,06	1,08	1,07	1,02	1,02	0,98	0,99
15	0,97	0,91	1,03	1,04	1,11	1,08	1,12	1,08	1,02	1,01	0,95	0,97
20	0,95	0,9	1,03	1,05	1,13	1,11	1,14	1,11	1,02	1	0,93	0,94
25	0,93	0,89	1,03	1,06	1,15	1,14	1,17	1,12	1,02	0,99	0,91	0,91
26	0,92	0,88	1,03	1,06	1,15	1,15	1,17	1,12	1,02	0,99	0,91	0,91
27	0,92	0,88	1,03	1,07	1,16	1,15	1,18	1,13	1,02	0,99	0,9	0,9
28	0,91	0,88	1,03	1,07	1,16	1,16	1,18	1,13	1,03	0,98	0,9	0,9
29	0,91	0,87	1,03	1,07	1,17	1,16	1,19	1,13	1,03	0,98	0,9	0,89
30	0,9	0,87	1,03	1,08	1,18	1,17	1,2	1,14	1,03	0,98	0,89	0,88
31	0,9	0,87	1,03	1,08	1,18	1,18	1,2	1,14	1,03	0,98	0,89	0,88
32	0,89	0,86	1,03	1,08	1,19	1,19	1,21	1,15	1,03	0,98	0,88	0,87
33	0,88	0,86	1,03	1,09	1,19	1,2	1,22	1,15	1,03	0,97	0,88	0,86
34	0,88	0,85	1,03	1,09	1,2	1,2	1,22	1,16	1,03	0,97	0,87	0,86
35	0,87	0,85	1,03	1,09	1,21	1,21	1,23	1,16	1,03	0,97	0,86	0,85
36	0,87	0,85	1,03	1,1	1,21	1,22	1,24	1,16	1,03	0,97	0,86	0,84
37	0,86	0,84	1,03	1,1	1,22	1,23	1,25	1,17	1,03	0,97	0,85	0,83
38	0,85	0,84	1,03	1,1	1,23	1,24	1,25	1,17	1,04	0,96	0,84	0,83
39	0,85	0,84	1,03	1,11	1,23	1,24	1,26	1,18	1,04	0,96	0,84	0,82
40	0,84	0,83	1,03	1,11	1,24	1,25	1,27	1,18	1,04	0,96	0,83	0,81
41	0,83	0,83	1,03	1,11	1,25	1,26	1,27	1,19	1,04	0,96	0,82	0,8
42	0,82	0,83	1,03	1,12	1,26	1,27	1,28	1,19	1,04	0,95	0,82	0,79
43	0,81	0,82	1,02	1,12	1,26	1,28	1,29	1,2	1,04	0,95	0,81	0,77
44	0,81	0,82	1,02	1,13	1,27	1,29	1,3	1,2	1,04	0,95	0,8	0,76
45	0,8	0,81	1,02	1,13	1,28	1,29	1,31	1,21	1,04	0,94	0,79	0,75
46	0,79	0,81	1,02	1,13	1,29	1,31	1,32	1,22	1,04	0,94	0,79	0,74
47	0,77	0,8	1,02	1,14	1,3	1,32	1,33	1,22	1,04	0,93	0,78	0,73
48	0,76	0,8	1,02	1,14	1,31	1,33	1,34	1,23	1,05	0,93	0,77	0,72
49	0,75	0,79	1,02	1,14	1,32	1,34	1,35	1,24	1,05	0,93	0,76	0,71
50	0,74	0,78	1,02	1,15	1,33	1,36	1,37	1,25	1,06	0,92	0,76	0,7

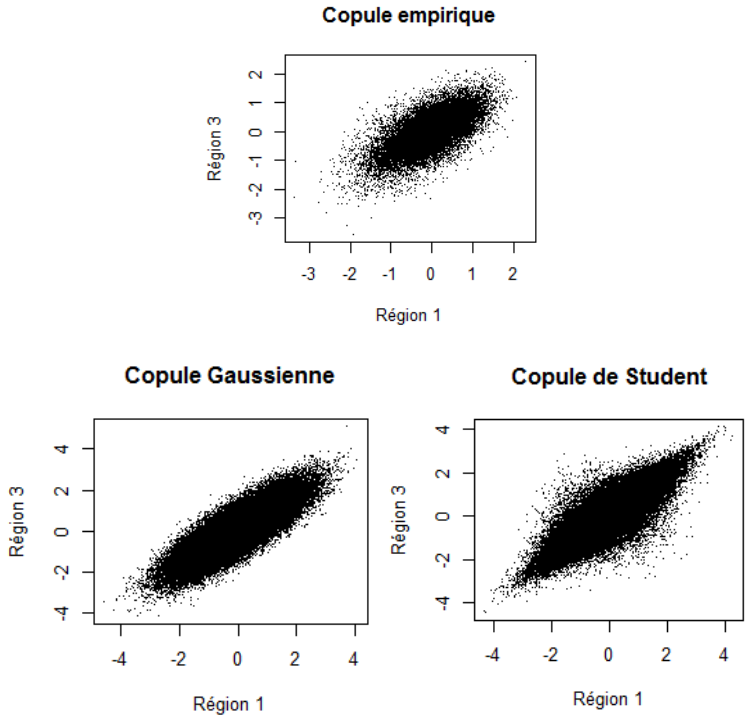
Latitude Sud	J	F	M	A	M	J	J	A	S	O	N	D
5	1,06	0,95	1,04	1	1,02	0,99	1,02	1,03	1	1,05	1,03	1,06
10	1,08	0,97	1,05	0,99	1,01	0,96	1	1,01	1	1,06	1,05	1,1
15	1,12	0,98	1,05	0,98	0,98	0,94	0,97	1	1	1,07	1,07	1,12
20	1,14	1	1,05	0,97	0,96	0,91	0,95	1,99	1	1,08	1,09	1,15
25	1,17	1,01	1,05	0,96	0,94	0,88	0,93	0,98	1	1,1	0,11	1,18
30	1,2	1,03	1,06	0,95	0,92	0,85	0,9	0,96	1	1,12	1,14	1,21
35	1,23	1,04	1,06	0,94	0,89	0,82	0,87	0,94	1	1,13	1,17	1,25
40	1,27	1,06	1,07	0,93	0,86	0,78	0,84	0,92	1	1,15	1,2	1,29
42	1,28	1,07	1,07	0,92	0,85	0,76	0,82	0,92	1	1,16	1,22	1,31
44	1,3	1,08	1,07	0,92	0,83	0,74	0,81	0,91	0,99	1,17	1,23	1,33
46	1,32	1,1	1,07	0,91	0,82	0,72	0,79	0,9	0,99	1,17	1,25	1,35
48	1,34	1,11	1,08	0,9	0,8	0,7	0,76	0,89	0,99	1,18	1,27	1,37
50	1,37	1,12	1,08	0,89	0,77	0,67	0,74	0,88	0,99	1,19	1,29	1,41

Tableau 5 - Evapotranspiration – Coefficients de correction

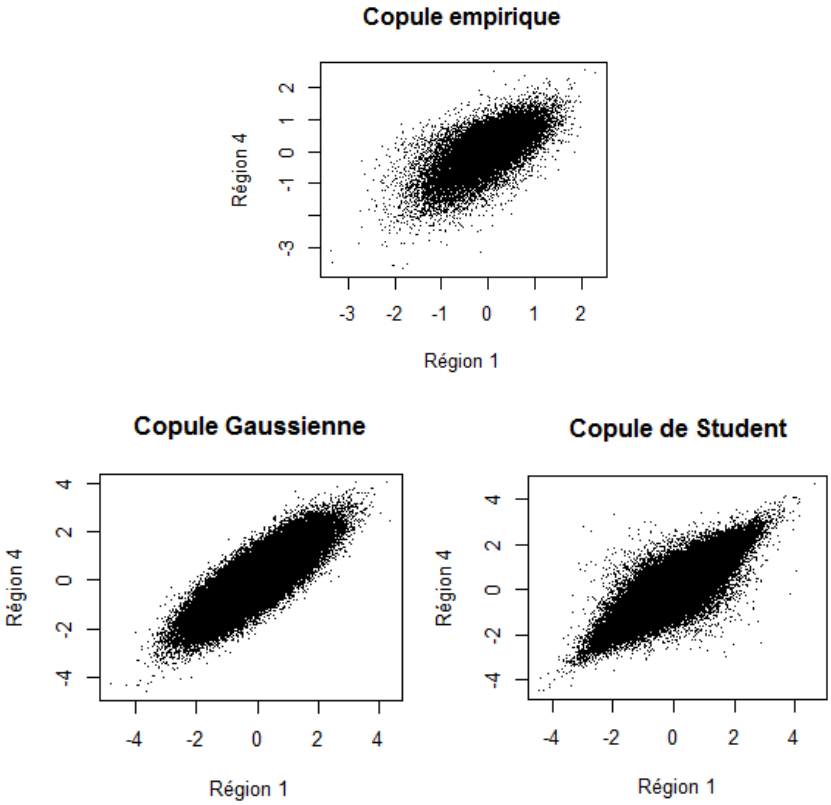
Annexe B : Dépendances entre régions – Comparaisons entre les copules empiriques et les copules elliptiques

- Région 1 : le sud de la France
- Région 2 : les zones montagneuses
- Région 3 : l'est de la France
- Région 4 : les pays de la Loire
- Région 5 : le nord de la France

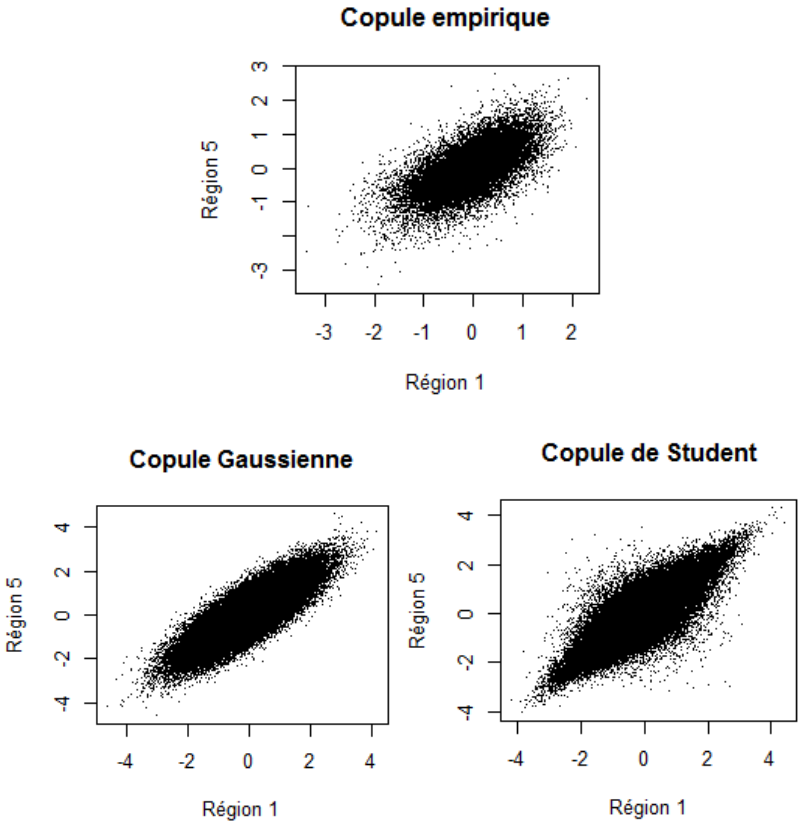
Région 1 / Région 3



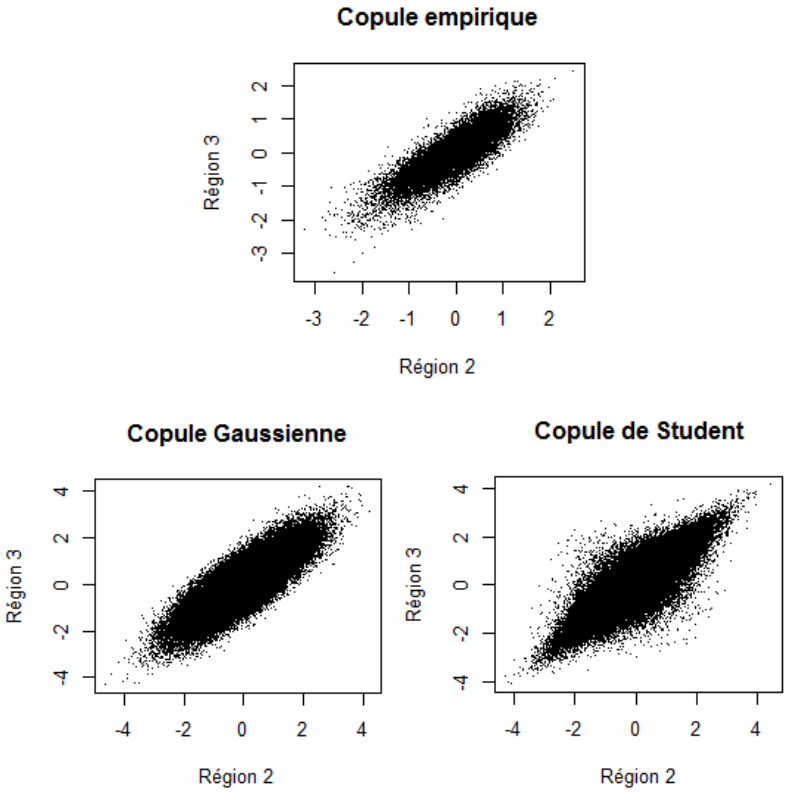
Région 1 / Région 4



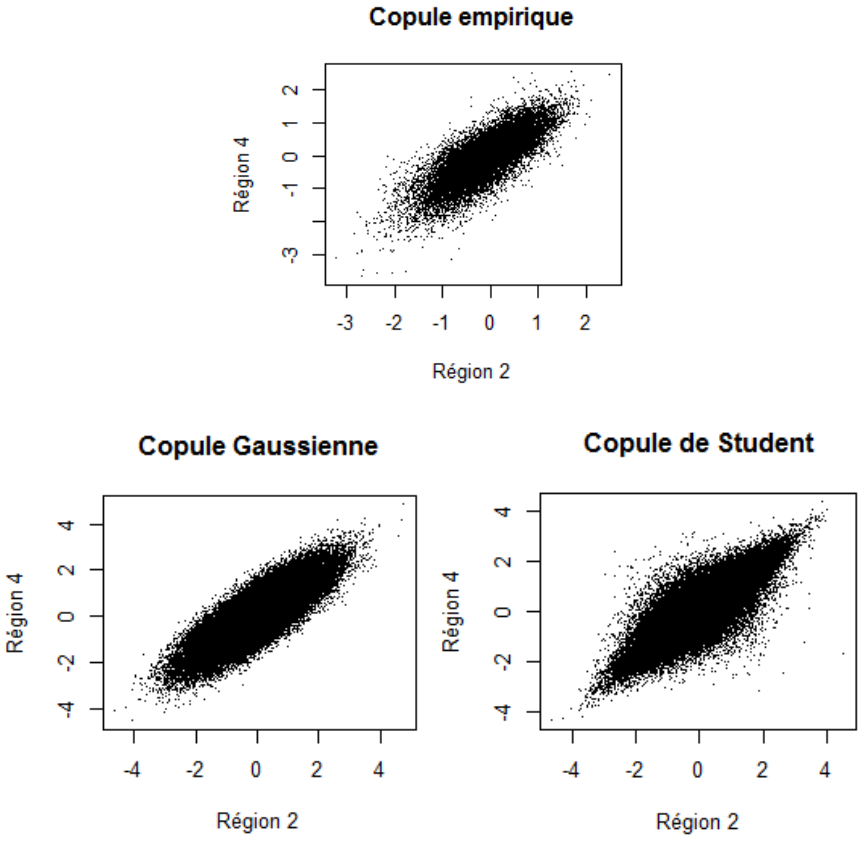
Région 1 / Région 5



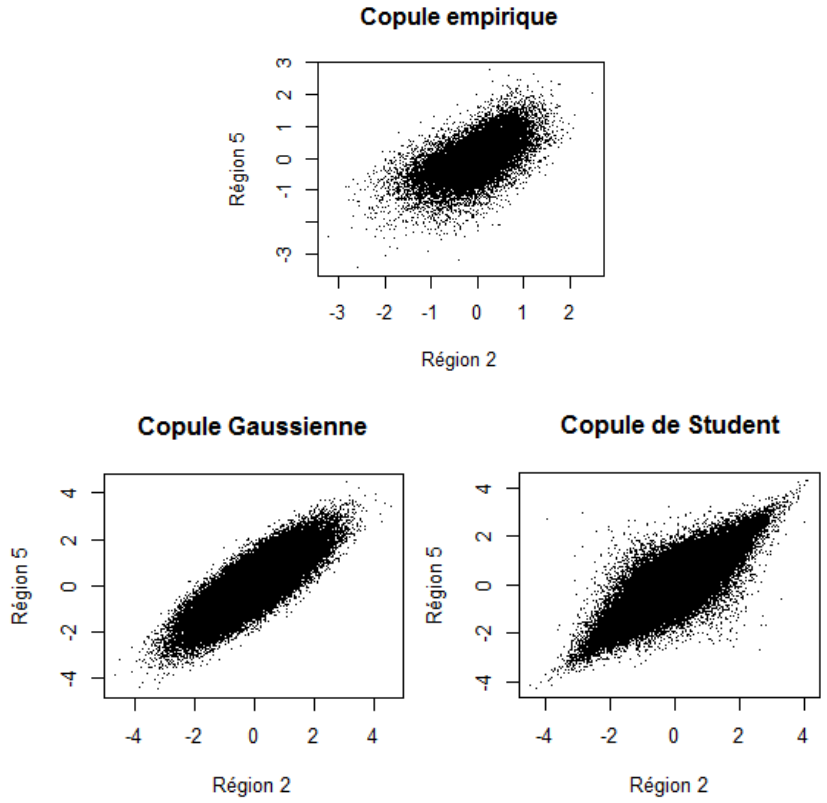
Région 2 / Région 3



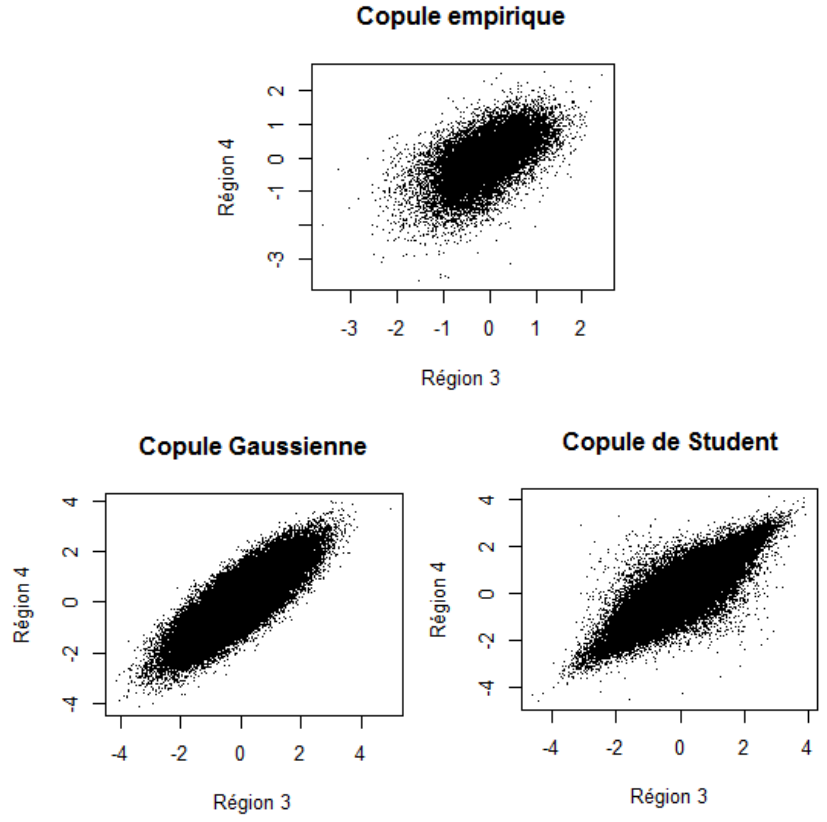
Région 2 / Région 4



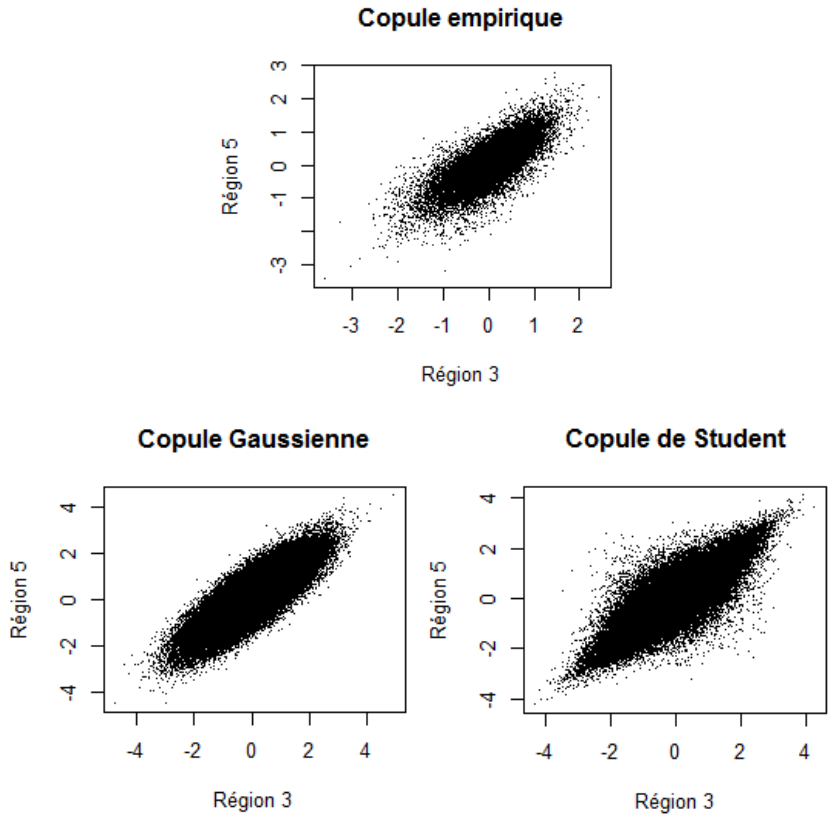
Région 2 / Région 5



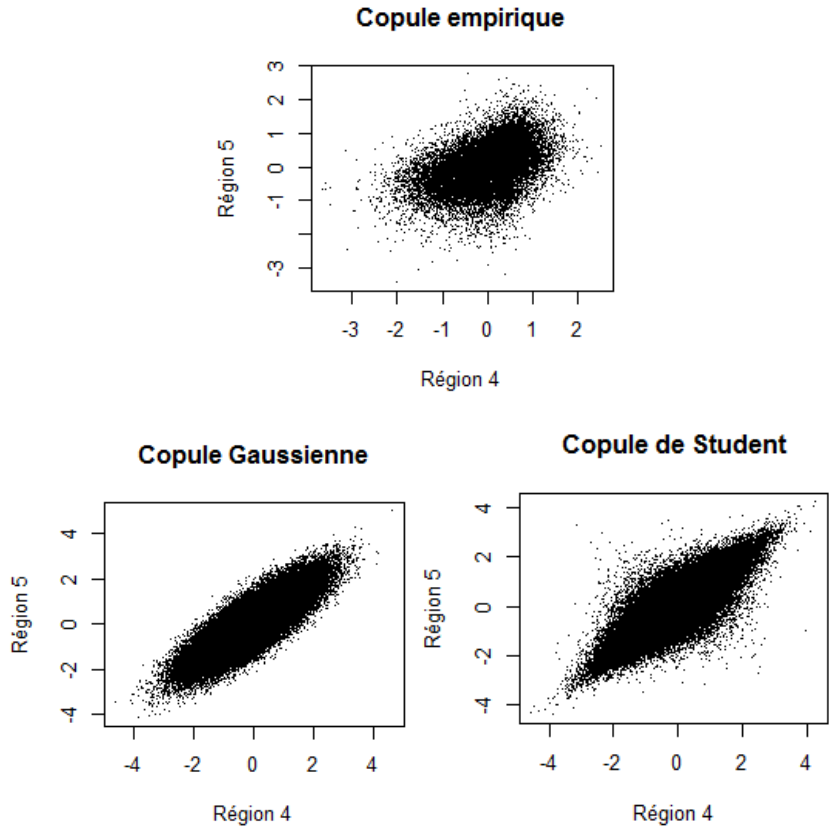
Région 3 / Région 4



Région 3 / Région 5



Région 4 / Région 5

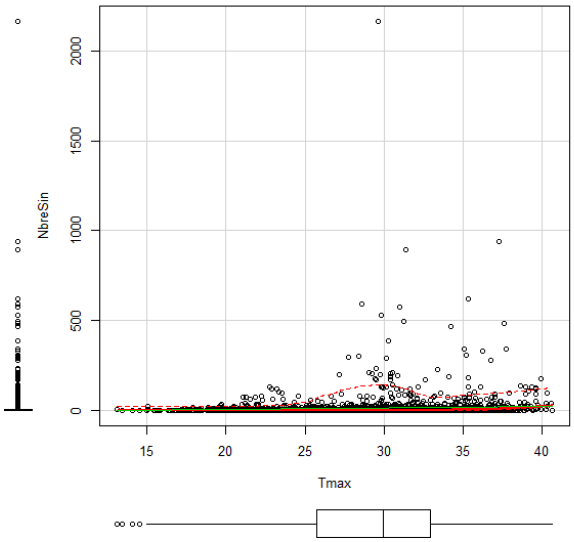


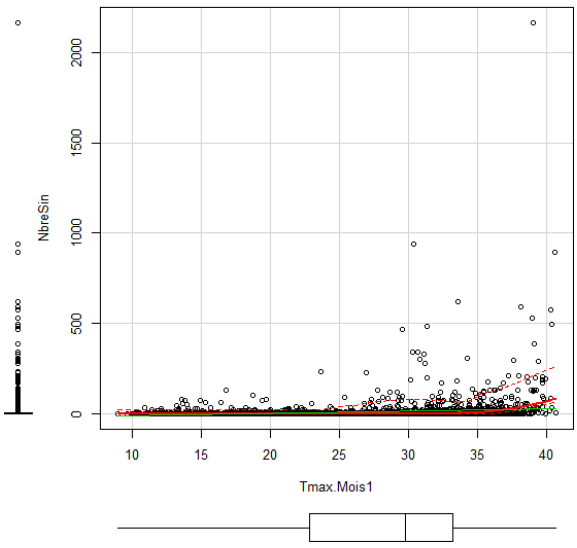
Annexe C : Variables explicatives – Graphiques

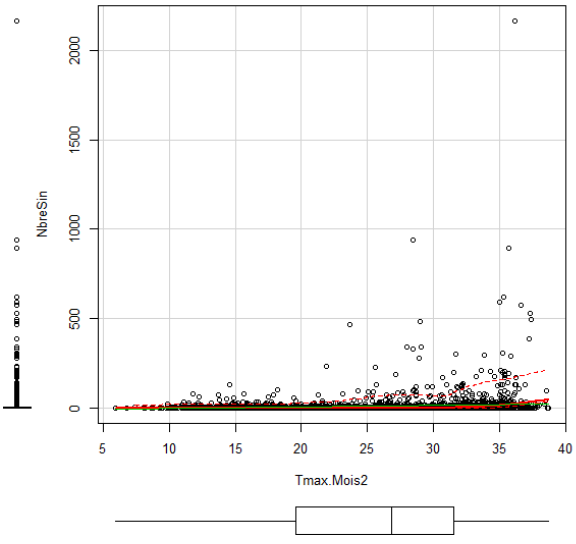
Nous souhaitons croiser le nombre de sinistres **mensuel** enregistré dans un *CRESTA* avec la valeur **mensuelle** d'un indicateur donné. On s'intéresse à la valeur du mois courant, de l'avant-dernier mois (. $Mois1$), et de l'avant-avant-dernier mois (. $Mois2$). On voit alors qu'il est plus intéressant de se concentrer sur les valeurs des deux dernier mois.

On s'intéressera ensuite à l'indicateur de dépassement de seuil grâce aux croisements avec les sinistres et avec différents seuils.

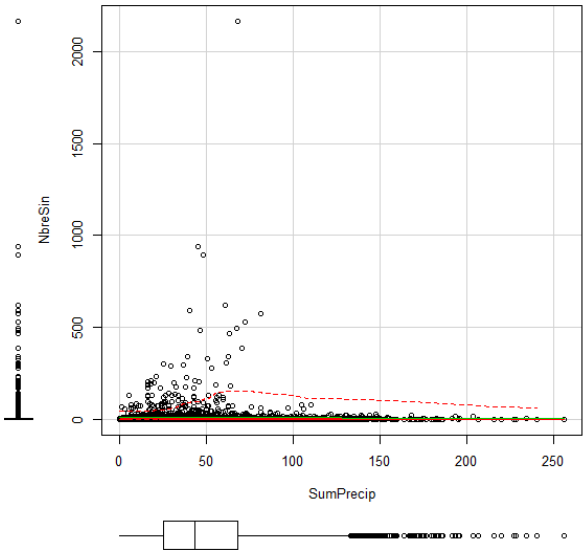
Température maximale mensuelle

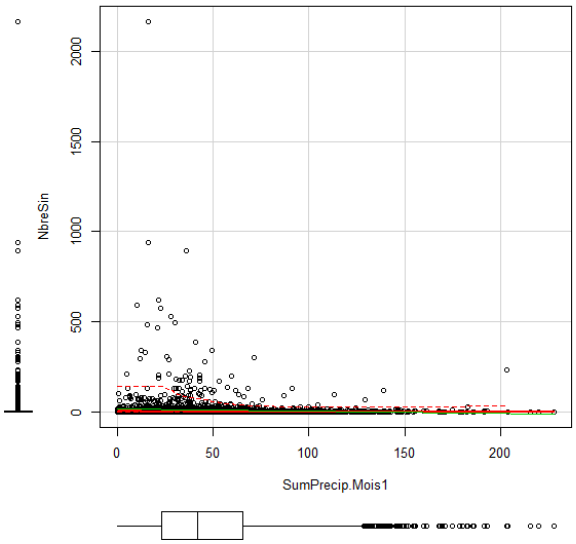


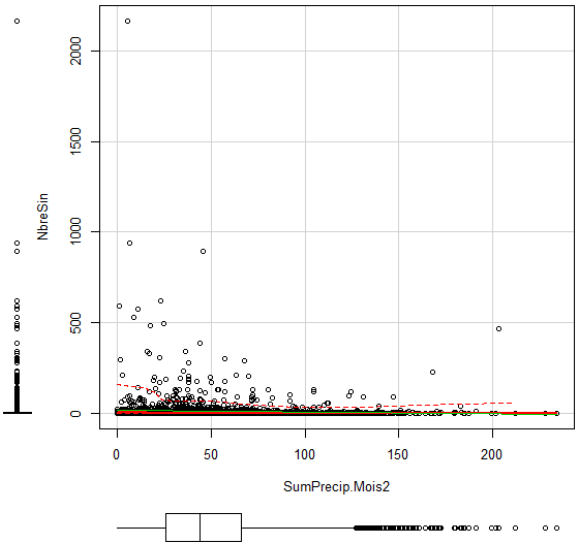




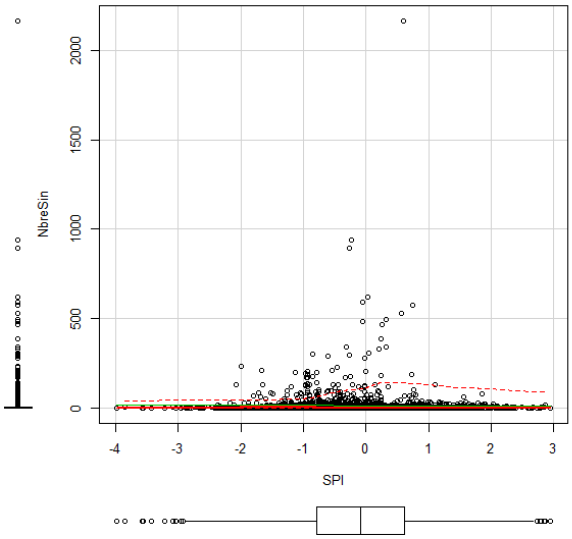
Précipitation cumulée mensuelle

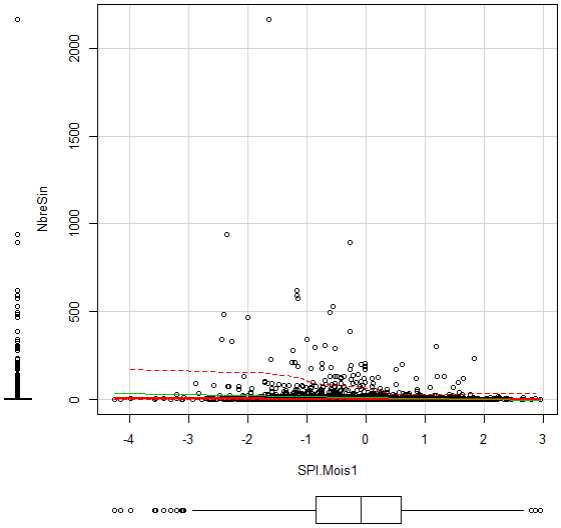


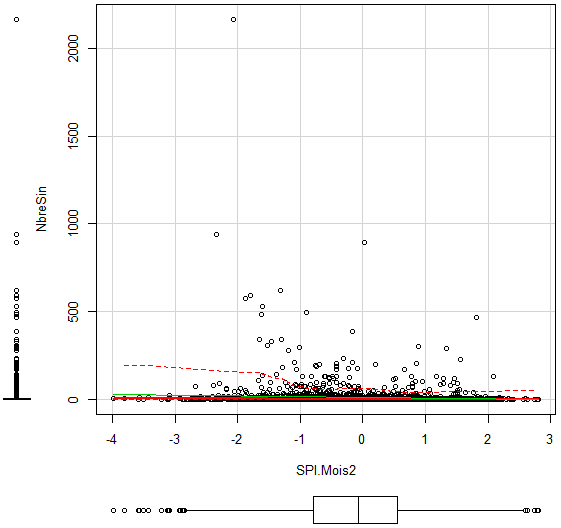




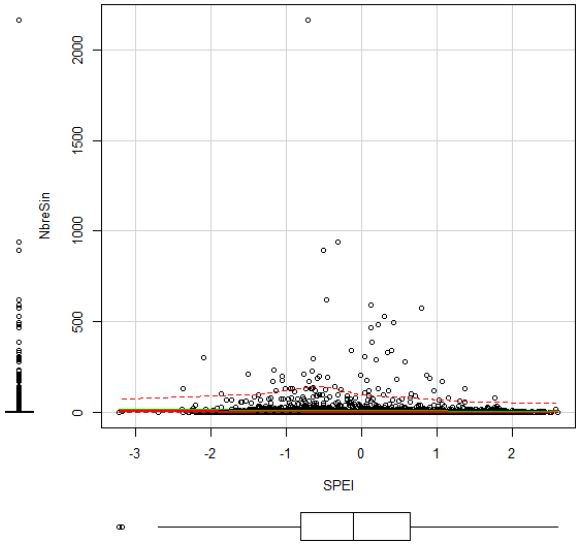
Indice SPI

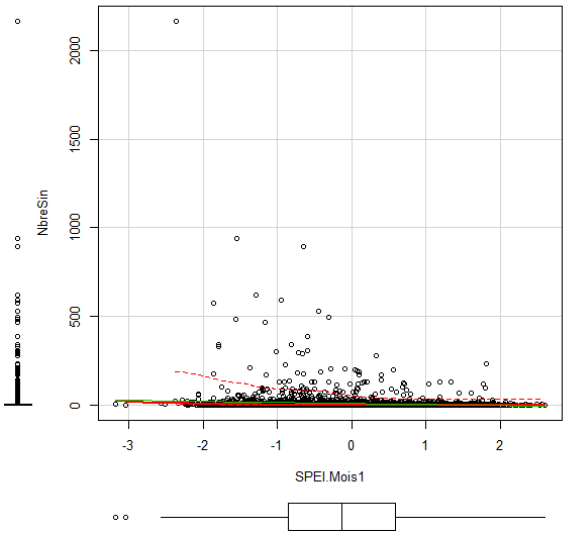

NbreSin
0
500
1000
1500
2000
-4
-3
-2
-1
0
1
2
3
SPI

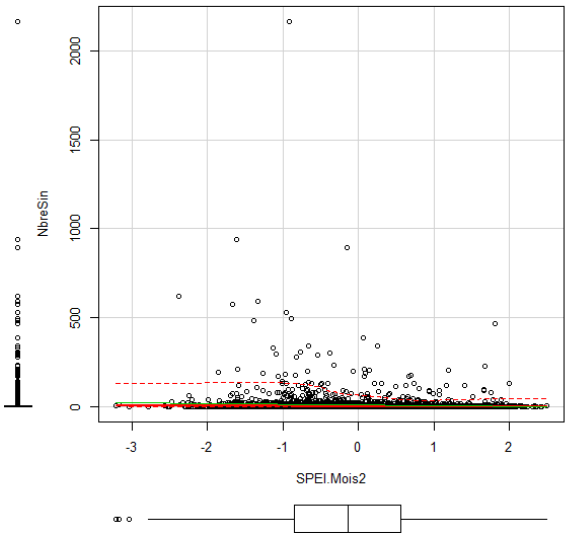

NbreSin
0
500
1000
1500
2000
-4
-3
-2
-1
0
1
2
3
SPI.Mois1


NbreSin
0
500
1000
1500
2000
-4
-3
-2
-1
0
1
2
3
SPI.Mois2

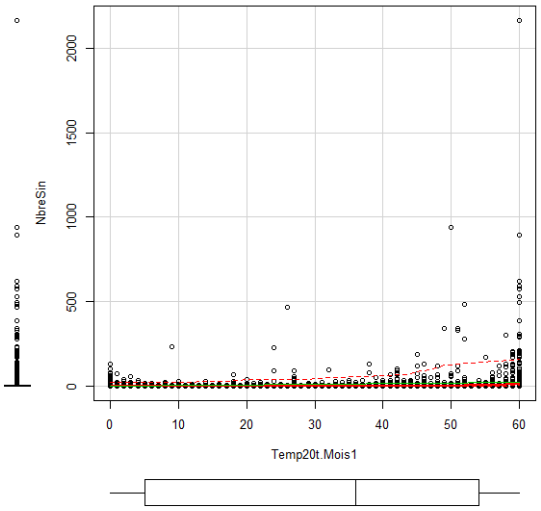
Indice SPEI

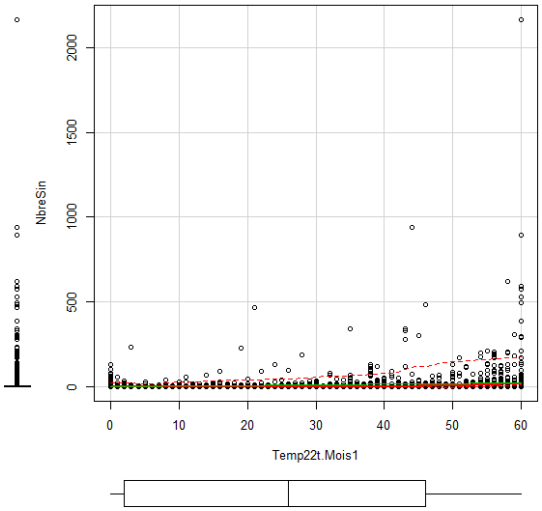

NbreSin
0
500
1000
1500
2000
-3
-2
-1
0
1
2
SPEI

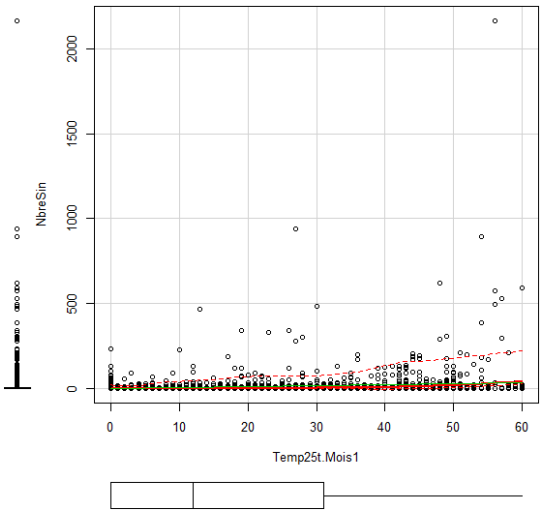

NbreSin
0
500
1000
1500
2000
-3
-2
-1
0
1
2
SPEI.Mois1

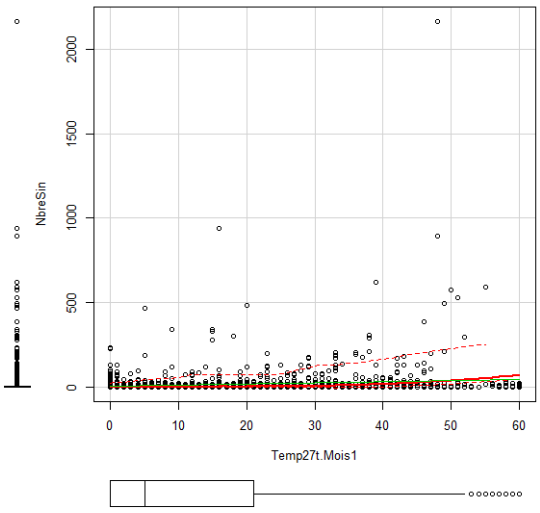

NbreSin
0
500
1000
1500
2000
-3
-2
-1
0
1
2
SPEI.Mois2

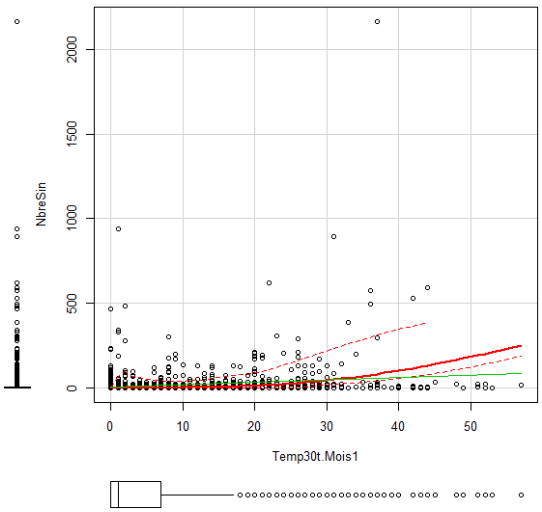
Indicateur à seuil

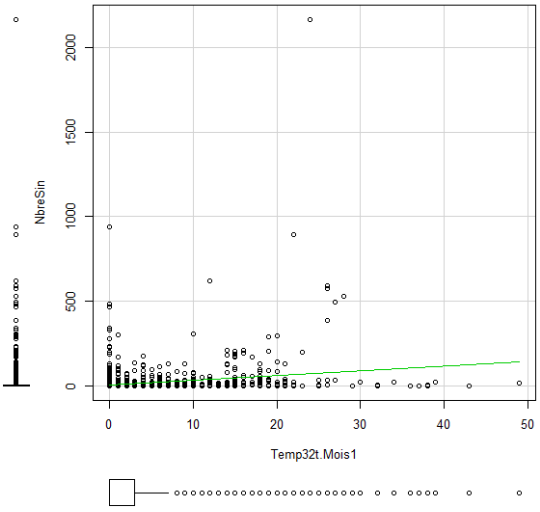












Annexe D : Tests statistiques

Test de Student

Le test de Student est souvent utilisé pour tester la significativité d'un coefficient dans le cadre d'une régression linéaire.

Soit β le coefficient étudié. Soit $\hat{\beta}$ son estimateur. On souhaite tester :

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

On introduit alors la statistique de test :

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\mathbb{V}(\hat{\beta})}}$$

Sous H_0, T suit asymptotiquement une loi de Student. Pour un niveau de confiance α, la zone de rejet de H_0 est donc :

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

où $t_{\alpha,h}$ est le quantile de niveau α d'une loi de Student avec d degrés de liberté (dépendant du nombre de paramètres en entrée).

Test de Ljung-Box

Le test de Ljung-Box est souvent utilisé pour tester l'indépendance des distributions dans une série temporelle.

On souhaite tester :

$$\begin{cases} H_0 : \text{Les données sont indépendamment distribuées} \\ H_1 : \text{Les données ne sont pas réparties de façon indépendante} \end{cases}$$

On introduit alors la statistique de test :

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

où n est la taille de l'échantillon, $\widehat{p_k}$ est l'autocorrélation de l'échantillon au lag k, et h est le nombre lags testés.

Sous H_0, Q suit une loi χ^2_h. Pour un niveau de confiance α, la zone de rejet de H_0 est donc :

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

où $\chi^2_{1-\alpha,h}$ est le quantile de niveau α d'une loi χ^2_h (h est le degré de liberté).

Test KPSS

Le test KPSS (Kwiatkowski–Phillips–Schmidt–Shin) est souvent utilisé pour tester la stationnarité d'une série temporelle.

On part du modèle suivant :

$$Y_t = \mu_t + \varepsilon_t$$

où ε est stationnaire, et μ vérifie :

$$\mu_t = \mu_{t-1} + u_t \quad avec\ u_t \sim idd(0, \sigma_u^2)$$

On souhaite tester :

$$\begin{cases} H_0 : \text{ Y est stationnaire} \\ H_1 : \text{ Y n'est pas stationnaire} \end{cases}$$

Cela revient à dire :

$$\begin{cases} H_0 : \sigma_u^2 = 0 \text{ ou } \mu \text{ est constante} \\ H_1 : \sigma_u^2 > 0 \end{cases}$$

On introduit alors la statistique de test :

$$LM = \frac{1}{\widehat{S^2}} \sum_{t=1}^{T} S_t^2$$

$S_t = \sum_{r=1}^{t} e_r$ où (e_t) est la série des résidus

S^2 est la « variance de long-terme ». Elle est égale à :

$$S^2 = \lim_{T \to +\infty} \frac{\mathbb{E}(S_T^2)}{T}$$

Elle est estimée par :

$$\widehat{S^2} = \frac{1}{T} \sum_{t=1}^{T} {e_t}^2 + \frac{2}{T} \sum_{i=1}^{l} \left(1 - \frac{i}{l+1}\right) \sum_{t=i+1}^{T} e_t e_{t-i}$$

On rejette H_0 lorsque LM est supérieure à sa valeur critique. Les valeurs critiques sont données par le tableau ci-dessous :

Valeurs critiques du test KPSS		
10 %	5 %	1 %
0,347	0,463	0,739
0,119	0,146	0,216

Test de Wald

Le test de Wald est souvent utilisé pour tester la significativité d'un ensemble de coefficients dans le cadre d'un modèle linéaire généralisé.

Soit β le vecteur des coefficients étudiés. Soit $\hat{\beta}$ son estimateur. On souhaite tester :

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

On introduit alors la statistique de test :

$$T_W = n(\hat{\beta} - \beta)' \boldsymbol{I}(\hat{\beta})(\hat{\beta} - \beta)$$

où n est la taille de l'échantillon et $\boldsymbol{I}$ est l'information de Fisher.

Sous H_0, T_W suit une loi χ^2_h. Pour un niveau de confiance α, la zone de rejet de H_0 est donc :

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

où $\chi^2_{1-\alpha,h}$ est le quantile de niveau α d'une loi χ^2_h (h est le degré de liberté).

Annexe E : Ajustement de modèle – Coefficients

Rappel

Variables	Nom dans le modèle
Nombre de sinistres enregistrés	$NbreSin$
Classe de sinistralité	$ClassSin$
Niveau moyen d'aléa retrait-gonflement des argiles	$nALEA$
Précipitation cumulée mensuelle	$SumPrecip$
Classe de température	$ClassTemp$
Température maximale mensuelle	$Tmax$
Nombre maximal mensuel du nombre de jours où la température est supérieure à 30°C	$Temp_{30}$
Indice *SPI*	SPI
Indice *SPEI*	$SPEI$

Tableau 6 - Les variables pour modéliser la sécheresse

Pour un mois donné, on s'intéresse aux valeurs prises par ces variables les deux derniers mois. Par exemple, la valeur du mois dernier de la température maximale enregistrée est notée $Tmax.Mois1$.

Modèle de fréquence de sinistralité – Poisson

Le tableau suivant donne les coefficients d'ajustement d'un modèle de Poisson **sans** indicateur à seuil.

	Coefficient	Erreur	Statistique de Wald z	Pr(>|z|)	
(Intercept)	-12,9	0,10	-122,85	< 2e-16	***
ClassSin2	2,21	0,043	50,63	< 2e-16	***
ClassSin3	3,61	0,044	82,11	< 2e-16	***
ClassSin4	3,85	0,046	82,88	< 2e-16	***
ClassSin5	4,60	0,047	98,70	< 2e-16	***
ClassTemp2	-0,33	0,065	-5,08	3,84e-07	***
ClassTemp3	-0,23	0,023	-10,01	< 2e-16	***
ClassTemp4	-0,35	0,036	-9,54	< 2e-16	***
ClassTemp5	-0,30	0,034	-8,75	< 2e-16	***
nALEA	0,020	0,0068	2,88	0,0040	**
Tmax.Mois1	0,20	0,0029	70,82	< 2e-16	***
Tmax.Mois2	0,094	0,0024	38,82	< 2e-16	***
SumPrecip.Mois1	0,019	0,00051	37,17	< 2e-16	***
SumPrecip.Mois2	0,0096	0,00046	20,86	< 2e-16	***
SPEI.Mois1	-0,26	0,0090	-28,96	< 2e-16	***
SPEI.Mois2	-0,051	0,0091	-5,58	2,48e-08	***
SPI.Mois1	-0,78	0,013	-59,99	< 2e-16	***
SPI.Mois2	-0,56	0,013	-43,94	< 2e-16	***

Tableau 7 - Paramètres de régression – Modèle de Poisson

Modèle de fréquence de sinistralité – Poisson avec indicateur à seuil

Le tableau suivant donne les coefficients d'ajustement d'un modèle de Poisson **avec** indicateur à seuil.

	Coefficient	Erreur	Statistique de Wald z	Pr(>|z|)	
(Intercept)	-11,14	0,11	-105,52	< 2e-16	***
ClassSin2	2,16	0,044	49,44	< 2e-16	***
ClassSin3	3,70	0,044	83,72	< 2e-16	***
ClassSin4	3,71	0,047	79,62	< 2e-16	***
ClassSin5	4,55	0,047	97,28	< 2e-16	***
ClassTemp2	-0,31	0,065	-4,84	1,28e-06	***
ClassTemp3	-0,083	0,023	-3,59	0,00033	***
ClassTemp4	-0,35	0,037	-9,55	< 2e-16	***
ClassTemp5	-0,19	0,034	-5,60	2,16e-08	***
nALEA	0,074	0,0069	10,71	< 2e-16	***
Tmax.Mois1	0,23	0,0029	78,09	< 2e-16	***
Tmax.Mois2	-0,0017	0,0026	-0,64	0,52	
Temp30t.Mois1	-0,047	0,0016	-29,15	< 2e-16	***
Temp30t.Mois2	0,12	0,0016	71,30	< 2e-16	***
SumPrecip.Mois1	0,017	0,00054	31,25	< 2e-16	***
SumPrecip.Mois2	0,013	0,00043	30,86	< 2e-16	***
SPEI.Mois1	-0,13	0,0091	-14,35	< 2e-16	***
SPEI.Mois2	-0,088	0,0092	-9,51	< 2e-16	***
SPI.Mois1	-0,83	0,014	-60,21	< 2e-16	***
SPI.Mois2	-0,58	0,013	-45,38	< 2e-16	***

Tableau 8 - Paramètres de régression – Modèle de Poisson avec indicateur à seuil

Modèle de fréquence de sinistralité – Binomiale négative

Le graphique suivant présente le résultat de l'ajustement d'un modèle linéaire généralisé avec les mêmes hypothèses sur les variables que dans le cas du modèle de Poisson.

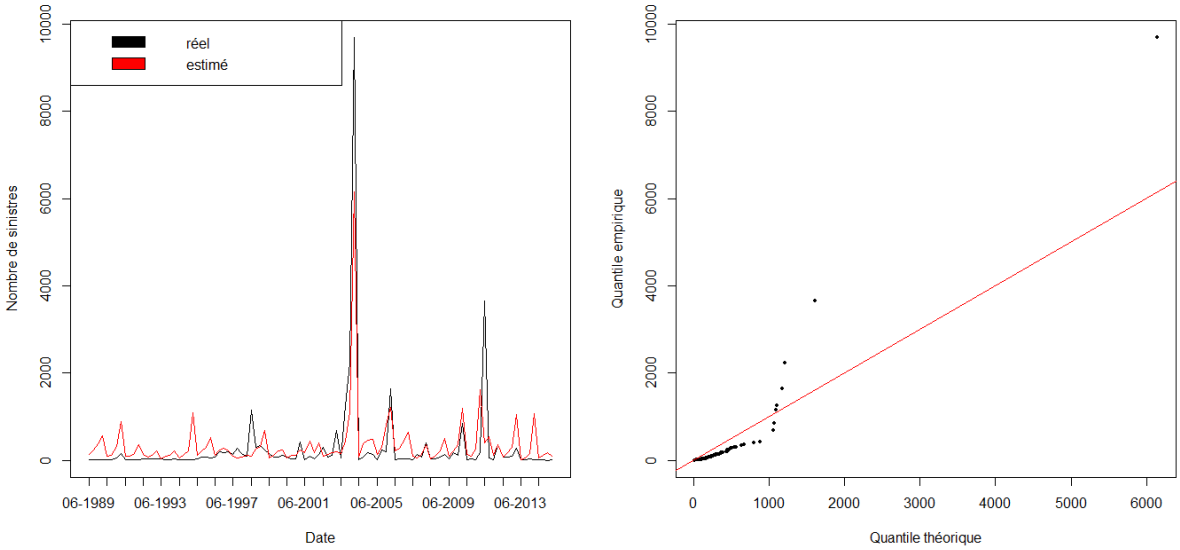


Figure 57 - Modèle de fréquence de sinistralité – Binomiale négative

Le tableau suivant donne les coefficients de l'ajustement.

	Coefficient	Erreur	Statistique de Wald z	Pr(>|z|)	
(Intercept)	-9,04	0,48	-18,97	< 2e-16	***
ClassSin2	2,24	0,091	24,59	< 2e-16	***
ClassSin3	3,69	0,11	33,87	< 2e-16	***
ClassSin4	3,98	0,15	26,65	< 2e-16	***
ClassSin5	4,89	0,21	23,35	< 2e-16	***
nALEA	-0,041	0,027	-1,51	0,13	
ClassTemp2	-0,73	0,16	-4,52	6,10e-06	***
ClassTemp3	-0,38	0,10	-3,77	0,00016	***
ClassTemp4	-0,31	0,14	-2,25	0,025	*
ClassTemp5	-0,22	0,13	-1,75	0,080	.
Tmax.Mois1	0,21	0,017	12,66	< 2e-16	***
Tmax.Mois2	-0,0013	0,013	-0,10	0,92	
SumPrecip.Mois1	0,0088	0,0025	3,54	0,00041	***
SumPrecip.Mois2	-0,0016	0,0022	-0,69	0,49	
Temp30t.Mois1	-0,041	0,011	-3,80	0,00015	***
Temp30t.Mois2	0,11	0,014	7,94	1,94e-15	***
SPEI.Mois1	0,074	0,049	1,52	0,13	
SPEI.Mois2	-0,058	0,049	-1,19	0,23	
SPI.Mois1	-0,65	0,078	-8,31	< 2e-16	***
SPI.Mois2	-0,18	0,076	-2,33	0,020	*

Tableau 9 - Paramètres de régression – Modèle Binomiale négative

Ce modèle n'est pas plus performant que celui de Poisson.

Table des figures

Figure 1 - Représentation schématique du bilan évapotranspiration – Source : Wikipédia 3
Figure 2 - Conséquences de la sécheresse 4
Figure 3 - Températures et précipitations au printemps de 1959 à 2015 – Source : Météo France 5
Figure 4 - Fréquence des catastrophes naturelles sur la période 1988-2006 – Source : FFSA 6
Figure 5 - Coût global des catastrophes naturelles en millions d'euros – Source : FFSA 7
Figure 6 - Impact de la réassurance sur le résultat technique – Source : FFSA 9
Figure 7 - Calcul du *SCR* par une combinaison de modules – Source : *EIOPA* 11
Figure 8 - Reconnaissance CAT NAT – Source : CCR 12
Figure 9 - Fréquence moyenne des sinistres – Source : CCR 13
Figure 10 - Coûts par départements – Source : CCR 13
Figure 11 - Coût moyen des sinistres – Source : CCR 14
Figure 12 - Phénomène de retrait-gonflement des argiles – Source : BRGM 17
Figure 13 - Exemple d'une courbe de vulnérabilité pour un type d'objet assuré – Source : EQUECAT 19
Figure 14 - Courbes *AEP* et *OEP* 21
Figure 15 - Sinistralité et précipitations mensuelles 24
Figure 16 - Sinistralité et précipitations mensuelles (2) 25
Figure 17 - Sinistralité et *SPI* 26
Figure 18 - Sinistralité et température maximale mensuelle 27
Figure 19 - Sinistralité et température maximale mensuelle (2) 28
Figure 20 - Sinistralité et évapotranspiration 30
Figure 21 - Sinistralité et *SPEI* 31
Figure 22 - Comparaison de la répartition de l'aléa retrait-gonflement des argiles avec la répartition du nombre de sinistres liés à la sécheresse 32
Figure 23 - Boîtes à moustaches des précipitations mensuelles par année, tous *CRESTA* et mois confondus 35
Figure 24 - Boîtes à moustaches des précipitations cumulées mensuelles, tous *CRESTA* et années confondus 35
Figure 25 - Tous *CRESTA* et mois confondus : *Q-Q plot* des précipitations mensuelles avec une loi Normale, Exponentielle et Gamma 38
Figure 26 - *CRESTA* 92 en juin : *Q-Q plot* des précipitations mensuelles avec une loi Gamma 39
Figure 27 - Taux d'inertie expliqué pour chacun des axes obtenus 42
Figure 28 - Dendrogramme issu de la CAH 45
Figure 29 - Résultats obtenus avec la CAH pour différents nombres de classes 46
Figure 30 - Evolution des températures maximales dans la région Sud 47
Figure 31 - Evolution de la tendance 48
Figure 32 - Evolution de la saisonnalité 49
Figure 33 - Evolution de la série résiduelle 50
Figure 34 - Boîtes à moustaches mensuelles de ($\boldsymbol{Zt}$) 51
Figure 35 - Ecart-type de $\boldsymbol{Zt}$ pour différents paramètres de lissage h 52
Figure 36 - Estimation de l'écart-type de $\boldsymbol{Zt}$ 52
Figure 37 - Boîte à moustaches des résidus $\boldsymbol{\varepsilon t}$ 54
Figure 38 - *Q-Q plot* de la distribution des résidus ($\boldsymbol{\varepsilon t}$) avec celle d'une normale standard 54

Figure 39 - *Backtesting* des températures dans la région Sud 55
Figure 40 - Copules empiriques entre régions 60
Figure 41 - Comparaison de la copule empirique avec les copules gaussiennes et de Student paramétrées 61
Figure 42 - Fonctions de Kendall empiriques et théoriques pour différentes copules 61
Figure 43 - Exemple de scénario de température 64
Figure 44 - Un scénario annuel des variables explicatives de la sécheresse 64
Figure 45 - Indépendance fréquence/coût 65
Figure 46 - Dendrogramme pour la fréquence de sinistralité 70
Figure 47 - Répartition des classes de sinistralité 70
Figure 48 - Modèle de fréquence de sinistralité – Poisson 71
Figure 49 - Sinistralité et indicateur à seuil 73
Figure 50 - Modèle de fréquence de sinistralité – Poisson avec indicateur à seuil 74
Figure 51 - Modèle de fréquence de sinistralité – Poisson croisé 75
Figure 52 - Modèle de fréquence de sinistralité – Forêt aléatoire 79
Figure 53 - Relation log-linéaire entre les taux de destruction et les sommes assurées 84
Figure 54 - Courbes de vulnérabilité 86
Figure 55 - Ajustement perte estivale/perte annuelle 88
Figure 56 - Courbes *AEP* et *OEP* avec une approche fréquence/coût 92
Figure 57 - Modèle de fréquence de sinistralité – Binomiale négative 113

Liste des tableaux

Tableau 1 - Variances intra-classe 46
Tableau 2 - Corrélations empiriques des résidus entre régions 60
Tableau 3 - Paramètres de régression – Taux de destruction et somme assurée 85
Tableau 4 - Paramètres de régression – Perte estivale et Perte annuelle 88
Tableau 5 - Evapotranspiration – Coefficients de correction 96
Tableau 6 - Les variables pour modéliser la sécheresse 110
Tableau 7 - Paramètres de régression – Modèle de Poisson 111
Tableau 8 - Paramètres de régression – Modèle de Poisson avec indicateur à seuil 112
Tableau 9 - Paramètres de régression – Modèle Binomiale négative 113

Bibliographie

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

BREIMAN L. (2001), *Random Forests.* Machine Learning 45: 5-32

CCR (Caisse Centrale de Réassurance) (2011)*, Le régime d'indemnisation des catastrophes naturelles*

CCR (Caisse Centrale de Réassurance), *https://erisk.ccr.fr/faces/erisk-generalites-perils-secheresse.jsp (synthèse des principaux sinistres liés à la sécheresse)*

COHEN-SALMON L., GNINGHAYE FONGANG D.J. (2014) *Modélisation du risque gel en France.* Mémoire d'actuariat

CHABOT M., *Concepts de dépendance et copule,* CaMUS 4: 48-71

CHARPENTIER A., S*éries temporelles: théorie et applications.* Cours de l'ENSAE et de l'Université Paris-Dauphine

CHARPENTIER A., DUTANG C. (2012), *L'Actuariat avec R*

D'ANTIN H. (2014) *Modélisation du risque de crue de la Seine en région parisienne.* Mémoire d'actuariat

DENUIT M., CHARPENTIER A. (2005), *Mathématiques de l'assurance non-vie.* Economica

EUROPEAN CLIMATE ASSESSMENT & DATASET, http://eca.knmi.nl/ *(données des précipitations et des températures journalières)*

FFSA (Fédération Française des Sociétés d'Assurances) (2009), *Synthèse de l'étude relative à l'impact du changement climatique et de l'aménagement du territoire sur la survenance d'événements naturels en France*

GEORISQUES, http://www.georisques.gouv.fr/dossiers/argiles/donnees#/ (*données de l'aléa retrait-gonflement des argiles*)

GROSSI P., KUNREUTHER H. (2005), *Catastrophe modeling: a new approach to managing risk.* Springer

HASTIE T., TIBSHIRANI R., FRIEDMAN J. (2008), *The Elements of Statistical Learning: data mining, inference and prediction (2nd ed.).* Springer

HUSSON F., LE S., PAGES J. (2009), *Analyse de données avec R.* Presses Universitaires de Rennes.

MRN (Association Mission Risques Naturels) (2015), *L'assurance des catastrophes naturelles en 2013*

Projet ARGIC (Analyse du retrait-gonflement et de ses Incidences sur les Constructions) (2009), *Rapport de synthèse finale.* Centre de Géosciences, BRGM, LMSSMat, CERMES, Fondasol, INERIS, LAEGO, INRA, LGCIE, LCPC, Météo-France, GHYMAC, Université de Poitiers

Projet ClimSec (2011), *Impact du changement climatique sur la sécheresse et l'eau du sol.* Météo-France, CNRS, CERFACS, UMR SISYPHE, CEMAGREF, FONDATION MAIF

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

SWISS RE (2003), *Catastrophes naturelles et réassurance*

THORNTHWAITE C. W. (1948), *An approach toward a rational classification of climate*. *Geographical Review* 38: 55–94

VICENTE-SERRANO S.M., BEGUERIA S., LOPEZ-MORENO J.I. (2010), *A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index.* Journal of Climate 23: 1696-1718