Institut des ACTUAIRES

Dauphine | PSL
UNIVERSITÉ PARIS

Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires

le

Par : Thomas KERMORVANT
Titre : Modélisation et projection du risque sécheresse : Étude de soutenabilité à horizon 2050

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise : Mazars Actuariat
Nom : Mazars Actuariat
Signature :

Mazars Actuariat
Société par Actions Simplifiée
61, rue Henri Regnault - 92075 Paris - La Défense Cedex
Siret : 342 405 321 00049 - APE 6920Z
RCS Nanterre 342 405 321
Siège social : 61, rue Henri Regnault - 92400 COURBEVOIE

Directeurs de Mémoire en entreprise :
Nom : Alex-Michel Ngningha
Pauline Berger
Signature :

Membres présents du Jury du Certificat d'Actuaire de Paris-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 :

Signature du candidat

Bibliothèque :

Résumé

Le changement climatique est une réalité qui s'impose comme un enjeu majeur du siècle. Les différentes études menées autour de cette problématique prévoient la multiplication des évènements climatiques extrêmes et le secteur de l'assurance est directement concerné. En effet, l'écosystème assurantiel est exposé à plusieurs risques découlant de ce dérèglement climatique, en particulier le risque sécheresse, identifié comme un risque en pleine croissance. Les montants des sinistres liés aux évènements de sécheresse ont fortement augmenté au cours des dernières années, et cette tendance devrait s'amplifier dans les années futures. Ces évolutions à la fois en fréquence et en sévérité interroge naturellement sur la pérennité du système assurantiel actuel. En particulier, cette croissance des sinistres climatiques amène les assureurs à réajuster les primes d'assurance afin de maintenir leur rentabilité.

Lors de son premier exercice pilote, l'ACPR (2021) signalait déjà un problème de viabilité des primes à moyen terme qui augmenteraient entre 130 et 200% pour couvrir l'accroissement des sinistres climatiques. Cette envolée des tarifs rendrait les primes inabordables pour un certain nombre de ménage. En réponse à ce premier constat, l'ACPR (2023) a intégré, dans son deuxième exercice pilote, la notion de soutenabilité de la prime du côté des assurés. Pour continuer dans ce sens et explorer davantage cette problématique, l'approche du mémoire consiste à étudier la soutenabilité de la prime à horizon 2050 à travers le péril sécheresse.

Concentrant son analyse sur le secteur de l'assurance habitation (MRH), ce mémoire propose dans un premier temps une modélisation du péril sécheresse à travers des techniques de *machine-learning* afin d'aboutir dans un second temps à une projection de la sinistralité associée à climat futur. Enfin, compte tenu des projections obtenues, le mémoire se consacre dans un dernier temps à une analyse concernant la progression et la soutenabilité des primes couvrant les sinistres liés à la sécheresse.

Cette étude a permis à la fois de quantifier l'évolution des primes en réponse à la croissance du risque, mais également de mettre en lumière, au moyen d'un indicateur de richesse, les territoires exposés au risque d'insoutenabilité de la prime du point de vue des assurés.

Mots-clés : *Sécheresse*, *Changement climatique*, *Soutenabilité*, *Crédibilité hiérarchique*, *Machine-learning*, *Mutualisation*.

Abstract

Climate change is a reality that is emerging as one of the major challenges of this century. The various studies carried out on this issue predict an increase in the number of extreme climatic events, and the insurance sector is directly concerned. Indeed, the insurance ecosystem is exposed to a number of risks arising from climate disruption, in particular subsidence, which is one of the consequences of drought, has been identified as a fast-growing risk. Losses due to subsidence events have risen sharply in recent years, and this trend is set to intensify in the years ahead. These trends in both frequency and severity naturally raise questions about the sustainability of the current insurance system. In particular, this increase in weather-related claims is leading insurers to readjust insurance premiums in order to maintain their profitability.

During its first pilot exercise, ACPR (2021) was already pointing to a problem with the viability of premiums in the medium term, which would rise by between 130 and 200% to cover the increase in weather-related claims. This surge in rates would make premiums unaffordable for a number of households. In response to this initial observation, the ACPR (2023) integrated the notion of premium sustainability on the policyholder side into its second pilot exercise. To continue in this vein and explore this issue further, the approach of this master thesis is to study premium sustainability to 2050 through the peril of subsidence.

Focusing its analysis on the home insurance sector, this master thesis firstly proposes a model of the peril of subsidence using machine-learning techniques, in order to arrive at a projection of the claims experience associated with a future climate. Finally, taking into account the projections obtained, the dissertation is devoted to an analysis of the progression and sustainability of premiums covering drought-related claims.

This study not only quantifies premium growth in response to increased risk, but also highlights, through a wealth indicator, the territories exposed to the risk of premium unsustainability from the policyholders' perspective.

Keywords : *Drought*, *Climate change*, *Sustainability*, *Hierarchical credibility*, *Machine-learning*, *Mutualization*.

Note de Synthèse

Introduction

En France, le risque de retrait-gonflement des argiles (RGA) est le risque climatique qui présente la plus importante croissance depuis 2016, tant en termes de fréquence que de sévérité. La sinistralité due au phénomène RGA s'est en effet accélérée ces dernières années avec une charge annuelle moyenne qui a dépassé le milliard d'euros sur la période 2016-2020 et des événements de plus en plus intenses comme en témoigne l'année 2022, record en termes de coût engendré (CCR (2023b)). Cette hausse de la sinistralité s'explique en grande partie par une dégradation des conditions climatiques sur le territoire métropolitain. De nombreuses études climatiques effectuées en France (Soubeyroux et al. (2011), Gourdier et Plat (2018)) et en Europe (Spinoni et al. (2015)) confirment l'influence du changement climatique sur la manifestation des événements de sécheresse et prévoient une aggravation des épisodes de sécheresse tout au long du XXième siècle.

Face à cette multiplication annoncée des événements de sécheresse, le secteur de l'assurance a cherché à quantifier cette sinistralité croissante à horizon 2050. Selon les hypothèses retenues, le montant des dommages relatifs à la sécheresse augmenterait entre 44% et 162%. Dans sa dernière étude, la CCR (2023a) qualifie même le péril sécheresse de "*péril le plus préoccupant à horizon futur*". Cette progression des risques climatiques prévue dans les années à venir, notamment celle du phénomène de sécheresse, interroge naturellement sur la pérennité de l'écosystème assurantiel actuel. En particulier, les assureurs sont amenés à réajuster les primes d'assurances pour garantir une certaine rentabilité. Lors de son premier exercice pilote climatique en 2021, l'ACPR (2021) pointait déjà un problème de viabilité des primes à moyen terme qui augmenteraient entre 130% et 200% pour couvrir l'accroissement des sinistres climatiques. En réponse à ce premier constat, l'ACPR (2023) a intégré, dans son deuxième exercice pilote, la notion de soutenabilité de la prime du côté des assurés. Pour continuer dans ce sens et explorer davantage cette problématique, le mémoire propose d'étudier la question de la soutenabilité de la prime à horizon 2050 à travers le péril sécheresse.

Pour répondre à la problématique, une première étape de modélisation du risque sécheresse est effectuée à l'aide de techniques de *machine learning*. Ensuite, les données relatives au risque sécheresse sont projetées afin d'aboutir à une estimation de la charge sinistre à horizon 2050. Enfin, la dernière partie de l'étude consiste à calculer la prime Multi-Risque Habitation (MRH) future et à analyser la capacité des assurés à supporter sa progression selon plusieurs scénarios de tarification.

Modèle sécheresse

L'objectif de cette première partie consiste à développer un modèle capable d'estimer la charge sinistre sécheresse annuelle départementale, notée Y, constituant ainsi une base cohérente pour la suite de l'étude. Pour cela, l'entraînement du modèle a nécessité le croisement de données de sinistres, de données d'exposition et de données climatiques.

La base de sinistre considérée pour l'étude comprend un historique de sinistres sécheresse survenus entre 2000 et 2020 d'un portefeuille MRH couvrant l'ensemble du territoire métropolitain. Afin de garantir une comparaison cohérente de la charge sur la période retenue, une démarche de traitement "*as-if*" des données a été effectuée. Ce traitement est double car il vise à corriger à la fois le montant des sinistres mais également l'exposition du portefeuille.

L'apparition d'un sinistre RGA résulte principalement de la combinaison de facteurs de prédisposition et de facteurs déclenchants. Le principal facteur de prédisposition est la nature du sol. Les sols argileux, de par leur structure malléable, vont se rétracter lors de périodes très sèches et se gonfler pendant les périodes humides. Ainsi, pour alimenter le modèle, une première étape consiste à intégrer des données d'exposition, telles que la nature géologique du sol, la proportion de logements dans des zones argileuses et également le type de bien assuré.

Par la suite, la succession de périodes sèches et humides constitue le facteur déclenchant et abouti à une déformation des sols argileux. De ce fait, la deuxième étape vise à construire des indices de sécheresse capables de faire le lien entre la variation des conditions climatiques et la manifestation du phénomène RGA. Le premier indice construit est l'indice météorologique SPEI (*Standardized Precipitation Evapotranspiration Index*) (Vicente-Serrano et al. (2010)) qui dépend uniquement des données de température, de durée d'ensoleillement et de précipitations. Le deuxième indice mis en place est l'indice de magnitude SWI (*Soil Wetness Index*). Cet indice hydrologique, élaboré spécifiquement pour cette étude, se base sur l'indice d'humidité des sols de Météo France. Ce dernier présente une courbe annuelle sinusoïdale caractérisée par des valeurs plus faibles pendant la période estivale, synonyme de sécheresse. L'indice de magnitude SWI est alors obtenu en ajustant une courbe sur les données mensuelles de SWI et en calculant l'intégrale de la courbe ajustée sous une valeur seuil γ. L'indice ainsi construit, correspond à une mesure d'intensité de sécheresse où une valeur positive élevée caractérisera un épisode de sécheresse intense.

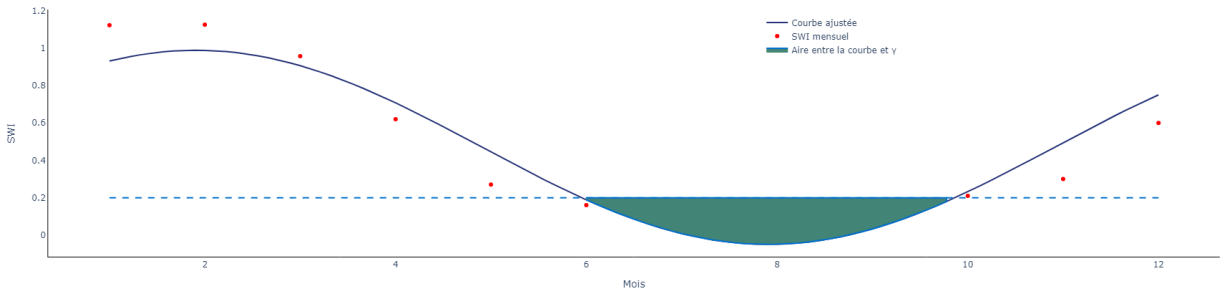


Figure 1 : Calcul de l'indice de magnitude SWI(γ).

Ensuite, la consolidation des indices de sécheresse, des données d'exposition et des données de sinistres a permis de constituer un jeu de données $(x_i, y_i)_{i\in[\![1,n]\!]}$, avec x_i le vecteur des variables explicatives et y_i la réalisation de la variable cible Y. A partir de ces observations, une méthode d'apprentissage supervisé a été employée pour élaborer un modèle prédictif répondant aux objectifs de l'étude. Le principe fondamental de ces méthodes réside dans la construction d'une fonction de prédiction $\hat{f}$ qui minimise le risque empirique

$$\hat{\mathcal{R}}_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \hat{f}(x_i)),$$

en utilisant une fonction de perte l.

Pour approcher la fonction de prédiction qui minimise le risque empirique, l'algorithme *Catboost* (Prokhorenkova et al. (2018)) a été utilisé. Afin d'évaluer l'efficacité du modèle élaboré, une approche de validation intégrant des métriques standard telles que le R^2 et la MAE (*Mean Absolute Error*) a été adoptée. De plus, pour évaluer la capacité du modèle à détecter les pics de sinistralité, la métrique MAPEX(q) (*Mean Absolute Percentage Error for eXtreme values*) a été introduite spécifiquement pour le mémoire. Cette dernière permet de mesurer uniquement l'erreur du modèle commise sur les événements extrêmes de sécheresse. Par la suite, ces métriques ont été calculées selon deux approches de validation croisée afin de prendre en compte le caractère spatial et temporel des données utilisées. Ces procédures de validation visent à limiter l'apparition d'un biais entre les données d'entraînement et de test et ainsi garantir que le modèle soit capable de généraliser d'une année à l'autre et d'un département à l'autre.

Il a par la suite été nécessaire d'optimiser le modèle. En effet, une des principales difficultés rencontrées lors de modélisation résidait dans le caractère déséquilibré de la variable cible, propre aux sinistres climatiques, où 68% des observations du jeu de données ne présentaient aucune charge sinistre. Ce déséquilibre a une incidence directe sur le processus d'apprentissage, conduisant le modèle à sous-estimer la charge sinistre et à mal identifier les événements extrêmes de sécheresse. Pour palier à ce problème, un poids w_i a été ajouté sur les observations de sorte à pénaliser davantage les erreurs commises sur les sinistres majeurs et ainsi forcer le modèle à se corriger sur ces données. Cette pénalisation de la fonction de perte a permis d'améliorer les performances du modèle par rapport à un premier modèle de référence.

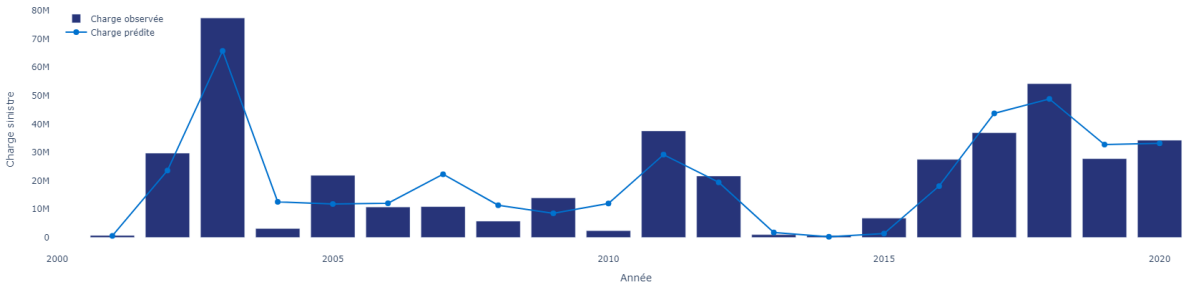
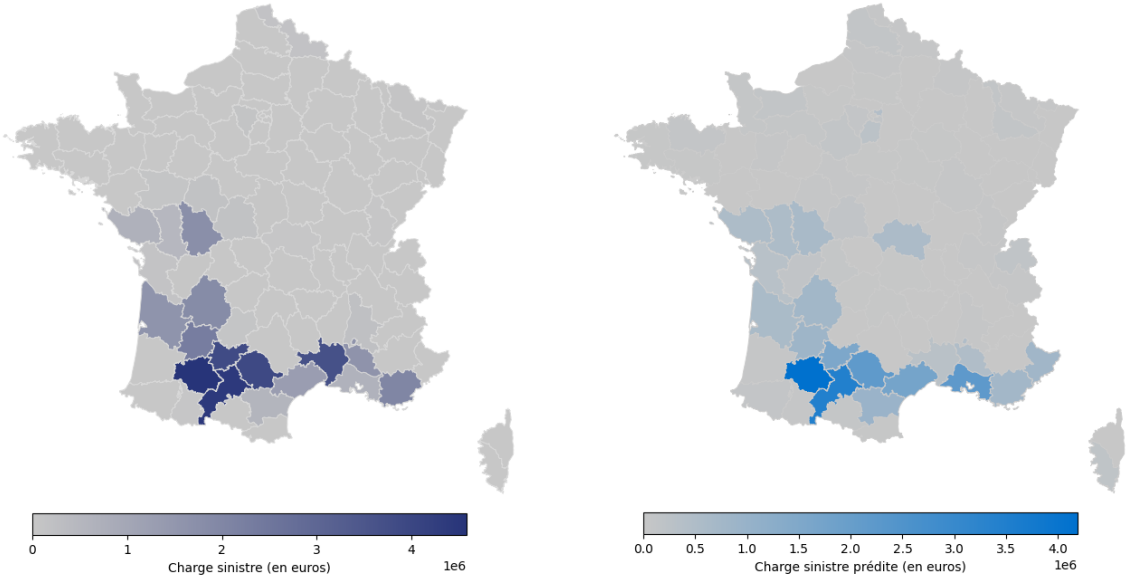


Figure 2 : Comparaison temporelle de la charge sinistre cumulée annuelle observée et prédite pour le modèle optimisé

L'analyse spatio-temporelle révèle que le modèle optimisé reproduit plus précisément les tendances de sinistralité, se démarquant par sa capacité à capter les pics de charge sinistre, notamment ceux de 2003, 2017 et 2018. Bien que l'optimisation ait réduit l'erreur dans les zones à forte sinistralité, elle a également entraîné une surestimation du montant des sinistres dans certains départements. Cependant, dans une logique prudentielle, les résultats du modèle optimisé restent préférables, justifiant son maintien pour la suite de l'étude.

Projection du risque

Afin d'aboutir à une estimation de la charge sinistre sécheresse à horizon 2050, la suite de l'étude consiste en la projection des éléments intervenant dans la prédiction du risque. Cette projection



(a) Charge sinistre observée - Année 2017 (b) Charge sinistre prédite - Année 2017

FIGURE 3 : Comparaison spatiale de la charge sinistre observée et prédite pour le modèle optimisé- Année 2017

concerne à la fois les données climatiques et les données d'exposition.

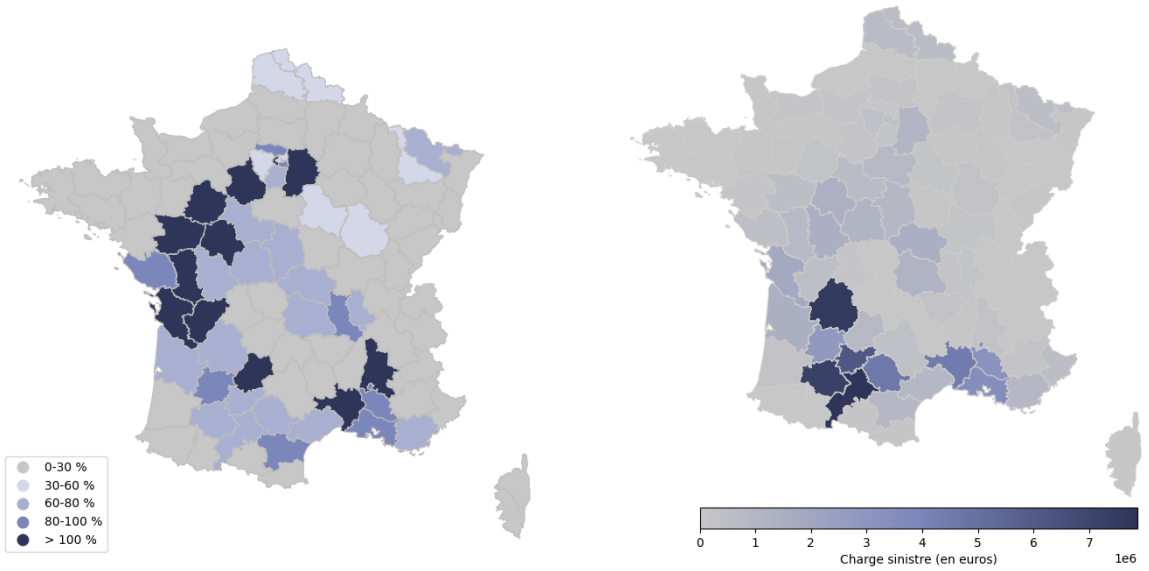
En vue de confronter le risque de sécheresse aux changements climatiques futurs, une approche pertinente consiste à utiliser des scénarios d'émissions futures de gaz à effet de serre. Les scénarios RCP du GIEC sont élaborés dans cette optique, offrant différentes trajectoires de forçage radiatif allant de la plus optimiste pour la RCP 2.6 à la plus pessimiste pour la RCP 8.5. Dans le cadre de l'étude, les variables climatiques et donc les indices de sécheresse ont été projetés à travers la trajectoire RCP 4.5. Ce choix permet notamment d'être cohérent avec les hypothèses du nouvel exercice pilote de l'ACPR. En parallèle des projections climatiques, les enjeux assurantiels du portefeuille sont modélisés à horizon 2050 en tenant compte à la fois de l'évolution démographique prédite par l'INSEE (2022) sur cette même période mais également de l'augmentation des valeurs assurées.

Les indices de sécheresse et les enjeux assurés projetés sont ensuite utilisés en entrée du modèle sécheresse construit précédemment pour obtenir une estimation de la sinistralité à horizon 2050.

Une première analyse, à exposition constante, met en évidence une augmentation de 61% de la perte moyenne annuelle due au changement climatique d'ici à 2050, selon le scénario RCP 4.5. Les évolutions les plus significatives se situent le long du croissant argileux, dans le Bassin parisien, le nord des Hauts-de-France, et le Centre. Cependant, la sinistralité future est concentrée, dans une large mesure, en Occitanie, en Nouvelle-Aquitaine et en Provence-Alpes-Côte d'Azur, où seulement six départements représentent 60% de la charge sinistre.

De plus, l'analyse des périodes de retour montrent que les événements extrêmes de sécheresse de type 2003, 2011 ou 2018 se produiront deux fois plus régulièrement à horizon 2050 sous le scénario RCP 4.5.

La suite de l'analyse consiste à appliquer cette fois-ci les hypothèses d'évolution démographique de l'INSEE et celles concernant l'augmentation des valeurs assurées. La charge finale projetée correspond alors à une augmentation de 160% de la charge initiale. Le principal facteur inflationniste à horizon



(a) Évolution de la charge sinistre à horizon 2050 (b) Montant de la charge sinistre à horizon 2050

FIGURE 4 : Évolution et montant de la charge sinistre sécheresse à horizon 2050 - RCP 4.5

	Charge sinistre (en euros)	Période de retour - Référence	Période de retour - RCP 4.5
Sécheresse 2003	80 M	20 ans	9 ans
Sécheresse 2018	55M	10 ans	4 ans
Sécheresse 2011	40M	4 ans	2 ans

TABLE 1 : Comparaison des périodes de retour de plusieurs événements de sécheresse entre le climat de référence et le climat projeté sous scénario 4.5

2050 est l'augmentation des valeurs assurées impliquant une hausse de près de 90% de la charge initiale. L'effet du changement climatique arrive ensuite en deuxième position avec une augmentation de 61% de la charge initiale, comme vu précédemment. En revanche, l'évolution démographique exerce un effet relativement modéré entraînant une hausse de seulement 10% de la charge moyenne de référence.

Etude de soutenabilité à horizon 2050

L'augmentation de la sinistralité détaillée précédemment pourrait amener les assureurs à ajuster leur prime d'assurance MRH pour couvrir l'évolution du risque sécheresse. L'objectif de cette partie consiste à étudier cette hypothèse et à questionner la capacité des assurés à supporter l'accroissement des primes MRH selon divers scénarios de tarification.

Pour mesurer la soutenabilité de la prime par les assurés, des données relatives au niveau de richesse ont été récupérées pour définir le ratio de soutenabilité

$$\kappa_{d,n} = \frac{\pi_{d,n}}{R_{d,n}} \times 100,$$

avec $\pi_{d,n}$ la prime MRH moyenne payée par le département d l'année n et $R_{d,n}$ le revenu médian du département associé pour l'année considérée. Ce ratio est utilisé dans la suite de l'étude pour appréhender le risque d'insoutenabilité, en appréciant sa déviation à horizon 2050

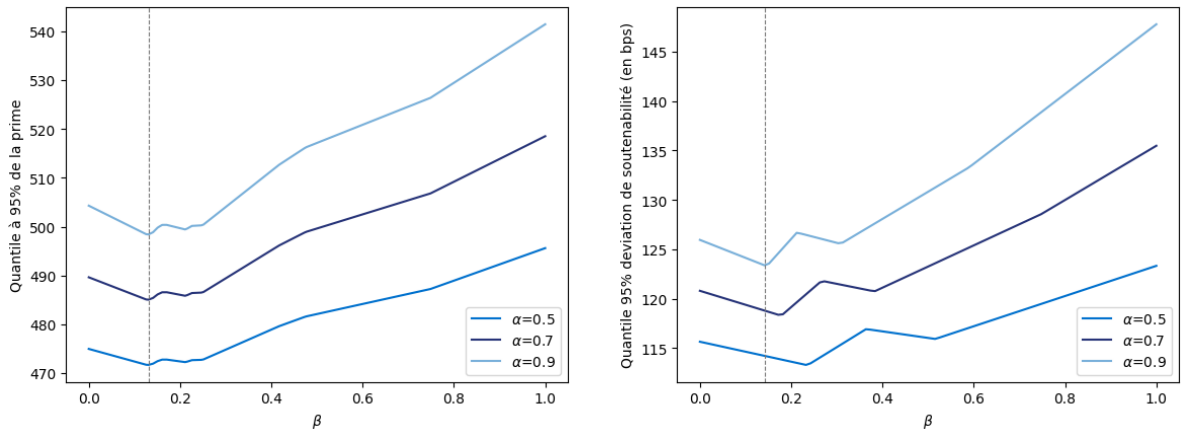
$$\Delta\kappa_d = (\kappa_{d,2050} - \kappa_{d,2020}) \times 100.$$

Afin d'estimer la prime MRH à horizon 2050, divers scénarios de tarification basés sur les méthodes de crédibilité hiérarchique sont explorés. Dans le cadre de l'étude, un modèle hiérarchique à deux niveaux a été considéré avec une première segmentation du portefeuille effectuée par région puis une deuxième par département. Le principe de tarification de ce modèle réside dans le fait que la prime départementale est obtenue à partir d'une pondération de l'expérience du portefeuille, de la région et du département. Cette approche vise à garantir une prime équitable d'un point de vue actuariel. A partir de ce modèle hiérarchique, un premier scénario de tarification "orienté risque" est étudié, dans le sens où les pondérations sont obtenues à partir de la projection des sinistres futurs. Dans ce contexte, la prime payée par un département correspond alors au vrai prix du risque sécheresse auquel il est exposé. L'application de cette méthode de tarification conduit à une augmentation globale de 82% de la prime moyenne MRH du portefeuille avec une croissance dépassant même 130% dans certains départements. Cette augmentation des primes s'accompagne également d'un accroissement du ratio de soutenabilité moyen du portefeuille qui double presque à horizon 2050, synonyme que les primes progressent plus vite que les revenus au cours de la période considérée.

Un deuxième scénario de tarification est implémenté, également basé sur le modèle hiérarchique et inspiré des travaux de Charpentier et al. (2022a). La prime est cette fois-ci paramétrée selon différents niveaux de solidarité à la fois nationale et régionale représentés par les coefficients $(\alpha, \beta) \in [0,1]^2$

$$\pi^{H}_{d,2050}(\alpha,\beta) = (1-\alpha)\pi_{N,2050} + \alpha[(1-\beta)\pi_{R,2050} + \beta\pi_{D,2050}]$$

avec $\pi_{N,2050}$ la prime collective (nationale) du portefeuille, $\pi_{R,2050}$ la prime régionale et $\pi_{D,2050}$ la prime départementale en 2050. Dans cette situation, les facteurs de pondérations ne sont plus liés au risque mais à un niveau de solidarité, introduisant ainsi une tarification basée sur la solidarité. Dans une logique de soutenabilité, l'étude s'est penchée sur les variations des quantiles extrêmes de la distribution des primes, considérant qu'une prime excessivement élevée dans un département peut engendrer un risque d'insoutenabilité.



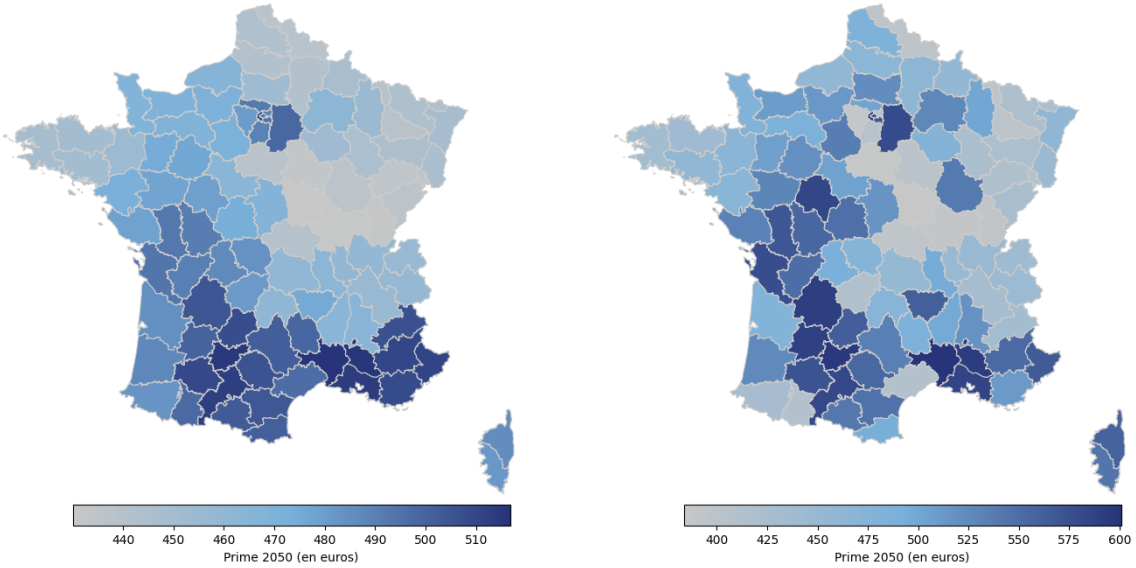
(a) Quantile à 95% des primes MRH (b) Quantile à 95% des déviations de soutenabilité

Figure 5 : Evolution du quantile à 95% de la prime MRH moyenne départementale π^{H}_{2050} et de la déviation de soutenabilité en fonction de β et de plusieurs α.

L'analyse révèle que le quantile n'est pas monotone en β. En effet, en présence d'hétérogénéité dans le portefeuille, il existe un niveau de partage optimal β qui réduit les primes injustement élevées tout en maintenant la stabilité des primes des départements fortement exposés.

De plus, l'absence totale de solidarité, caractérisée par le cas pathologique où $\beta = 1$, se traduit

par une forte augmentation des primes dans de nombreux départements, entraînant ainsi un risque d'insoutenabilité. Cette observation souligne la fragilité de certains départements, dont la soutenabilité repose sur une prise en charge solidaire du risque.



(a) Primes paramétriques estimées à horizon 2050 pour $\beta = 0.14$

(b) Primes paramétriques estimées à horizon 2050 pour $\beta = 1$

FIGURE 6 : Comparaison des primes paramétriques estimées à horizon 2050 pour différents β .

Conclusion

L'étude menée a permis de quantifier l'évolution de la charge sinistre sécheresse à horizon 2050, laquelle connaîtrait une augmentation de 61% sous le scénario RCP 4.5, corroborant ainsi la tendance haussière annoncée par d'autres études. Les projections du modèle construit prévoient également une multiplication des événements extrêmes, avec une fréquence d'apparition deux fois plus élevée par rapport à la période de référence.

A l'issue de l'étude, pour conserver une rentabilité constante et faire payer le coût de cette sinistralité croissante, la prime MRH moyenne du portefeuille augmenterait de 82%, avec des évolutions de primes dépassant 130% dans les zones à risques. De plus, le ratio de soutenabilité a permis d'identifier les départements où la prime progresse plus vite que les revenus à horizon 2050. Le mémoire a également montré que face à l'hétérogénéité du risque sécheresse, la soutenabilité des territoires les plus exposés repose sur une prise en charge solidaire du risque et qu'elle profite d'avantage aux départements caractérisés par des niveaux de revenu plus faibles.

Enfin, il est important de souligner que le mémoire présente un certain nombre de limites. Tout d'abord, l'utilisation d'une résolution spatiale plus fine, à l'échelle communale, aurait pu améliorer la précision des résultats et s'alignerait avec le système d'indemnisation du régime Cat Nat. De plus, les projections de sinistralité excluent tout plan de prévention et modification réglementaire. La dimension politique entourant le régime Cat Nat complique son analyse à moyen terme et la formulation d'hypothèses concernant son évolution. Enfin, l'étude de la soutenabilité a été réalisée uniquement via le risque sécheresse, mais une analyse multi-périls avec une prise en compte de l'ensemble des risques physiques pourrait enrichir les conclusions de l'étude.

Synthesis note

Introduction

In France, the risk of subsidence (or more specifically clay-shrinkage-induced subsidence), which is one of the consequences of drought, is the climate risk that has shown the greatest growth since 2016, both in terms of frequency and severity. Claims due to subsidence have accelerated in recent years, with the average annual cost exceeding one billion euros over the period 2016-2020, and events becoming increasingly intense, as demonstrated by 2022, a record year in terms of costs incurred (CCR (2023b)). This rise in claims is largely due to worsening weather conditions in mainland France. Numerous climate studies carried out in France (Soubeyroux et al. (2011), Gourdier and Plat (2018)) and in Europe (Spinoni et al. (2015)) confirm the influence of climate change on the occurrence of drought events, and predict a worsening of drought episodes throughout the 20th century.

Given this predicted increase in drought-related events, the insurance industry has sought to quantify this growing loss experience by 2050. Depending on the assumptions made, the amount of drought-related damage is expected to rise between 44% and 162%. In its latest study, CCR (2023a) even describes subsidence as "the most worrying peril on the future horizon". The increase in climate risks predicted for the coming years, particularly drought, naturally raises questions about the sustainability of the current insurance industry. In particular, insurers have to adjust insurance premiums to guarantee a certain level of profitability.. During its first climate pilot exercise in 2021, ACPR (2021) was already pointing to a problem with the viability of premiums in the medium term, which would rise between 130% and 200% to cover the increase in climate-related claims. In response to this initial observation, ACPR (2023) included the notion of premium sustainability on the policyholder side in its second pilot exercise. TTo continue in this vein and explore this issue further, the master thesis proposes to study the question of premium sustainability to 2050 through the peril of subsidence.

To address the problem, a first stage of subsidence risk modeling is carried out using machine learning techniques. Next, subsidence risk data is projected to produce an estimate of the claims cost by 2050. Finally, the last part of the study consists of calculating the future home insurance premium and analyzing policyholders' ability to bear its increase according to several pricing scenarios.

Subsidence model

The aim of this first part of the study was to develop a model capable of estimating the annual subsidence claims cost for a department, denoted Y, thus providing a coherent basis for the rest of the study. To achieve this, the model was trained by cross-referencing claims data, exposure data and climate data.

The claims database considered for the study contains a history of drought-related claims occurring between 2000 and 2020 for a property and casualty portfolio covering the whole of mainland France.

In order to guarantee a consistent comparison of the cost over the period in question, the data was processed using the "as-if" method. This treatment is twofold, as it aims to correct both the amount of claims and the portfolio's exposure.

The occurrence of a subsidence disaster is mainly the result of a combination of predisposing and triggering factors. The main predisposing factor is the nature of the soil. Clay soils, by virtue of their malleable structure, will shrink during very dry periods and swell during wet ones. So, to input the model, the first step is to integrate exposure data, such as the geological nature of the soil, the proportion of properties in clay zones and also the type of property insured.

Subsequently, the succession of dry and wet periods is the triggering factor, leading to deformation of the clay soils. As a result, the second stage aims to construct drought indexes capable of linking variations in climatic conditions to the manifestation of the subsidence phenomenon. The first index is the SPEI (*Standardized Precipitation Evapotranspiration Index*) meteorological index (Vicente-Serrano et al. (2010)), which depends solely on temperature, sunshine duration and precipitation data. The second index is the SWI magnitude index (*Soil Wetness Index*). This hydrological index, developed specifically for this study, is based on the Météo France soil moisture index. The latter shows a sinusoidal annual curve, with lower values during the summer period, synonymous with drought. The SWI magnitude index is then obtained by fitting a curve to the monthly SWI data and calculating the integral of the fitted curve under a threshold value γ. The index thus constructed corresponds to a measure of drought intensity, with a high positive value characterizing an intense drought episode.

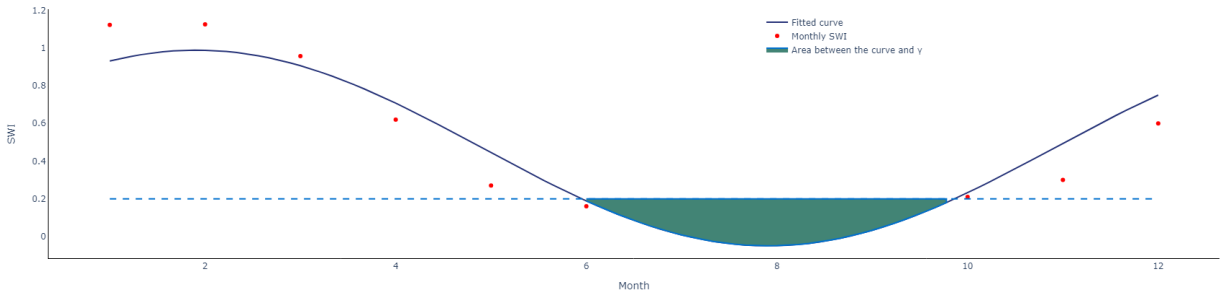


Figure 7: Calculation of the SWI magnitude index SWI(γ).

Then, by consolidating drought indexes, exposure data and claims data, a dataset $(x_i, y_i)_{\in[\![1,n]\!]}$ was created, with x_i the vector of explanatory variables and y_i the realization of the target variable Y. Based on these observations, a supervised learning method was employed to develop a predictive model to meet the study objectives. The fundamental principle of these methods lies in the construction of a prediction function $\hat{f}$ that minimizes the empirical risk

$$\hat{\mathcal{R}}_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \hat{f}(x_i)),$$

using a loss function l.

To approximate the prediction function that minimizes the empirical risk, the Catboost algorithm (Prokhorenkova et al. (2018)) was used. In order to assess the effectiveness of the model developed, a validation approach incorporating standard metrics such as R^2 and MAE (*Mean Absolute Error*) was adopted. In addition, to assess the model's ability to detect claims peaks, the metric MAPEX(q)

(*Mean Absolute Percentage Error for eXtreme values*) was introduced specifically for the master thesis. The latter measures only the model error committed on extreme drought events. These metrics were then calculated using two cross-validation approaches to take into account the spatial and temporal features of the data used. These validation procedures aim to limit the occurrence of bias between training and test data, and thus ensure that the model is able to generalize from one year to the next and from one department to the next.

It was then necessary to optimize the model. Indeed, one of the main difficulties encountered during the modeling process was the unbalanced nature of the target variable, specific to weather-related claims, where 68% of the observations in the dataset had no claims load at all. This imbalance had a direct impact on the learning process, leading the model to underestimate the claims load and misidentify extreme drought events. To overcome this problem, a weight w_i was added to the observations to further penalize errors made on major claims, and thus force the model to correct itself on these data. This penalization of the loss function improved the model's performance compared with a first reference model.

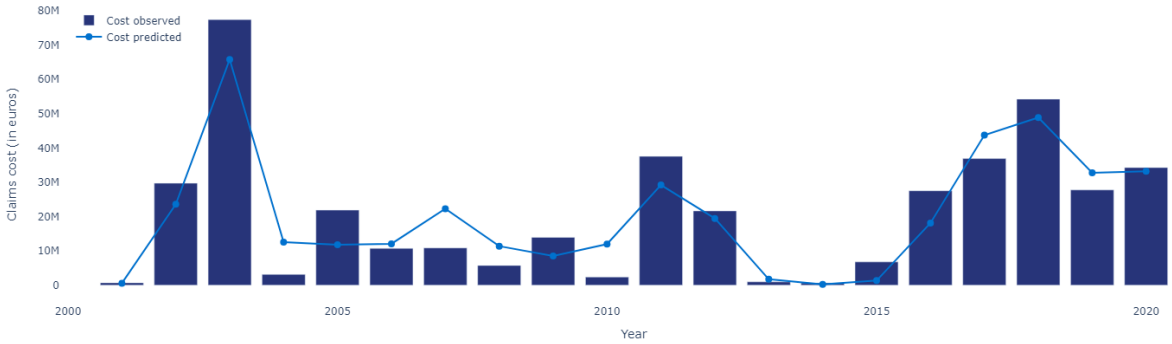


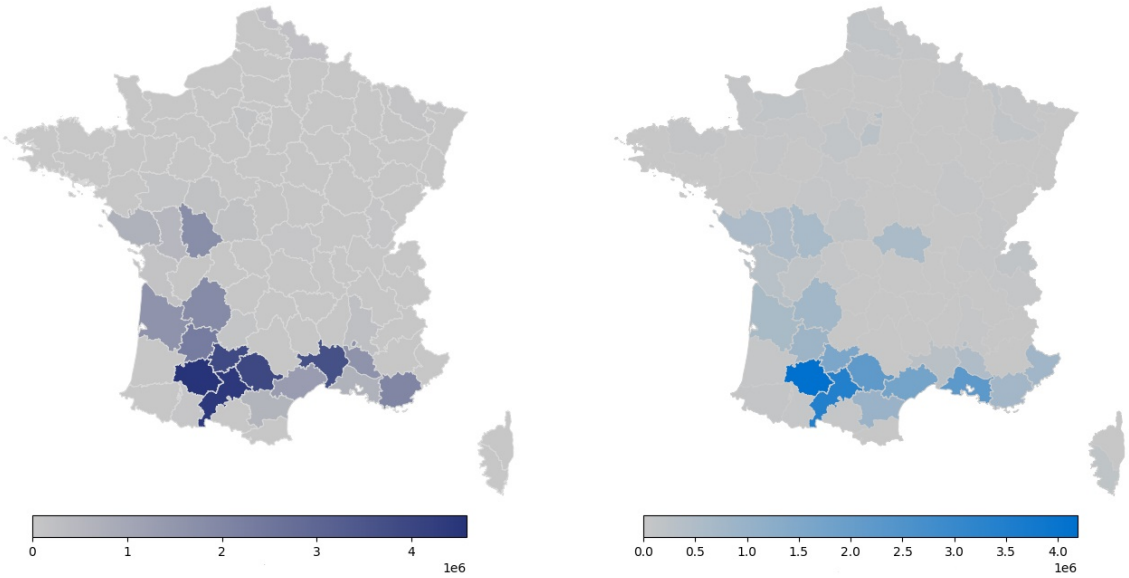
Figure 8: Temporal comparison of observed and predicted annual cumulative claims cost (in euros) for the optimized model

Spatiotemporal analysis reveals that the optimized model more accurately reproduces claims trends, standing out for its ability to capture claims cost peaks, notably those of 2003, 2017 and 2018. Although optimization has reduced the error in areas with high claims experience, it has also led to an overestimation of the amount of claims in certain departments. However, from a prudential point of view, the results of the optimized model remain preferable, justifying its use for the rest of the study.

Risk projection

To estimate the claims cost for subsidence in 2050, the rest of the study consists in projecting the elements involved in risk prediction. This projection concerns both climatic and exposure data.

In order to confront subsidence risk with future climate change, one relevant approach is to use scenarios of future greenhouse gas emissions. The IPCC RCP scenarios are developed with this in mind, offering different radiative forcing trajectories ranging from the most optimistic for RCP 2.6 to the most pessimistic for RCP 8.5. For the purposes of this study, climate variables and hence drought indexes have been projected through the RCP 4.5 trajectory. This choice enables us to be consistent



(a) Observed claims cost (in euros)- Year 2017 (b) Predicted claims cost (in euros) - Year 2017

Figure 9: Spatial comparison of observed and predicted claims cost for the optimized model - Year 2017

with the assumptions of the ACPR's new pilot exercise. In parallel with the climate projections, the portfolio's insurance stakes are modeled for 2050, taking into account both demographic trends predicted by INSEE (2022) over the same period and the increase in insured values.

The projected drought indexes and insured stakes are then used as inputs to the subsidence model built earlier to obtain an estimate of the claims experience by 2050.

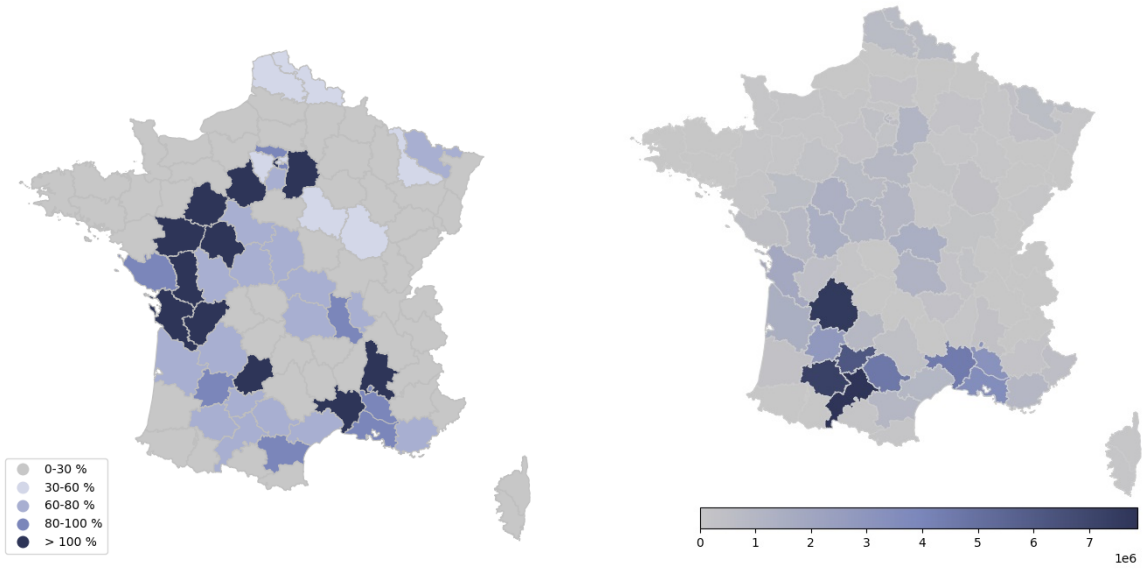
An initial analysis, assuming constant exposure, shows a 61% increase in average annual loss due to climate change by 2050, according to the RCP 4.5 scenario. The most significant changes are to be found along the clay crescent, in the Paris Basin, northern Hauts-de-France and the Centre region. However, future claims costs are largely concentrated in Occitanie, Nouvelle-Aquitaine and Provence-Alpes-Côte d'Azur, where just sixt departments account for 60% of the claims cost.

Furthermore, analysis of return periods shows that extreme drought events of the 2003, 2011 or 2018 type will occur twice as often by 2050 under the RCP 4.5 scenario.

	Claims cost (in euros)	Return period - Reference	Return period - RCP 4.5
Drought of 2003	80 M	20 years	9 years
Drought of 2018	55M	10 years	4 years
Drought of 2011	40M	4 years	2 years

Table 2: Return periods comparison of several drought events between the reference climate and the climate projected under scenario RCP 4.5

The next step in the analysis is to apply the INSEE demographic assumptions and those concerning the increase in insured values. The projected final cost then corresponds to a 160% increase in the initial cost. The main inflationary factor up to 2050 is the increase in insured values, implying an increase of almost 90% in the initial load. The effect of climate change comes in second place, with an increase of 61% in the initial load, as seen above. Demographic change, on the other hand, has a



(a) Percentage change in claims costs to 2050 (b) Predicted claims cost (in euros) in 2050

Figure 10: Evolution and amount of claims costs to 2050 - RCP 4.5

relatively moderate effect, increasing the reference load by only 10%.

Sustainability analysis to 2050

The increase in claims experience described above could lead insurers to adjust their home insurance premiums to cover the increase in subsidence risk. The aim of this section is to study this hypothesis and to question the ability of policyholders to bear the rise in home insurance premiums under various pricing scenarios.

To measure the sustainability of policyholders, data on the level of wealth have been collected to define the sustainability ratio

$$\kappa_{d,n} = \frac{\pi_{d,n}}{R_{d,n}} \times 100,$$

with $\pi_{d,n}$ the average home insurance premium paid by department d in year n and $R_{d,n}$ the median income of the associated department for the year in question. This ratio is used in the rest of the study to assess the risk of unsustainability, by evaluating its deviation by 2050.

$$\Delta\kappa_d = (\kappa_{d,2050} - \kappa_{d,2020}) \times 100.$$

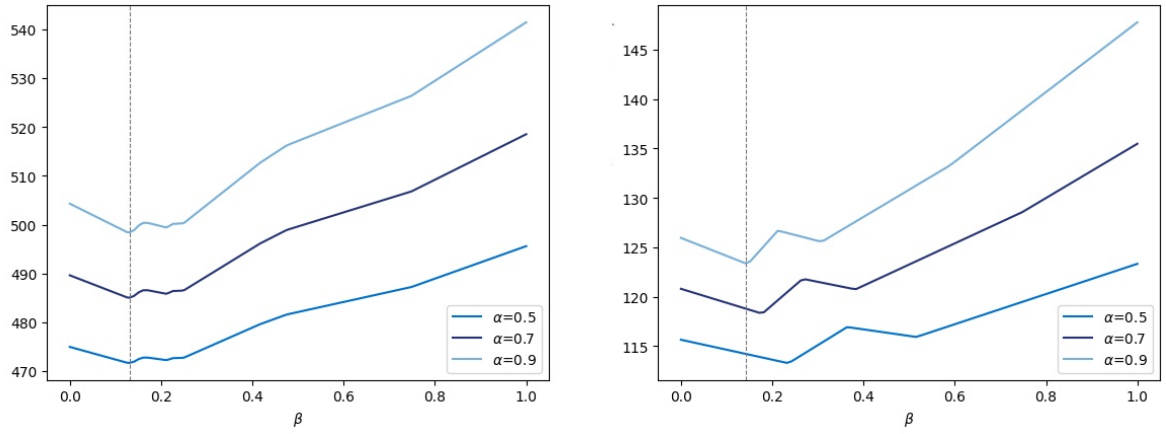
In order to estimate the home insurance premium by 2050, various pricing scenarios based on hierarchical credibility methods are explored. For the purposes of this study, a two-level hierarchical model was considered, with the portfolio segmented first by region and then by department. The pricing principle of this model lies in the fact that the departmental premium is obtained from a weighting of portfolio, region and department experience. This approach is designed to ensure an actuarial fair premium. Based on this hierarchical model, a first "risk-oriented" pricing scenario is studied, in the sense that weights are obtained from the projection of future claims. In this context, the premium paid by a department corresponds to the true price of the subsidence risk to which it is exposed. The application of this pricing method leads to an overall increase of 82% in the portfolio's average home insurance premium, with growth exceeding 130% in some departments. This increase

in premiums is also accompanied by a rise in the portfolio's average sustainability ratio, which almost doubles by 2050, This rise in premiums is also accompanied by an increase in the portfolio's average sustainability ratio, which almost doubles by 2050, meaning that premiums will grow faster than revenues over the period.

A second pricing scenario is implemented, also based on the hierarchical model and inspired by the work of Charpentier et al. (2022a). This time, the premium is adjusted according to different levels of solidarity, both national and regional, represented by the coefficients $(\alpha, \beta) \in [0,1]^2$.

$$\pi^{H}_{d,2050}(\alpha,\beta) = (1-\alpha)\pi_{N,2050} + \alpha[(1-\beta)\pi_{R,2050} + \beta\pi_{D,2050}]$$

with $\pi_{N,2050}$ the collective (national) portfolio premium, $\pi_{R,2050}$ the regional premium and $\pi_{D,2050}$ the department premium in 2050. In this situation, the weighting factors are no longer linked to risk but to a level of solidarity, thus introducing solidarity-based pricing. With a view to sustainability, the study looked at variations in the extreme quantiles of the premium distribution, considering that an excessively high premium in one department could give rise to a risk of unsustainability.



(a) 95% quantile of home insurance premium (in euros)

(b) 95% quantile of sustainability deviations (in basis points)

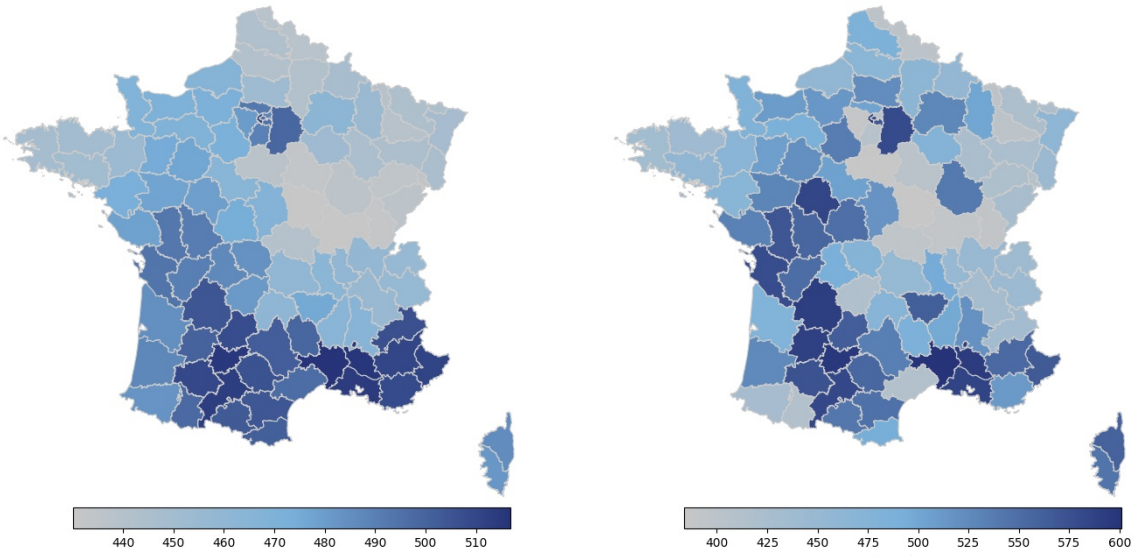
Figure 11: Evolution of the 95% quantile of the departmental average home insurance premium π^{H}_{2050} and of the sustainability deviation as a function of β and several α.

The analysis reveals that the quantile is not monotonic in β. Indeed, in the presence of heterogeneity in the portfolio, there is an optimal sharing level β that reduces unfairly high premiums while maintaining premium stability for highly exposed departments.

Moreover, the total absence of solidarity, characterized by the pathological case where $\beta = 1$, results in a sharp increase in premiums in many departments, leading to a risk of unsustainability. This observation underlines the fragility of certain departments, whose sustainability depends on solidarity in the assumption of risk.

Conclusion

The study enabled us to quantify the evolution of subsidence claims costs up to 2050, which would increase by 61% under the RCP 4.5 scenario, thus corroborating the upward trend predicted by other studies. The model's projections also predict an increase in the number of extreme events, with a frequency of occurrence twice as high as in the reference period.



(a) Estimated parametric premiums (in euros) in 2050 for $\beta = 0.14$.

(b) Estimated parametric premiums (in euros) in 2050 for $\beta = 1$.

Figure 12: Comparison of estimated parametric premiums in 2050 for different β .

At the end of the study, in order to maintain constant profitability and pay the cost of this growing claims experience, the average home insurance premium for the portfolio would increase by 82%, with premium growth exceeding 130% in high-risk areas. In addition, the sustainability ratio has enabled us to identify the departments where premiums will grow faster than revenues by 2050. By opting for a solidarity-based pricing system, the report showed that, given the heterogeneous nature of subsidence risk, the sustainability of the most exposed territories depends on the assumption of risk on the basis of solidarity, and that this benefits departments with lower income levels.

Finally, it is important to point out that the master thesis has a number of limitations. First of all, the use of a finer spatial resolution, on a communal scale, could have improved the accuracy of the results and would have aligned them with the Cat Nat compensation system.In addition, the loss experience projections exclude any prevention plans or regulatory changes. The political dimension surrounding the Cat Nat scheme complicates its medium-term analysis and the formulation of hypotheses concerning its evolution. Lastly, the sustainability study was carried out solely on the basis of drought risk, but a multi-risk analysis taking into account all physical risks could enrich the study's conclusions.

Remerciements

Je tenais à remercier avant tout mes tuteurs de stage Pauline BERGER et Alex-Michel NGNINGHA pour leur soutien, leur disponibilité et leurs conseils réguliers qui m'ont permis de comprendre les enjeux du sujet et ainsi me guider dans la rédaction de ce mémoire.

J'adresse notamment ma reconnaissance aux associés de Mazars Actuariat Grégory BOUTIER, Alice THOU et Alexandre GUCHET pour m'avoir fait confiance et permis de réaliser mon mémoire au sein du cabinet.

J'aimerais également remercier les équipes de Mazars Actuariat pour leur accueil chaleureux et leur aide pendant le stage. Plus particulièrement, Ismaël TAHRI HASSANI, Fabrice TANO, Sixte JONGLEZ de LIGNE et Auguste DERREAL pour leur bonne humeur et le temps qu'ils m'ont accordé pour répondre à mes interrogations techniques.

Je souhaite, de plus, exprimer mes meilleurs sentiments aux autres stagiaires de Mazars Actuariat qui ont contribué au bon déroulement de mon stage en développant un cadre de travail enrichissant sous le signe de l'entraide.

Mes remerciements vont enfin à l'ensemble de l'équipe pédagogique du M2 Actuariat et plus particulièrement à M. Quentin GUIBERT pour la qualité de ses enseignements, son écoute et sa bienveillance tout au long de l'année ainsi qu'à M. Christophe DUTANG pour son suivi attentif et la pertinence de ses remarques tout au long de mon stage .

Sans oublier, Elisabeth de VANDIERE et ma famille pour leurs précieux conseils, leur soutien et leurs encouragements à chaque étape de mon parcours académique.

Table des matières

Liste des abréviations et sigles utilisés

Dans le présent mémoire, l'utilisation récurrente de certains sigles et abréviations vise à faciliter la compréhension et la lecture du contenu. Leurs définitions sont exposées de manière exhaustive ci-dessous. Par ailleurs, les sigles tels que PDSI, SPEI, SPI, SSWI, et SWI, tous représentant des indices de sécheresse, sont explicitement détaillés dans la section 1.1.3. De même, les sigles MAE, MAPEX et MSE font référence à des métriques décrites en section 2.3.2.

ACPR : Autorité de Contrôle Prudentiel et de Résolution

BRGM : Bureau de Recherche Géologiques et Minières

Cat Nat : Catastrophes Naturelles

CCR : Caisse Centrale de Réassurance

CNRM : Centre National de Recherches Météorologiques

DRIAS : Donner accès aux scénarios climatiques Régionalisés français pour l'Impact et l'Adaptation de nos Sociétés et environnement

FFA : Fédération Française des Assureurs

GIEC : Groupe d'experts Intergouvernemental sur l'Évolution du Climat

INSEE : Institut National de la Statistique et des Études Économiques

MRH : Multi-Risques Habitation

PACA : Provence-Alpes-Côte d'Azur

RCP : *Representative Concentration Pathways*. Trajectoires Représentatives de Concentration

RGA : Retrait-Gonflement des sols Argileux

SIM : Safran-Isba-Moscou

Introduction

Dans un contexte de dérèglement climatique, la France subit ces dernières années l'expansion des catastrophes naturelles de grande ampleur. Concernant la sécheresse géotechnique ou phénomène de retrait-gonflement des sols argileux (RGA), qui provoque des dégâts sur le bâti, les successions de sécheresses exceptionnelles entre 2016 et 2022 témoignent de sa progression alarmante sur le territoire métropolitain. En effet, la charge sinistre associée à ce phénomène est passée de 400 millions d'euros par an en moyenne sur la période 1989-2015 à 1 milliard d'euros par an en moyenne sur la période 2016-2020.

L'année 2022 fait notamment figure d'ovni climatique au regard des épisodes de sécheresse géotechnique, avec un coût des sinistres estimé à plus de 3 milliards selon une évaluation récente de la CCR en marge de l'annonce du projet "Initiative Sécheresse"(CCR (2023b)) en collaboration avec France Assureurs et la Mission Risques Naturels. Cette somme représente une augmentation d'environ 1 milliard d'euros par rapport à 2003, qui détenait le précédent record en termes de coûts pour ce type de sinistre.

En France, le projet CLIMSEC (Soubeyroux et al. (2011)) et en Europe les travaux de Spinoni et al. (2015), ont apporté des preuves scientifiques de l'influence du changement climatique sur la manifestation des épisodes de sécheresse. Les résultats obtenus au travers de ces études prévoient également l'augmentation des événements de sécheresses tout au long du XXIè siècle, en particulier lors de sa deuxième moitié.

La sinistralité observée durant la dernière décennie ainsi que l'accroissement annoncée des épisodes de sécheresses font du phénomène RGA un enjeu majeur pour le marché de l'assurance. Forts de ce constat, les acteurs de l'industrie de l'assurance ont réalisé de nombreuses études visant à quantifier cette sinistralité croissante à horizon 2050. L'ACPR (2021), la CCR (2018) mais également France Assureurs (FFA (2021)) et Covéa (2022) se sont penchés sur les conséquences du changement climatiques sur le coût des catastrophes naturelles, notamment celui du risque RGA. Selon l'auteur des travaux et les hypothèses retenues, le montant des dommages en lien avec la sécheresse géotechnique augmenterait entre 60 et 70%. Dans sa dernière étude parue en septembre, la CCR (2023a) qualifie le péril sécheresse de "*le péril le plus préoccupant à horizon futur*" en matière de charge sinistre.

Cette progression des risques climatiques, notamment celle du phénomène RGA, prévue dans les années à venir, interroge naturellement sur la pérennité du système assurantiel actuel. En France, le régime Cat Nat, qui est chargé d'indemniser les sinistres rattachés à des arrêtés de catastrophes naturelles, est déficitaire depuis 5 ans et est au centre de nombreux débats concernant sa sauvegarde. Du côté des assureurs, l'augmentation de la charge sinistre des événements climatiques, amène les compagnies d'assurance à réajuster les primes d'assurance afin de maintenir leurs résultats. Ce réajustement des primes d'assurance, qui est voué à croître dans le futur, a une influence directe sur les assurés qui subissent ces augmentations. Lors de son premier exercice pilote, l'ACPR (2021) pointait déjà un problème de viabilité des primes d'assurance, qui augmenteraient entre 130% et 200% du fait de

l'accroissement significatif des sinistres climatiques à horizon 2050 si les assureurs souhaitent maintenir le même niveau de rentabilité. Dans son prochain exercice l'ACPR (2023) intègre cette fois-ci dans ses hypothèses, cette notion de soutenabilité de la prime du côté des assurés.

Pour continuer dans ce sens et approfondir la problématique d'accessibilité de la prime, le mémoire propose d'étudier la question de sa soutenabilité à horizon 2050 à travers le prisme du phénomène RGA.

Pour répondre à cette problématique, une présentation du péril sécheresse et du contexte assurantiel qui l'entoure est effectuée dans un premier temps au sein du chapitre 1. Dans un deuxième temps, la modélisation du risque sécheresse est effectuée dans le chapitre 2 à l'aide des techniques de *machine learning*. Enfin, dans un dernier temps, la charge sinistre sécheresse est projetée à horizon 2050 pour aboutir au calcul de la prime associée et étudier sa soutenabilité dans le cadre du chapitre 3. Pour analyser l'exposition des territoires métropolitains au risque d'insoutenabilité de la prime, plusieurs scénarios de tarification sont explorés, basés sur différents de niveau de solidarité.

Chapitre 1

Le risque sécheresse dans un environnement incertain

Ce premier chapitre vise à présenter les caractéristiques du péril sécheresse, le contexte qui l'entoure ainsi que les enjeux assurantiels nécessaires à la pleine compréhension du mémoire.

Dans **une première section**, les origines et les conséquences du phénomène de sécheresse, plus particulièrement la subsidence induite par le retrait-gonflement des argiles (RGA), sont décrites.
Dans **une deuxième section**, la prise en charge au sein du système assurantiel français des dommages causés par cet aléa climatique ainsi que le cadre réglementaire qui le régit sont détaillés.
Enfin, dans **une dernière section**, les défis futurs auxquels sont confrontés les acteurs de l'assurance face à la progression de ce phénomène climatique sont exposés.

1.1 Présentation du risque sécheresse

La présente section s'attache tout d'abord à explorer le phénomène de sécheresse, en mettant en exergue les diverses modalités de son expression à l'échelle mondiale et nationale. Par la suite, une analyse détaillée est consacrée au phénomène de subsidence, dont l'incidence perturbe le secteur assurantiel français. Enfin, une introduction des divers indices de sécheresse, utilisés dans la littérature scientifique est proposée. Ces derniers joueront un rôle primordial dans la gestion et la quantification du risque en question.

1.1.1 La sécheresse : un risque global aux multiples conséquences

La sécheresse est un phénomène climatique complexe qui fait référence de manière générale à un déficit en eau dans une zone particulière. Au sein du cycle hydrologique terrestre, diverses formes de sécheresse se manifestent, et leurs conséquences varient en fonction de leur ampleur et des régions géographiques touchées. Dans la littérature scientifique quatre types de sécheresse existent :

- La **sécheresse météorologique** qui se définit par un déficit pluviométrique prolongé. Usuellement, ce déficit est calculé par comparaison aux moyennes de précipitations historiques sur une profondeur de 30 ans dans la zone en question.

- La **sécheresse édaphique** également appelée sécheresse agricole, qui se produit lorsque le manque de précipitations, combiné à d'autres facteurs, affecte progressivement les sols en surface et compromet le développement de la végétation.

- La **sécheresse hydrologique** qui désigne quant à elle le déficit hydrique des milieux aquatiques (cours d'eau, lacs et nappes phréatiques). Au cours de l'été, le niveau des nappes souterraines et le débit des cours d'eau diminuent de manière saisonnière. Cette diminution peut être amplifiée en cas de sécheresse météorologique associée.
- La **sécheresse géotechnique** qui se réfère à une période de déficit de précipitation induisant une altération de la teneur en eau du sous-sol. Selon les caractéristiques du sol, cette variation de l'état hydrique du sol peut provoquer des mouvements de terrain. Une description plus fine de la sécheresse géotechnique est effectuée dans la section 1.1.2.

Ces différentes formes de sécheresses engendrent des conséquences multiples dans le monde. Chaque année, environ 55 millions de personnes subissent les effets de la sécheresse à l'échelle mondiale. L'Afrique est particulièrement touchée avec plus de 300 événements enregistrés au cours des 100 dernières années, représentant ainsi 44% du total mondial. L'Europe est également affectée, avec pas moins de 45 épisodes de sécheresse comptabilisés au siècle dernier (Guha-Sapir et al. (2021)).

Un déficit hydrique dans les circuits naturels provoque l'assèchement voire la désertification de certaines zones, pénalisant l'équilibre des écosystèmes. Dans les circuits artificiels, le faible volume d'eau impose aux agriculteurs des limitations quant à l'irrigation de leurs cultures ainsi que des problèmes d'approvisionnement en eau potable. Selon l'Organisation Mondiale de la Santé (OMS), la sécheresse constitue le plus grand danger pour le bétail et les cultures dans pratiquement toutes les régions du globe (WHO (2021)). Son impact sur la santé est également majeur. D'après l'Organisation Météorologique Mondiale, la sécheresse a engendré plus de 650 000 décès sur la période 1970-2019 (WMO (2021)), principalement dans des pays en développement.

Outre ses effets néfastes sur la santé, la biodiversité et sa mise en danger des cultures et du bétail, la sécheresse influe également sur la taille et l'intensité des feux de forêts. En effet, la propagation des incendies dépend de l'état de sécheresse de la végétation, de l'humidité de l'air, de la température et du vent (ONF (2022)). Chaque année, entre 3000 et 4000 feux de forêts sont recensés.

En France, la sécheresse se manifeste par un ensemble de caractéristiques observables. Elle favorise notamment le phénomène de subsidence (ou retrait-gonflement des argiles) qui affecte les habitations, entraînant ainsi de lourdes pertes économiques. Le système assurantiel français, qui couvre les dégâts causés par les événements climatiques sur le bâti, est particulièrement touché par la survenance de ce type de sécheresse géotechnique. La présentation du mécanisme de ce phénomène physique, central dans ce mémoire, est effectuée dans la section ci-après.

1.1.2 Le phénomène de subsidence

La subsidence, également connue sous les termes de sécheresse géotechnique ou encore de retrait-gonflement des argiles, désigne le phénomène physique résultant des variations de teneur en eau des terrains argileux. Ce processus de dessiccation et de réhydradation successives des sols entraîne des mouvements différentiels de terrain significatifs, causant des dommages directs sur la structure des bâtiments si celle-ci n'est pas suffisamment rigide, comme illustré sur la figure 1.1. Les maisons individuelles sont principalement affectées par ce phénomène, étant donné leurs fondations moins profondes et leur structure plus légère par rapport aux immeubles collectifs.

Comme décrit dans Covéa (2022), l'apparition de ce phénomène est liée à une combinaison multifactorielle que l'on peut distinguer en trois catégories : les facteurs de prédisposition qui caractérisent la sensibilité de la zone géographique au risque, les facteurs déclenchant qui permettent de provoquer le phénomène et enfin les facteurs aggravants qui amplifient son impact.

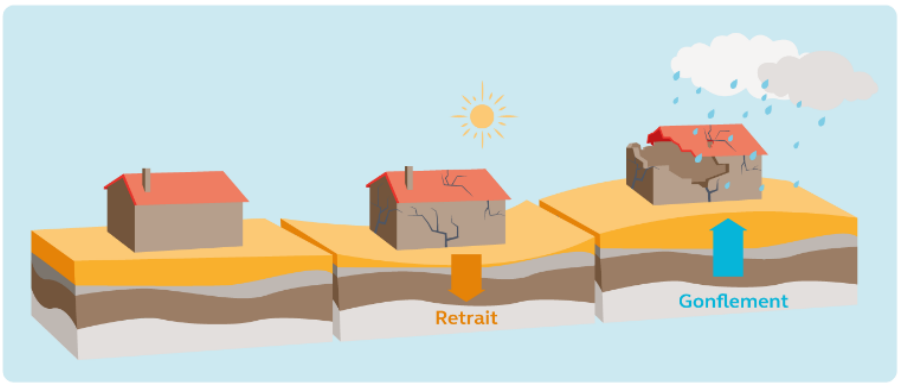


FIGURE 1.1 : Illustration du phénomène retrait-gonflement des argiles - COUR DES COMPTES (2022)

Facteur de prédisposition

Le principal facteur de prédisposition est la nature du sol. Les sols argileux, de par leur structure minéralogique particulière en feuillet, contribuent à l'occurrence du phénomène. En effet, les espaces entre les différentes couches peuvent contenir de l'eau et des ions, ce qui confère aux argiles leur capacité à se dilater et à se rétracter. Les minéraux qui composent l'argile influent de manière directe sur sa malléabilité. La montmorillonite ainsi que la vermiculite sont qualifiées "d'argiles gonflantes" car elles possèdent des propriétés de déformation potentielle importantes.

Depuis quelques années, le BRGM (Bureau de Recherches Géologiques et Minière) développe une cartographie de l'exposition au RGA en France métropolitaine. Cette dernière est issue d'une analyse des cartes géologiques de la France, qui a permis d'identifier plus de 2000 formations argileuses affleurantes ou sub-affleurantes. Ces formations ont été classées en trois catégories de susceptibilité croissante (faible, moyenne et forte) en se basant sur la combinaison de trois critères géologiques pour caractériser les formations : leur nature lithologique, leur composition minéralogique et leur comportement géotechnique. Par la suite, le niveau de susceptibilité est croisé avec la sinistralité effectivement observée pour obtenir le degré d'exposition (faible, moyen et fort). La carte ainsi conçue, disponible sur le portail GEORISQUE (2023a) , est présentée sur la figure 1.2.

Le BRGM (2023) stipule également que 48% du territoire métropolitain se retrouve en zone d'exposition moyenne ou forte au RGA et 93% des sinistres recensés se concentrent dans des zones d'exposition moyenne (38%) ou forte (93%). Ces chiffres montrent donc que le territoire métropolitain est exposé de manière concrète au risque RGA. Les zones présentant une forte exposition sont regroupés principalement le long d'un arc qualifié de croissant argileux. Ce dernier s'étend du nord-est de la France jusqu'à la région PACA en passant par l'Occitanie et la façade atlantique. Dans des départements comme le Tarn ou le Gers, plus de 80% des communes contiennent des maisons individuelles en zone à risque RGA fort.

Facteurs déclenchants

Les facteurs déclenchants se manifestent dans le cadre de variations climatiques exceptionnelles. Plus précisément, l'humidité des sols affecte directement les caractéristiques physiques des argiles. Les deux éléments primordiaux qui interviennent sur l'état d'hydratation et donc sur la déformation structurelle du sol sont les précipitations et l'évapotranspiration. Cette dernière désigne le mécanisme

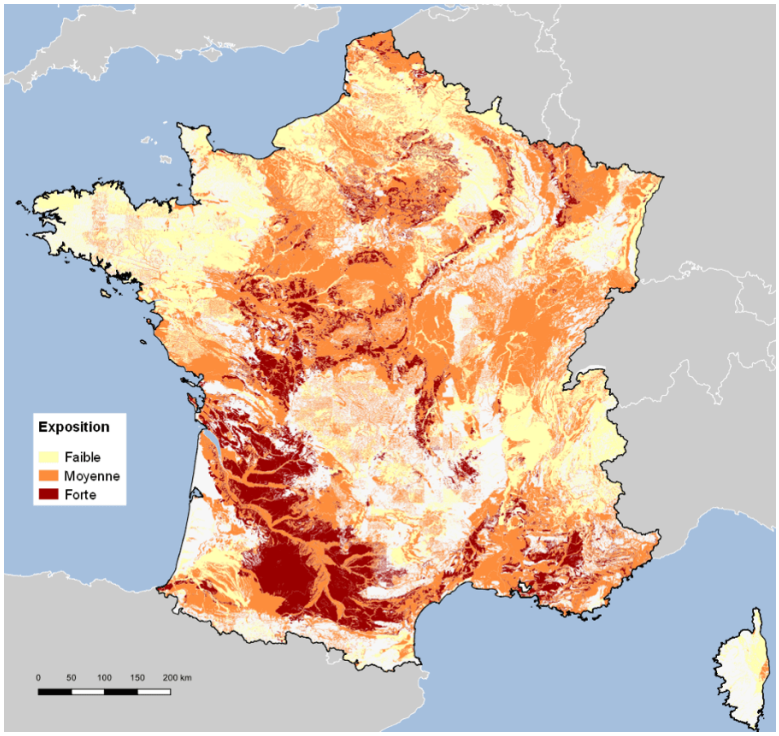


FIGURE 1.2 : Cartographie de l'exposition au retrait-gonflement des argiles actualisée en 2019 - BRGM (2023)

biophysique par lequel une quantité d'eau est transférée de la surface du sol et des plantes vers l'atmosphère, à la fois par évaporation et par transpiration végétale.

Dans un contexte de sécheresse extrême, marquée par une une insuffisance de précipitations et une évaporation anormalement élevée, la couche superficielle du sol subit une déformation en réponse à ces conditions. Les molécules d'eau emprisonnées dans les espaces entre les feuillets sont ainsi relâchées, entraînant le phénomène de rétractation des argiles. A l'inverse, lors de périodes humides, les sols se remplissent d'eau et les terrains argileux connaissent ainsi des processus de gonflement. Ces différents mécanismes sont illustrés sur la figure 1.1.

Facteurs aggravants

Outre les facteurs de prédisposition et déclenchants, des actions d'origine humaine ou environnementale peuvent contribuer à l'aggravation du phénomène de subsidence. Les travaux d'aménagement, de nature anthropique, nécessitant notamment des actions de drainage du sol, de pompage ou encore de plantation, altèrent directement la circulation de l'eau dans le sol. En ce qui concerne les facteurs environnementaux, la simple présence d'une végétation arbustive conséquente autour d'une zone d'exposition sensible peut suffire à aggraver le processus de gonflement des argiles. Les racines des plantes aspirent l'eau du sol par succion, affectant son équilibre hydrique.

1.1.3 Les indices de sécheresse

Consciente des conséquences dévastatrices de la sécheresse sur la population, les écosystèmes, l'agriculture ou encore sur l'approvisionnement en eau, la communauté scientifique s'est efforcée de

développer ces dernières années des indices afin de mieux appréhender et évaluer l'ampleur de ce phénomène. De manière générale, ces indices sont standardisés et présentent donc une indépendance vis-à-vis des particularités climatiques locales. Cette caractéristique permet notamment une évaluation plus objective de l'intensité et de la sévérité des événements, facilitant ainsi la comparaison et l'analyse des phénomène de sécheresse à différentes échelles géographiques. Dans le cadre de ce mémoire, une liste non exhaustive de ces indices est détaillée. L'objectif est de mettre en évidence les principaux outils et approches exploités par les scientifiques pour mesurer la sécheresse.

SPI

Le *Standardized Precipitation Index* (McKee et al. (1993)) est un premier indice standardisé qui se base uniquement sur les données de précipitations et permet donc de mesurer essentiellement la sécheresse météorologique plutôt que la sécheresse touchant les sols. Sa construction s'effectue à partir d'un historique de précipitations relativement profond, idéalement 50-60 ans (Guttman (1999)). Ceci permet notamment d'évaluer les conditions de précipitations par rapport à une période de référence. Pour définir complètement l'indice, il est nécessaire de choisir une échelle de temps sur laquelle les données vont être évaluées. Pour respecter les spécificités du phénomène de sécheresse, des périodes de 3 ou 6 mois sont considérées (les indices sont notés respectivement SPI-3 et SPI-6). Pour calculer le SPI, le volume de précipitation reçu au cours d'une période spécifique (par exemple, trois mois) est comparé à celui des mêmes mois pour les années précédentes. Une estimation de la distribution des valeurs observées est réalisée, ce qui permet de déterminer le quantile auquel se situe la période étudiée par rapport à la distribution choisie. En général, la distribution Gamma est utilisée pour représenter la pluviométrie. Le SPI est ensuite déterminé comme étant la valeur de ce quantile pour la loi normale centrée réduite.

SPEI

Le *Standardized Precipitation-Evapotranspiration Index* (Vicente-Serrano et al. (2010)) est une amélioration de l'indice SPI mentionné précédemment. Alors que le SPI ne prend en compte que les précipitations et est donc principalement lié à la sécheresse météorologique, le SPEI intègre le phénomène de l'évapotranspiration pour une appréciation plus précise de l'état d'hydratation des sols. Le principe de construction est sensiblement similaire à celui du SPI, à la différence que celui ci se distingue par l'utilisation des précipitations nettes, c'est à dire les précipitations auxquelles l'évapotranspiration a été soustraite. Une description plus fine et rigoureuse de l'élaboration de cet indice est effectuée en partie 2.2.

SWI

Le *Soil Wetness Index* (Soubeyroux et al. (2011)) est un autre indice largement utilisé pour mesurer la sécheresse des sols. Il joue un rôle essentiel dans le régime Cat Nat en permettant notamment de définir un critère pour la reconnaissance d'une commune en état de catastrophe naturelle (*cf.* partie 1.2.3). Le SWI fournit une estimation de la disponibilité en eau pour les plantes par rapport à la capacité maximale de rétention du sol. L'indice est défini à partir des concepts de point de flétrissement et de capacité du champ. Le point de flétrissement correspond au niveau d'humidité du sol à partir duquel les plantes ne sont plus capables d'extraire suffisamment d'eau pour leur survie. Au-delà de ce seuil et en l'absence d'arrosage, les plantes se dessèchent et finissent par mourir. La capacité du champ, également appelée point de ressuyage, représente la teneur en eau du sol après un drainage complet par gravité. Au-delà de ce seuil, le sol est saturé en eau. En notant donc la teneur en eau du sol W, la teneur en eau du point de flétrissement W_f et la capacité du champ W_c, l'indice

est obtenu l'aide de la formule

$$\text{SWI} = \frac{W - W_f}{W_c - W_f}.$$

Le SWI est généralement compris entre 0 et 1. D'après Météo-France, un sol est qualifié de sec lorsque son SWI est inférieure à 0.5, tandis qu'il est considéré comme très humide au-delà de 0.8. Par ailleurs, le type de sol influe directement sur l'indice. Les sols de nature argileuse, qui sont associés au phénomène de subsidence décrit dans la partie 1.1.2, ont généralement des seuils de saturation et de flétrissement plus élevés que d'autres types de sols tels que les sols sableux ou limoneux. Dans la section 2.2, un nouvel indice de sécheresse, fondé sur le SWI, sera proposé.

SSWI

Le *Standardized Soil Wetness Index* (Vidal et al. (2010)) est un indice dérivé du SWI présenté ci-dessus qui, à l'instar des précédent indices standardisés, applique une méthode de projection sur une distribution normale centrée réduite, similaire à celle utilisée pour le SPI ou le SPEI. Ses valeurs s'étendent désormais dans une plage allant de 0.5, correspondant à une sécheresse légère, jusqu'à -2, représentant des conditions de sécheresse extrêmes. Contrairement au SWI qui peut être difficilement comparable entre différentes régions, l'indice SSWI permet de comparer l'humidité des sols indépendamment des caractéristiques climatiques de la zone considérée.

PDSI

Le Palmer Drought Severity Index (Palmer (1965)) est un indice sophistiqué qui prend en compte l'humidité des sols, ce qui le rend plus complexe à calculer. Cet indice est déterminé, pour un mois donné, par une formule récursive impliquant un autre indice mensuel qui mesure l'anomalie d'humidité. Son calcul implique des variables telles que les précipitations mensuelles, l'évapotranspiration potentielle et d'autres facteurs spécifiques au mois considéré. Le PDSI est centré autour de 0, où une valeur nulle indique des conditions climatiques normales. Les valeurs négatives du PDSI indiquent la présence de sécheresse, et plus la valeur s'éloigne de 0, plus le phénomène de sécheresse est sévère. Un PDSI inférieur à -3 caractérise un épisode de sécheresse extrême.

Autres indices communs

Il existe d'autres indices usuels de sécheresse au sein de la littérature scientifique, chacun apportant une approche unique pour évaluer et apprécier les conditions de sécheresse. Parmi ces indices, le *Keetch Byram Drought Index* (Keetch et Byram (1968)) tient compte de l'humidité et des températures, offrant ainsi une mesure pour prévenir notamment le risque d'incendie. *L'Effective Drought Index* (Byun et Wilhite (1999)) tout comme le *Rainfall Anomaly Index* [Van Rooy (1965)], sont des indices qui permettent d'identifier les variations anormales de précipitations par rapport aux données historiques. Par ailleurs, certains indices se concentrent exclusivement sur les températures, comme le *Temperature Condition Index* (Kogan (1995a)), quantifiant l'impact de la chaleur sur la survenance des événements de sécheresse. Enfin, une dernière catégorie d'indices est basée sur des mesures de la réflectance du sol et de la végétation, tels que le *Normalized Difference Vegetation Index* ou le *Vegetation Condition Index* (Kogan (1995b)). Ces indices exploitent les variations dans la couverture végétale pour estimer l'état hydrique des écosystèmes et fournir des informations sur la sécheresse potentielle. En combinant différentes mesures et perspectives, ces indices complémentaires contribuent à une meilleure compréhension et à une évaluation plus précise des conditions de sécheresse dans divers contextes environnementaux.

1.2 Prise en charge du risque sécheresse et cadre réglementaire

Cette section se concentre sur la prise en charge du risque sécheresse par le système assurantiel français. Le produit d'assurance Multirisques Habitation est présenté car il joue un rôle central dans la protection des habitations touchées par un événement de sécheresse. Ensuite le fonctionnement du régime d'indemnisation des catastrophes naturelles sera abordé. Enfin, l'application de ce dernier dans le contexte du risque sécheresse sera analysée, en examinant notamment les critères d'éligibilité des sinistres liés à ce phénomène.

1.2.1 Le produit Multirisques Habitation

Afin de se protéger contre les aléas, les risques et les conséquences ne relevant pas de la vie humaine, un individu peut souscrire un contrat d'assurance dommages. Ce type de contrat vise à protéger le patrimoine de l'assuré en indemnisant les pertes financières résultant de la détérioration ou de la destruction de ses biens, ainsi que des dommages causés à des tiers. En France, il existe divers produits d'assurance dommages offrant des garanties spécifiques pour répondre à ces besoins.

Dans le cadre de ce mémoire, le risque sécheresse est étudié à travers le produit Multirisques Habitations (MRH) puisqu'il joue un rôle essentiel dans la prise en charge des dégâts sur le bâti. Ce produit d'assurance, offre une couverture multi-garantie protégeant le patrimoine familial (habitation et mobilier) lorsque l'assuré est responsable ou victime d'un sinistre. Les biens assurables dans le cadre ce produit sont :

- Les **bâtiments** (maison, appartements,etc.) appartenant à l'assuré ainsi que leurs aménagements (garages, abris de jardins, etc.) qui ne peuvent être dissociés sans être affectés ou sans détériorer la construction.
- Le **mobilier personnel** qui correspond aux meubles et objets personnels appartenant à l'assuré.
- Les **biens à usage professionnel** qui sont tous les meubles, équipements, outils et machines utilisées dans le cadre de l'activité professionnel de l'assuré.

Il existe de nombreuses garanties constituant le contrat MRH. Les principales sont :

- La garantie **incendie-explosion** qui couvre les dommages matériels résultant d'un incendie, d'une explosion ou d'une implosion.
- La garantie **dégâts des eaux** qui prend en charge les conséquences d'un dégât des eaux, sans toutefois inclure l'indemnisation des réparations de la partie de la construction ou de l'appareil responsables du sinistre.
- La garantie **vol** qui couvre la disparition et les pertes matérielles causées par les vols ainsi que les tentatives de vol et/ou les actes de vandalisme conformément aux conditions stipulées dans le contrat. Pour jouir de cette garantie, l'assuré est tenu de fournir les preuves nécessaires pour justifier les circonstances des événements.
- La garantie **bris de glace** qui couvre les dommages matériels (bris, fissures, etc.) subis par les vitres, les fenêtres, les baies vitrées, les vélux, les garde-corps, les parois séparatives de balcons, ainsi que les verres et glaces du mobilier.

L'une des caractéristiques importantes du produit MRH est notamment la présence de la garantie **catastrophes naturelles**. Cette dernière est une garantie obligatoire, prévue par la loi, qui permet

d'offrir à l'assuré une couverture contre les dégâts provoqués par des catastrophes naturelles.Toutefois, compte tenu de l'ampleur potentielle des sinistres liés aux événements climatiques, les assureurs ne peuvent assumer seuls la charge financière qui en découle. Ainsi, depuis l'adoption de la loi du 13 juillet 1982, l'État a mis en place le régime d'indemnisation des catastrophes naturelles, plus connu sous le nom de régime Cat Nat.

En instaurant ce régime complémentaire, la France se distingue de nombreux pays en garantissant à ses citoyens une indemnisation des dommages résultant de phénomènes naturels considérés comme non assurables. Parmi les nombreux aléas naturels couverts par ce régime, la sécheresse est prise en compte lorsqu'elle implique des tassements différentiels sur des sols argileux (sécheresse géotechnique) et que celle-ci revêt un caractère exceptionnel. Les parties suivantes visent justement à présenter le fonctionnement du régime Cat Nat ainsi que son application dans le cadre du péril sécheresse.

1.2.2 Le régime Cat Nat

Le régime d'indemnisation Cat Nat est une assurance de biens qui intervient lorsque le sinistre est causé par une catastrophe naturelle, défini selon l'article L125-1 du Code des Assurances. Ce dernier stipule que les effets des catastrophes naturelles sont "*les 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*". Pour entamer la procédure d'indemnisation d'un sinistre, il est nécessaire que l'état de catastrophe naturelle soit constaté, à l'échelle de la commune, au travers d'un arrêté ministériel publié au Journal Officiel. Celui-ci spécifie les communes impliquées, la période, les dangers ainsi que les dommages associés. Il est donc primordial que le maire de la commune sinistrée fasse une demande de reconnaissance de l'état de catastrophe naturelle au préalable.

Les périls habituellement couverts sont les inondations, la sécheresse, les mouvements de terrain, les cyclones, les séismes ou encore les tsunamis. Certains événements naturels ne sont pas pris en charge dans la garantie Cat Nat comme les tempêtes, les chutes de grêle ou de neige. Les garanties relatives à ces trois types d'aléas peuvent être incluses de manière obligatoire dans les contrats d'assurance dommages aux biens (pour les tempêtes) ou proposées en option (pour la grêle et la neige).

La garantie catastrophe naturelle n'est pas automatiquement incluse dans les contrats d'assurances. En revanche, c'est une extension de garantie obligatoire pour l'ensemble des contrats d'assurance de dommages, tels que l'assurance multirisque habitation (présentée en partie 1.2.1), l'assurance tous risques en automobile ou encore l'assurance local professionnel. Il convient de noter que les contrats d'assurance pour les bateaux peuvent ne pas contenir cette extension spécifique.

Le financement de cette garantie repose sur le paiement d'une surprise uniforme à l'échelle nationale, dont le montant, exprimé en pourcentage de la prime de départ, est déterminé par l'Etat. Cette surprime correspond à :

- 12% de la prime associée aux garanties dommages du contrat de base pour les biens autres que véhicules à moteur.
- 6% des primes vol et incendie pour les véhicules terrestres à moteur.

Ce fonctionnement est fondé sur le principe de solidarité, assurant ainsi une couverture à un prix accessible, même en présence d'expositions inégales au risque de catastrophes naturelles.

La garantie catastrophe naturelle possède un mécanisme de franchise et de limites. Le plafond

est contractuel tandis que la franchise est fixée par le gouvernement et définies en annexe de l'article A125-1 du Code des Assurances. Celle ci s'élève à :

- 380 euros pour les habitations et autres bien à usage non professionnels.
- 1520 euros dans le cas de dommages causés par un mouvement de terrain consécutif à la sécheresse ou une réhydratation du sol.

Par définition, ces montants sont à la charge de l'assuré et peuvent dans certains cas faire l'objet de majoration. La franchise peut être multipliée par 4 en fonction du nombre de reconnaissance en état de catastrophe naturelle d'une commune durant une période de 5 ans, si le dispositif de prévention pour le risque en question n'a pas été appliqué.

L'indemnisation d'un sinistre est effectuée conjointement entre l'assureur et la Caisse Centrale de Réassurance, qui est détenue par l'Etat français. Celle ci a la possibilité d'offrir une couverture illimitée à l'assureur, tandis que l'Etat joue le rôle de garant en cas de faillite. La couverture se décompose en deux sous-traités inextricablement liés :

- Un quote-part de 44%.
- Un stop-loss à 200% de la surprime Cat Nat.

Pour rappel, le quote-part est un traité de réassurance proportionnelle dans lequel l'assureur cède une partie fixe des risques couverts au risque au réassureur en contrepartie de la même partie de ses primes. Le stop-loss est quant à lui un traité non proportionnel dans lequel l'assureur transfère les risques au delà d'un certain seuil au réassureur. L'assureur est alors responsable de tous les sinistres jusqu'à ce seuil, mais une fois ce seuil atteint, l'excédent est à la charge du réassureur.

Dans la suite du fonctionnement du système d'indemnisation, une répartition est opérée au sein des 56% cédés dans le cadre du traité en quote-part, où 44% sont effectivement assumés par la CCR, tandis que les 12% restants sont pris en charge par le Fonds de Prévention des Risques Naturels Majeurs (FPRNM), également connu sous le nom de Fonds Barnier. En contrepartie de sa couverture auprès de la CCR, l'Etat francais est rémunéré sous forme de dividendes. Le mécanisme complet du régime Cat Nat est schématisé dans la figure 1.3.

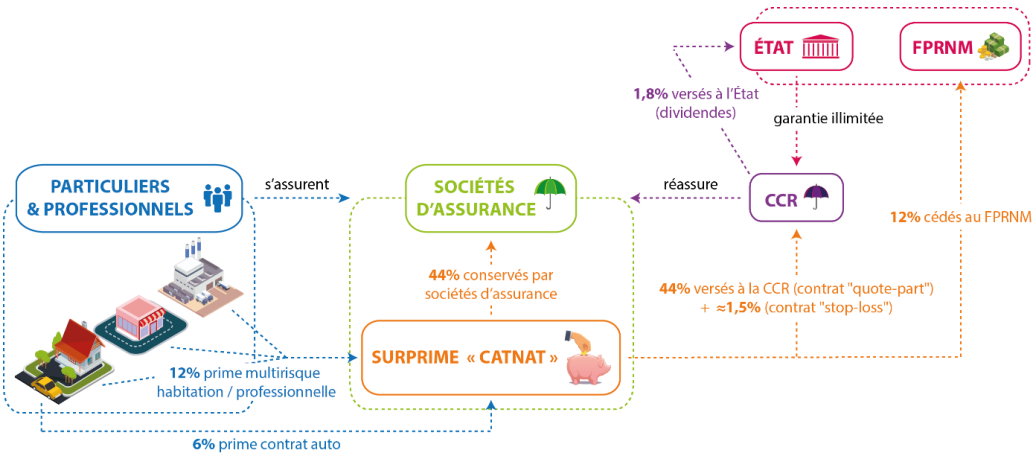


FIGURE 1.3 : Schéma du fonctionnement complet du régime Cat Nat - CAZAUX et al. (2019)

Dans le contexte spécifique du risque sécheresse, le régime d'indemnisation des catastrophes naturelles s'applique sur les dégâts liés au phénomène de subsidence décrit dans la partie 1.1.2, lorsque les

critères d'éligibilité définis au préalable sont respectés. En raison de sa cinétique lente, la sécheresse ne s'inscrit pas directement dans la catégorie des événements liés à une intensité anormale d'un événement naturel. Cette particularité rend plus complexe l'établissement d'un lien de causalité entre les dommages subis et le phénomène, qui constitue un élément essentiel pour la reconnaissance de l'état de catastrophe naturelle. La partie suivante présente les conditions retenues par les autorités pour caractériser un événement de sécheresse anormal ainsi que la place qu'elle occupe au sein du régime.

1.2.3 La sécheresse au sein du régime Cat Nat

Critères de reconnaissance Cat Nat

La prise en charge des périls sécheresse dans le régime Cat Nat a connu de nombreuses évolutions réglementaires ces dernières années. En effet, depuis 1989 huit ajustements ont été apportés aux critère de reconnaissance de l'état de catastrophe naturelle. En 2019, date de la dernière réforme, deux nouveaux critères, plus pertinents scientifiquement, sont désormais requis spécifiquement pour la sécheresse afin d'obtenir l'éligibilité d'une commune sinistrée (Circulaire (2019)).

Le **premier critère, géotechnique**, lié à la présence d'argiles sensibles au phénomène de retrait-gonflement des sols, en place depuis 1989, est maintenu. Pour que ce premier critère soit validé, il est nécessaire que la surface exposée au risque de retrait-gonflement des argiles d'une commune, quelque soit le niveau d'exposition, soit supérieure à 3% de la surface totale de la commune. Ce critère, qui permet d'identifier les sols présentant une prédisposition, se base sur les données ainsi que la cartographie produites par le BRGM, présentées en section 1.1.2. Néanmoins, ce dernier n'est pas suffisant pour évaluer l'intensité d"un épisode de sécheresse et la réglementation impose qu'il soit associé à un critère météorologique.

Le **second critère météorologique**, utilisé pour caractériser l'ampleur de la déshydratation du sol, se base sur une seule variable hydrométéorologique le SWI, présenté dans la partie 1.1.3, et une période de retour. Plus précisément, comme expliqué dans Delorme (2022), l'Etat utilise les données du modèle hydro-météorologique Safran-Isba-Moscou (SIM) de Météo-France qui simule numériquement au pas de temps journaliers l'indice SWI sur une grille avec une résolution de 8 kilomètres. A partir de l'indice journalier, un SWI mensuel est calculé pour chaque mois m comme suit

$$\overline{\mathrm{SWI}}_m = \frac{1}{n_m} \sum_{k=1}^{n_m} \mathrm{SWI}_{k,m},$$

avec n_m le nombre de jours du mois m et $\mathrm{SWI}_{k,m}$ la valeur de l'indice pour le $k^{\text{ème}}$ jour du mois m.

Puis, afin de rester cohérent avec la temporalité des épisodes de sécheresse, l'approche choisie consiste à utiliser une autre série mensuelle qui s'obtient, pour un mois donnée, à partir de la moyenne de son $\overline{\mathrm{SWI}}$ et de ceux des deux mois qui le précèdent. Ceci conduit à retenir l'indice

$$\overline{\mathrm{SWI}}^C_m = \frac{\overline{\mathrm{SWI}}_m + \overline{\mathrm{SWI}}_{m-1} + \overline{\mathrm{SWI}}_{m-2}}{3}.$$

En se basant sur l'indice élaboré précédemment, un seuil unique est retenu pour qualifier le caractère anormal d'une sécheresse au sens de l'article L.125-1 du Code des assurances. Ce dernier correspond à une période de retour supérieure ou égale à 25 ans, sur un historique de 50 ans. Ainsi, l'autorité administrative compare une valeur calculée pour un mois et une année donnés aux valeurs

des 49 années précédentes pour ce même mois. Cette méthode qui se réfère à une période "glissante" et intègre les années les plus récentes permet de tenir compte de l'évolution du climat. Le deuxième critère est donc déclenché dès lors que l'indice $\overline{\text{SWI}}^C_m$ est inférieure ou égale à la deuxième plus petite valeur de la série. Avec des notations mathématiques, cela revient à dire que le critère est validé pour le mois m de l'année N si

$$\overline{\text{SWI}}^C_{m,N} \leq \min_{j=0,\dots,49}{}^{(2)} \overline{\text{SWI}}^C_{m,N-j}$$

,

où $\min^{(2)}$ correspond à la deuxième plus petite valeur de l'ensemble considéré.

Une autre manière de caractériser ce critère est de dire que le seuil retenu correspond à un quantile d'ordre 4% ou moins de l'indice $\overline{\text{SWI}}^C_m$.

Par ailleurs, le critère sera apprécié pour chaque saison d'une année civile avec une catégorisation de la saisonalité particulière : durant l'hiver (Janvier à Mars), le printemps (avril à juin), l'été (juillet à septembre) et l'automne (octobre à décembre). Ainsi, si une commune est reconnue en état de catastrophe naturelle pour un mois d'une saison, elle gardera son éligibilité tout au long de la saison en question.

Critiques autour des critères d'éligibilité

Le mécanisme retenu pour l'éligibilité des communes présenté ci-dessus a fait l'objet de nombreuses critiques. Une première critique est faite quant à la caractérisation de l'état de catastrophe naturelle au niveau communal. Des communes limitrophes peuvent présenter la même exposition au risque de subsidence sans pour autant connaître le même traitement en cas de sinistres. En effet, le découpage administratif ne reflétant pas la prédisposition des sols, il est possible qu'un événement de sécheresse touche plusieurs communes qui ne remplissent pas l'ensemble des critères de reconnaissance. Cela donne lieu à des situations injustes où les communes les plus sinistrées ne sont pas indemnisés. Entre 2013 et 2021, seulement la moitié des demandes de reconnaissances ont été favorables.

Par ailleurs, Charpentier et al. (2022b) indique que le choix de retenir l'exceptionnalité des causes pourrait provoquer à terme une diminution du taux de demande favorable. Si une commune effectue des demandes de reconnaissance de manière récurrente, l'événement de sécheresse perd son caractère anormal et la probabilité que les critères soient déclenchés devient de plus en plus faible. Dans ce cas, sans connaître des événements de sécheresse extrême, la commune subit des pertes récurrentes, sans être indemnisée. Cette limite du système d'indemnisation est également souligné par le Sénat (2023). Ce dernier évoque la possibilité de substituer cette notion d'exceptionnalié des causes à celle d'exceptionnalité des conséquences, donc des dommages constatés. La Caisse Centrale de Réassurance a effectué des scénarios pour évaluer les impacts financiers d'une réforme de cette nature sur le régime Cat Nat. Les conclusions de ces estimations indiquent que, à moins d'imposer des contraintes drastiques limitant considérablement le nombre de sinistrés éligibles et les montants d'indemnisation, une telle réforme entraînerait des coûts trop importants susceptibles de bouleverser profondément l'équilibre financier du régime.

Face aux inadéquations mentionnées ci-dessus, une ordonnance a été présentée en Conseil des ministres le 8 février 2023, conformément à l'article 161 de la loi "3DS". Cette ordonnance ne visait pas initialement à modifier la logique de la prise en charge du risque RGA. Cependant, le gouvernement s'est engagé, sans que cela soit prévu dans l'ordonnance elle-même, à modifier les critères de reconnaissance de l'état de catastrophe naturelle pour le risque RGA par voie réglementaire. Ces modifications comprennent la simplification du critère météorologique, l'éligibilité automatique d'une commune li-

mitrophe d'une commune reconnue en état de catastrophe naturelle, ainsi que la reconnaissance de l'état de catastrophe naturelle en raison d'une succession de sécheresses d'ampleur moyenne.

Enfin, le 6 avril 2023, une proposition de loi (Assemblee Nationale (2023)) visant à "*mieux indemniser les dégâts sur les biens immobiliers causés par le retrait-gonflemet de l'argile*" a été adoptée par l'Assemblée nationale et doit désormais être examinée par le Sénat. Le texte a pour objectif d'assouplir les critères de reconnaissance de l'état de catastrophe naturelle par le biais de nouvelles mesures, avec notamment l'abaissement de la période de retour à 10 ans pour caractériser une sécheresse extrême. L'effet de ces directives sur l'éligibilité des communes à horizon futur est présenté en annexe A.1.

1.3 Changement climatique et enjeux assurantiels

Cette section se consacre en premier lieu à l'analyse des répercussions du changement climatique sur l'apparition des événements de sécheresse. L'anticipation de la multiplication des périodes de sécheresse dans les années à venir représente un enjeu majeur pour l'industrie de l'assurance. Les acteurs de ce domaine ont effectué de nombreuses projections concernant la sinistralité associée à la sécheresse, dont les résultats sont exposés par la suite. Parallèlement, l'examen de la viabilité à long terme du régime Cat Nat s'impose, étant donné la croissance significative du risque sécheresse. Enfin, afin d'aborder de manière proactive la gestion du péril sécheresse, les modèles employés pour appréhender concrètement ce type de risques sont présentés.

1.3.1 Les effets du changement climatique sur la sécheresse

Dès le début des années 2000, la communauté scientifique et notamment Bradford (2000), ont cherché à mieux comprendre le potentiel lien qui existait entre la sécheresse et le changement climatique. Un consensus se dessinait, prévoyant une augmentation future de la fréquence de ces événements climatiques. Plus récemment, des preuves supplémentaires sur l'influence du changement climatique sur les sécheresses en Europe a été apporté par Iglesias et al. (2018).

De même, dans un article paru en 2021, Ionita et Nagavciuc (2021) étudie l'évolution de trois indices de sécheresses (le SPI, le SPEI et le PDSI présentés en partie 1.1.3) sur la période 1901-2019. Cette étude met dans un premier temps en évidence la corrélation significative entre l'occurence des événements de sécheresse et les variables climatiques, telles que l'évapotranspiration potentielle et la température. Dans un second temps, elle souligne également que ces variables sont susceptibles d'augmenter sous l'effet du réchauffement climatique, suscitant ainsi des préoccupations quant à l'évolution future de ce risque. La figure 1.4 montre l'évolution temporelle des zones touchées par un événement de sécheresse dans la région de l'Europe centrale et région méditerranéenne à travers les indices considérés.

De manière similaire, les études menées par Spinoni et al. (2015) ont apporté des éclairages sur l'évolution des conditions climatiques en Europe, en se focalisant sur la survenance des sécheresses. Dans leur travail initial, qui couvrait l'ensemble du globe, Spinoni et al. (2014) ont observé des changements significatifs dans la fréquence et la gravité des phénomènes de sécheresse.

Concentrant par la suite leurs recherches sur l'Europe, Spinoni et al. (2017) ont constaté que ces changements étaient plus marqués au cours des trois dernières décennies. A partir des indices de sécheresses comme le SPI ou le SPEI, une analyse de l'évolution du nombre d'événements de sécheresse par décennie à travers l'indicateur DRF (Drought Frequency and Trend Values) a été réalisée sur la

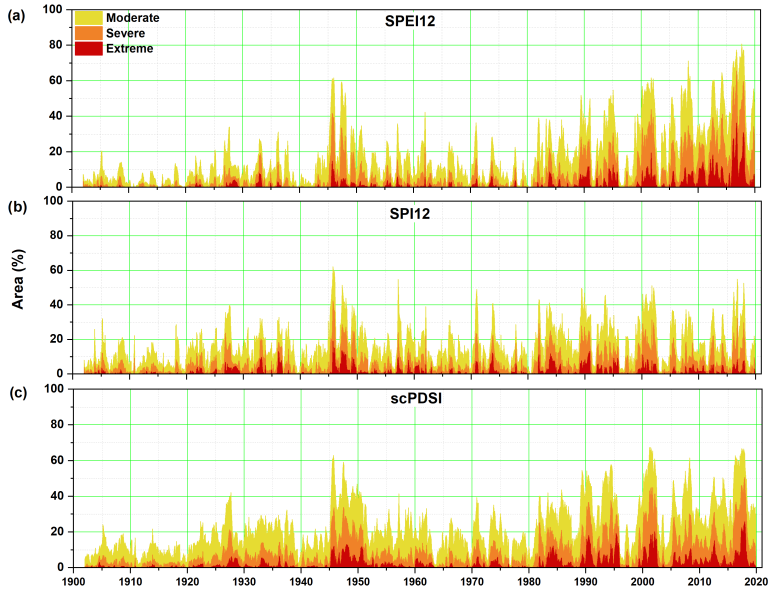


FIGURE 1.4 : Evolution temporelle des zones touchées par la sécheresse - IONITA et NAGAVCIUC (2021)

période 1950-2015. En particulier, une tendance vers des conditions plus sèches a été observée au printemps en Europe centrale, en été dans la région méditerranéenne, et en automne en Europe de l'Est (*cf.* figure 1.5).

En France, le projet CLIMSEC (SOUBEYROUX et al. (2011)) avait pour objectif de caractériser la sécheresse sur le territoire métropolitain et d'analyser ses potentiels liens avec le changement climatique. En se basant sur les précipitations ainsi que sur l'indice d'humidité des sols SWI, il dresse la progression attendue de la sécheresse à travers divers scénarios socio-économiques. Les résultats principaux mettent en évidence une croissance de l'intensité et des fréquences de la sécheresse, en particulier lors de la deuxième moitié du XXIè siècle (*cf.* figure 1.6). Cette aggravation du phénomène, provoquée par un déficit hydrique du sol plutôt que par un déficit pluviométrique, résulte du bouleversement des conditions climatiques dans les différents scénarios de projection considérés.

1.3.2 Projection de la sinistralité à climat futur

La multiplication annoncée des événements de sécheresse au cours du siècle est un enjeu préoccupant pour le marché de l'assurance. En effet, comme expliqué dans la partie 1.1.2, le phénomène de subsidence, cause des dégâts importants sur le bâti et vient influer directement sur l'activité "dommages aux biens" des assureurs. Dans cette perspective, de nombreuses études ont été réalisées par l'industrie de l'assurance afin de quantifier la sinistralité croissante liée notamment aux catastrophes naturelles.

En 2020, l'ACPR (2021) a réalisé un exercice pilote avec l'aide de la CCR, visant à estimer, pour les organismes d'assurance concernés, les dommages subis sur la période 2020-2050 pour l'ensemble des périls couverts par le régime Cat Nat en France. Les résultats de ces projections, qui se basent sur le scénario RCP 8.5 du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), indique une croissance de la sinistralité de 174% entre 2019 et 2050 pour les branches prises en compte dans le calcul de la contribution au régime Cat Nat. Lors de cette étude, les participants avaient la possibilité de revoir leur politique de souscription en fonction de l'évolution de la sinistralité, en

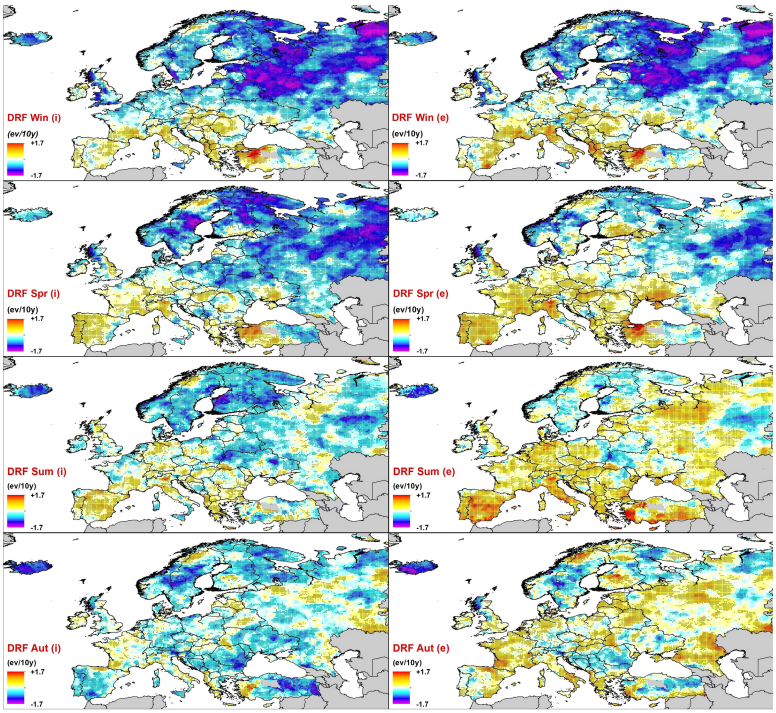


FIGURE 1.5 : DRF par saison sur la période 1950-2015 - SPINONI et al. (2017)

réallouant géographiquement leur portefeuille, en augmentant les primes, en révisant les programmes de réassurance ou encore en adaptant les produits offerts. Le rapport mentionne que les participants ont principalement choisi de maintenir le ratio sinistres sur primes constant, ce qui entraîne une augmentation des primes d'assurance entre 130% et 200% sur 30 ans. Par ailleurs l'ACPR indique que les assureurs n'ont pas exploité la possibilité de modifier leur stratégie de souscription pour sortir des zones les plus impactées par l'augmentation de la sinistralité ou refuser d'assurer les zones les plus exposées au changement climatique.

La CCR (2018) s'est elle aussi penchée sur les conséquences du changement climatique sur le coût des catastrophes naturelles. Spécifiquement pour les dommages consécutifs à la sécheresse géotechnique, la sinistralité augmenterait entre 20 à 60% sur tout le territoire métropolitain à horizon 2050. Cette progression serait d'autant plus prononcée dans la partie méridionale du pays en raison de l'intensification du phénomène, tandis que sur toute la côte atlantique, elle serait influencée par l'évolution significative des biens assurés. Dans une dernière étude parue en octobre 2023, la CCR (2023a) alerte à nouveau sur l'expansion du risque sécheresse à horizon 2050, avec une croissance du montant annuel moyen des dommages due à l'aléa comprise entre 44% et 162% selon les scénarios du GIEC.

En 2015, France Assureur a entrepris une évaluation prospective des effets du climat sur l'industrie de l'assurance jusqu'à l'horizon 2040 (FFA (2015)). Une étude plus récente, publiée en 2021 (FFA (2021)), vient actualiser ces travaux en se projetant jusqu'en 2050. Cette dernière étude fournit des résultats détaillés pour chaque type de catastrophe, mettant en évidence une tendance à la hausse des coûts des événements naturels, confirmant ainsi la croissance alarmante attendue au cours des prochaines années. Toujours en se basant sur le scénario les trajectoires climatiques du GIEC, le rapport indique une augmentation significative du coût cumulé estimé à 43 milliards d'euros d'ici à 2050 pour la sécheresse. Cette estimation dépasse de plus de trois fois la charge moyenne annuelle observée au cours des trois dernières décennies, soulignant l'ampleur croissante de l'impact financier associé

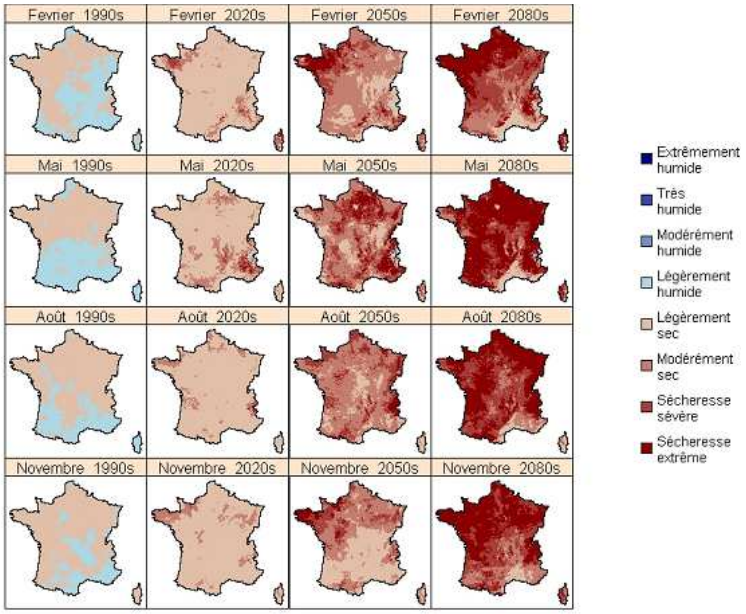


FIGURE 1.6 : Evolution temporelle des sécheresses (SWI) au cours du XXIème siècle selon les saisons - SOUBEYROUX et al. (2011)

aux événements de sécheresse. Dans ces simulations, la hausse de la sinistralité est principalement attribuée à cinq départements, à savoir la Haute-Garonne, la Gironde, les Bouches-du-Rhône, le Tarn-et-Garonne et le Tarn, qui à eux seuls représentent les deux tiers de cette augmentation.

Enfin COVÉA (2022) propose dans son livre blanc une projection du risque RGA à travers une nouvelle fois le scénario le plus pessimiste du GIEC. L'étude prévoit une augmentation de 70% de la fréquence d'éligibilité Cat Nat pour ce risque à horizon 2050. Cette hausse touche l'ensemble du territoire métropolitain avec des zones plus sensibles comme le croissant argileux, le Grand Est ou encore la Bretagne. Toujours selon l'analyse effectuée par Covea, une croissance de l'ordre de 60% de la sinistralité sécheresse est attendue en 2050. Sous l'effet du changement climatique, le croissant argileux, le bassin parisien, le Centre-Est ou encore les Hauts-de-France seraient particulièrement affectés.

1.3.3 Un régime cat nat menacé

Un déséquilibre financier

Le risque de sécheresse représente une part significative des prestations du régime Cat Nat. En juin 2022, d'après les chiffres clés de la CCR (2022), il se positionne comme le deuxième poste d'indemnisation des sinistres Cat Nat, représentant 37% de l'ensemble de la sinistralité, juste après les inondations qui représentent 53% (*cf.* figure 1.7). Depuis 1989, année d'intégration de la sécheresse dans le régime Cat Nat, 8 des 20 événements naturels les plus importants en France sont liés à la sécheresse. D'après la CCR, entre 1989 et 2020, le coût total de la sécheresse s'élève à près de 15,2 milliards d'euros.

Le risque sécheresse se démarque des autres catastrophes naturelles non seulement par la place qu'il

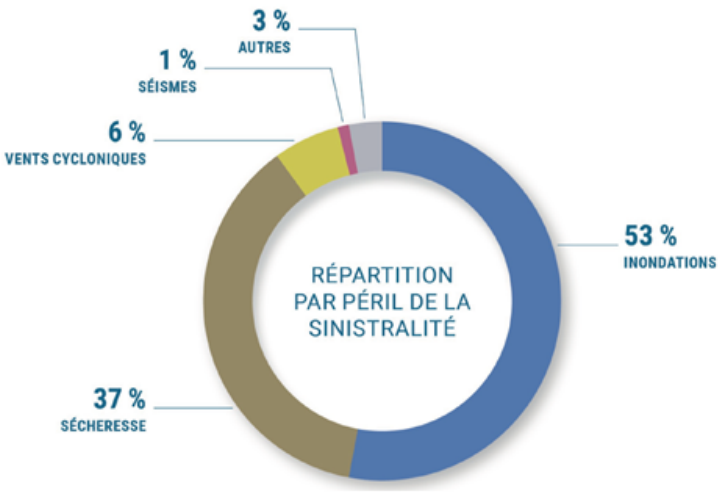


FIGURE 1.7 : Répartition par péril de la sinistralité - CCR (2022)

occupe au sein du régime, mais aussi par son caractère progressif. Cette dimension évolutive, provoquée par le dérèglement climatique, est vouée à s'amplifier et compromet notamment la pérennité du régime Cat Nat. Dans son rapport d'information, le SÉNAT (2023) souligne l'évolution déjà perceptible du risque sécheresse sur la période plus récente. Entre 2017 et 2020, la charge annuelle moyenne associée au RGA a dépassé 1 milliard d'euros, tandis qu'elle s'élevait seulement à 445 millions d'euros depuis 1982. Au sein du régime Cat Nat, le péril sécheresse est l'un des risques naturels qui progresse le plus vite.

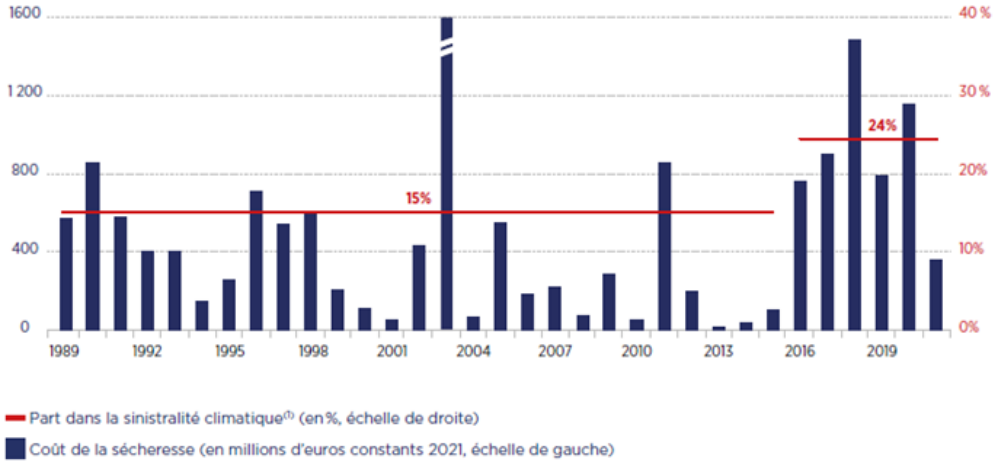


FIGURE 1.8 : Evolution du coût de la sécheresse - France Assureur (FFA (2022))

De même, le coût de la sécheresse survenue en 2022 est estimée à plus de 3 milliard d'euros (CCR (2023b)), ce qui la situe largement au dessus de l'événement de sécheresse le plus coûteux, celui de 2003, évalué à 1,6 milliard d'euros. Le rapport indique également que depuis 5 ans, le régime Cat Nat est déficitaire. En 2017, le régime a enregistré son déficit le plus élevé depuis sa création, s'élevant à 439 millions d'euros. Sur la période allant de 2015 à 2019, le déficit cumulé du régime a atteint près

d'un milliard d'euros. Selon la Cour des comptes (2022), les réserves de la CCR ont baissé de 44% entre 2015 et 2020, passant de 3,85 milliard d'euros à 2,67.

Comme évoqué dans la partie précédente, la sinistralité liée au risque RGA est vouée à considérablement augmenter ces prochaines décennies. En se basant sur son scénario "optimiste", la Caisse Centrale de Réassurance estime que le régime Cat Nat ne sera plus en mesure de constituer suffisamment de réserves pour couvrir les sinistres d'ici 2040, principalement en raison du coût croissant lié au risque RGA. Elle estime également que le déficit de financement du régime catastrophes naturelles s'élèvera à 420 millions d'euros par an à horizon 2050.

Ainsi, la sécheresse apparaît donc comme un facteur majeur du déséquilibre du régime. Afin de garantir sa soutenabilité , de nombreuses pistes d'évolutions sont alors envisagées.

Mesures de sauvegarde du régime

Dans son rapport au ministère, la CCR alerte dans un premier temps sur la nécessité de réajuster le niveau de la surprime Cat Nat. Cette piste est également suggérée par Franck Le Vallois, directeur général de France Assureurs, dans un article de l'argus de l'assurance du 22 mars 2023 (Argus (2023)). Il y a un consensus autour du fait que cette hausse de la surprime permettrait de rééquilibrer assez rapidement le régime des catastrophes naturelles.

Dans son rapport, la Cour des Comptes proposait également à l'Etat de se questionner sur une éventuelle sortie du risque sécheresse du régime. Néanmoins, cette piste viendrait supprimer le principe de solidarité, propre au régime, et une telle sortie ne permettrait plus de procéder à la mutualisation financière entre risque qui existe au sein du régime et qui bénéficie actuellement au risque RGA. Le Sénat indique également dans son rapport que le risque RGA ne pourrait pas être assumé par le système assurantiel privé. Les primes d'assurance deviendraient alors excessives pour les individus résidant dans des zones à risque, ce qui rendrait l'assurance inaccessible à de nombreux particuliers. De ce fait, cette perspective fut rapidement écartée par le gouvernement.

Comme souligné par Franck Le Vallois dans l'Argus (2023) ou encore par la Cour des comptes (2022), la notion de prévention est également primordiale pour la survie du régime. Dans son enquête, la Cour des comptes souligne l'importance de la prévention du risque dans le domaine de la construction. Elle met l'accent sur la nécessité de trouver des solutions de révisions efficaces pour réduire les dommages liés au phénomène de RGA, aussi bien pour les constructions existantes que pour celles à venir. Elle recommande la mise en place d'un dispositif de contrôle et de sanction des mesures prévues par la loi *Evolution du logement, de l'aménagement et du numérique* (ÉLAN) pour les nouvelles constructions en zones exposées au risque RGA. Enfin, elle encourage l'accélération des projets de recherche et développement pour développer des mesures de révisions adaptées aux constructions antérieures à 2020 et exposées au risque RGA, en privilégiant leur efficacité et leur coût.

1.3.4 Vers une explosion des primes ?

Lors de son exercice pilote, l'ACPR (2021) constate que la croissance des sinistres climatique conduirait à une augmentation des primes d'assurance entre 130% et 200% à horizon 2050 si les assureurs souhaitent maintenir une rentabilité constante. Une telle progression pourrait rendre les primes inabordables pour un certain nombre de ménages et soulève naturellement des questions de soutenabilité du point de vue des assurés.

Pour son prochain exercice, l'ACPR (2023) cherche à intégrer cette notion de soutenabilité en définissant des seuils de résiliation, basés sur le rapport entre la prime dommages, telle que définie à l'article L. 125-2 du Code des assurances, et la valeur totale assurée,

$$\frac{\text{Prime dommages}}{\text{Valeurs assurées (en K euros)}}.$$

Au delà d'un certain seuil, défini à la maille départementale par l'ACPR, les contrats de particuliers pour lesquels l'assuré est propriétaire du bien assuré sont supposés résiliés (l'assurance habitation étant obligatoire pour les locataires mais facultative pour les propriétaires). Cette méthodologie permet de prendre en compte le comportement de l'assuré face à une augmentation significative de sa prime dans la gestion du passif de l'assureur.

Pour continuer dans ce sens, le mémoire propose d'explorer cette question de soutenabilité à travers le prisme du phénomène RGA exposé précédemment. Pour aboutir à une estimation de la prime nécessaire pour couvrir le risque à climat futur et ainsi étudier sa viabilité, l'étude s'effectue en différentes étapes.

Il est tout d'abord essentiel de modéliser le risque pour pouvoir faire le lien entre la manifestation d'un phénomène de sécheresse et les pertes potentielles sur le portefeuille de l'assureur. Dans le cadre du mémoire, des données climatiques et géologiques sont exploitées pour comprendre les aspects physiques du risque et entraîner un modèle de *machine learning* capable de prédire la charge sinistre liée au phénomène RGA.

Ensuite, afin de quantifier le risque RGA à horizon futur, une projection des différents éléments composant le risque est effectuée. Cette projection requiert la considération de plusieurs facteurs, notamment l'évolution des enjeux assurés, comme la croissance démographique et l'accroissement des valeurs assurées du portefeuille, conjuguée à l'évolution du climat, laquelle est déterminée par les trajectoires d'émissions de gaz à effet de serre.

Une fois ces premières étapes effectuées, la sinistralité projetée permet de calculer l'évolution de la prime MRH et ainsi confronter la tarification du risque sécheresse à la capacité des assurés à assumer l'évolution des primes. Enfin, pour quantifier cette notion de soutenabilité, une métrique définie à partir du niveau de richesse des assurés est utilisée.

La méthodologie décrite ci-dessus est synthétisée dans la figure 1.9 et intègre une première étape de modélisation du risque sécheresse, suivi d'une phase de projections visant à appréhender l'évolution de la tarification de la prime MRH et, par conséquent, à questionner sa viabilité. Cette approche est employée dans ce mémoire pour répondre à l'objectif de l'étude, à savoir évaluer la soutenabilité d'une éventuelle explosion des primes MRH due à l'accroissement du phénomène RGA à horizon futur.

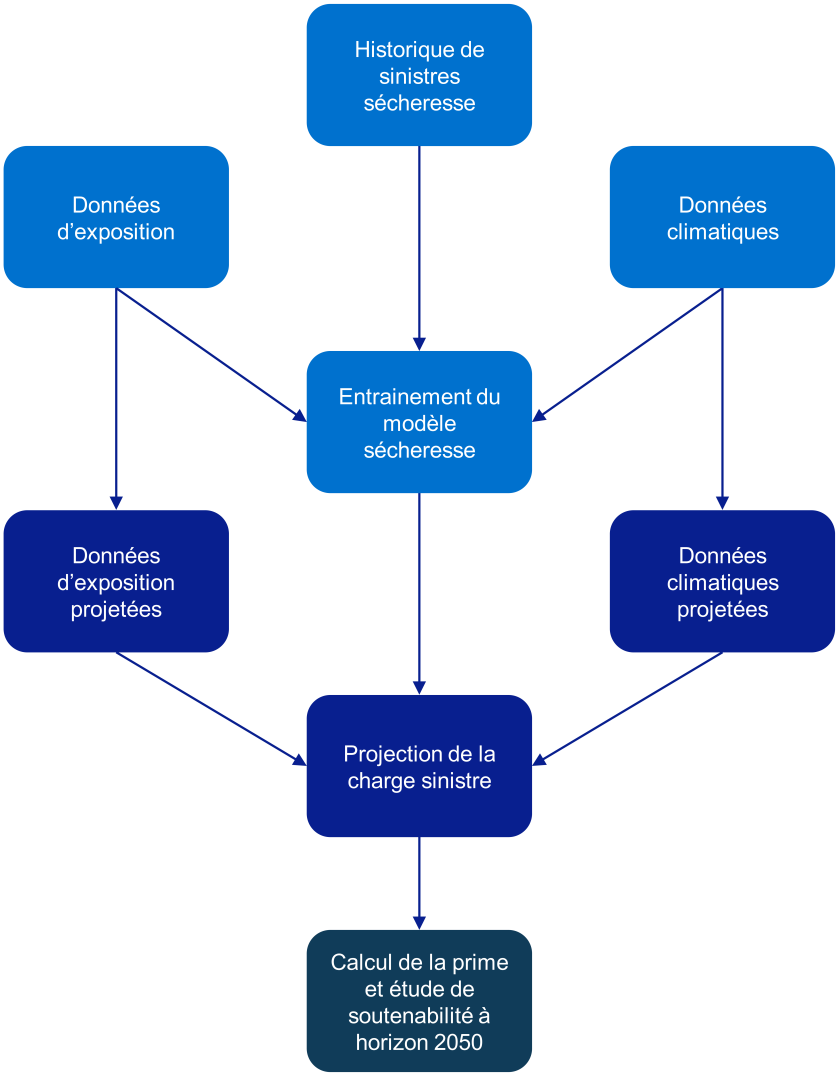


FIGURE 1.9 : Architecture de la méthodologie employée pour l'étude

Chapitre 2

Modélisation du risque sécheresse

Ce deuxième chapitre s'articule autour de l'élaboration d'un modèle de *machine learning*, relatif à la prédiction du péril sécheresse, qui sera utilisé dans la suite de l'étude. L'objectif consiste à développer un modèle reflétant au mieux la réalité des sinistres liés à la sécheresse géotechnique tout en gardant une complexité de modélisation climatique raisonnable afin de permettre une étude actuarielle pertinente.

Dans **une première section**, les caractéristiques du portefeuille d'assurance utilisé pour l'entraînement du modèle sont présentées.
Dans **une deuxième section**, la construction des indices de sécheresse, éléments constitutifs principaux du modèle, est effectuée.
Dans **une troisième section**, l'objectif du modèle ainsi que les outils théoriques employés pour l'étude sont détaillés.
Enfin, dans **une dernière section**, les étapes d'optimisation puis de validation du modèle sont abordées.

2.1 Présentation du portefeuille

Les données considérées dans cette étude correspondent à des données réelles d'un assureur dont le portefeuille est représentatif du marché français, autant en termes de distribution des sinistres qu'en répartition géographique des biens assurés. Ainsi, les observations du portefeuille en question vont être supposées comme généralisables pour toute la France métropolitaine. Le portefeuille employé pour le mémoire est uniquement constitué de contrats d'assurance Multi-Risque habitation (MRH). Comme décrit dans la partie 1.2.1, ce contrat comprend la garantie Cat Nat qui prend en charge les sinistres liés à la sécheresse, sous réserve du respect des critères d'éligibilité énoncés dans la section 1.2.3. Pour modéliser le risque sécheresse, les informations des sinistres RGA survenus sur la période 2000-2020 sont retenues puis agrégées à la maille départementale. Seules les données de sinistres rattachés à un arrêté de catastrophe naturelle du Journal Officiel sont recensées dans la base et servent à l'entraînement du modèle. Au sein de cette section, une étude descriptive de la sinistralité sécheresse du portefeuille est présentée ainsi que les données entrant dans la modélisation du risque.

2.1.1 Sinistralité sécheresse du portefeuille

Traitement "*as if*" de la charge sinistre

Afin de garantir une comparaison cohérente de la charge sinistre sur la période 2000-2020, une démarche de traitement "*as-if*" des données a été effectuée. Ce traitement est double car il intervient à la fois sur le montant des sinistres mais également sur l'exposition du portefeuille. L'objectif sous-jaçent est de comparer les charges sinistres annuelles en euros constants et à exposition constante.

La correction du montant de la charge sinistre a été réalisée en prenant en compte l'inflation des prix par rapport à une année de référence. Dans ce contexte, chaque montant de sinistre a été actualisé en euros 2020 sur la base de l'évolution de l'indice FFB du coût de la construction (ICC) (*cf.* annexe A.4). Cet indice, mis à disposition par la Fédération Française du Bâtiment, est déterminé à partir du coût de revient d'un immeuble de rapport de type courant à Paris. Le choix de cet indice se justifie par le fait que les sinistres sécheresse touchent le bâti.

Un autre ajustement est également apporté à la charge sinistre, visant cette fois-ci la correction de l'exposition du portefeuille. Compte tenu des évolutions observées dans le nombre de contrats et les valeurs assurées tout au long de la période considérée, une correction a été appliquée, basée une nouvelle fois sur l'année de référence 2020. Cette approche assure ainsi l'homogénéisation du montant total des sinistres, contribuant à la cohérence méthodologique qui guide la suite de l'étude.

Etude de la charge sinistre

La démarche précédemment exposée facilite la réalisation d'une analyse comparative annuelle des charges sinistres associées à la sécheresse géotechnique. Ceci permet notamment d'identifier et de mettre en relief les événements marquants de sécheresse survenus dans le portefeuille au cours de la période d'étude.

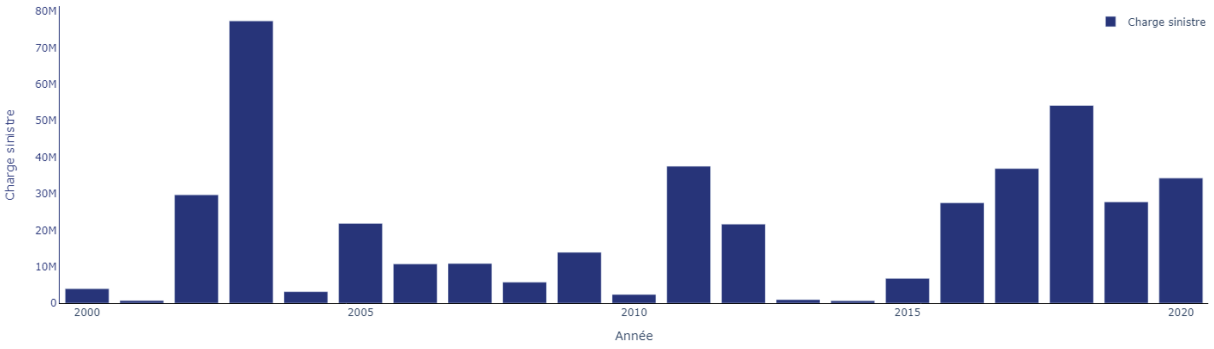


FIGURE 2.1 : Charge sinistre annuelle "Sécheresse" du portefeuille - Période 2000-2020.

L'évolution historique des charges sinistres annuelles, illustrée par la figure 2.1, révèle une dynamique marquée par une certaine volatilité. L'année 2003 se détache avec une charge sinistre exceptionnellement élevée, suivie par les années 2018, 2011 et 2017. Par ailleurs, le graphique met en évidence une tendance à l'accentuation de la sinistralité au cours des cinq dernières années, traduisant ainsi une intensification des événements de sécheresse sur le territoire métropolitain, comme évoqué en partie 1.3.3. En effet, la charge sinistre enregistrée sur la période de 2015 à 2020 représente à elle seule 43% de la charge sinistre totale du portefeuille.

Il est essentiel de souligner qu'un biais inhérent persiste dans la comparaison de la sinistralité annuelle, en raison des ajustements apportés aux critères d'éligibilité aux états de catastrophe naturelle

liés au péril sécheresse. Étant donné que ces critères ont évolué au cours des deux dernières décennies, la comparaison de la charge sinistre ne peut être effectuée en considérant une réglementation constante. Cependant, il est important de noter que les modifications dans les critères n'exercent pas une influence significative sur la tendance générale de la sinistralité et ne compromettent pas la lecture des différents niveaux de sinistralité des années.

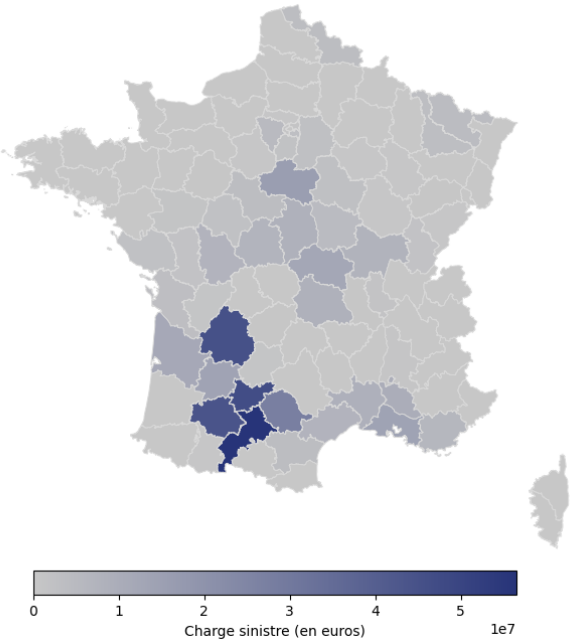


FIGURE 2.2 : Répartition de la charge sinistre cumulée "Sécheresse" du portefeuille - Période 2000-2020.

Sur le plan spatial, l'expression du péril sécheresse se matérialise, au sein du portefeuille, dans des régions bien spécifiques. Comme décrite dans la section 1.1.2, la manifestation du phénomène de sécheresse géotechnique prédomine principalement dans les zones caractérisées par la présence de sols argileux.

La figure 2.2 représente la répartition de la charge sinistre cumulée du portefeuille sur la période de l'étude. Cette répartition évoque celle du croissant argileux décrit en partie 1.1.2 et se superpose aux zones à forte exposition au phénomène de retrait-gonflement des argiles, telles qu'illustrées dans la figure 1.2. Les départements les plus sinistrés sont localisés dans le sud-ouest du territoire métroplitain. En effet, les départements de la Haute-Garonne, du Tarn-et-Garonne, de la Dordogne, du Gers et du Tarn regroupent à eux seuls 50% de la charge sinistre totale du portefeuille..

D'autres régions géographiques se distinguent également comme des zones à risque. Le long de la côte méditerranéenne, les départements du Gard, des Bouches-du-Rhône, du Var et de l'Hérault sont également sujets à de fréquents épisodes de sécheresse. Dans la région centrale de la France, les départements du Loiret et du Cher présentent une sinistralité marquée. Dans le nord du pays, des sinistres sont répertoriés dans le département du Nord, tandis qu'au nord-est, le département de la Moselle est également touché. Enfin, au sein dela région Ile-de-France, quelques sinistres atteignent les départements des Yvelines, des Hauts-de-Seine et de la Seine-et-Marne.

L'analyse de la répartition de la charge sinistre peut être enrichie par l'utilisation d'une courbe de Lorenz, un outil graphique qui offre une représentation concise et visuelle de la concentration des valeurs au sein d'une distribution. La construction de cette courbe suit plusieurs étapes précises.

Tout d'abord, les départements sont triés en fonction de leur charge sinistre, de la moins élevée à la plus élevée. Ensuite, la part cumulée de la charge sinistre Y_i est calculée pour chaque département, relativement à la somme totale des charges sinistres

$$Y_i = \frac{y_1 + y_2 + \ldots + y_i}{y_1 + y_2 + \ldots + y_{96}} \quad \text{pour } i = 1, 2, \ldots, 96,$$

avec $y_1, \ldots, y_{96}$ les charges sinistres triées par ordre croissant. Parallèlement, la part cumulée des départements $Q_i = \frac{i}{96}$ est calculée.pour $i = 1, 2, \ldots, 96$. Enfin, la courbe de Lorenz est tracée à l'aide des points (Y_i, Q_i), qui correspond à la part cumulée de la charge sinistre en fonction de la part cumulée des départements.

De manière similaire, la part cumulée idéale est établie, représentant une distribution égale de la charge sinistre entre les départements et correspondant à la ligne droite en pointillé montante à 45 degrés depuis l'origine sur la figure 2.3. La distance entre la courbe de Lorenz et la ligne de référence diagonale permet d'apprécier la répartition de la distribution de la variable étudiée, où une courbe s'éloignant davantage de cette ligne traduit une distribution plus inégale.

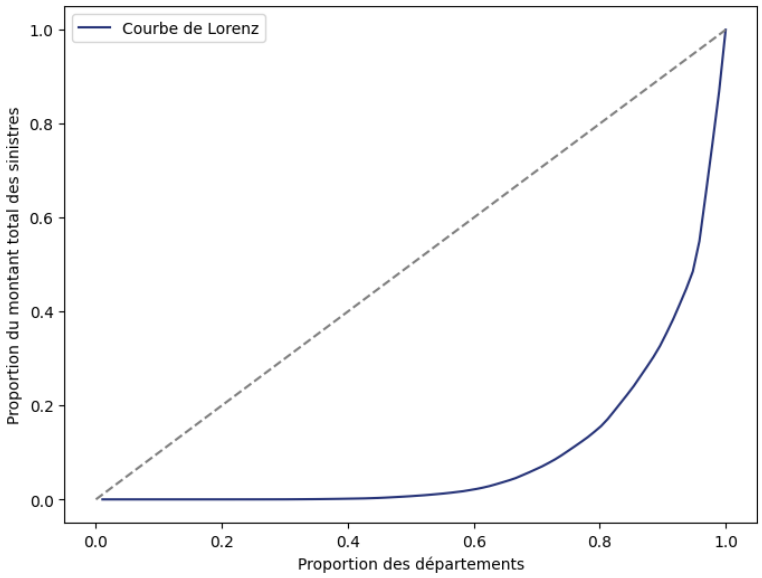


FIGURE 2.3 : Courbe de Lorenz de la charge sinistre cumulée "Sécheresse" du portefeuille - Période 2000-2020.

La figure 2.3 illustre bien la concentration de la charge sinistre sécheresse sur une faible proportion des départements. Le point de coordonnées $(0.80, 0.15)$ met en évidence que 85% de la charge sinistre est supporté par seulement 20% des départements.

2.1.2 Données d'exposition

Données d'exposition au RGA

L'exploitation des données d'exposition des territoires au risque de retrait-gonflement des argiles (RGA), fournies par le portail Georisque (2023b), permet d'élaborer des variables explicatives relatives à l'exposition du portefeuille. Ces variables sont vouées à enrichir ultérieurement la précision du modèle. Comme expliqué en partie 1.2, le fichier d'exposition au risque RGA couvre l'ensemble du territoire français métropolitain, englobant plus de 100 000 polygones géocodés. Ces polygones identifient les territoires classés selon les niveaux de risque "faible", "moyen" et "fort" en termes d'exposition au retrait-gonflement des argiles.

Ces données permettent de calculer la superficie de chaque département exposée à chaque niveau de risque. Cette valeur est ensuite mise en relation avec la superficie totale du département, permettant ainsi d'obtenir la proportion de la surface départementale exposée à chaque classe de risque. Cette donnée peut être interprétée comme une mesure de l'exposition potentielle d'un département. Plus précisément, un département dont la proportion de la surface exposée au risque RGA est élevée et dont la population est vouée à augmenter de manière importante peut voir son parc de logement s'étendre et venir se superposer dans des zones à risques.

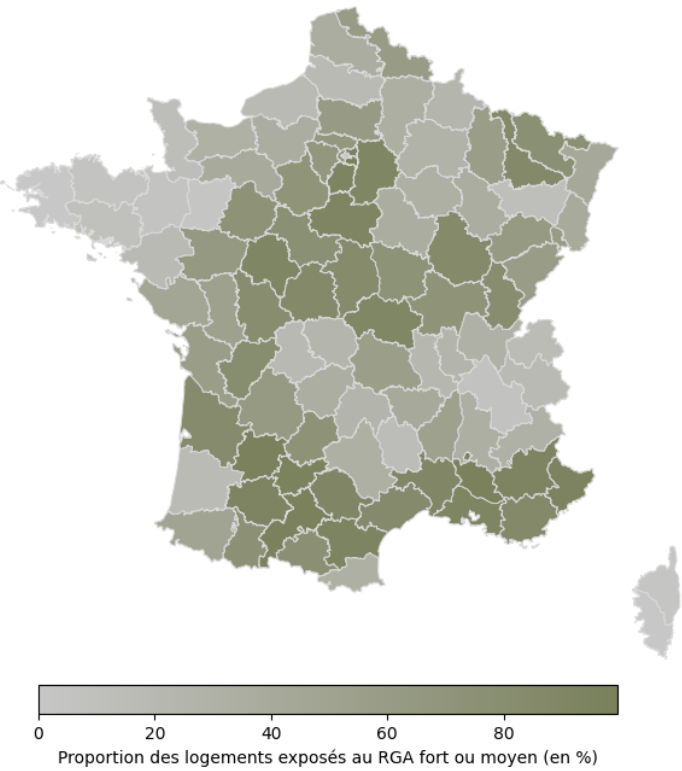


Figure 2.4 : Part des logements exposés au risque fort et moyen RGA

De plus, avec l'aide du BRGM, le Service des données et études Statistiques (SDES (2021)) des ministères chargés de l'environnement, de l'énergie, de la construction, du logement et des transports a mis à jour en 2021 un indicateur d'exposition des maisons individuelles au retrait gonflement des argiles

par commune. L'exploitation et l'agrégation de ces données ont permis de quantifier la proportion des logements par département se situant dans les différentes zones d'exposition aux risques.

La représentation graphique fournie par la figure 2.4 illustre l'exposition du portefeuille au risque RGA. Celle-ci montre que les départements du croissant argileux présentent naturellement une part importante de leur logements dans des zones à risque. Une exposition significative est également marquée en Île-de-France ainsi que dans la région Provence-Alpes-Côte d'Azur.

Au sein de la base sinistre historique, seulement 5% des sinistres ont été enregistrés dans des zones à risque RGA faible, tandis que 95% sont observés dans des zones à risque RGA moyen ou fort. Cette distribution souligne de manière significative l'importance prépondérante des sols argileux dans l'apparition des sinistres liés à la sécheresse, renforçant ainsi la nécessité d'intégrer une dimension géologique dans le modèle.

Le lien entre les sinistres dus à la sécheresse dans un département et la proportion de logements exposés à un risque RGA moyen ou fort peut être quantifié à l'aide du coefficient de corrélation de Kendall. Ce dernier, noté τ (tau), permet de mesurer l'association entre deux variables et repose sur le principe de concordance. Une description plus détaillée du calcul de cette mesure statistique est effectuée en annexe A.5.

Dans notre cas, le tau de kendall entre la proportion de logement par département se situant en zone de risque RGA fort ou moyen et le montant de la charge sinistre est de 0.62. Cette valeur met une nouvelle fois en évidence le lien fort qui relie les deux variables.

Données sur les logements

L'impact prédominant du phénomène de subsidence sur les maisons, en raison de leurs fondations relativement moins robustes par rapport aux appartements, confère une importance accrue à la nature du bien. Au sein de la base sinistre historique, 92% des sinistres concernent des dégâts sur des maisons. De ce fait, il est donc judicieux de distinguer au sein du modèle les différents types de bien assurés afin d'affiner ses prédictions.

Ainsi, pour chaque année de la période considérée et pour chaque département, le portefeuille intègre des informations telles que le montant de la charge sinistre, les valeurs assurées et le nombre de contrats. Cette base est ensuite enrichie avec les données d'exposition précédemment évoquées, comprenant la proportion de la surface départementale exposée à divers niveaux de risque RGA, la proportion de logements dans le portefeuille également exposée au risque RGA pour chaque département, ainsi que des détails concernant la répartition des maisons et des appartements au sein du portefeuille.

Après avoir consolidé les données d'exposition, la base d'entraînement du modèle doit être complétée en intégrant des informations climatiques afin d'obtenir une meilleure compréhension des facteurs météorologiques sous-jacents aux phénomènes de sécheresse. Dans cette optique, la prochaine section se consacre à la création d'indices de sécheresse, conçus pour établir des liens entre les conditions climatiques d'une zone donnée et l'occurrence des sinistres liés à la sécheresse.

2.2 Construction des indices de sécheresse

Cette section se focalise sur l'élaboration de deux indices de sécheresse distincts : le premier, de nature météorologique, le SPEI-3, préalablement introduit dans la partie 1.1.3 ; le second, de nature hydrologique, reposant sur l'indice SWI utilisé dans le cadre du dispositif Cat Nat, explicitement présenté dans la partie 1.2.3. La mise en place de ces indices s'opère par l'acquisition des données climatiques nécessaires à leurs calculs respectifs. Pour la phase d'apprentissage du modèle, la plage temporelle allant de 2000 à 2020 a été retenue.

Par ailleurs, l'objectif de ce mémoire réside dans l'analyse de l'évolution du péril sécheresse à l'échelle intégrale du territoire métropolitain. Ainsi, pour l'entraînement du modèle, il est nécessaire de disposer des indices recouvrant la France métropolitaine pour la période considérée. Afin de concilier cet impératif et les enjeux engendrés par le volume de données requis, le choix d'une granularité départementale et d'une fréquence annuelle pour les indices de sécheresse a été retenu. Cette démarche permet d'obtenir un compromis entre la rigueur de l'analyse souhaitée et les défis induits par la gestion des données à large échelle.

2.2.1 Récupération et traitement des données climatiques

Données météorologiques

Les données employées pour l'élaboration de l'indice météorologique SPEI-3 sont extraites des fichiers messages CLIMAT de Météo France (2023). Ces fichiers sont accessibles en tant que données publiques et fournissent à une cadence mensuelle de nombreux paramètres climatologiques issus des stations de Métropole et d'Outre-Mer appartenant au Réseau Climatologique Régional de Base (RCBN) de l'Organisation Météorologique Mondiale (OMM) (*cf.* figure 2.5). Parmi la panoplie des paramètres disponibles, la température (exprimée en degrés celsius) ainsi que les précipitations (mesurées en millimètres) sont utilisées pour la construction de l'indice SPEI-3.

Une fois les données des stations recueillies, l'objectif est d'obtenir les données météorologiques à la maille départementale. Pour ce faire, le processus consiste tou d'abord à sélectionner un point géographique de référence pour chaque département. Par la suite, un lissage géospatial est effectué en vue de calculer les valeurs des variables climatiques à ce point de référence utilisant les données issues des différentes stations. Le recours à ce point de référence propre à chaque département vise à réduire la dimensionnalité des données employées pour l'étude, tout en cherchant à préserver la représentativité des caractéristiques climatiques inhérentes à chaque département. En ce qui concerne l'indice météorologique, le point de référence retenu correspond au centroïde du département, comme illustré dans la figure 2.6.

Sur le plan géométrique, un département peut être appréhendé comme un polygone à n sommets. Pour un polygone avec n sommets, représenté par les coordonnées (x_i, y_i) pour i allant de 1 à n, les coordonnées du centroïde $C = (C_x, C_y)$ sont données par

$$C_x = \frac{1}{6A} \sum_{i=0}^{n-1} (x_i + x_{i+1}) \times (x_i \times y_{i+1} - x_{i+1} \times y_i)$$

$$C_y = \frac{1}{6A} \sum_{i=0}^{n-1} (y_i + y_{i+1}) \times (x_i \times y_{i+1} - x_{i+1} \times y_i)$$

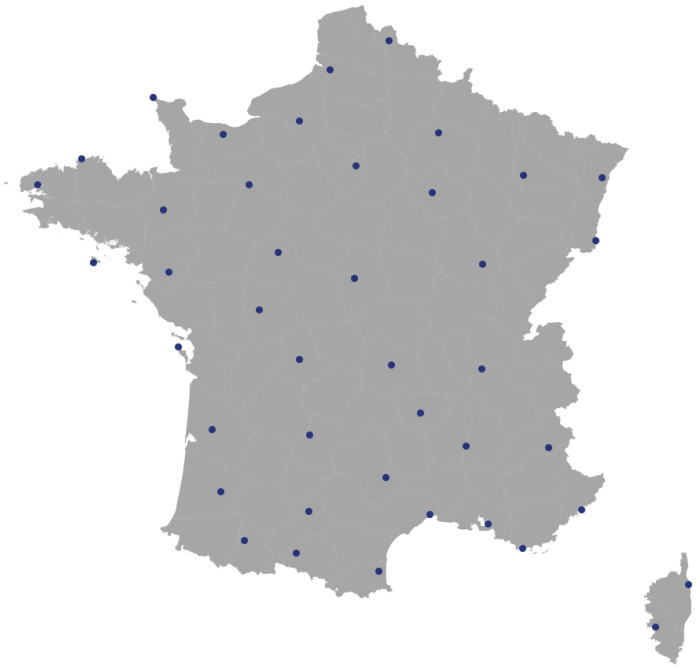


FIGURE 2.5 : Stations Météo France appartenant au RCBN

où A représente l'aire signée du polygone, telle que décrite par la formule du lacet

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i \times y_{i+1} - x_{i+1} \times y_i).$$

.

Cette formule, valable pour les polygones convexes et non convexes, permet de calculer le centroïde de l'ensemble des départements métropolitain. Par la suite, le processus de lissage sera appliqué à chaque centroïde, permettant ainsi de disposer des données relatives aux températures et aux précipitations pour chaque département.

Le lissage a pour vocation de palier au manque de données et par conséquent à disposer des mesures de températures et précipitations sur l'ensemble des points de référence du territoire métropolitain. Dans le cadre de ce mémoire, l'application de la méthode de lissage *Bi-Weight* (SHEPARD (1968)), qui se traduit par une interpolation spatiale, a été effectuée. Cette dernière lisse les données climatiques des stations Météo France, notée s_i, au point de référence C, au moyen de la formule suivante

$$\hat{C} = \frac{\sum_i w_i \cdot s_i}{\sum_i w_i},$$

avec les poids

$$w_i = \begin{cases} \left(1 - \left(\frac{d_i}{D}\right)^2\right)^2 & \text{si } d_i \leq D \\ 0 & \text{si } d_i > D. \end{cases}$$

Dans la formule des poids w_i, d_i correspond à la distance entre le centroïde du département C et la

FIGURE 2.6 : Centroïde du département de la Haute Garonne

station s_i et D est un paramètre du lissage. Cette valeur permet de délimiter une distance au-delà de laquelle les stations employées pour l'interpolation n'exercent plus d'influence sur la valeur estimée. La définition de ce paramètre requiert un compromis délicat : une valeur faible renforce l'ajustement du processus de lissage tout en réduisant le volume de données, tandis qu'une valeur élevée préserve une proportion significative des données mais peut introduire davantage de fluctuations indésirables. Dans le cadre de l'étude, le paramètre D a été obtenu de manière itérative de sorte à minimiser l'erreur d'estimation sur un échantillon de points de référence.

Afin de déterminer la distance d_i entre le centroïde C et une station s_i, la distance géodésique est utilisée car cette dernière tient compte de la courbure de la Terre et permet de considérer la sphéricité du globe terrestre. La formule de la distance géodésique entre deux points x et y peut être exprimée en utilisant la formule de la distance sur une sphère :

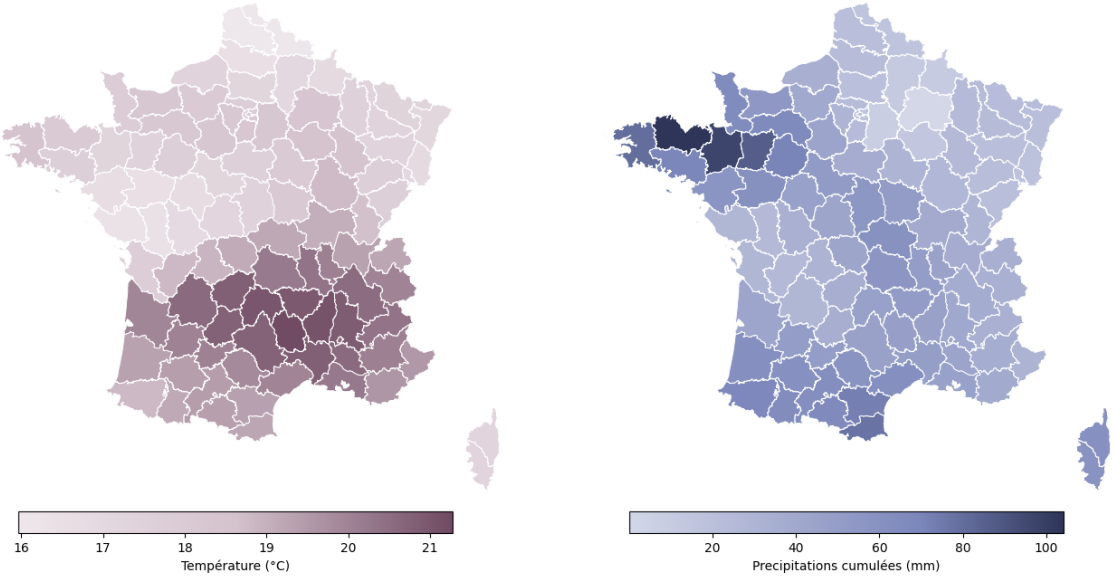
$$d(x, y) = R \cdot \arccos(\sin(\varphi_x) \cdot \sin(\varphi_y) + \cos(\varphi_x) \cdot \cos(\varphi_y) \cdot \cos(\Delta\lambda))$$

où

- d est la distance géodésique entre les deux points x et y,
- R est le rayon de la terre (= 6371 km),
- φ_x et φ_y sont les latitudes des points x et y, respectivement,
- $\Delta\lambda$ est la différence de longitude entre les points x et y.

Ainsi, l'application du lissage *Bi-Weight* permet l'acquisition des données mensuelles relatives aux températures et aux précipitations, couvrant intégralement la période allant de 2000 à 2020 pour chaque département de la France métropolitaine, conformément à l'illustration de la figure 2.7 . Par

ailleurs, en vue de garantir la cohérence du processus de lissage, une représentation graphique des variables lissées a été entreprise.



(a) Températures moyennes lissées - Mai 2011 (b) Précipitations cumulées lissées - Février 2007

FIGURE 2.7 : Exemple de données climatiques mensuelles lissées

Données hydrologiques

Pour le calcul du second indice de sécheresse, les données de l'indice SWI, décrit en partie 1.1.3 ont été récupérées pour la période 2000-2020 sur le site data.gouv (DATA.GOUV (2023)). Ce dernier fournit les données mensuelles de l'indice d'humidité des sols utilisé pour le régime Cat Nat, disponibles à une résolution spatiale de 8km, recouvrant l'ensemble du territoire métropolitain.

Dans le but de réduire la dimensionnalité des données tout en générant une cartographie précise de l'indice SWI par département, une approche consiste à déterminer le parangon propre à chaque département. Le parangon d'un ensemble de points correspond à son point le plus représentatif. De manière plus rigoureuse, il est défini comme le point le plus proche, au sens de la norme euclidienne $\|.\|$, de son barycentre (*cf.* figure 2.8). Ainsi, en notant $(D_i)_{i\in[\![1,96]\!]}$, les quatre-vingt seize départements métropolitains, $(n_i)_{i\in[\![1,96]\!]}$ leurs nombres de points, $(S_{ij})_{(i,j)\in[\![1,96]\!]\times j\in[\![1,n_i]\!]}$ la série des SWI du département D_i pour la maille j, et $(g_i)_{i\in[\![1,96]\!]}$ les isobarycentres de chaque département, le parangon p_i d'un département D_i pour $i \in [\![1, 96]\!]$ peut être obtenu de la façon suivante :

$$p_i = \underset{k\in[\![1,n_i]\!]}{\arg\min} \|S_{ik} - g_i\|$$

Cette méthode permet de réduire considérablement le volume de données, avec un passage de 8981 à 96 points mensuels, tout en exploitant l'intégralité des caractéristiques de chaque département. De plus, Le parangon obtenu pour chaque département sert par la suite de point de référence pour la projection des variables climatiques effectuée en section 3.1. Une fois que toutes les données climatiques requises ont été rassemblées, à savoir les mesures de précipitations et de températures pour l'indice SPEI-3 ainsi que les valeurs mensuelles de SWI pour l'évaluation de l'indice de magnitude SWI, la

section suivante se penche sur la procédure détaillée de calcul des indices de sécheresse utilisés dans cette étude.

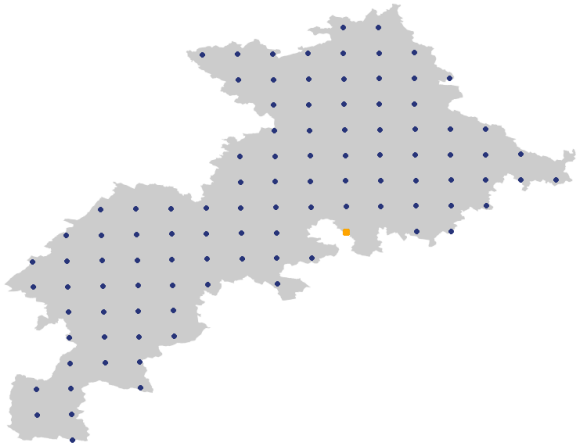


FIGURE 2.8 : Parangon du département de la Haute-Garonne

2.2.2 Indice SPEI-3

Calcul de l'évapotranspiration

Comme décrit en partie 1.1.3, l'indice SPEI intègre non seulement les précipitations, mais aussi les données de températures et de latitude afin de prendre en compte le rayonnement solaire incident et ainsi l'évapotranspiration. Cette dernière peut être approximée par la méthode de THORNTHWAITE (1948) qui offre une estimation simple et robuste sous différentes latitudes. L'évapotranspiration mensuelle, exprimée en millimètres, à la latitude φ au mois j est donnée par

$$\mathrm{ETP}(\varphi, j) = 16 \times K(\varphi, j) \times \left(10 \times \frac{T_j}{I}\right)^m,$$

où T_j est la température moyenne du mois j (en °C) et I est un indice de chaleur, calculé à l'aide de la somme des 12 indices thermiques mensuels

$$I = \sum_i^{12} \left(\frac{T_i}{5}\right)^{1.514}.$$

Le coefficient m dépend directement de l'indice de chaleur I avec la formule $m = 6.75 \times 10^{-7} \times I^3 - 7.71 \times 10^{-5} \times I^2 + 1.79 \times 10^{-2} \times I + 0.492$; et $K(\varphi, j)$ est un coefficient de correction qui dépend de la latitude φ et du mois j

$$K(\varphi, j) = \left(\frac{S(\varphi)}{12}\right)\left(\frac{N_j}{30}\right).$$

Dans l'expression précédente, N_j est le nombre de jours du mois j et $S(\varphi)$ correspond à la durée maximale d'ensoleillement en heure qui est obtenue en utilisant la relation

$$S(\varphi) = (\frac{24}{\pi}) \arccos(-\tan\varphi \tan\delta),$$

avec δ, l'inclinaison solaire en radians estimée par

$$\delta = 0.409 \cdot \sin\left(2\pi \frac{J}{365} - 1.405\right),$$

où J est le jour julien moyen du mois.

Ajustement de la loi log-logistique

Une fois l'ETP calculée, les précipitations nettes d'un département D_i, notée Δ_i, sont obtenues en soustrayant l'évapotranspiration aux précipitations (P_i)

$$\Delta_i = P_i - \text{ETP}_i.$$

La quantité Δ_i offre ainsi une mesure du surplus ou du déficit hydrique du département et de la période considérés.

Pour cet indice, la profondeur temporelle retenue est de 3 mois. Cette base temporelle a notamment été employée lors du projet ClimSec (Soubeyroux et al. (2011)) et permet de définir plusieurs indices saisonniers. Les précipitations nettes sont alors considérées par saison

$$\Delta_{i,s} = P_{i,s} - \text{ETP}_{i,s},$$

avec $s = \{\text{Automne}, \text{Hiver}, \text{Printemps}, \text{Ete}\}$.

Ainsi pour les précipitations nettes estivales $\Delta_{i,\text{Ete}}$, l'évapotranspiration des mois de juin, juillet et aout est soustraite aux précipitations de ces mêmes mois.

Usuellement, les précipitations P_i sont modélisées à l'aide d'une loi $\Gamma(\alpha, \beta)$ lors du calcul de l'indice SPI-3. Cependant, avec la prise en compte de l'évapotranspiration ETP_i, la série considérée Δ_i peut prendre des valeurs négatives, ce qui rend la modélisation par une loi $\Gamma(\alpha, \beta)$ inadaptée. L'article de Vicente-Serrano et al. (2010) propose des distributions candidates comme la loi log-logistique, la loi Pearson III, la loi Log-normale ou encore les lois d'extremum généralisées. Les résultats fournis par l'étude suggèrent pour la modélisation de Δ_i l'utilisation de la loi log-logistique(α, β, γ) dont la densité s'écrit

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1}\left[1 + \left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right]^{-2}$$

où,

- α est le paramètre d'échelle,
- β est le paramètre de forme,
- γ est le paramètre de dispersion.

De ce fait, pour chaque saison et pour chaque département, une loi log-logistique est ajustée sur les données de précipitations nettes historiques sur une profondeur de 40 ans (*cf.* annexe A.2).

Calcul de l'indice

Une fois les paramètres de la loi log-logistique calibrés, l'indice SPEI-3 est finalement obtenu en standardisant les valeurs de la fonction de répartition $F(x)$ de la loi log-logistique selon l'approximation d'Abramowitz et Stegun (1968). En notant $P(x) = 1 - F(x)$ la fonction de survie de la loi log-logistique calibrée sur les données historiques, si $P(x) \leq 0.5$, l'indice SPEI-3, qui prend en argument les précipitations nettes, est donné par

$$\text{SPEI-3}(x) = W(x) - \frac{c_0 + c_1 W(x) + c_2 W(x)^2}{1 + d_1 W(x) + d_2 W(x)^2 + d_3 W(x)^3}$$

avec $W(x) = \sqrt{-2\ln(P(x))}$. En revanche, si $P(x) > 0.5$, le calcul devient

$$\text{SPEI-3}(x) = -W(x) + \frac{c_0 + c_1 W(x) + c_2 W(x)^2}{1 + d_1 W(x) + d_2 W(x)^2 + d_3 W(x)^3}$$

avec $W(x) = \sqrt{-2\ln(1 - P(x))}$.

Les constantes utilisées dans les formules précédentes sont

- $c_0 = 2.515517$,$c_1 = 0.802583$, $c_2 = 0.010328$,
- $d_1 = 1.432788$, $d_2 = 0.189269$, $d_3 = 0.001308$.

Analyse spatio-temporelle de l'indice

Afin d'évaluer la validité de l'indice élaboré antérieurement, une analyse initiale du coefficient de corrélation de Kendall est entreprise. Cette analyse vise à établir des liens potentiels entre les divers indices saisonniers SPEI-3 et la charge sinistre annuelle par département. Par construction, le SPEI peut être interprété comme une mesure de l'écart par rapport aux données historiques des précipitations nettes. Par conséquent, lorsque les précipitations nettes sont inférieures aux valeurs historiques, l'indice SPEI prend des valeurs négatives. Dans notre contexte, étant donné que des précipitations nettes faibles traduisent un état de sécheresse pour la zone d'intérêt, une dépendance négative entre les valeurs de l'indice SPEI et le montant de la charge sinistre annuelle doit être observée.

	SPEI Hivernal	SPEI Printanier	SPEI Estival	SPEI Automnale
Charge sinistre	-0.21	- 0.29	- 0.34	- 0.19

Table 2.1 : τ de Kendall entre la charge sinistre annuelle par département et les indices saisonniers.

L'analyse du tau de Kendall pour chaque indice saisonnier, synthétisée dans le tableau 2.1, confirme la présence d'une corrélation négative entre les valeurs de l'indice SPEI et la charge sinistre annuelle. Les indices SPEI estival et hivernal semblent fournir une explication plus significative des événements de sécheresse, avec des coefficients de Kendall respectifs de -0.34 et -0.21. Cette corrélation est également illustrée visuellement dans la figure 2.9, où l'on observe une tendance croissante des montants de sinistres à mesure que l'indice SPEI estival s'éloigne de zéro.

Du point de vue temporel, l'indice SPEI-3 calculé reflète en grande partie les tendances observées en matière de sinistralité. La figure 2.10 révèle que l'indice tend à se rapprocher de zéro lors des années caractérisées par une faible sinistralité, tandis qu'il présente une nette diminution lors des épisodes de sécheresse majeurs tels que ceux de 2003, 2016 et 2018. Cependant, certaines années font exception

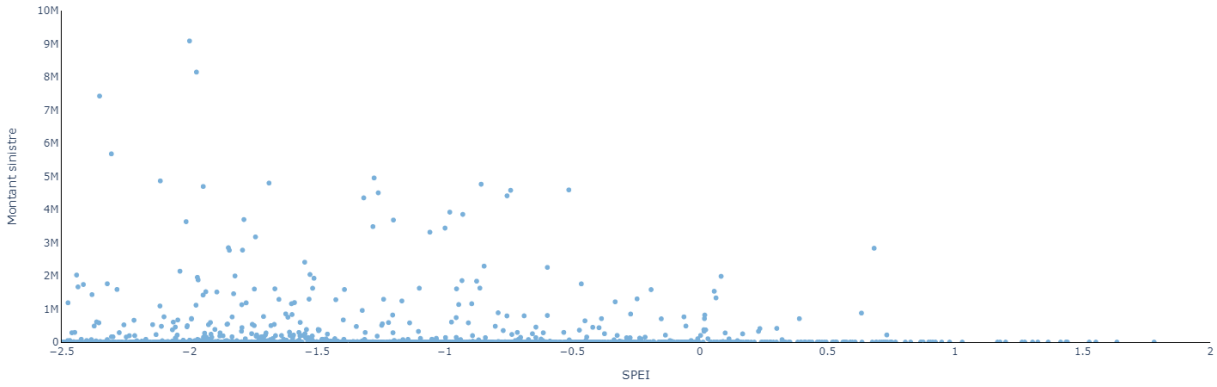


FIGURE 2.9 : Nuage de points entre la valeur du SPEI estival et le montant de la charge sinistre

à cette corrélation, comme c'est le cas pour 2011 et 2017, où l'indice SPEI-3 ne semble pas déceler d'anomalie dans les valeurs de précipitations nettes.

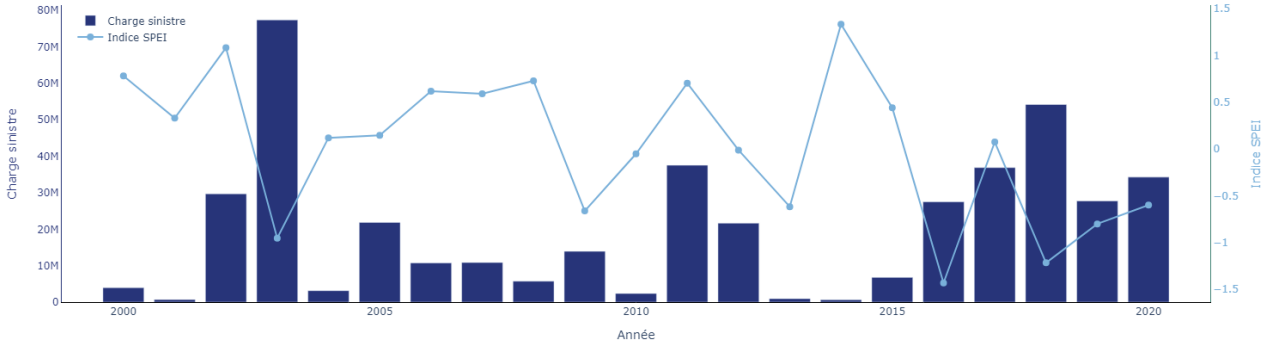
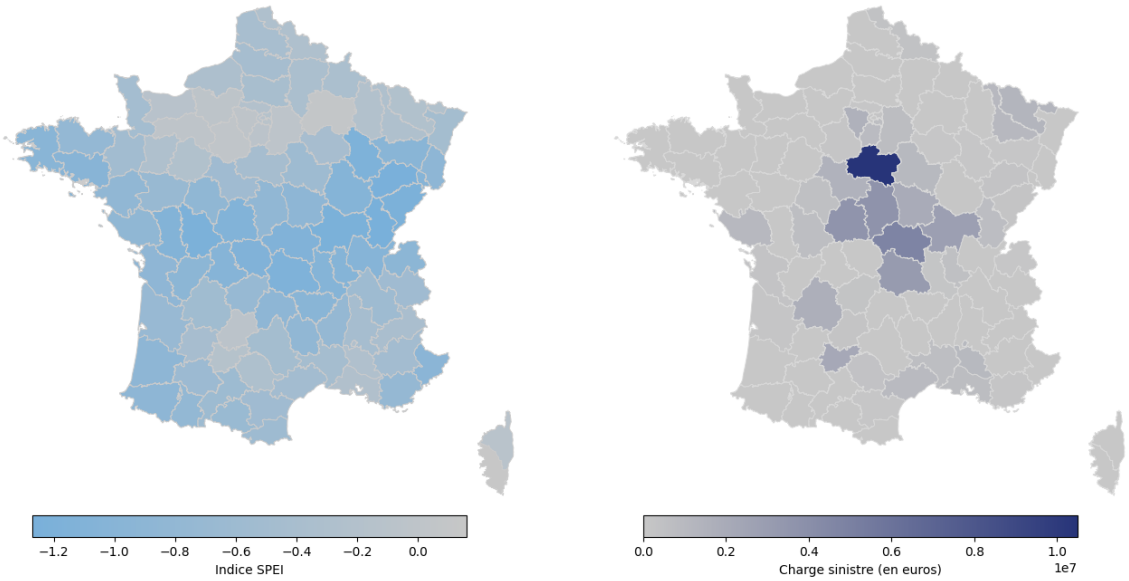


FIGURE 2.10 : Comparaison temporelle de la charge sinistre cumulée annuelle et de l'indice SPEI-3

Sur le plan spatial, même si l'indice ne se superpose pas parfaitement avec la répartition de la charge sinistre, il exhibe généralement des valeurs faibles dans les zones sinistrées. La comparaison entre la configuration géographique de l'indice SPEI estival et la répartition des sinistres pour l'année 2018, illustrée par la figure 2.11, est instructive. Cette année a enregistré une sinistralité atypique, principalement concentrée dans les départements du Loiret, du Cher et de l'Allier. L'indice SPEI parvient pour cette même année à traduire de manière cohérente les épisodes exceptionnels de précipitations et d'évapotranspiration dans ces mêmes régions.

Il convient de noter que cet indice, bien qu'informatif, ne se suffit pas à lui-même pour expliquer la prévalence des sinistres liés à la sécheresse. Il reflète les conditions climatiques caractérisées par un déficit significatif de précipitations dans une zone spécifique. Toutefois, une faible valeur de l'indice SPEI-3 n'implique pas automatiquement la survenue de sinistres, car la sinistralité dépend également de l'exposition de la zone en question. Par conséquent, une analyse exhaustive de la sinistralité du



(a) Indice SPEI estival - Année 2018 (b) Charge sinistre par département - Année 2018

FIGURE 2.11 : Comparaison spatiale de la charge sinistre avec l'indice SPEI estival - Année 2018

portefeuille nécessite la combinaison de cet indice avec les facteurs d'exposition présentée en partie 2.1.2.

2.2.3 Indice de magnitude SWI

Dans cette section, un autre indice est construit à partir de l'indice d'humidité des sols SWI fourni par Météo France présenté en partie 1.1.3. Ce nouvel indice de sécheresse s'inspire des travaux réalisés dans le cadre du rapport scientifique de la CCR (BARTHELEMY ET AL. (2022)). L'objectif de cette approche est d'utiliser les données d'humidité du sol pour extraire une mesure annuelle de l'intensité de la sécheresse. Cette mesure est obtenue en évaluant l'intégrale annuelle de l'indice d'humidité du sol en dessous d'un seuil préalablement déterminé.

Pour ce faire, les valeurs des $\overline{\text{SWI}}^{C}_{m,N}$ sont extraites pour chaque mois m de l'année N. Ces valeurs suivent un schéma périodique, caractérisé par des SWI faibles pendant les mois printaniers et estivaux, et des SWI élevés pendant les mois automnaux et hivernaux. En se basant sur ces données, il devient possible de modéliser une courbe sinusoïdale ajustée, comme illustré à travers la figure 2.12, de la forme suivante

$$f(x) = A \cdot \sin\left(\frac{2\pi}{T} \cdot x - c\right) + d$$

avec,

- A, l'amplitude,
- T, la période, qui vaut 12 ici,
- c, le *shift* horizontal,
- d, le *shift* vertical.

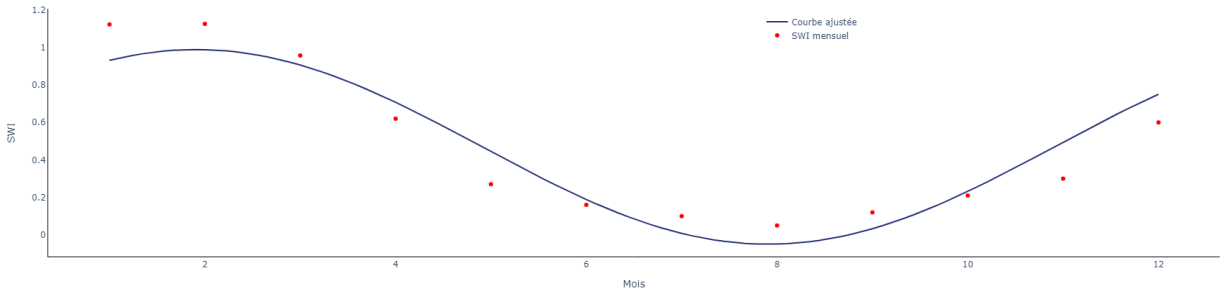


FIGURE 2.12 : Exemple d'ajustement de la courbe sinusoïdale aux données SWI mensuelle pour le département du Cher.

L'ensemble des paramètres de la courbe sinusoïdale est noté $\theta = \{A, c, d\}$. Les paramètres optimaux θ^* sont obtenus en résolvant le problème de minimisation suivant

$$\min_{\theta} \sum_{m=1}^{12} \left(\overline{\text{SWI}}^{C}_{m,N} - f_\theta(x_m) \right)^2.$$

Une fois la courbe ajustée, un seuil γ est fixé afin de calculer l'intégrale de la courbe obtenue sous cette valeur, comme illustré sur la figure 2.13.. La valeur *a priori* retenue pour ce seuil correspond au 10-ème percentile des SWI. Ceci permet de définir le nouvel indice de magnitude SWI

$$\text{SWI}(\gamma) = \int_0^{12} \max(0, \gamma - f_{\theta^*}(x))\, dx.$$

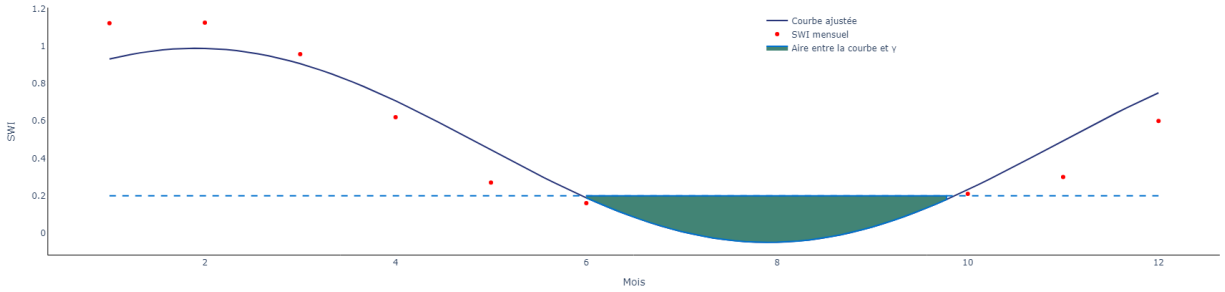


FIGURE 2.13 : Calcul de l'indice de magnitude SWI(γ).

Optimisation du seuil

Par la suite, une démarche d'optimisation du seuil γ a été effectuée. Dans la perspective de construire un indice capable de faire le lien entre les conditions hydrologiques d'une zone et la survenance des sinistres relatifs à la sécheresse, la valeur optimale γ^* a été déterminée en comparant l'indice ainsi obtenu avec la charge sinistre historique du portefeuille, de la manière suivante

$$\gamma^* = \arg\max_{\gamma} \tau(\mathrm{SWI}(\gamma), Y)$$

où

- τ désigne le tau de Kendall,
- Y la charge sinistre annuelle par département liée au phénomène de sécheresse géotechnique.

L'application de cette approche de maximisation a conduit à une valeur optimale de 0.2 pour γ^*.

Analyse spatio-temporelle de l'indice

De manière analogue au premier indice construit, il est impératif d'évaluer la pertinence de l'indice de magnitude SWI en le comparant à la variable d'intérêt à modéliser. À l'opposé du SPEI-3, l'indice de magnitude SWI présente une valeur plus élevée lorsque l'humidité du sol atteint des niveaux anormalement bas, induisant ainsi des conditions de sécheresse pour la région concernée. Par conséquent, une corrélation positive entre la valeur de cet indice et la charge sinistre est à anticiper.

Le tau de kendall entre la valeur de SWI(γ) et la charge sinistre sécheresse s'établit à 0.57. Ceci témoigne de la dépendance positive entre ces deux variables et valide le lien significatif qui réside entre les caractéristiques hydrologiques d'un département et l'apparition des sinistres sécheresse.

Par la suite, une analyse comparative à la fois spatiale et temporelle est faite entre la valeur de l'indice SWI(γ) et la charge sinistre historique annuelle liée à la sécheresse.

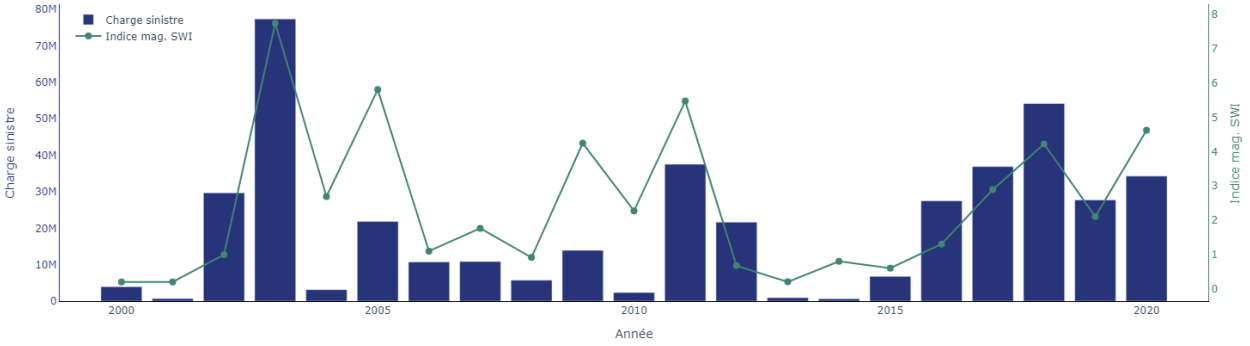
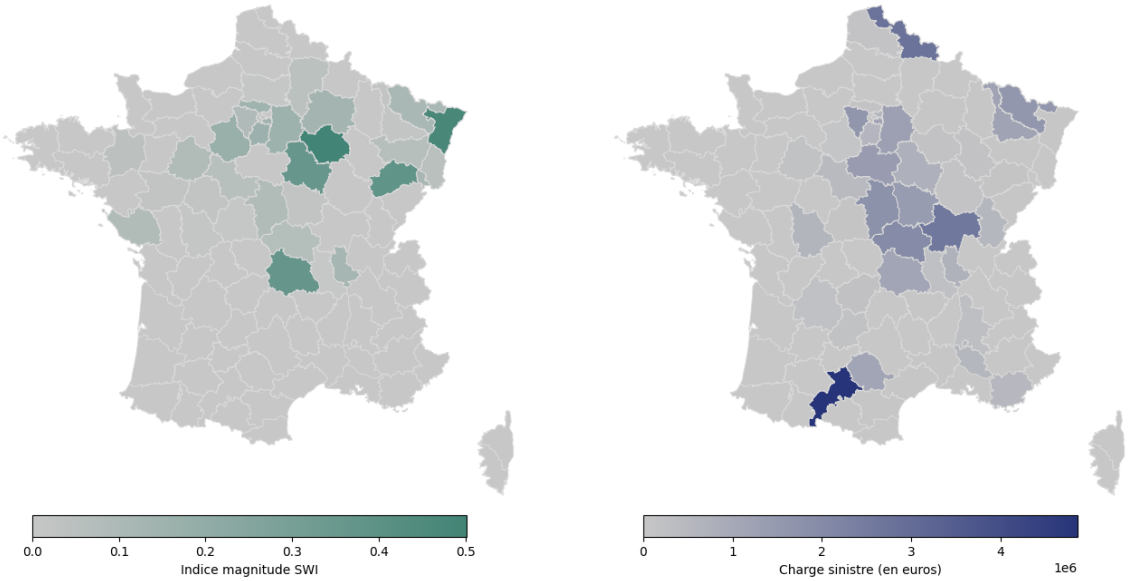


FIGURE 2.14 : Comparaison temporelle de la charge sinistre cumulée annuelle et de l'indice SWI(γ)

L'examen temporel, à travers la figure 2.14, révèle la capacité de l'indice de magnitude SWI à reproduire les pics de sinistralité. Les années exceptionnelles telles que 2003, 2011, 2018 et 2017 se caractérisent par des valeurs élevées de cet indice. En plus de sa reproduction satisfaisante des épisodes de sécheresse extrêmes, l'indice élaboré suit les tendances observées en matière de sinistralité. Plus précisément, il présente des valeurs modérées durant les années épargnées par la sécheresse et une tendance à la hausse entre les années 2015 et 2020.

Ensuite, d'un point de vue spatial, même si l'indice ne parvient pas à reproduire annuellement l'entièreté de la répartition de la charge sinistre, il réussit néanmoins à détecter les régions affectées

par les sinistres, comme le met en évidence la figure 2.15. Cette représentation confronte la distribution de la charge sinistre et les valeurs de l'indice de magnitude SWI pour l'année 2020. Elle révèle que l'indice présente des valeurs élevées pour les départements touchés en Île-de-France ainsi que pour les départements du Cher et de l'Allier, mais montre des valeurs relativement faibles pour les départements du Nord et de la Haute-Garonne, bien que ceux-ci aient également été touchés cette année-là.



(a) Indice de magnitude SWI(γ) - Année 2020 (b) Charge sinistre par département - Année 2020

FIGURE 2.15 : Comparaison spatiale de la charge sinistre avec l'indice de magnitude SWI(γ) - Année 2020

De manière similaire à l'indice SPEI-3, l'indice de magnitude SWI ne permet pas à lui seul d'identifier de manière exhaustive les sinistres sécheresse. Toutefois, sa forte corrélation avec l'occurrence des sinistres, conjuguée aux données d'exposition, en fait une variable d'intérêt pour le modèle prédictif.

2.3 Cadre théorique de modélisation

2.3.1 Formalisme général

L'objectif du modèle employé lors de l'étude est de répondre à une problématique d'apprentissage supervisé. Un échantillon de taille n est considéré et pour chaque observation $i \leq n$ de celui-ci, un vecteur aléatoire $X_i \in \mathcal{X} \subset \mathbb{R}^p$, contenant des variables explicatives, est associé à une variable aléatoire $Y_i \in \mathcal{Y} \subset \mathbb{R}$, désignée comme la variable cible. Dans le cadre du mémoire, la variable cible que le modèle cherche à prédire est la charge sinistre annuelle par département liée aux événements de sécheresse géotechnique. Par conséquent, la variable Y_i prend ses valeurs dans $\mathcal{Y} = \mathbb{R}_+$, l'ensemble des nombres réels positifs. Il est également supposé que la suite $(X_i, Y_i)_{i\in[1,n]}$ est indépendante et identiquement distribuée ($i.i.d$) et par souci de notation, le couple associé à cette distribution est désigné par (X, Y). Par la suite, l'objectif du modèle étudié consiste à approcher la fonction F telle que

$$\mathbb{E}[Y|X] = F(X).$$

Dans la pratique, les valeurs des variables et des vecteurs aléatoires sont disponibles pour chaque observation $i \leq n$. Un vecteur de taille n est ainsi formé, noté $\mathbf{Y}$, accompagné d'une matrice de

dimensions n × p, notée $\mathbf{X}$. Dans le reste du mémoire, les valeurs du vecteur $\mathbf{Y}$ sont notées y_i et les lignes et colonnes de la matrice $\mathbf{X}$ sont respectivement notées x_i et x^j, avec $1 \leq j \leq p$. A partir des ces données, l'objectif est de construire une fonction $\hat{F}$ telle que, pour un vecteur donné x_i

$$\hat{F}(x_i) \approx F(x_i) = \mathbb{E}\left[Y|X = x_i\right].$$

Le modèle d'apprentissage supervisé utilisé au cours de cette étude est basé sur les techniques de *Boosting*, dont les fondements théoriques sont exposés en détail dans la section 2.3.3. La spécificité des modèles d'apprentissage supervisé réside dans leur absence d'hypothèse concernant la loi de $\mathbb{E}\left[Y|X\right]$ ainsi que sur la forme de la fonction F. De ce fait, ces derniers font appel à une fonction de perte l et une fonction de prédiction $f \in \mathcal{F}\left(\mathcal{X}, \mathcal{Y}\right)$ qui permet de quantifier l'erreur de prédiction avec Y. Dans cette perspective, le risque du modèle, noté $\mathcal{R}(f)$, est introduit et constitue ainsi le principal critère d'évaluation pour ces modèles

$$\mathcal{R}(f) = \mathbb{E}\left[l(Y, f(X))\right].$$

Le meilleur modèle est obtenu en cherchant la fonction de prédiction f qui minimise le risque, c'est à dire

$$f^* = \underset{f \in \mathcal{F}(\mathcal{X},\mathcal{Y})}{\operatorname{argmin}} \mathcal{R}(f).$$

Cependant, la recherche d'un modèle optimal peut se révéler complexe puisqu'il n'existe aucune hypothèse établie quant à la distribution de $l(Y, f(X))$, et il n'y a aucune garantie concernant l'existence d'une telle fonction. Par conséquent, le risque empirique est considéré en pratique

$$\hat{\mathcal{R}}_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \hat{f}(x_i)),$$

où y_i est une réalisation de la variable aléatoire Y.

2.3.2 Critères d'évaluation

Au cours de l'élaboration d'un modèle prédictif, la sélection des méthodes de validation ainsi que des critères pour son évaluation représente un élément essentiel. Dans le contexte de données spatio-temporelles, tel que le cas présent, des approches de validation croisée spécifiques sont adoptées en vue de maintenir une cohérence avec les objectifs de l'étude

Validation croisée spatiale et temporelle

La première étape nécessaire pour évaluer l'efficacité d'un modèle implique la séparation des données en deux ensembles distincts : une base d'entraînement et une base de test. Cette approche est notamment conçue pour mettre en évidence le phénomène de sur-apprentissage, caractérisé par un ajustement excessif du modèle aux données d'entraînement et, en conséquence, une incapacité du modèle à généraliser. De manière concrète, cette démarche facilite la comparaison des performances du modèle sur la base d'entraînement (où le modèle apprend) avec celles obtenues sur la base de test, qui font office de nouvelles données.

En vue d'améliorer la fiabilité de la validation et de prévenir toute occurrence de sur-apprentissage, il est fréquent de segmenter les données en K sous-ensembles distincts (appelés *"folds"*) et de procéder à K cycles d'entraînement du modèle. Pendant chaque itération, l'un des sous-ensembles est réservé pour la validation tandis que les autres données sont utilisées pour l'apprentissage du modèle. Cette approche correspond au concept de la validation croisée à K sous-ensembles.

Toutefois, en raison de la nature aléatoire inhérente au principe de séparation, cette approche peut montrer certaines limites et faire apparaître un biais dans les cas où les données comportent des dimensions temporelles et spatiales. Pour atténuer les dépendances spatiales et temporelles entre les ensembles d'entraînement et de validation, une validation croisée à la fois temporelle et spatiale est mise en œuvre.

La validation croisée temporelle se réalise en excluant les données futures de l'ensemble d'entraînement. Afin d'évaluer les performances du modèle pour l'année n, seules les observations jusqu'à l'année $n-1$ sont utilisées pour l'entraînement. Ceci permet notamment d'empêcher, que des données d'un département soient employées pour prédire la charge sinistre d'un autre pour une année donnée. Parallèlement, cette procédure vise à garantir que le modèle soit capable de généraliser d'une année à l'autre.

La validation croisée spatiale (Pohjankukka et al. (2017)) consiste quant à elle à retirer un sous échantillon de plusieurs départements de la base d'entraînement. Comme illustré dans la figure 2.16, pour valider les performances du modèle sur un département spécifique, seuls les départements qui ne sont pas limitrophes au département ciblé sont inclus dans la phase d'apprentissage. Cette stratégie vise à restreindre les corrélations spatiales, propres aux données géographiques, et à éviter l'utilisation de données provenant d'un département présentant des caractéristiques similaires à celles du département d'intérêt pour l'estimation de la charge sinistre.

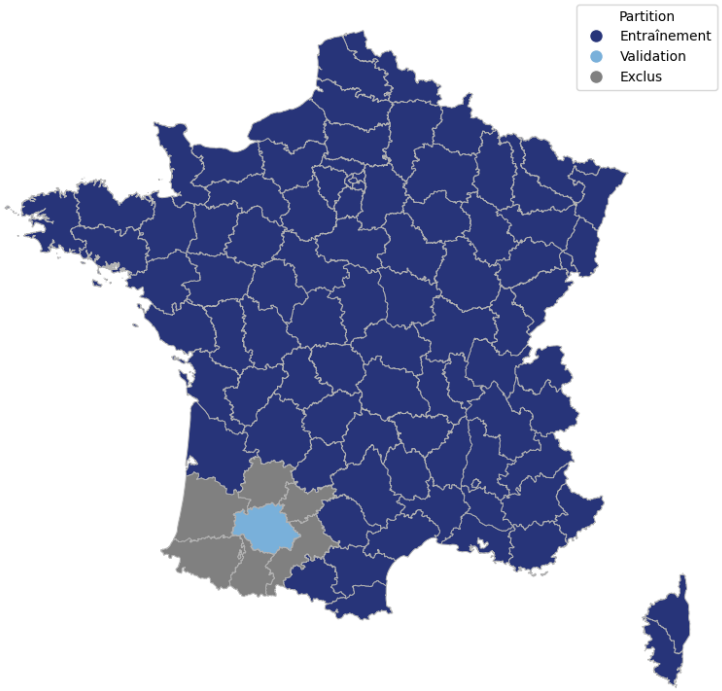


Figure 2.16 : Exemple de validation croisée spatiale pour le département du Gers.

Pour cette étude, les deux approches sont employées et le modèle retenu est celui qui obtient les meilleurs résultats sur ces deux méthodes de validation. Il est à noter que les deux approches peuvent être utilisées simultanément mais cela présente l'inconvénient de réduire considérablement les données d'entraînement et par conséquent d'introduire un autre biais dans les prédictions. L'étude présentée ci dessous traite donc ces deux méthodes séparément.

Pour l'optimisation des hyperparamètres du modèle détaillée en partie 2.4.2, seule la validation croisée temporelle est utilisée.

Métriques en lien avec la problématique

Lors des procédures de validation présentées précédemment, un score est généralement calculé sur les données de validation au terme de chaque entraînement basé sur une métrique de performance. Ces métriques permettent de mesurer l'efficacité du modèle sur des données nouvelles et sont à distinguer des fonctions de pertes utilisées pour l'entraînement du modèle.

Dans le cas d'une régression, les métriques de validation couramment employées sont issues des fonctions de pertes suivantes

- La perte brute $Y - f(X)$, dont l'inconvénient majeur réside dans la possibilité que le risque associé puisse demeurer négligeable sans pour autant garantir les performances prédictives du modèle en question.
- La perte absolue $|Y - f(X)|$ qui mesure l'écart de prédiction sans distinguer la sous-estimation de la surestimation.
- La perte quadratique $(Y - f(X))^2$ qui pénalise plus fortement les grands écarts de prédiction que les petits.

Les propriétés inhérentes à la perte absolue et à la perte quadratique sont telles que le risque associé est nul si et seulement si $Y = f(X)$ presque sûrement. Cette caractéristique ne se retrouve cependant pas pour la perte brute.

En s'appuyant sur ces fonctions de perte, il est possible de constituer des mesures de risques empiriques. Tout d'abord, la perte quadratique est associée à l'erreur quadratique moyenne (MSE), définie par la formule

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

La MSE, qui calcule la moyenne des carrés des erreurs offre l'avantage d'une sensibilité accrue envers les écarts significatifs. Par conséquent, elle s'avère pertinente pour pénaliser les erreurs de grande ampleur. De plus, la nature différentiable de la MSE permet son utilisation lors de la phase d'entraînement des algorithmes reposant sur des méthodes de gradient. En conséquence, dans le contexte de l'étude, étant donné que la variable cible présente une variance importante et peut atteindre des valeurs élevées, le choix de la MSE comme fonction d'optimisation pour l'entraînement du modèle est justifié. Pour l'évaluation du modèle post-entrainement, d'autres métriques présentées ci-dessous, jouissant d'une meilleure explicabilité, sont employées.

A partir de la MSE, il est possible de définir le coefficient de détermination linéaire de Pearson, noté R^2, qui s'écrit

$$R^2 = 1 - \frac{\text{MSE}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

avec $\bar{y}$ la moyenne des observations y_i. Cette métrique permet de confronter l'erreur du modèle avec celle d'un modèle de base qui prédit systématiquement la moyenne de la variable cible. Le R^2 est d'autant plus élevé que le modèle est performant, atteignant l'optimalité avec une valeur de 1 lorsqu'il

réalise des prédictions exactes. Si R^2 prend une valeur négative, cela indique que le modèle se révèle moins performant que le modèle simple qui prédit la moyenne de la variable à prédire. Cependant, cette métrique présente le désavantage de ne pas fournir d'indications concernant l'erreur moyenne du modèle et de croître avec le nombre de variables explicatives

La perte absolue est quant à elle reliée à l'erreur absolue moyenne (MAE), qui s'écrit

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{f}(x_i)|.$$

Contrairement à la MSE, la MAE s'exprime dans la même unité que la variable à prédire, ce qui la rend plus interprétable. Cependant, cette métrique n'amplifie pas les écarts entre les erreurs faibles et les erreurs fortes. C'est pourquoi elle est complétée dans l'étude par une autre métrique.

Dans le but de quantifier la capacité du modèle à discerner les occurrences exceptionnelles de sécheresse, il peut être judicieux de mettre en place une mesure d'erreur pour les prédictions impliquant une charge sinistre observée qui présente un caractère extrême. En conséquence, pour chaque échantillon de validation V, un sous-ensemble est défini comme $\mathcal{V}(q) = \{(x, y) \in V \mid y \geq q\}$, où q équivaut au 80 centile de la distribution des charges sinistres historiques du portefeuille. Ceci permet de définir la nouvelle métrique MAPEX(q) (*Mean Absolute Percentage Error for eXtreme values*)

$$\text{MAPEX}(q) = \frac{100}{|\mathcal{V}(q)|}\sum_{(x,y)\in\mathcal{V}(q)} \frac{|y - \hat{f}(x)|}{y},$$

qui évalue l'exactitude des prédictions relatives à ces valeurs exceptionnelles. Cette métrique, qui s'exprime comme le pourcentage moyen d'erreur entre les observations et les valeurs prédites, a l'avantage de considérer l'erreur du modèle en proportion de la valeur prédite. Ceci permet de réduire l'effet des erreurs sur les valeurs aberrantes et de conduire à une mesure facilement interprétable.

Conformément à ce qui a été énoncé précédemment, le modèle est soumis à deux validations, une temporelle et une spatiale. Ainsi, toutes les métriques détaillées ci-dessous, à l'exception de la MSE employée pour la phase d'apprentissage du modèle, sont déclinées à travers les deux méthodes de validations, constituant donc deux types de mesure de risques. La première, notée $\mathcal{T}(.)$, permet de mesurer les performances prédictives temporelles du modèle et est définie, pour une mesure de risque empirique $\hat{\mathcal{R}}(f)$, par

$$\mathcal{T}(\hat{\mathcal{R}}(f)) = \frac{1}{|T|}\sum_{t\in T} \hat{\mathcal{R}}_t(f)$$

où $|.|$ désigne le cardinal, T l'ensemble des subdivisions de la validation croisée temporelle et $\hat{\mathcal{R}}_t(f)$ le risque obtenue pour la subdivision t.

La deuxième, notée $\mathcal{S}(.)$, quantifie quant à elle la capacité du modèle à généraliser d'un département à l'autre et s'écrit

$$\mathcal{S}(\hat{\mathcal{R}}(f)) = \frac{1}{|S|}\sum_{s\in S} \hat{\mathcal{R}}_s(f)$$

avec S correspondant à l'ensemble des partitionnement de la validation croisée spatiale et $\hat{\mathcal{R}}_s(f)$ le risque calculé sur le partitionnement s.

2.3.3 Théorie du *Gradient Boosting*

L'algorithme d'apprentissage supervisé utilisé pour l'étude est l'algorithme *Catboost* (Prokhorenkova et al. (2018)), qui fait partie de la famille des méthodes de *Gradient Tree Boosting* dont la structure utilise des arbres de décisions. Cette partie se concentre sur le fonctionnement théorique et les spécificités de ces algorithmes.

Arbre de décision

Un arbre de décision (Morgan et Sonquist (1963)) est un modèle d'apprentissage supervisé introduit dans les années 60 et dont la fonction de prédiction est de la forme

$$f(x) = \sum_{j=1}^{t} \alpha_j I_j(x).$$

Dans l'expression précédente, les I_j correspondent à des indicatrices sur des "hyper-rectangles" (produit cartésien d'intervalles) et les α_j à des constantes associées. En notant R_j ces régions, la fonction de prédiction est donnée par :

$$f(x) = \sum_{j=1}^{t} \alpha_j \mathbb{1}_{x \in R_j}.$$

La notion d'arbre provient à la fois de la méthode de construction du modèle et de sa représentation. En effet, lorsque l'espace des variables est découpé selon des critères de décision, la représentation arborescente est adaptée. Chaque sommet consiste en une condition qui s'accumule aux précédentes, les feuilles étant l'intersection de toutes ces conditions et correspondent aux régions R_j.

Le concept d'arbre fait référence à la fois à la méthodologie de construction du modèle et à sa configuration graphique. En effet, lors de la segmentation de l'espace des variables selon des critères de décision, la représentation arborescente, telle qu'illustrée dans la figure 2.17, s'avère appropriée. Chaque nœud interne se compose d'une condition qui s'ajoute aux précédentes, les feuilles étant l'intersection de l'ensemble de ces conditions et correspondent aux régions R_j, comme le montre la figure 2.18.

La méthode la plus largement adoptée pour construire de tels arbres est la méthode CART (*Classification and Regression Trees*) (Breiman et al. (2017)). Le principe central des arbres CART consiste à diviser de manière récursive et binaire l'espace des variables explicatives, en vue de couvrir toutes les valeurs possibles de la variable à prédire. Ce processus se déroule en couches successives, où à chaque itération, pour chaque nœud de la couche courante, l'espace est divisé en deux régions. Cette séparation s'opère en utilisant une variable x^j et une valeur seuil a spécifiques comme critères de division. Il est à noter que lors de cette procédure de partitionnement, l'utilisation répétée d'une même variable n'est pas exclue, pouvant ainsi être employée sur différentes couches de l'arbre.

La méthode CART opère ainsi en sélectionnant, à chaque itération et à chaque nœud, la variable x^j ainsi que le seuil a qui réalise la meilleure segmentation de l'espace des variables en deux sous-espaces. Ce choix s'effectue selon la procédure suivante : l'hyper-rectangle associé au nœud où la division est réalisée, est noté R, et pour $a \in \mathbb{R}$, les ensembles $R_1(j,a) = \{x \in \mathbb{R}^p | x^j \leq a\} \cap R$ et $R_2(j,a) = \{x \in \mathbb{R}^p | x^j > a\} \cap R$ sont introduits. Lorsque la variable x^j est qualitative, il est possible de choisir a comme tout sous-ensemble non vide de l'ensemble des valeurs possibles de x^j, et ainsi définir $R_1(j,a)$ et $R_2(j,a)$ par

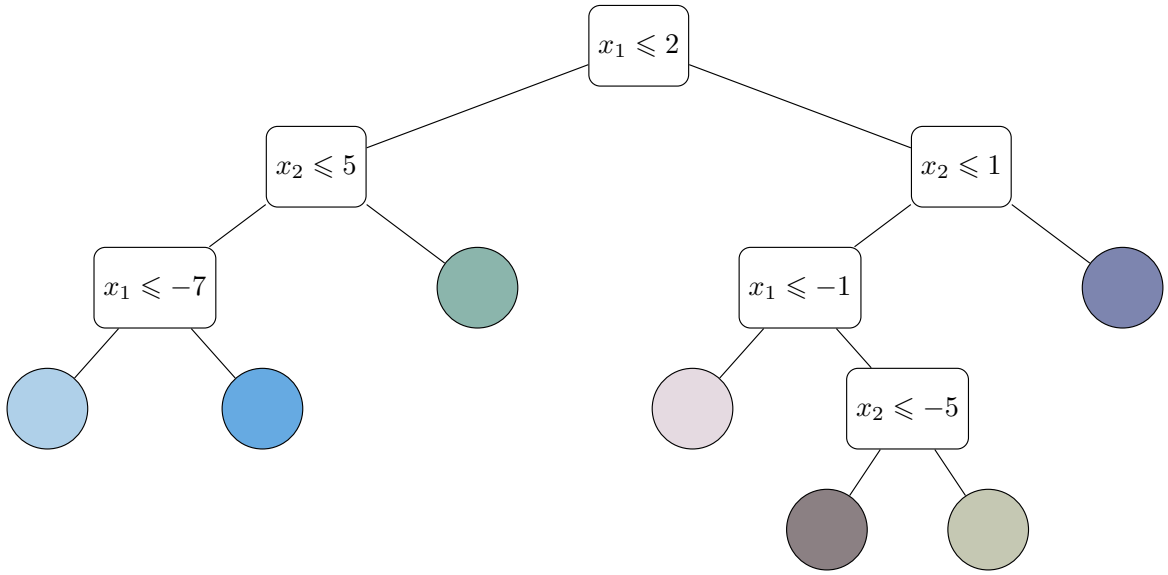


FIGURE 2.17 : Exemple d'arbre de décision

$$R_1(j,a) = \{x \in \mathbb{R}^p | x^j \in a\} \quad \text{et} \quad R_1(j,a) = \{x \in \mathbb{R}^p | x^j \notin a\}.$$

Le choix de l'indice j (de la variable x^j) et du seuil a est donné par la résolution du problème suivant

$$\min_{j,a} \left[\min_{\hat{f}_1} \sum_{x_i \in R_1(j,a)} l(y_i, \hat{f}_1) + \min_{\hat{f}_2} \sum_{x_i \in R_2(j,a)} l(y_i, \hat{f}_2) \right]$$

avec $\hat{f}_1$ et $\hat{f}_2$, les prédictions faites pour les données appartenant respectivement à $R_1(j,a)$ et $R_2(j,a)$ et l une fonction de perte adaptée au problème considéré. En effet, dans un cas de régression, l'erreur quadratique peut être employée.

Au sein de cette classe d'algorithmes, il est envisageable d'interrompre la phase de partitionnement durant la construction de l'arbre en régulant sa dimension. Dans le processus de modélisation, le choix de la taille de l'arbre revêt une importance capitale : un arbre trop grand peut engendrer un phénomène de sur-apprentissage, tandis qu'un arbre trop petit peut négliger des aspects essentiels des données, menant à du sous-apprentissage.

Approche générale des algorithmes de *boosting*

Les méthodes de *boosting* sont des techniques d'apprentissage ensembliste destinées à résoudre des problèmes de classification ou de régression. L'idée fondamentale qui réunit ces algorithmes est d'optimiser les performances d'un ensemble de modèles de prédiction qualifiés de "faibles", en les combinant entre eux afin de créer un modèle dit "fort", aux performances prédictives supérieures. Un modèle de prédiction "faible" désigne une méthode de classification ou de régression légèrement plus performante qu'un tirage aléatoire. Au sein de ce type de méthode, il est impératif que les modèles faibles ne présentent pas une efficacité trop importante pour écarter tout risque de sur-apprentissage. La forme de la fonction de prédiction globale du modèle peut être exprimée de la manière suivante

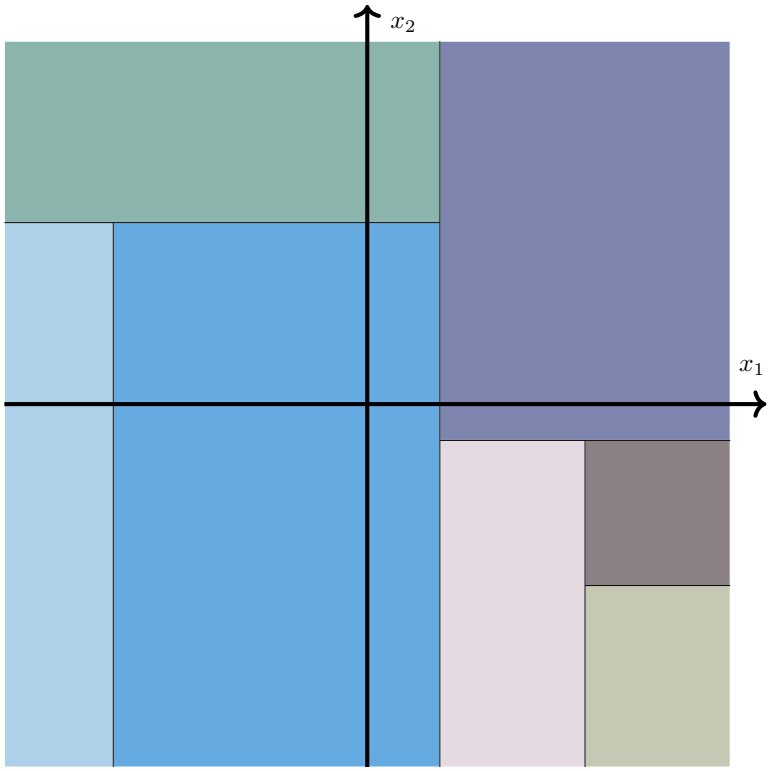


FIGURE 2.18 : Région correspondant à l'arbre de la figure 2.17

$$F_M(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

avec h_m la fonction de prédiction du modèle faible m et β_m le coefficient qui lui est associé.

Le mécanisme sous-jacent au *boosting* consiste en la construction récursive de modèles faibles : une fois entraîné, les sorties du modèle h_1 sont utilisées en entrée du modèle h_2, et ainsi de suite. La démarche vise à ce que, à chaque itération, les prédictions erronées se voient attribuées une pondération supérieure, afin que le modèle faible suivant puisse y accorder davantage d'attention. Ce mode de prédiction se distingue particulièrement des méthodes de *bagging*, comme l'algorithme *Random Forest* (BREIMAN (2001)), où les modèles faibles sont employés séparément sur des subdivisions des données avant que leurs résultats ne soient agrégés.

Pour l'étude, le modèle de *boosting* considéré repose sur les arbres de décision (*Tree Boosting*), qui représentent généralement le modèle de prédiction faible choisi. Dans ce contexte, la fonction de prédiction du modèle faible est formulée comme suit

$$h_m(x) = \sum_{j=1}^{t_m} \alpha_{m,j} \mathbf{1}_{x \in R_{m,j}}$$

et la minimisation du risque empirique d'un tel arbre peut être formulée ainsi

$$\min_{h_m \in \mathcal{H}} \quad \frac{1}{n} \sum_{i=1}^{n} l(y_i, h_m(x_i)).$$

En utilisant le fait que la fonction de prédiction finale s'écrit

$$F_M(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

et en supposant que $m-1$ arbres ont déjà été construits avec F_{m-1} la fonction de prédiction obtenue, il vient, que le m-ième arbre est obtenu par la résolution du problème suivant

$$\min_{h_m \in \mathcal{H}} \quad \frac{1}{n} \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h_m(x_i)). \tag{2.1}$$

Une fois que le problème de minimisation est formulé, le concept des algorithmes *gradient boosting* (Friedman (2001)) peut maintenant être introduit.

Gradient Tree Boosting

Pour résoudre le problème de minimisation précédemment énoncé, les algorithmes de boosting exploitent le principe de la descente de gradient. Il est pertinent de rappeler que, étant donné une fonction $f : \mathbb{R}^d \mapsto \mathbb{R}$ différentiable où $x, h \in \mathbb{R}^d$, la relation suivante s'établie

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).$$

L'objectif à la descente de gradient réside dans la recherche du minimum de la fonction f par l'annulation de son gradient, une condition qui se révèle suffisante lorsque f est strictement convexe. À cet effet, un coefficient $\eta > 0$ est défini, et h est choisi tel que $h = -\eta \nabla f(x)$. Ainsi, l'expression suivante est obtenue

$$f(x - \eta \nabla f(x)) = f(x) - \eta \|\nabla f(x)\|^2 + o(\eta \|\nabla f(x)\|).$$

Étant donné que $|\nabla f(x)|^2 \geq 0$, deux cas se présentent : soit $|\nabla f(x)| = 0$, ce qui indique que le point est déjà un minimum (pouvant être local si f n'est pas convexe), soit pour η suffisamment petit, l'inégalité $f(x - \eta \nabla f(x)) < f(x)$ s'établit. Cette démarche assure ainsi une convergence progressive vers le minimum de la fonction f.

Une analogie peut être dressée à présent avec le problème de minimisation 2.1 caractérisant l'algorithme de *boosting*. L'hypothèse est donc faite que $m-1$ arbres ont été construits et que la fonction de prédiction F_{m-1} est connue. Dans ce cadre, l'objectif consiste à résoudre la minimisation suivante

$$\min_{h_m \in \mathcal{H}} \quad \frac{1}{n} \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h_m(x_i)).$$

En tirant inspiration de l'approche de minimisation exposée ci-dessous, la résolution du problème 2.1 peut être abordée au moyen d'une méthode de descente de gradient. En d'autres termes, l'objectif est de configurer l'arbre de décision de manière à satisfaire la condition

$$h_m(x) = -\eta \left[\frac{\partial l(y_i, F(x_i)))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}.$$

En pratique, pour obtenir une telle relation et paramétrer l'arbre de décision, une approximation par les moindres carrés est employée

$$h_m = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \quad \sum_{i=1}^{n} \left(g_m(x_i, y) - h(x_i)\right)^2$$

avec $g_m(x_i, y) = \left[\frac{\partial l(y_i, F(x_i)))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$.

Ainsi, le rôle du modèle faible consiste à prédire cette valeur, assurant par conséquent de manière indirecte la minimisation du risque empirique.

Algorithm 1 Gradient Tree Boosting

1. Initialiser $F_0(x) = \arg\min_{\alpha} \sum_{i=1}^{n} l(y, \alpha)$ avec une valeur constante
2. Pour m allant de 1 à M
   (a) Pour tout $i \in [\![1, n]\!]$, calculer :
   $$r_{m,i} = -\left[\frac{\partial l(y_i, F(x_i)))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$
   (b) Entraîner un arbre de décision sur r_m, on note $(R_{m,j})_{j \in [\![1, J_m]\!]}$ les régions obtenues
   (c) Pour tout $j \in [\![1, J_m]\!]$, calculer :
   $$\alpha_{m,j} = \arg\min_{\alpha} \sum_{x_i \in R_{m,j}} l(y_i, F_{m-1}(x_i) + \alpha)$$
   (d) Définir $F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^{J_m} \alpha_{m,j} \mathbb{1}_{x \in R_{m,j}}$
3. Retourner la prédiction $F_M(x)$

L'algorithme *Catboost*

Le modèle *Catboost*, élaboré par Yandex (Prokhorenkova et al. (2018)), est un algorithme de *gradient boosting* qui se caractérise par plusieurs propriétés distinctives. L'un de ses aspects majeurs réside dans sa capacité à gérer efficacement les variables catégorielles.

Dans un contexte de modélisation, traiter les variables catégorielles avec un grand nombre d'occurrences peut s'avérer périlleux. Les méthodes courantes telles que le *"one-hot-encoding"* ou le *"label-encoding"* rencontrent des limites pratiques. La première méthode, générant $d-1$ nouvelles variables pour une variable à d modalités, peut amplifier le problème de dimensionnalité. De plus, le *"label-encoding"*, qui attribue des valeurs numériques aux modalités, souffre de l'absence d'ordre entre les étiquettes.

Pour surmonter ces problèmes, l'approche de *"Target Statistics"* est employée. Pour une variable catégorielle x^j, l'idée est de substituer x_i^j par $\hat{x}_i^j = \mathbb{E}[Y | X^j = x_i^j]$. Cette quantité est estimée par la formule suivante

$$\hat{x}_k^j = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{x_i^j = x_k^j\}} y_i + ap}{\sum_{i=1}^{n} \mathbb{1}_{\{x_i^j = x_k^j\}} + a}$$

où a, p sont des paramètres de lissage (la valeur usuelle pour le paramètre p est $\frac{1}{n}\sum_{i=1}^{n} y_i$).

Cependant, l'inconvénient est le risque de *"target leakage"*, car $\hat{x}_i^j$ est calculé à partir de $\mathbf{Y}$, engendrant potentiellement un sur-apprentissage. Pour prévenir cela, l'approche innovante proposée par l'algorithme *Catboost* est d'introduire une méthode nommée *"Ordered Target Statistics"* inspirée des méthodes d'*Online Learning*. Elle ajoute un "temps" artificiel en permutant les lignes du jeu de données, formant ainsi un ensemble $\mathbb{D}_k = \{x_i \mid \sigma(i) < \sigma(k)\}$, avec σ une permutation aléatoire des lignes, ce qui élimine le risque de *"target leakage"*. La formule pour obtenir la valeur de la variable catégorielle s'écrit

$$\hat{x}_k^j = \frac{\sum_{x_i \in \mathbb{D}_k} \mathbb{1}_{\{x_i^j = x_k^j\}} y_i + ap}{\sum_{x_i \in \mathbb{D}_k} \mathbb{1}_{\{x_i^j = x_k^j\}} + a}.$$

De plus dans le contexte des algorithmes de *gradient boosting* précédemment exposés, une autre source de sur-apprentissage émerge lors de la phase d'entraînement. Celle ci résulte du calcul du gradient à chaque itération, utilisant les mêmes données ayant contribué à la construction du modèle. Cette absence d'indépendance entre les arbres successifs peut déformer la distribution conditionnelle, en particulier après une itération complète des données. Idéalement, il serait préférable de recalculer le gradient à chaque étape avec de nouvelles données non encore utilisées. Néanmoins, les limites en taille des données rendent cette approche difficile à mettre en œuvre.

Dans la lignée de la méthode *"Ordered Target Statistics"*, *Catboost* propose la solution du *"Ordered Boosting"*. Cette approche implique l'introduction de $S+1$ permutations $(\sigma)_{s\in[\![0,S]\!]}$ des observations et la notation $F_{s,m}$ qui correspond à la fonction de prédiction obtenue à la m-ième itération avec la permutation σ_s. À chaque étape m de l'algorithme et pour une permutation σ_s , n modèles $F_{s,m,j}$ sont entraînés sur les observations précédentes jusqu'à $j \in [\![1, n]\!]$. De plus, une permutation $r \in [\![1, S]\!]$ est aléatoirement sélectionnée, et le modèle associé à cette permutation $F_{r,m-1}$ est utilisé pour structurer le m-ième arbre. Ceci brise la dépendance entre les arbres construits de manière successive. Cependant, pour contenir la complexité induite par l'entraînement de n modèles, seules les prédictions des 2^i observations pour $1 < i < \log_2(n)$, satisfaisant $\sigma_r(i) < 2^{j+1}$, sont exploitées.

L'algorithme *Catboost* présente également une particularité dans la structure des arbres utilisés pour la prédiction. Premièrement, les arbres employés sont conçus de manière symétrique : chaque nœud possède uniquement deux choix de décision, comme représenté dans la figure 2.19. De plus, pour chaque couche de l'arbre, le même critère est appliqué à tous les nœuds. Cette configuration présente l'avantage d'un encodage de l'arbre simplifié. En effet, seules deux listes sont nécessaires pour caractériser l'ensemble de l'arbre : une liste de critères et une liste de valeurs correspondant à chaque feuille. En évaluant chaque critère et en traduisant la réponse en binaire, l'accès aux valeurs des feuilles s'effectue de manière efficiente. Par exemple, dans le cas d'un arbre de profondeur 3 où les critères 1 et 3 sont validés, l'indice binaire $101_2 = 5_{10}$ permet d'extraire la cinquième valeur dans la liste des feuilles.

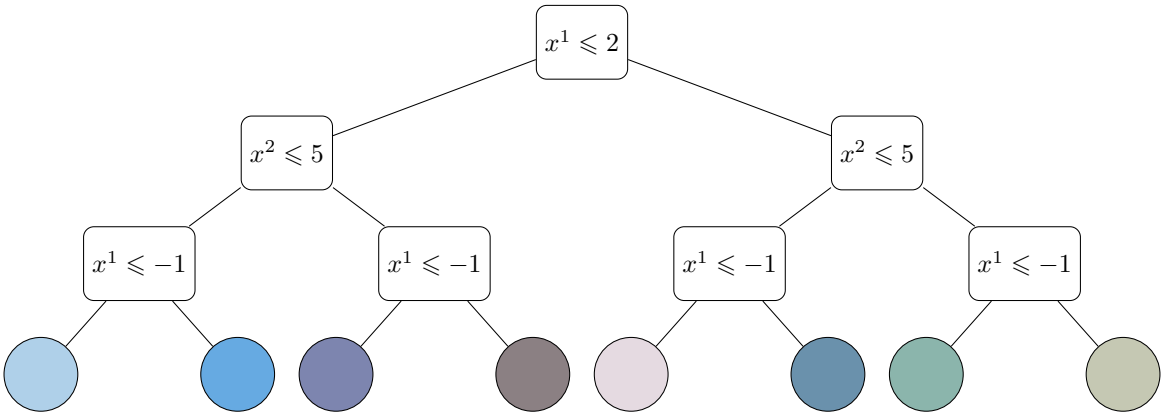


FIGURE 2.19 : Arbre symétrique employé par l'algorithme Catboost

Ainsi, ces attributs confèrent à l'algorithme *Catboost* une efficacité notable par rapport à d'autres méthodes de *boosting*, tant dans la construction des arbres (grâce à la réduction du nombre de critères) que dans les opérations de prédiction (via un accès particulièrement rapide aux valeurs). Parallèlement, les exigences en matière d'espace pour le stockage de ces modèles sont également réduites. Enfin, *Catboost* offre également une souplesse dans la modélisation à travers des hyperparamètres, décrit en partie 2.4.2, qui permettent de répondre aux exigences de l'étude.

2.4 Modèle sécheresse

La partie suivante s'intéresse à l'analyse et l'optimisation des performances du modèle *Catboost* implémenté pour répondre à la problématique. La mise en pratique du modèle est également suivie d'une étude concernant l'explicabilité du modèle.

Pour rappel l'objectif du modèle est de prédire la charge sinistre départementale liée à un phénomène de sécheresse géotechnique, notée Y. Les variables explicatives utilisées au sein du modèle peuvent se distinguer en deux catégories. D'une part les variables qui concernent l'exposition des départements assurés, caractérisant leur susceptibilité à subir un événement de sécheresse, telles que

- la proportion des logements exposés aux différents niveau de risque RGA,
- la proportion de surface du département associés aux différents niveau de risque RGA,
- la proportion de maison et d'appartements,
- Le montant des valeurs assurées.

D'autre part les indices de sécheresse, construits en partie 2.2, permettant de faire le lien entre les conditions climatiques d'une zone et la manifestation d'un événement de sécheresse

- l'indice météorologique SPEI-3, qui se décline par saison,
- l'indice hydrologique de magnitude SWI, défini annuellement.

2.4.1 Etude des performances

Dans cette partie, une première exploration des performances du modèle, mettant en évidence notamment ses limites, est effectuée. Pour rappel, l'objectif majeur est de développer un modèle qui

puisse généraliser efficacement les prédictions de la charge sinistre, tant du point de vue temporel (pour une nouvelle année) que spatial (d'un département à un autre). À cet effet, l'évaluation du modèle se base sur les métriques définies dans la section 2.3.2, lesquelles émanent des techniques de validation croisée, à la fois spatiale et temporelle. Il convient de noter que les résultats exposés dans la section suivante sont issus d'un modèle *Catboost* non optimisé.

Performance spatiale et temporelle

Pour se convaincre de la pertinence du modèle, une première étape de validation croisée temporelle est effectuée. Lors de cette procédure, les métriques $\mathcal{T}(R^2)$ et $\mathcal{T}(\text{MAE})$ sont calculées et parallèlement, la charge sinistre prédite est tracée sur la figure 2.20 pour une évaluation temporelle.

Les résultats indiquent que le modèle a obtenu un $\mathcal{T}(R^2)$ de 0.51. Il s'agit d'une mesure de performance initiale encourageante, en considérant le fait que le modèle n'ait pas encore été optimisé. De même, le $\mathcal{T}(\text{MAE})$ révèle une erreur d'environ 170 k euros. Il est important de noter que le montant moyen d'un sinistre de sécheresse est de 16 k euros. Ainsi, cette erreur équivaut à une moyenne d'approximativement 11 sinistres de sécheresse par département lorsque le modèle effectue des prédictions concernant la charge sinistre pour une nouvelle année.

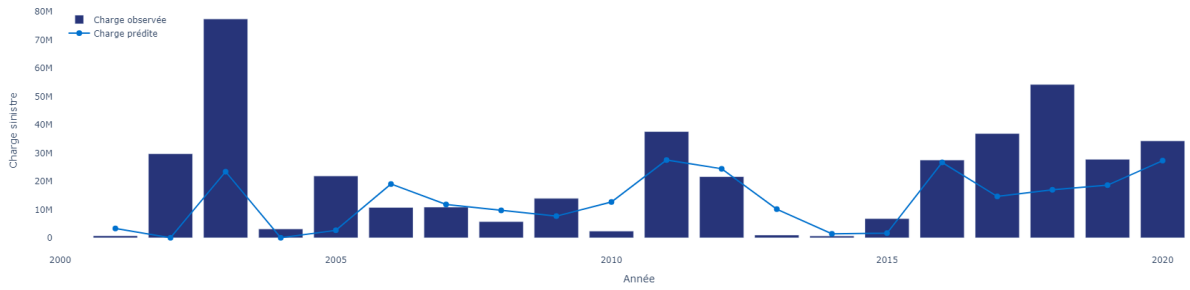
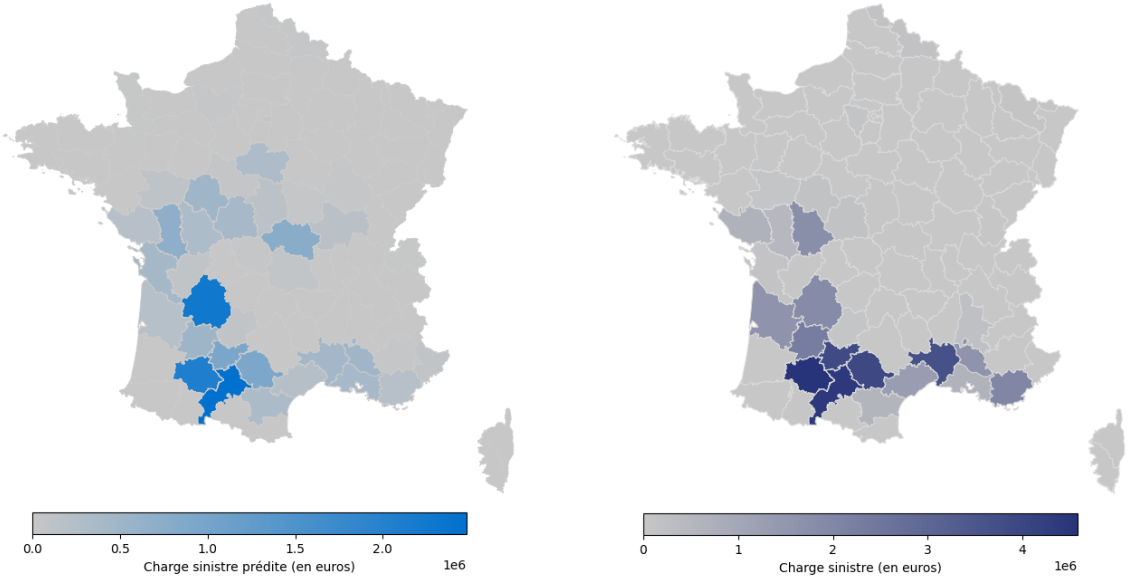


FIGURE 2.20 : Comparaison temporelle de la charge sinistre cumulée annuelle observée et prédite

L'observation de la figure 2.20 révèle que le modèle réussit à saisir les tendances générales de la sinistralité. Toutefois, il présente une sous-estimation significative de la charge sinistre pour les années marquées par des événements majeurs de sécheresse, notamment en 2003, 2005, 2011, 2017, ainsi qu'en 2018. Cette constatation souligne que le modèle ne parvient pas à appréhender pleinement l'ampleur des événements de sécheresse lorsque ces derniers se démarquent de manière exceptionnelle.

Une autre approche consiste à examiner la répartition spatiale des sinistres par année, afin d'évaluer la capacité du modèle à identifier les événements de sécheresse. Comme le montre la figure 2.21, le modèle parvient à reproduire de manière cohérente la configuration spatiale des sinistres pour une année donnée. Cependant, il est à noter que, de façon récurrente, le modèle présente une tendance à sous-estimer la magnitude des sinistres.

En ce qui concerne les résultats issus de la validation croisée spatiale, il est important de noter que les performances du modèle sont légèrement atténuées. Cette diminution de performance découle de la restriction imposée par la procédure de validation, qui restreint l'accès aux données d'entraînement pour certaines zones géographiques à risque. En conséquence, cela engendre une déficience dans l'apprentissage du modèle pour ces zones spécifiques, entraînant ainsi un biais dans les estimations.



(a) Charge sinistre prédite - Année 2017 (b) Charge sinistre observée - Année 2017

FIGURE 2.21 : Comparaison spatiale de la charge sinistre observée et prédite - Année 2017

Les métriques spatiales $\mathcal{S}(R^2)$ et $\mathcal{S}(\text{MAE})$ révèlent un score de 0.31 pour la première et une erreur de 220 000 pour la seconde. Ces résultats mettent en évidence la complexité du modèle à reproduire de manière précise la sinistralité d'un département lorsque celui-ci est exclu totalement de la base d'entraînement. Ces observations soulignent également la nature intrinsèquement singulière de la structure géographique de chaque zone.

Performances sur les zones à risques

Conformément aux constatations précédentes, les premières analyses du modèle montrent que le modèle présente des lacunes dans sa capacité à détecter des événements exceptionnels de sécheresse. Pour quantifier précisément ces erreurs, l'utilisation de la métrique MAPEX(q) s'avère pertinente, car elle permet d'évaluer les performances du modèle spécifiquement sur les zones caractérisées par une sinistralité significative.

Les calculs de $\mathcal{T}(\text{MAPEX}(q))$ et $\mathcal{S}(\text{MAPEX}(q))$ révèlent des valeurs de 21% et 33%, respectivement. Ces résultats indiquent que, au cours des processus de validation, l'erreur relative moyenne en pourcentage sur les événements majeurs de sécheresse est de 21% pour la validation temporelle et de 33% pour la validation spatiale, laquelle semble être une nouvelle fois plus pénalisante.

La valeur ajoutée de cette approche consiste en l'expression de l'erreur en pourcentage par rapport à la valeur cible, ce qui limite l'amplification des erreurs causée par les valeurs extrêmes. De plus, elle permet de comparer des ensembles de validation présentant des ordres de grandeur différents.

Les résultats obtenus à travers cette métrique mettent en évidence les difficultés intrinsèques du modèle, qui ont été précédemment identifiées en ce qui concerne les zones à forte sinistralité. L'objectif à terme est de parvenir à développer un modèle capable de mieux détecter les épisodes de sécheresse extrême, même au prix d'une légère surestimation des événements mineurs, dans une perspective de gestion prudente des risques. Pour ce faire, une étape importante consiste à optimiser les

hyperparamètres du modèle de manière à minimiser la métrique MAPEX(q), tout en maintenant des performances satisfaisantes sur les autres métriques relatives à la dimension temporelle et spatiale.

	Métriques temporelles			Métriques spatiales		
Métriques	$\mathcal{T}(R^2)$	$\mathcal{T}$(MAE)	$\mathcal{T}$(MAPEX(q))	$\mathcal{S}(R^2)$	$\mathcal{S}$(MAE)	$\mathcal{S}$(MAPEX(q))
Modèle de référence	0.51	170×10^3	21%	0.31	220×10^3	33%

TABLE 2.2 : Métriques pour le modèle non optimisé

2.4.2 Optimisation du modèle

Traitement des données déséquilibrées

La variable cible affiche un déséquilibre notable, où environ 68% des observations du jeu de données se caractérisent par une absence totale de charge sinistre. Ce déséquilibre a une incidence sur le processus d'apprentissage du modèle, puisque ce dernier est exposé à un grand nombre de données où la charge sinistre est nulle. Par conséquent, le modèle a tendance à sous-estimer le montant des sinistres. De plus, ce déséquilibre dans les données entrave la capacité des modèles à identifier et à prédire avec précision les événements de sinistralité élevée au sein du portefeuille. Or, ces événements ont de fait un poids important dans la sinistralité globale du portefeuille, justifiant ainsi la nécessité de les modéliser de manière plus efficace.

Une stratégie pour aborder le problème des données déséquilibrées consiste à intervenir directement au niveau de la fonction de perte l, qui joue un rôle clé dans le processus d'entraînement de l'algorithme. Comme indiqué précédemment, pendant la phase d'entraînement, l'algorithme *Catboost* cherche à minimiser cette fonction de perte, ce qui se traduit par une diminution progressive des valeurs du gradient associées à cette fonction. Une valeur faible de gradient indique que le modèle présente peu d'erreurs et nécessite donc peu de corrections. En revanche, un gradient élevé suggère que le modèle comporte de nombreuses erreurs, qui doivent être corrigées de manière significative.

Les méthodes de *boosting*, y compris *Catboost*, ont été conçues de manière à utiliser ces valeurs de gradient comme base pour entraîner successivement de nouveaux arbres, afin de corriger les erreurs accumulées par les arbres précédemment construits. Par conséquent, afin de donner une priorité ou pénaliser spécifiquement certaines erreurs, il est possible d'ajouter un vecteur de poids, noté w, à l'intérieur de la fonction de perte. Ce vecteur de poids permet d'assigner un poids particulier, noté w_i, à chaque prédiction $\hat{f}(x_i)$ dans l'ensemble d'entraînement.

Dans ce contexte, la MSE décrite précédemment, utilisée pour l'optimisation pendant la phase d'apprentissage de l'algorithme, subit une transformation. Elle devient alors

$$\text{MSE}(w) = \frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \, w_i}{\sum_{i=1}^{n} w_i}.$$

Pour détecter les charges liées à des événements anormaux de sécheresse, il est donc judicieux de pénaliser davantage l'erreur de prédiction lorsque celle ci est associée à une charge sinistre importante. Pour ce faire, un seuil ϵ est défini de telle sorte que

$$w_i = \begin{cases} 1 & \text{si } y_i = 0 \\ w^* & \text{si } y_i \geq \epsilon \end{cases}$$

avec w^*, le poids optimal attribué aux erreurs commises sur les grandes valeurs prédites, obtenu lors de l'optimisation des hyperparamètres.

D'autres approches existent pour traiter des données caractérisées par une distribution déséquilibrée. Parmi celles-ci, les techniques d'échantillonnage visent à réajuster la base d'entraînement afin de compenser cette disparité. Deux catégories principales se distinguent : les techniques de "sur-échantillonnage" (*oversampling*), dont le principe est de créer de nouvelles observations appartenant à la classe minoritaire, et de "sous-échantillonnage" (*undersampling*), qui consistent à éliminer les éléments de la classe majoritaire.

L'approche la plus simple de sur-échantillonage est le *Random Oversampling*. Cette dernière implique la création de nouvelles observations issues de la classe minoritaire en effectuant un tirage avec remise. Cependant, cette méthode n'est pas optimale et présente un risque de sur-apprentissage dû à la duplication d'observations.

Une méthode plus avancée, SMOTE (*Synthetic Minority Oversampling Technique*) (Chawla et al. (2002)), génère de nouvelles observations en combinant de manière convexe des éléments de la classe minoritaire, en veillant à ce qu'au moins l'un des éléments combinés soit l'un des k plus proches voisins de l'autre. Cette technique crée ainsi de nouvelles observations de la classe minoritaire, positionnées volontairement à proximité des observations déjà existantes dans la base d'entraînement.

En contexte de régression, l'application de ces méthodes est moins directe en raison de l'absence de classes pour catégoriser les observations. L'algorithme SMOGN (*Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise*)(Branco et al. (2017)), utilisé dans cette étude, propose une méthode dérivée du principe de SMOTE et adaptée à des variables à prédire continues.

Cependant ces techniques de *sampling* ne se sont pas avérées pertinentes au cours de l'amélioration du modèle, conduisant à retenir la méthode de pénalisation du gradient de la fonction de perte.

Pondération des variables explicatives et application des contraintes de monotonie

Afin de répondre aux objectifs du modèle, notamment lorsqu'il existe une connaissance métier *a priori* des variables et de leur influence sur le modèle, il peut être judicieux de contraindre le modèle pour améliorer ses performances. Cette approche est particulièrement pertinente lorsqu'il y a existence d'une relation de monotonie entre une variable explicative donnée et la variable cible du modèle. Dans ce cas, il est possible d'imposer une contrainte de monotonie sur cette variable explicative en utilisant l'hyperparamètre *monotone constraints* du modèle *Catboost*. Cette contrainte peut être soit croissante, soit décroissante. En considérant f, la fonction de prédiction du modèle, et $\{x^1, ..., x^p\}$ l'ensemble des p variables explicatives du modèle, cette contrainte s'exprime comme suit

$$f(x^1, x^2, ..., x, ..., x^{p-1}, x^p) < f(x^1, x^2, ..., x', ..., x^{p-1}, x^p)$$

pour $x < x'$, dans le cas d'une contrainte croissante, ou encore

$$f(x^1, x^2, ..., x, ..., x^{p-1}, x^p) > f(x^1, x^2, ..., x', ..., x^{p-1}, x^p)$$

pour $x > x'$, dans le cas d'une contrainte décroissante.

Dans le contexte de l'étude, l'analyse des indices de sécheresse élaborés en partie 2.2 a permis de mettre en évidence une relation décroissante entre l'indice SPEI-3 et la charge sinistre et une relation croissante entre l'indice de magnitude SWI et la charge sinistre. Partant de ce postulat, une contrainte décroissante a été imposée pour l'indice SPEI-3 et croissante pour l'indice de magnitude SWI.

Parallèlement, il est envisageable de renforcer l'influence d'une variable sur les sorties du modèle en attribuant des poids spécifiques à certaines variables explicatives à l'aide de l'hyperparamètre *feature weights* de l'algorithme *Catboost*. Ces poids interviennent lors de la construction de l'arbre de décision, comme exposé dans la section 2.3.3, en augmentant artificiellement le score de la variable concernée lors du processus de découpage, afin d'accroître sa capacité discriminante. L'intérêt de cette approche réside dans le fait que si une variable joue un rôle déterminant dans la prédiction, il peut être judicieux de forcer son utilisation.

Pour l'optimisation du modèle, plusieurs valeurs de poids ont été testées, et ces poids ont été appliqués aux variables sur lesquelles avaient été préalablement imposée une contrainte de monotonie. De plus, des poids ont été associés aux variables d'exposition au RGA dans le but d'amplifier leur effet bénéfique sur les prédictions réalisées.

Optimisation des hyperparamètres

Comme indiqué précédemment, les algorithmes de *boosting*, notamment *Catboost*, offrent une flexibilité considérable en ce qui concerne la sélection et l'optimisation des hyperparamètres. De ce fait, l'optimisation du modèle s'est concentrée sur une fraction précise des hyperparamètres et selon une certaine approche méthodique.

Dans une première étape, l'objectif est de déterminer la structure optimale du modèle. Ceci comprenait la fixation d'un taux d'apprentissage (pas de la descente de gradient) approprié, déterminé par l'hyperparamètre *learning rate*, la sélection du nombre optimal d'itérations (nombre d'arbres) à l'aide de l'hyperparamètre *num trees*, ainsi que le choix de la profondeur maximale pour chaque arbre construit grâce à l'hyperparamètre *depth*. Par ailleurs, l'hyperparamètre *border count* a été optimisé, ce dernier contrôlant le nombre de divisions autorisées pour les variables numériques lors du processus de découpage.

Cette première phase d'optimisation visait à établir les paramètres structurels du modèle de manière à ce qu'il puisse apprendre de manière efficace et appropriée à partir des données.

Dans une seconde étape et dans le but d'améliorer la robustesse du modèle dans un contexte de données déséquilibrées, l'attention s'est portée sur les hyperparamètres visant à prévenir le surapprentissage. À cet effet, l'hyperparamètre *random strength* a été utilisé pour rajouter de l'aléa lors de la construction de l'arbre. De plus, le nombre minimal d'observations requis dans une feuille a été spécifié grâce à l'hyperparamètre *min data in leaf*, et enfin, un coefficient de régularisation L2 a été appliqué à la fonction de perte en utilisant le paramètre *L2 leaf reg*.

Enfin, en dernière étape, l'optimisation a porté sur les poids utilisés pour la pénalisation du gradient de la fonction de perte, noté w^*, ainsi que sur les poids utilisés pour la pondération des variables explicatives.

L'ensemble de ces hyperparamètres a été optimisé à l'aide la procédure de validation croisée temporelle. Ainsi, les hyperparamètres retenus sont ceux minimisant l'erreur des métriques temporelles définie en partie 2.3.2.

Analyse du modèle optimisé

Après l'optimisation de tous les paramètres, une évaluation comparative entre le modèle optimisé et le modèle de référence (non optimisé) est effectuée.

L'analyse temporelle, illustrée dans la figure 2.22, révèle que les prédictions de charge sinistre générées par le modèle optimisé se rapprochent davantage des valeurs de charge sinistre observées. Outre la capacité du modèle optimisé à reproduire avec précision les tendances de la charge de sinistre, il se démarque également du modèle de référence en étant capable d'estimer plus précisément les charges sinistres exceptionnelles des années 2003, 2017, et 2018.

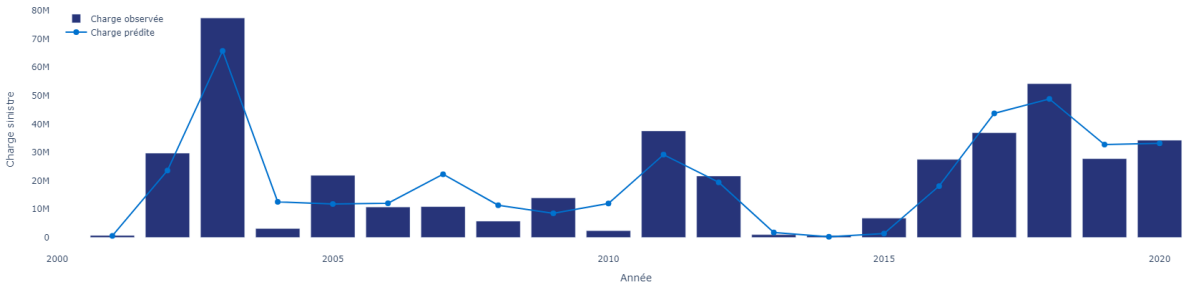


FIGURE 2.22 : Comparaison temporelle de la charge sinistre cumulée annuelle observée et prédite pour le modèle optimisé

Cette faculté à reproduire les pics de sinistralité se manifeste également par une amélioration significative des métriques $\mathcal{T}(\text{MAPEX}(q))$ et $\mathcal{S}(\text{MAPEX}(q))$, qui passent de 21% à 12% et de 33% à 26% respectivement. La MAE ainsi que le R^2 ne connaissent pas d'amélioration majeure à l'issue de l'optimisation. En ce qui concerne le R^2, $\mathcal{T}(R^2)$ augmente de 0.51 à 0.55 $\mathcal{S}(R^2)$ de 0.31 à 0.36. Pour la MAE, l'erreur moyenne est désormais de 140 000 pour la validation croisée temporelle et de 193 000 pour la validation croisée spatiale.

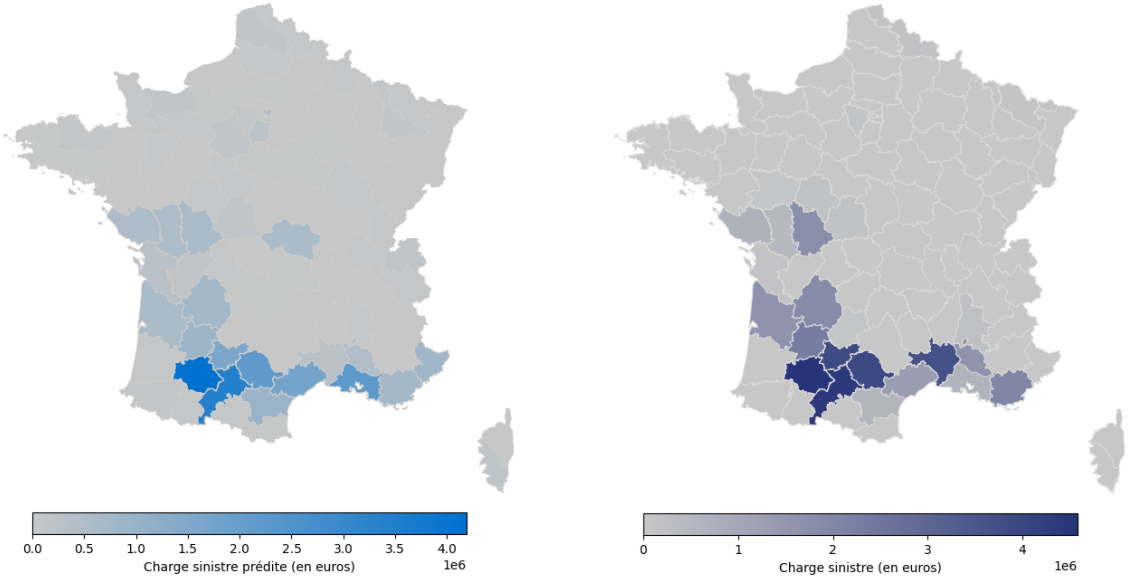
L'optimisation des paramètres, principalement axée sur l'amélioration de la détection des événements majeurs de sinistralité sécheresse, a abouti à une réduction globale de l'erreur dans les zones à forte sinistralité, comme en témoignent les résultats de la métrique MAPEX. Cependant, cette optimisation a également entraîné une surestimation de la charge sécheresse dans certains département caractérisés par une sinistralité moins prononcée.

La figure 2.23 met notamment en évidence ce phénomène de manière explicite. La charge sinistre prédite se rapproche considérablement de la charge observée dans les départements fortement touchés par les sinistres. Toutefois, cette correction apportée grâce à l'optimisation du modèle entraîne simultanément une surestimation de la charge sinistre dans certains départements. Dans une perspective prudentielle, les résultats du modèle optimisé demeurent préférables à ceux du modèle de référence.

	Métriques temporelles			Métriques spatiales		
Métriques	$\mathcal{T}(R^2)$	$\mathcal{T}(\text{MAE})$	$\mathcal{T}(\text{MAPEX}(q)$	$\mathcal{S}(R^2)$	$\mathcal{S}(\text{MAE})$	$\mathcal{S}(\text{MAPEX}(q)$
Modèle de référence	0.51	170×10^3	21%	0.31	220×10^3	33%
Modèle optimisé	0.55	140×10^3	12%	0.36	193×10^3	25%

TABLE 2.3 : Comparaison des métriques entre le modèle de référence et optimisé

Les résultats obtenus, bien que globalement satisfaisants en ce qui concerne l'estimation annuelle de la charge sinistre, souffre d'un manque de précision quant à la répartition géographique des sinistres



(a) Charge sinistre prédite - Année 2017 (b) Charge sinistre observée - Année 2017

FIGURE 2.23 : Comparaison spatiale de la charge sinistre observée et prédite pour le modèle optimisé- Année 2017

prédits, mettant ainsi en lumière les défis inhérents à la modélisation du risque physique associé au phénomène de subsidence. Comme discuté dans l'article de CHARPENTIER et al. (2022b), l'incertitude des prédictions peut être en partie attribuée à l'hétérogénéité des données utilisées pour calibrer les modèles. Les critères de déclaration des catastrophes naturelles ont subi de multiples modifications au fil de la période de calibration, ce qui a introduit des biais dans les données historiques. De plus, les prédictions auraient pu bénéficier d'une résolution spatiale plus fine des données pour prendre en compte les caractéristiques individuelles de chaque propriété, telles que la proximité des arbres ou encore la topographie du terrain.

Le modèle permet néanmoins d'établir un lien cohérent entre les conditions géologiques et météorologiques du territoire métropolitain et l'occurrence des phénomènes de sécheresse au cours de la période d'observation, fournissant des informations utiles pour la compréhension de la gravité des sinistres au fil des années.

Au terme de l'analyse, ce modèle constitue une base satisfaisante pour appréhender le risque sécheresse dans un contexte climatique futur, une étude qui sera approfondie dans le prochain chapitre.

2.4.3 Interprétabilité

La construction du modèle présenté précédemment revêt un double intérêt. D'une part, il offre la possibilité d'obtenir des prédictions cohérentes concernant la sinistralité liée aux épisodes de sécheresse. D'autre part, il constitue un outil de pilotage permettant d'identifier les variables importantes susceptibles d'expliquer l'évolution de la sinistralité dans des zones spécifiques touchées par la sécheresse. De surcroît, dans le contexte de l'assurance, il est impératif que les modèles soient interprétables, de manière à ce que l'assureur puisse justifier auprès de l'assuré toute augmentation tarifaire résultant d'une hausse de sinistralité. Pour parvenir à cette interprétation du modèle, le concept des valeurs SHAP (*SHapley Additive exPlanations*) (LUNDBERG et LEE (2017)), emprunté à la théorie des jeux,

se révèle particulièrement adapté aux modèles d'apprentissage automatique, comme celui employé lors de l'étude.

Introduction aux valeurs SHAP

Alors que les modèles linéaires jouissent d'une bonne interprétabilité grâce à l'attribution explicite de coefficients à chaque variable, les modèles basés sur des arbres de décision et les techniques de *boosting* présentent un défi en termes de transparence en raison de leur complexité intrinsèque. Cependant, des avancées récentes ont permis de développer des méthodes visant à rendre ces modèles plus explicables et à les sortir de la catégorie des "boîtes noires". L'une de ces avancées est l'introduction des valeurs SHAP. Cette approche permet non seulement d'évaluer l'importance des variables au sein du modèle, mais elle offre également la possibilité d'analyser comment une variable influence le modèle, ce qui s'avère particulièrement utile lorsqu'il s'agit de valider des hypothèses basées sur une intuition métier.

Le concept de valeurs SHAP découle de la théorie des jeux où chaque observation constitue un nouveau jeu et chaque variable est appréhendée comme un joueur. L'objectif consiste à mesurer l'effet de chacun des joueurs (variables) sur le résultat final (prédcition du modèle). De manière très simplifiée, la valeur SHAP d'une variable x^j correspond précisément à la contribution marginale de cette dernière à la prédiction finale du modèle.

Afin d'obtenir cette contribution, les valeurs SHAP utilise la notion de coalition, définie comme un sous-ensemble non vide d'indices des variables explicative $S \subseteq P$, où P représente l'ensemble complet des indices. Une coalition S reflète une combinaison spécifique des variables explicatives, impliquant ainsi une interaction simultanée entre elles lors de la contribution à la prédiction du modèle. Lors du calcul de la valeur SHAP, toutes les coalitions possibles sont considérées afin de capturer les relations entre les variables explicatives.

La valeur SHAP de la variable x^j, pour la fonction de prédiction f est finalement donnée par la formule

$$\text{SHAP}(x^j, f) = \sum_{S \subseteq P \setminus \{j\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} (f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S))$$

avec

- $x_{S\cup\{j\}}$ et x_S les valeurs des variables explicatives dans l'ensemble $S \cup \{j\}$ et S respectivement.
- $f_{S\cup\{j\}}$ et f_S les fonctions de prédictions entraînés sur les variables explicatives de l'ensemble $S \cup \{j\}$ et S respectivement.

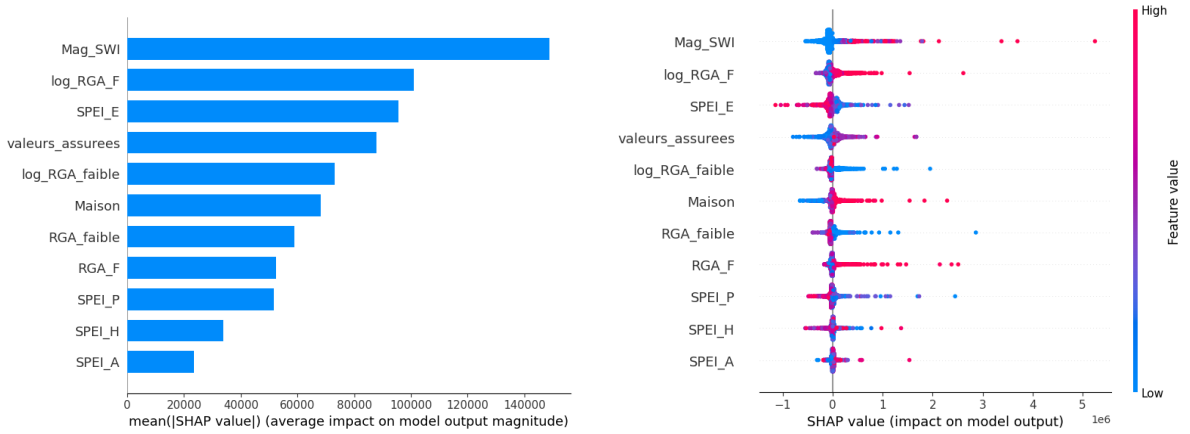
Dans la formule précédente $f_{S\cup\{j\}}(x_{S\cup\{j\}})$ correspond à la contribution marginale de la coalition S lorsque x^j est ajoutée, et $f_S(x_S)$ quantifie la contribution marginale de la coalition S sans x^j.

Cette approche offre une méthodologie rigoureuse pour décomposer les prédictions du modèle, distinguant avec précision l'apport de chaque variable tout en tenant compte de ses dépendances avec les autres.

Pour obtenir les valeurs SHAP de l'ensemble des variables du modèle, ce processus doit être généralisé sur chaque observation et pour tous les jeux. Cependant, le nombre de coalition possible pour un ensemble de p variables explicatives est de 2^p, ce qui rend impossible le calcul dans la plupart des cas. De ce fait, c'est une valeur SHAP légèrement simplifiée, moins coûteuse, qui est calculée en pratique.

Valeurs SHAP du modèle

Dans une logique d'interprétation du modèle obtenu, les valeurs SHAP des variables sont tracées et hiérarchisées par leur ordre d'importance, qui traduit leur influence sur les prédictions. Pour compléter l'explicabilité du modèle un graphique *Beeswarm* est également tracé, qui est une fonctionnalité de la librairie SHAP de Python permettant de déterminer la relation des variables explicatives avec la variable cible. Ce dernier se lit de la manière suivante : pour chaque variable explicative, les différentes valeurs prises par la variable sont représentées de manière graphique (rouge pour les grandes valeurs, bleu pour les petites valeurs) et si pour un type de valeur prise par la variable on s'écarte de l'origine vers la droite, cela veut dire que pour ce type de valeur, la variable impacte positivement la prédiction. Symétriquement, si l'on s'écarte vers la gauche, cela montre que la variable impacte négativement la variable cible.



(a) Importance des variables selon SHAP (b) Graphique *beeswarm*

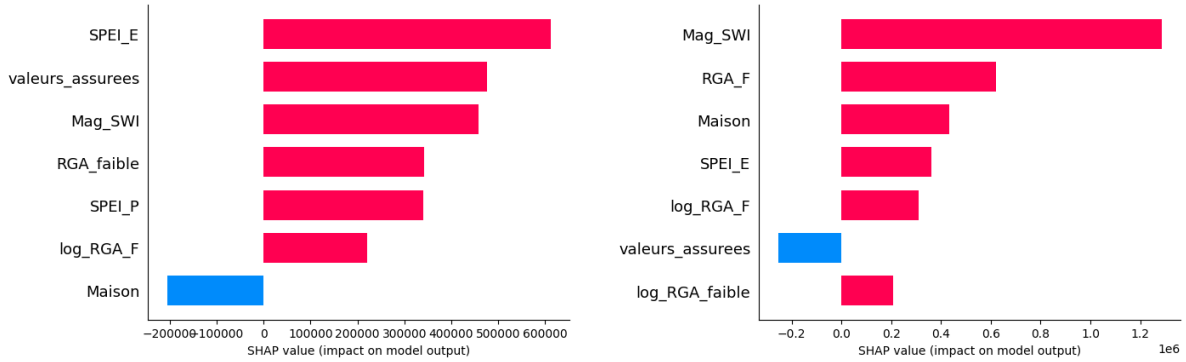
FIGURE 2.24 : Résultat des valeurs SHAP pour le modèle optimisé

Le graphique (a) de la figure 2.24 révèle que les indices de sécheresse, en particulier l'indice de magnitude SWI et le SPEI estival, revêtent une importance cruciale dans les prédictions, comme en témoigne leur valeur moyenne SHAP élevée. Le graphique (b) de la figure 2.24 met en évidence la relation positive entre l'indice de magnitude SWI et le montant de la charge sinistre, bien que cette influence semble plus marquée lorsque l'indice atteint des valeurs élevées. Cette observation peut s'expliquer en partie par la nature même de cet indice, qui mesure davantage l'intensité de la sécheresse d'une zone donnée plutôt que son état général d'humidité ou de sécheresse.

En ce qui concerne l'indice SPEI estival, il présente une relation négative significative avec la charge sinistre et joue un rôle clé pour expliquer la survenance d'un sinistre. Le SPEI printanier montre également une relation similaire, bien que son impact sur les prédictions du modèle semble moins prononcé. Quant aux indices SPEI hivernal et automnal, leur lien avec la variable cible apparaît moins évident. Il est important de noter que la forte présence de maison dans le département, ainsi que la proportion élevée de logements situés dans des zones à risque RGA élevé, constituent naturellement des facteurs aggravants dans l'occurrence des sinistres liés à la sécheresse.

Par ailleurs, une analyse locale des valeurs SHAP, c'est-à-dire à l'échelle de chaque observation, permet de mettre en évidence la complémentarité des différents indices de sécheresse dans la modélisation.

En examinant deux événements de sécheresse majeurs présentés dans la figure 2.25, il est possible



(a) Valeur SHAP - Gironde 2017

(b) Valeur SHAP - Cher 2003

FIGURE 2.25 : Comparaison des valeurs SHAP entre deux zones fortement sinistrées

de constater que, pour la charge sinistre survenue en Gironde en 2017, l'indice SPEI estival exerce une influence plus marquée sur les prédictions. En revanche, pour la sécheresse de 2003, qui a touché entre autres le département du Cher, l'information concernant les conditions de sécheresse est davantage véhiculée par l'indice de magnitude SWI.

Chapitre 3

Projection et soutenabilité du risque sécheresse à climat futur

Ce troisième chapitre aborde la projection des éléments intervenant dans la prédiction du risque sécheresse afin d'aboutir à une estimation de la charge sinistre à climat futur. Cette démarche vise à étudier, pour l'ensemble du territoire métropolitain, l'évolution de la prime d'assurance qui est directément liée à ce risque. Du point de vue de l'assuré, la question de la soutenabilité de la prime au regard de sa progression attendue à horizon 2050, est par conséquent étudiée. Le choix de l'horizon 2050 permet de garder une incertitude raisonnable dans les projections, tout en appréciant l'évolution du climat, et de comparer les résultats avec d'autres études menées par les acteurs de l'industrie de l'assurance.

Dans **une première section**, la projection des variables climatiques, permettant le calcul des indices de sécheresse, ainsi que l'évolution des enjeux des assurés sont détaillées.
Dans **une deuxième section**, la projection de la charge sinistre sécheresse à climat futur est présentée, incluant une analyse des différents facteurs responsables de son augmentation.
Dans **une dernière section**, la question de la soutenabilité de la prime d'assurance est étudiée par le biais de différents scénarios de tarification du risque sécheresse.

3.1 Projection des variables climatiques et d'exposition

Cette partie traite tout d'abord la projection des variables climatiques, à savoir les températures, les précipitations et l'indice d'humidité des sols SWI, qui constituent la base pour la construction des indices de sécheresse. Par la suite, l'évolution des enjeux des assurés, à travers la répartition géographique du portefeuille mais également les valeurs des biens assurés, est explorée.

3.1.1 Projection des indices de sécheresse

Afin de confronter le risque sécheresse au changement climatique, il est pertinent de recourir à des scénarios d'émissions futures de gaz à effet de serre, étant donné que ce changement est essentiellement d'origine anthropique et résulte de ces émissions. Dans cette perspective, les scénarios RCP élaborés par le GIEC s'appuient sur cette observation fondamentale et sont donc employés comme base dans cette étude pour projeter les variables climatiques et, par conséquent, les indices de sécheresse.

Les scénarios RCP

Les scénarios RCP (*Representative Concentration Pathway*) correspondent à des scénarios socio-économiques associés à des trajectoires de forçage radiatif sur une période s'étendant de 2006 à 2300. La

définition de ces scénarios a été effectué par le GIEC (Groupement d'experts Intergouvernemental sur l'Evolution du Climat). Le GIEC est un groupe crée en 1988 relevant de l'Organisation Météorologique Mondiale et du Programme des Nations Unies pour l'Environnement. Il rassemble une communauté international de scientifiques, climatologues et socio-économistes dans le but de mener à bien sa mission, à savoir l'évaluation objective du phénomène de réchauffement climatique et la synthèse des connaissances afférentes.

Quatre trajectoires, illustrées dans le schéma 3.1, ont été présentées dans le cinquième rapport du GIEC (2014), communément appelé "AR5". Ces trajectoires diffèrent de par les hypothèses sur la quantité émise de gaz à effet de serre et notamment sur les stratégies d'adaptation et d'atténuation mises en place pour réduire leur émission. Les gaz à effets de serre de nature anthropique, responsable du rechauffement climatique, sont directement reliés au concept de forcage radiatif, utilisé par le GIEC. Le forcage radiatif se définit comme l'équilibre entre les rayonnements solaires entrants et les émissions de rayonnements infrarouges sortants de l'atmosphère, notamment au niveau du sommet de la troposphère, qui est la couche de l'atmosphère en contact direct avec la surface terrestre. En d'autres termes, un forçage radiatif positif indique un excès de rayonnement solaire absorbé, entraînant ainsi un réchauffement du système. Les RCP sont dénommés d'après leur niveau de forcage radiatif (exprimé en W/m^2) atteint en 2100 :

- **RCP 2.6** : scénario le plus optimiste qui caractérise un monde à très faibles émissions avec un point culminant avant 2050, suivi d'une diminution ;
- **RCP 4.5** et **RCP 6.0** : scénarios intermédiaires dans lesquels les émissions de gaz à effet de serre sont stabilisées avant la fin du XXI^{e} siècle ;
- **RCP 8.5** : scénario le plus pessimiste au cours duquel les émissions de gaz à effet de serre continuent d'augmenter au rythme actuel, souvent appelé *business as usual*.

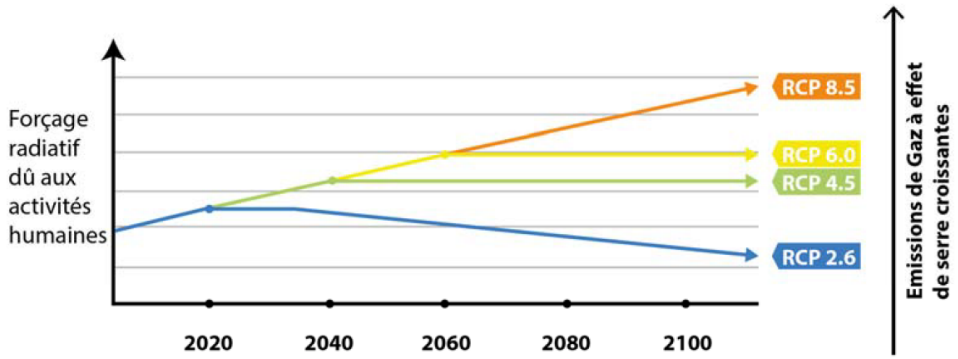


FIGURE 3.1 : Scénarios d'évolution du climat RCP

Dans le cadre de l'étude, le scénario RCP 4.5 a été retenu pour projeter les variables climatiques. Bien que le scénario 8.5 ait été envisagé au départ, le choix du scénario 4.5 s'explique par le fait que celui ci ne présente pas de différence notable avec le 8.5 pour l'horizon considéré. En effet, comme illustré dans la figure 3.1, les trajectoires des scénarios RCP 4.5 et 8.5 se confondent jusqu'en 2040, et les mesures visant à réduire les émissions de gaz à effet de serre dans le scénario RCP 4.5 sont mises en place progressivement à partir de cette date. De plus, la faisabilité du scénario RCP 8.5 fait l'objet de débats, principalement en raison des hypothèses sous-jacentes concernant l'évolution de l'utilisation des énergies fossiles sur lesquelles il repose. Par conséquent, sa pertinence en tant que scénario représentatif

d'une évolution "*business as usual*" est sujette à des interrogations, comme expliqué par Hausfather (2019). Enfin le choix de ce scénario permet d'être cohérent avec les hypothèses du nouvel exercice pilote sur les stress test climatique qui sera mené par l'ACPR (2023).

Afin de générer des projections de variables climatiques cohérentes avec les trajectoires sélectionnées, les scénarios RCP sont incorporés au sein de modèles climatiques. Le projet DRIAS, dont les principes sont exposés ci-dessous, offre notamment un accès aux projections de ces modèles climatiques, tenant compte des scénarios d'émissions de gaz à effet de serre.

Projections climatiques du DRIAS

Le projet DRIAS (2012) (Donner accès aux scénarios climatiques Régionalisés francais pour l'impact et l'Adaptation de nos sociétés et environnement) a abouti en 2012 à la mise en place d'un portail web, en partenariat avec de nombreux laboratoires français de modélisation du climat (CERFACS, CNRM, IPSL). A l'initiative de Météo-France, ce portail donne au grand public l'accès à des projections climatiques sur le territoire français sous la forme de cartes interactives ou de données brutes à télécharger.

Le DRIAS regroupe différents modèles climatiques, présentés dans la figure 3.2, composés chacun d'un modèle climatique global et régional. Les modèles globaux, qui ont pour vocation à représenter les interactions physiques entre océans, continents et atmosphère, présentent des limitations en raison de leurs résolutions géographiques grossières, ce qui limite leur précision pour étudier les phénomènes locaux et les impacts régionaux. Pour pallier ces insuffisances, les modèles régionaux sont souvent utilisés en tandem avec les modèles globaux, offrant une résolution plus fine, généralement autour de 10 kilomètres. Ils se basent sur les données de sortie des modèles globaux, permettant ainsi une modélisation plus précise à l'échelle régionale, soit par une approche de désagrégation dynamique, soit par une désagrégation statistique. Toutefois, ces modèles régionaux doivent rester guidés par les prévisions du modèle global, ce qui nécessite un équilibre subtil pour déterminer l'influence des prévisions à plus petite échelle.

Les modèles climatiques diffèrent des modèles de prévision en ce sens qu'ils ne sont pas directement contraints par des observations. Le système climatique évolue de manière autonome, recevant de l'énergie solaire et perdant de l'énergie sous forme de rayonnement infrarouge vers l'espace. Le climat simulé résulte de l'équilibre entre ces échanges d'énergie, le forçage radiatif (défini précédemment). Par conséquent, il est essentiel de modéliser avec précision ces échanges d'énergie pour assurer le bon fonctionnement des modèles climatiques. Les modèles présentés ici, utilisés pour différentes trajectoires socio-économiques, ne tiennent compte que des variations du forçage radiatif dues à l'activité humaine, négligeant ainsi les variations naturelles de l'énergie globale du système Terre. De ce fait, les projections employées pour l'étude n'intègrent pas la variabilité naturelle de l'énergie globale terrestre.

Le DRIAS constitue donc une ressource permettant d'accéder aux projections climatiques associées aux trajectoires socio-économiques RCP évoquées précédemment. À l'intérieur de cette plateforme, dix modèles sont mobilisés pour simuler le climat futur selon le scénario RCP 4.5, qui est le scénario privilégié pour l'étude. Ces données de projection couvrent la période de 2006 à 2100, à une fréquence quotidienne, et sont disponibles sur la grille SIM2, laquelle divise le territoire métropolitain en mailles de 8 kilomètres.

Le modèle climatique retenu pour l'étude est le CNRM-ALADIN, choisi en raison de sa représentation d'un réchauffement climatique moyen par rapport à d'autres modèles. Ce modèle projette au pas de

Nom de la simulation	Institution	GCM	RCM	Scénarios	Périodes disponibles
CNRM-CERFACS-CNRM-CM5_CNRM-ALADIN63	CNRM	CNRM-CM5	**ALADIN63**	RCP8.5, RCP4.5, RCP2.6	1951-2100
MPI-M-MPI-ESM-LR_CLMcom-CCLM4-8-17	CLMcom	MPI-ESM	**CCLM4-8-17**	RCP8.5, RCP4.5, RCP2.6	1950-2100
MOHC-HadGEM2-ES_ICTP-RegCM4-6	ICTP	HadGEM2	**RegCM4-6**	RCP8.5, — , RCP2.6	1970-2099
ICHEC-EC-EARTH_SMHI-RCA4	SMHI	EC-EARTH	**RCA4**	RCP8.5, RCP4.5, RCP2.6	1970-2100
IPSL-IPSL-CM5A-MR_IPSL-WRF381P	IPSL	IPSL-CM5A	**WRF381P**	RCP8.5, RCP4.5, —	1951-2100
NCC-NorESM1-M_GERICS-REMO2015	GERICS	Nor-ESM1	**REMO2015**	RCP8.5, — , RCP2.6	1950-2100
MPI-M-MPI-ESM-LR_MPI-CSC-REMO2009	CSC	MPI-ESM	**REMO2009**	RCP8.5, RCP4.5, RCP2.6	1970-2100
MOHC-HadGEM2-ES_CLMcom-CCLM4-8-17	CLMcom	HadGEM2	**CCLM4-8-17**	RCP8.5, RCP4.5, —	1950-2099
ICHEC-EC-EARTH_KNMI-RACMO22E	KNMI	EC-EARTH	**RACMO22E**	RCP8.5, RCP4.5, RCP2.6	1950-2100
IPSL-IPSL-CM5A-MR_SMHI-RCA4	SMHI	IPSL-CM5A	**RCA4**	RCP8.5, RCP4.5, —	1970-2100
CNRM-CERFACS-CNRM-CM5_KNMI-RACMO22E	KNMI	CNRM-CM5	**RACMO22E**	RCP8.5, RCP4.5, RCP2.6	1950-2100
NCC-NorESM1-M_DMI-HIRHAM5	DMI	Nor-ESM1	**HIRHAM5 v3**	RCP8.5, RCP4.5, —	1951-2100

FIGURE 3.2 : Liste des projections climatiques disponibles sur le DRIAS

temps journalier un ensemble de sept variables climatiques, à savoir la température moyenne, minimale et maximale, l'humidité spécifique près de la surface, les précipitations totales et neigeuses, et la vitesse du vent. Le modèle CNRM-ALADIN est complété par le modèle hydrologique SIM2 de Météo-France afin d'obtenir des variables dépendantes des interactions entre le sol et l'eau. Dans le cadre de la projection des indices de sécheresse, les températures moyennes et les précipitations sont utilisées pour calculer le SPEI-3, tandis que l'indice d'humidité des sols (SWI) est employé pour évaluer l'indice de magnitude SWI. Le SWI est récupéré via le portail "eau" du système DRIAS, qui constitue un module complémentaire au site principal. Il convient de noter qu'une approche multi-modèle aurait pu être effectuée pour rendre les résultats des projections plus robustes.

Evolution spatio-temporelle des indices de sécheresses

L'évolution des indices de sécheresse est étudiée à la fois dans un contexte spatial et temporel, en comparaison avec la période historique de référence (2000-2020). Pour l'indice SPEI-3, l'indice SPEI estival présente la plus significative variation sous le scénario RCP 4.5 par rapport à la période de référence. Malgré des fluctuations, cet indice évolue de manière notable, passant d'une moyenne annuelle de 0.02 entre 2000 et 2020, puis de -0.26 entre 2030 et 2040, pour enfin atteindre une moyenne annuelle de -0.95 sur la période de 2041 à 2050.

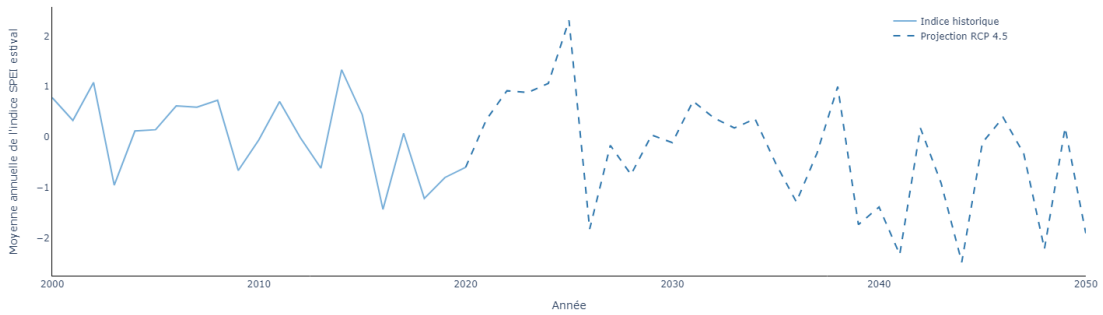
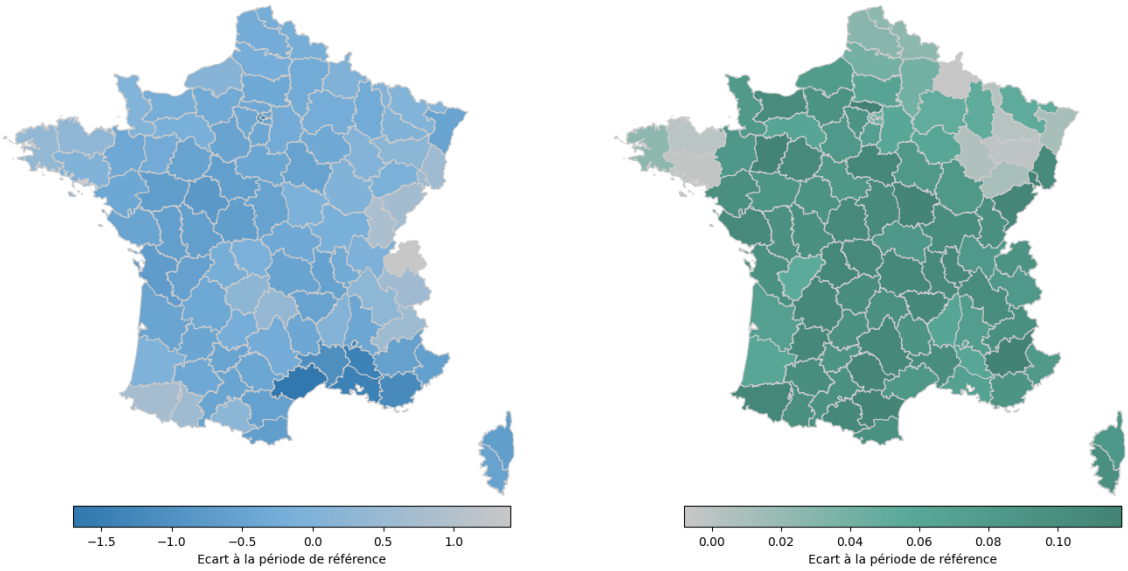


FIGURE 3.3 : Projection selon le scénario RCP 4.5 de la moyenne annuelle de l'indice SPEI estival

La Figure 3.3, qui illustre l'évolution de la moyenne annuelle de l'indice SPEI estival, témoigne de la dégradation de la valeur de cet indice, reflétant ainsi l'aggravation des conditions de sécheresse. Il convient de rappeler que l'analyse précédemment conduite sur l'indice SPEI-3, exposée dans la section 2.2, a révélé que l'indice SPEI estival était étroitement lié aux sinistres sécheresse, présentant le tau de Kendall le plus élevé. Par conséquent, cette diminution de l'indice constitue un risque tangible pour l'occurrence future de sinistres liés à la sécheresse.

D'un point de vue géographique, la carte (a) de la figure 3.4 met en évidence une détérioration du SPEI estival sur l'ensemble du territoire, à l'exception de la Bretagne et de certains territories dotés de reliefs montagneux, tels que la Savoie, les Hautes-Alpes, les Pyrénées-Atlantiques, les Hautes-Pyrénées, le Cantal, ainsi que le Jura. Les départements situés le long du bassin méditerranéen subissent une diminution sévère du SPEI, tout comme ceux de la façade atlantique et de la région Centre-Val de Loire. En définitive, à l'horizon 2050, ce sont les départements du croissant argileux qui sont particulièrement touchés par une baisse significative de l'indice SPEI estival.



(a) Écart de moyenne annuelle pour l'indice SPEI-3

(b) Écart de moyenne annuelle pour l'indice de magnitude SWI

FIGURE 3.4 : Écart de moyenne annuelle de chaque indice sécheresse entre la période de référence et la période 2041-2050

En ce qui concerne l'indice de magnitude SWI, il enregistre également une augmentation significative sous le scénario RCP 4.5. La moyenne annuelle de cet indice, qui est de 0.02 entre 2000 et 2020, double entre 2031 et 2040 pour finalement tripler et atteindre 0.066 entre 2041 et 2050. Cette progression graduelle jusqu'à la moitié du XXIe siècle est visible sur la Figure 3.5, qui trace l'évolution de la moyenne annuelle de l'indice de magnitude SWI. Compte tenu de son importance au sein du modèle prédictif de la charge sinistre sécheresse, l'accroissement prévu de l'indice de magnitude SWI sous le scénario RCP 4.5 suggère une augmentation des sinistres à horizon futur.

Sur le plan géographique, la carte (b) de la figure 3.4 révèle une augmentation globale de l'indice de magnitude SWI sur l'ensemble du territoire métropolitain à l'exception de la partie nord-est, de la Bretagne ainsi que des départements des Landes et de la Gironde.

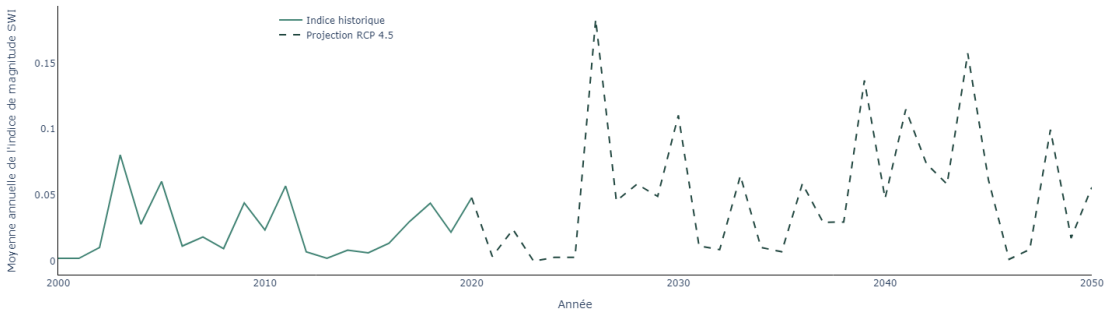


FIGURE 3.5 : Projection selon le scénario RCP 4.5 de la moyenne annuelle de l'indice de magnitude SWI

3.1.2 Projection des enjeux assurés

Cette partie est dédiée à la projection des enjeux assurés composant le portefeuille considéré. Cette projection recouvre plusieurs aspects : l'évaluation de l'accroissement du nombre de risques assurés jusqu'à l'horizon 2050, la projection de leur répartition spatiale et l'estimation de l'augmentation des valeurs assurées.

Projection démographique

Pour faire évoluer le nombre de risques assurés, les scénarios de projections démographique de l'INSEE ont été utilisés. Les projections reposent sur le modèle Omphale de l'INSEE (2022), qui se base sur les données démographiques par sexe et âge au 1er janvier 2013 issues du recensement de la population. Ce modèle effectue des projections pour toutes les zones géographiques comptant plus de 50 000 habitants. En pratique, il applique des ratios d'émigration, de fécondité, et de mortalité spécifiques à chaque sexe et âge, calculés en 2013 sur la zone concernée. Par la suite, ces ratios évoluent selon les tendances nationales, en suivant les dernières données disponibles. Trois scénarios (bas, central et haut) d'évolution de la population à l'horizon 2050 ont été envisagés, parmi lesquels le scénario "central" a été retenu. Ce dernier est caractérisé par des indicateurs de fécondité et de mortalité stables, ainsi que des taux migratoires constants sur toute la période de projection, reflétant les mouvements de population entre différentes zones géographiques. L'exploitation de ces données a permis de déterminer des taux de croissance pour chaque année et pour chaque département, qui ont ensuite été appliqués à l'ensemble du portefeuille.

La figure 3.6 montre des dynamiques territoriales hétérogènes, se caractérisant par une croissance significative de la population le long de la façade atlantique, sur le littoral méditerranéen, ainsi que dans la région Île-de-France. Les tendances illustrées par la figure 3.6 revêtent une importance particulière étant donné que de fortes augmentations sont observées notamment dans les départements situés dans le croissant argileux, tels que la Gironde, la Haute-Garonne, le Tarn-et-Garonne et l'Hérault, qui présentent une vulnérabilité élevée à la sécheresse. Entre 2020 et 2050, la population augmente de 5% dans les vingt départements les plus exposés à la sécheresse, contre seulement 1% dans le reste des départements. L'accroissement de la population dans ces zones à risque pourrait par conséquent avoir des répercussions notables sur la charge sinistre du portefeuille à l'avenir. En parallèle, certaines régions touchées par la sécheresse connaîtront une diminution du nombre de risques, notamment les départements de l'Allier, de la Nièvre, du Cher et de la Dordogne.

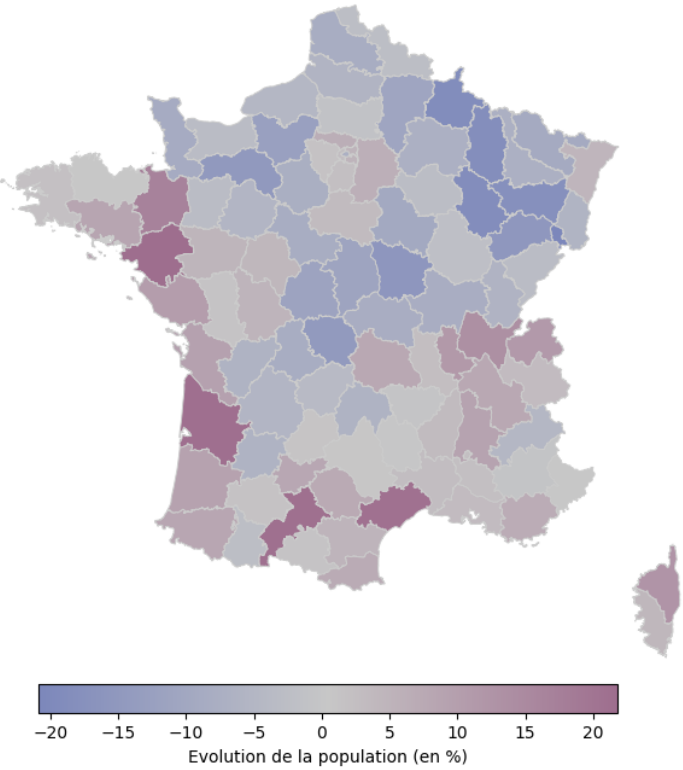


FIGURE 3.6 : Evolution de la population en % entre 2050 et l'année de référence 2020

Il est néanmoins impératif de noter que ces projections n'intègrent pas les effets de plan de prévention contre les risques climatiques physiques. La prise de conscience accrue de la population concernant ces risques ainsi que la mise en place de mesures de prévention par le gouvernement qui se profilent dans les années à venir peuvent conduire à une atténuation des évolutions démographiques au sein des zones à risques. De plus, il est concevable que les personnes qui migrent ne s'installent pas systématiquement dans les zones du département présentant une forte vulnérabilité. Par conséquent, une augmentation du nombre de risques dans un département ne signifie pas nécessairement une augmentation concomitante des sinistres liés à la sécheresse dans ce même département. Dans cette étude, l'hypothèse d'une répartition constante du nombre de logements entre les zones à risques et les zones non à risques lors des projections a été retenue, mais une approche plus précise aurait pu être considérée, prenant en compte des scénarios de concentration ou de dispersion de la population au sein d'un département.

Projection des valeurs assurées

Pour projeter les valeurs assurées à horizon 2050, deux éléments d'inflation sont pris en compte : d'une part, une composante associée à l'indice FFB du coût de la construction (ICC) présenté en 2.1, et d'autre part, une composante d'enrichissement ρ_d spécifique au département d. Ainsi pour la valeur d'un bien en 2020 dans le département d, notée $V_{d,2020}$, la relation suivante est établie

$$V_{d,2020} = V_{d,2000}\,(1 + \overline{\text{ICC}})^{20}\,(1 + \overline{\rho_d})^{20}$$

avec $\overline{\text{ICC}}$) le taux de croissance annuel moyen de l'indice FFB du coût de la construction entre 2000 et 2020 et $\overline{\rho_d}$ le taux de croissance annuel moyen de l'indice d'enrichissement du département.

La prise en compte de ces deux éléments revêt un intérêt particulier car elle permet d'isoler et de mettre en lumière les effets de l'enrichissement, induisant ainsi une augmentation plus prononcée de la valeur des biens dans certains départements du portefeuille, indépendamment de l'inflation attribuée à l'indice ICC. Étant donné que $\overline{\text{ICC}}$ est connu et que les valeurs des biens par département sont accessibles dans l'historique du portefeuille, le taux de croissance annuel moyen de l'indice d'enrichissement peut être calculé en utilisant la formule suivante

$$\overline{\rho_d} = \left(\frac{V_{d,2020}}{V_{d,2000}}\right)^{1/20} \times (1 + \overline{\text{ICC}})^{-1} - 1.$$

Ainsi, la valeur projetée du bien du département d, pour l'année $k \geq 2020$, est donnée par

$$V_{d,k} = V_{d,2020}\,(1 + \overline{\text{ICC}})^{k-2020}\,(1 + \overline{\rho_d})^{k-2020}$$

L'utilisation de cette approche suppose que les taux de croissance annuel moyen de l'ICC et de ρ_d entre 2000 et 2020 sont les même pour la période de projection.

La formule précédente permet donc de projeter les valeurs assurées de chaque département jusqu'en 2050 en prenant en considération l'indice ICC et les diverses dynamiques d'enrichissement constatées jusqu'à la période actuelle.

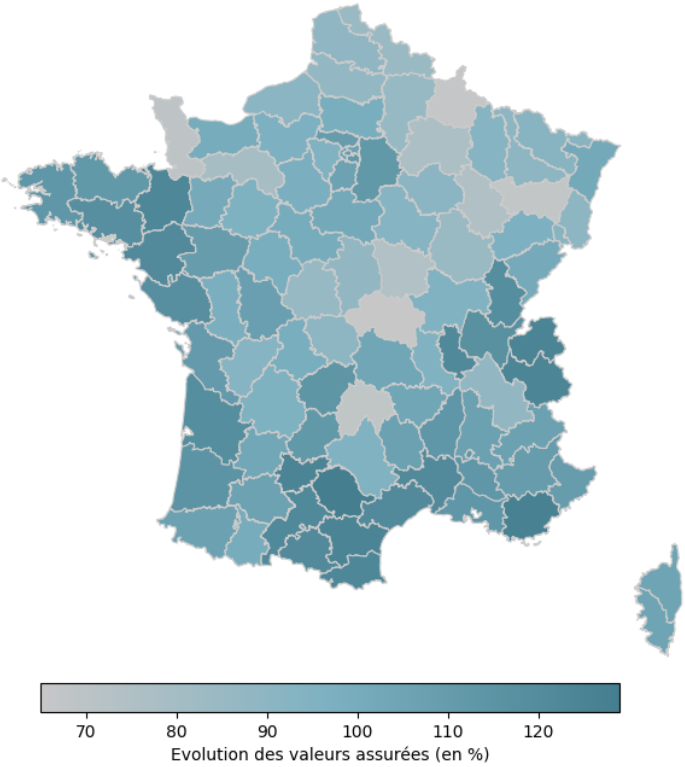


FIGURE 3.7 : Evolution des valeurs assurées en % entre 2050 et l'année de référence 2020

L'application de cette méthode de projection révèle que, dans l'ensemble, les valeurs assurées du portefeuille augmentent de manière significative, avec une croissance de 99% entre 2020 et 2050. Cependant, une analyse plus approfondie de la figure 3.7, met en évidence une nouvelle fois d'importantes disparités territoriales.

Les départements situés le long des littoraux atlantique et méditerranéen, ainsi que dans le sud-ouest, caractérisés par une croissance démographique marquée, connaissent également une augmentation importante des valeurs assurées. Plus précisément, l'évolution des valeurs assurées dans ces régions dépasse 110%, voire atteint 127% dans les départements du Var, du Tarn et de la Haute-Savoie. En revanche, les territoires confrontés à un déclin démographique dans la figure 3.6 enregistrent une augmentation des valeurs assurées, mais à des taux moins élevés. Par exemple, la région du quart nord-est connaît une hausse plus modérée.

Par ailleurs, il est à noter que les départements les plus touchés par les sinistres liés à la sécheresse enregistrent une augmentation moyenne de 105% des valeurs assurées entre 2020 et 2050, comparée à une hausse de 97% dans les autres départements.

3.2 Projection de la sinistralité

Cette section a pour objectif d'analyser la projection de la sinistralité liée à la sécheresse en considérant, dans un premier temps, uniquement l'évolution climatique selon le scénario RCP 4.5, puis dans un second temps, en prenant en compte l'augmentation des enjeux assurés. Cette distinction permet de décomposer la charge sinistre future et d'approfondir la compréhension des facteurs qui influencent l'accroissement des montants des sinistres. Il convient de souligner qu'aucune hypothèse relative à une modification du régime d'indemnisation des catastrophes naturelles n'a été incorporée dans la poursuite de cette étude. Par conséquent, la projection de la charge sinistre s'opère en considérant une réglementation constante. De plus, les projections ne tiennent pas compte de la mise en œuvre de mesures préventives ni d'une politique d'amélioration du bâti, facteurs susceptibles d'atténuer la sinistralité à moyen terme.

3.2.1 Analyse à exposition constante

Une première analyse est réalisée en maintenant l'exposition constante, ce qui signifie que la répartition géographique, la taille du portefeuille ainsi que la valeur des biens assurés demeurent inchangées. Cette approche permet de mettre en lumière exclusivement les impacts du changement climatique sur l'évolution du montant des sinistres, à travers le scénario RCP 4.5, et de comparer la charge sinistre projetée à celle de la période de référence.

Les conséquences du réchauffement climatique sur la charge sinistre associée à la sécheresse révèlent une augmentation progressive jusqu'au milieu du XXIe siècle. Cette hausse est évaluée par rapport à la période de référence (2000-2020) sur plusieurs horizons, selon la formule suivante

$$\Delta Y = \left(\frac{\overline{Y}_H}{\overline{Y}_{\text{REF}}} - 1\right) \times 100$$

où $\overline{Y}_H$ correspond à la moyenne des pertes annuelles de l'horizon futur et $\overline{Y}_{\text{REF}}$ à la moyenne des pertes annuelles de la période de référence.

À l'échelle globale du portefeuille, l'augmentation de la moyenne des pertes annuelles se chiffre à 50% sur la période allant de 2021 à 2040, puis s'accroît à 61% à l'approche du milieu du XXIe siècle, pour la période de 2041 à 2050. Ces données illustrent clairement la tendance à la hausse de la sinistralité liée à la sécheresse à l'horizon 2050 du fait du réchauffement climatique. Sous le scénario RCP 4.5, les périodes de retour des événements de sécheresse sont divisées par 2. Selon les projections, à horizon 2050, la période de retour est de 9 ans pour un événement sécheresse de type 2003, de 4 ans pour celui de 2018, et de 2 ans pour celui de 2011.

	Charge sinistre (en euros)	Période de retour - Référence	Période de retour - RCP 4.5
Sécheresse 2003	80 M	20 ans	9 ans
Sécheresse 2018	55M	10 ans	4 ans
Sécheresse 2011	40M	4 ans	2 ans

TABLE 3.1 : Comparaison des périodes de retour de plusieurs événements de sécheresse entre le climat de référence et le climat projeté sous scénario 4.5

D'un point de vue spatial, l'évolution de la charge sinistre présente de fortes inégalités. La figure 3.8 révèle que les évolutions significatives se concentrent principalement le long du croissant argileux, ainsi qu'au niveau du Bassin Parisien, du nord des Hauts-de-France, de l'Auvergne et dans les départements de la Moselle.

Les départements les plus impactés par ces évolutions à l'horizon 2050 sont la Charente, les Deux-Sèvres en Nouvelle-Aquitaine, le Maine-et-Loire et la Sarthe dans les Pays de la Loire, ainsi que le Gard en Occitanie et la Seine-et-Marne en Île-de-France. Il convient de noter que ces départements sont principalement situés dans des zones à risque RGA de niveau moyen, comme indiqué sur la carte 1.2. Cette observation reflète que les zones avec une exposition moyenne connaissent les plus importantes évolutions avec l'aggravation des conditions climatiques.

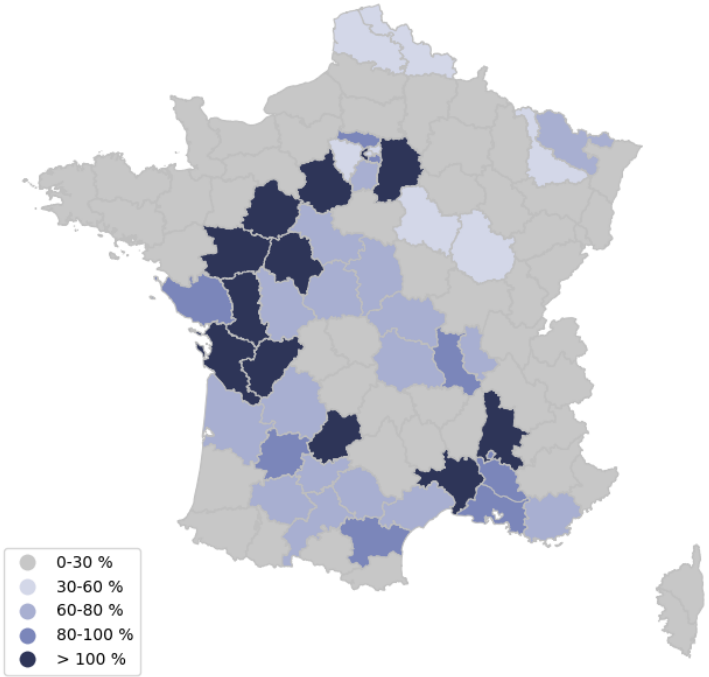


FIGURE 3.8 : Evolution de la charge sinistre annuelle moyenne entre la période de référence (2000-2020) et la période 2041-2050

Cependant les départements où les plus fortes évolutions sont attendues ne sont pas forcément ceux qui pèsent le plus dans la sinistralité future. La figure 3.9 révèle que les départements présentant la charge sinistre annuelle moyenne la plus élevée sur la période 2041-2050 sont la Haute-Garonne, le Tarn-et-Garonne, la Dordogne, le Gers, ou encore le Vaucluse et le Gard, qui à eux seuls regroupent

60% de la charge sinistre totale future.

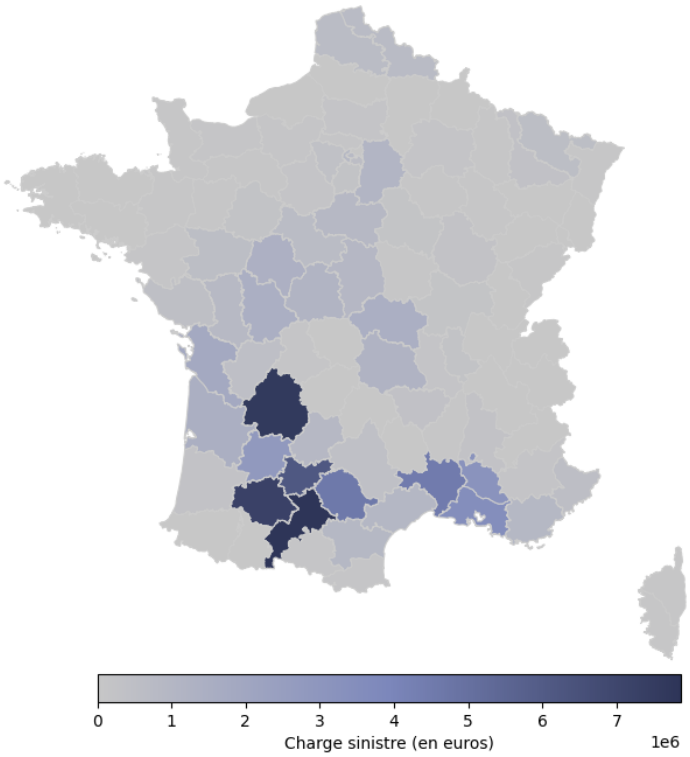


FIGURE 3.9 : Charge sinistre annuelle moyenne à climat futur (2041-2050)

3.2.2 Projection de l'exposition et décomposition de la charge sinistre

Dans cette partie, les hypothèses d'évolutions démographiques de l'INSEE sont appliquées au portefeuille ainsi que celles concernant l'augmentation de la valeur des biens assurés. Ces hypothèses sont mises en place séparément de sorte à décomposer la charge sinistre et ainsi identifier l'influence spécifique de chaque facteurs inflationniste sur l'évolution de la sinistralité sécheresse.

La décomposition de la charge sinistre annuelle moyenne entre 2041 et 2050 représentée par le graphique 3.10 indique que le principal facteur inflationniste reste l'augmentation de la valeur des biens assurés qui implique une hausse de près de 100% de la charge sinistre initiale (charge sinistre moyenne de la période 2000-2020). Ceci s'explique notamment par le graphique 3.7 qui indique une augmentation globale des valeurs assurées sur tout le territoire, marquée par des évolutions fortes dans les zones à risque, notamment au niveau du bassin méditérranéen et de la façade atlantique.

L'effet du changement climatique arrive en deuxième position avec un impact conséquent sur l'augmentation de la charge sinistre. A lui seul, il provoque une croissance de plus de 60% de la charge sinistre initiale.En revanche, la répartition démographique exerce un effet relativement modéré, entraînant une hausse de seulement 10% de la charge moyenne de référence.

En excluant l'influence de l'augmentation des valeurs assurées, l'accroissement de la sinistralité engendrée par le changement climatique et l'évolution démographique se traduit par une augmentation

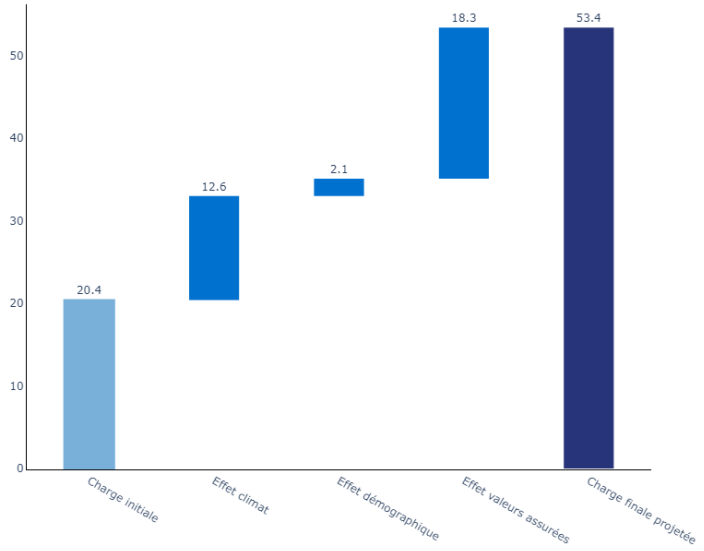


FIGURE 3.10 : Décomposition de la charge sinistre annuelle moyenne pour la période 2041-2050

de 72% de la perte annuelle moyenne entre la période de référence et la période de projection 2041-2050. Ce résultat s'aligne notamment avec les précédentes études menées par COVÉA (2022) et la CCR (2018) qui prévoyaient respectivement 60% et 72% d'augmentation de la charge sinistre sécheresse à horizon 2050. Dans une dernière étude parue en octobre 2023, la CCR (2023a) estimait l'évolution des dommages relatifs à la sécheresse à 59% sous le scénario RCP 4.5.

3.3 Soutenabilité du risque sécheresse à climat futur

L'augmentation de la sinistralité détaillée dans la partie précédente amène les assureurs à ajuster les primes d'assurances. Dans le cadre du risque sécheresse qui affecte le bâti, c'est la prime MRH, qui comprend la garantie Cat Nat, qui est concernée. L'objectif de cette partie consiste à projeter l'évolution des primes proposées par l'assureur, issue de l'augmentation des sinistres sécheresse, et à questionner la capacité des assurés à supporter l'accroissement des primes MRH selon divers scénarios de tarification.

3.3.1 Notion de soutenabilité

Base FILOSOFI

Pour introduire la notion de supportabilié de la prime, le revenu médian départemental est utilisé. Les données sur le revenu médian sont obtenues à partir du dispositif FiLoSoFi de l'INSEE (2020) qui regroupe un ensemble d'indicateurs sur la distribution des revenus disponibles à l'échelle communale, supra-communale et infra-communale. Ces bases sont mises à jour chaque année depuis 2014 et remplacent les anciens dispositifs Revenus Fiscaux Localisés (RFL) et Revenus Disponibles Localisés (RDL).

Dans le cadre de l'étude, le revenu médian est récupéré à la maille IRIS (Ilots Regroupés pour l'Information Statistique), qui constitue la plus fine résolution géographique des bases à disposition,

puis agrégé à l'échelle départementale. Ceci permet d'obtenir des données sur les revenus médians par département pour l'année de référence 2020.

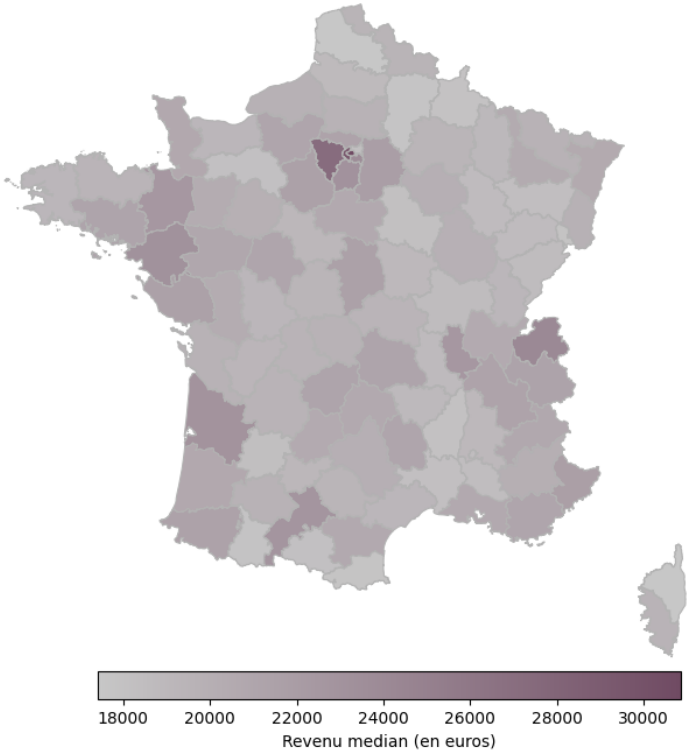


FIGURE 3.11 : Revenu median départemental (en euros) en 2020

La figure 3.11 met en évidence une distribution contrastée des niveaux de revenus. Les départements présentant les revenus les plus élevés englobent les Hauts-de-Seine, Paris, les Yvelines et la Haute-Savoie. Ils sont suivis par la plupart des départements abritant des capitales régionales, ainsi que par ceux localisés dans la région parisienne, le long de la façade atlantique et le long des frontières suisse et italienne. En revanche, les revenus médians sont considérablement plus bas en Seine-Saint-Denis, et de manière moins marquée dans les départements de l'Aude, des Pyrénées-Orientales, du Pas-de-Calais, de la Creuse, de l'Aisne, des Ardennes ou de l'Ariège.

Les données historiques des revenus provenant des bases FILOSOFI sont également employées pour calculer le taux d'évolution annuel moyen pour chaque département. Cette démarche permet de discerner les tendances d'accroissement des revenus à travers le territoire métropolitain, en vue de projeter le revenu médian par département à l'horizon 2050. Cette projection sert notamment à confronter, dans la section suivante, le niveau de richesse et l'exposition au risque de sécheresse.

Risque sécheresse et richesse

A partir des revenus médians projetés en 2050, il est possible d'étudier le potentiel lien entre la richesse d'un département et le niveau de risque sécheresse auquel il est exposé à horizon futur. Pour cela, en se basant sur les projections de sinistralité sécheresse à horizon 2050, les départements sont initialement regroupés selon trois groupes de catégories de risques : Non risqué (58%), risqué (17%)

et très risqué (25%). De manière analogue, les quartiles de la distribution des revenus projetés sont utilisés pour créer quatre niveaux de richesse (notées Q1, Q2, Q3 et Q4) s'échelonnant de Q1 (le premier quartile, avec un revenu médian inférieur à 21 221 euros) à Q4 (le dernier quartile, supérieur 26 068 euros).

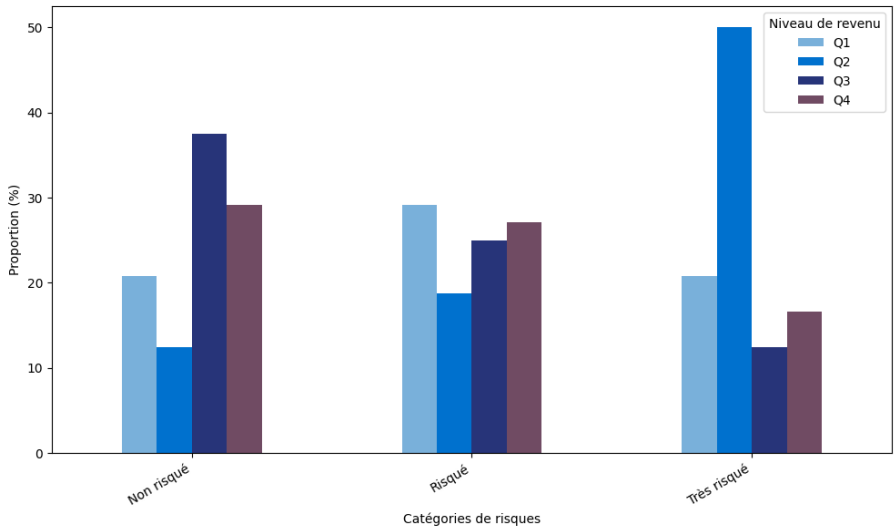


FIGURE 3.12 : Distribution des départements en fonction de leur catégorie de risque et de leur niveau de richesse à horizon 2050

La répartition des départements en fonction de leur catégorie de risque et de leur niveau de richesse, illustrée dans la figure 3.12, met en évidence que les départements à sinistralité élevée sont associés à des revenus médians plus modestes. Concernant les départements classés comme "très risqués", ils appartiennent majoritairement aux niveaux de richesse Q1 (20%) et Q2 (49%), qui représentent les niveaux de revenus les plus bas. En revanche, parmi les départements non risqués, ce sont principalement ceux affichant une richesse plus élevée qui prédominent.

Un test d'indépendance du khi-deux est également effectué entre le niveau de richesse et les catégories de risque. Ce dernier présente une p-valeur de 2×10^{-12} qui rend difficile l'acceptation de l'hypothèse d'indépendance (H_0) et amène par conséquent une preuve statistique supplémentaire d'un lien entre le revenu médian d'un département et le risque sécheresse.

Ce constat, selon lequel la plupart des départements fortement impactés par le risque sécheresse présentent des niveaux de revenus plus bas, suscite légitimement des interrogations quant à la viabilité de la prime MRH pour ces populations, particulièrement si la tarification du produit suit de près l'accroissement du risque.

Mesurer la soutenabilité

Afin de quantifier la notion de soutenabilité des assurés face à la prime MRH proposée par l'assureur, le ratio de soutenabilité défini par département et par année est introduit

$$\kappa_{d,n} = \frac{\pi_{d,n}}{R_{d,n}} \times 100$$

avec $\pi_{d,n}$ la prime MRH moyenne payée par le département d l'année n et $R_{d,n}$ le revenu médian

du département associé pour l'année considérée. Le revenu médian est ici préféré au revenu moyen, en raison de sa robustesse aux valeurs extrêmes de la variable considérée.

Dans la suite de l'étude, le ratio $\kappa_{d,n}$ est utilisé pour appréhender la déviation de soutenabilité d'un département entre l'année de référence 2020 et l'horizon 2050 avec la formule

$$\Delta\kappa_d = (\kappa_{d,2050} - \kappa_{d,2020}) \times 100.$$

La quantité $\Delta\kappa_d$, exprimée en points de base, permet ainsi de mettre en évidence les départements fortement exposés au risque sécheresse dont l'augmentation significative de la prime MRH conduit à un risque d'insoutenabilité chez les assurés. En effet, une déviation positive du ratio indique que la prime du département augmente plus vite que le revenu au cours de la période. La question de la tarification de la prime à horizon futur est abordée au travers des méthodes de crédibilité dont la théorie est exposée dans la partie suivante.

3.3.2 Théorie de la crédibilité

Formalisme général

La théorie de la crédibilité s'appuie sur les travaux de Whitney (1918) , qui suggère dans une perspective de tarification équitable la nécessité de pondérer d'un côté l'expérience collective et de l'autre l'expérience individuelle. Pour cela, il introduit la prime individuelle

$$\hat{\pi}_i = zY + (1 - z)m$$

où Y correspond à l'expérience individuelle (reposant sur l'historique des sinistres propres à l'assuré) et m l'expérience collective (se basant sur l'historique des sinistres de tous les assurés du portefeuille), pondérées par un facteur de crédibilité z de la forme

$$z = \frac{n}{n + K}$$

avec n la taille de l'expérience individuelle et K une constante qui dépend du paramètre du modèle. Une prime de cette forme est appelée prime de crédibilité.

Cependant cette première approche présentait l'inconvénient d'avoir un facteur de crédibilité ne dépendant que de la taille des expériences individuelles, conduisant ainsi à une estimation peu satisfaisante de la prime. Par la suite, Bailey (1950) ou encore Bühlmann (1967) ont apporté des améliorations à ces travaux préliminaires, contribuant ainsi à l'expansion des méthodes de la crédibilité.

D'une manière plus générale, le but de la crédibilité est de calculer la "meilleure" prime individuelle $\hat{\pi}_{i,n+1}$ étant donné le niveau de risque du contrat. Pour faire intervenir la notion d'individualité expliquée précédemment, la variable aléatoire Θ est introduite pour modéliser les profils de risques des assurés θ_i. Si le risque est connu, alors la meilleure estimation de la prime de risque (au sens des moindres carrés) est

$$\mu(\theta_i) = \mathbb{E}\left[Y_{i,n+1} \mid \Theta = \theta_i\right]$$

.

Une première estimation de la prime de risque peut être la moyenne pondérée de toutes les primes de risques possibles $\mathbb{E}[\mu(\Theta)] = m$. Dès lors, la prime calculée est la même pour tous les contrats et

correspond à la prime collective. Cette dernière est adéquate mais n'est pas nécessairement équitable au regard des différents profils de risque.

En pratique, le profil de risque n'est pas connu de l'assureur mais ce dernier dispose de données historiques de sinistres $Y_{i,1}, \ldots, Y_{i,n}$. Lorsque ces données sont disponible, la meilleure approximation de la prime $\mu(\theta_i)$ est la fonction des observations $g(Y_{i,1}, \ldots, Y_{i,n})$ minimisant l'erreur quadratique moyenne

$$g^* = \arg\min_g \left\{ \mathbb{E}\left[\mu(\Theta) - g(Y_{i,1}, \ldots, Y_{i,n})\right]^2 \right\}.$$

La fonction $g^*(Y_{i,1}, \ldots, Y_{i,n})$ est appelée la prime de Bayes

$$B_{i,n+1} = \mathbb{E}\left[\mu(\Theta) \mid (Y_{i,1}, \ldots, Y_{i,n})\right].$$

Modèle de crébibilité de Bulhman et Bulhman Straub

La prime de Bayes $B_{i,n+1}$ est donc le meilleur estimateur possible de la prime individuelle $\mu(\theta_i)$. Cependant, dans la majorité des cas, elle ne peut s'exprimer de façon analytique. La théorie de crédibilité de Bühlmann (1967) consiste à restreindre l'estimation de la prime aux seules fonctions linéaires en les observations. Les hypothèses du modèles sont

- Chaque contrat $i \in [\![1, I]\!]$ du portefeuille est représenté par un vecteur aléatoire $(\Theta_i, Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n})$. Ces vecteurs sont supposés indépendants et identiquement distribués.
- Conditionnellement à $\Theta_i = \theta_i$, les variables $(Y_{i,k})_{k=1,\ldots,n}$ sont indépendantes et identiquement distribuées.

En considérant $\overline{Y_i} = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{i,k}$, la prime est alors obtenue en résolvant le problème de minimisation

$$\min_{\alpha_1,\alpha_2} \left\{ \mathbb{E}\left[(\mu(\Theta) - \alpha_1 - \alpha_2 \overline{Y_i})^2\right] \right\}.$$

La minimisation conduit à $\alpha_1^* = (1 - \alpha_2^*)\,\mathbb{E}[\mu(\Theta)] = (1 - \alpha_2^*)\,m$ et

$$\alpha_2^* = \frac{n}{n + s^2/a}$$

avec $s^2 = \mathbb{E}[\sigma^2(\Theta)] = \mathrm{Var}[Y_{i,n} \mid \Theta]$ et $a = \mathrm{Var}[\mu(\Theta)]$.

En posant $z := \alpha_2^*$, la prime de Bulhman pour l'année n+1 est obtenue sous la forme d'une prime de crédibilité

$$\hat{\pi}^B_{i,n+1} = z\overline{Y_i} + (1 - z)m.$$

Les quantités m, s^2 et a sont appelées paramètres de structure puisqu'ils caractérisent la structure interne du portefeuille. Le facteur de crédibilité z dépend positivement du paramètre a, lequel reflète l'hétérogénéité du portefeuille (la dispersion inter-assuré). Lorsque s^2 augmente, c'est à dire les fluctuations au sein du portefeuille (les dispersions intra-assuré) augmente, z diminue. Une grande valeur de s^2 traduit de fortes différences de sinistralité mais principalement dues au hasard et non à l'hétérogénéité du portefeuille. De plus, lorsque la période d'observation, notée n, croît, la prime collective m a moins de sens car il devient possible de tarifer le risque individuel en se basant sur

un historique de sinistres suffisamment étendu. En pratique, les paramètres de structure sont estimés empiriquement à partir des données du portefeuille.

Une généralisation du modèle de Buhlmann est introduite par Bühlmann et Straub (1970), de sorte à tenir compte de l'exposition au risque des assurés. Pour ce faire, une pondération $w_{i,t}$ est attribuée à chaque observation. L'hypothèse selon laquelle les observations sont (conditionnellement) indépendantes et identiquement distribuées est remplacée par l'hypothèse que pour tout $s, t = 1, \dots, n$,

$$\begin{aligned} \mathbb{E}\left[N_{it} \mid \Theta_i\right] &= \mu\left(\Theta_i\right) \\ \operatorname{Cov}\left(Y_{i,s}, Y_{i,t} \mid \Theta_i\right) &= \begin{cases} \frac{\sigma^2(\Theta_i)}{w_{it}}, & s = t \\ 0, & s \neq t \end{cases}. \end{aligned}$$

Il est important de noter que pour s'assurer que la relation soit vérifiée, les observations $Y_{i,.}$ doivent être homogènes et correspondent souvent à des ratios. En pratique, les observations employées dans le modèle de Buhlmann-Straub sont des montants de sinistres divisés par le volume. La meilleure estimation linéaire de la prime de risque est la prime de crédibilité de Buhlman-Straub

$$\hat{\pi}_{i,n+1}^{BS} = z_i Y_{iw} + (1 - z_i)m$$

avec $z_i = \frac{w_i}{w_i + s^2/a}$ et $Y_{iw} = \sum_{t=1}^{n} \frac{w_{it}}{w_{i\Sigma}} N_{it}$, où $w_{i\Sigma} = \sum_{t=1}^{n} w_{it}$. La formule de la prime de crédibilité précédente permet de mettre en évidence que le modèle de Buhlman est un cas particulier du modèle de Buhlman-Straub où $w_{i,t} = 1$. De manière analogue au modèle simple de Bulhman, les paramètres de structures sont estimés de façon empirique.

Modèle hiérarchique de Jewell

Une autre généralisation des deux premiers modèles précédents est apportée par Jewell (1975) qui propose le concept de crédibilité hiérarchique. Dans le cadre de ce mémoire, les aspects théoriques d'un modèle de crédibilité hiérarchique à deux niveaux sont présentés. Le principe de ce dernier consiste à diviser un portefeuille de départ en sous-portefeuilles, eux même divisés en cohortes.

Les données de sinistres sont alors notées Y_{ijn} où l'indice $i \in [\![1, I]\!]$ identifie le sous portefeuille, l'indice $j \in [\![1, J]\!]$ identifie la cohorte au sein du sous portefeuille et enfin l'indice n_{ij} correspond à la période d'observation. A l'instar du modèle de Buhlman-Straub, un poids w_{ijn} est attribué aux observations. L'hétérogénéité du portefeuille est alors modélisé à l'aide des variables aléatoires Φ_i et Θ_{ij} représentant respectivement les niveaux de risques des sous-portefeuilles et des cohortes. Les hypothèses du modèle sont

- Les variables aléatoires $\Phi_1, \dots, \Phi_I$ sont i.i.d..
- Les variables aléatoires $\Theta_{i1}, \dots, \Theta_{i,J_i}$ sont conditionnellement i.i.d. sachant $\Phi_i, i = 1, \dots, I$.
- Les variables aléatoires $Y_{ij1}, \dots, Y_{ijn_{ij}}$ sont conditionnellement i.i.d. sachant Θ_{ij} (et Φ_i) , $j = 1, \dots, I_j$. Leur variance est finie.
- Pour tout $s, t = 1, \dots, n_{ij}$,

$$\begin{aligned} \mathbb{E}\left[Y_{ijt} \mid \Theta_{ij}, \Phi_i\right] &= \mu\left(\Theta_{ij}, \Phi_i\right) \\ \operatorname{Cov}\left(Y_{ijs}, N_{ijt} \mid \Theta_{ij}, \Phi_i\right) &= \begin{cases} \frac{\sigma^2(\Theta_{ij}, \Phi_i)}{w_{ijt}}, & s = t \\ 0, & s \neq t \end{cases} \end{aligned}$$

Le modèle de crédibilité hiérarchique comprend quatre paramètres de structure qui sont

- la moyenne collective : $m = \mathbb{E}[\mathbb{E}[\mu(\Theta_{ij}, \Phi_i)] \mid \Phi_i]$,
- la variance intra-cohorte moyenne : $s^2 = \mathbb{E}[\mathbb{E}[\sigma^2(\Theta_{ij}, \Phi_i)] \mid \Phi_i]$,
- la variance inter-cohorte (ou intra-sous-portefeuile) : $a = \mathbb{E}[\mathrm{Var}[\mu(\Theta_{ij}, \Phi_i)] \mid \Phi_i]$,
- la variance inter-sous-portefeuille : $b = \mathrm{Var}[\mathbb{E}[\mu(\Theta_{ij}, \Phi_i)] \mid \Phi_i]$.

L'objectif est donc d'estimer les primes de risque $\mu(\Theta_{ij}, \Phi_i)$. Dans ce type de modèle, la crédibilité de la cohorte est sous la même forme que dans les modèles précédents mais elle est cette fois-ci complétée par la prime de crédibilité du sous-portefeuille. Cette dernière correspond à la moyenne pondérée de l'expérience du sous-portefeuille et de celle du portefeuille. Ainsi, la prime de risque est donnée par les équations récursives

$$\begin{aligned}\hat{\pi}_{ij}^H &= z_{ij} Y_{ijw} + (1 - z_{ij})\hat{\pi}_i^H, \\ \hat{\pi}_i^H &= z_i Y_{izw} + (1 - z_i)\, m\end{aligned}$$

avec les facteurs de crédibilité

$$\begin{aligned} z_{ij} &= \frac{w_{ij.}}{w_{ij.} + s^2/a}, \quad & w_{ij.} &= \sum_{n=1}^{n_{ij}} w_{ijn} \\ z_i &= \frac{z_{i.}}{z_{i.} + a/b}, & z_{i.} &= \sum_{j=1}^{J_i} z_{ij} \end{aligned}$$

et les moyennes pondérées

$$\begin{aligned} Y_{ijw} &= \sum_{n=1}^{n_{ij}} \frac{w_{ijn}}{w_{ij.}} Y_{ijn}, \\ Y_{izw} &= \sum_{j=1}^{J_i} \frac{z_{ij}}{z_{i.}} Y_{ijw}. \end{aligned}$$

3.3.3 Etude de différents scénario de tarification à climat futur

Dans cette partie, plusieurs scénarios de tarification de la prime MRH moyenne par département sont étudiés par le biais de modèles de crédibilité hiérarchique, exposés précédemment. Pour l'étude, l'hypothèse que les sinistres hors sécheresse couverts par la garantie MRH évoluent au taux moyen annuel de l'indice ICC ($\overline{\mathrm{ICC}}$) est faite. Concernant les sinistres relatifs à la sécheresse, ces derniers sont estimés à l'aide du modèle élaboré dans le chapitre 2.

Tarification actuariellement juste / basée sur le risque

Pour ce premier scénario de tarification, le modèle de crédibilité hiérarchique de Jewell, présenté précédemment en section 3.3.2, est utilisé pour déterminer la prime MRH moyenne de chaque département à horizon 2050. Comme illustré dans la figure 3.13, lors de l'application du modèle, les sous portefeuilles correspondent aux régions de la France métropolitaine et les cohortes aux départements.

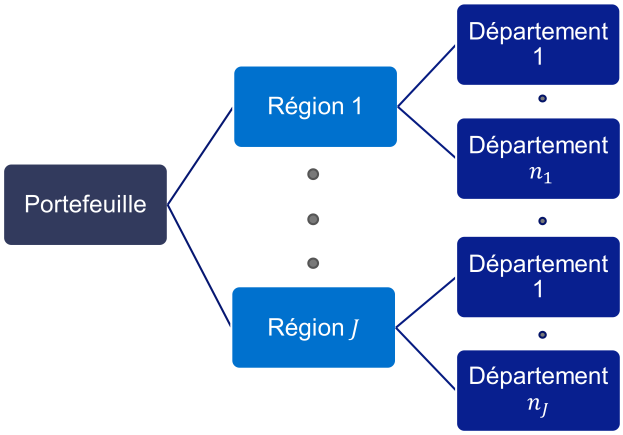


FIGURE 3.13 : Segmentation hiérarchique du portefeuille en deux niveaux

Pour les poids w_{ijn}, les valeurs assurées des départements sont employées. Afin d'obtenir la prime hiérarchique, il est primordial de calculer l'ensemble des paramètres de structure du modèle (m, s^2, a et b). Le premier paramètre de structure nécessaire pour le calcul de la prime est la moyenne collective m, estimé par

$$\hat{m} = Y_{zzw} = \sum_{i=1}^{96} \frac{z_i}{z_.} Y_{izw}.$$

L'obtention des facteurs de crédibilité passe tout d'abord par le calcul la variance intra-cohorte moyenne s^2 estimée par la formule

$$\hat{s}^2 = \frac{1}{\sum_{i=1}^{96}\sum_{j=1}^{J_i}(n_{ij}-1)} \sum_{i=1}^{96}\sum_{j=1}^{J_i}\sum_{n=1}^{n_{ij}} w_{ijn}(Y_{ijn}-Y_{ijw})^2$$

En ce qui concerne les variances inter-cohorte a et inter-sous-portefeuille b, leur estimation repose sur les quantités suivantes

$$A_i = \sum_{j=1}^{J_i} w_{ijn}(Y_{ijw}-Y_{iww})^2 - (J_i-1)s^2, \quad c_i = w_{i..} - \sum_{j=1}^{J_i}\frac{w_{ij}^2}{w_{i..}}$$
$$B = \sum_{i=1}^{96} z_i(Y_{izw}-\overline{Y}_{zzw})^2 - (96-1)a, \qquad d = z_{..} - \sum_{i=1}^{96}\frac{z_i^2}{z_{..}}$$

avec

$$\overline{Y}_{zzw} = \sum_{i=1}^{96}\frac{z_{i.}}{z_{..}} Y_{izw}.$$

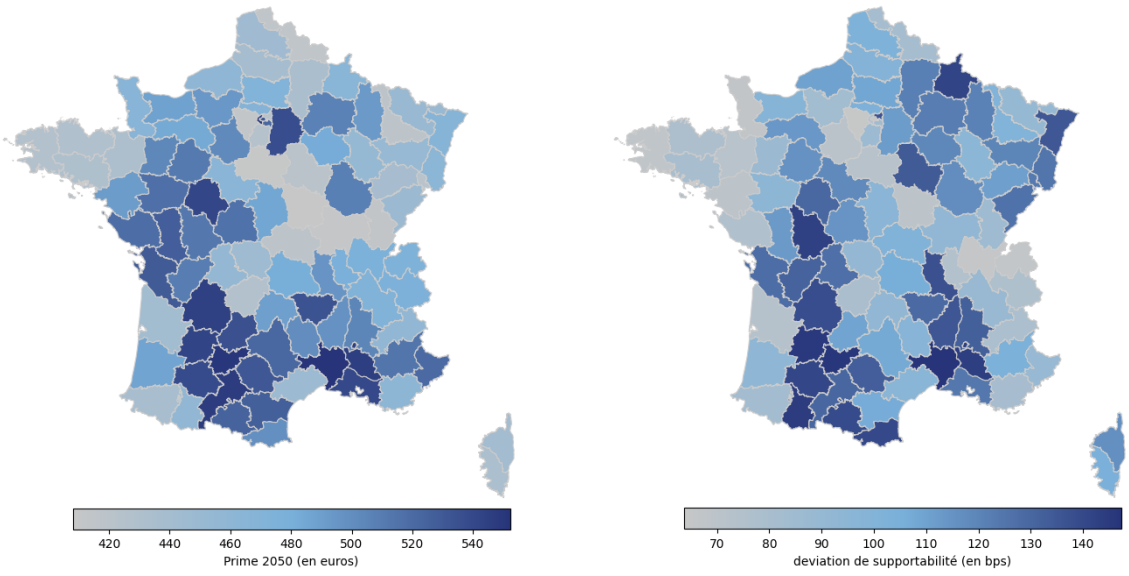
Puis les estimateurs retenus pour l'étude sont les estimateurs itératifs

$$\hat{a} = \frac{1}{\sum_{i=1}^{96}(J_i - 1)} \sum_{i=1}^{96} \sum_{j=1}^{J_i} z_{ij}(Y_{ijw} - Y_{izw})^2$$

$$\hat{b} = \frac{1}{96 - 1} \sum_{i=1}^{96} z_i(Y_{izw} - Y_{zzw})^2.$$

L'utilisation du modèle permet d'apprécier l'évolution des primes causée par l'augmentation des sinistres liés à la sécheresse en fournissant une prime actuariellement "juste" au sens de la théorie de la crédibilité. La prime payée par le département est ainsi représentative de son influence sur la sinistralité du portefeuille. Les premiers résultats montrent une augmentation de la prime moyenne du portefeuille de 82% entre l'année de référence 2020 et 2050 qui s'accompagne d'un accroissement du ratio de soutenabilité moyen passant de 1.01% en 2020 à 1.89% en 2050.

Ces évolutions globales cachent une forte disparité au sein du portefeuille. Face à des niveaux de sinistralité hétérogènes sur le territoire métropolitain, l'application de la méthode de crédibilité hiérarchique fait apparaître l'explosion de certaines primes départementale et par conséquent des déviations importantes du ratio de soutenabilité ($\Delta\kappa$). La figure 3.14 montre des évolutions de primes très élevées dans le département du Gard, du Vaucluse, du Tarn et Garonne, du Lot-et-Garonne, mais également le Gers qui s'accompagne d'une forte déviation de soutenabilité.



(a) Primes hiérarchiques estimées en 2050 (b) Deviation de soutenabilité $\Delta\kappa$ en 2050

FIGURE 3.14 : Représentation spatiale des primes et des déviations de soutenabilité en 2050 selon le modèle hiérarchique

Les département qui connaissent les plus fortes déviations de soutenabilité à horizon 2050 du fait de l'évolution des sinistres sécheresse sont détaillés dans le tableau 3.2. Ce dernier montre que la question de la soutenabilité touche principalement les régions du bassin méditerranéen ainsi que la nouvelle région Aquitaine où de nombreux départements sont exposés au risque d'insoutenabilité sous l'effet combiné d'un faible niveau de revenu et d'une sinistralité forte en 2050.

Departement	**Région**	π^H_{2050}	κ_{2050}	$\Delta\kappa$
Gard	Occitanie	552 (+ 132%)	2.56	147
Tarn-et-Garonne	Occitanie	523 (+ 109%)	2.34	130
Lot-et-Garonne	Nouvelle-Aquitaine	484 (+ 98%)	2.27	118
Vaucluse	PACA	493 (+ 86%)	2.19	111
Gers	Occitanie	481 (+ 85%)	2.10	108

TABLE 3.2 : Primes et métriques de soutenabilité obtenues à horizon 2050 après application du modèle hiérarchique

Ainsi, une tarification actuariellement juste du risque sécheresse à horizon 2050 soulève donc des questions de soutenabilité dans certains territoire où les revenus augmentent significativement moins vite que la prime. Pour apprécier la fragilité de ces départements exposés au risque d'insoutenabilité, la tarification de la prime MRH est abordée dans la section suivante à travers différent niveau de solidarité.

Tarification avec solidarité paramétrique

Dans cette partie, le modèle hiérarchique pour la tarification à horizon 2050 est conservé et s'appuie sur l'article de CHARPENTIER et al. (2022a) en introduisant les paramètres $(\alpha, \beta) \in [0,1]^2$ qui correspondent au niveau de solidarité respectivement à l'échelle nationale et à l'échelle régionale. La prime de crédibilité du département d pour l'année 2050 s'écrit finalement

$$\pi^H_{d,2050}(\alpha,\beta) = (1-\alpha)\pi_{N,2050} + \alpha[(1-\beta)\pi_{R,2050} + \beta\pi_{D,2050}] \tag{3.1}$$

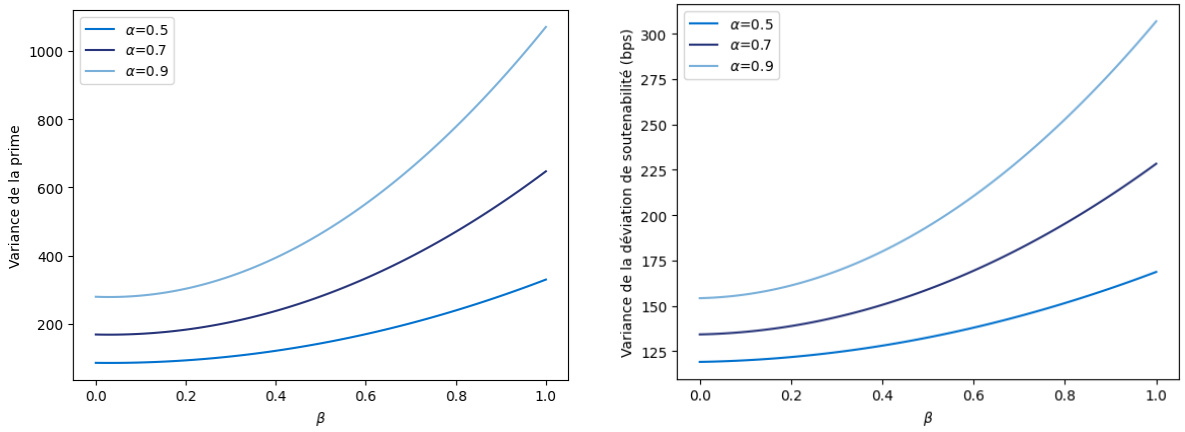
avec $\pi_{N,2050}$ la prime collective (nationale) du portefeuille, $\pi_{R,2050}$ la prime régionale et $\pi_{D,2050}$ la prime départementale en 2050. La prime collective ainsi que les primes régionales et départementales ont été calculées de sorte à respecter l'hypothèse d'un maintien du *loss ratio* cible de l'assureur à horizon 2050.

L'intérêt de paramétriser la prime selon différents niveaux de solidarité permet d'étudier la sensibilité de certains territoires aux effets de segmentation et de mutualisation dans le but d'identifier des risques d'insoutenabilité au sein du portefeuille. Dans la formule 3.1, le coefficient $(1-\alpha)$ correspond à la part de solidarité nationale, $\alpha\,(1-\beta)$ à la part de solidarité régionale et $\alpha\,\beta$ à la part de solidarité départementale.

De manière assez intuitive, la figure 3.15 montre que la variance de la prime et de la déviation du ratio de soutenabilité est une fonction croissante de α et de β. Une valeur élevée de α et de β traduit une faible solidarité au niveau national (entre les régions) et au niveau régional (entre les départements) ce qui implique une segmentation plus fine du risque et par conséquent une prime plus élevée pour les départements à risque et une prime plus faible pour les départements les moins touchés.

Plutôt que de regarder la dispersion des primes et du ratio de soutenabilité, il est judicieux, dans une logique d'étude de soutenabilité , de s'intéresser aux valeurs des quantiles élevés de la distribution des primes. Se concentrer sur les valeurs extrêmes de la distribution fait sens car une prime trop élevée dans un département peut conduire à un risque d'insoutenabilité.

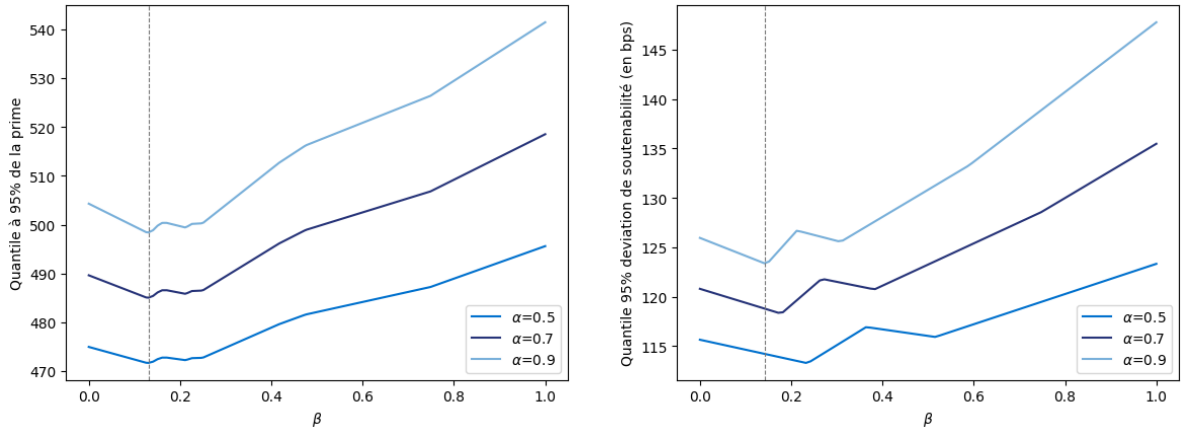
Pour cela, la figure 3.16 trace le quantile à 95% de la distribution de $\pi^H_{2050}(\alpha,\beta)$ et de $\Delta\kappa$ en fonction de β et plusieurs α. Cette dernière montre que pour les deux distributions, le quantile est bien croisant en α mais n'est pas monotone en β. Pour les primes, le quantile à 95% est en effet



(a) Variance des primes MRH (b) variance des deviations de soutenabilité

FIGURE 3.15 : Evolution de la variance de la prime MRH moyenne départementale π^H_{2050} et de la déviation de soutenabilité en fonction de β et de plusieurs α.

minimal lorsque $\beta = 0.14$ peu importe le α. Pour la déviation de soutenabilité, le quantile à 95% est minimal en 0.14 pour $\alpha = 0.9$, puis obtient des valeurs plus élevés quand la solidarité nationale ou inter-région est importante.



(a) Quantile à 95% des primes MRH (b) Quantile à 95% des déviations de soutenabilité

FIGURE 3.16 : Evolution du quantile à 95% de la prime MRH moyenne départementale π^H_{2050} et de la déviation de soutenabilité en fonction de β et de plusieurs α.

Dans le cas $(\alpha, \beta) = (0.9, 0)$, l'ensemble des départements d'une même région payent la même prime, car $\beta = 0$ et il n'y a donc aucun partage du coût du risque entre les départements (*cf.* carte (a) de la figure 3.17). En présence d'hétérogénéité du risque sécheresse au sein des régions, cette configuration conduit donc à faire payer une prime injustement élevée pour des départements qui ne sont pourtant pas sinistrés. Lorsque le coefficient β varie entre 0 et 0.14, la prime hiérarchique se détache légèrement de la prime régionale permettant au département épargnés par la sécheresse de baisser leur prime. Dans un même temps, les départements qui pèsent lourdement sur la sinistralité de la région voient leur prime légèrement augmenter mais pas suffisamment pour croître la valeur du quantile (*cf.* carte (c) de la figure 3.17). Pour $\beta \geq 0.14$, la prime hiérarchique tend vers la prime individuelle (départementale) qui fait naturellement exploser la valeur du quantile (*cf.* carte (e) de la

figure 3.17). L'optimum obtenu pour $\beta = 0.14$ permet donc de se rapprocher d'une prime équitable (dans le sens où la prime payée converge vers la prime départementale, fidèle au risque) tout en minimisant le quantile à 95%. En conclusion, le niveau de partage optimal inter-département à 0.14 fait baisser la prime des départements qui payent trop cher sans faire exploser celles de ceux fortement exposés. Un raisonnement analogue peut être utilisé pour justifier la non-monotonie en β du quantile de la distribution des déviations de soutenabilité.

Les figures 3.16 et 3.17 témoignent de la fragilité des départements fortement touchés par la sinistralité sécheresse à horizon 2050. En effet, le cas pathologique où $\beta = 1$, caractérisé par une absence totale de solidarité, conduit à une augmentation drastique de la prime MRH pour de nombreux départements en Occitanie, en Nouvelle Aquitaine, dans la région PACA mais également dans les Pays de la Loire. Pour poursuivre l'étude et appuyer cet argument d'insoutenabilité, il peut être intéressant d'analyser comment se décompose l'impact de la solidarité sur la prime en fonction des niveaux de revenus définis en section 3.3.1.

Pour décomposer l'effet de la solidarité sur les primes, le graphique (a) de la figure 3.18 trace la moyenne des différences entre la prime $\pi(\alpha, \beta)$, payée par le département en présence de solidarité, et la prime actuariellement "juste" π^H qu'il serait censé payer, basée sur son risque et calculée dans la section 3.3.3. Cette moyenne des différences, notée $\overline{\Delta}_i$ est calculée pour plusieurs niveaux de solidarité avec $\alpha = 0.9$ et $\beta \in [0, 1]$ ($\beta = 0$: solidarité maximale inter-département, $\beta = 1$: solidarité minimale inter-département) et pour chaque niveau de revenu (Q1, Q2, Q3 et Q4, introduits en section 3.3.1) avec la formule

$$\overline{\Delta}_i(\alpha, \beta) = \frac{1}{|D_i|} \sum_{d \in D_i} (\pi_d^H(\alpha, \beta) - \pi_d^H)$$

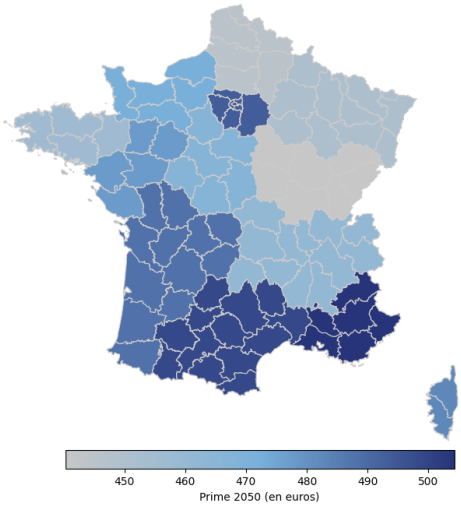
où D_i est l'ensemble des départements ayant un niveau de revenu Q_i.

Le graphique (b) trace quant à lui le gain maximal, c'est à dire la plus grande différence entre la prime en présence de solidarité et la prime actuariellement juste en fonction du niveau de solidarité et du niveau de revenu. Ce gain maximal, noté Δ_i^{max}, est obtenu avec la formule

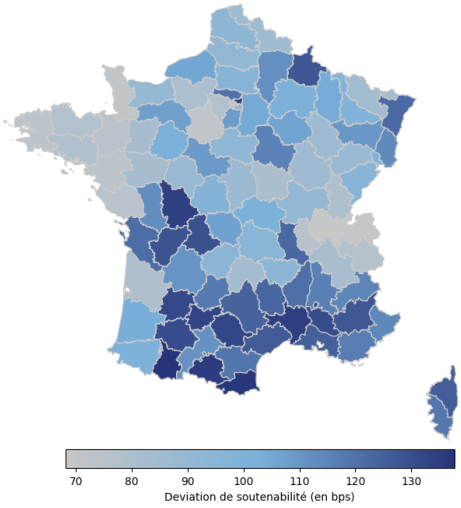
$$\Delta_i^{max}(\alpha, \beta) = \max_{d \in D_i} \left\{ (\pi_d^H(\alpha, \beta) - \pi_d^H) \right\}.$$

Le graphique (a) indique qu'un niveau élevé de solidarité profite davantage aux départements du niveau de revenu Q2 avec un gain moyen de 8 euros sur la prime moyenne MRH. Dans le cas d'une solidarité maximale entre les départements, ce sont les départements les plus riches qui en moyenne contribuent à payer le coût du risque sécheresse des départements aux bas niveaux de revenu. L'impact de la solidarité sur les départements avec un niveau de richesse Q1 est relativement faible en moyenne mais est considérable au niveau du gain maximal sur le graphique (b), indiquant un gain de 100 euros sur la prime MRH moyenne pour le département le plus exposé au risque sécheresse. Ceci révèle donc une nouvelle fois que la capacité des départements aux revenus faible à supporter la prime repose sur l'apport de la solidarité.

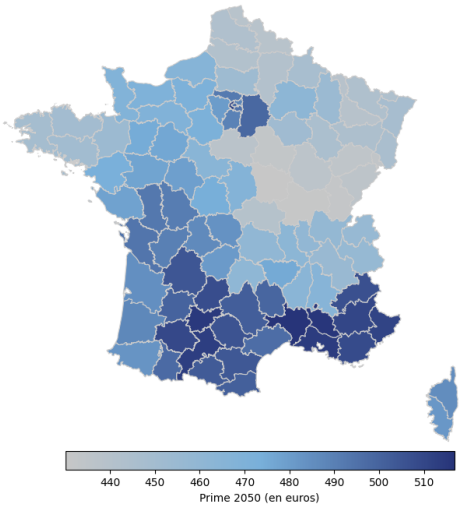
Cependant, au vu des différentes prévisions des acteurs de l'assurance concernant l'augmentation des sinistres climatiques (inondations, submersions marines, tempêtes et grêles) et l'expansion des zones touchées sur le territoire français, l'hypothèse d'une tarification de plus en plus segmentée est envisagée ($\alpha \to 1$, $\beta \to 1$ dans le modèle hiérarchique) afin de maintenir la rentabilité des assureurs.



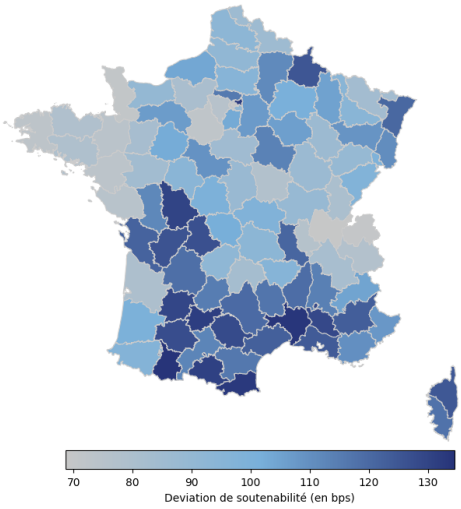
(a) Estimation des primes pour $\alpha = 0.9$ et $\beta = 0$



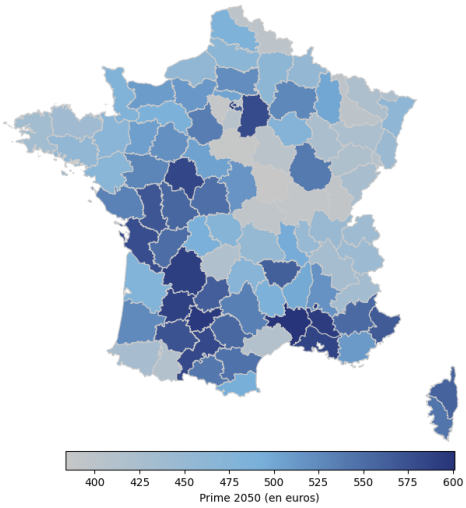
(b) Deviation de soutenabilité pour $\alpha = 0.9$ et $\beta = 0$



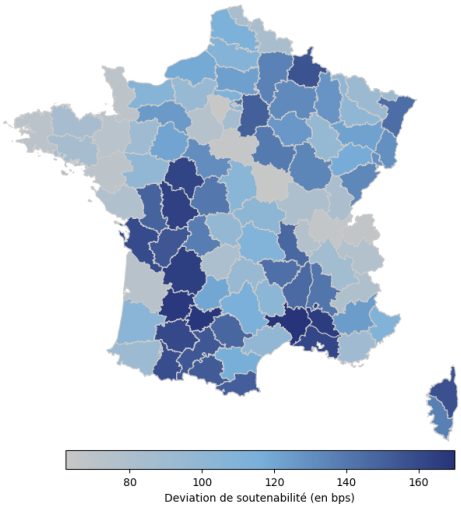
(c) Estimation des primes pour $\alpha = 0.9$ et $\beta = 0.14$



(d) Deviation de soutenabilité pour $\alpha = 0.9$ et $\beta = 0.14$

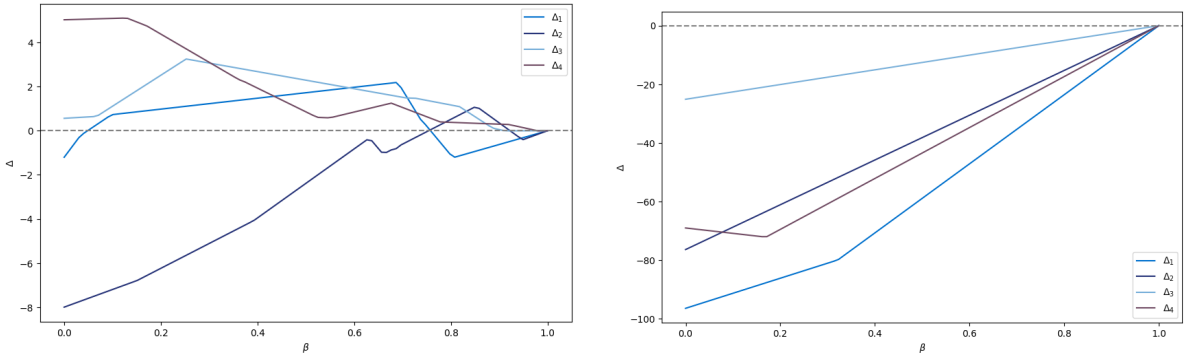


(e) Estimation des primes pour $\alpha = 0.9$ et $\beta = 1$



(f) Deviation de soutenabilité pour $\alpha = 0.9$ et $\beta = 1$

FIGURE 3.17 : Représentation spatiale de la prime MRH moyenne départementale π^H_{2050} et de la déviation de soutenabilité en fonction de β et α.



(a) Gain/coût moyen sur la prime moyenne MRH en fonction de β et par niveau de revenu

(b) Gain/coût maximal sur la prime moyenne MRH en fonction de β et par niveau de revenu

FIGURE 3.18 : Evolution du gain/coût moyen et maximal sur la prime moyenne MRH en fonction de β et par niveau de revenu

Sous cette hypothèse, la mutualisation des risques au niveau national et régional serait compromise et conduirait, comme le montre la présente étude, à une explosion de la prime MRH dans de nombreux départements et par conséquent à un risque d'insoutenabilité chez les assurés à horizon 2050.

3.4 Rappels des limites de l'étude et extensions possibles

Cette section a pour objectif de rappeler les diverses limites inhérentes au périmètre défini par ce mémoire, tout en mentionnant les orientations envisageables pour des travaux ultérieurs.

3.4.1 Autour de la modélisation du risque sécheresse

Pour des raisons liées à l'accessibilité des données et aux problèmes de complexité computationnelle, la modélisation du risque de sécheresse a été abordée à travers une maille départementale. Au regard de la complexité du phénomène modélisé, l'utilisation d'une échelle communale pour l'entraînement du modèle aurait potentiellement amélioré la précision et renforcé la robustesse de celui-ci. Avec une telle résolution spatiale, des variables propres aux habitations auraient pu enrichir le modèle. En effet, la proximité de la végétation ou encore la topologie du terrain jouent un rôle crucial en tant que facteurs aggravants dans la manifestation du phénomène. De plus, l'utilisation d'une résolution communale aurait permis d'être cohérent avec le dispositif Cat Nat, où la reconnaissance de l'état de catastrophe naturelle s'effectue par commune, et ainsi intégrer une dimension réglementaire au modèle. Toutefois, le modèle élaboré comprend de manière partielle cet aspect réglementaire car l'indice de magnitude SWI, construit spécifiquement pour l'étude, dépend de l'indice SWI qui est utilisé comme critère par le régime Cat Nat.

Par ailleurs, comme mentionné dans l'article de Charpentier et al. (2022b), l'incertitude des prédictions peut être en partie attribuée à l'hétérogénéité des données utilisées pour calibrer le modèle. Les critères de déclaration des catastrophes naturelles ont subi de multiples modifications au fil de la période de calibration, avec huit ajustements apportés depuis 1989, ce qui a introduit des biais dans les données historiques. En effet, sous des conditions climatiques et d'exposition équivalentes, la détection d'un même événement de sécheresse par le modèle peut varier d'une année à l'autre en raison de modifications intervenues dans le mécanisme d'indemnisation du régime. En ce sens, bien que l'exercice apparaisse délicat en pratique, procéder à un retraitement *"as-if"* des données, de sorte à obtenir un historique de charge sinistre à réglementation constante aurait permis de construire une base de données reflétant davantage la sinsitralité du phénomène modélisé, bénéficiant ainsi à l'entraînement du modèle.

3.4.2 Autour de la projection du risque

Une première limite relative à la projection du risque réside dans l'absence d'intégration, dans le scénario de projection démographique de l'INSEE (2020) utilisé dans l'étude, d'un effet de prise de conscience de la population face au risque climatique. Cet effet potentiel aurait pu entraîner une réduction des déplacements de populations vers les zones à risques. De plus, il est concevable que les personnes qui migrent vers ces territoires exposés ne s'installent pas systématiquement dans les zones du département présentant une forte vulnérabilité. Par conséquent, une augmentation du nombre de risques dans un département ne signifie pas nécessairement une augmentation concomitante des sinistres liés à la sécheresse dans ce même département. Dans ce mémoire, l'hypothèse d'une répartition constante du nombre de logements entre les zones à risques et les zones non à risques lors des projections a été retenue, mais une approche plus précise aurait pu être considérée, prenant en compte des scénarios de concentration ou de dispersion de la population au sein d'un département. Cette influence du mode d'urbanisation sur le coût relatif au phénomène RGA a notamment été traitée dans les travaux de Gourdier et Plat (2018).

De surcroît, la projection de la sinistralité effectuée dans ce mémoire est réalisée sans prendre en considération l'hypothèse d'une mise en place de plans de prévention ou d'une politique d'amélioration

du bâti. En septembre 2023, France Assureurs, CCR et la Mission Risques Naturels (MRN) ont lancé le projet « Initiative Sécheresse » CCR (2023b), visant à analyser des solutions de prévention et de protection des habitations individuelles contre la sécheresse. Ces mesures pourraient potentiellement atténuer, à moyen terme, la charge sinistre liée au phénomène RGA. Ainsi, quantifier l'impact de ces mesures préventives sur la sinistralité future, comme cela a été réalisé dans l'étude de Gourdier et Plat (2018), pourrait apporter un nouvel éclairage à l'étude et à la problématique de l'assurabilité du risque.

Enfin la projection de la charge sinistre future a été effectuée à réglementation constante et ne comprend donc pas de modifications du mécanisme d'indemnisation et des critères d'éligibilité du régime Cat Nat. La dimension politique qui entoure le régime complique son analyse et la mise en place d'hypothèses concernant l'évolution de son fonctionnement. Cependant, le 6 avril 2023, une proposition de loi (Assemblee Nationale (2023)) visant à "*mieux indemniser les dégâts sur les biens immobiliers causés par le retrait-gonflemet de l'argile*" a été adoptée par l'Assemblée Nationale et doit désormais être examinée par le Sénat. Le texte a pour objectif d'assouplir les critères de reconnaissance de l'état de catastrophe naturelle au moyen de nouvelles mesures, incluant notamment la réduction de la période de retour à 10 ans pour caractériser une sécheresse extrême. Cet allégement du critère d'éligibilité pourrait potentiellement exercer une influence sur l'indemnisation des communes sinistrées. Bien qu'une première évaluation de l'impact de cette réforme sur le nombre de reconnaissances de l'état de catastrophe naturelle ait été réalisée en A.1, une analyse plus approfondie pour évaluer son influence sur l'évolution du montant de la charge sinistre future serait opportune.

3.4.3 Autour de l'étude de soutenabilité

Concernant les résultats de l'étude de soutenabilité, il convient de souligner que le mémoire a focalisé son analyse sur l'évolution de la prime MRH en examinant exclusivement le risque sécheresse. Dans le cadre du mémoire, la croissance des autres risques climatiques couverts par le produit MRH repose sur l'évolution de l'indice FFB. Cependant, les diverses études conduites par Covéa (2022), France Assureurs (FFA (2021)), ainsi que la CCR (2023a), prévoient une augmentation des sinistres climatiques d'ici 2050, avec notamment une progression importante de la charge sinistres liée aux inondations, pouvant atteindre jusqu'à 110% selon les hypothèses retenues. Par conséquent, une modélisation globale de l'ensemble des risques climatiques aurait permis de fournir une réponse plus exhaustive à la question de la soutenabilité de la prime MRH. Cette approche aurait été d'autant plus pertinente que les risques se superposent dans certaines régions, renforçant ainsi l'hypothèse d'une évolution drastique de la prime et les préoccupations relatives à l'accessibilité de celle-ci.

Enfin, les conclusions de l'étude auraient pu être améliorées par une résolution géographique plus fine. L'adoption d'une maille communale, à l'instar de l'approche adoptée dans Charpentier et al. (2022a), aurait permis de mettre en évidence de manière plus précise l'hétérogénéité du risque sécheresse et des disparités économiques sur le territoire métropolitain, conduisant ainsi à des résultats plus significatifs.

Conclusion

Afin d'appréhender la soutenabilité des primes MRH à horizon futur, la première partie du mémoire a consisté en l'élaboration d'un modèle sécheresse capable de projeter de manière fidèle la sinistralité du phénomène RGA. La construction de ce modèle a nécessité la prise en compte de plusieurs facteurs clés intervenant dans l'apparition des sinistres liés à la sécheresse géotechnique. D'une part, les paramètres décrivant la susceptibilité d'un territoire au phénomène de subsidence ont été exploités. Ces facteurs de prédisposition concernent à la fois les caractéristiques des logements, le phénomène affectant principalement les maisons, mais également leur localisation. Ainsi, une variable sur la proportion de logements situés en zone argileuse a été considérée, la nature du sol étant reconnu à ce jour comme un facteur de prédisposition majeur.

D'autre part, des données de nature climatique ont été intégrées au modèle afin de faire le lien entre une variation des conditions météorologiques et hydrologiques d'une zone et le déclenchement d'un sinistre sécheresse. À cette fin, deux indices de sécheresse ont été mis en place : l'indice SPEI-3 (Vicente-Serrano et al. (2010)), qui se base sur l'évapotranspiration, et l'indice de magnitude SWI, qui est un nouvel indice spécifiquement développé pour le mémoire et dérivé de l'indice SWI uniforme établi par Météo France.

Le choix du modèle de *machine learning* utilisé pour la prédiction de la charge sinistre sécheresse s'est porté sur l'algorithme *Catboost* (Prokhorenkova et al. (2018)), fondé sur les méthodes de *boosting*. Pour évaluer les performances du modèle construit et prendre en compte l'aspect temporel et spatial des données d'entraînement, deux méthodes de validations ont été appliquées : une validation croisée temporelle et une validation croisée spatiale (Pohjankukka et al. (2017)). L'emploi de ces deux approches a notamment permis de limiter l'apparition d'un biais lors des phases d'apprentissage, aboutissant à une évaluation rigoureuse du modèle retenu, tout en pointant ses faiblesses. La principale difficulté résidait sur le caractère déséquilibré de la variable cible dans la base d'entraînement, une caractéristique commune aux sinistres liés aux phénomènes climatiques, où seule une fraction réduite de la base présentait des charges extrêmes. La flexibilité de calibration offerte par l'algorithme *Catboost* a rendu possible la gestion de ce déséquilibre lors de la phase d'entraînement, en pénalisant le gradient de la fonction de perte pour des valeurs de sinistres élevées. Cette pénalisation ainsi que l'optimisation des paramètres du modèle ont conduit à une amélioration globale des métriques temporelles et spatiales définies pour l'étude, avec notamment une meilleure détection des événements de sinistralité de grande ampleur. Il convient de noter également que l'indice de magnitude SWI, affiche la plus forte importance relative sur les sorties du modèle au regard des valeurs SHAP (Lundberg et Lee (2017)), confirmant la pertinence de son utilisation pour la prédiction du risque sécheresse.

Le modèle ainsi construit et validé a été ensuite utilisé pour projeter la sinistralité sécheresse à horizon 2050. La projection des variables climatiques, et plus précisément des indices de sécheresse, s'est effectuée au travers des trajectoires des émissions de gaz à effet de serre du scénario RCP 4.5 défini par le GIEC. Bien que le scénario RCP 8.5 ait été envisagé préalablement, sa faisabilité concernant

l'utilisation des énergies fossiles rend sa pertinence en tant que scénario "*business as usual*" discutable (Hausfather (2019)). De plus le choix du scénario RCP 4.5 a été motivée par un souci de cohérence avec le prochain exercice pilote de l'ACPR (2023). Les résultats issus de l'application de la trajectoire 4.5 montre au travers des indices de sécheresse, une dégradation des conditions de sécheresse jusqu'à la moitié du XXIè siècle sur l'ensemble du territoire métropolitain, notamment autour du bassin méditerranéen.

Pour la projection des enjeux assurés, deux aspects ont été traités. D'une part l'évolution du nombre de risques assurés à horizon 2050 à travers les projections démographiques du modèle Omphale de l'INSEE (2022). D'autre part l'évolution des valeurs assurées, basée sur l'indice FFB du coût de la construction (ICC) ainsi que sur une composante d'enrichissement par département calculé à partir de l'historique du portefeuille. En ce qui concerne les départements fortement impactés par la sécheresse, une croissance démographique de 5% est observée entre 2020 et 2050, comparée à une augmentation de 1% dans les autres départements. De plus, la valeur des biens assurés connaît une augmentation de 105%, dépassant ainsi de 8% la croissance enregistrée dans les autres départements.

Le modèle développé dans le mémoire ainsi que la projection des différents éléments le composant confirme la tendance haussière de la sinistralité sécheresse à horizon 2050 annoncée par d'autres études des acteurs de l'assurance. Les effets du changement climatique sous le scénario RCP 4.5 et l'évolution démographique provoquent une croissance de 72% de la perte annuelle moyenne entre la période de référence (2000-2020) et la période de projection (2041-2050). Ce résultat s'aligne notamment avec les précédentes études menées par Covéa (2022) et la CCR (2023a) qui prévoyaient entre 60 et 70% d'augmentation de la charge sinistre sécheresse à horizon 2050. La projection de la progression des valeurs assurés accentue d'autant plus ces évolutions et constitue le principal facteur inflationniste à horizon futur (101%) de la présente étude. D'un point de vue spatial, les territoires situés actuellement dans des zones à risque RGA moyen, comme les départements de la Charente, de la Sarthe ou encore de la Seine-et-Marne, connaissent à climat futur les plus fortes évolutions. Les départements qui pèsent le plus dans la sinistralité future sont la Haute-Garonne, le Tarn-et-Garonne, la Dordogne, le Gers, ou encore le Vaucluse et le Gard, qui concentrent à eux seuls 60% de la charge sinistre totale future.

A partir de la sinistralité sécheresse projetée, le mémoire poursuit son analyse en croisant la catégorie de risque des départements à horizon 2050 avec leur niveau de richesse. Pour caractériser le niveau d'un richesse d'un département, la variable du revenu médian a été retenue, permettant d'avoir un indicateur agrégé et plus robuste que la moyenne. Cette étude a permis de mettre en évidence que parmi les départements classés comme "très risqués", un très grand nombre d'entre eux était associé à des revenus médians relativement bas. Un test d'indépendance du khi-deux amène également à rejeter l'hypothèse d'indépendance entre la catégorie de risque sécheresse d'un département et son niveau de richesse. Ce constat soulève naturellement des questions quant à la fragilité de certains départements pour faire face au coût de la protection des phénomènes de sécheresse et invite donc à les traiter au travers de la tarificiation de la prime MRH à horizon 2050.

Pour continuer dans ce sens, un ratio de soutenabilité est défini comme métrique pour étudier la soutenabilité d'un département face à l'augmenation de la prime MRH horizon futur. De plus, deux scénarios de tarification sont étudiés par le biais de la théorie de la crédibilité. Le premier correspond à une tarification actuariellement juste, basée sur le risque, où la prime payée par le département est en accord avec le risque auquel il est exposé. L'application d'une telle méthode de tarification conduit à une augmentation significative des primes de 82% en moyenne sur l'ensemble du territoire, avec des montant de primes dépassant 530 euros dans certains départements. Cette progression majeure des primes s'accompagne d'une déviation de 90 points de base du ratio de soutenabilité moyen du

portefeuille, confirmant une augmentation plus importante de la sinistralité sécheresse par rapport au revenu médian. Les départements du Gard, du Tarn-et-Garonne, du Lot-et-Garonne du Vaucluse mais également du Gers font figure de *hot-spot* concernant le risque d'insoutenabilité de la prime MRH du fait de l'augmentation des phénomène de sécheresse.

Le deuxième scénario repose quant à lui sur une tarification paramétrique basée sur différents niveaux de solidarité (nationale et régionale). L'application de cette approche tarifaire permet d'analyser la variation des primes et de la soutenabilité des départements en fonction de différents degrés de solidarité et ainsi mettre en évidence la fragilité de certains territoires en cas d'absence de mutualisation du risque sécheresse. L'étude de la tarification paramétrique permet également de quantifier l'apport de la solidarité sur le montant de la prime moyenne MRH et de montrer que celle-ci bénéficie le plus aux départements caractérisés par de bas niveaux de revenus. Ce constat renforce une nouvelle fois l'argument qu'en absence de solidarité, de nombreux départements seraient exposés au risque d'insoutenabilité à horizon futur.

Néanmoins, ces conclusions autour de la soutenabilité doivent être nuancées par les limites de l'approche utilisée et remises dans le contexte de l'étude. En effet, celle-ci analyse l'évolution de la prime MRH exclusivement par le biais du risque sécheresse. Or, une modélisation multi-périls, intégrant d'autres risques climatiques préoccupant tels que les inondations, tendrait à accentuer les résultats obtenus sur la déviation de soutenabilité à horizon futur. Une telle approche aurait été plus pertinente, compte tenu de la superposition des risques dans certaines régions, offrant ainsi une réponse plus exhaustive à la question de la soutenabilité. De plus, les projections de sinistralité ont été réalisées en maintenant la réglementation constante et ne tiennent pas compte de l'implémentation de plans de prévention ou de modifications des critères d'éligibilité Cat Nat, des facteurs qui pourraient potentiellement influencer les coûts à la charge des assureurs à moyen terme. Enfin, bien que les résultats aient été obtenus à l'échelle départementale, une analyse plus fine au niveau communal révélerait de manière significative des territoires exposés à un risque d'insoutenabilité, renforçant ainsi les préoccupations relatives à l'accessibilité de la prime.

Bibliographie

Abramowitz, M. et Stegun, I. A. (1968). Handbook of mathematical functions with formulas, graphs, and mathematical tables. T. 55. US Government printing office.

ACPR (2021). Une première évaluation des risques financiers dus au changement climatique - les principaux résultats de l'exercice pilote climatique 2020. Analyses et synthèses.

ACPR (2023). Scénarios et hypothèses principales de l'exercice de stress test climatique 2023. url : https://acpr.banque-france.fr/scenarios-et-hypotheses-principales-de-lexercice-de-stress-test-climatique-2023.

Argus (2023). Sécheresse : le Sénat s'inquiète pour l'équilibre du régime Cat Nat. L'Argus de l'assurance. Article. url : https://www.argusdelassurance.com/les-assureurs/secheresse-le-senat-s-inquiete-pour-l-equilibre-du-regime-cat-nat.213041.

Assemblee Nationale (2023). Proposition de loi n°103, visant à mieux indemniser les dégâts sur les biens immobiliers causés par le retrait-gonflement de l'argile. url : https://www.assemblee-nationale.fr/dyn/16/textes/l16t0103_texte-adopte-seance.

Bailey, A. L. (1950). Credibility Procedures: Laplace's generalization of Bayes' Rule and the combination of collateral knowledge with observed data. New York State Insurance Department.

Barthelemy et al. (2022). Comprendre le phénomène de retraitgonflement des argiles par le biais d'un indicateur agrégé à la commune : la magnitude des sécheresses. Rapport Scientifique CCR 2022 - CCR Paris 2022, pp. 20-22.

Bradford, R. (2000). Drought events in Europe. *Drought and drought mitigation in Europe*. Springer, p. 7-20.

Branco, P., Torgo, L. et Ribeiro, R. P. (2017). SMOGN: a pre-processing approach for imbalanced regression. *First international workshop on learning with imbalanced domains: Theory and applications*. PMLR, p. 36-50.

Breiman, L. (2001). Random forests. *Machine learning* 45.1, p. 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (2017). Classification and regression trees. Routledge.

BRGM (mai 2023). Cartographie de l'exposition au phénomène retrait-gonflement des argiles. Georisques. url : https://www.georisques.gouv.fr/articles-risques/retrait-gonflement-des-argiles/exposition-du-territoire-au-phenomene (visité le 12/05/2023).

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin: The Journal of the IAA* 4.3, p. 199-207.

Bühlmann, H. et Straub, E. (1970). Glaubwürdigkeit für schadensätze. *Bulletin of the Swiss Association of Actuaries* 70.1, p. 111-133.

Byun, H.-R. et Wilhite, D. A. (1999). Objective quantification of drought severity and duration. *Journal of climate* 12.9, p. 2747-2756.

Cazaux, E., Meur-Férec, C. et Peinturier, C. (2019). Le régime d'assurance des catastrophes naturelles à l'épreuve des risques côtiers. Aléas versus aménités, le cas particulier des territoires littoraux. *Cybergeo: European Journal of Geography*.

CCR (2018). Conséquences du changement climatique sur le coût des catastrophes naturelles en France à horizon 2050.

CCR (2022). Les catastrophes naturelles en France - Chiffres clés 2021. (Visité le 16/03/2023).

CCR (2023a). Conséquences du changement climatique sur le coût des catastrophes naturelles en France à horizon 2050 - Septembre 2023.

CCR (2023b). Projet Initiative sécheresse. URL : https://www.ccr.fr/-/ccr-projet-initiative-secheresse-1.

Charpentier, A., Barry, L. et James, M. R. (2022a). Insurance against natural catastrophes: balancing actuarial fairness and social solidarity. *The Geneva Papers on Risk and Insurance-Issues and Practice* 47.1, p. 50-78.

Charpentier, A., James, M. et Ali, H. (2022b). Predicting drought and subsidence risks in France. *Natural Hazards and Earth System Sciences* 22.7, p. 2401-2418.

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, p. 321-357.

Circulaire (2019). Ministère de l'intérieur - Procédure de reconnaissance de l'état de catastrophe naturelle. URL : https://www.legifrance.gouv.fr/circulaire/id/44648 (visité le 12/05/2023).

Cour des comptes (2022). Sols argileux et catastrophes naturelles. Rapport. Cour des Comptes.

Covéa (2022). Changement climatique et assurance : quelles conséquences sur la sinistralité à horizon 2050. Livre blanc. Covéa et Risk Weather Tech.

Data.Gouv (2023). Données mensuelles de l'indice d'humidité des sols pour le dispositif Cat Nat. URL : https://www.data.gouv.fr/fr/reuses/indice-swi-uniforme/.

Delorme, A. (2022). Assurance contre la sécheresse au XXIe siècle : perspectives d'évolution. Mémoire d'actuariat. Paris : Université Paris-Dauphine.

DRIAS (2012). DRIAS, les futurs du cliat - Ministère de la transition écologique. URL : https://www.drias-climat.fr/.

FFA (2015). Changement climatique et assurance. Rapport d'étude horizon 2040.

FFA (2021). Impact du changement climatique sur l'assurance à l'horizon 2050. Rapport.

FFA (2022). Le risque sécheresse et son impact sur les habitations. Master Class.

FFB (mai 2023). Indice FFB du coût de la construction (ICC). URL : https://www.outils.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html (visité le 16/05/2023).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, p. 1189-1232.

Georisque (2023a). M'informer sur le retrait-gonflement des argiles - Cartographie de l'exposition. URL : https://www.georisques.gouv.fr/articles-risques/retrait-gonflement-des-argiles/exposition-du-territoire-au-phenomene.

Georisque (2023b). Retrait gonflement des argiles - base de données. URL : https://www.georisques.gouv.fr/donnees/bases-de-donnees/retrait-gonflement-des-argiles.

GIEC (2014). AR5 Synthesis Report: Climate Change 2014. URL : https://www.ipcc.ch/report/ar5/syr/.

Gourdier, S. et Plat, E. (2018). Impact du changement climatique sur la sinistralité due au retrait-gonflement des argiles. *Journées Nationales de Géotechnique et Géologie de l'Ingénieur (JNGG) 2018*.

Guha-Sapir et al. (2021). EM-DAT: The CRED/OFDA International. Disaster Database. URL : https://www.emdat.be (visité le 12/11/2023).

Guttman, N. B. (1999). Accepting the standardized precipitation index: a calculation algorithm 1. *JAWRA Journal of the American Water Resources Association* 35.2, p. 311-322.

Hausfather (2019). Explainer: The high-emissions 'RCP8.5' global warming scenario. URL : https://www.carbonbrief.org/explainer-the-high-emissions-rcp8-5-global-warming-scenario/.

Iglesias, A., Assimacopoulos, D. et Van Lanen, H. A. (2018). Drought: Science and policy. John Wiley & Sons.

INSEE (2020). Dispositif sur les revenus localisés sociaux et fiscaux - FILSOFI. url : https://www.insee.fr/fr/metadonnees/source/serie/s1172.

INSEE (2022). Projections Démographiques Omphale. url : https://www.insee.fr/fr/information/1303412 (visité le 12/07/2023).

Ionita, M. et Nagavciuc, V. (2021). Changes in drought features at the European level over the last 120 years. *Natural Hazards and Earth System Sciences* 21.5, p. 1685-1701.

Jewell, W. S. (1975). The use of collateral data in credibility theory: a hierarchical model.

Keetch, J. J. et Byram, G. M. (1968). A drought index for forest fire control. T. 38. US Department of Agriculture, Forest Service, Southeastern Forest Experiment.

Kogan, F. N. (1995a). Application of vegetation index and brightness temperature for drought detection. *Advances in space research* 15.11, p. 91-100.

Kogan, F. N. (1995b). Droughts of the late 1980s in the United States as derived from NOAA polar-orbiting satellite data. *Bulletin of the American Meteorological Society* 76.5, p. 655-668.

Lundberg, S. M. et Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

McKee, T. B., Doesken, N. J., Kleist, J. et al. (1993). The relationship of drought frequency and duration to time scales. *Proceedings of the 8th Conference on Applied Climatology*. T. 17. 22. Boston, p. 179-183.

Morgan, J. N. et Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association* 58.302, p. 415-434.

Météo France (2023). Données climatologiques mensuelles pour les stations de métropole et d'outre-mer appartenant au Réseau Climatologique Régional de Base (RBCN) de l'Organisation Météorologique Mondiale (OMM). url : https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=117&id_rubrique=39.

ONF (2022). "Feux de forêt : le risque s'étend partout en France". url : www.onf.fr.

Palmer, W. C. (1965). Meteorological drought. T. 30. US Department of Commerce, Weather Bureau.

Pohjankukka, J., Pahikkala, T., Nevalainen, P. et Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science* 31.10, p. 2001-2019.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. et Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31.

SDES (2021). Indicateur d'exposition des maisons individuelles au retrait gonflement des argiles par commune. url : https://www.statistiques.developpement-durable.gouv.fr/nouveau-zonage-dexposition-au-retrait-gonflement-des-argiles-plus-de-104-millions-de-maisons.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 23rd ACM national conference*, p. 517-524.

Soubeyroux, J., Vidal, J., Najac, J, Kitova, N, Blanchard, M, Dandin, P, Martin, E, Pagé, C et Habets, F (2011). Projet ClimSec Impact du changement climatique en France sur la sécheresse et l'eau du sol. *Météo-France, Toulouse, France*.

Spinoni, J., Naumann, G., Carrao, H., Barbosa, P. et Vogt, J. (2014). World drought frequency, duration, and severity for 1951–2010. *International Journal of Climatology* 34.8, p. 2792-2804.

Spinoni, J., Naumann, G., Vogt, J. et Barbosa, P. (2015). European drought climatologies and trends based on a multi-indicator approach. *Global and Planetary Change* 127, p. 50-57.

Spinoni, J., Naumann, G. et Vogt, J. V. (2017). Pan-European seasonal trends and recent changes of drought frequency and severity. *Global and Planetary Change* 148, p. 113-130.

Sénat (2023). Financement du risque de retrait gonflement des argiles et de ses conséquences sur le bâti. Rapport d'information. Sénat.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical review* 38.1, p. 55-94.

Van Rooy, M. (1965). A rainfall anomally index independent of time and space, notos.

Vicente-Serrano, S. M., Beguería, S. et López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate* 23.7, p. 1696-1718.

Vidal, J.-P., Martin, E, Franchistéguy, L, Habets, F., Soubeyroux, J.-M., Blanchard, M et Baillon, M (2010). Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcou hydrometeorological suite. *Hydrology and Earth System Sciences* 14.3, p. 459-478.

Whitney, A. W. (1918). Theory of experience rating.

WHO (2021). Organisation Mondiale de la Santé - Drought Overview. url : www.who.int/health-topics/drought#.

WMO (2021). Organisation Météorologique Mondiale - WMO Atlas of mortality and Economic Losses From Weather, Climate and Water Extremes (1970-2019).

Annexe A

Annexes

A.1 Évolution des critères d'éligibilité

Face aux critiques autour des critères d'éligibilité mentionnées en partie 1.2.3 et à la suite du rapport du Sénat (2023) sur le financement du risque RGA, une proposition de loi (Assemblee Nationale (2023)) visant à "*mieux indemniser les dégâts sur les biens immobiliers causés par le retrait-gonflement de l'argile*" a été adoptée par l'Assemblée nationale le 6 Avril 2023 et doit désormais être examinée par le Sénat. Le texte a pour objectif d'assouplir les critères de reconnaissance de l'état de catastrophe naturelle par le biais de nouvelles mesures, avec notamment l'abaissement de la période de retour à 10 ans pour caractériser une sécheresse extrême.

L'analyse menée dans cette section vise à évaluer les répercussions d'une possible modification du système d'indemnisation des sinistres RGA sur le nombre de reconnaissances potentielles à horizon 2050. À cette fin, une cohorte de 1006 communes, réparties sur l'ensemble du territoire métropolitain et ayant fait l'objet d'au moins une reconnaissance de l'état de catastrophe naturelle au cours de la période 2000-2020, a été sélectionnée pour être incluse dans cette étude.

A.1.1 Projection du seuil d'éligibilité

Pour rappel, à la suite des dernières réformes (Circulaire (2019)), l'éligibilité d'une commune est basée sur la variable SWI uniforme de Météo France ($\overline{\text{SWI}}_m^C$), défini en partie 1.2.3. A partir de cet indice, un seuil d'éligibilité unique est retenu pour qualifier le caractère anormal d'une sécheresse, défini comme la deuxième plus petite valeur de SWI sur une profondeur d'historique de 50 ans. La proposition de loi actuellement à l'étude consiste à changer ce seuil en considérant dorénavant la cinquième plus petite valeur sur une profondeur d'historique de 50 ans.

Afin d'apprécier l'évolution du seuil d'éligibilité à l'horizon 2050, la figure A.1 trace initialement la projection du seuil selon le scénario RCP 4.5, en tenant compte de la réglementation en vigueur. Ce choix permet notamment de mettre en évidence l'impact du climat sur la valeur du seuil d'éligibilité. Simultanément, la projection du seuil sous le même scénario RCP 4.5 est réalisée, mais avec l'incorporation des nouvelles directives précédemment évoquées afin d'analyser l'effet croisé à moyen terme du climat et des évolutions réglementaires. En représentant la moyenne annuelle du seuil pour l'ensemble du portefeuille, la figure A.1 illustre une réduction du seuil à l'horizon 2050 sous l'influence du climat, caractérisée par une baisse significative à partir de 2026 suivie d'une stabilisation jusqu'au milieu du siècle. Concernant l'effet croisé, la mise en place de nouvelle réglementation conduit à un assouplissement du critère, perceptible à travers des valeurs moyennes plus élevées sur toute la période projetée. L'application de cette réforme entraîne en effet une augmentation moyenne de 42% du seuil

d'éligibilité tout au long de la période de projection

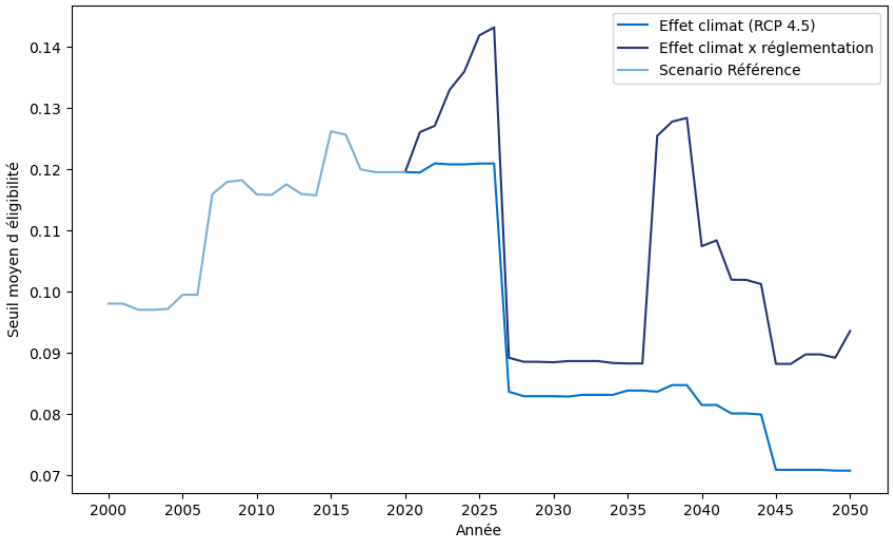


FIGURE A.1 : Projection du seuil moyen d'éligibilité avec effet climat et réglementation

Par la suite, l'évolution moyenne du seuil de chaque commune est représentée géographiquement sur la figure A.2 avec la formule

$$\Delta(\overline{\text{SWI}}) = \frac{\overline{\text{SWI}}^{C^*} - \overline{\text{SWI}}^{C}}{\overline{\text{SWI}}^{C}} \times 100,$$

où $\overline{\text{SWI}}^{C}$ correspond au seuil d'éligibilité obtenu avec la réglementation actuelle et $\overline{\text{SWI}}^{C^*}$ celui calculé avec la réglementation proposée par l'ASSEMBLEE NATIONALE (2023).

D'un point de vue spatial, la carte présentée dans la figure A.2 met en évidence la sensibilité accrue des communes localisées dans les départements septentrionaux de la Nouvelle-Aquitaine (Gironde, Charente-Maritime et Deux-Sèvres), ainsi que celles de l'Indre et du Loiret, à l'égard de la modification réglementaire envisagée. En contraste, les communes situées dans les régions de l'Occitanie, de l'Ile-de-France et des Hauts-de-France semblent être moins touchées par cette évolution normative.

A.1.2 Impacts sur le nombre de reconnaissances potentielles

Le critère de reconnaissance Cat Nat est donc déclenché dès lors que l'indice SWI est inférieur ou égal au seuil décrit précédemment. Pour mesurer l'impact de la réglementation sur l'éligibilité des communes à horizon 2050, une étude comparative du nombre de reconnaissances potentielles sur la période projetée 2020-2050 est effectuée sur l'ensemble des communes retenues. Pour cela, le nombre de reconnaissances potentielles est simulé à horizon 2050 à travers le scénario RCP 4.5 avec la réglementation actuelle et avec la nouvelle réglementation présentée ci-dessus. Le terme "potentielle" fait ici référence au fait que les communes concernées par l'analyse ne sont pas forcément sinistrées au moment où elles sont éligibles. Ainsi, l'évolution du nombre de reconnaissances potentielles se traduit par une évolution de la probabilité que les communes soient déclarées éligibles et non pas par une évolution du nombre de communes déclarées éligibles. La figure A.3 représente cette évolution par commune sur le territoire métropolitain avec la formule

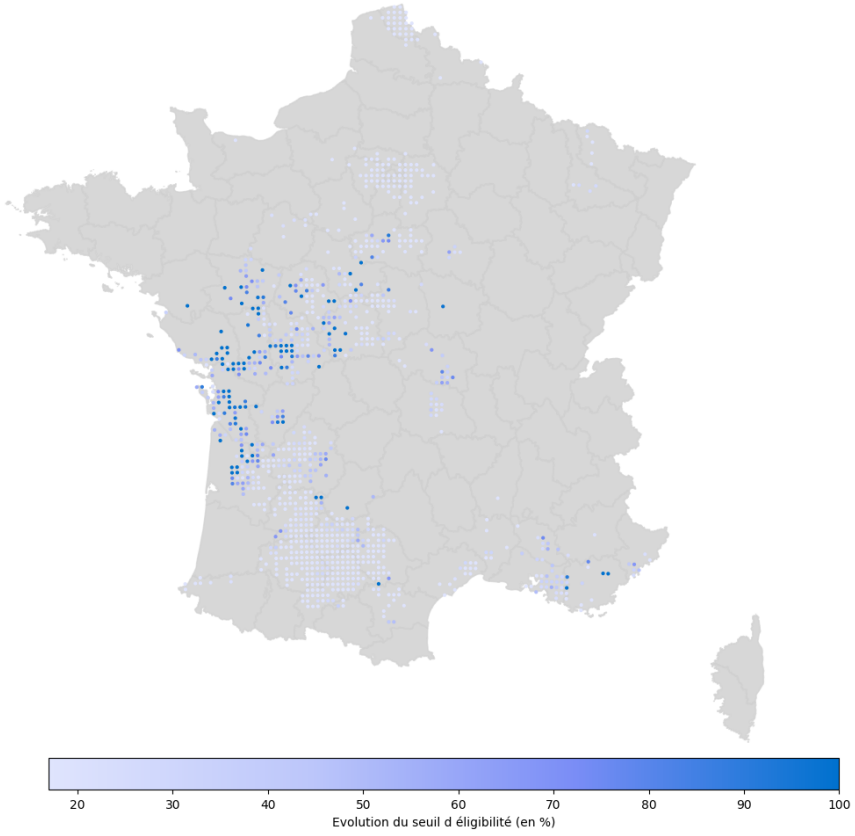


FIGURE A.2 : Impact de la nouvelle réglementation sur l'évolution du seuil d'éligibilité pendant la période projetée (2020-2050) sous le scénario RCP 4.5

$$\Delta(n_i) = \frac{n_i^* - n_i}{n_i} \times 100,$$

avec n_i^*, le nombre de reconnaissances sous la nouvelle réglementation de la commune i sur la période projetée et n_i, le nombre de reconnaissances sous la réglementation actuelle de cette même commune.

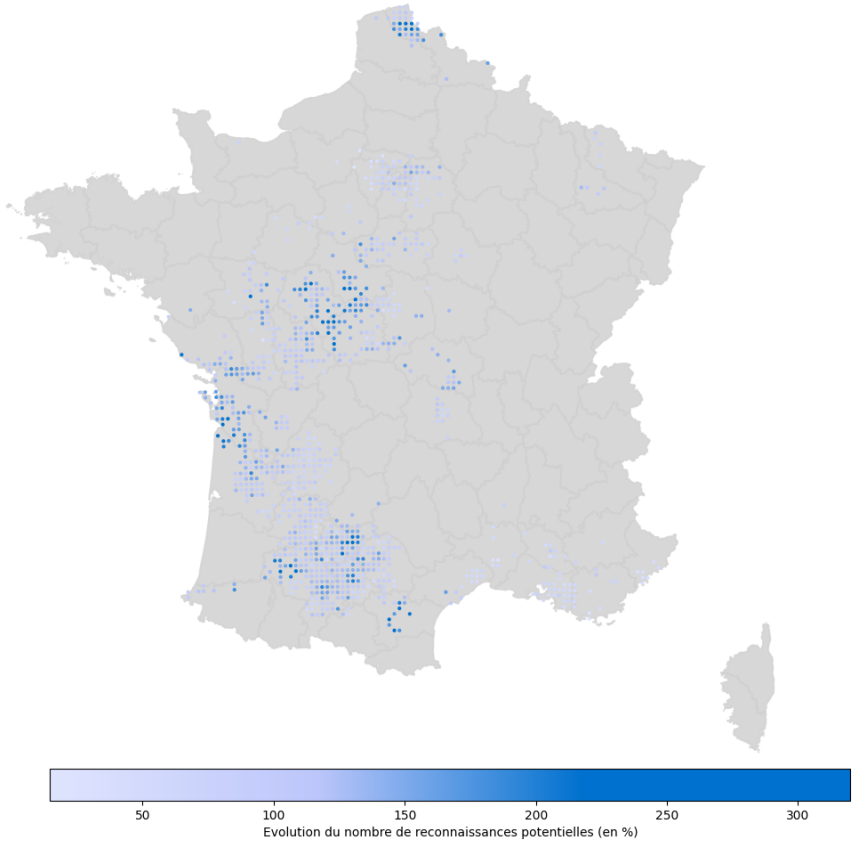


FIGURE A.3 : Impact de la nouvelle réglementation sur le nombre de reconnaissances potentielles sur la période projetée (2020-2050) sous le scénario RCP 4.5

Sur l'ensemble du territoire, le nombre de reconnaissances potentielles augmenteraient de 108% sous l'effet de la nouvelle réglementation. L'impact de cette nouvelle directive s'observe principalement au sein des départements de l'Aude, du Gers, du Tarn-et-Garonne, de la Charente-Maritime, de l'Indre, et du Nord, se traduisant par des évolutions dépassant les 320% dans certaines communes. En revanche, les effets de la réforme se révèlent relativement modérées pour les communes situées dans les départements de la Dordogne, du Var, des Bouches-du-Rhône, ainsi que dans la région Île-de-France.

A.2 Ajustement de la loi log-logistique

Pour rappel, le calcul de l'indice SPEI-3 fait intervenir la calibration d'une loi sur les données historiques saisonnières de précipitations nettes $\Delta_{i,s}$ définies par

$$\Delta_{i,s} = P_{i,s} - \mathrm{ETP}_{i,s},$$

avec $s = \{\text{Automne}, \text{Hiver}, \text{Printemps}, \text{Ete}\}$, i le département, $P_{i,s}$ les précipitations et $\mathrm{ETP}_{i,s}$ l'évapotranspiration.

Pour modéliser les précipitations nettes $\Delta_{i,s}$, Vicente-Serrano et al. (2010) suggère l'utilisation de la loi log-logistique(α, β, γ) dont la densité s'écrit

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1}\left[1+\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right]^{-2}$$

où,

- α est le paramètre d'échelle
- β est le paramètre de forme
- γ est le paramètre de dispersion.

Ainsi, pour construire l'indice à la maille souhaitée, pour chaque saison et pour chaque département, une loi log-logistique est ajustée sur les données de précipitations nettes historiques sur une profondeur de 40 ans, comme illustré dans la figure A.4.

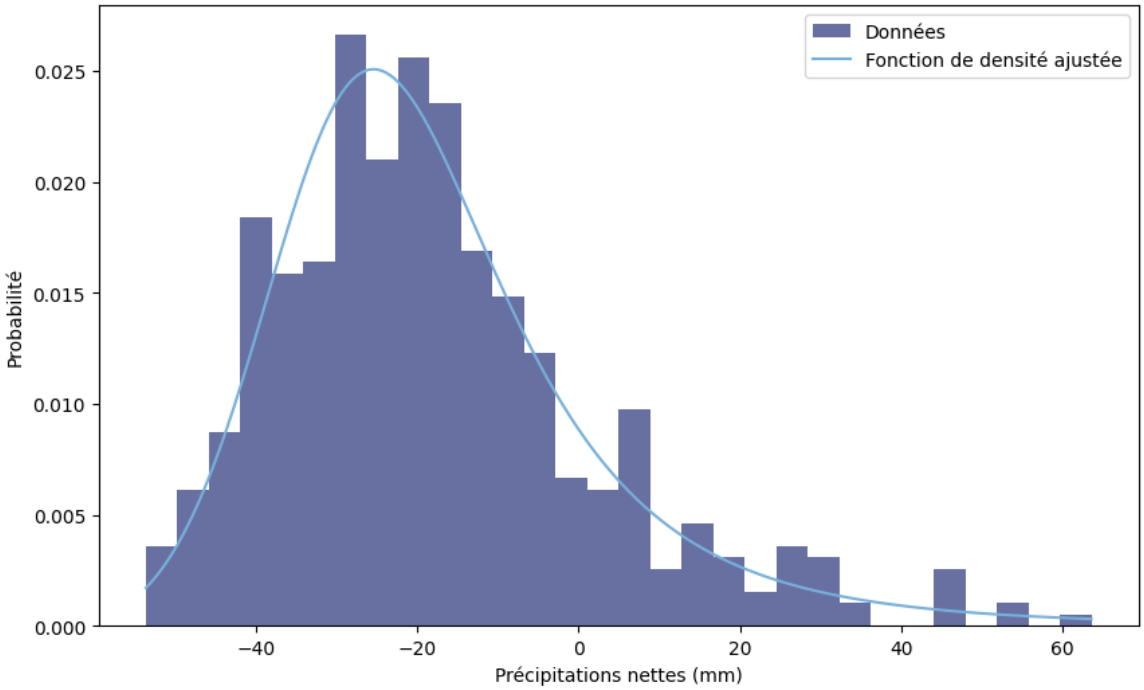
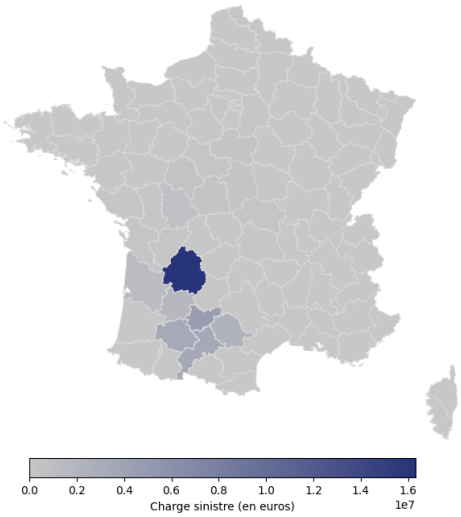


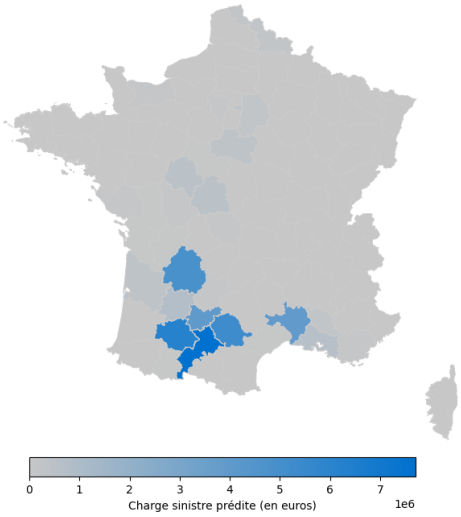
Figure A.4 : Ajustement de la fonction de densité de la loi log logistique pour le calcul du SPEI estival du département du Gers

A.3 Comparaison spatiale de la charge sinistre annuelle observée et prédite

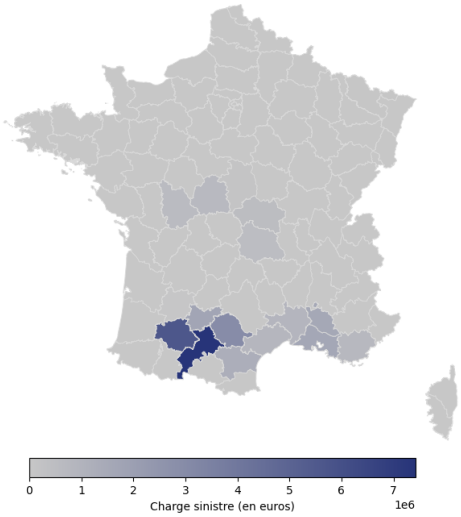
Les figures A.5 et A.6 comparent la charge sinistre annuelle observée et prédite par le modèle optimisé construit en section 2.4.2 pour différents épisodes majeurs de sécheresse en France métropolitaine.



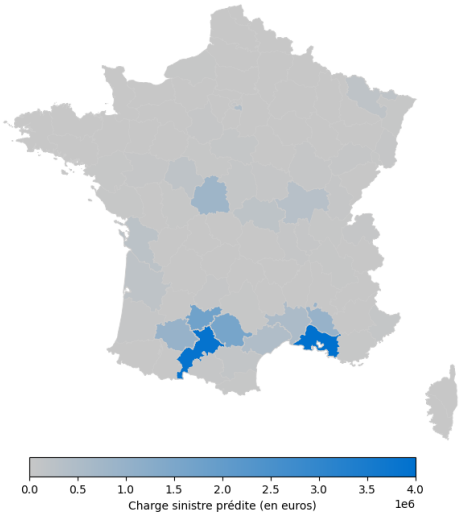
(a) Observée - Année 2011



(b) Prédite - Année 2011

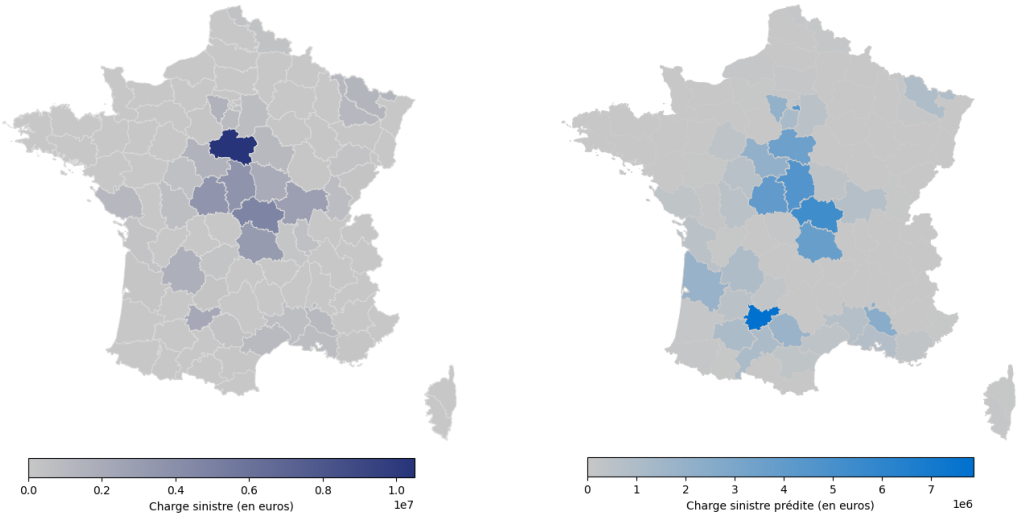


(c) Observée - Année 2016



(d) Prédite - Année 2016

FIGURE A.5 : Comparaison spatiale de la charge sinistre annuelle observée et prédite



(a) Observée - Année 2018

(b) Prédite - Année 2018

(c) Observée - Année 2020

(d) Prédite - Année 2020

FIGURE A.6 : Comparaison spatiale de la charge sinistre annuelle observée et prédite

A.4 Valeurs historiques de l'indice FFB du coût de la construction (ICC)

Pour le traitement *"as-if"* du portefeuille ainsi que la projection des valeurs assurées, l'indice FFB du coût de la construction (ICC) a été utilisé. Cet indice est publié trimestriellement, avec une base 1 au 1er janvier 1941. Son calcul repose sur le prix de revient moyen d'un immeuble à Paris, intégrant implicitement plusieurs composants tels que les coûts des matériaux et de la main-d'œuvre. La valeur du terrain n'est pas prise en compte dans ce calcul. L'indice FFB du coût de la construction est employé pour l'indexation des polices d'assurance habitation, notamment pour l'ajustement des primes, justifiant ainsi son utilisation dans cette étude. La figure A.7 présente l'évolution trimestrielle de cet indice depuis l'année 2000.

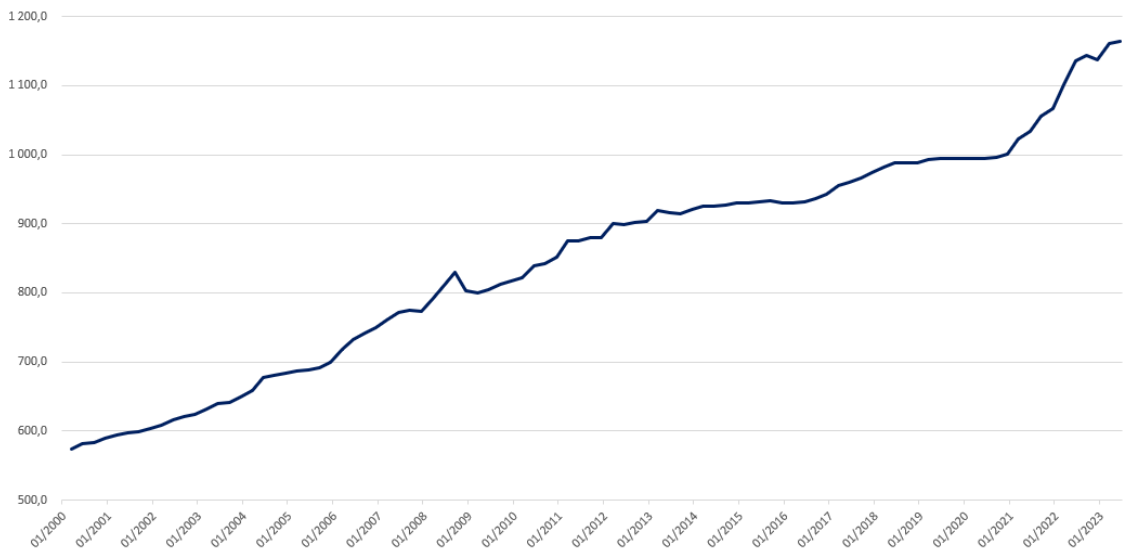


FIGURE A.7 : Valeurs historiques trimestrielles de l'indice FFB du coût de la construction - Base 1941 (FFB (2023))

A.5 Rappel sur le tau de Kendall

En statistique, le coefficient de corrélation de Kendall, plus communément appelé τ (tau) de Kendall, est une approche non paramétrique permettant de mesurer la corrélation entre deux variables continues, transformées en variables des rangs, ou ordinales appariées. Ce dernier correspond à une mesure de la corrélation des rangs (*i.e.* la similarité des ordres des données lorsqu'elles sont classées par chacune des quantités) et repose sur le principe de concordance (*cf.* figure A.8) .

Formellement, ce principe peut être illustré en considérant n paires de données (x_i, y_i), avec $i = 1, 2, \ldots, n$. Une paire (x_i, y_i) est dite concordante avec une autre paire (x_j, y_j) si les ordres de x_i et x_j sont les mêmes et les ordres de y_i et y_j sont également les mêmes, ou si les ordres de x_i et x_j sont différents et les ordres de y_i et y_j sont également différents (*i.e.* $x_i < x_j$ et $y_i < y_j$ ou $x_i > x_j$ et $y_i > y_j$). Sinon, la paire est considérée comme discordante. Dans le cas où $x_i = x_j$ ou $y_i = y_j$, la paire n'est ni concordante ni discordante. Finalement, le τ de Kendall est alors calculé par la formule suivante

$$\tau = \frac{\text{Nombre de paires concordantes} - \text{Nombre de paires discordantes}}{\frac{1}{2}\, n\,(n-1)}$$

où n est le nombre total de paires de données.

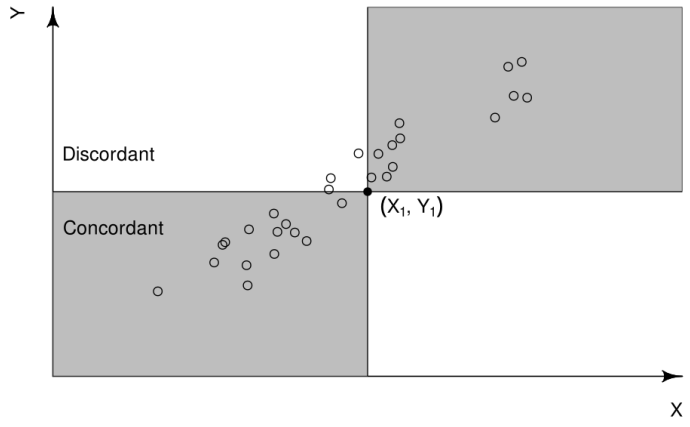


FIGURE A.8 : Illustration du principe de concordance

La valeur de τ varie entre -1 et 1. Un τ proche de 1 indique une forte corrélation positive, un τ proche de -1 indique une forte corrélation négative, et un τ proche de 0 indique une corrélation faible voire nulle entre les variables.

A.6 Proportion départementale des maisons

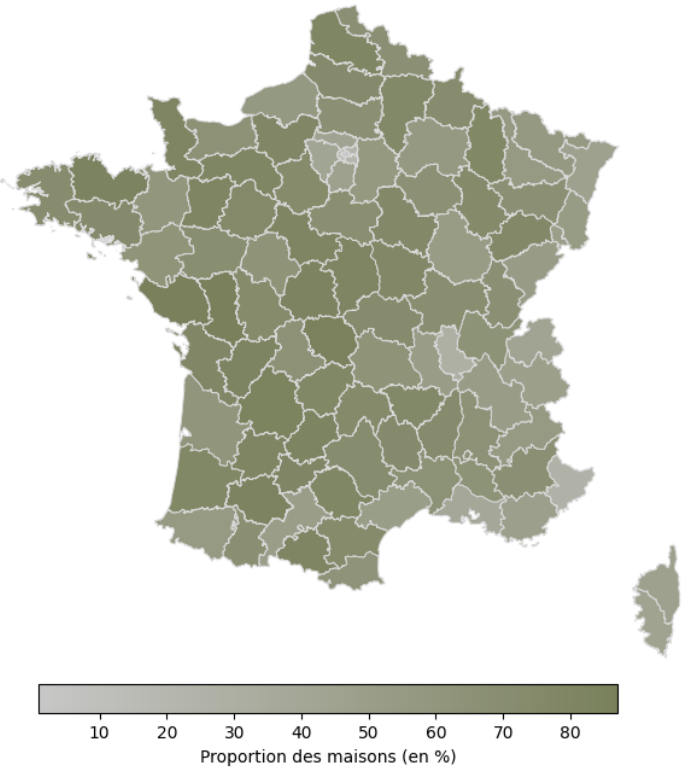


FIGURE A.9 : Proportion départementale des maisons au sein du portefeuille