Mémoire présenté devant le jury de l'EURIA en vue de l'admission à l'Institut des Actuaires

9 septembre 2022

Par : Hamadi Bahri
Titre : **Estimation de la charge sinistres Sécheresse pour Groupama d'OC en période d'arrêté des comptes**

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Romain NOBIS
Cécile VIGOUROUX

Signatures :
Membre présent du jury de l'EURIA :
Philippe LENCA
Signature :

Entreprise :
Groupama d'OC
Signature :

Groupama d'Oc
Direction Technique
14 Rue Vidailhan - CS 93105
31131 BALMA Cedex

Directeur de mémoire en entreprise :
David GUILLEMOT
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Mots clefs: Sécheresse, CatNat, Provisionnement, Arrêté des comptes, GLM, Arbres de décisions, Forêts aléatoires, Multi-Risques Climatiques sur Récoltes.

Dans un contexte de sinistralité croissante liée aux aléas naturels, probablement en lien avec le dérèglement climatique, la Cour des comptes [1] a proposé au gouvernement de sortir la Sécheresse du régime d'indemnisation des catastrophes naturelles. L'intensification et le coût très élevé de ces événements Sécheresse en font un risque de plus en plus prégnant. Des études plus importantes au sein des compagnies d'assurance doivent être réalisées sur ce risque afin de s'en prémunir.

A ce jour, la Sécheresse géotechnique est caractérisée par le phénomène de retrait gonflement des sols argileux, via l'alternance de fortes chaleurs qui entraînent le retrait du sol, suivies de fortes pluies qui entraînent son gonflement. Ces successions de mouvements, du fait des variations de la teneur en eau du sol, génèrent des dégâts sur les bâtiments (pouvant les rendre non habitables ou inexploitables).

A cet effet, les assureurs interviennent dans le remboursement de travaux de solidification [2] ou de relocalisation [3] dès lors qu'un arrêté catastrophe naturelle déclarant l'état de la Sécheresse est publié pour la commune dans le Journal Officiel (noté également JO pour la suite).

En lien avec l'actualité de cette année 2022, la Sécheresse touche aussi durement les exploitations agricoles et représente un des risques majeurs auquel les agriculteurs sont confrontés. Pour protéger leurs cultures et leurs récoltes, ces derniers peuvent souscrire une assurance Multi-Risques Climatiques. Ce contrat d'assurance, comme son nom le précise, couvre différents aléas climatiques, dont la Sécheresse. Dès lors, et sans aucune publication au JO (à la différence des bâtiments), une indemnisation peut être réalisée en cas de perte de rendement sur la culture agricole ayant été altérée par l'évènement.

Le but de ce mémoire est d'étudier la corrélation entre la Sécheresse qui touche les bâtiments et la Sécheresse affectant les cultures agricoles. Une corrélation qui pourrait permettre ainsi d'anticiper, lors des arrêtés des comptes, les publications au JO et le potentiel coût de la Sécheresse sur les bâtiments pour le portefeuille de Groupama d'OC.

La première partie du mémoire s'attachera à présenter l'impact de la Sécheresse dans les résultats techniques de Groupama d'OC.

La deuxième partie est consacrée à l'étude de la corrélation des évènements Sécheresse sur les cultures agricoles et la parution des arrêtés Sécheresse au JO. Une estimation prédictive des arrêtés catastrophes naturelles Sécheresse pour le territoire de Groupama d'OC sera ainsi réalisée. Avec ces modélisations et en estimant le coût moyen d'un arrêté, une évaluation de la charge sinistres de la Sécheresse pour le portefeuille de Groupama d'OC sera présentée pour une année donnée.

Enfin, un regard critique sur les travaux réalisés sera proposé dans la dernière partie du mémoire.

1. Cour des comptes. Sols argileux et Catastrophes Naturelles. Février 2022. (5)
2. Agrafes des mûrs fissurés, installation de micropieux, ...
3. En cas de risque d'effondrement, sans attendre la publication d'un arrêté dans le JO

Abstract

Keywords: Drought, CatNat, Reserving, Closing of accounts, GLM, Decision trees, Random forests, Multi-Climatic Risks on Harvests.

In a context of increasing claims due to natural hazards, probably linked to climate change, the Court of Auditors [4] proposed to the government to remove drought from the natural disaster compensation scheme. The intensification and very high cost of these Drought events make them an increasingly significant risk. More important studies within insurance companies must be carried out on this risk in order to guard against it.

To date, geotechnical drought is characterized by the phenomenon of shrinkage swelling of clay soils, via the alternation of high temperatures which cause the soil to shrink, followed by heavy rains which cause it to swell. These successions of movements, due to variations in the water content of the soil, generate damage to the buildings (which may make them uninhabitable or unusable).

To this end, insurers intervene in the reimbursement of solidification work [5] or relocation [6] when a natural disaster order declaring the state of Drought is published for the municipality in the Official Journal (also noted OJ for the following).

In connection with the news of this year 2022, the Drought is also hard on the farms and represents one of the major risks that farmers face. To protect their crops and harvests, they can take out Multi-Risk Climate insurance. This insurance contract, as its name indicates, covers various climatic hazards, including drought. Therefore, and without any publication in the OJ (unlike buildings), compensation can be made in the event of loss of yield on the agricultural crop that has been altered by the event.

The purpose of this dissertation is to study the correlation between Drought affecting the buildings and Drought affecting agricultural crops. A correlation which could thus make it possible to anticipate, during the closing of the accounts, the publications in the Official Journal and the potential cost of the Drought on the buildings for the portfolio of Groupama d'OC.

The first part of the thesis will focus on presenting the impact of the Drought on the technical results of Groupama d'OC.

The second part is devoted to the study of the correlation of drought events on agricultural crops and the publication of drought decrees in the Official Journal. A predictive estimate of drought natural disaster decrees for the territory of Groupama d'OC will thus be carried out. With these models and by estimating the average cost of a statement, an assessment of the Drought claims expense for the Groupama d'OC portfolio will be presented for a given year.

Finally, a critical look at the work carried out will be proposed in the last part of the thesis.

4. Court of Auditors. Clay Soils and Natural Disasters. February 2022. (5)
5. Staples of cracked walls, installation of micropiles, etc.
6. In the event of risk of collapse, without waiting for the publication of an order in the OJ

Note de synthèse

Mots clefs: Sécheresse, CatNat, Provisionnement, Arrêté des comptes, GLM, Arbres de décisions, Forêts aléatoires, Multi-Risques Climatiques sur Récoltes.

Introduction

En France, pour lutter contre les évènements climatiques, deux régimes d'indemnisation existent. L'un, mis en place depuis 1982, est basé sur le principe de solidarité nationale devant les catastrophes naturelles. Dans le cadre de ce régime, un sinistre de type catastrophe naturelle (noté CatNat par la suite) est indemnisé dès lors qu'un arrêté reconnaissant la survenance du péril est publié dans le Journal Officiel (JO).

La Sécheresse fait partie des nombreux aléas climatiques couverts par ce régime. Ce risque est caractérisé par le phénomène de retrait gonflement des sols argileux, via l'alternance de fortes chaleurs qui entraînent le retrait du sol, suivies de fortes pluies qui entraînent son gonflement. Ces successions de mouvements, du fait des variations de la teneur en eau du sol, génèrent des dégâts sur les bâtiments (pouvant les rendre non habitables ou inexploitables).

Dans un contexte de sinistralité croissante, probablement en lien avec le dérèglement climatique, et l'impact récurrent des sécheresses sur les résultats de GOC, une réflexion a été engagée quant aux méthodes utilisées pour estimer la charge sinistres de ce péril en période d'inventaire et donner une meilleure appréciation de la rentabilité économique de certains produits couvrant ce risque.

Contexte et Problématique

Aujourd'hui, pour que l'état de Sécheresse soit déclaré dans une commune, deux critères doivent être respectés :

— un critère de prédisposition géologique (disponible à tout moment, via les travaux de cartographie réalisés par le Bureau de Recherche Géologique et Minières),

— un critère déclenchant météorologique (le SWI [7] uniforme).

Ce second critère n'est pas disponible immédiatement. En effet, Météo France ne le diffuse qu'après la publication des arrêtés dans le Journal Officiel (JO). Malheureusement, les parutions des arrêtés CatNat Sécheresse se font essentiellement après l'exercice d'inventaire N (généralement après juillet N+1).

Ce délai génère une incertitude dans l'évaluation du provisionnement de ce risque. L'objectif de ce mémoire est d'apporter une réponse à cette problématique en période d'arrêté des comptes.

En parallèle et au cœur de l'actualité 2022, la Sécheresse touche aussi durement les exploitations agricoles et représente un des risques majeurs auquel les agriculteurs sont confrontés. Pour protéger

7. Soil Wetness Index

leurs cultures et leurs récoltes, ces derniers peuvent souscrire une assurance Multi-Risques Climatiques. Ce contrat d'assurance (comme son nom le précise) couvre différents aléas climatiques, dont la Sécheresse. Dès lors, et sans aucune publication au JO (à la différence des bâtiments), une indemnisation peut être réalisée en cas de perte de rendement sur la culture agricole ayant été altérée par l'évènement.

Objectif

Le but de ce mémoire et son originalité est d'étudier la corrélation entre la Sécheresse qui touche les bâtiments et la Sécheresse affectant les cultures agricoles. Une corrélation qui pourrait permettre ainsi d'anticiper, lors des arrêtés des comptes, les publications au JO et le potentiel coût de la Sécheresse sur les bâtiments pour le portefeuille de Groupama d'OC.

Construction d'une base de données

Pour chaque individu (commune), la base d'étude a été constituée de trois types de données :

— les données historiques des arrêtés CatNat Sécheresse publiés sur le territoire de GOC,

— les données géologiques et géographiques de chaque commune,

— les données sinistres en assurance Multi-Risques Climatiques sur Récoltes.

Pour une année et commune donnée, il est possible qu'aucune donnée MRC ne soit disponible. Cette indisponibilité peut être due à la non présence de GOC dans la commune ou bien à la non présence de cultures agricoles. La majorité des valeurs manquantes représentent des communes purement urbaines. Par exemple, la commune de Toulouse, ne compte aucune culture agricole donc logiquement n'a aucune donnée MRC.

Afin de pallier cette problématique, la notion de Région Agricole [8] (RA) a été exploitée. Ainsi, par analogie, il a été affecté aux communes sans information, la moyenne des données présentes dans la RA.

Les premières analyses ont permis de s'interroger sur le maillage de modélisation. Entre Commune (maillage qui serait logique puisque les arrêtés sont publiés par commune, mais la robustesse des données peut être faible pour plusieurs d'entre elles) et Département (prédire un nombre d'arrêtés dans un département génère finalement peu « d'individus » pour faire un travail de modélisation), d'autres niveaux existent. Par inclusion, ils peuvent être présentés ainsi (le nombre entre parenthèses représente le nombre d'individus possibles) :

Commune (4471) $\subset$ Code Postal (597) $\subset$ PRA (114) $\subset$ RA (72) $\subset$ Département (14)

Compte tenu des 10 années d'historique utilisées, le Code Postal [9] s'est vu être le maillage optimal.

Modélisation de la fréquence Sécheresse

Afin de modéliser la publication d'un arrêté CatNat Sécheresse favorable, la base de données est scindée en deux sous bases : une base d'apprentissage et une base de test.

Pour prédire la publication d'au moins un arrêté CatNat Sécheresse dans un Code Postal, trois modèles statistiques ont été proposés et élaborés sur la base d'apprentissage : un modèle de régression

8. D'après GéoConfluences, "Les régions agricoles (RA) constituent, en France, deux entités d'échelle différente du zonage statistique, géré par l'Insee et lancé en 1949 par le Commissariat général au Plan. Il s'agit de zones agricoles homogènes, tant par la nature des sols que pour les conditions climatiques et la vocation dominante des exploitations agricoles. " (7)

9. Dès lors, un code postal qui regroupe 5 communes et qui peut avoir 2 ou 3 arrêtés CatNat ressortira dans la classe 1 (au moins un arrêté CatNat Sécheresse publié dans le Code Postal).

logistique, un modèle d'arbres de décision et un modèle de forêts aléatoires.
Le tableau ci-dessous synthétise les résultats obtenus sur la base de test :

Modèle	Rappel	Précision	F1-score	AUC
Régression Logistique	78%	61%	69%	84%
Arbre de décision	72%	58%	65%	79%
Forêt aléatoire	78%	56%	65%	76%

TABLE 1 – Tableau récapitulatif des métriques obtenus.

Le modèle de régression logistique affiche les meilleurs résultats. Ce modèle est plus précis que les deux autres modèles et repère aussi bien les arrêtés Sécheresse que le modèle de forêt aléatoire. Il a logiquement le meilleur F1 Score. Le modèle de forêt aléatoire surestime le nombre d'arrêté et manque de précision par rapport aux autres modèles. Le modèle d'arbres de décision, quant à lui, se rapproche de la précision du modèle de régression logistique, mais ne repère pas autant de Sécheresse que ce dernier. En plus de ces éléments, l'AUC du modèle de régression logistique est la plus élevée. Il a donc été décidé de choisir le modèle de régression logistique pour la suite des travaux.

Afin de mieux tester le modèle de régression logistique, un backtesting [10] est réalisé sur les données d'apprentissage. La figure ci-dessous présente les résultats obtenus :

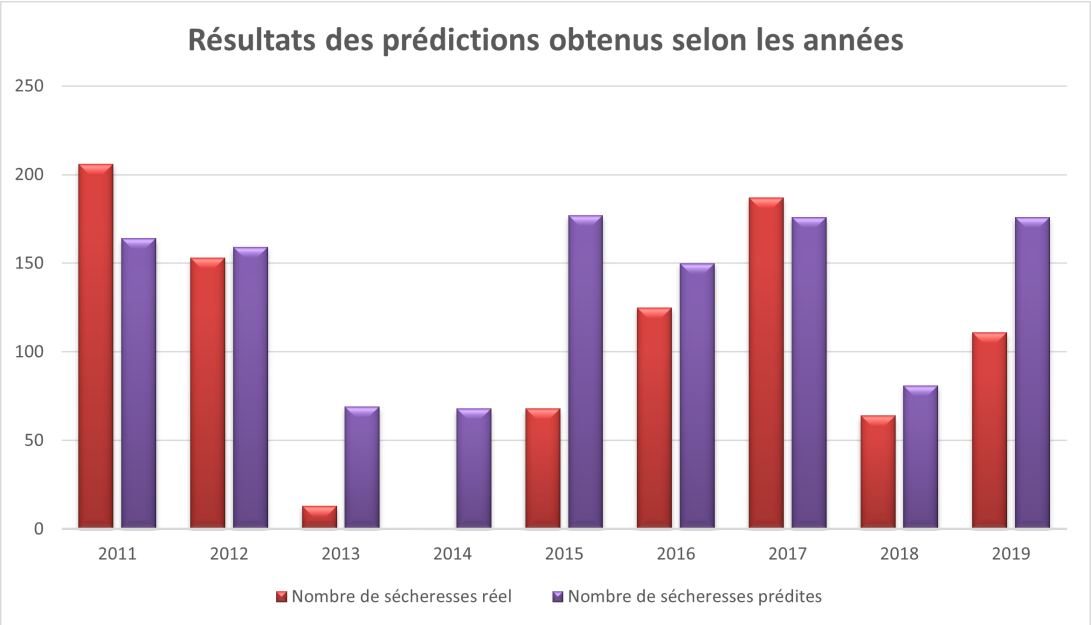

FIGURE 1 – Résultats des prédictions du modèle par année

Le backtesting réalisé montre que les prédictions du modèle s'approchent de la réalité avec toutefois des écarts sur certains exercices. Cependant, il faut faire attention à la lecture de ce graphique. En effet, le nombre de sécheresses prédites est défini selon un seuil unique à l'ensemble des années. C'est un biais qui sera pallier dans la suite de l'étude en prenant la probabilité elle-même dans l'estimation de la charge sinistres.

10. évaluer la précision du modèle en utilisant les données sur lesquelles celui-ci a été élaboré

Estimation du CTP [11] Sécheresse

Après avoir réalisée la modélisation de la fréquence des évènements Sécheresse sur le territoire de GOC, et dans le but d'estimer la charge sinistres relative à ce péril, un coût moyen d'une Sécheresse doit être défini. Différentes méthodes ont été testées :

Méthode	**Description**	**Formule**
Méthode 1	Coût moyen global pour un arrêté CatNat Sécheresse	$CM_{Sécheresse} = \frac{CTP_{sécheresse}}{Nombres\_total\_arreté\_sécheresse}$
Méthode 2	Coût moyen qui dépend du département	$CM_{Sécheresse}(Dept) = \frac{CTP_{sécheresse,Dept}}{Nombres\_total\_arreté\_dans\_le\_dept}$
Méthode 3	Coût moyen par Entité Assurée (EA)	$CM_{sécheresse,EA} = \frac{CTP_{scheresse}}{Nombre\_total\_EA}$
Méthode 4	Coût moyen en fonction du S/C	$CM_{Sécheresse}(CP) = \frac{S}{C}(Dept_{CP}) * Cotisations\_CatNat$
Méthode 5	Coût moyen selon un S/C de destruction fonction des probabilités de survenance d'une sécheresse	$CM_{Sécheresse} = Cotisation\_CatNat\_CP * \frac{S}{C}de\_destruction$

Table 2 – Formules d'estimation du coût moyen selon les méthodes

Le calcul des CTP prédit pour chaque méthode est présenté dans le tableau suivant :

Méthode	**CTP Prédit**
Méthode 1	$CTP_{prédit} = CM_{sécheresse} * Nombre\_Totale\_Sécheresses\_Prédites$
Méthode 2	$CTP_{prédit} = \sum_{Dept} CM_{sécheresse,CP}(Dept) * Nombre\_Sécheresses\_Prédites\_par\_dept$
Méthode 3	$CTP_{prédit} = \sum_{CP} \mathbb{1}_{P(CP=1>Seuil)} * Nombre\_de\_EA\_dans\_le\_CP * CM_{Scheresse,EA}$
Méthode 4	$CTP_{prédit} = \sum_{CP} P(CP=1) * CM_{Scheresse}(CP)$
Méthode 5	$CTP_{prédit} = \sum_{CP} P(CP=1) * Cotisation\_CatNat\_CP * \frac{S}{C}de\_destruction$

Table 3 – Formules d'estimation du CTP Sécheresse

Avec :

— $CM_{Sécheresse}$: coût moyen d'une Sécheresse sur bâtiments,

— $P(CP=1)$: probabilité, renvoyée par le modèle, de la publication d'au moins un arrêté CatNat dans le code postal

— $\sum_{CP} \mathbb{1}_{P(CP=1>Seuil)}$: la somme des indicatrices fonction de la valeur des probabilités

11. Charge Totale Probable (CTP = Règlements + Provisions - Recours - Prévisions de recours)

— S/C : Quotient des montants des sinistres CatNat payés (plus précisément du CTP) et les cotisations CatNat perçues par GOC

Dès lors, la charge sinistres de chaque méthode peut être comparée ci-dessous :

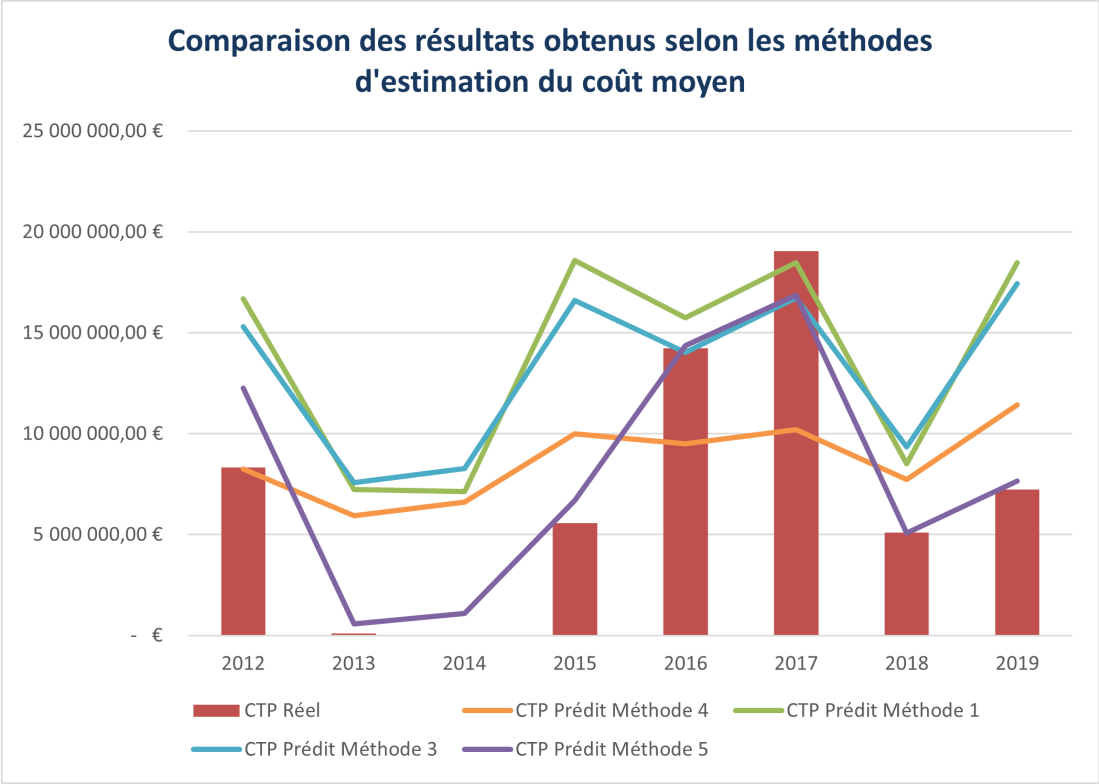

FIGURE 2 – Comparaison des résultats obtenus selon les méthodes d'estimation du coût moyen

Au regard des résultats, la méthode 5 parait être la méthode la plus adéquate pour estimer le CTP Sécheresse en période d'arrêtés des comptes. En effet, cette méthode tient à la fois compte :

— des probabilités de survenance d'un arrêté CatNat Sécheresse (aucune définition de seuil),

— des sommes assurées selon les codes postaux,

— de la sévérité des évènements.

Limites et sensibilités du modèle

Les estimations obtenues ne sont pas parfaites. En effet, les modèles développés présentent quelques faiblesses :

— Tout d'abord, un historique de 10 ans peut sembler important, mais selon différents échanges, c'est un historique relativement court en climatique. Les modèles gagneraient (et gagneront) en performance avec l'ajout d'années complémentaires,

— La non-intégration dans le modèle du caractère social et économique de la sollicitation de la commission interministérielle,

— Enfin, et le point n'est pas neutre, le modèle permet d'évaluer une charge à une date N, mais cette évaluation n'évolue plus par la suite. En effet, les données en entrée de modèles sont figées une fois l'exercice d'inventaire fini. Les déclarations de sinistres arrivant en juillet N+1, GOC utilise, à ce moment-là, différentes méthodes d'évaluations de charge sinistres, dont la méthode Chain-Ladder afin d'évaluer un CTP potentiellement plus adéquate.

Perspectives

Face à la forte progression des évènements Sécheresse et de leurs impacts sur les compagnies d'assurance, de réassurance et les pouvoirs publics, la Cour des comptes [12] a été saisie par le Comité d'Evaluation et de Contrôle (CEC) des politiques publiques de l'Assemblée Nationale.
A l'issue de son enquête, la Cour a publié un rapport en février 2022 dans lequel elle présente ses remarques et recommandations quant à ce phénomène. De plus, une réforme sur la loi nommée "ELAN" [13] a également été introduite. Ces travaux de réformes (toujours en cours) laissent entrevoir de meilleures perspectives dans la maîtrise et la gestion de ce risque.

En parallèle, sous l'égide du Ministère de l'Agriculture et de la Souveraineté Alimentaire, l'ordonnance n°2022-1075 a été circularisée le 29 juillet 2022. Dans cette ordonnance, les subventions prévues dans le cadre de l'assurance Multi-Risques Climatiques sur Récoltes sont en cours de réforme. L'idée et l'objectif de cette évolution est d'inciter davantage les exploitants agricoles à se protéger face à la croissance et l'intensité des évènements climatiques de toute nature.
Cette réforme laisse entrevoir, pour le modèle mis en place, une plus grande robustesse dans les données MRC utilisées et de fait, de meilleures évaluations à venir.

Conclusion

Les incertitudes liées au changement climatique de la planète et l'impact récurrent des sécheresses sur les résultats de GOC ont conduit à engager une réflexion quant aux méthodes utilisées pour estimer la charge sinistres de ce péril en période d'inventaire.

Jusqu'à présent, des évaluations externes et des méthodes basées sur des triangles de liquidation étaient utilisées pour provisionner le risque Sécheresse. Le fait de ne pas maîtriser les évaluations externes était une problématique pour GOC.

L'objectif de ce mémoire était de proposer de nouvelles approches plus adaptées à ce risque en période d'inventaire N (sans connaissance des arrêtés CatNat Sécheresse publiés à partir de juillet N+1).
Ainsi, l'originalité de l'étude résidait non seulement dans l'utilisation de nouveaux types de modèles, mais, avant tout, dans le remplacement du second critère de déclenchement (le SWI uniforme), par les données sinistres en MRC.

Aux vus des différents éléments apportés dans ce rapport, et même si des points mériteraient d'être améliorer, cette nouvelle approche répond globalement bien à l'évaluation d'un CTP en période d'inventaire pour un risque très erratique.
Le modèle est en cours d'utilisation depuis l'arrêté des comptes du premier semestre 2022.
L'estimation de la charge sinistres sur les exercices plus récents 2020 et 2021 semble cohérente avec les évaluations actuelles.

Ce mémoire s'est aussi intéressé à de possibles évolutions de ce risque, notamment à travers les recommandations de la Cour des comptes et la réforme en cours de la loi ELAN qui probablement pourraient conduire à externaliser la Sécheresse du régime CatNat. Au-delà des points d'améliorations mentionnés dans ce rapport, la perspective d'une élaboration tarifaire spécifique constituerait également un prolongement intéressant à ces travaux et à l'équilibre économique des produits assurantiels couvrant ce péril.

12. Cour des comptes. Sols argileux et catastrophes naturelles. 2022. (5)
13. Évolution du Logement de l'Aménagement et du Numérique

Synthesis note

Keywords: Drought, CatNat, Reserving, Closing of accounts, GLM, Decision trees, Random forests, Multi-Climatic Risks on Harvests.

Introduction

In France, to fight against climatic events, two compensation schemes exist. One, in place since 1982, is based on the principle of national solidarity in the face of natural disasters. Under this scheme, a natural disaster type claim (noted CatNat thereafter) is compensated when a decree recognizing the occurrence of the peril is published in the Official Journal (JO).

Drought is one of the many climatic hazards covered by this regime. This risk is characterized by the phenomenon of shrinkage and swelling of clay soils, via the alternation of high temperatures which cause the soil to shrink, followed by heavy rains which cause it to swell. These successions of movements, due to variations in the water content of the soil, generate damage to the buildings (which may make them uninhabitable or unusable).

In a context of increasing claims, probably linked to climate change, and the recurring impact of droughts on GOC's results, a reflection has been initiated on the methods used to estimate the claims charge of this peril during the inventory period and give a better appreciation of the economic profitability of certain products covering this risk.

Context and Issue

Today, for the state of Drought to be declared in a municipality, two criteria must be met :

— a geological predisposition criterion (available at any time, via the mapping work carried out by the Bureau of Geological and Mining Research),

— a meteorological triggering criterion (the uniform SWI [14]).

For this second criterion, the uniform SWI is not known. Indeed, Météo France does not broadcast it only after the publication of the decrees in the Official Journal (JO). Unfortunately, Drought decrees are published mainly after the N inventory exercise (generally after July N+1).

This delay generates uncertainty in the assessment of the provisioning of this risk. This is the heart of the matter. An answer will be provided to this problem through this thesis.

At the same time and at the heart of the news for 2022, drought is also hitting farms hard and represents one of the major risks that farmers face. To protect their crops and harvests, they can take out Multi-Risk Climate insurance. This insurance contract, as its name indicates, covers various climatic hazards, including drought. Therefore, and without any publication in the OJ (unlike buildings), compensation can be made in the event of loss of yield on the agricultural crop that has been altered

14. Soil Wetness Index

by the event.

Objective

The purpose of this dissertation and its originality is to study the correlation between drought affecting buildings and drought affecting agricultural crops. A correlation that could thus make it possible to anticipate, during the closing of the accounts, the publications in the Official Journal and the potential cost of the Drought on the buildings for the Groupama d'OC portfolio.

Building a database

For each individual (municipality), the study base was made up of three types of data :

— the historical data of CatNat Drought decrees published on the territory of GOC,

— the geological and geographical data of each municipality,

— claims data in Multi-Risk Climate Insurance on Crops.

For a given year and municipality, it is possible that no MRC data is available. This unavailability may be due to the low presence of GOC in the municipality or to the low presence of agricultural crops. The majority of missing values represent purely urban municipalities. For example, the municipality of Toulouse has no agricultural crops and therefore logically has no MRC data.

In order to overcome this problem, the concept of Agricultural Region [15] (RA) was exploited. Thus, by analogy, it was assigned to the municipalities without information, the average of the data present in the AR.

The first analyzes made it possible to question the modeling mesh. Between Municipality (mesh which would be logical since the decrees are published by municipality, but the robustness of the data may be weak for several municipalities) and Department (Predicting a number of decrees in a department ultimately generates few "individuals" to make a modeling work), other levels exist. By inclusion, they can be presented as follows (the number in brackets represents the number of possible individuals) :

Municipality (4471) $\subset$ ZIP Code (597) $\subset$ PRA (114) $\subset$ RA (72) $\subset$ Department (14)

Given the 10 years of history used, the Zip Code [16] was found to be the optimal mesh.

Drought Frequency Modeling

In order to model the publication of a favorable CatNat Drought decree, the database is divided into two sub-bases : a learning data and a test data.

In order to predict the publication of at least one CatNat drought decree in a Zip Code, three statistical models were proposed and developed on the learning basis : a logistic regression model, a decision tree model and a random forests.
The table below summarizes the results obtained on the test data.

15. According to GéoConfluences, "Agricultural regions (RA) constitute, in France, two entities of different scales of statistical zoning, managed by INSEE and launched in 1949 by the General Planning Commission. These are homogeneous agricultural areas, both in terms of the nature of the soil and the climatic conditions and the dominant vocation of the farms. " (7)

16. From then on, a Zip code which includes 5 municipalities and which may have 2 or 3 CatNat decrees will appear in class 1 (at least one CatNat Drought decree published in the Zip Code).

Model	Recall	Precision	F1-score	AUC
Logistic Regression	78%	61%	69%	84%
Decision tree	72%	58%	65%	79%
Random Forest	78%	56%	65%	76%

TABLE 4 – Summary table of the metrics obtained.

The logistic regression model shows the best results. This model is more precise than the other two models and identifies both Drought decrees and the random forest model. It logically has the best F1 Score. The random forest model overestimates the number of decrees and lacks precision compared to other models. The accuracy of the decision tree model is close to the accuracy of the logistic regression model but does not detect as much decrees as the latter. In addition to these, the AUC of the logistic regression model is the highest. It was therefore decided to choose the logistic regression model for the rest of the work.

In order to better test the logistic regression model, a backtesting [17] is performed on the training data.
The figure below shows the results obtained.

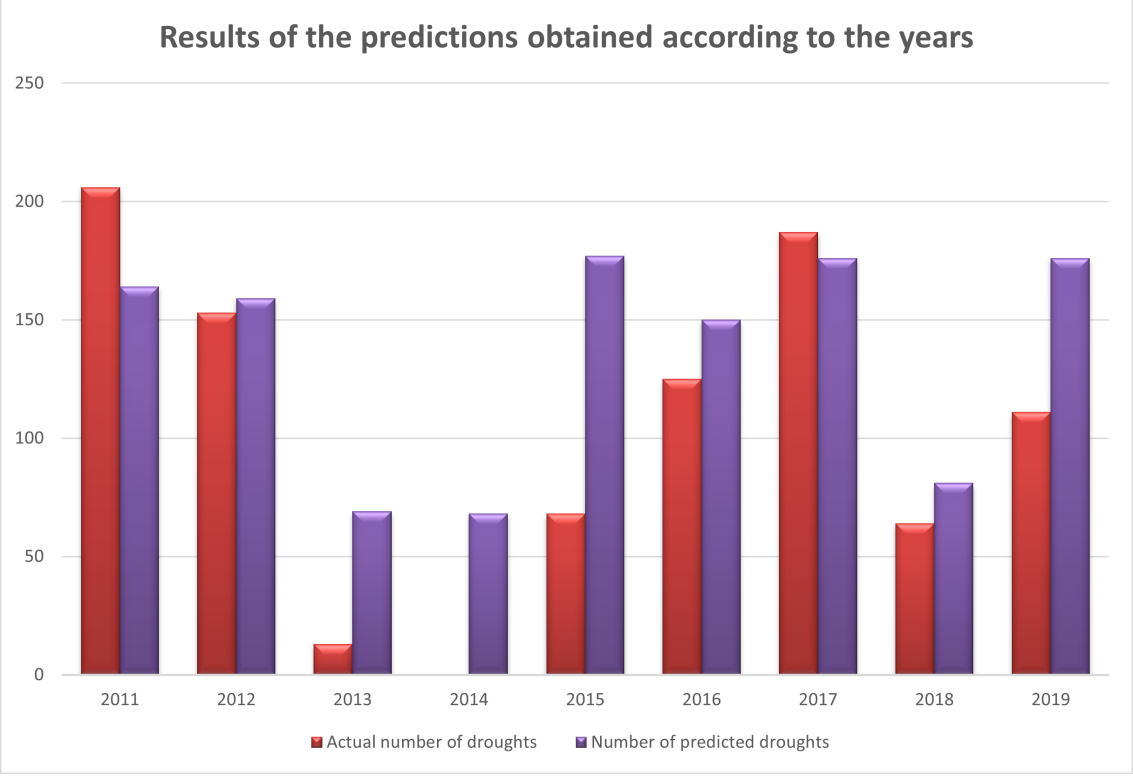

FIGURE 3 – Model prediction results by year

The backtesting carried out shows that the model's predictions are close to reality with, however, deviations in certain exercises. However, one must be careful when reading this graph. Indeed, the number of predicted droughts is defined according to a single threshold for all years. This is a bias that will be corrected in the rest of the study by taking the probability itself into the estimate of the claims charge.

17. evaluate the accuracy of the model using the data on which it was developed

Drought PTL [18] estimate

After modeling the frequency of Drought events on GOC territory, and in order to estimate the claims burden relating to this peril, an average cost of a Drought must be defined. Different methods have been tested :

Method	Description	Formula
Method 1	Overall average cost for a CatNat Drought decree	$AC_{Drought} = \frac{PTL_{Drought}}{total\_number\_of\_decreed\_drought}$
Method 2	Average cost which depends on the department	$AC_{Drought}(Dept) = \frac{PTL_{Drought,Dept}}{total\_number\_of\_decreed\_in\_dept}$
Method 3	Average cost per Insured Entity (IE)	$AC_{Drought,IE} = \frac{PTL_{Drought}}{total\_number\_of\_EA}$
Method 4	Average cost according to Loss Ratio	$AC_{Drought}(CP) = Loss\_Ratio(Dept_{ZIP}) * Contribution$
Method 5	Average cost according to a Loss Ratio destruction	$AC_{Drought} = ContributionP * Loss\_Ratio\_destruction$

Table 5 – Formulas for estimating the average cost according to the methods

The calculation of the predicted PTL for each method is presented in the following table :

Method	Predicted PTL
Method 1	$PTL_{predicted} = AC_{Drought} * number\_of\_predicted\_drought$
Method 2	$PTL_{predicted} = \sum_{Dept} AC_{Drought,ZIP}(Dept) * number\_of\_predicted\_drought\_per\_Dept$
Method 3	$PTL_{predicted} = \sum_{ZIP} \mathbb{1}_{P(ZIP=1>threshold)} * number\_of\_IE\_per\_ZIP * AC_{Drought,IA}$
Method 4	$PTL_{predicted} = \sum_{ZIP} P(ZIP = 1) * AC_{Drought}(ZIP)$
Method 5	$PTL_{predicted} = \sum_{ZIP} P(ZIP = 1) * Contribution\_CatNat\_ZIP * Loss\_Ratio\_destruction$

Table 6 – Drought PTL Estimation Formulas

With :

— $AC_{Drought}$: average cost of a Drought on buildings,

— $P(ZIP = 1)$: probability, returned by the model, of the publication of at least one CatNat order in the ZIP

— $\sum_{ZIP} \mathbb{1}_{P(CP=1>Threshold)}$: model prediction (0/1).

18. Probable Total Charge (PTL = Settlements + Provisions - Recourse - Recourse forecasts)

— Loss Ratio : Quotient Quotient of the amounts of CatNat claims paid (more precisely of the CTP) and the CatNat contributions collected by GOC

Therefore, the claims charge of each method can be compared below :

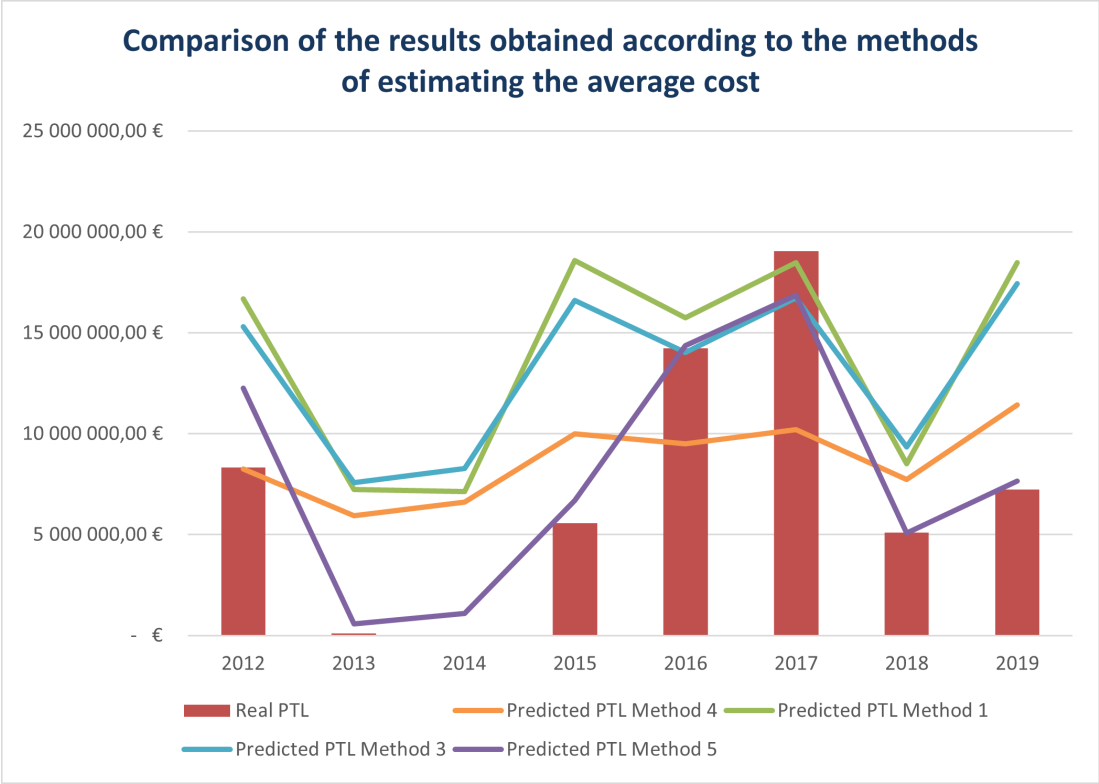

FIGURE 4 – Comparison of the results obtained according to the methods of estimating the average cost

In view of the results, method 5 seems to be the most appropriate method for estimating the PTL Drought during the period of closing the accounts. Indeed, this method takes into account both :

— of the probabilities of occurrence of a CatNat Drought order (no threshold definition),

— of the sums insured according to the postal codes,

— of event severity.

Model limits and sensitivities

The estimates obtained are not perfect. Indeed, the developed models have some weaknesses :

— First of all, a history of 10 years may seem important, but according to different exchanges, it is a relatively short history in climate. The models would gain (and will gain) in performance with the addition of additional years,

— The non-integration in the model of the social and economic nature of the solicitation of the interministerial commission,

— Finally, and the point is not neutral, the model makes it possible to evaluate a load on a date N, but this evaluation no longer evolves thereafter. Indeed, the model input data is frozen once the inventory exercise is finished. As claims declarations arrive in July N+1, GOC uses, at that time, different loss load assessment methods, including the Chain-Ladder method in order to assess a potentially more adequate PTL.

Outlook

Faced with the sharp increase in drought events and their impact on insurance and reinsurance companies and public authorities, the Court of Auditors [19] was seized by the Evaluation and Control Committee (CEC) of public policies of the National Assembly.
At the end of its investigation, the Court published a report in February 2022 in which it presents its observations and recommendations regarding this phenomenon. In addition, a reform on the law named "ELAN" was also introduced. This reform work (still in progress) suggests better prospects in the control and management of this risk.

At the same time, under the aegis of the Ministry of Agriculture and Food Sovereignty, ordinance no. 2022-1075 was circularized on July 29, 2022. In this ordinance, the subsidies provided for under the Multi -Climatic Risks on Harvests are being reformed. The idea and objective of this development is to further encourage farmers to protect themselves against the growth and intensity of climatic events of all kinds.
This reform suggests, for the model put in place, greater robustness in the MRC data used and, in fact, better evaluations to come.

The CatNat compensation scheme has shown some limitations after 40 years of existence. Changes seem necessary to alleviate the impact on insurance and reinsurance companies and public authorities. The ongoing reform work suggests better prospects for controlling and managing this risk.

Conclusion

The uncertainties related to global climate change and the recurring impact of droughts on GOC's results have led to a reflection on the methods used to estimate the loss burden of this peril during the inventory period.

Until now, external valuations and methods based on liquidation triangles were used to provision for drought risk. Not mastering external evaluations was a problem for GOC.

The objective of this dissertation was to propose new approaches more suited to this risk during the N inventory period (without knowledge of the CatNat Drought decrees published from July N+1).

Thus, the originality of the study resided not only in the use of new types of models, but, above all, in the replacement of the second triggering criterion (the uniform SWI), by disaster data in MRC.

In view of the various elements provided in this report, and even if some points deserve to be improved, this new approach generally responds well to the evaluation of a CTP during the inventory period for a very erratic risk.
The model has been in use since the closing of the accounts for the first half of 2022.
The estimate of the claims charge for the most recent years 2020 and 2021 seems consistent with the current assessments.
This thesis is also interested in possible evolutions of this risk, in particular through the recommendations of the Court of Auditors and the current reform of the ELAN law which probably could lead to outsourcing the Drought of the CatNat regime. Beyond the points for improvement mentioned in this report, the prospect of a specific tariff development would also constitute an interesting extension to this work and to the economic balance of insurance products covering this peril.

19. Court of Auditors. Clay soils and natural disasters. 2022. (5)

Remerciements

Je souhaite tout d'abord remercier toutes les personnes qui m'ont permis de mener à bien ce stage et ce projet de fin d'études.

Mes remerciements vont particulièrement à mon tuteur en entreprise **David Guillemot**, actuaire et responsable du pôle Pilotage Technique et Actuariat (PTA) chez Groupama d'OC, pour ses précieux conseils, et son accompagnement tout au long de ce stage. Je tiens aussi à remercier l'ensemble de l'équipe PTA, pour m'avoir accueilli parmi eux. Je remercie aussi **Yohan Fouet** Directeur Technique pour m'avoir accepté au sein de la direction. J'exprime également ma reconnaissance envers **William Nastro** ancien étudiant à l'EURIA et actuaire responsable du provisionnement, pour nos échanges et notre collaboration durant mon stage.

Je tiens également à remercier **Dominique Abgrall**, mon tuteur pédagogique pour son suivi, sa disponibilité et ses conseils. Je suis également reconnaissant envers **Franck Vermet**, Directeur de l'EURIA, ainsi que toute l'équipe pédagogique et académique de l'EURIA pour la qualité de la formation et des cours.

Enfin, merci à ma famille et mes amis pour leur soutien et leur aide.

Table des matières

Introduction Générale

Selon le rapport du « Global Risk Report 2022 » [20], la crise climatique est la plus grande menace de long terme à laquelle fait et fera face le monde. Selon ce rapport, ce risque se classe à la deuxième place des plus grandes menaces face auxquelles les compagnies d'assurance sont confrontées derrière le risque de cyberattaque et devant le risque de pandémie.
En France, pour lutter contre les évènements climatiques, deux régimes d'indemnisation existent. L'un, mis en place depuis 1982, est basé sur le principe de solidarité nationale devant les catastrophes naturelles. Dans le cadre de ce régime, un sinistre de type catastrophe naturelle (noté CatNat par la suite) est indemnisé dès lors qu'un arrêté reconnaissant la survenance du péril est publié dans le Journal Officiel (JO).

La Sécheresse fait partie des nombreux aléas climatiques couverts par ce régime [21]. Ces dernières années, la fréquence de ce péril est en nette croissance. Ces retours épisodiques plus récurrents remettent en question la couverture de ce risque. La Cour des comptes [22] a d'ailleurs proposé à l'état de sortir cet aléa du régime d'indemnisation des catastrophes naturelles.
A titre d'élément chiffré, selon la Caisse Centrale de Réassurance (appelé également CCR), la charge imputable des sinistres Sécheresse en 2019 s'établit déjà à 415 millions d'euros tandis que la charge sinistres de cet aléa sur le territoire français pourrait atteindre le milliard d'euros en 2020.
Comme d'autres assureurs, Groupama d'OC n'est pas épargné et est également impacté par le phénomène.

Au-delà des bâtiments impactés, la Sécheresse représente également un risque agricole. En effet, la faible humidité du sol, associée à la rareté de l'eau, arrête la croissance végétale et diminue les rendements des cultures agricoles. Pour protéger leurs récoltes, les agriculteurs peuvent souscrire à une assurance Multi-Risques Climatiques [23] (noté MRC pour la suite).

En parallèle, en période d'arrêté des comptes, si la charge sinistres Sécheresse en MRC peut être évaluée (voire même connue, car les indemnisations sont réalisées quasiment dans l'année d'inventaire), la charge sinistres sur bâtiments est une estimation avec une forte incertitude car aucune donnée sinistre n'est disponible sur l'exercice courant N. En effet, l'une des spécificités de ce péril réside dans le fait que les sinistres sont déclarés par les assurés dès lors qu'un arrêté CatNat est publié au JO. Or, ces arrêtés commencent à être publiés généralement en juillet N+1. C'est ainsi qu'une idée à émerger au sein de GOC : n'y aurait-il pas une corrélation entre la Sécheresse en MRC et la Sécheresse sur bâti ? Si une relation existe entre ces deux risques, alors une anticipation des publications au JO pourrait

20. The global Risks Reports 2022, 17ème édition, Forum Économique Mondial

21. la Sécheresse géotechnique est caractérisée par le phénomène de retrait gonflement des sols argileux, via l'alternance de fortes chaleurs qui entraînent le retrait du sol, suivies de fortes pluies qui entraînent son gonflement. Ces successions de mouvements, du fait des variations de la teneur en eau du sol, génèrent des dégâts sur les bâtiments (pouvant les rendre non habitables ou inexploitables).

22. Cour des comptes. Sols Argileux et Catastrophes Naturelles. Février 2022. (5)

23. Ce contrat d'assurance couvre différents aléas climatiques, dont la Sécheresse. Dès lors, et sans aucune publication au JO (à la différence du régime CatNat pour les bâtiments), une indemnisation peut être réalisée en cas de perte de rendement sur une culture agricole dès lors que la croissance ait été altérée par un des aléas climatiques mentionnés au contrat d'assurance.

être réalisée, permettant par la suite d'évaluer une charge sinistres de la Sécheresse des bâtiments.

Le but de ce mémoire est donc de permettre à Groupama d'OC de provisionner le plus justement possible le risque Sécheresse, en période d'arrêté des comptes, à partir d'un modèle prédictif qui s'appuierait sur les données spécifiques de la MRC.

Pour répondre à cet enjeu, le mémoire s'articulera en 3 sections :

— La première partie du mémoire s'attachera à présenter le cadre et le contexte dans lesquels cette étude a été menée. Les fondements et principes du régime d'indemnisation des catastrophes naturelles ainsi que le mécanisme d'indemnisation des sinistres seront dressés.
En parallèle, l'impact de la Sécheresse dans les résultats techniques de Groupama d'OC sera exposé.

— La deuxième section sera consacrée à la construction de la base de données, les méthodes et les modèles prédictifs qui ont été testés. Chaque modèle utilisé fera l'objet en amont d'une présentation théorique.
Les différents résultats obtenus seront confrontés.

— Enfin, dans la dernière partie du mémoire, un regard critique sera proposé en présentant les forces et les faiblesses des travaux réalisés.
En pleine actualité 2022, des perspectives sur le péril Sécheresse seront également dressés.

Première partie

Cadre de l'étude et impact de la Sécheresse sur le portefeuille de GOC

Chapitre 1

Le risque Sécheresse

Introduction

Une catastrophe naturelle est un évènement résultant de l'exposition d'une population humaine et de leurs infrastructures (enjeu) à un risque naturel (aléa). Les catastrophes naturelles se divisent en trois sous-groupes de catastrophes [1] :

1. **Les catastrophes météorologiques :**
   - Tempête
   - Température extrême
   - Brouillard.
2. **Les catastrophes climatiques :**
   - Sécheresse
   - Feu incontrôlé
   - Vidange brutale de lac glaciaire
3. **Les catastrophes hydrologiques :**
   - Crue
   - Glissement de terrain
   - Action de la houle

Ce chapitre commence par présenter les catastrophes naturelles dans le monde, l'évolution de leurs fréquences et coûts au cours des dernières décennies. Ensuite, il propose de s'intéresser aux fondements et principes du régime d'indemnisation des CatNat en France. Enfin, il se focalise, dans sa dernière partie, sur le risque Sécheresse et ses spécificités.

1.1 Les catastrophes naturelles dans le monde

Selon le rapport "Atlas de la mortalité et des pertes économiques dues à des phénomènes météorologiques, climatiques et hydrologiques extrêmes (1970-2019)" publié par l'Organisation Météorologique Mondiale en 2021, le monde a enregistré 11 000 catastrophes naturelles entre 1970 et 2019.

La figure ci-dessous représente la répartition des évènements naturels selon l'aléa climatique dans le monde sur cette période.

1. ORGANISATION MÉTÉOROLOGIQUE MONDIALE. Atlas de la mortalité et des pertes économiques dues à des phénomènes météorologiques, climatiques et hydrologiques extrêmes (1970-2019). 2021. (10)

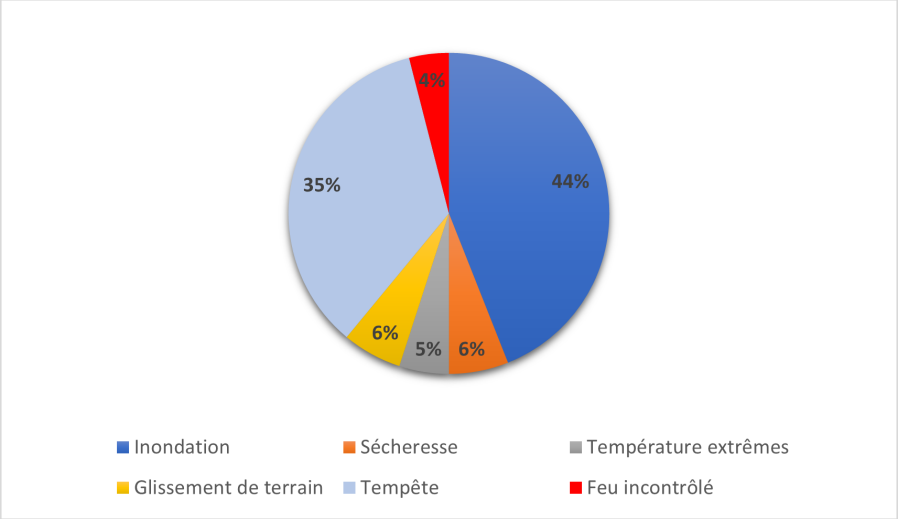

FIGURE 1.1 – Répartition des évènements naturels dans le monde selon l'aléa entre 1970 et 2019 (Source : Organisation Météorologique Mondiale)

Constats :

— Les inondations et les tempêtes représentent à elles seules 79% des aléas climatiques dans le monde.

— Les évènements Sécheresse sont moins fréquents et représentent seulement 6% des catastrophes naturelles.

1.1.1 Évolution de la fréquence et des coûts des catastrophes naturelles dans le monde

Les constats précédents sont également à interpréter de leurs évolutions dans le temps.
Dans un contexte de réchauffement climatique, la fréquence des catastrophes naturelles ne cesse d'augmenter ces dernières décennies.
La figure ci-dessous représente l'évolution du nombre de catastrophes naturelles dans le monde par décennie entre 1970 et 2019.

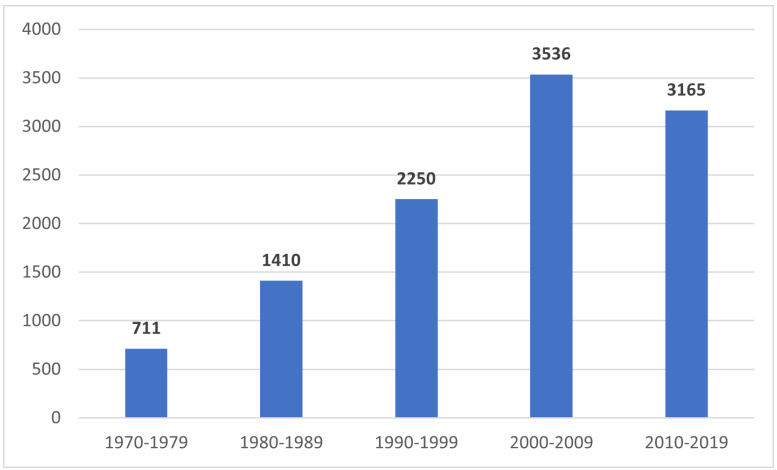

FIGURE 1.2 – Évolution du nombre de catastrophes naturelles par décennie dans le monde entre 1970 et 2019 (Source : Organisation Météorologique Mondiale)

Toujours selon le même rapport, durant la décennie 1970-1979, la moyenne du nombre d'évènements naturels dans le monde était de 71 par an. Ce chiffre a doublé durant la décennie suivante. Entre 2000

et 2019, plus de 670 catastrophes naturelles en moyenne par an ont frappé la planète et ses habitants. En plus des fréquences qui augmentent, le coût de ces évènements est conséquent et représente un enjeu économique important pour les gouvernements et les assureurs. A titre d'exemple, selon Statista [2], la Sécheresse de 2012 qui a touché le centre ouest américain a coûté 35 milliards de dollars aux assureurs américains, soit 1% de la croissance économique du pays.
La figure ci-dessous illustre l'évolution des coûts, en milliards de dollars, des catastrophes naturelles dans le monde par décennie entre 1970 et 2019.

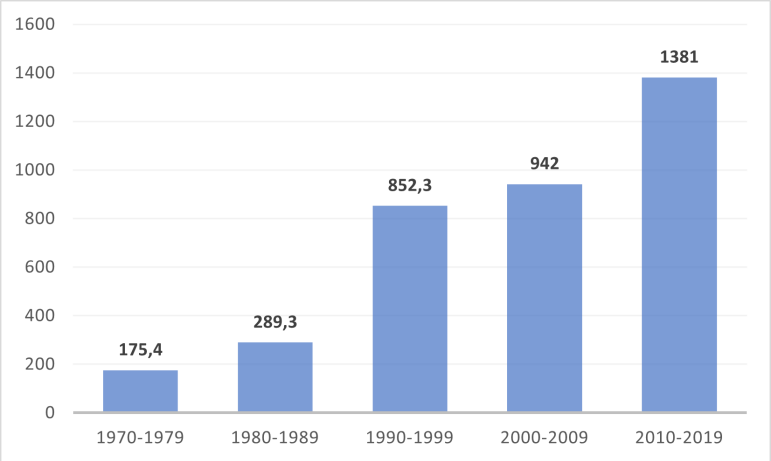

FIGURE 1.3 – Évolution des coûts des évènements naturels dans le monde par décennie en milliards de dollars entre 1970 et 2019 (Source : Organisation Météorologique Mondiale)

Constats :

— Les coûts des évènements naturels dans le monde ont été multipliés par 7 entre la première décennie (1970-1979) et la dernière (2010-2019) pour atteindre les 1381 milliards de dollars.

— L'année 2011 a été une année record avec une charge totale de 364 milliards de dollars.

1.2 Les catastrophes naturelles en France

La France n'est pas épargnée par les évènements naturels. Du fait de sa situation géographique, elle est exposée à de multiples scénarios naturels. D'après un rapport du Sénat en 2019 [3], un Français sur quatre est exposé à un risque climatique. Ce rapport fait état, d'ici à 2050, d'une augmentation probable de 50% du coût des sinistres liés aux évènements climatiques.
La couverture des évènements naturels, en France, se base sur trois garanties :

1. **La garantie Catastrophes Naturelles** : nommée aussi régime d'indemnisation CatNat. Une présentation plus détaillée de ce régime sera réalisée dans la section suivante.
2. **La garantie Tempête Grêle Neige** : notée aussi TGN, elle couvre les assurés face aux évènements tempête, grêle et neige.
3. **la garantie Climatiques sur récoltes** : appelée aussi assurance Multi-Risques Climatiques (MRC). Elle a pour objectif de couvrir les cultures des agriculteurs contre la destruction des récoltes.

D'après France Assureurs [4] (anciennement connu sous le nom de FFA), pour faire face aux catastrophes naturelles, le volume des cotisations perçues par l'ensemble des assureurs en 2020 a atteint 4 milliards d'euros. La charge des sinistres a, quant à elle, dépassé 3 milliards d'euros pour la même année.

2. STATISTA. Les catastrophes naturelles dans le monde. 2021. (11)
3. N. BONNEFOY. Rapport d'information fait au nom de la mission d'information sur la gestion des risques climatiques et l'évolution de nos régimes d'indemnisation. 2019 (2)
4. FRANCE ASSUREURS. L'assurance des évènements naturels en 2020. Étude - 2022. (6)

La figure ci-dessous illustre la répartition des cotisations et des charges sinistres selon le type de garantie.

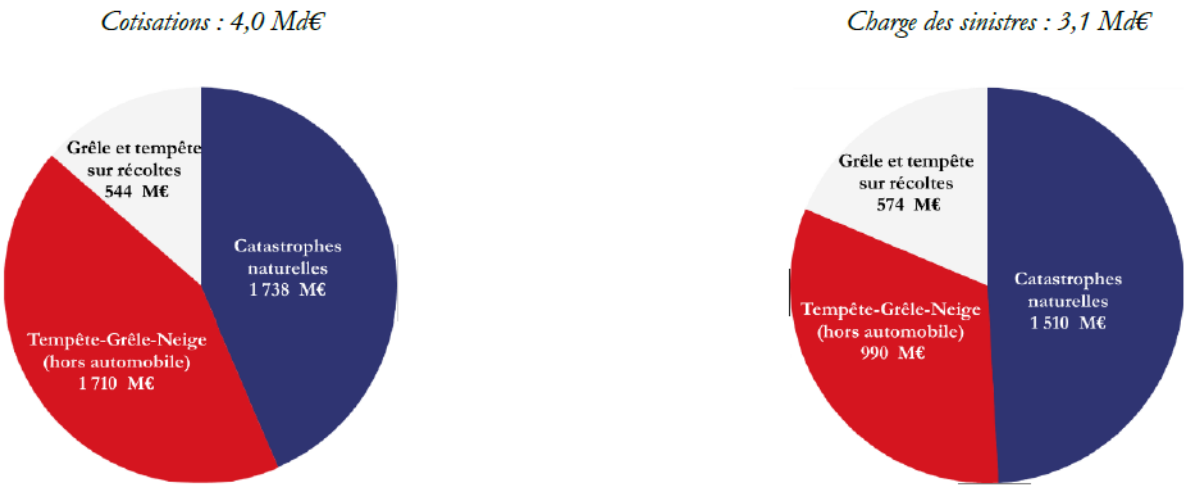

FIGURE 1.4 – Répartition des sinistres et cotisations de couverture selon la garantie (Source : France Assureurs)

Les cotisations sont réparties de la façon suivante : 43% à la garantie CatNat, 43% à la garantie TGN et 14% pour les récoltes. La moitié de la charge des sinistres est destinée à la garantie CatNat. En complément du dernier graphique et à titre d'information, la Figure 1.5 présente l'évolution des coûts des évènements climatiques en France depuis 1984.

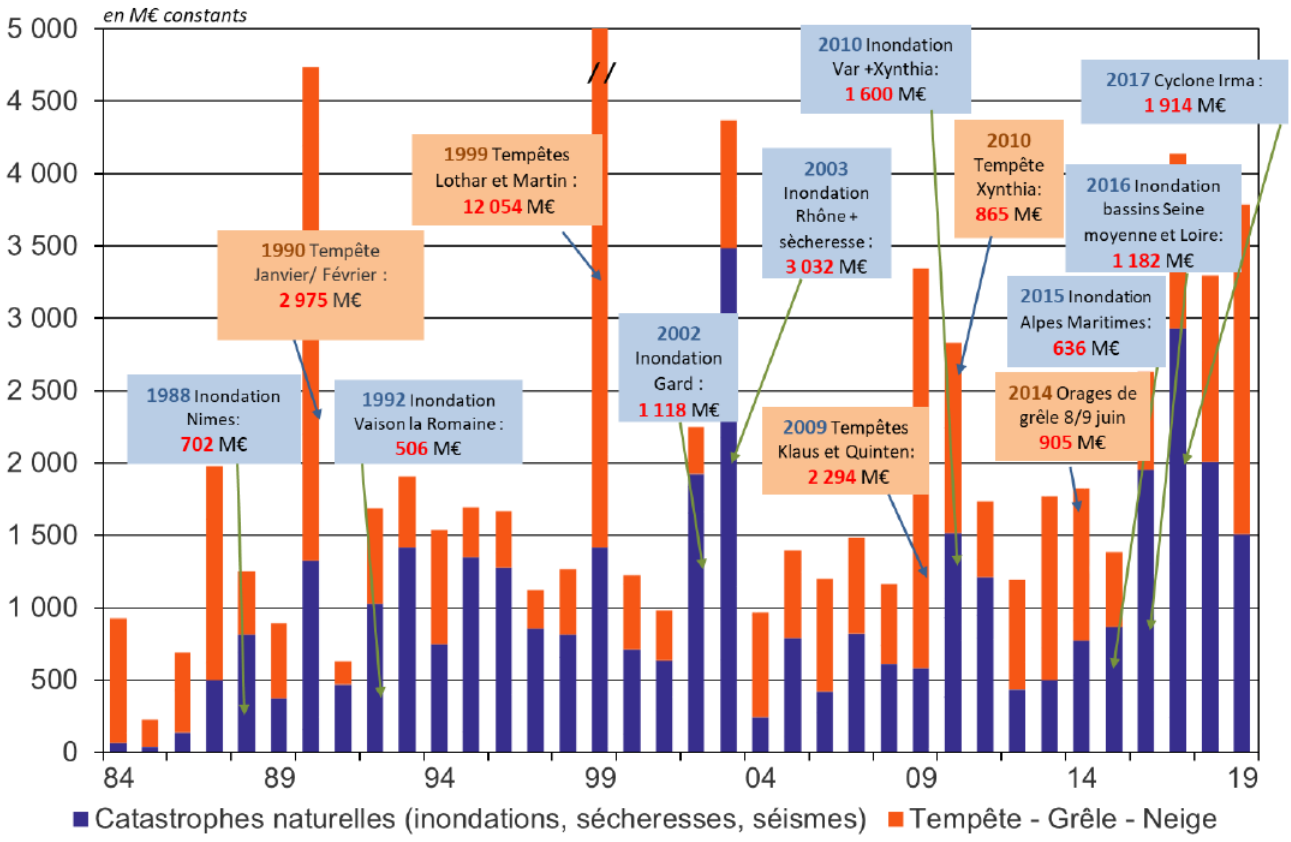

FIGURE 1.5 – Coût des évènements climatiques depuis 1984 (Source : France Assureurs)

Ces charges sinistres ne cessent d'augmenter au fil des années avec parfois des évènements atypiques tels que la Sécheresse de 2003 qui a coûté 3 milliards d'euros ou les tempêtes de Lothar et Martin de 1995 (12 milliards d'euros).

1.3 Le régime d'indemnisation des catastrophes naturelles [5]

Au début des années 80, la France a connu plusieurs événements naturels qui ont causé des dommages importants sur tout le territoire métropolitain. Jusque-là, aucune législation ne permettait d'indemniser les victimes de ces dégâts. Les risques climatiques, autres que les tempêtes et grêles, n'étaient pas encore assurables. Les pouvoirs publics ont alors pris la décision de créer un régime obligatoire d'indemnisation des catastrophes naturelles.

1.3.1 Fondement et principes du régime

Le régime d'indemnisation des catastrophes naturelles se base sur l'alinéa 12 du préambule de la Constitution de 1947 qui stipule : "La Nation proclame la solidarité et l'égalité de tous les Français devant les charges qui résultent des calamités nationales". C'est dans cet esprit que ce régime a été fondé sur deux piliers essentiels : Solidarité et Responsabilité.
Dans son rapport, la CCR (4) présente les enjeux de la mise en place de ce régime :

- couvrir tous les périls possibles et tous ceux impactés par les catastrophes naturelles : Particuliers, Entreprises et Collectivités,
- prôner le principe de solidarité en permettant que les charges globales des dégâts soient supportables par tous,
- accorder une importance à la prévention face aux risques climatiques pour essayer d'anticiper au mieux les évènements qui en découlent et éviter les dérives des coûts,
- mettre en place un système mixte qui implique les acteurs publics (comme les collectivités, l'état) et les acteurs privés comme les compagnies d'assurances,
- garantir une solvabilité et une pérennité du régime.

Pour financer ce régime et mettre en avant le principe de solidarité, le législateur a décidé d'instaurer un taux unique de prime additionnelle d'assurance. Le taux de surprime est de 12% pour les contrats dommages aux biens autre que les véhicules à moteur et de 6% des primes vol et incendie pour les véhicules à moteur.
Pour rendre solvable ce régime, la Caisse Centrale de Réassurance joue un rôle primordial. En effet, cette dernière propose aux assureurs des contrats de réassurance spécifiques aux catastrophes naturelles. De son coté, pour se couvrir, la CCR se base sur une réassurance illimitée de la part de l'État.

1.3.2 Mécanisme d'indemnisation des sinistres

Dans le cadre du régime CatNat, lors de la survenance d'un sinistre, l'assuré n'est pas indemnisé de façon classique, à savoir : déclaration d'un sinistre, suivi de la venue d'un expert, suivi du paiement du sinistre. Le processus d'indemnisation suit les étapes suivantes :

- L'assuré doit déclarer le sinistre CatNat à son assureur et prévenir la mairie,
- La mairie doit demander la reconnaissance de l'état de catastrophe naturelle auprès de la préfecture,
- La préfecture centralise les demandes communales et sollicite les rapports techniques,

5. Cette section est basée sur le rapport de la Cour des comptes : "Sols Argileux et Catastrophes Naturelles. 2022." (5) et les différents articles fournis par la CCR : "Caisse Centrale de Réassurance. Le régime d'indemnisation des catastrophes naturelles. 2011." (4)

— La direction générale de la sécurité civile et de la gestion des crises instruit et présente les dossiers à la commission interministérielle qui statue sur l'intensité anormale de l'agent naturel et émet un avis favorable, défavorable ou d'ajournement dans un arrêté interministériel qui est publié au Journal Officiel (JO),

— La préfecture communique aux mairies qui communiquent à leur tour aux sinistrés. Ces derniers ont 10 jours pour déclarer un sinistre,

— L'assureur a trois mois pour prendre en charge le dossier.

En cas d'extrême urgence, une procédure accélérée qui permet aux assurés d'être indemnisés plus rapidement peut être possible. Seul le gouvernement a le pouvoir de déclencher la nouvelle procédure. La figure ci-dessous présente avec plus de détails toutes les étapes du processus d'indemnisation des assurés dans le cadre du régime d'indemnisation des catastrophes naturelles.

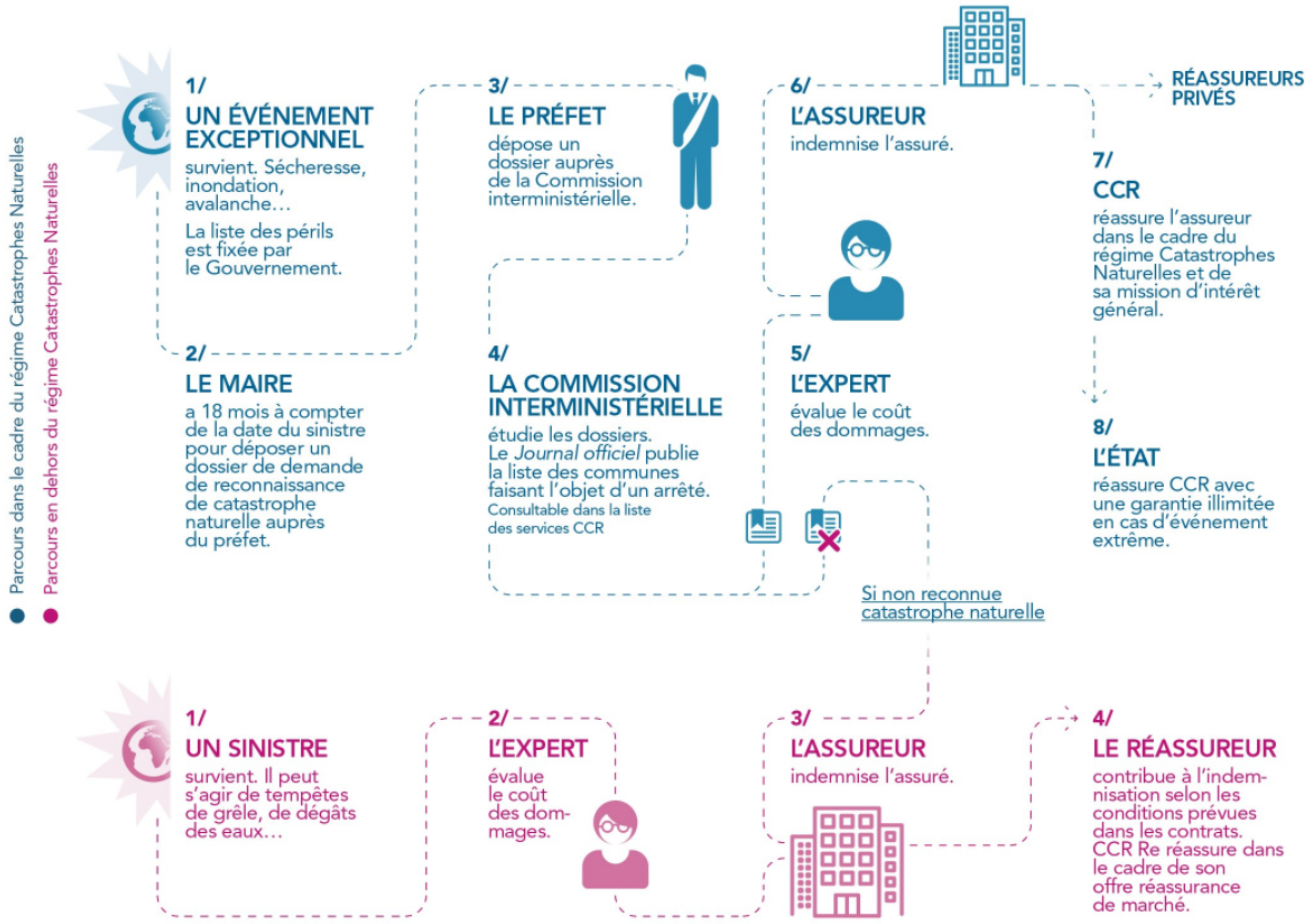

FIGURE 1.6 – Processus d'indemnisation des catastrophes naturelles (Source : CCR)

1.3.3 Les périls couverts

La CCR propose une liste non exhaustive des périls habituellement couverts dans le cadre de ce régime. Les aléas qui composent cette liste sont nombreux mais les plus importants d'entre eux sont :

— l'inondation,

— la sécheresse,

— le mouvement de terrains,

— le séisme,

— le raz de marée,

— l'avalanche.

Les tempêtes, la grêle et la neige ne sont pas couverts par ce régime.
La figure ci-dessous illustre la répartition des demandes des arrêtés CatNat acceptées par nature de phénomène entre 1984 et 2012.

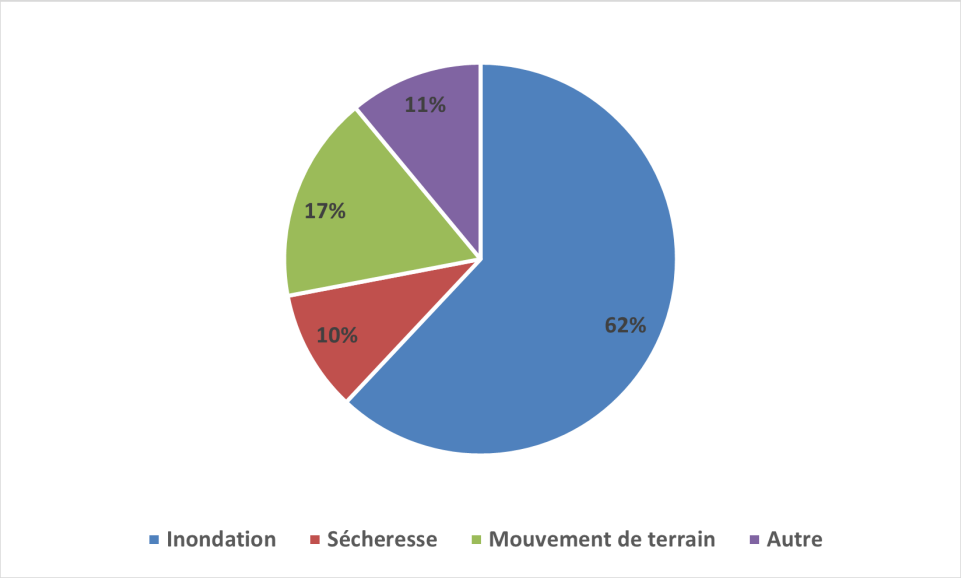

FIGURE 1.7 – Répartition des demandes d'arrêtés CatNat par nature de phénomène entre 1984 et 2012 (Source : CCR)

Constat :
La Sécheresse se classe à la troisième place des demandes acceptées.

1.4 Le risque Sécheresse

Le risque Sécheresse est classé comme un risque climatique. En effet, la vulnérabilité des bâtiments (l'enjeu) est liée à la variation des indices climatiques comme la température, la pluviométrie ou l'humidité (l'aléa).

1.4.1 La Sécheresse géotechnique

La Sécheresse géotechnique est caractérisée par le phénomène de Retrait Gonflement des sols Argileux (RGA), via l'alternance de fortes chaleurs qui entraînent le retrait du sol, suivies de fortes pluies qui entraînent son gonflement. Ces successions de mouvements, du fait des variations de la teneur en eau du sol, génèrent des dégâts sur les bâtiments (pouvant les rendre non habitables ou inexploitables). L'apparition de ce phénomène sous les fondations d'un bâtiment peut causer des fissures, voire un effondrement de toute la structure.

La figure ci-dessous illustre le phénomène de retrait gonflement des sols argileux.

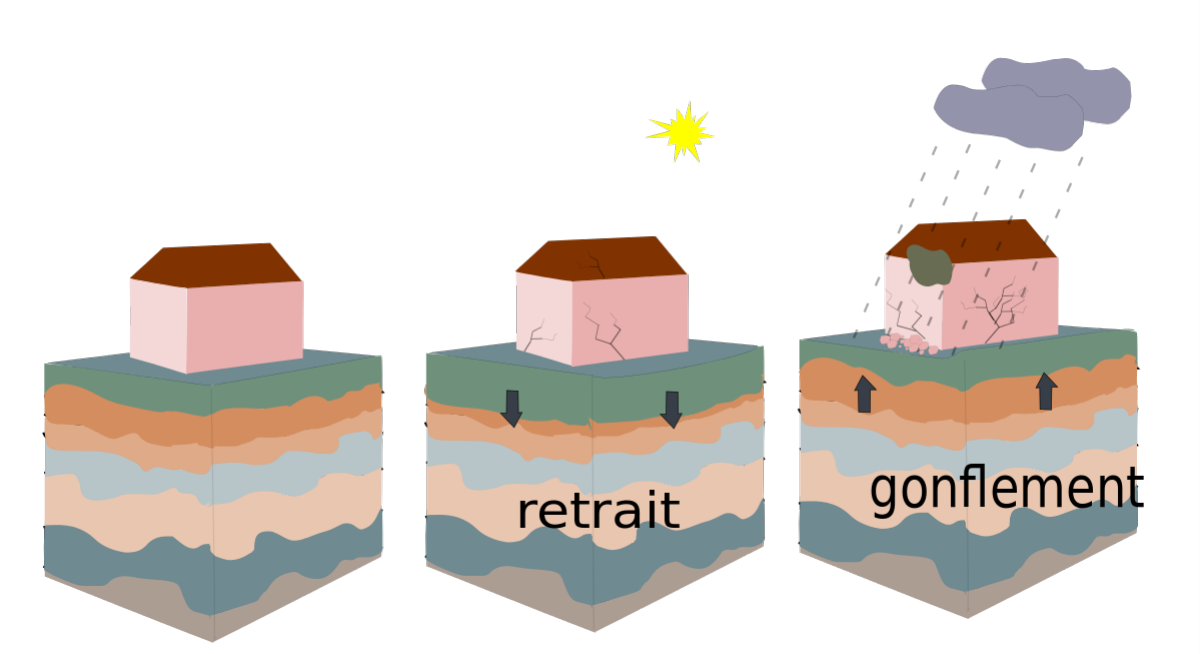

FIGURE 1.8 – Retrait Gonflement des sols Argileux (RGA)

1.4.2 Arrêté CatNat Sécheresse

Un arrêté catastrophe naturelle Sécheresse est un arrêté qui déclare un avis favorable ou non favorable d'un état de Sécheresse dans une commune donnée, pour une période donnée, après l'étude par la commission interministérielle de la demande.
Pour que l'état de Sécheresse soit déclaré dans une commune, il faut que deux critères soient respectés :

— un critère de prédisposition géologique,

— un critère déclenchant météorologique.

Premier Critère : Critère de prédisposition géologique [6]

A travers ce premier critère géologique, le but est d'identifier les territoires de la France métropolitaine qui sont plus susceptibles d'être concernés par le phénomène de RGA. Ces travaux de cartographie pour identifier les zones à risques, ont été menés par le Bureau de Recherche Géologique et Minières (BRGM).
Grâce à cette étude, quatre types de zones peuvent être identifiés sur la carte de l'Hexagone :

— une zone non argileuse (soit une exposition du sol face au phénomène RGA nulle),

— une zone d'exposition faible,

— une zone d'exposition moyenne,

— une zone d'exposition forte.

La figure ci-dessous est la cartographie de ces quatre zones sur le territoire métropolitain.

6. BRGM. Cartographie de l'aléa retrait-gonflement des sols argileux dans le département de la marne. 2008. (3)

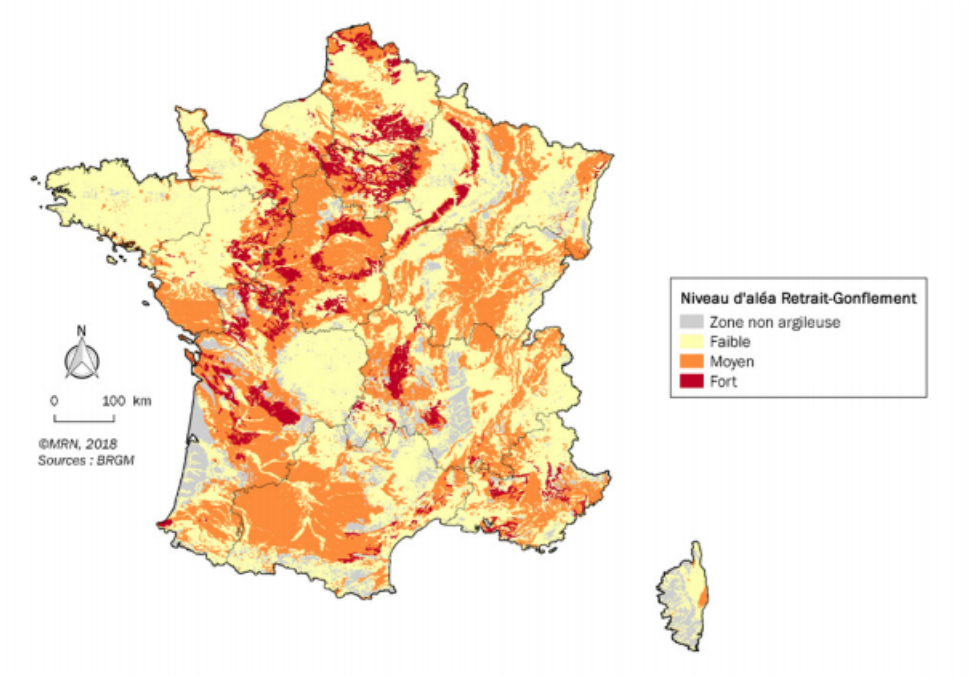

FIGURE 1.9 – Carte d'exposition au phénomène RGA (Source : BRGM)

Afin d'obtenir la publication d'un arrêté CatNat Sécheresse, le quotient de la surface communale par la surface exposée au phénomène RGA doit dépasser les 3%. L'association Mission Risques Naturels [7] propose cette figure pour illustrer le mécanisme de calcul de ce critère.

FIGURE 1.10 – Mécanisme de calcul de la surface RGA (Source : Mission Risques Naturels)

Deuxième Critère : Critère de déclenchement météorologique : SWI [8]

D'après Météo France, le « Soil Wetness Index » ou SWI uniforme est un indice d'humidité des sols. Il représente sur une profondeur de deux mètres, l'état de réserve en eau du sol par rapport à la réserve utile. C'est un indice qui est compris entre zéro et un. Si ce dernier est égal à zéro, cela veut dire que le sol est très sec, si le SWI est égal à 1, le sol est alors saturé en eau et a atteint sa réserve utile.

7. Mission Risques Naturels (MRN). La sécheresse géotechnique. 2018. (9)
8. Météo France, Descriptif du SWI uniforme CatNat. (24)

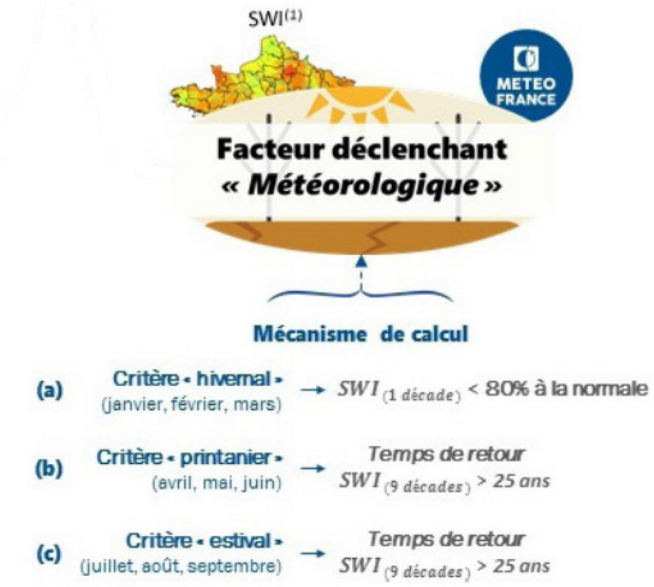

FIGURE 1.11 – Critère SWI uniforme (Source : Mission Risques Naturels (MRN))

Chaque mois, un indicateur d'humidité des sols moyens est calculé. Cet indicateur est la moyenne journalière des SWI uniformes. Ensuite, pour chaque saison, trois indicateurs mensuels moyens sont définis. L'indicateur s'appuie sur les moyennes calculées sur les trois derniers mois. Par exemple, l'indicateur du mois d'août s'appuie sur la moyenne des SWI des mois d'août, juillet et juin.
L'intensité d'un épisode de Sécheresse est anormale dès lors que l'indicateur présente une durée de retour supérieure ou égale à 25 ans. Autrement dit, il y a 4% de chance qu'un tel indicateur soit calculé pour une année donnée.
La durée de retour est calculée à partir des indicateurs des années passées. Chaque saison, la commission interministérielle retient l'indicateur présentant la durée de retour la plus élevée. Si l'indicateur d'un mois dépasse une durée de retour de 25 ans, alors le critère est validé.
La figure ci-dessous représente l'évolution de l'indicateur entre 2003 et 2020.

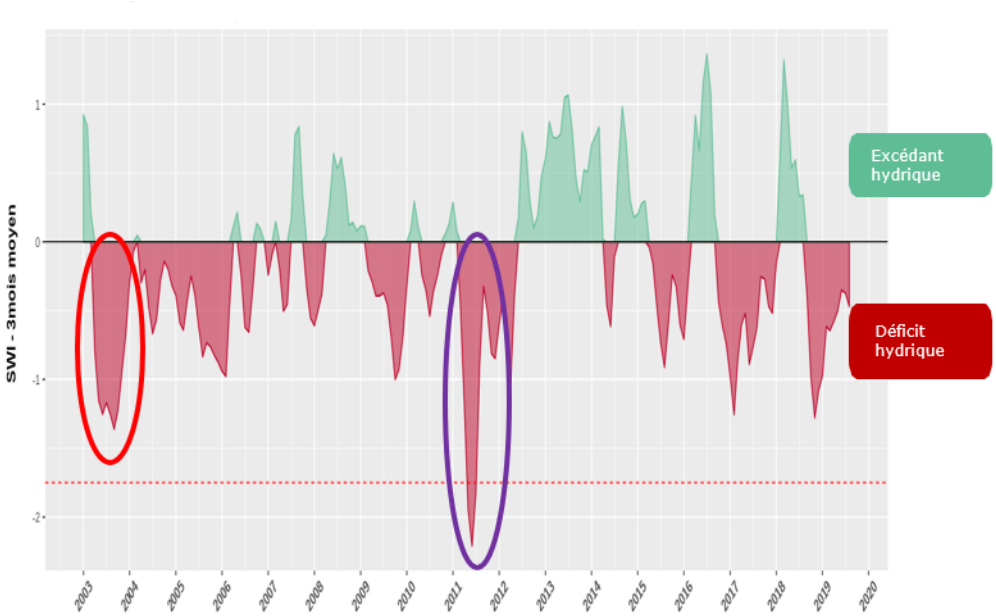

FIGURE 1.12 – Évolution du SWI uniforme (Source : Météo France)

Le déficit hydrique en 2003 a induit un grand nombre de communes en état de Sécheresse. En effet, la CCR comptabilise 4500 arrêtés CatNat Sécheresse pour cette année. Pour rappel, le coût a été de 3 milliards d'euros (Figure 1.5). De même, l'année 2011 a connu un déficit hydrique important qui a engendré l'apparition de près de 2500 arrêtés CatNat Sécheresse dans le JO.

1.4.3 Évolution des coûts de la Sécheresse en France

Les coûts de la Sécheresse sont de plus en plus onéreux. D'année en année, le nombre d'arrêtés publiés augmente et génère des charges. La figure ci-dessous représente l'évolution de la charge sinistres (le coût total) de la Sécheresse en France depuis la création du régime CatNat.

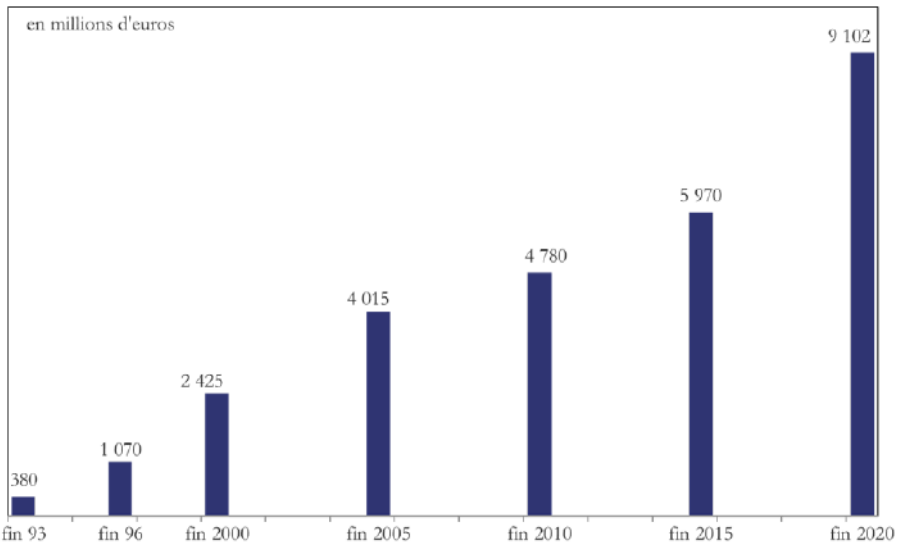

FIGURE 1.13 – Coût sinistres de la Sécheresse en France (Source : France Assureurs)

Selon France Assureurs, les coûts de la Sécheresse pour les années 2019 et 2020 s'élèveraient respectivement à 1,3 milliards et 779 millions d'euros. Tandis que la charge sinistres de la Sécheresse s'élèverait à 9.1 milliards d'euros en fin 2020 depuis l'instauration du régime des CatNat. (6)

Conclusion

L'évolution des fréquences des catastrophes naturelles dans le monde lors des dernières décennies laisse penser que la survenance de ces évènements devrait malheureusement continuer à croître.

Le risque Sécheresse, caractérisé par le phénomène de retrait gonflement des sols argileux, représente un des aléas climatiques majeurs. Ses fréquences de plus en plus rapprochées ainsi que l'intensité des évènements font croître son coût.

A la lumière des risques présentés, les compagnies d'assurances, dont Groupama d'OC, doivent se protéger contre la dérive de ce phénomène.
La prochaine étape consiste à présenter la problématique de cette étude ainsi que les objectifs et les enjeux de ce mémoire. Ensuite, l'impact de la Sécheresse sur le portefeuille de l'entreprise sera étudié.

Chapitre 2

Problématique, objectifs et enjeux du mémoire

2.1 Problématique

Lors d'un arrêté des comptes, la charge sinistres du risque Sécheresse sur bâtiments est difficile à évaluer pour les assureurs. Deux principales raisons à cette difficulté :

1 - La non connaissance du SWI [1] uniforme

Les compagnies d'assurances n'ont pas accès au SWI uniforme. Météo France ne publie ce critère qu'après la publication des arrêtés CatNat Sécheresse dans le Journal Officiel (JO). Généralement ces parutions se font après l'exercice d'inventaire. De fait, il est primordial pour les compagnies d'assurances de développer des méthodes prédictives de publications des arrêtés et d'anticiper une charge sinistres correspondante.

2 - La parution tardive des arrêtés CatNat Sécheresse dans le JO

Le délai d'attente entre la fin d'un évènement Sécheresse et le jour de publication de l'arrêté dans le JO est parfois très long. Par exemple en 2008, deux communes ont été reconnues en état de Sécheresse pour l'année 1989, c'est à dire 19 ans après. Cet exemple reste néanmoins très rare. Selon la MRN [2], la durée moyenne entre la fin de survenance d'une Sécheresse et la date de publication d'un arrêté est de 18 mois.
La figure ci-dessous représente la répartition du nombre d'arrêtés parus en 2020 selon l'aléa et le délai de parution.

1. Soil Wetness Index : critère météorologique pour la reconnaissance des arrêtés
2. L'association Mission Risques Naturels (MRN) est une mission des sociétés d'assurances, pour la connaissance et la prévention des risques naturels. (9)

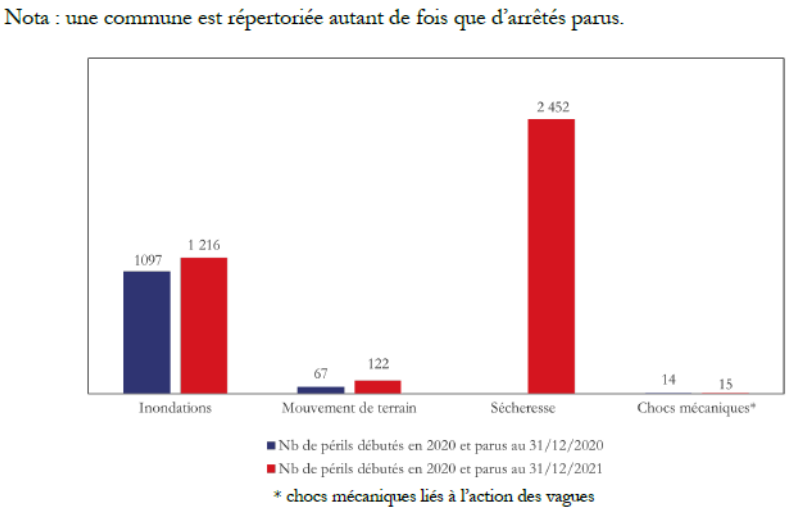

FIGURE 2.1 – Répartition du nombre d'arrêtés parus en 2020 selon l'aléa et le délai de parution (Source : France Assureurs)

Constats :

— 2452 arrêtés CatNat ont été publiés pour des évènements Sécheresse en 2020,

— Néanmoins, aucun de ces arrêtés n'a été publié l'année même.

Ce délai provoque une grosse incertitude dans l'évaluation du provisionnement de ce risque pour les assureurs.

2.2 Objectifs

Pour répondre à cette problématique d'évaluation, l'approche décrite dans ce mémoire s'articulera essentiellement en 3 étapes :

— dans un premier temps, les fondements et principes du régime d'indemnisation des catastrophes naturelles ainsi que le mécanisme d'indemnisation des sinistres seront dressés. A la suite, l'impact de la Sécheresse dans les résultats techniques de Groupama d'OC sera exposé.

— la deuxième étape aura pour objectif de modéliser la survenance des arrêtés sur le territoire de GOC en créant une base de données constituée de l'historique des arrêtés, de la constitution géologiques du sol et des données sinistres en assurance Multi-Risques Climatiques sur Récoltes. Puis, différents modèles prédictifs seront testés. Le modèle le plus adapté sera retenu.

— enfin, la dernière étape consistera à donner une évaluation de la charge sinistres de la Sécheresse sur bâtiments sur l'année d'inventaire en appliquant aux évaluations précédemment calculées un coût moyen. Là aussi, différentes méthodes seront employées.

2.3 Enjeux

Ce mémoire présente deux enjeux importants pour Groupama d'OC :

— Premièrement, étudier la corrélation entre la Sécheresse qui touche les bâtiments et la Sécheresse affectant les cultures agricoles. Une corrélation qui pourrait permettre d'anticiper les publications au JO des arrêtés CatNat. Ainsi, lors des arrêtés des comptes, une évaluation du coût de la Sécheresse sur bâti pourrait être réalisée.

— Deuxièmement, présenter les perspectives possibles pour améliorer la gestion du risque Sécheresse.

Chapitre 3

Impact de la Sécheresse sur le portefeuille de GOC

Introduction

Groupama d'OC (GOC) est un assureur généraliste qui fait partie du groupe Groupama. Comme toutes les compagnies d'assurance, l'entreprise n'est pas épargnée par le risque Sécheresse. Afin d'apporter tous les éléments nécessaires au lecteur, il est important de bien contextualiser l'étude. Ce chapitre commence par présenter le groupe Groupama, l'entreprise GOC, son secteur d'activité et ses chiffres clés pour l'année 2021. Dans un second temps, il sera présenté l'impact de la Sécheresse sur son portefeuille.

3.1 Présentation de l'entreprise

3.1.1 Présentation du groupe Groupama

Cette section est basée sur les informations du groupe citées sur le site de l'entreprise [1].
Groupama est une société d'assurance mutuelle française, créé en 1900 à la suite de la loi du 4 juillet, qui autorisait la création de caisses d'assurances mutuelles agricoles en France.
Groupama est régie par le code Français des assurances et possède 2 grandes marques commerciales :

— Groupama, marque d'assurance généraliste distribuée par le réseau des Caisses Régionales (dont Groupama d'OC). Elle est également implémentée dans 10 pays dans le monde.

— Gan, assureur des entrepreneurs porté par Gan Assurances, un réseau généraliste, ainsi que Gan Patrimoine et Gan Prévoyance, deux réseaux spécialisés.

Le groupe Groupama

Groupama est un groupe plus que centenaire qui s'est progressivement internationalisé au cours de la décennie 2000-2010. Son chiffre d'affaires est de 15.5 milliards d'euros dont 12.9 milliards d'euros en France avec un résultat net de 493 millions d'euros.
Le groupe Groupama emploie aujourd'hui 31 000 collaborateurs, dont 6 000 à l'international. Groupama couvre 12 millions de sociétaires et de clients dans le monde.
Selon le site de l'entreprise, le groupe Groupama est aujourd'hui en France :

— 6ème assureur généraliste,

— 1er assureur agricole,

— 1er assureur des communes,

— 2ème assureur en santé individuelle (en chiffre d'affaires),

1. https://www.groupama.com/fr/notre-modele/le-groupe-en-chiffres/

— 3ème assureur habitation,

— 4ème assureur prévoyance individuelle,

— 4ème assureur auto,

De la même façon, le groupe est bien présent dans le monde en étant :

— 4ème assureur non-vie en Hongrie,

— 5ème assureur en Roumanie,

— 10ème assureur non-vie en Italie.

Organisation du groupe Groupama

Le groupe Groupama est composé de 4 entités juridiques :

— **Les caisses locales :** Le groupe comprend 2 750 caisses locales réparties sur tout le territoire français. Dans le système mutualiste, les sociétaires élisent leurs représentants au niveau local qui élisent à leur tour leurs représentants au niveau régional et national. Chaque décision importante fait l'objet d'une réflexion collective en amont dans le cadre des assemblées générales, conseils d'administration, commissions et réunions de travail. A la différence d'autres entreprises, dans le monde mutualiste, ce sont les sociétaires qui sont les "propriétaires" de l'entreprise.

— **Les caisses régionales :** 9 caisses métropolitaines, 2 caisses d'outre-mer et 2 caisses spécialisées qui proposent une offre complète d'assurance et de produits financiers.

— **Groupama Assurances Mutuelles :** joue le rôle de réassureur pour les caisses adhérentes, les caisses régionales de Groupama et GAN.

— **Les filiales du groupe :** La plupart des filiales de services et d'assurances françaises et internationales est portée par une holding, Groupama Holding Filiales et Participations, détenue entièrement par Groupama Assurances Mutuelles.

L'organigramme de ces 4 entités est représenté par la figure ci-dessous.

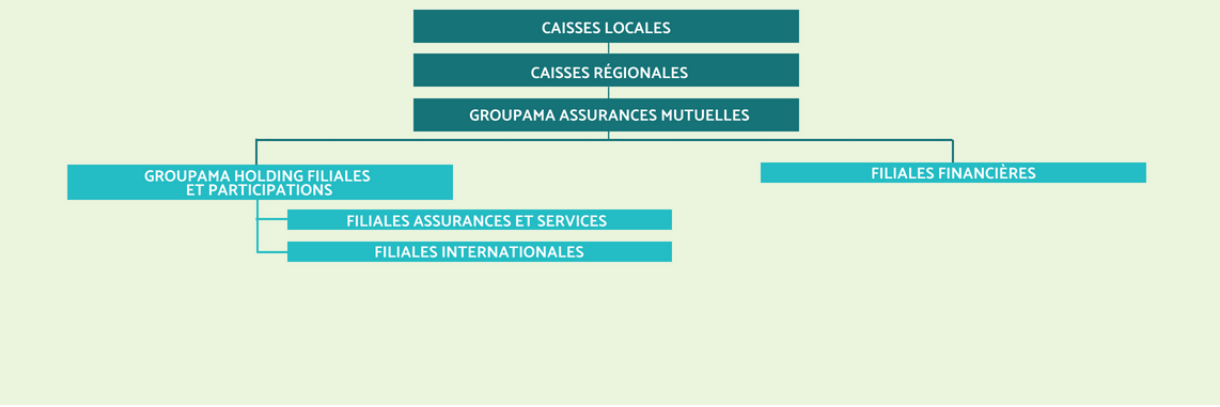

FIGURE 3.1 – Organisation du groupe Groupama (Source : Groupama)

Les caisses régionales

Les caisses régionales sont des entreprises d'assurances mutuelles agréées qui proposent une offre complète d'assurance et de produits financiers. Ces caisses disposent de leurs propres réseaux commerciaux de salariés et assurent la gestion des relations avec leurs clients.
Les caisses régionales ont deux objectifs principaux :

— distribuer les produits créés par l'entreprise.

— réassurer les caisses locales d'assurances mutuelles adhérentes au groupe et faciliter leur fonctionnement.

Le groupe Groupama s'appuie sur un réseau de 9 caisses régionales métropolitaines, 2 caisses d'outre-mer (Antilles-Guyane et Océan Indien) et 2 caisses spécialisées (Groupama Forêts Assurances (Misso) et Producteurs de tabac).

3.1.2 Présentation de Groupama d'OC (GOC)

Groupama d'OC est un assureur mutualiste généraliste qui prend en charge les besoins de chacun : particuliers, professionnels, exploitants agricoles, entreprises, associations et collectivités locales.
Groupama d'OC est aussi une caisse régionale du groupe Groupama, qui a été créée en 2003 suite à la fusion de 3 caisses : Pays Verts (Cantal, Creuse, Corrèze) ; Petit Oc (Lot, Aveyron, Lozère, Tarn et Tarn-et-Garonne) et Groupama Sud-Ouest (Haute-Garonne, Gers, Ariège, Landes, Hautes-Pyrénées et Pyrénées-Atlantiques). Cette fusion a permis de renforcer la caisse régionale pour pouvoir faire face à des risques plus lourds, mais aussi de se doter de nouvelles compétences humaines et de nouveaux moyens technologiques. Une taille plus importante qui a aussi permis de prendre une nouvelle place au niveau national.
Située dans un territoire qui compte 14 départements et représente plus de 4 millions d'habitants, à ce jour, Groupama d'OC protège 501 100 clients sociétaires. Pour la commercialisation de ces produits, GOC peut compter sur un réseau de 304 agences implantées dans les communes du territoire. Ce dernier est illustré dans la figure ci-dessous.

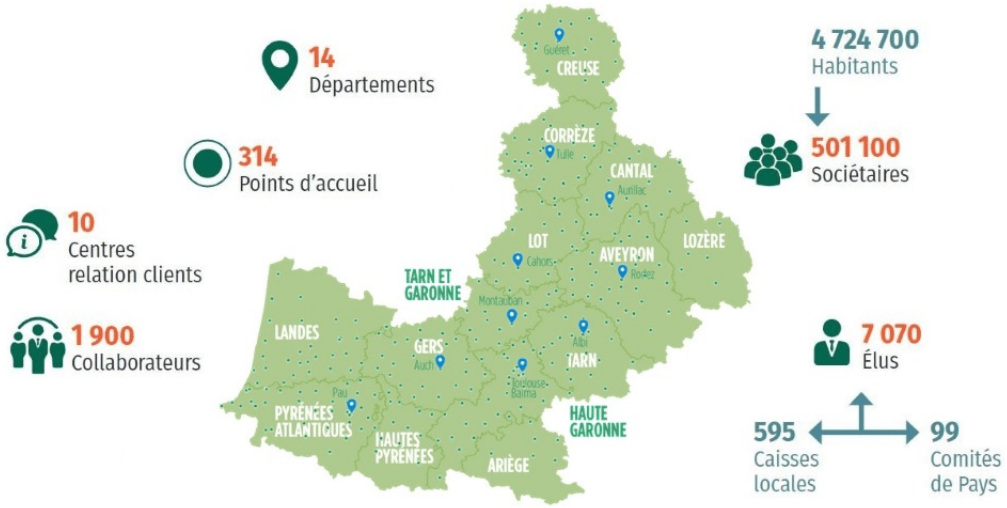

FIGURE 3.2 – Carte du territoire de GOC

Groupama d'OC en chiffres

En 2021, le chiffre d'affaires de GOC a atteint 785.4 millions d'euros en assurance IARD sur un volume de 1 512 700 contrats.
La répartition des cotisations perçues par marché se présente de la façon suivante :

— 32% Retraités,

— 27% Agricole,

— 18% Particuliers,

— 14% Entreprises,

— 9% Professionnels.

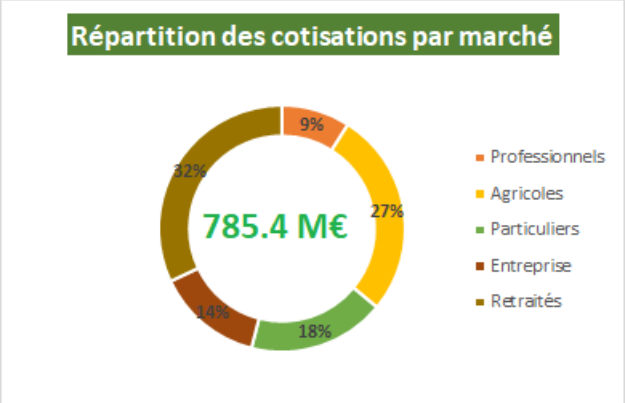

FIGURE 3.3 – Répartition des cotisations par marché

La répartition des cotisations par marché montre que le portefeuille de l'entreprise est le reflet de la région. En effet, l'Occitanie est la 2ème région agricole de France, derrière la Nouvelle Aquitaine alors que 85% du territoire est situé en zone de piémont et montagne, dites "zones de handicap naturel". La répartition des cotisations par risque est comme suit :

— 28% Véhicules à moteurs (Auto, TMA, 2/3 roues),

— 34% Assurance de Personnes (Santé, Prévoyance),

— 4% Climatiques sur récoltes,

— 34% Dommages aux Biens et Responsabilité Civile (Habitation, TNS, Dommages Agri, ...).

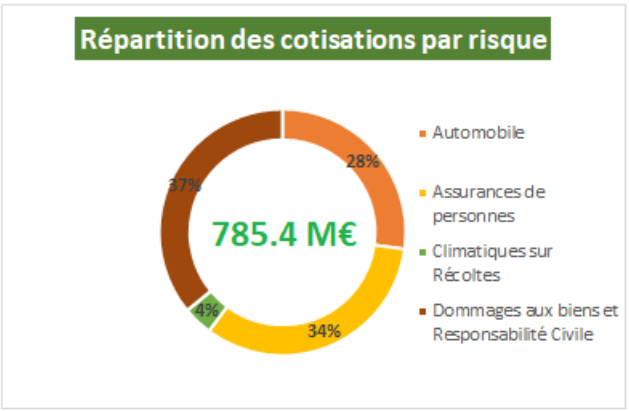

FIGURE 3.4 – Répartition des cotisations par risque

A titre d'information, selon le SFCR [2] de l'entreprise :

— les ratios de couverture SCR [3] et MCR [4] réglementaires sont respectivement de 299% et 1197% au 31 décembre 2021,

— le total bas de bilan est de 2.8 milliards d'euros.

2. Rapport sur la solvabilité et la situation financière de l'entreprise (8)
3. Capital de solvabilité requis
4. Capital minimum requis

3.2 Impact de la Sécheresse sur le portefeuille de GOC

Le territoire de GOC est fortement exposé aux catastrophes naturelles. Qu'il s'agisse de :

— sécheresse,

— inondation,

— mouvement de terrain.

GOC doit donc se prémunir contre les sinistres engendrés par ces risques.
Pour se protéger, l'entreprise perçoit des cotisations (appelées cotisations CatNat) de la part de ses sociétaires. Ces montants récoltés sont supposés couvrir GOC contre tous les sinistres CatNat.
Plusieurs questions se posent :

— Ces cotisations sont-elles suffisantes pour protéger l'entreprise contre l'ensemble des sinistres des aléas climatiques ?

— Quelle est la part de l'aléa Sécheresse dans la charge sinistres CatNat subie par GOC ?

— Quels sont les métiers les plus impactés par la Sécheresse ?

Pour répondre à ces questions et afin de matérialiser les enjeux du risque Sécheresse, une étude d'impact de cet aléa sur le portefeuille est réalisée.

3.2.1 Répartition de la charge sinistres CatNat selon l'aléa

Tout d'abord, l'histogramme ci-dessous représente la répartition de la charge de ces sinistres selon l'aléa qui les a engendrés.

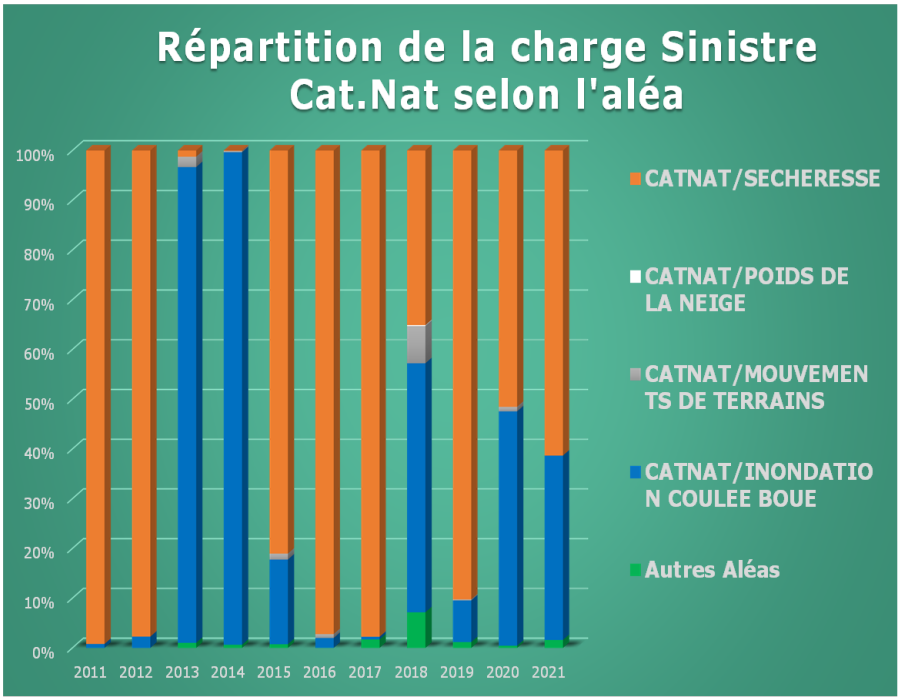

FIGURE 3.5 – Répartition de la charge sinistres CatNat selon l'aléa

Interprétation Graphique :

— La Sécheresse (en orange) suivie de l'inondation (en bleue) sont les aléas qui engendrent le plus de charge sinistres.

— Pour les années 2011, 2012, 2016, 2017 et 2019, la charge des sinistres Sécheresse représente plus de 90% de la charge sinistres globale des évènements CatNat.

— En 2013 et 2014, la part des sinistres Sécheresse est quasi-nulle. Au cours de ces deux années, le portefeuille n'a pas été impacté par la Sécheresse mais par l'inondation qui englobe presque toute la charge sinistres CatNat.

— Pour les années 2018, 2020 et 2021, la charge sinistres CatNat est partagée entre l'aléa Sécheresse et l'inondation.

Constat : Sur la période présentée, la Sécheresse représente l'aléa CatNat qui engendre le plus de charge sinistres à l'entreprise.

3.2.2 Répartition de la charge sinistres Sécheresse selon les métiers

Au sein de GOC, le terme métier est employé pour les différentes typologies de contrat. Comme précisé dans la Figure 3.3, quatre grandes typologies de risques ont été définies. Seuls deux risques détiennent une garantie CatNat. Les contrats d'assurance de la famille DAB/RC et véhicules à moteurs. Pour ces risques, les contrats d'assurance détiennent la garantie CatNat et couvre l'assuré lors des évènements climatiques publiés au JO.
L'histogramme ci-dessous présente la répartition de la charge sinistres Sécheresse selon les différents types de contrat.

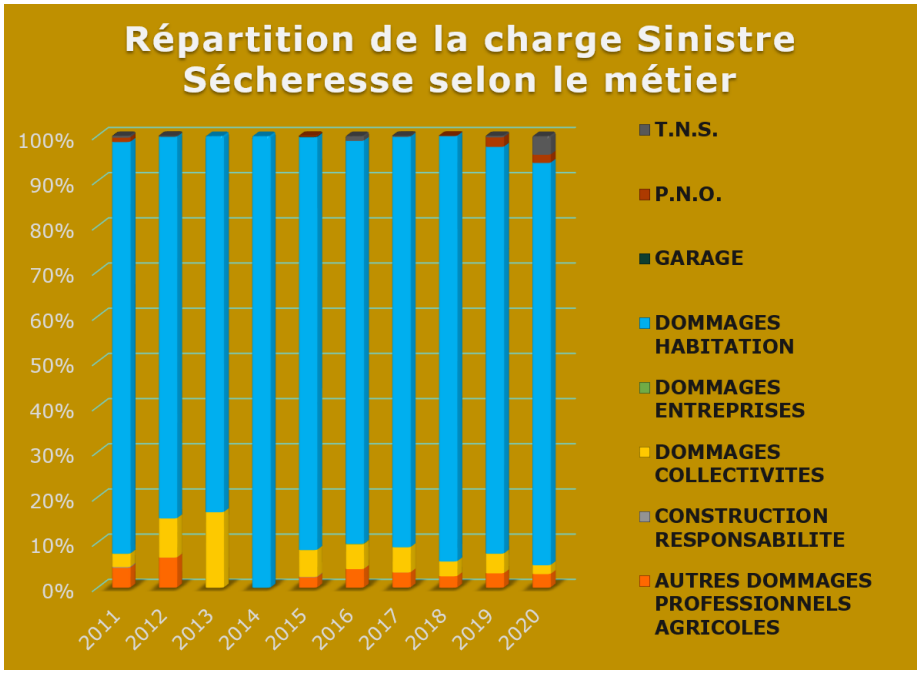

FIGURE 3.6 – Répartition de la charge sinistres Sécheresse selon le métier

Interprétation Graphique :

— La Sécheresse impacte plusieurs métiers du portefeuille. Les métiers "Dommages Habitation", "Dommages Collectivités", "Dommages Entreprises", "Autres Dommages Professionnels Agricoles" sont les branches principalement touchées par cet aléa,

— Le métier "Dommages Habitation" est la branche d'assurance la plus impactée par la Sécheresse. Au minimum, 85% de la charge sinistres Sécheresse par an représente une charge pour ce métier,

— Les collectivités sont aussi impactées par la Sécheresse. En 2013, les sinistres Sécheresse du métier "Dommages Collectivités" ont contribué à plus de 10% de l'ensemble de la charge sinistres,

— Les véhicules terrestres à moteur sont globalement épargnés par le risque Sécheresse,

Compte tenu des résultats précédents, un "focus" spécifique est réalisé sur le métier Habitation.

3.2.3 Présentation des résultats techniques en Habitation

Dans le cadre d'un contrat d'assurance, le sociétaire paye une cotisation (pour une mutuelle) ou une prime (pour les compagnies d'assurances) : celle-ci est notée C. Lors de la survenance d'un sinistre, l'assureur paye le montant des dégâts : ce montant est noté S.
La rentabilité d'un contrat d'assurance est alors mesurée par le ratio S/C.
Si le ratio S/C est supérieur à 100%, l'entreprise a une marge technique négative. Sinon, l'entreprise a une marge technique positive. Le produit d'assurance n'est cependant pas techniquement rentable. En effet, il faut couvrir les frais généraux également. Ces derniers sont différents selon les métiers.
Pour des raisons de confidentialité, le S/C cible en Habitation ne sera pas précisé.
Sur la Figure 3.7, la courbe en pointillé représente l'évolution du S/C brut du métier (en prenant en compte toute la charge sinistres) et la courbe en continu représente le S/C des sinistres CatNat Sécheresse.

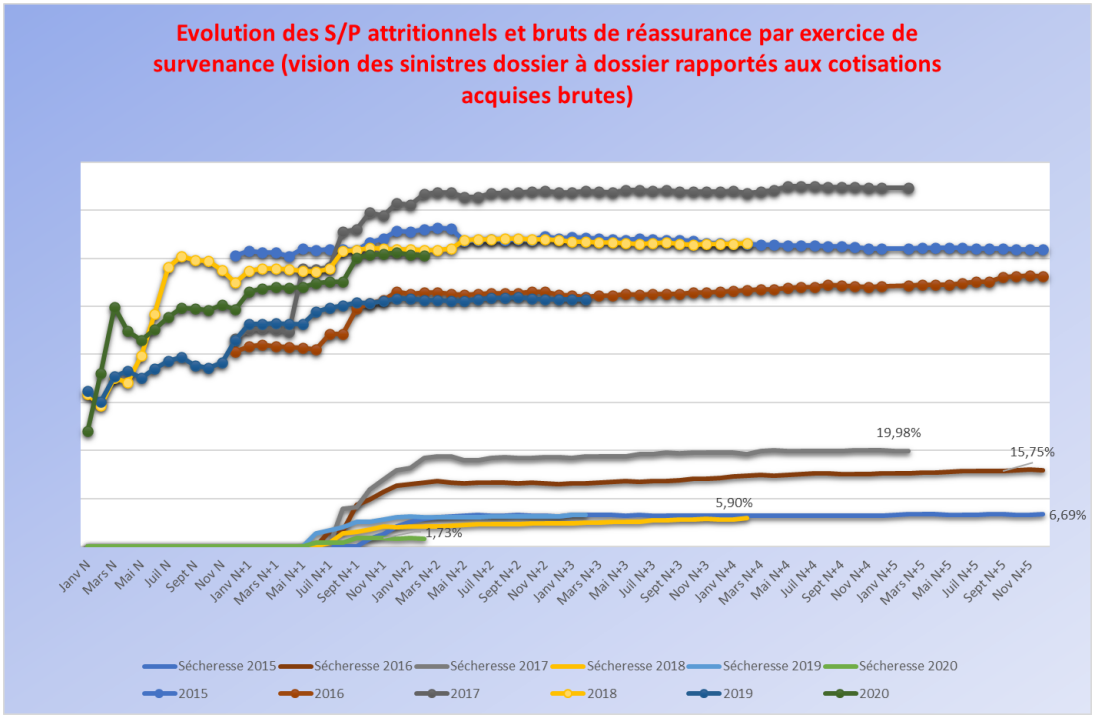

FIGURE 3.7 – Évolution des S/C en Dommages Habitation

Interprétation graphique :

- Le S/C brut (en pointillé) est volatil lors de la première année. De janvier à décembre N, les courbes évoluent. Ce qui est assez logique car dépendant de la déclaration de sinistre (Incendie, Tempête, Dégâts des eaux,...) des sociétaires tout au long de l'année. Puis entre janvier N+1 et juin N+1 elles sont relativement stables,
- L'ensemble des courbes en pointillé évolue à partir d'un an et demi après. Ces évolutions tardives sont liées à la déclaration d'arrêtés CatNat Sécheresse,
- En 2017, les S/C des sinistres Sécheresse a atteint 20%. C'est-à-dire que la charge des sinistres Sécheresse pour cette année représente à elle seule, un cinquième des cotisations perçues,

Constat :
La Sécheresse impacte fortement le métier "Dommages Habitation".
Ces niveaux de S/C ont été calculés en prenant en compte l'ensemble des cotisations du métier. Quels seraient ces niveaux si la prime dédiée à la garantie CatNat était uniquement considérée ?
Zoom sur la Sécheresse :
Dans la figure ci-dessous, seule la cotisation dédiée à la garantie CatNat est prise en compte.

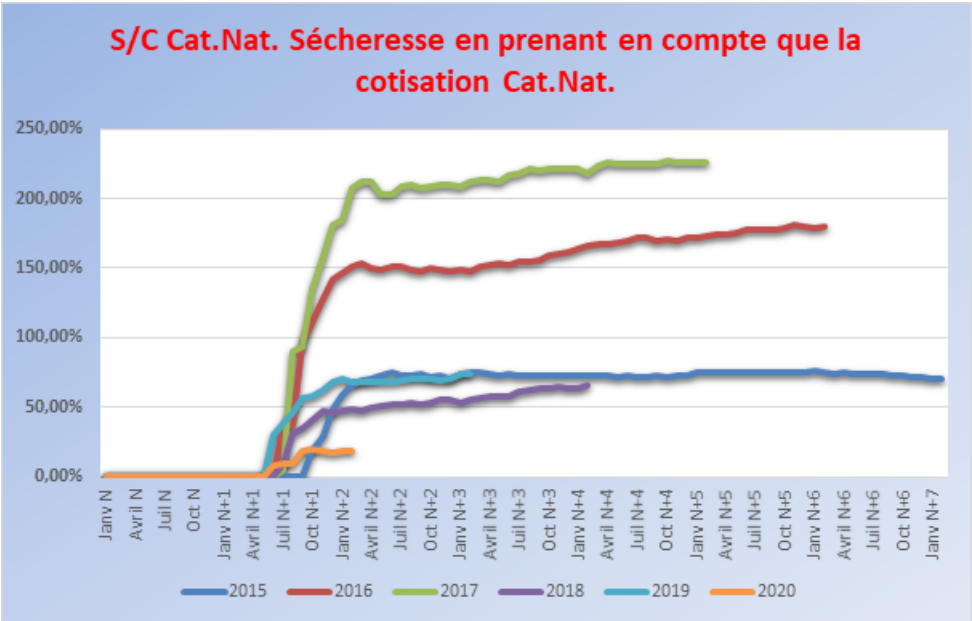

FIGURE 3.8 – Évolution des S/C Sécheresse en Dommages Habitation

Interprétation graphique :

- Les exercices 2017 et 2016 sont les plus touchés par la Sécheresse avec des S/C respectivement de 225% et de 180%,
- La courbe de l'exercice 2019 suit, à peu près, la même allure que la courbe de l'exercice 2015,
- L'année 2020 semble être pour l'instant une année relativement faible en Sécheresse (cependant, de nouveaux arrêtés peuvent encore être publiés).

Constat :
Sur deux exercices, les niveaux de S/C CatNat Sécheresse sont largement supérieurs à 150%. De plus, ces cotisations sont supposées protéger GOC contre tous les autres aléas CatNat. Cela veut dire que la prime dédiée à la garantie CatNat n'est pas suffisante pour faire face à l'ensemble des périls CatNat

Conclusion

La Sécheresse représente un risque majeur pour GOC. Cet aléa touche plusieurs métiers du portefeuille.

La charge sinistres liée à cet unique péril est nettement supérieure aux cotisations perçues et montre là une des faiblesses du régime d'indemnisation des CatNat.

Deuxième partie

Estimation de la charge sinistres de la Sécheresse pour le portefeuille de Groupama d'OC en arrêté des comptes

Chapitre 4

Bases de données

Introduction

Avant de commencer un quelconque travail de modélisation, il faut constituer une base de travail. Cette base de données est formée par trois types de données :

— Les arrêtés CatNat Sécheresse publiés sur le territoire de GOC,

— Les données géologiques et géographiques des communes du territoire,

— Les données sinistres en assurance Multi-Risques Climatiques sur Récoltes.

Tout d'abord, la base historique des arrêtés CatNat sera présentée et analysée.
Les deux autres jeux de données feront également l'objet d'une section afin d'y présenter les différents indicateurs.

4.1 Les arrêtés CatNat Sécheresse

4.1.1 Présentation des données

Les données Sécheresse ont été extraites sur le site de la Caisse Centrale de Réassurance (CCR). La table est constituée de 7 variables comme suit :

1. Le numéro INSEE qui représente l'identifiant d'une commune,
2. Le département de la commune,
3. Le nom de la commune,
4. Le début de l'évènement Sécheresse,
5. La fin de l'évènement Sécheresse,
6. La date de publication de l'arrête dans le Journal Officiel (JO).
7. La décision de l'arrêté : Favorable ou Non Favorable

Filtration des données pour le territoire de GOC

La base extraite est une base qui regroupe toutes les communes de France. Or, dans le cadre de ce mémoire, seuls les arrêtés CatNat Sécheresse publiés sur le territoire de GOC sont étudiés.
Pour ce faire, une filtration sur la base a été réalisée. Les arrêtés CatNat Sécheresse qui ont été pris en compte sont les arrêtés publiés sur le territoire de GOC entre 2011 et 2019.

4.1.2 Analyse descriptive de la base de données

Dans l'objectif de décrire et résumer les données, une analyse descriptive de la base est effectuée.

Part des arrêtés favorables par rapport à tous les arrêtés publiés

Parmi les arrêtés CatNat Sécheresse publiés sur le territoire de GOC, nombreux ne sont pas favorables, c'est à dire, qu'ils ne reconnaissent pas l'état de Sécheresse dans la commune.
La figure ci-dessous propose de classifier les arrêtés CatNat Sécheresse selon le type de la décision.

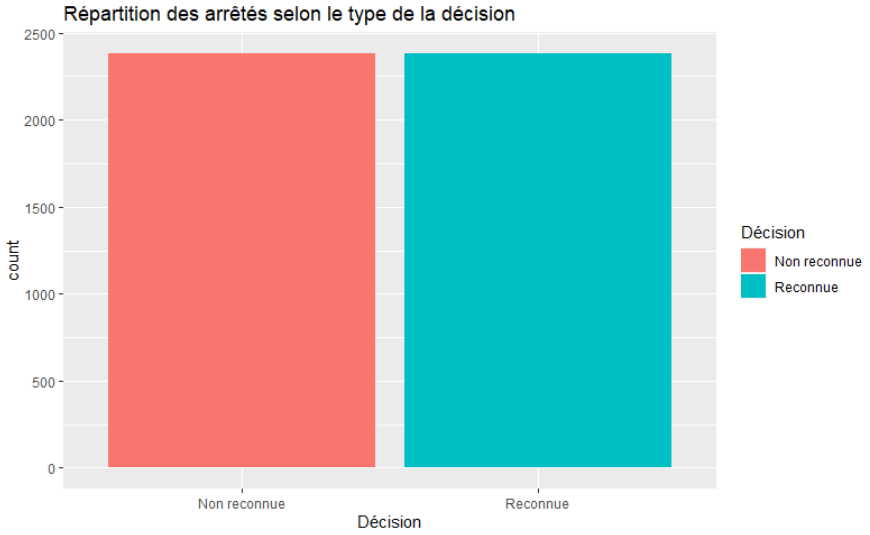

FIGURE 4.1 – Répartition de la publication des arrêtés selon la nature de la décision

Constat :
Un arrêté CatNat Sécheresse publié sur deux est non favorable. Néanmoins, ce pourcentage dépend du nombre de demandes par année.
Afin de mieux illustrer ces propos, la Figure 4.2 propose d'associer la décision de l'arrêté à l'année de survenance.

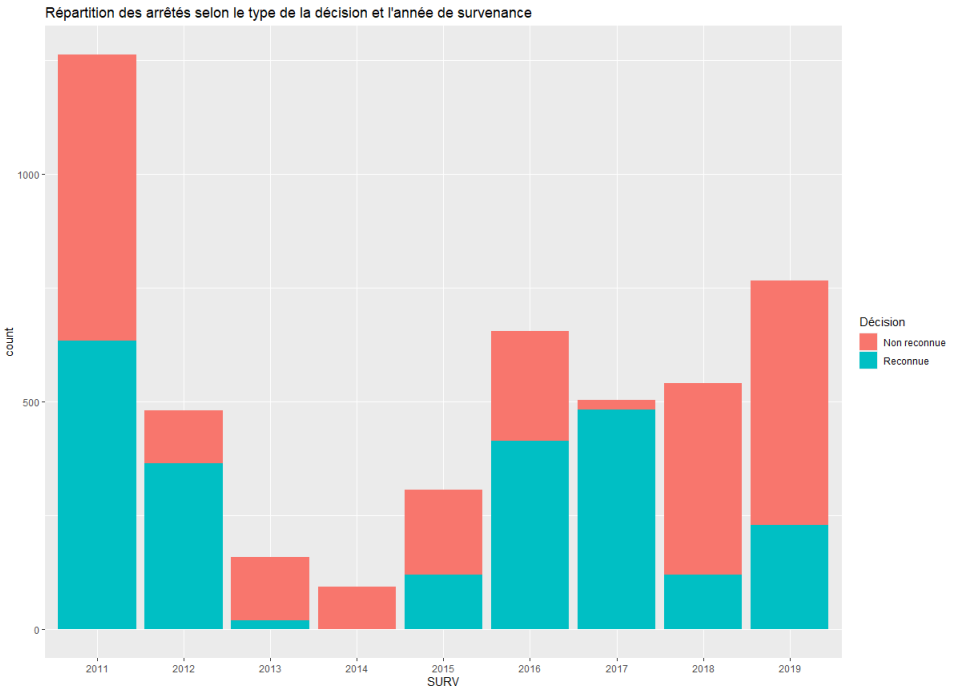

FIGURE 4.2 – Répartition de la publication des arrêtés selon la nature de la décision et l'année de survenance

Constats :

— Le taux d'arrêtés CatNat Sécheresse non favorables varie selon les années.

— En 2011, un nombre élevé de demandes ont été formulées. La moitié de ces demandes a abouti à des arrêtés non favorables. Malgré ce taux élevé de refus, l'année 2011 représente l'année qui compte le plus d'arrêtés favorables.

— En 2013, la majorité des arrêtés publiés est non favorable. En 2014, aucun arrêté favorable n'a été publié.

— Les années 2018 et 2019 ont connu un nombre important d'arrêtés non favorables. Ceci est dû à un nouveau durcissement des règles de reconnaissance après une année 2017 forte en Sécheresse.

Dans l'optique d'estimer la charge sinistres Sécheresse, seuls les arrêtés CatNat **favorables** sont pris en compte.

Écart entre la survenance d'une Sécheresse et la publication de l'arrêté dans le journal officiel

Une des problématiques de l'étude réside dans le retard conséquent de la publication des arrêtés CatNat.
La Figure 4.3 illustre cette problématique d'écart entre la fin d'un évènement Sécheresse et la parution de l'arrêté relative à celle-ci dans le JO.

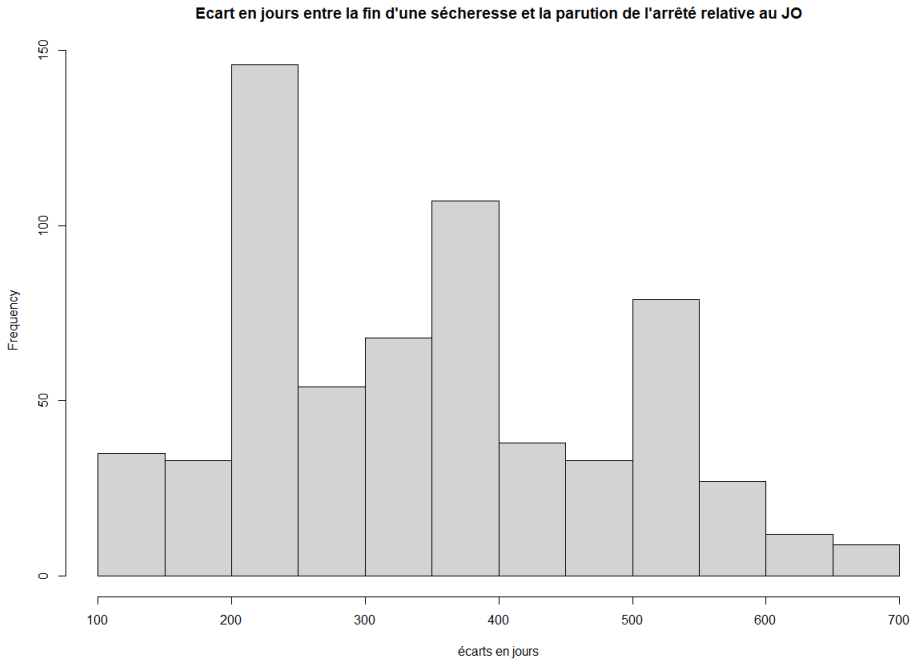

FIGURE 4.3 – Écarts en jours entre la survenance et la publication d'une Sécheresse

Constats :

— Le temps moyen de cet écart est de 351 jours, près d'une année après la fin de l'évènement Sécheresse.

— Cet écart peut atteindre 700 jours sur certains évènements.

Le graphique ci-dessous présente la part des arrêtés publiés très tardivement dans le JO par année de survenance.

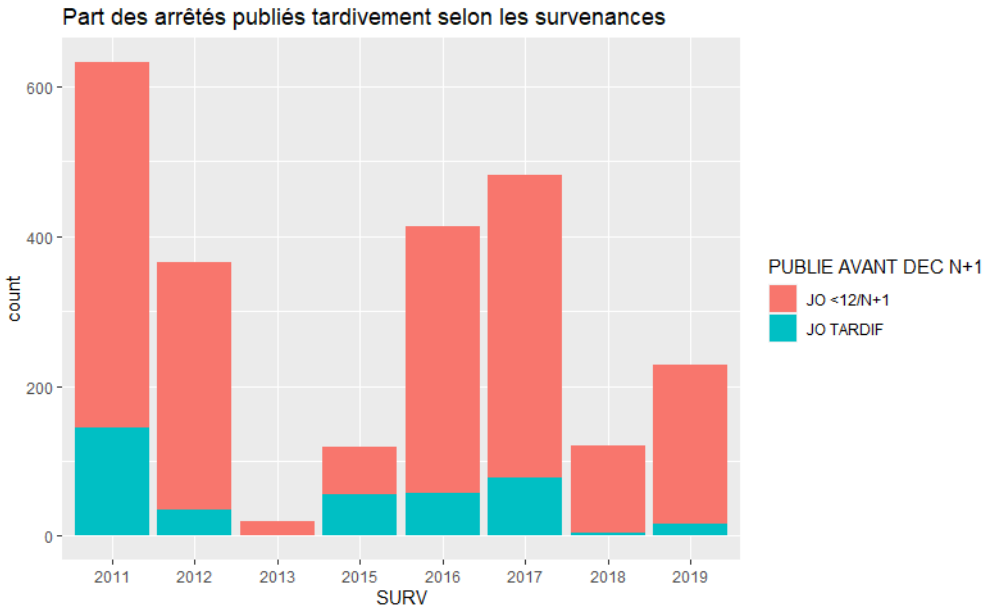

FIGURE 4.4 – Part des arrêtés publiés tardivement selon les survenances

Constats :

— Chaque année, sauf pour l'année 2013, une part importante d'arrêtés est publiée tardivement.

— Néanmoins, ce pourcentage diminue au fil des années. Ceci est dû à une nouvelle réglementation : à partir de 2018, les mairies ont un délai maximal de 18 mois pour demander la reconnaissance d'un état de Sécheresse.

Ces analyses confirment la première problématique. Prédire le nombre d'arrêtés CatNat Sécheresse favorables avant la parution de ces derniers dans le JO est un des enjeux majeurs dans la définition des provisions de l'arrêté des comptes.

4.2 Cartographie des arrêtés CatNat Sécheresse

La réalisation des cartes avec le logiciel R est basée sur le manuel : T. GIRAUD et H. PECOUT. Cartographie avec R. 2021. (21)

Afin d'étudier les zones à risques sur le territoire de GOC, la répartition des arrêtés CatNat Sécheresse favorables est cartographiée et présentée dans le chapitre suivant.

4.2.1 Méthodologie : Cartographie avec R

Pour créer les différentes cartes de l'étude, le logiciel R a été utilisé, avec notamment certaines librairies existantes (les "packages" : "Leaflet" [1], "rgdal" [2] et "sp" [3]).
Ensuite, les étapes suivantes ont été réalisées :

— Télécharger un fichier de données nommé "GEOFLA COMMUNE" [4] constitué par les données géographiques (latitude, longitude) des différentes communes de France et leurs délimitations dans l'espace.

— Importer ce fichier sur le logiciel R.

1. Package R, permet de créer des cartes interactives
2. Package qui permet de créer des cartes à partir du logiciel Qgis
3. Permet de transformer les données R en données spatiales
4. Fichier disponible en libre accès sur le site du gouvernement dédié à l'Open Data

— Créer un fond de carte du territoire voulu.

La figure ci-dessous présente deux exemples d'un fond de carte du territoire de GOC. La carte à gauche est une carte à délimitation départementale. Celle de droite est une carte communale.

FIGURE 4.5 – Fonds de carte du territoire de GOC selon les délimitations départementales et communale

Ensuite, la seconde étape consiste à colorier ces fonds de cartes avec les données des arrêtés CatNat Sécheresse. Pour ce faire, une jointure est réalisée entre les données des communes et la base des arrêtés CatNat Sécheresse favorables.
Les Figures 4.6 et 4.7 présentent l'évolution des arrêtés CatNat Sécheresse entre 2016 et 2019.

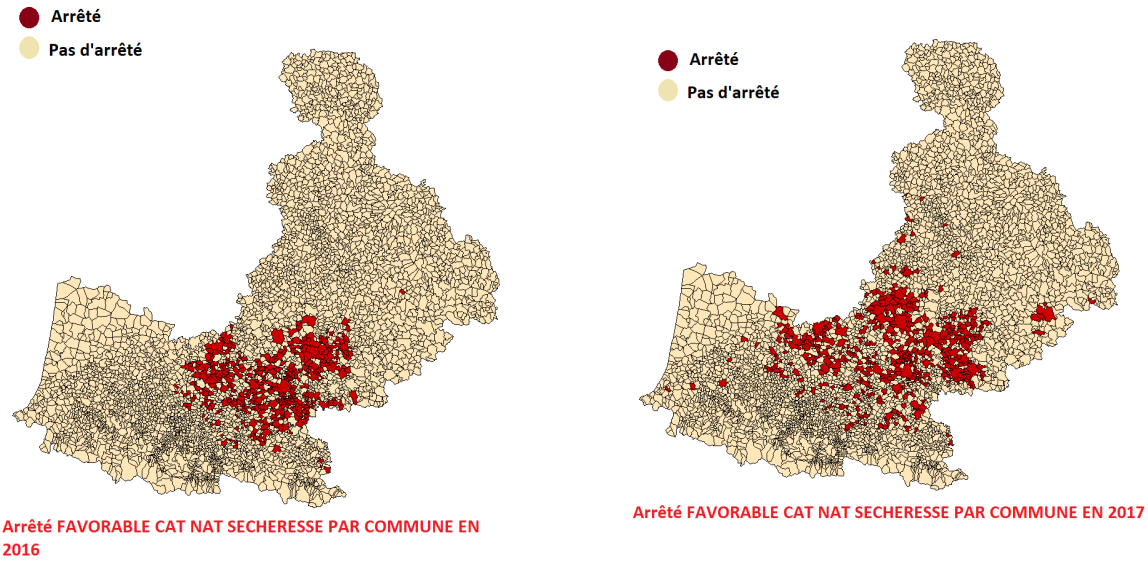

FIGURE 4.6 – Cartographie des arrêtés CatNat Sécheresse favorables sur le territoire GOC pour les années 2016 et 2017

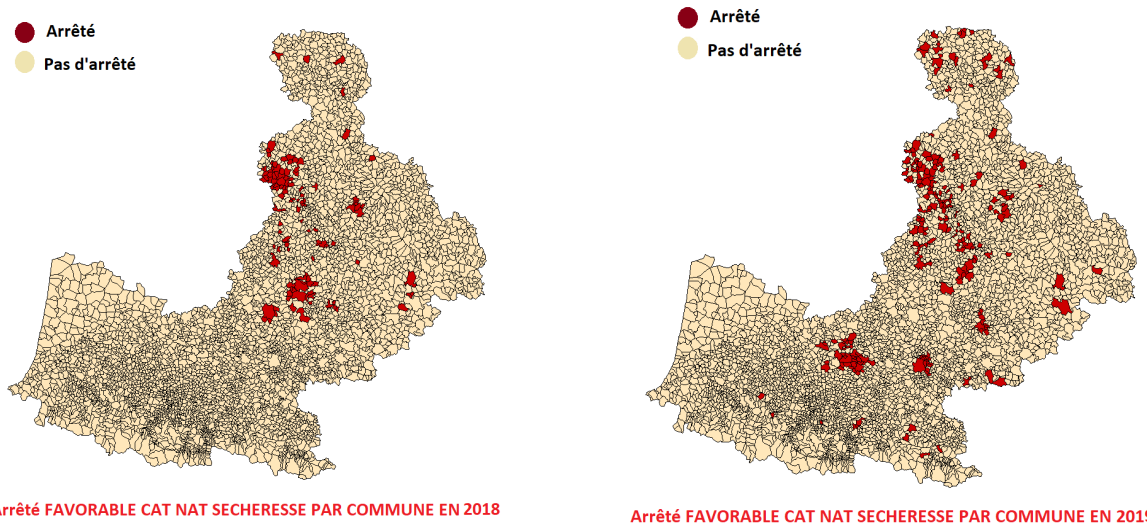

FIGURE 4.7 – Cartographie des arrêtés CatNat Sécheresse favorables sur le territoire GOC pour les années 2018 et 2019

Constats :

— En 2016 et 2017, un grand nombre d'arrêtés a été publié au centre du territoire.

— En 2018 et 2019, le nombre d'arrêtés publiés a diminué. Ces derniers ont plutôt touché la zone nord-est du territoire.

— Les communes touchées par la Sécheresse lors d'une année ne sont pas forcément touchées l'année d'après.

4.3 Données géologiques : Nature des sols des communes de GOC

4.3.1 Notion d'exposition du sol face au phénomène RGA

Pour rappel, pour qu'un arrêté soit favorable, il faut qu'au moins 3% de la superficie de la commune soit exposée au phénomène de retrait gonflement des sols argileux (RGA). C'est le premier critère d'attribution pour un arrêté CatNat Sécheresse. Il est donc pertinent d'ajouter à la base de données, l'exposition de chaque commune face à ce phénomène.
Ces données sont fournies par le BRGM [5] et l'exposition est exprimée sous forme d'aléa.
Selon le BRGM, *"Le terme d'aléa désigne la probabilité qu'un phénomène naturel d'intensité donnée survienne sur un secteur géographique donné et dans une période de temps donnée. Cartographier l'aléa retrait-gonflement des argiles reviendrait donc à définir, en tout point du territoire, quelle est la probabilité qu'une maison individuelle soit affectée d'un sinistre par exemple dans les dix ans qui viennent."* [6]
Quatre zones d'exposition ont été créés :

— **Zone d'exposition forte :** La nature du sol de toute commune dans cette zone est fortement exposé au phénomène de RGA. La probabilité de survenance d'un sinistre CatNat Sécheresse est très élevé et l'intensité des phénomènes attendues est forte.

— **Zone d'exposition moyenne**

— **Zone d'exposition faible**

— **Zone d'exposition nulle :** La nature du sol n'est pas argileuse. Une zone où le sol n'est pas exposé au phénomène RGA.

5. Bureau de Recherches Géologiques et Minières
6. Citation issue du rapport du BRGM sur la cartographie de l'aléa RGA dans le département du marne

La surface de la commune est répartie selon ces quatre zones.
Pour rappel, les communes avec un taux d'exposition supérieur à 3% sont éligibles au premier critère de reconnaissance des arrêtés CatNat Sécheresse.

Carte de l'exposition du sol :

Le graphe ci-dessous présente la carte de l'exposition du sol du territoire Français face au retrait gonflement des sols argileux.

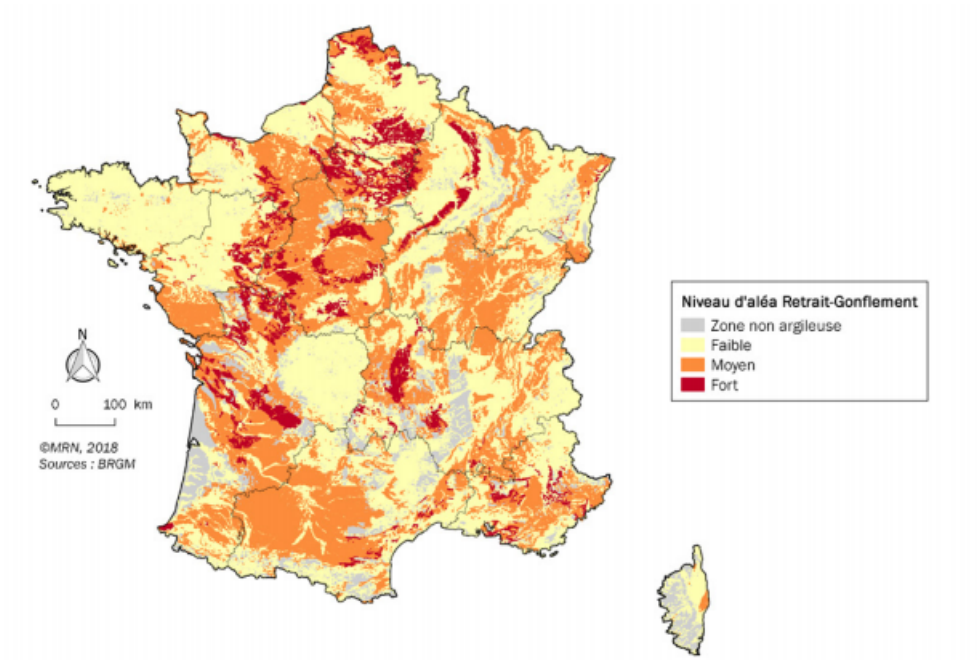

FIGURE 4.8 – Carte de l'exposition du sol face au phénomène RGA (Source : BRGM)

Constat : GOC a une exposition qualifiée de moyenne sur la quasi-totalité de son territoire (sud-ouest de la France).

Présentation des données

Cette base de données est constituée des variables suivantes :

— Le code INSEE de la commune,

— La surface d'exposition forte, faible, moyenne et nulle au phénomène RGA,

— Le taux d'exposition fort, moyen, faible et nul au même phénomène,

— La population de la commune,

— La superficie de la commune.

4.4 Données sinistres en assurance Multi-Risques Climatiques

Pour faire face à la deuxième problématique de l'étude [7], les données sinistres en assurance Multi-Risques Climatiques (MRC) ont été utilisées.

7. La non connaissance du critère météorologique SWI (Soil Wetness Index) uniforme

4.4.1 Principe de l'assurance Multi-Risques Climatiques sur Récoltes

Pour protéger leurs cultures et leurs récoltes, les agriculteurs peuvent souscrire une assurance MRC. Ce contrat d'assurance, comme son nom le précise, couvre différents aléas climatiques, dont la Sécheresse. Dès lors, et sans qu'aucune publication au JO soit faite (à la différence des bâtiments), une indemnisation peut être réalisée en cas de perte de rendement sur la culture agricole ayant été altérée par l'évènement.

Hypothèse :

Si pour une année donnée, GOC a enregistré un nombre important de sinistres MRC d'aléas Sécheresse et inondations [8], ceci peut être considéré comme un indicateur, dans la prédiction des arrêtés CatNat Sécheresse et donc dans l'estimation de la charge sinistres (provisionnement) de la Sécheresse sur bâtiments.

Avant de présenter la base de données et ses variables, la connaissance de certaines définitions de l'assurance MRC est nécessaire.

Définitions

- **charge sinistres Brut :** Le montant total du sinistre, brut de réassurance et de franchise suite à l'aléa climatique
- **Capital Assuré :** La valeur totale des récoltes assurés dans le contrat.
- **Taux de destruction :** Un taux qui indique le pourcentage de récolte qui a été détruit suite à la survenance d'un aléa climatique. Ce taux est le quotient de la charge du sinistre brut par le capital assuré.

4.4.2 Présentation des données

Cette base de données est constituée de 9 variables :

Nom de la variable	Description
Année	Exercice de survenance
CODE INSEE	Code INSEE de la commune
Capital Assuré	Capital assuré
Surface Sin	Surface sinistrée
Nbr Sin	Le nombre de sinistre MRC
Charge Sin Brut	Le montant des charges sinistres brut
Aléa	L'aléa du sinistre : Sécheresse/Inondation
Espèce	L'espèce de la culture sinistrée
Tx dest	Le taux de destruction

TABLE 4.1 – Présentation des données MRC

4.5 Construction de la base de données finale

4.5.1 Base finale à la maille "Commune"

Cette base a été construite sur des données entre 2011 et 2019 (9 années). Le nombre de communes dans le territoire de GOC est de 4791. Ainsi, la base finale est constituée de 43 120 (9 x 4791) "indi-

8. Le phénomène de retrait gonflement des sols argileux est la conséquence de fortes chaleurs suivies de fortes pluies ou l'inverse

vidus".

Pour chaque individu, les variables explicatives associées sont :

1. le département,
2. le nombre de sinistre MRC survenu,
3. l'aléa des sinistres MRC : Une variable qualitative qui prend 4 modalités possibles : Sécheresse", s'il y a eu que des sinistres MRC d'aléa Sécheresse, "Inondation", s'il y a eu que des sinistres MRC d'aléa inondation, "Sécheresse + Inondation" si les deux évènements sont survenus. En effet, une parcelle peut durant l'année subir plusieurs aléas climatiques différents (gel, grêle, vent,...) et "Pas d'aléa" si aucun sinistre MRC n'est survenu.
4. la charge brute des sinistres MRC,
5. le capital assuré,
6. la surface totale sinistrée,
7. le taux de destruction,
8. l'exposition du sol de la commune,
9. la superficie,
10. la population,
11. **Arrêté :** Variable qualitative, qui indique la publication, ou pas, d'un arrêté CatNat Sécheresse dans la commune et dans l'année. Cette variable prend donc deux valeurs (1 ou 0).

Problématique 1 : Un nombre élevé de valeurs manquantes

Pour une année et une commune donnée, il est possible qu'aucune donnée MRC ne soit disponible. Cette indisponibilité peut être due à la faible présence de GOC dans la commune ou bien à la faible présence de cultures agricoles.
La majorité des valeurs manquantes représentent des communes purement urbaines. Par exemple, la commune de Toulouse, ne compte aucune culture agricole donc logiquement n'a aucune donnée MRC. Pour pallier cette problématique, un autre découpage du territoire a été utilisé.

Notion de RA et de PRA

D'après GéoConfluences[9], *"Les régions agricoles (RA) et petites régions agricoles (PRA) constituent, en France, deux entités d'échelle différente du zonage statistique, géré par l'Insee et lancé en 1949 par le Commissariat général au Plan.*
Il s'agit dans les deux cas de zones agricoles homogènes, tant par la nature des sols que pour les conditions climatiques et la vocation dominante des exploitations agricoles. Ce zonage sert de base à la production de nombreuses statistiques agricoles."

Pour une commune et une année donnée, la procédure pour attribuer une valeur manquante d'une variable MRC est la suivante :

1. Calculer une moyenne de la variable pour chaque PRA du territoire.
2. Attribuer à la commune, la moyenne de la PRA à laquelle elle appartient.

Problématique 2 : Déséquilibre de la variable "Arrêté"

La variable "Arrêté" est une variable binaire. La figure ci-dessous présente une répartition des deux classes au sein des individus.

9. Géoconfluences est une publication géographique numérique française à caractère scientifique (7)

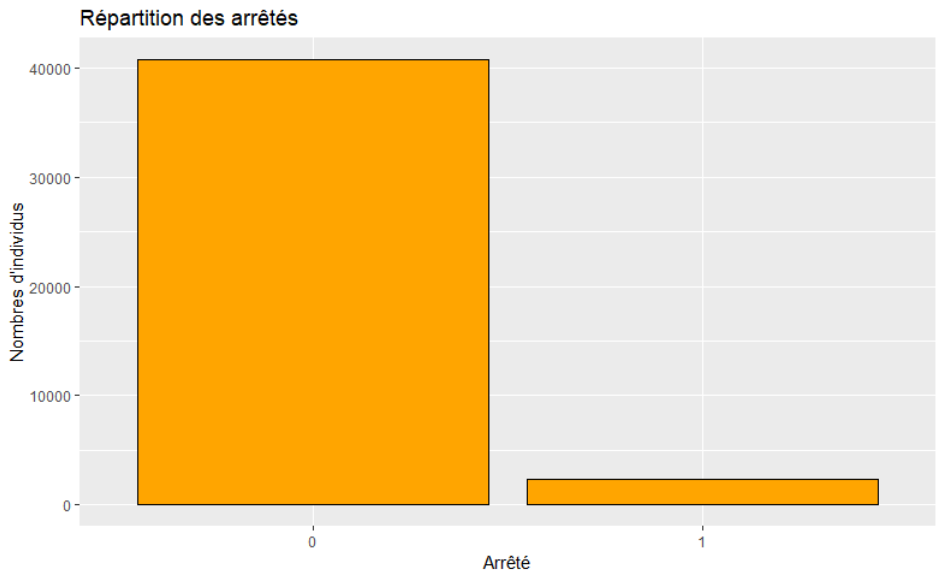

FIGURE 4.9 – Répartition de la variable arrêtés selon les individus de la base

Moins de 5% des individus de la base de données sont de classe "1". Ce déséquilibre pose problème dans l'apprentissage des modèles.

En réponse à cette problématique, deux solutions sont possibles :

1. **Le rééchantillonnage de la base de données :** Cette méthode consiste à créer synthétiquement des individus de la classe "1" et les rajouter dans la table. Plusieurs techniques de rééchantillonnage existent (sur-échantillonnage, sous-échantillonnage, méthode SMOTE [10]).
2. **Considérer un autre maillage pour la base de données** : Cette méthode consiste à réduire le spectre des individus. Par exemple, un individu de la base ne représenterait plus une commune mais un code postal.

4.5.2 Base finale à la maille "Code Postal"

Plusieurs raisons ont favorisé la constitution d'une base au maillage "Code Postal" :

— Un code postal peut regrouper plusieurs communes. Donc, dans un code postal, la probabilité d'avoir au moins un arrêté, dans une des communes, est plus élevé. Par conséquence, le déséquilibre de la variable "Arrêté" va diminuer.

— La méthode de rééchantillonnage est intéressante pour équilibrer la base d'apprentissage et ainsi augmenter la proportion d'individus de classe 1. Néanmoins, cette méthode n'apporte pas de solution à la première problématique (nombre de valeurs manquantes due aux manques d'exploitations agricoles dans les communes). En effet, dans les codes postaux, le nombre d'exploitations agricoles est plus important que pour les communes. La qualité des données MRC n'en sera que plus robuste.

D'autres maillages ont été considérés. Par exemple, le maillage "Département", mais la base finale serait constituée de 126 individus (14 départements x 9 années). Cette population est faible et ne permet pas de réaliser un travail de modélisation.

Répartition du nombre d'arrêtés dans la base

En effectuant ce changement de maillage, la variable "Arrêté" n'est plus binaire. Par exemple, un code postal qui regroupe 5 communes, peut avoir 2 ou 3 arrêtés CatNat Sécheresse publiés.
La figure ci-dessous illustre ces propos.

10. Synthetic Minority Over-sampling Technique

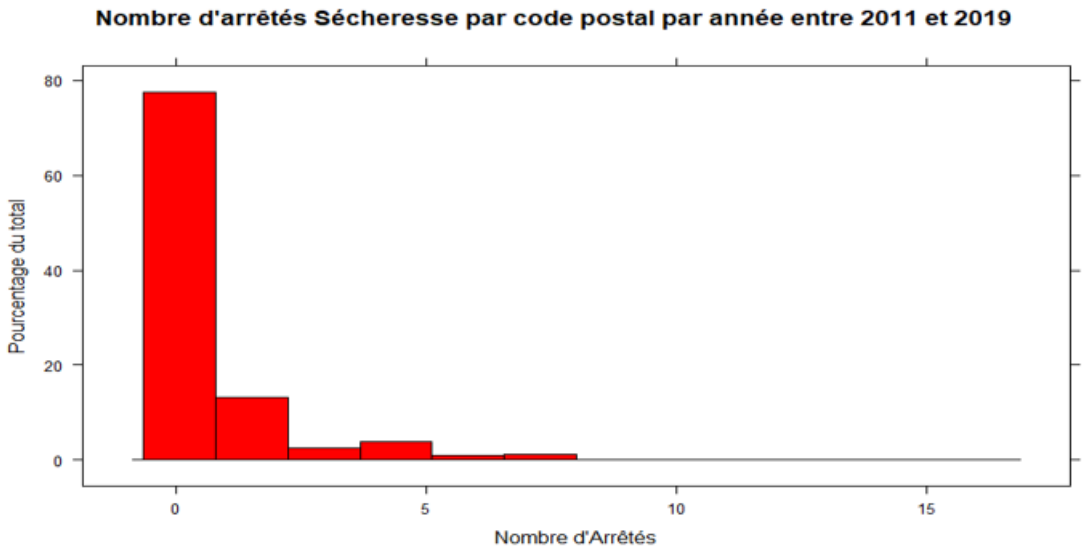

FIGURE 4.10 – Nombre d'arrêtés par code postal entre 2011 et 2019

Constat :
80% des codes postaux n'ont pas connu d'arrêté CatNat Sécheresse.

Hypothèse :
Pour simplifier le problème, deux classes seront considérés :

— La classe 1 : Au moins un arrêté CatNat Sécheresse publié dans le code postal.

— La classe 0 : Aucun arrêté n'a été publié dans le code postal et dans l'année.

Conclusion

La base de données étant construite, le prochain chapitre s'attachera à présenter la modélisation prédictive des arrêtés Sécheresse.

Chapitre 5

Modélisation de la fréquence des Sécheresse

"Les parties théoriques de ce chapitre sont basées sur des notes de cours et des livres de statistiques. Les références sont mentionnées dans la Bibliographie."

Introduction

L'objectif de ce chapitre est de présenter de quelle façon la publication d'au moins un arrêté CatNat Sécheresse favorable est prédite.
Afin de modéliser la fréquence des arrêtés CatNat Sécheresse, le processus suivant a été suivi :

1. Création d'une base de données,
2. Analyse descriptive de la base de données,
3. Division de la base de données en deux sous bases : Une base d'apprentissage et une base de test,
4. Apprentissage des différents modèles sur la base d'apprentissage.
5. Application des modèles sur la base de test,
6. Calcul des critères de performance selon les prédictions de chaque modèle,
7. Comparaison des résultats et choix du meilleur modèle.

Les deux premières étapes ont été illustrées dans le chapitre précédent. Ce chapitre se focalisera sur les présentations théoriques des modèles, la définition des critères de performances, l'apprentissage et le calibrage de chaque modèle testé et utilisé, suivi des résultats obtenus.

5.1 Métriques [1] [2]

Afin de comparer les résultats de chaque modèle, la définition de critères de performance est nécessaire. Ces critères sont appelés "métriques" [3].
Afin d'optimiser le choix du meilleur modèle, plusieurs métriques ont été considérées :

— Accuracy (taux de bien classés),

— La précision,

— Le rappel,

— Le F1 score,

1. B. NATANECLIC. Machine Learning : Précision, F1 Score, Courbe ROC, que choisir ?. (25)
2. MACHINE LEARNING GOOGLE COURSES. Classification : ROC Curve and AUC. (23)
3. Une métrique a pour but d'évaluer un modèle de prédiction en quantifiant la qualité des prédictions obtenus.

— Sensibilité, Spécificité et Courbe ROC.

Avant d'entrer plus en détail dans les caractéristiques de chaque métrique, il faut classifier les prédictions obtenues. En effet, une prédiction peut être soit correcte, soit fausse. Il est donc nécessaire de créer un tableau ou une matrice qui classifie ces prédictions.

5.1.1 Matrice de confusion

La définition d'une matrice de confusion est proposée par (25) : *"Une Confusion Matrix (matrice de confusion) ou tableau de contingence est un outil permettant de mesurer les performances d'un modèle de Machine Learning en vérifiant notamment à quelle fréquence ses prédictions sont exactes par rapport à la réalité dans des problèmes de classification."* Le tableau ci-dessous illustre une matrice de confusion.

Classes prédites	Classes réelles	
	0	1
0	Vrais négatifs	Faux négatifs
1	Faux positifs	Vrais positifs

TABLE 5.1 – Matrice de confusion

Interprétation de la matrice :

Les "Vrais Négatifs" (VN) : Le modèle ne prédit pas de Sécheresse et aucune Sécheresse n'est survenue.

Les "Faux Négatifs" (FN) ; Le modèle ne prédit pas de Sécheresse mais une Sécheresse a été publiée au JO.

Les "Faux Positifs" (FP) : Le modèle prédit une Sécheresse mais aucune Sécheresse n'est survenue.

Les "Vrais Positifs" (VP) : Le modèle prédit une Sécheresse et, effectivement, une Sécheresse est survenue.

5.1.2 Accuracy

L'Accuracy est la proportion des individus bien prédits par le modèle par rapport à l'ensemble des individus. Elle mesure donc le taux de prédictions correctes.

La formule de l'Accuracy est la suivante :

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN} = \frac{Biens\,prédits}{Total\,des\,individus}$$

Cette métrique est intuitive mais n'est pas adaptée à la base de données utilisée dans cette étude. En effet, la proportion des individus de classe 0 par rapport à la classe 1 est de 80%. Si le modèle prédit qu'aucune sécheresse n'aura lieu (prédire 0 pour tous les individus), ce dernier aura une Accuracy de 80% c'est-à-dire un taux de réussite de 80%. Ce dernier semble élevé mais ne donne réellement aucune information quant à la qualité des prédictions.

L'objectif de cette étude est de détecter les sécheresses donc de prédire les individus de classe 1. Cette métrique n'est alors pas intéressante et inadaptée dans ce cadre. Pour cela, d'autres métriques ont été considérées.

5.1.3 La précision, le rappel et le F1 Score

La précision

La précision correspond au nombre d'individus correctement attribués à la classe 1 par rapport au nombre total d'individus **prédits comme appartenant** à cette classe.

La formule de la précision est la suivante :

$$Précision = \frac{VP}{VP + FP} = \frac{Vrais\, positifs}{Total\, des\, prédictions\, positives}$$

Dans le cas de l'étude, la précision représente la proportion des codes postaux où le modèle a prédit correctement des arrêtés.

La précision est une métrique comprise entre 0 et 1. Plus elle s'approche de 1, plus le modèle est précis et donc meilleur.

Cette métrique permet de détecter les individus de classe 0 mal prédits (Les Faux Positifs). Elle accorde donc une importance à la classe 0. En effet, lorsque la précision vaut 1, cela veut dire que tous les individus prédits en classe 1 sont réellement des individus qui appartiennent à cette classe. Par conséquent, aucun individu de la classe 0 n'a été mal prédit.

Le rappel

Le rappel correspond au nombre d'individus correctement attribués à la classe 1 par rapport au nombre total d'individus **appartenant** à cette classe.

La formule du rappel est la suivante :

$$Rappel = \frac{VP}{VP + FN} = \frac{Vrais\, positifs}{Total\, des\, réels\, positifs}$$

Dans le cas de l'étude, le rappel représente la proportion des Sécheresse prédites par le modèle qui ont été réellement publiées.

Le rappel est aussi une métrique qui est comprise entre 0 et 1. Plus le rappel s'approche de 1 plus le modèle est meilleur.

Cependant, le rappel est une métrique qui accorde une importance à la classe 1. En effet, si le rappel vaut 1, cela veut dire que tous les individus des classes 1 ont été bien prédit et donc aucun individu de cette classe n'a été mal prédit.

L'objectif de l'étude est de prédire correctement les Sécheresse dans les codes postaux. L'analyse de la base a montré un déséquilibre de la variable à prédire. Donc, le modèle aura tendance à prédire des individus de classe 0 (classe majoritaire) et donc d'engendrer un nombre important de FN. L'objectif est donc de maximiser la proportion de Sécheresse détectées, ce qui revient à maximiser le rappel.
Néanmoins, si le rappel vaut 1 avec une précision médiocre, un nombre important de FP vont apparaître et le nombre de Sécheresse prédites sera surestimé. De ce fait, il est nécessaire de trouver une métrique qui harmonise la précision et le rappel.

Le F1 Score

Le F1 Score est une combinaison subtile de la précision et du rappel. Cette métrique correspond à une moyenne harmonique entre la précision et le rappel :

$$F1 = \frac{2 \times Rappel \times Précision}{Rappel + Précision}$$

L'avantage majeure du F1 Score est qu'il ne prend pas en compte les Faux Négatifs (FN).

Dans le cadre de l'étude, le F1 Score donne une importance à la précision du modèle (le nombre de Sécheresse prédites correctement) et sa robustesse (un nombre minime de Sécheresse non prédites).

5.1.4 Sensibilité, Spécificité, Courbe ROC

D'après la proposition de (23), *"Une courbe ROC (Receiver Operating Characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification"*. Pour bien comprendre le concept d'une courbe ROC, il faut définir le Seuil, la Sensibilité et la Spécificité d'un modèle.

Seuil

Un modèle prend en compte un nombre de variables et renvoie une probabilité, pour un individu, d'appartenir à une classe. Le seuil représente la probabilité à partir de laquelle l'individu est considéré appartenir à cette classe.

Sensibilité

La sensibilité mesure la proportion des individus de classes 1 qui a été correctement prédit. La sensibilité a donc la même formule et les mêmes caractéristiques que le rappel.

$$Sensibilité = \frac{VP}{VP + FN} = \frac{Vrais\, positifs}{Total\, des\, réels\, positifs}$$

Spécificité

La spécificité est l'inverse de la sensibilité. Elle mesure la proportion d'individus de classes 0 qui a été correctement prédit.
Sa formule est la suivante :

$$Spécificité = \frac{VN}{VN + FP} = \frac{Vrais\, négatifs}{Total\, des\, réels\, négatifs}$$

La formule de la Spécificité est complémentaire avec la formule de la Sensibilité. Si la spécificité augmente, la sensibilité diminue et vice versa. Par exemple, si la spécificité vaut 1, la sensibilité sera très faible et le modèle ne sera pas bon. Pour bâtir un modèle robuste, il faut qu'il soit sensible et spécifique et donc un équilibre entre les deux métriques.

Courbe ROC

La Courbe ROC est une courbe qui trace la sensibilité par rapport à la spécificité à différents seuils de classification.

Si le seuil varie, la classification des individus varie également et donc les sensibilités et spécificités varient à leurs tours. La courbe ROC permet de représenter ces variations. La sensibilité sera représentée sur l'axe des ordonnées et (1-Spécificité) sur l'axe des abscisses.

La figure ci-dessous permet d'illustrer un exemple d'une courbe ROC.

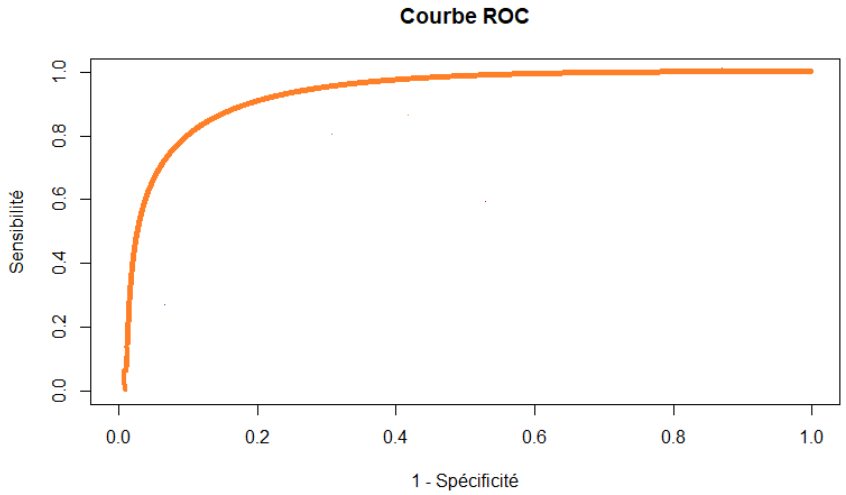

FIGURE 5.1 – Exemple d'une Courbe ROC

Grâce à cette courbe, il est possible de comparer les modèles et de choisir un seuil optimal qui maximise le meilleur compromis entre la spécificité et la sensibilité.

Pour comparer des modèles en utilisant la courbe ROC, il faut calculer l'aire sous la courbe (AUC : Area Under Curve). Plus l'aire sous la courbe est grande, plus le modèle est meilleur. Ceci est illustré avec la figure ci-dessous. Le modèle de la courbe orange est meilleur que le modèle de la courbe bleue qui est à son tour meilleur que le modèle de la courbe rouge.

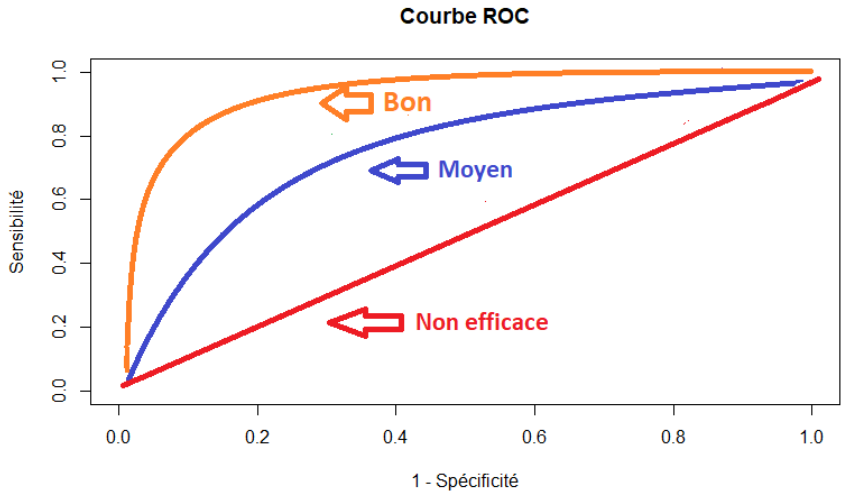

FIGURE 5.2 – Comparaison de modèle avec une courbe ROC

Après la définition de toutes ces métriques, il est désormais possible de mesurer les performances des modèles testés.

5.2 Régression Logistique [4] [5] [6]

Le modèle de régression logistique est une procédure de modélisation mathématique très populaire dans l'analyse des données d'épidémiologie et l'apprentissage automatique. Cette section s'appuie sur la théorie développée dans la littérature et développée dans (13), (15), (17).

5.2.1 Principe Général

Ce modèle est un cas particulier des modèles linéaires généralisés (GLM) qui sont une extension des modèles linéaires. Pour bien comprendre le modèle de régression logistique, des rappels sur les modèles linéaires et les modèles linéaires généralisés sont proposés.

Modèles linéaires

Un modèle linéaire est un modèle qui cherche à expliquer une variable $Y = \mu + \epsilon$, dite variable réponse, à l'aide de p variables explicatives $X_1, ..., X_p$ avec $\mu = E[Y]$ et ϵ une variable aléatoire.
Le modèle linéaire suppose que μ s'écrit comme une fonction linéaire des X_i. Le modèle s'écrit alors de la façon suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

avec : $(\beta_0, ..., \beta_p) \in \mathbb{R}^{p+1}$, les paramètres à estimer du modèle et ϵ suit une loi normale centrée réduite, i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Sous ces hypothèses, $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Si la variable à expliquer Y est une variable qualitative (sexe, genre, département,...) et appartient à un nombre défini de classes, il est impossible de modéliser directement cette variable par une relation linéaire. De même, si la variable Y est une variable discrète ou appartient à un sous ensemble de $\mathbb{R}$, le modèle linéaire à résidus gaussiens ne serait pas adapté. Ces modèles supposent que la variable à prédire Y soit à valeur continue dans $\mathbb{R}$. Afin de pallier ces problématiques, les modèles linéaires généralisées ont été créés.

Modèles linéaires généralisés

Les modèles linéaires généralisés (GLM) sont une généralisation des modèles linéaires. Les modèles GLM permettent aux modèles linéaires d'être reliés à la variable réponse à travers une fonction lien.

A l'opposé des modèles linéaires qui supposent la normalité de la variable réponse, dans les modèles généralisés linéaires, la loi de la variable réponse Y est supposée appartenir à un ensemble restreint de lois dites lois exponentielles GLM.

Trois éléments sont proposés par (17) pour caractériser un modèle linéaire généralisé :

— La variable réponse à expliquer Y, appelée aussi composante aléatoire et qui suit une loi qui appartient à la famille de lois exponentielles.

— La composante déterministe appelée aussi prédicteur linéaire $\eta : \eta(x) = \sum_{j=0}^{p} x_j \beta j$

— La fonction lien g supposée bijective et dérivable, qui a pour rôle de lier les paramètres de la loi à une combinaison linéaire des prédicteurs : $g(E[Y]) = \eta$

4. P-A. CORNILLON et E. MATZNER-LOBBER. Régression : Théorie et applications. Édition PUR. 2007. (17)
5. P. AILLIOT. Notes de cours sur les méthodes de régression. EURIA. 2016. (13)
6. L.ROUVIERE. Notes de cours : Régression Logistique avec R. Université de Rennes 2. (15)

L'approche d'un modèle GLM consiste alors à choisir une loi pour la variable réponse Y parmi un ensemble restreint de lois, ensuite de choisir une fonction lien g bijective et dérivable et enfin de modéliser la transformation de $E[Y]$ par la fonction g avec une combinaison linéaire grâce à la fonction η :

$$g(E(Y|X=x)) = \eta(x) = \sum_{j=0}^{p} x_j \beta j$$

Dans le tableau ci-dessous, quelques exemples de modèles GLM sont présentés.

Loi	**Nom du lien**	**Fonction lien**
Poisson	lien log	$g(x) = log(x)$
Normale	lien identité	$g(x) = x$
Gamma	lien réciproque	$g(x) = -1/x$

TABLE 5.2 – Exemples de fonctions liens pour les modèles GLM

Critères de sélection des variables

Dans un modèle linéaire généralisé, plusieurs variables explicatives font partie du modèle. Certaines de ces variables peuvent ne pas être significatives, d'autres au contraire peuvent avoir une plus grande importance. Pour améliorer la qualité d'un modèle, il faut optimiser le choix des variables utilisées.

Les paramètres du modèle sont estimés grâce à la théorie du maximum de vraisemblance. Pour tester la significativité d'une ou plusieurs variables explicatives, il faut tester l'hypothèse H_0 qui suppose la nullité des coefficients devant ces variables. Si le coefficient devant la variable explicative est nul, cela veut dire que la variable n'est pas significative. Par exemple, pour tester la significativité des q premières variables explicatives, avec $q >= 0$ alors l'hypothèse H_0 s'écrit :

$H_0 : \beta_0 = \beta_1 = ... = \beta_{q-1} = 0$ contre $H_1 : \exists k \in [|0, q-1|] : \beta_k \neq 0$

Plusieurs tests pour valider ou rejeter cette hypothèse sont possibles. Les tests suivants sont proposées par (15) :

Test de Wald

On note $\beta_{0,..,q-1}$ le vecteur composé des q premiers coefficients et $\sum^{-1}_{0,..,q}$ la matrice bloc composée des q premières lignes et colonnes de $\sum^{-1}$, alors sous H_0 :

$$\hat{\beta}'_{0,...,q-1} \hat{\sum}_{0,..,q} \hat{\beta}_{0,...q-1} \longrightarrow \chi^2$$

avec $\sum^{-1} = (X'WX)^{-1}$ l'inverse de la matrice d'information observée.

Test du rapport de vraisemblance

Le rapport de vraisemblance, dit aussi différence de déviance, entre un modèle avec $q1$ variables explicatives et un autre modèle avec $q2$ variables explicatives avec $q2 > q1$ permet d'évaluer l'apport des variables explicatives ajoutés dans le deuxième modèle.

Le test du rapport de vraisemblance compare le modèle sous l'hypothèse H0 et le modèle complet qui prend en compte toutes les variables.

$$2(L_n(\hat{\beta}) - L_n(\hat{\beta_{H_0}})) \longrightarrow \chi^2_q$$

avec : $\hat{\beta_{H_0}}$ l'estimateur de maximum de vraisemblance sous l'hypothèse H_0

Pour ces deux tests, l'hypothèse H_0 est à rejeter si la valeur observée de la statistique du test dépasse le quantile d'ordre $1 - \alpha$ de la loi χ^2_q.

Critère de sélection des modèles

Similairement à la comparaison des variables explicatives et de leur importance, une comparaison de la qualité des modèles est aussi possible.
Plusieurs critères pour estimer qu'un modèle est mieux qu'un autre sont proposés par (17). Parmi eux :

L'Akaike Information Criterion (AIC)

L'AIC est un critère introduit par Akaike en 1973. Il est défini par :

$$AIC = 2p - 2L$$

avec : p le nombre de paramètres du modèle et L le maximum de vraisemblance du modèle.
L'AIC représente une pénalisation de la vraisemblance. Plus un AIC d'un modèle est faible, plus le modèle est de meilleure qualité.

Le critère Bayesian Information Criterion (BIC)

Ce critère a été introduit par Schwartz en 1978 et est défini par :

$$BIC = -2 \ln L + \ln(n).p$$

avec : L la vraisemblance, p le nombre de paramètres et n le nombre d'observations.
Le BIC est également un critère de pénalisation de la vraisemblance mais qui rajoute au nombre de paramètres le nombre d'observations utilisées par le modèle.
Plus le nombre d'observations n augmente plus le BIC diminue. Pour choisir le meilleur modèle, il faut choisir celui avec le BIC le plus faible.

Déviance

Le troisième critère proposé par (17) est la Déviance.
Ce critère se définit de la façon suivante :

$$D = 2(L - L_{sat})$$

avec : L la vraisemblance du modèle et L_{sat} la déviance du modèle saturé. Le modèle saturé est un modèle qui possède autant de paramètres que de données.
D'après (17) : *"La déviance représente l'écart en terme de log-vraisemblance entre le modèle saturé d'ajustement maximal et le modèle considéré. Plus la déviance est faible, plus le modèle est bien ajusté."*

Procédure de sélection des modèles

Après la définition des critères qui permettent de choisir un modèle de qualité, il faut sélectionner ce modèle parmi tous les modèles possibles.

Deux types de recherches existent pour trouver le modèle optimal : La recherche exhaustive et la recherche pas à pas. Le principe de ces deux recherches est proposé par (15) :

La recherche exhaustive évalue tous les modèles possibles quand le nombre de ces modèles est fini. Si tous les modèles avec p variables sont possibles, il y a 2^p possibilités, donc si p est grand cette méthode n'est pas possible.

La recherche pas à pas est utilisée lorsque p est grand. Cette méthode ne permet pas d'obtenir une optimisation parfaite mais permet d'obtenir un optimum local. Le principe de cette procédure est de partir d'un point de départ ensuite de répéter la procédure jusqu'à arriver au meilleur modèle.

Toujours selon (15), trois types de recherche pas à pas existent :

Méthode ascendante (forward selection)

Lors de cette procédure, le modèle de départ est un modèle sans variable explicative. Ensuite, à chaque étape, la variable rajoutée est la variable dont le rajout augmente au maximum les critères de choix. Cette étape se répète jusqu'à ce que toutes les variables soient rajoutées ou que les critères de choix n'augmentent plus.

Méthode descendante (backward selection)

Avec cette méthode, le modèle de départ est le modèle complet avec toutes les variables descriptives. A chaque étape de la procédure, la variable retirée est la variable dont le retrait augmente le plus les critères de choix. Cette étape se répète jusqu'à ce que toutes les variables soient retirées ou qu'aucune augmentation des critères ne soit plus possible.

Méthode progressive (stepwise selection)

La méthode progressive est une méthode hybride de la méthode ascendante et descendante. Lors de chaque étape de la procédure, une variable est rajoutée et une variable est retirée.

Régression Logistique [7]

Cette section s'appuie sur la théorie développée dans (13).

Le modèle de régression logistique est un modèle de régression qui a pour objectif d'expliquer une variable qualitative Y, le plus souvent binaire (0 ou 1, oui ou non, survenance d'une Sécheresse ou pas,...) à partir d'un certain nombre de variables explicatives $X = (X_1, ..., X_p)$.

Dans le modèle de régression logistique, la variable réponse $Y = (Y_1, ..., Y_n)$ est un n-uplet de variables aléatoires indépendantes et pour tout $i \in 1, ..., n$, Y_i suit une loi de Bernouilli dont les paramètres dépendent des variables explicatives :

$$P[Y_i = 1] = \pi(x_{i,1}, ..., x_{i,p})$$

Selon (13), pour trouver une formule simple de la fonction π, le modèle linéaire a été pris en considération :

$$g(\pi(x_{i,1}, ..., x_{i,p})) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}$$

7. P. AILLIOT. Notes de cours sur les méthodes de régression. EURIA 2016. (13)

avec $g : [0,1] \mapsto \mathbb{R}$ la fonction lien du modèle. Cette fonction a pour objectif de transformer le paramètre de la loi de Bernouilli en un paramètre variant dans $\mathbb{R}$. Généralement, la fonction lien du modèle de régression logistique est la fonction logit.

Fonction logit

La fonction logit est une fonction bijective et dérivable de $]0,1[$ dans $\mathbb{R}$ telle que : $p \mapsto \ln(\frac{p}{1-p})$
Les deux figures ci-dessous illustrent les courbes de la fonction logit et de son inverse.

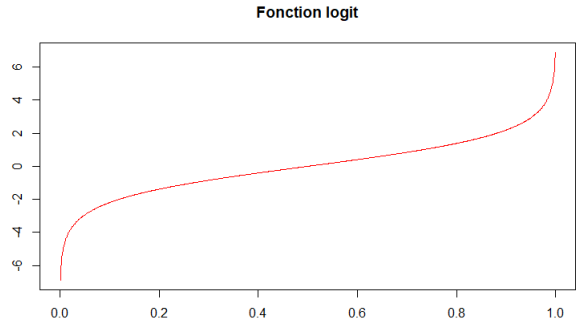

FIGURE 5.3 – Fonction logit

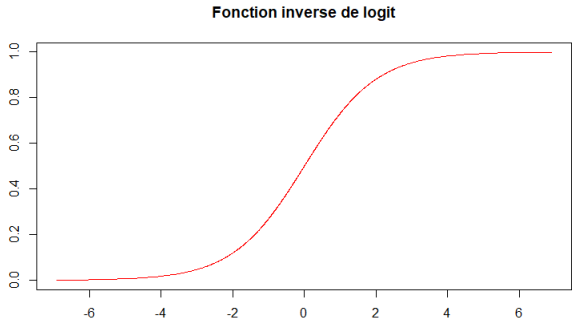

FIGURE 5.4 – Fonction inverse de logit

Odds Ratio (OR)

D'après la définition de la fonction logit :

$$logit(P[Y_i = 1]) = \ln(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}) = \ln(\frac{P[Y_i = 1]}{P[Y_i = 0]})$$

Le modèle de régression logistique s'écrit alors :

$$\ln(\frac{P[Y_i = 1]}{P[Y_i = 0]}) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}$$

avec $\frac{P[Y_i=1]}{P[Y_i=0]}$ le rapport entre la probabilité de succès et la probabilité d'échec. Ce rapport est aussi est appelé "Odds Ratio" (OR). Les coefficients $\beta_1, ..., \beta_p$ sont estimés par la méthode du maximum de vraisemblance. L'équation du modèle peut aussi s'écrire de la façon suivante :

$$P[Y_i = 1] = \frac{\exp(\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p})}$$

La définition suivante est proposée par (13) :

$$OR = \frac{P[Y_i = 1]}{P[Y_i = 0]}$$

Le rapport entre la probabilité de succès et la probabilité d'échec est dite Odds Ratio (OR). Pour une variable explicative X_i, son odds ratio correspond à l'exponentielle du coefficient β_i. Un odds ratio de 1 signifie l'absence d'effet de la variable sur le phénomène étudié. Quand l'odds ratio augmente et devient largement supérieur à 1, cela correspond à l'augmentation du phénomène étudié. Si au contraire l'odds ratio diminue et devient largement inférieur à 1, cela correspond à une diminution du phénomène étudié.

5.2.2 Application à la problématique [8]

L'application du modèle avec le logiciel R est basée sur un exemple fourni par (22).

Le modèle de régression logistique est utilisé dans cette étude pour expliquer et prédire la variable "Arrêté" qui est une variable binaire.
La variable "Arrêté" prend 0 si aucun arrêté n'a été publié dans le code postal et dans l'année, elle prend 1 si au moins un arrêté a été publié. Le but de cette étude est donc d'expliquer et de prédire cette variable par les 16 variables descriptives dont 2 variables qualitatives.
Le tableau ci-dessous rappelle les variables explicatives de la base de données :

Variable	**Type**	**Code**
Le département	Qualitative	Département
L'aléa des sinistres MRC	Qualitative	Aléa
Charge brute des sinistres MRC	Quantitative	Charge_Sinistre_Brut
Le capital assuré	Quantitative	Capital_Assuré
La surface sinistrée	Quantitative	Surface_Sinistrée
Le taux de destruction	Quantitative	Tx_Dest
La surface où l'exposition du sol est forte	Quantitative	Surface_Expo_Forte
La surface où l'exposition du sol est moyenne	Quantitative	Surface_Expo_Moy
La surface où l'exposition du sol est faible	Quantitative	Surface_Expo_Faible
La surface où l'exposition du sol est nulle	Quantitative	Surface_Non_Expo
Le pourcentage du sol où l'exposition est forte	Quantitative	Tx_Expo_Forte
Le pourcentage du sol où l'exposition est moyenne	Quantitative	Tx_Expo_Moy
Le pourcentage du sol où l'exposition est faible	Quantitative	Tx_Expo_Faible
Le pourcentage du sol où l'exposition est nulle	Quantitative	Tx_Non_Expo
Le nombre de sinistres MRC	Quantitative	Nbr_Sinistre_MRC
La population	Quantitative	Population
La superficie du code postal	Quantitative	Superficie

TABLE 5.3 – Présentation des variables explicatives

La variable "Département" peut prendre 14 valeurs : 09, 12, 15, 19, 23, 31, 32, 40, 46, 48, 64, 65, 81, 82 (les départements du territoire de GOC). La variable "Aléa" peut prendre 4 valeurs : "Inondation", "Sécheresse", "Sécheresse + Inondation" et "Pas d'aléa"

Pour appliquer le modèle, 5382 individus sont disponibles dans la base. La modélisation sera réalisée sur le logiciel R. La base de données a été séparée aléatoirement en deux sous bases, une base d'apprentissage sur laquelle le modèle va apprendre et une base de test sur laquelle le modèle va être testé. L'objectif est que le modèle arrive à bien prédire les individus de la base de test.

8. J. LARMARANGE. Régression logistique binaire, multinomiale et ordinale. (22)

Premier modèle : Modèle Complet

Pour commencer, le modèle complet est appliqué sur la base d'apprentissage et toutes les variables explicatives sont prises en considération.
Dans le modèle logistique, à la place d'étudier les coefficients du modèle, c'est les Odds Ratio (OR) qui sont étudiés. Si un OR diffère de 0, c'est équivalent à dire que le coefficient de la variable est différent de 0.
Le graphique ci-dessous permet d'illustrer les résultats du modèle.

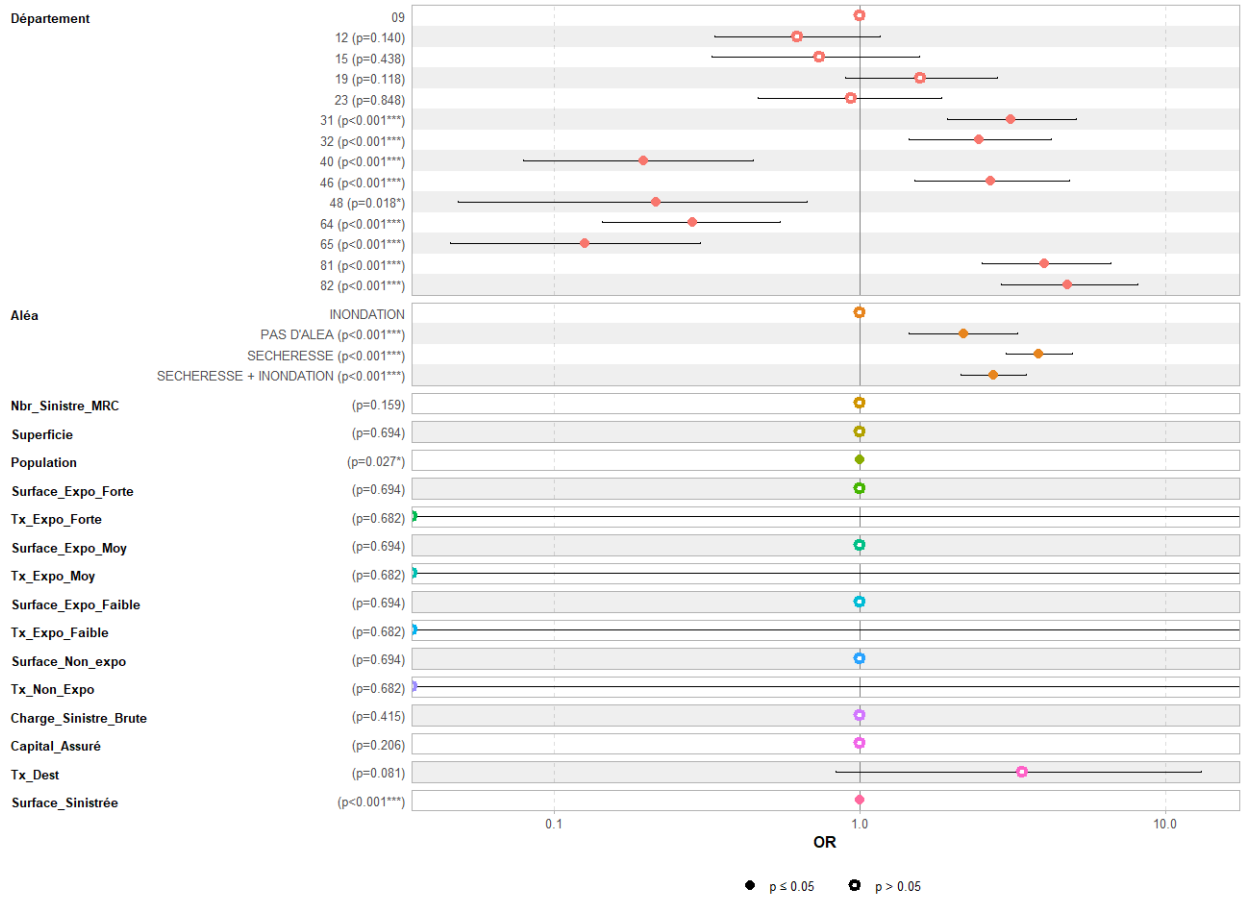

FIGURE 5.5 – Représentation graphique du modèle complet

Interprétation graphique :

— Les variables qui ont un OR supérieur à 1 sont le taux de destruction, le département et l'aléa. Cela veut dire que ces variables ont un effet sur la probabilité de la publication d'un arrêté.

— Pour la variable "Département", certains départements augmentent la probabilité qu'il y'ait un arrêté (OR>1) et d'autres au contraire diminuent cette probabilité (OR<1).

— Les variables avec un cercle pointillé ont une p-value supérieur à >5%. Un nombre important de variables ont une p-value supérieur à 5%. Ces variables ne sont donc pas significatives et peuvent être enlevées du modèle.

Effet des variables explicatives

La variable Département peut prendre 14 valeurs. Chaque valeur n'a pas le même effet sur la probabilité finale. Le graphe ci-dessous permet de voir les différents effets de cette variable.

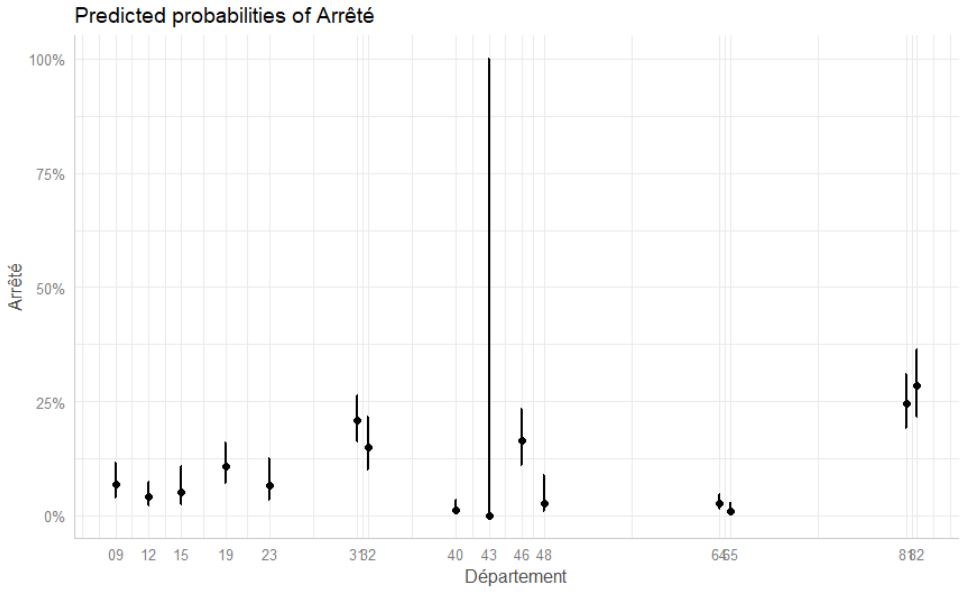

FIGURE 5.6 – Représentation graphique de l'effet des départements

D'après la figure ci-dessus les départements qui ont le plus d'effet sur la probabilité sont le 31, 32, 81 et le 82. En pratique ces départements sont les départements les plus urbains du territoire, donc avec un nombre élevé de maisons (cela implique plus de sinistrés donc plus de pression pour la mairie de demander la reconnaissance auprès de la commission).

La figure ci-dessous permet d'illustrer les différents effets de la variable "Aléa" sur la probabilité.

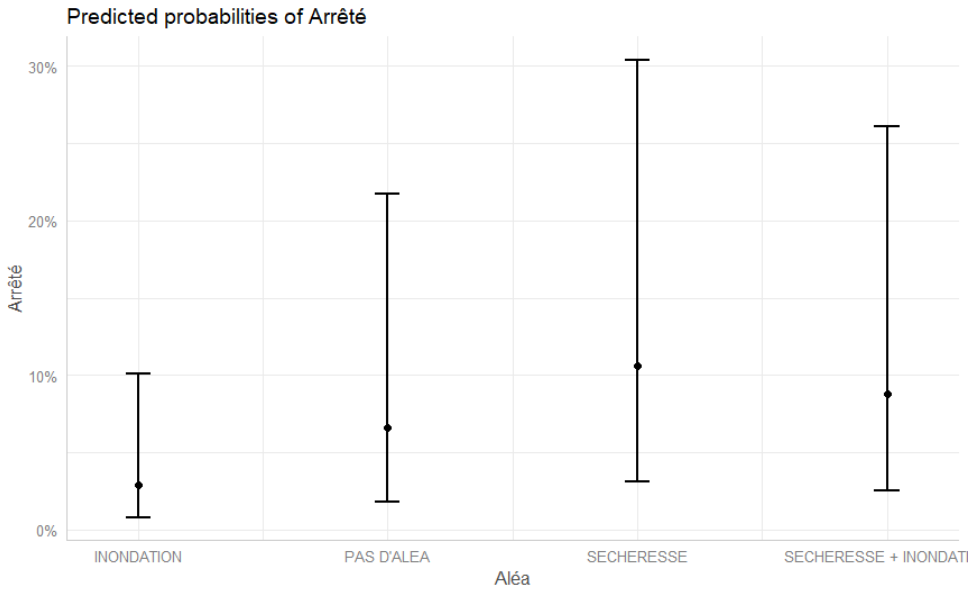

FIGURE 5.7 – Représentation graphique de l'effet de l'aléa

Selon cette figure, quand il y a de la Sécheresse (en MRC) dans les codes postaux, ça augmente la probabilité de publication d'arrêtés CatNat augmente. De même, quand il y a une survenance de Sécheresse et d'inondations ensemble.

La figure ci-dessous permet d'illustrer l'effet de la variable "Taux de Destruction" sur la probabilité qu'il y ait au moins un arrêté publié.

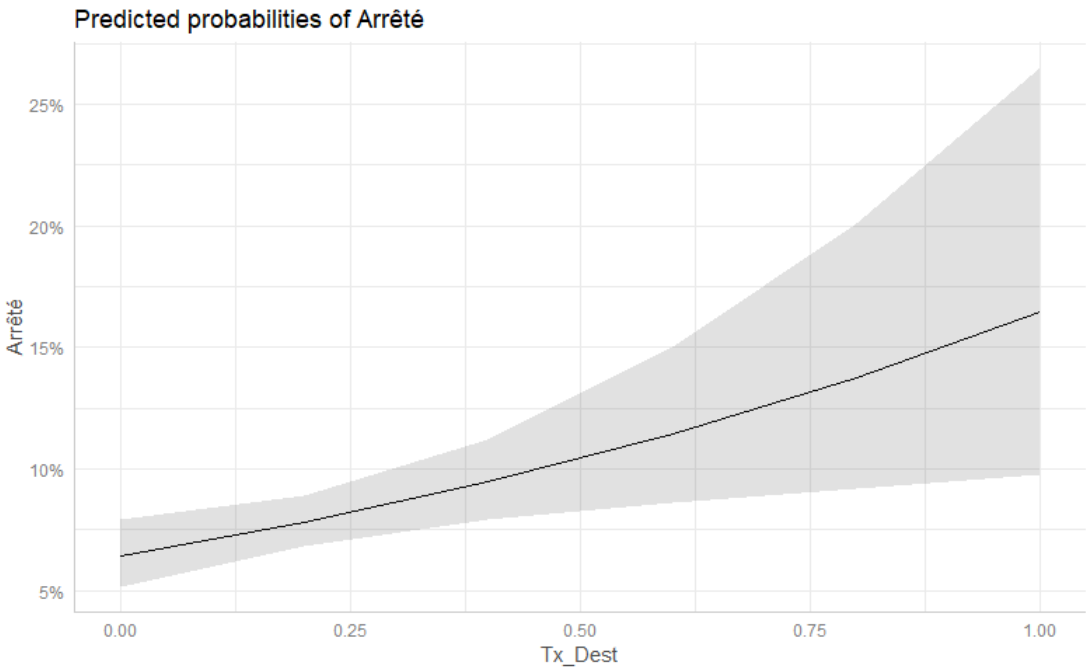

FIGURE 5.8 – Représentation graphique de l'effet du taux de destruction

D'après cette figure, si dans un code postal le taux de destruction en assurance MRC augmente, la probabilité qu'un arrêté soit publié augmente.

Sélection de modèle

Clairement le modèle complet n'est pas un bon modèle, plusieurs variables ne sont pas significatives et n'ont pas d'effet sur la probabilité. Il faut choisir un meilleur modèle via la méthode de recherche pas à pas descendante. Le critère de choix utilisé est l'AIC.
Le tableau ci-dessous permet de comparer les AIC du modèle complet et du modèle réduit.

Modèle	**Variables utilisés**	**AIC**
Modèle Complet	Toutes les variables explicatives	3563
Modèle réduit final	Département + Aléa + Population + Tx_Expo_Moy + Tx_Expo_Faible + Tx_Dest + Surface_Sinistrée + Superficie	3550.6

TABLE 5.4 – Sélection du modèle avec le AIC le plus faible

Deuxième modèle : Modèle réduit

Le graphe ci-dessous permet d'illustrer les résultats du deuxième modèle qui est composé des 8 variables sélectionnées.

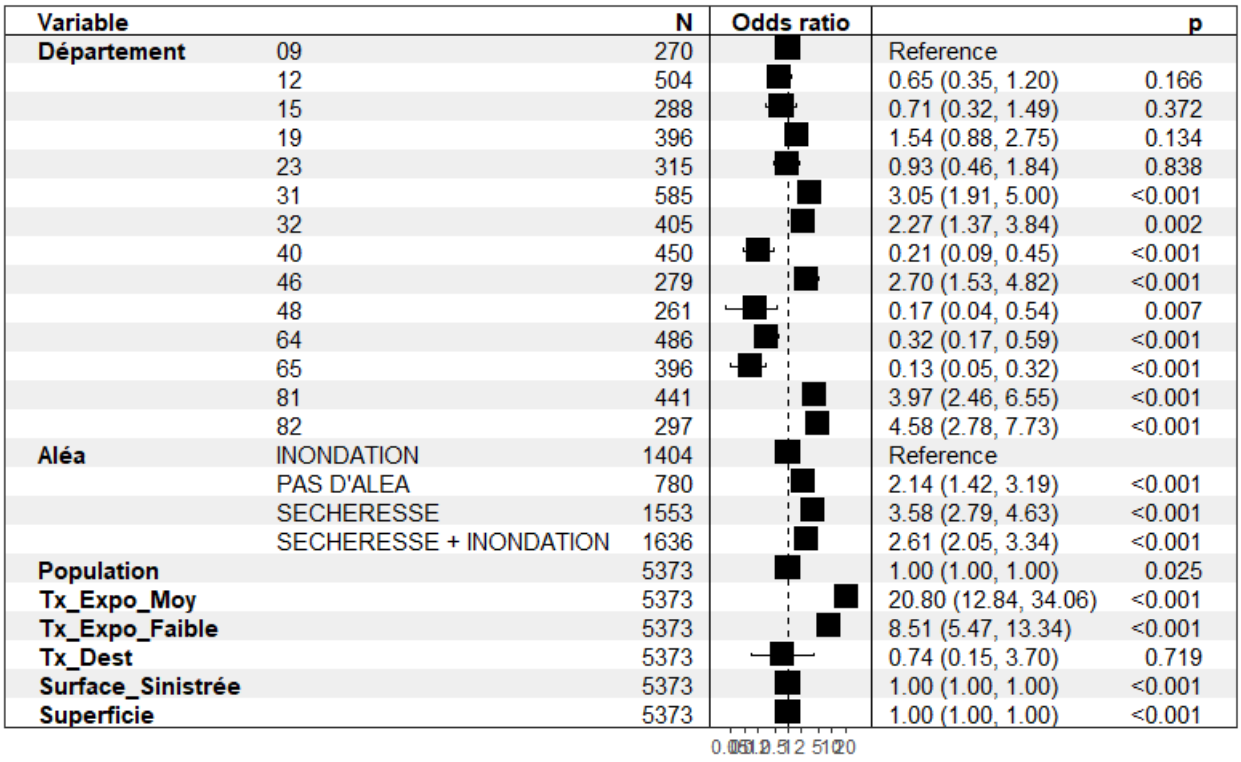

FIGURE 5.9 – Représentation graphique du modèle réduit

Interprétation Graphique :

— Le nombre de variables avec une p-value supérieur à 5% a diminué.

— La variable "Tx_Expo_Moy" (le pourcentage du sol qui est exposé moyennement face au phénomène de RGA) est la variable qui a le plus grand OR (variable qui a le plus d'effet sur la probabilité qu'il y ait une Sécheresse).

— L'effet de la variable "Tx_Dest" a clairement diminué.

5.2.3 Résultats du modèle

Après l'analyse et le calibrage du modèle, il faut tester ce dernier sur la base de test.
Le matrice ci-dessous est la matrice de confusion obtenue après les prédictions du modèle sur les individus de la base de test. La courbe ROC est aussi représentée.

Classes prédites	Classes réelles	
	0	1
0	473	55
1	128	198

TABLE 5.5 – Matrice de confusion de la régression logistique pour la base de test

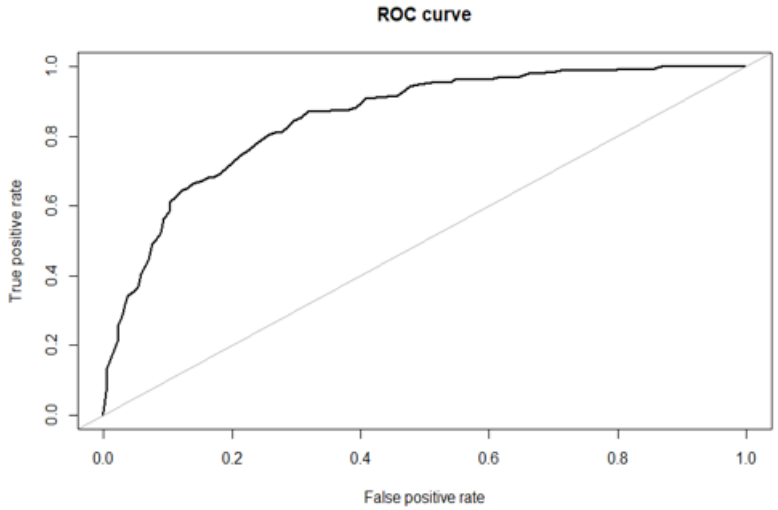

FIGURE 5.10 – Courbe ROC du modèle d'arbre de décision

Avec cette matrice et la courbe ROC, les métriques suivantes peuvent être calculées :

Rappel : 78%

Précision : 61%

F1-score : 69%

AUC : 84%

Interprétation des résultats :

- Le modèle repère 79% de Sécheresse (198 sur 253 Sécheresse réellement observées).
- Sur 100 Sécheresse prédites par le modèle, il y en a seulement 61 qui sont réellement survenues (198 sur 326 Sécheresse réellement observées).

5.3 Arbre de Décision [9] [10]

Toujours dans l'objectif de prédire la variable binaire "Arrêté", le deuxième modèle proposé dans ce mémoire est le modèle d'arbre de décision. La présentation théorique de ce modèle est basée sur (18) et (19).

Un modèle d'arbre de décision est un outil d'aide à la décision qui permet d'expliquer et de prédire des variables quantitatives ou qualitative à partir de variables explicatives. Si la variable à expliquer et prédire est une variable quantitative, l'arbre est dit arbre de régression, sinon l'arbre est dit arbre de classification. Cette méthode statistique a été introduite par Breiman *et al.* (1984)

5.3.1 Principe Général

D'après (18), le principe d'un arbre de décision est de créer des classes d'individus homogènes pour une base de données. Ces classes sont élaborées grâce à des règles binaires construites à partir des variables explicatives. L'objectif est que ces classes soient les plus homogènes possibles.
Un arbre de décision est un arbre à l'envers. La racine d'un arbre de décision est en haut. Le modèle commence par considérer l'ensemble des individus de la base de données avant de les séparer. Chaque embranchement est appelé *nœud*. Il existe deux types de nœuds, les nœuds sans descendant qui sont appelés les *feuilles* et les nœuds non terminaux qui ont des nœuds fils. Chaque nœud non terminal est

9. P-A. CORNILLON et al. Statistiques avec R : 2ème version augmentée. Édition PUR. 2010. (18)
10. R. GENUEUR et J-M. POGGI. Les forêts aléatoires avec R. Édition PUR. 2019. (19)

représenté par une condition sur une variable explicative. La réponse à cette condition va permettre de créer d'autres nœuds jusqu'à arriver à une feuille (nœud sans descendant). Chaque feuille est étiquetée par une modalité de la variable réponse. Pour classifier un individu, il suffit de parcourir l'arbre en répondant chaque fois à la question posée lors des nœuds jusqu'à arriver à une feuille.

Par exemple, un arbre de décision commence par tous les individus et pose une première question : "Le taux de destruction survenu dans le code postal est-il supérieur à 0.5 ?" Suite à cela, deux branches vont être créées. Dans chaque branche, il y aura deux proportions d'individus ayant chacune une modalité pour la variable réponse. Lorsque suite à une coupure, les individus qui composent une branche ont toutes la même modalité, une feuille est créée. Par exemple, à la suite de 3 questions posées, les individus obtenus ont tous subi une Sécheresse (1 pour la variable réponse), ces individus constituent donc une feuille. La coupure est aussi appelée découpe ou "split".
Cet exemple est illustré par le graphe ci-dessous.

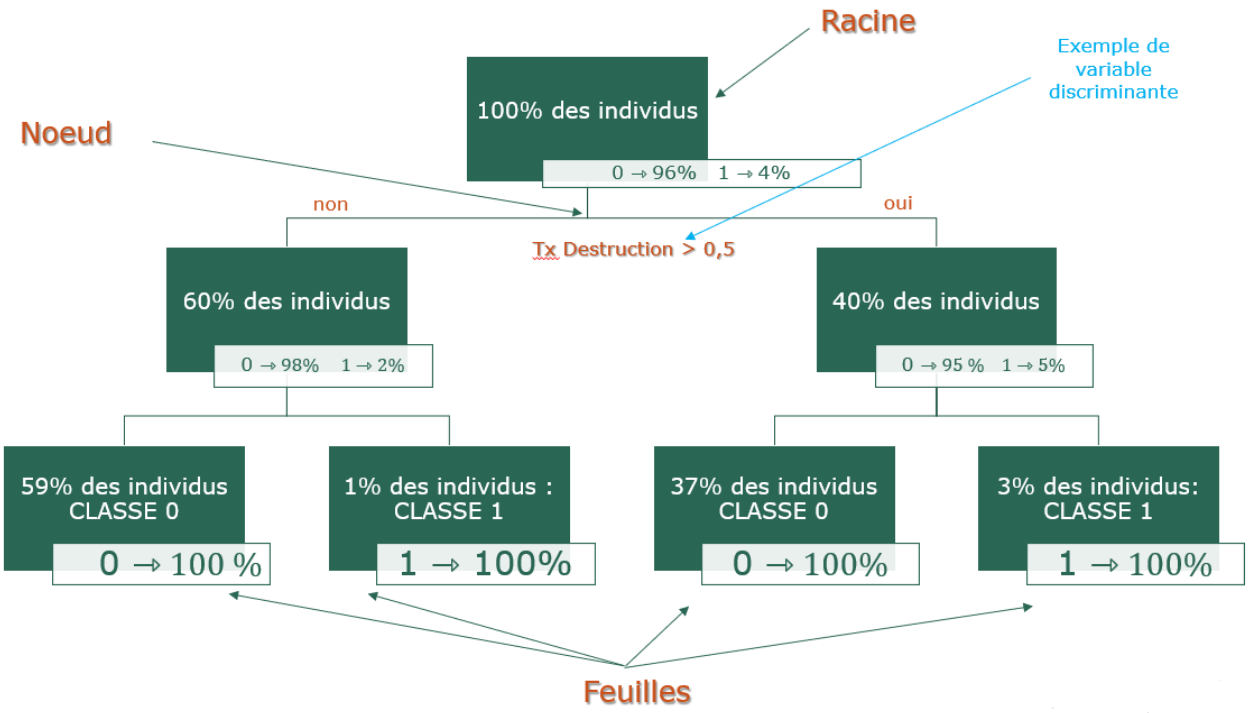

FIGURE 5.11 – Illustration d'un arbre de décision

Dans la suite, les notations suivantes représentent :

— p le nombre de variables explicatives du modèle

— L_n : l'échantillon d'apprentissage

— X_j la j-ème variable descriptive, $j \in \{1, ..., p\}$

— Un Split (ou découpe) : $\{X_j \leq d\} \cup \{X_j > d\}$ avec $d \in \mathbb{R}$

Toujours selon (18), un arbre de décision commence par découper la racine en deux nœuds fils. Ensuite, ce découpage est réitéré récursivement. Découper selon $\{X_j \leq d\} \cup \{X_j > d\}$ implique que tous les individus dont la valeur de la j-ème est supérieure à d vont dans le nœud fils de droite et les autres vont dans le nœud fils de gauche. Or, le modèle d'arbre de décision a pour objectif d'avoir un découpage homogène. Le modèle recherche alors la meilleure découpe possible. Ceci est obtenue en choisissant le couple (j, d) qui minimise l'impureté des nœuds fils. Cette impureté est mesurée par l'indice de Gini.

La définition de l'indice de Gini est proposée par (19) :

L'indice de Gini d'un nœud t est défini par

$$\phi(t) = \sum_{c=1}^{C} \hat{p_t^c}(1 - \hat{p_t^c})$$

avec $\{1, ..., C\}$ *l'ensemble des classes pour la variable réponse (*$\{0, 1\}$ *si classification binaire) et* $\hat{p_t^c}$ *la proportion d'individus de classes c dans le nœud t*
L'objectif est de diminuer la fonction de pureté de Gini, ce qui va permettre d'augmenter l'homogénéité des nœuds.

Arbre maximal

Après le découpage de la racine, la procédure se répète suivant le même principe pour les nœuds fils jusqu'à atteindre une condition d'arrêt. Quand toutes les découpes possibles ont été réalisées, l'arbre obtenu est l'arbre maximal. Un arbre maximal, noté T_{max}, est l'arbre pleinement développé, contenant le maximum de branches possibles. Si le nombre de variables explicatives est assez élevé (le cas échéant), l'arbre maximal devient illisible comme peut le montrer cette figure ci-dessous.

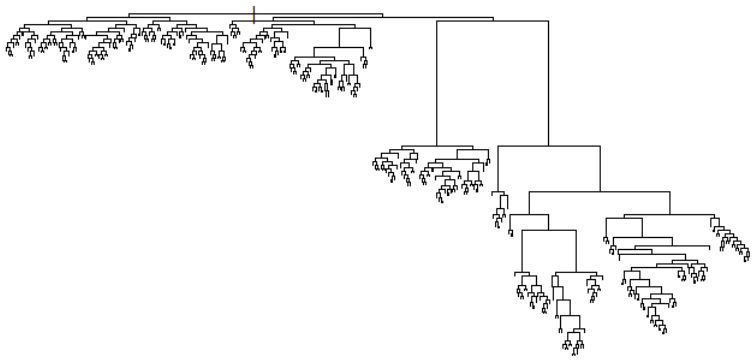

FIGURE 5.12 – L'arbre maximal associé à la problématique

Arbre optimal

L'arbre maximal est un arbre qui possède une très grande variance mais un biais faible. Si un arbre de décision est trop profond, il y a un risque de sur-apprentissage (bien prédire la base d'apprentissage mais avoir des prédictions mauvaises pour les autres bases). L'arbre optimale est un sous-arbre entre l'arbre maximal et l'arbre composé seulement de la racine, qui minimise l'erreur de classification.

Validation Croisée

Pour obtenir un arbre optimal, une procédure appelée "Validation Croisée" [11] existe. Les étapes de la validation croisée sont proposées par (19) :

1. Diviser aléatoirement la base d'apprentissage en V sous échantillons (par exemple $V = 10$),
2. Pour chaque sous échantillon $v = 1, ...V$, un arbre est construit et appliqué pour classifier les autres sous échantillons,

11. V-fold cross-validation

3. Calculer l'erreur obtenue,
4. Répéter l'opération 2 et 3 pour les autres sous échantillons,
5. Choisir l'arbre qui minimise l'erreur.

5.3.2 Application à la problématique

Ce modèle va être appliqué à la même base de données sur laquelle a été appliquée le modèle de régression logistique, avec toujours la même variable à prédire "Arrêté" selon les 17 variables explicatives. La base de données a été divisée en une base d'apprentissage sur laquelle le modèle va apprendre et une base de test.

Construction de l'arbre par défaut

La base de données est constituée de 17 variables explicatives, ce nombre est considéré élevé. Il est donc impossible de lire l'arbre maximal. Dans un premier temps, l'arbre de décision par défaut, c'est à dire obtenu sans l'optimisation d'aucun paramètre sera considéré et analysé.
La figure ci-dessous illustre l'arbre par défaut obtenu grâce au logiciel R.

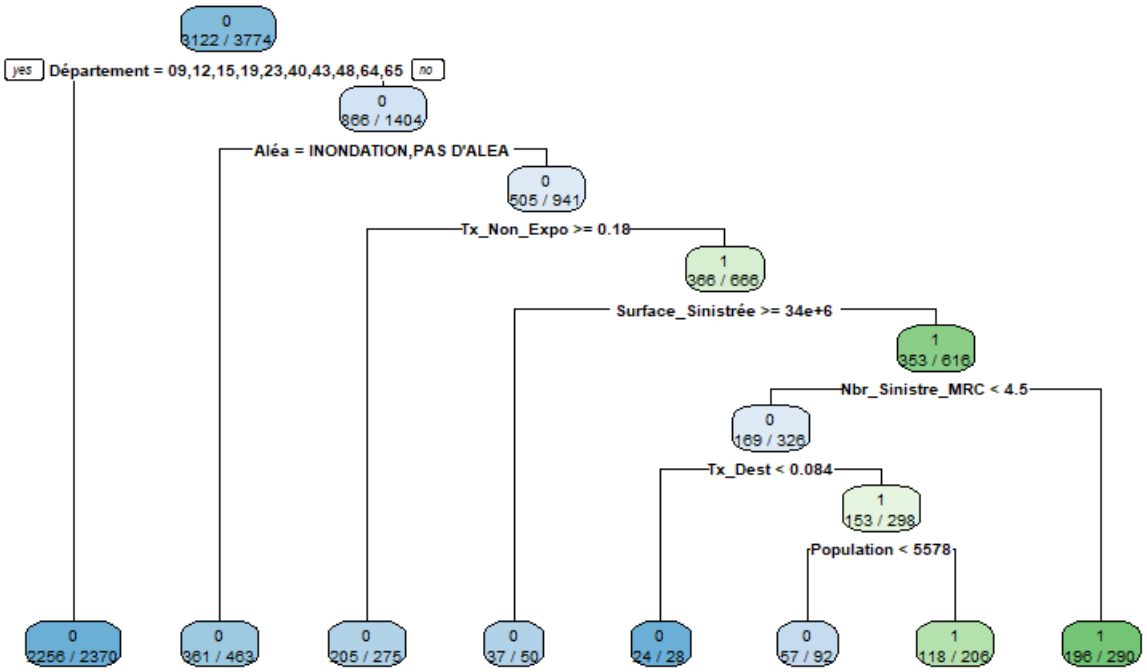

FIGURE 5.13 – Arbre par défaut obtenu

Dans le cas du premier nœud, le modèle a divisé selon le département du code postal. Comme il a été constaté lors de l'analyse descriptive de la base, la répartition des arrêtés n'est pas équilibrée entre les départements. 3122 sur 3774 codes postaux qui appartiennent aux départements 09, 12, 15, 19, 23, 40, 43, 48, 54, et 65, n'ont pas subi une Sécheresse.

Le deuxième nœud s'intéresse à la variable "Aléa". Avec le modèle de régression logistique, il a été constaté que les modalités "Inondation" et "Pas d'Aléa" (absence de sinistres MRC dans le code postal), diminuent la probabilité qu'il y ait un arrêté dans le code postal. Ceci vient se confirmer avec cet arbre. Cet arbre indique aussi que lorsque pour un code postal, le taux d'exposition nulle face aux phénomène RGA est élevé, il est peu probable qu'un arrêté soit publié. Enfin toujours d'après

cet arbre, si la variable nombre de sinistre MRC est supérieure à 4,5 par an, il est probable qu'une Sécheresse surviendra dans le code postal.

Construction de l'arbre optimal

Pour améliorer l'arbre obtenu, la méthode de validation croisée sera réalisée afin d'obtenir l'arbre optimal.
La figure ci-dessous trace l'évolution de l'erreur relative en fonction de la taille de l'arbre lors de chaque itération de la validation croisée.

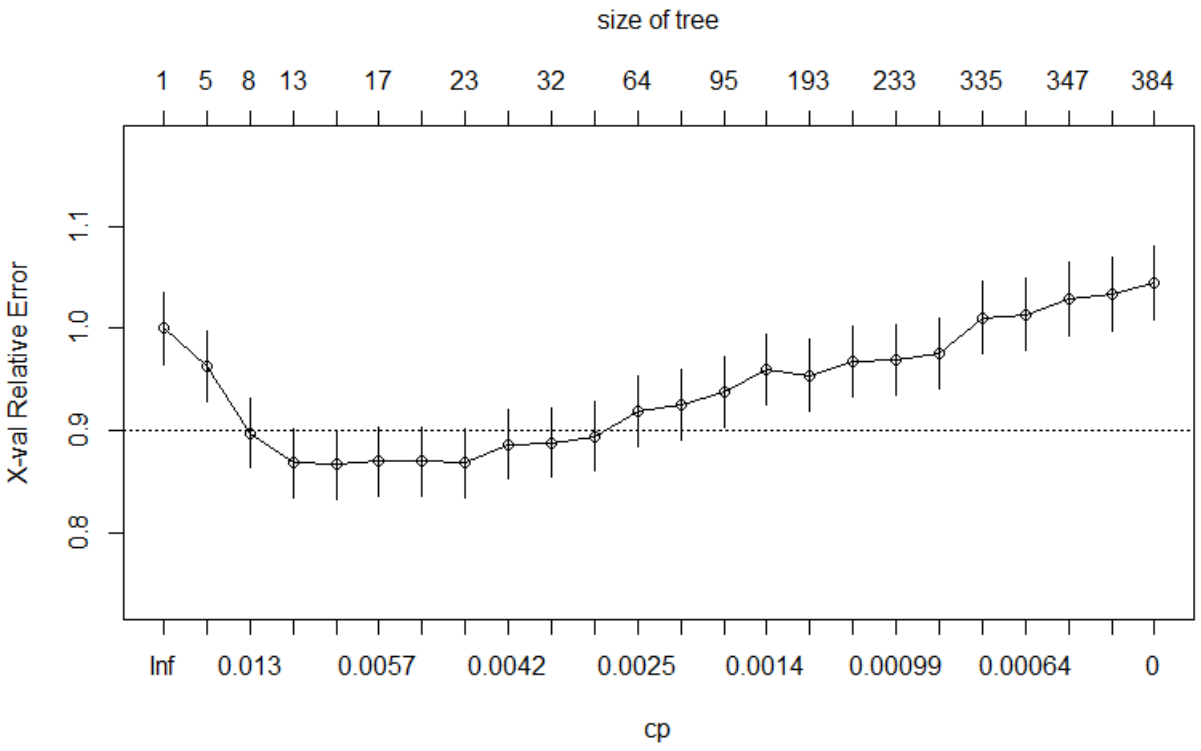

FIGURE 5.14 – Taille de l'arbre en fonction de l'erreur relative

La figure ci-dessous indique que l'erreur relative diminue avant d'augmenter à partir d'un nouveau seuil. A partir de cette courbe, la taille de l'arbre optimal est de 15 et la complexité (notée cp sur le graphique) optimale est de 0.016.
Ces paramètres vont être appliqués au modèle pour obtenir l'arbre optimal qui est illustré par le graphe ci-dessous.

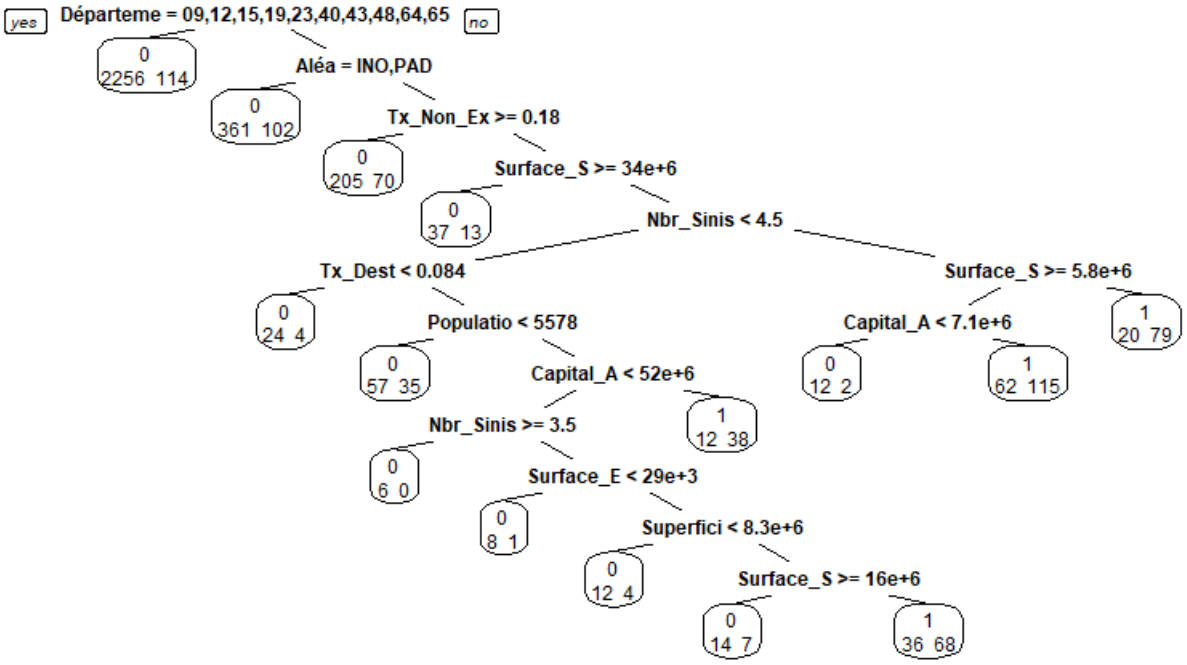

FIGURE 5.15 – Arbre optimal obtenu

L'arbre optimal est plus compact que l'arbre par défaut et fait intervenir de nouvelles variables comme la superficie du code postal ou bien le capital assuré.

Importance des variables

Pour avoir une idée sur l'importance des variables explicatives dans la base, le graphe ci-dessous illustre chaque variable et son importance dans l'arbre optimal.

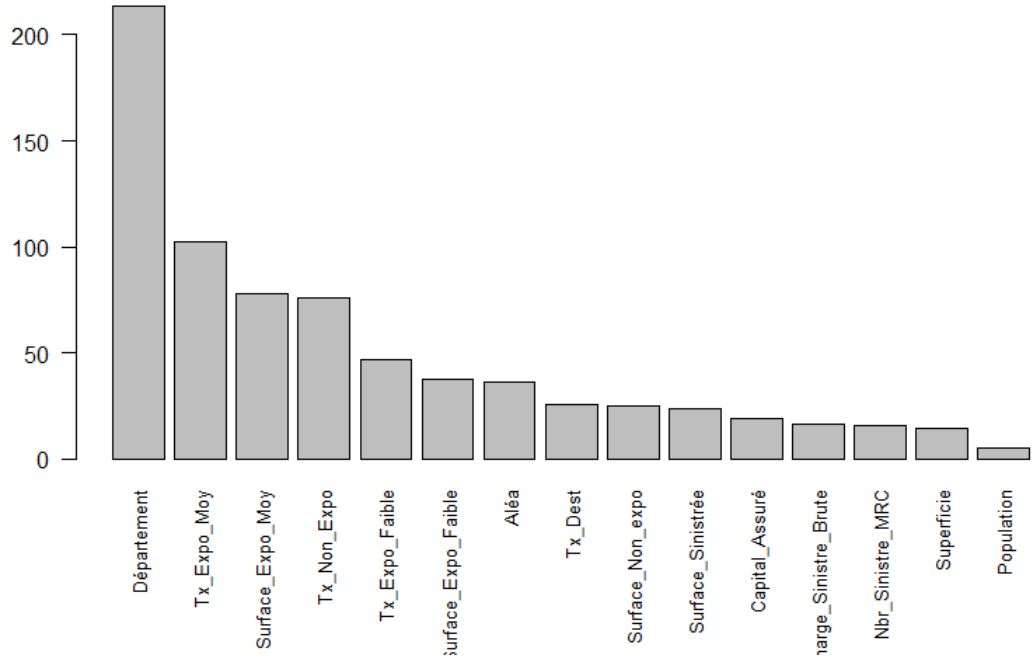

FIGURE 5.16 – Importance des variables pour l'arbre optimal

Dans ce modèle, le département est la variable la plus importante, suivie des variables d'exposition du sol face au phénomène de RGA, l'aléa et du taux de destruction.

5.3.3 Résultats du modèle

Une fois l'arbre optimal créé et calibré, il faut le tester sur la base de test. La matrice de confusion ainsi que la courbe ROC obtenue sont représentées ci-dessous :

Classes prédites	Classes réelles	
	0	1
0	473	71
1	128	182

TABLE 5.6 – Matrice de confusion de l'arbre de décision pour la base de test

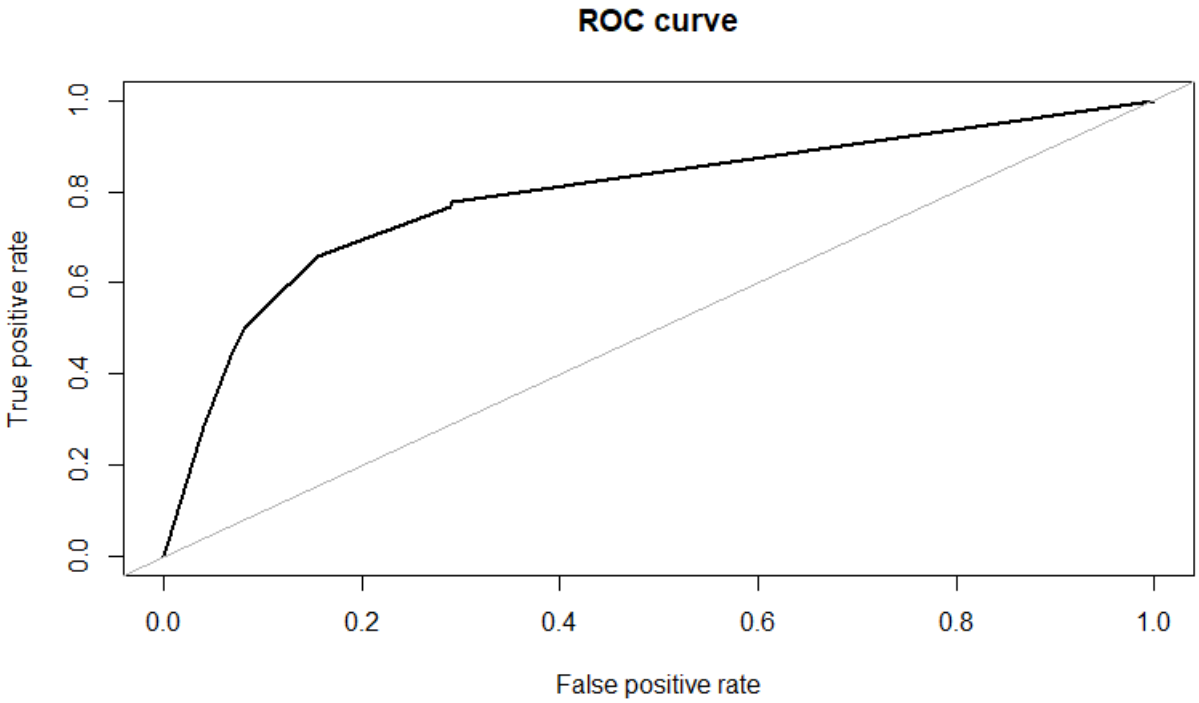

FIGURE 5.17 – Courbe ROC du modèle d'arbre de décision

Avec cette matrice et la courbe ROC, les métriques suivantes peuvent être calculées :

Rappel : 72%

Précision : 58%

F1-score : 65%

AUC : 79%

Interprétation des résultats :

— Le modèle repère 72% de Sécheresse.

— Sur 100 Sécheresse prédites par le modèle, il y en a seulement 58 qui sont réellement survenues.

5.4 Forêts aléatoires [12] [13]

Le troisième et dernier modèle qui sera appliqué à la base de données pour prédire la variable "Arrêté" est le modèle de forêts aléatoires.
Cette section s'appuie sur la théorie développée dans (19),(20).

5.4.1 Principe Général

Un modèle de forêts aléatoires est un outil d'aide à la décision qui utilise les arbres de décisions, qui a pour objectif de prédire une variable réponse selon des variables explicatives. Cette méthode a été introduite par Breiman et al (2001).

Selon (19), une forêt aléatoire est l'agrégation d'un certain nombre d'arbres de décisions aléatoires. L'idée consiste à réaliser plusieurs arbres de décisions indépendants en choisissant aléatoirement à chaque fois un nombre de variables explicatives et d'agréger les résultats obtenus pour arriver à une prédiction finale. En classification, la prédiction finale est la classe qui est la plus présentée parmi les arbres de décisions.
Pour obtenir une forêt aléatoire performante, les arbres de décisions doivent être diversifiés donc ils ne doivent pas tous apprendre sur la même base d'apprentissage. C'est pour cela que le modèle de forêts aléatoires s'appuie sur les échantillons "Bootstrap".

La définition d'un échantillon "Bootstrap" est alors proposée par (19) :

Définition : Échantillon Bootstrap

Un échantillon bootstrap d'un échantillon d'apprentissage L_n de taille n est obtenu en tirant aléatoirement n observations parmi les observations de L_n avec remise, chaque observation (X_i, Y_i) de L_n ayant une probabilité $1/n$ d'être sélectionné à chaque tirage.

Pour construire une forêt aléatoire il faut alors suivre ces étapes :

1. Prendre un échantillon bootstrap de la base d'apprentissage,
2. Choisir un nombre de variables explicatives à utiliser,
3. Entraîner un modèle d'arbre de décision sur l'échantillon bootstrap,
4. Répéter les étapes un certain nombre de fois (nombre d'arbres dans la forêt).

Pour classifier un nouvel individu, il suffit de parcourir tous les arbres et à chaque fois et enregistrer la classe obtenue. La classe retenue sera la classe majoritaire dans les prédictions des arbres.

Echantillon Out Of Bag (OOB)

L'échantillon Out Of Bag (OOB) représente l'échantillon en dehors du bootstrap, c'est à dire tous les individus qui appartiennent à la base d'apprentissage mais qui n'ont pas été tirées dans l'échantillon Bootstrap.
Cet échantillon peut jouer le rôle d'une base de test. En effet, lors de l'application du modèle, chaque arbre apprend sur une base Bootstrap, donc le modèle n'a pas d'information concernant les individus de l'échantillon OOB.
L'Erreur Out Of Bag (Erreur OOB) est l'erreur de prédiction de la forêt aléatoire sur l'échantillon Out Of Bag.

12. R. GENUER et J-M. POGGI. Les forêts aléatoires avec R. Édition PUR. 2019. (19)
13. B. BOEHMKE et B. GREENWELL. Hands-On Machine Learning with R. 2020. (20)

Paramètres à optimiser

Pour avoir un modèle de forêt aléatoire robuste, efficace et qui présente de bonnes prédictions, il faut optimiser un certain nombre de paramètres.
Les paramètres à optimiser proposés par (20) sont :

1. Le nombre d'arbres (ntree dans R).
2. Le nombre de variables explicatives choisies aléatoirement dans la construction des arbres (mtry dans R).
3. La complexité des arbres, c'est à dire, contrôler les tailles des nœuds (node size) ou bien la profondeur maximale des arbres (max nodes).

Ces paramètres doivent être choisis de façon à minimiser l'erreur OOB renvoyée par le modèle ou bien maximiser les métriques (par exemple F1 Score) sur le même échantillon.

5.4.2 Application à la problématique

Le modèle de forêts aléatoires sera appliqué à la même base de données que pour les modèles de régression logistique et d'arbre de décision. La variable à prédire est toujours la variable "Arrêté"

Optimisation du nombre d'arbres

Pour optimiser le nombre d'arbres dans la forêt, il faut simuler un grand nombre de forêts avec différents nombres d'arbres et noter à chaque fois le F1 Score obtenu sur l'échantillon Out Of Bag. Le nombre d'arbres optimal est celui qui maximise le F1 score.
La figure ci-dessous illustre l'évolution du F1 score selon le nombre d'arbres dans les forêts.

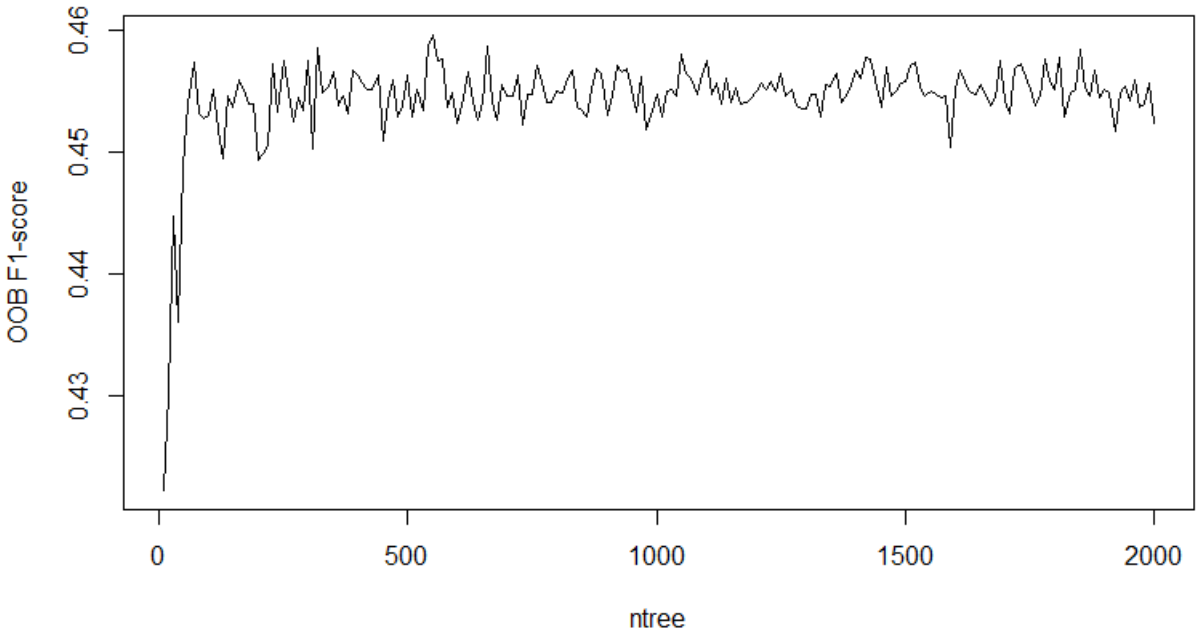

FIGURE 5.18 – Optimisation du nombre d'arbres dans la forêt aléatoire

Optimiser la complexité du modèle permet de diminuer le temps de calcul. D'après la figure, le F1 score se stabilise à partir de 200 arbres dans la forêt.

Optimisation du nombre de variables explicatives utilisées

Après l'optimisation du nombre d'arbres dans le modèle, il faut optimiser le nombre de variables explicatives choisies pour construire les arbres. La même procédure sera appliquée mais en variant, cette fois-ci, ce nouveau paramètre.

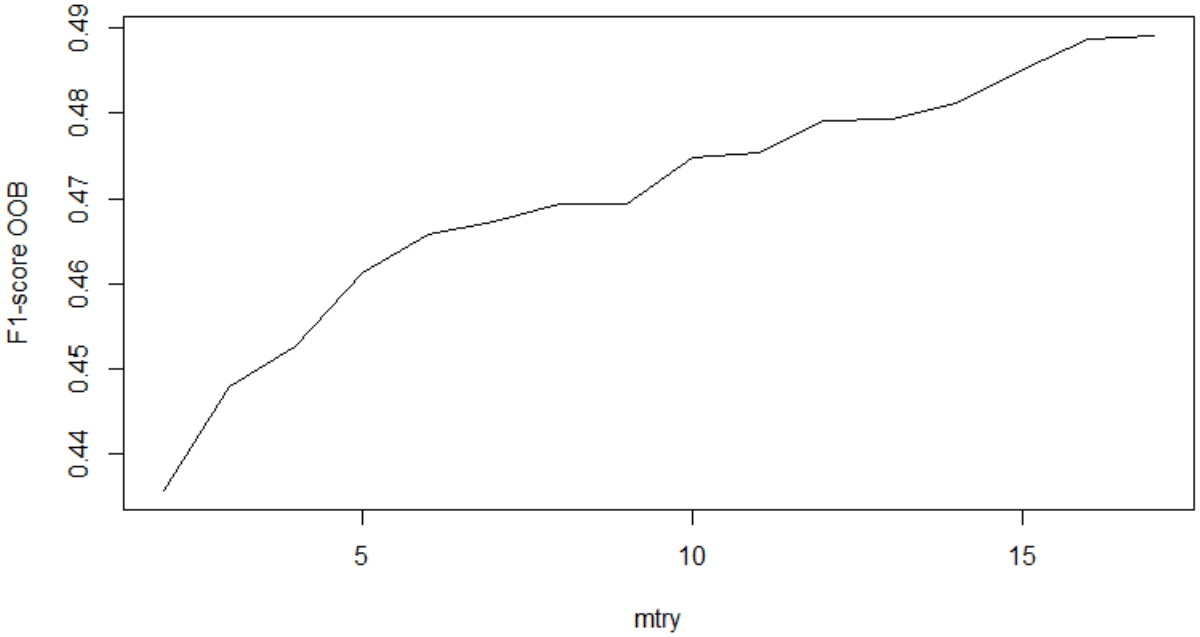

FIGURE 5.19 – Optimisation du nombre de variables explicatives utilisés lors de chaque nœud

La figure ci-dessus montre que plus le paramètre augmente, plus le F1 score augmente. C'est à dire, le modèle prédit mieux quand il prend à chaque fois toutes les variables explicatives pour construire les arbres.

5.4.3 Résultats du modèle

Une fois le modèle calibré et optimisé, il faut l'appliquer sur la base de test. La matrice de confusion et la courbe ROC (figure ci-dessous) permettent d'illustrer les résultats du modèle ainsi que de calculer les métriques.

Classes prédites	Classes réelles	
	0	1
0	448	57
1	153	196

TABLE 5.7 – Matrice de confusion du modèle de forêt aléatoire pour la base de test

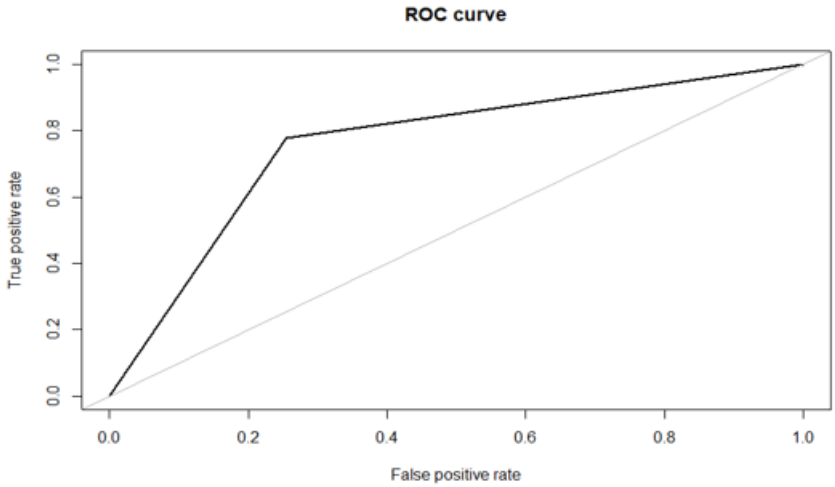

FIGURE 5.20 – Courbe ROC du modèle de forêt aléatoire

Rappel : 78%

Précision : 56%

F1-score : 65%

AUC : 76.2%

Interprétation des résultats :

— Le modèle repère 78% de Sécheresse.

— Sur 100 sécheresses prédites par le modèle, il y en a seulement 56 qui sont réellement survenues.

5.5 Comparaison des résultats et choix du meilleur modèle

Pour choisir le meilleur modèle entre les trois modèles utilisés, il est proposé de synthétiser les résultats le tableau ci-dessous :

Modèle	Rappel	Précision	F1-score	AUC
Régression Logistique	78%	61%	69%	84%
Arbre de décision	72%	58%	65%	79%
Forêt aléatoire	78%	56%	65%	76%

TABLE 5.8 – Tableau récapitulatif des métriques obtenues.

Le modèle de régression logistique affiche les meilleurs résultats. Ce modèle est plus précis que les deux autres modèles et repère aussi bien les Sécheresse que le modèle de forêt aléatoire. Il a logiquement le meilleur F1 Score. Le modèle de forêt aléatoire donne une assez grande importance à la classe 1, en repérant les Sécheresse, mais en manquant de précision par rapport aux autres modèles. Le modèle d'arbres de décision quant à lui se rapproche de la précision du modèle de régression logistique mais ne repère pas autant de Sécheresse que ce dernier.
En plus de ces éléments, l'AUC du modèle de régression logistique est la plus élevée. Il a été donc décidé de choisir le modèle de régression logistique pour la suite des travaux.

5.6 Mise en application du modèle sur les années antérieures

Afin de mieux tester le modèle de régression logistique, un backtesting [14] est réalisée sur les années antérieures entre 2011 et 2019.
La figure ci-dessous présente les résultats obtenus :

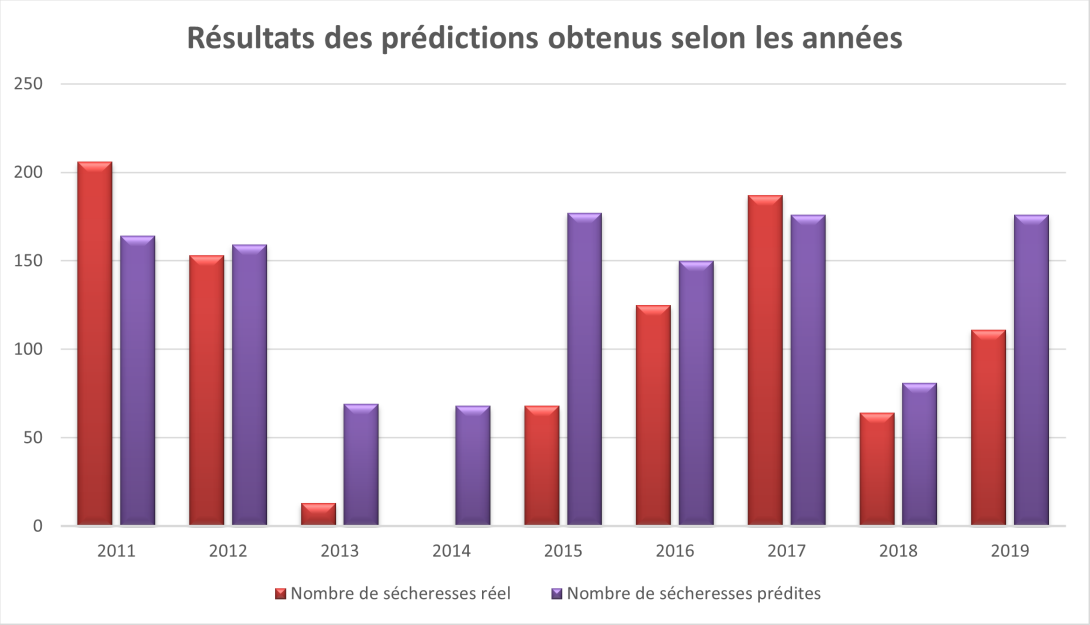

FIGURE 5.21 – Résultats des prédictions du modèle par année

Constats :

— Les années 2012, 2016 et 2018 sont bien prédites par le modèle.

— Les années 2013 et 2014 sont deux années atypiques. Le modèle prédit moins de Sécheresse mais ces prédictions restent assez élevées par rapport au nombre de Sécheresse réel.

— Le modèle surestime le nombre de Sécheresse au cours des années 2015 et 2019.

Précision importante : Un seul seuil a été défini pour l'ensemble des années. C'est peut-être là un biais (qui est confirmé par la suite de l'étude).

Conclusion

Dans le cadre de ce chapitre, trois modèles statistiques ont été proposés et élaborés.
Tout d'abord un modèle de régression logistique a été calibré et optimisé. Ce modèle présente un F1 score de 69% et un AUC de 84%.

En second lieu, un modèle d'arbre de décision et de forêts aléatoires ont aussi été calibrés et optimisés. Néanmoins, les résultats renvoyés par ces derniers étaient inférieurs aux résultats du premier modèle.
Enfin, le "backtesting" réalisé montre que les prédictions du modèle s'approchent de la réalité avec toutefois des écarts sur certains exercices.

La prochaine étape consistera à estimer le coût moyen de ces évènements prédits pour obtenir une charge sinistres de la Sécheresse pour le portefeuille.

14. Mise en application du modèle sur les années qui constituent la base

Chapitre 6

Estimation de la charge sinistres de la Sécheresse

Introduction

Après la modélisation de la fréquence des évènements Sécheresse sur le territoire de GOC et dans le but d'estimer une charge sinistres relative à ce péril, l'objectif de ce chapitre est :

— Tout d'abord, de calculer un coût moyen pour ces évènements selon différentes méthodes.

— Ensuite, d'appliquer ce coût aux évènements prédits par le modèle.

— Enfin, de "backtester" la méthode en comparant la charge de sinistre estimée aux valeurs réelles.

Dans la suite, les notations suivantes représentent :

— $CM_{Sécheresse}$: Coût moyen d'une Sécheresse sur bâtiments

— CP : Code postal du territoire de GOC.

— Dept : Département du territoire de GOC.

— CTP : Charge Totale Probable (CTP = Règlements + Provisions - Recours - Prévisions de recours)

— S/C : Quotient des montants des sinistres payés (plus précisément du CTP) et les cotisations perçues par GOC.

— $P(CP = 1)$: Probabilité, renvoyée par le modèle, de la publication d'au moins un arrêté CatNat Sécheresse dans le code postal.

6.1 Estimation des charges sinistres de la Sécheresse entre 2012 et 2019

6.1.1 Méthode 1 : Coût moyen global

La première méthode consiste à estimer un coût moyen pour un arrêté CatNat Sécheresse.
La formule ci-dessous présente cette méthode :

$$CM_{Sécheresse} = \frac{CTP_{sécheresse}}{Nombres\_total\_arreté\_sécheresse}$$

Le coût moyen d'un arrêté Sécheresse pour le territoire de GOC, selon cette méthode, est estimé à **105 000 €**.

Estimation du CTP Sécheresse

Ce coût moyen calculé est appliqué aux fréquences prédites par le modèle sur les années antérieurs (de 2012 à 2019) tel que :

$$CTP_{prédit} = CM_{sécheresse} * Nombre\_Totale\_Sécheresses\_Prédites$$

avec : $Nombres\_Total\_Sécheresse\_Prédites = \sum_{CP} \mathbb{1}_{P(CP=1>Seuil)}$
Le graphe ci-dessous présente l'estimation de la charge prédite par cette méthode comparée à la charge réelle.

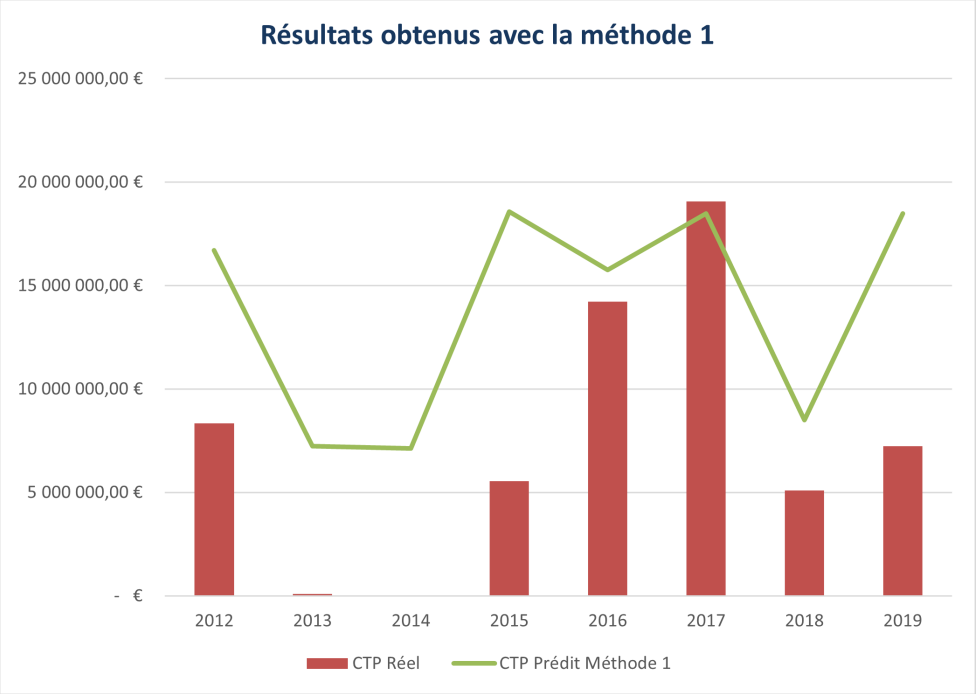

FIGURE 6.1 – Charge sinistres de la Sécheresse avec un coût moyen globale

Constat :

— Les années 2013 et 2014 représentent deux années atypiques avec une charge sinistres Sécheresse réelle presque nulle. En utilisant cette méthode, le modèle surestime le CTP pour ces deux années.

— La charge prédite pour les années 2016 et 2017 s'approche de la charge réelle.

— En 2012, 2015 et 2019, la charge estimée est très supérieure à la charge réelle.

Problématique :
Cette méthode ne tient pas compte d'une valeur de bien différente selon la localisation.

6.1.2 Méthode 2 : Coût moyen selon le département

La deuxième méthode consiste à estimer, cette fois, un coût moyen qui dépend du département. La formule du coût moyen s'écrit alors :

$$CM_{Sécheresse}(Dept) = \frac{CTP_{sécheresse,Dept}}{Nombres\_total\_arreté\_dans\_le\_dept}$$

Le tableau ci-dessous illustre les coûts moyens obtenus par département.

Département	Coût Moyen
09	59 000 €
12	90 000 €
15	70 000 €
19	47 000 €
23	10 000 €
31	57 000 €
32	159 000 €
40	50 000 €
46	110 000 €
48	40 000 €
64	78 500 €
65	100 000 €
81	115 000 €
82	86 000 €

TABLE 6.1 – Coûts moyens par département

Estimation du CTP Sécheresse

De la même façon que pour la première méthode, ce coût moyen est appliqué aux évènements prédits dans les différents départements.
La charge sinistres est alors calculée avec cette formule :

$$CTP_{prédit} = \sum_{Dept} CM_{sécheresse,CP}(Dept) * Nombre\_Sécheresses\_Prédites\_par\_dept$$

avec : $Nombre\_Sécheresses\_Prédites\_par\_dept = \sum_{CP \in Dept} \mathbb{1}_{P(CP=1>Seuil)}$

La figure 6.2 présente les résultats obtenus en utilisant cette méthode.

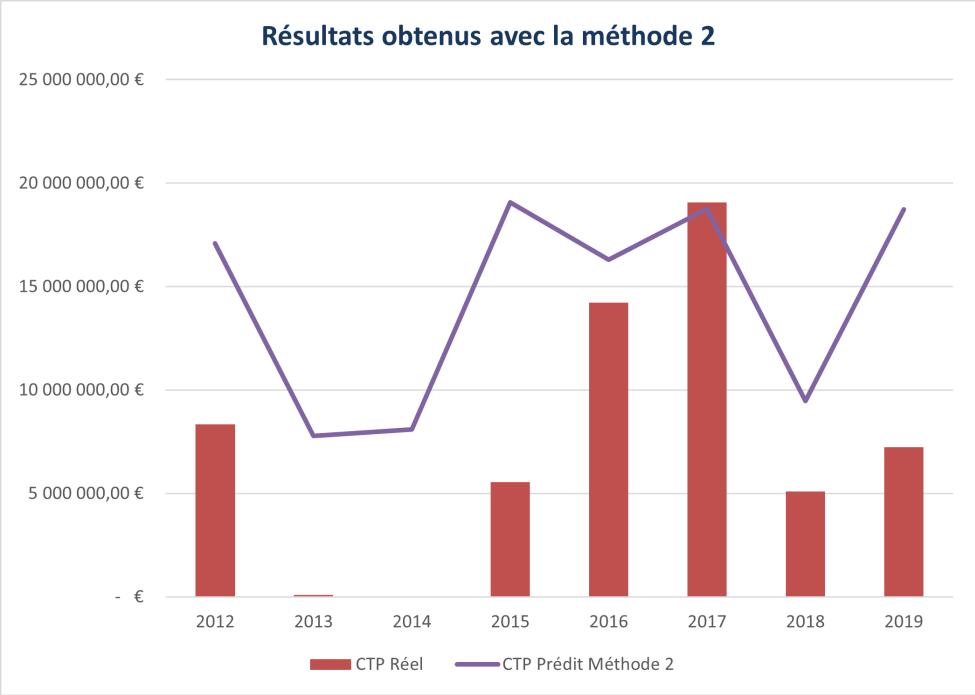

FIGURE 6.2 – Charge sinistres de la Sécheresse avec un coût moyen qui dépend du département

Constats :

— Cette méthode met en évidence la différence des coûts entre les départements.

— Le courbe du CTP prédit avec cette méthode suit la même allure que la courbe du CTP prédit avec la première méthode.

— Le nombre d'évènements Sécheresse dans certains départements est parfois faible rendant l'estimation peu juste. Par exemple, le département 48 a connu seulement une Sécheresse durant toute la période d'observation.

Problématique :

Les méthodes 1 et 2 ne tiennent pas compte du nombre d'habitation, plus précisément, du taux de pénétration de GOC selon le territoire.

6.1.3 Méthode 3 : Coût moyen par entité assurée

Cette méthode consiste à estimer un coût moyen d'une sécheresse selon le nombre d'entités assurées [1] dans le code postal. Attention, il s'agit ici d'un coût moyen calculé sur l'ensemble du territoire de GOC.

La formule du coût moyen s'écrit alors :

$$CM_{sécheresse,EA} = \frac{CTP_{scheresse}}{Nombre\_total\_EA}$$

Estimation du CTP Sécheresse

La charge sinistre est alors calculée avec cette formule :

$$CTP_{prédit} = \sum_{CP} \mathbb{1}_{P(CP=1>Seuil)} * Nombre\_de\_EA\_dans\_le\_CP * CM_{Scheresse,EA}$$

Les résultats obtenus sont illustrés par la figure ci-dessous :

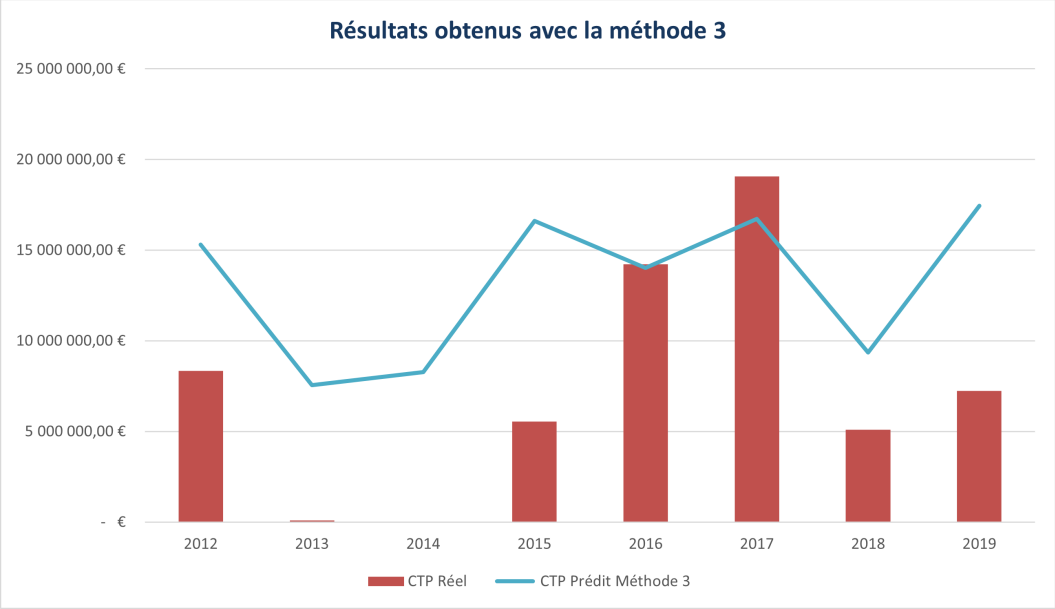

FIGURE 6.3 – Charge sinistres de la Sécheresse avec un coût moyen qui dépend du nombre d'entités assurées

1. Au sein de Groupama, le terme entité assurée est souvent employé. Ce terme peut aussi bien décrire une habitation, une automobile, un homme, une femme, un chien, un chat, ... En l'occurrence, il s'agit ici des habitations.

Constats :

— La charge prédite pour l'année 2016 s'approche de la charge réelle.

— Similairement aux deux premières méthodes, les années 2012, 2015 et 2019 sont surestimés.

Problématique :

Les trois premières méthodes ne tiennent pas compte des valeurs assurées (valeurs des habitations) par GOC selon le code postal. En effet, un bien assuré sur Toulouse n'a pas la même valeur qu'une habitation assurée à Brandonnet (petite commune de 320 habitants).

6.1.4 Méthode 4 : Coût moyen en fonction du S/C du département

Pour faire face aux problématiques des trois premières méthodes, une quatrième approche a été imaginée. Cette approche se base sur :

— la présence de biens assurés par GOC sur le code postal ciblé (nombre de bâtiments par CP),

— la valeur de ces biens transcrite par la cotisation appelée aux assurés,

— un taux/facteur multiplicatif sur la somme des cotisations afférentes à ces biens est par la suite appliqué.

La formule du coût moyen pour cette méthode s'écrit alors :

$$CM_{Sécheresse}(CP) = \frac{S}{C}(Historique, Dept_{CP}) * Cotisations\_CatNat\_du\_CP$$

Cette méthode 4 s'approche par analogie à un modèle de réassurance : Capitaux assurés x Taux de destruction.

Estimation du CTP Sécheresse

Le CTP Sécheresse est alors estimé avec cette formule :

$$CTP_{prédit} = \sum_{CP} P(CP = 1) * CM_{Scheresse}(CP)$$

Les résultats obtenus par cette méthode sont présentés par la figure ci-dessous.

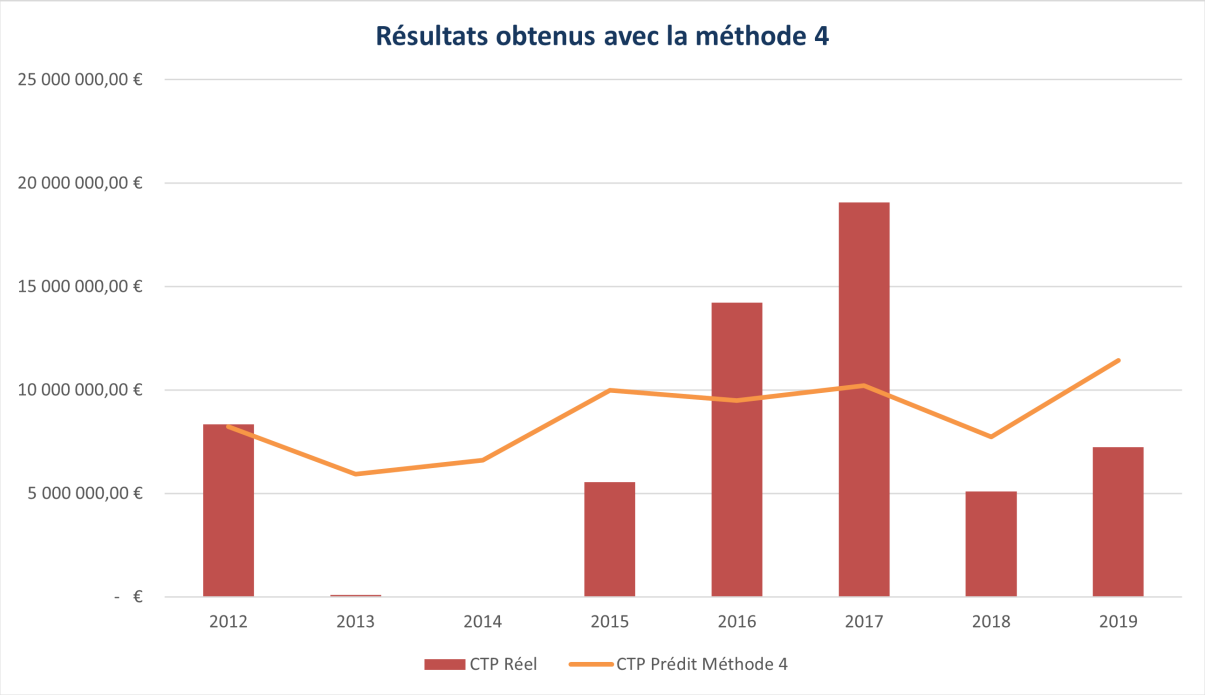

FIGURE 6.4 – Charge sinistres de la Sécheresse avec un coût moyen qui dépend du code postal

Constats :

— L'année 2012 est bien estimée par cette méthode.

— Les années 2013 et 2014 sont moins surestimées que par les autres méthodes.

— Les estimations des années 2015, 2018 et 2019 s'approchent des CTP réels.

— Cette méthode n'estime pas bien les évènements extrêmes. En effet, les années 2016 et 2017 ont été sous-estimées.

Problématique :

Cette méthode semble performante mais elle ne tient pas compte de l'intensité d'un évènement Sécheresse qui peut varier à chaque exercice.

6.1.5 Méthode 5 : Coût moyen qui dépend de l'intensité des Sécheresses

Les résultats obtenus avec les quatre méthodes ne sont pas suffisamment concluants. Deux raisons peuvent expliquer cela :

— Le seuil de prédiction est commun à chaque exercice et les résultats sur certains exercices de survenances montrent un écart entre la valeur prédite et la valeur réelle du nombre d'arrêtés (cf. Figure 5.21).

— Le coût d'un sinistre ne peut pas être, ou du moins l'intensité, ne peut être identique tous les ans.

Une 5ème méthode a donc été définie. Cette méthode consiste à multiplier les cotisations CatNat perçues dans le CP par la probabilité renvoyée par le modèle et un facteur (nommé S/C de destruction). Ce facteur dépend de l'intensité dans le sens où il somme toutes les probabilités, renvoyées par le modèle, de la publication d'au moins un arrêté pour chaque code postal.

$$Somme\_probas = \sum_{CP} P(CP = 1)$$

— si la somme des probabilités est faible, le facteur est accentué à la baisse,

— si la somme des probabilités est forte, le facteur est accentué à la hausse.

Afin de montrer que cette méthode 5 a du sens, le tableau ci-dessous est présenté :

Exercice de survenance	**Somme des probabilités**	**CTP Sécheresse réel**
2012	111.93	8.3 M€
2013	67.84	95 k€
2014	69.05	21 k€
2015	125.13	5.5 M€
2016	109.71	14.2 M€
2017	120.51	19 M€
2018	76.84	5.1 M€
2019	125.23	7.2 M€

TABLE 6.2 – Relation entre la somme des probabilités et le CTP réel

En lecture rapide, une certaine corrélation apparaît. Plus la somme des probabilités est importante, plus le CTP est élevé, avec des niveaux différents certes mais une tendance se dégage malgré tout.
Ensuite, cette non totale linéarité est à mettre en parallèle avec la localisation des arrêtés CatNat Sécheresse (cf. les différentes cartographies).
C'est donc assez naturellement que l'idée d'utiliser une fonction polynomiale est venue et pourrait permettre de répondre à la problématique.

Des équations de 2 à 6 degrés ont été testées et calculées à l'aide de la méthode des moindres carrées (régression). Finalement, une simple équation de second degré est suffisante :

$$\frac{S}{C}de\_destruction = -0.001151 * (Somme\_probas)^2 + 0.226792 * Somme\_probas - 9.962918$$

Ainsi, le CTP Sécheresse est estimée par la formule suivante :

$$CTP_{prédit} = \sum_{CP} P(CP = 1) * Cotisation\_CatNat\_CP * \frac{S}{C}de\_destruction$$

Résultats obtenues

Le graphe ci-dessous présente les résultats du CTP prédit avec cette méthode.

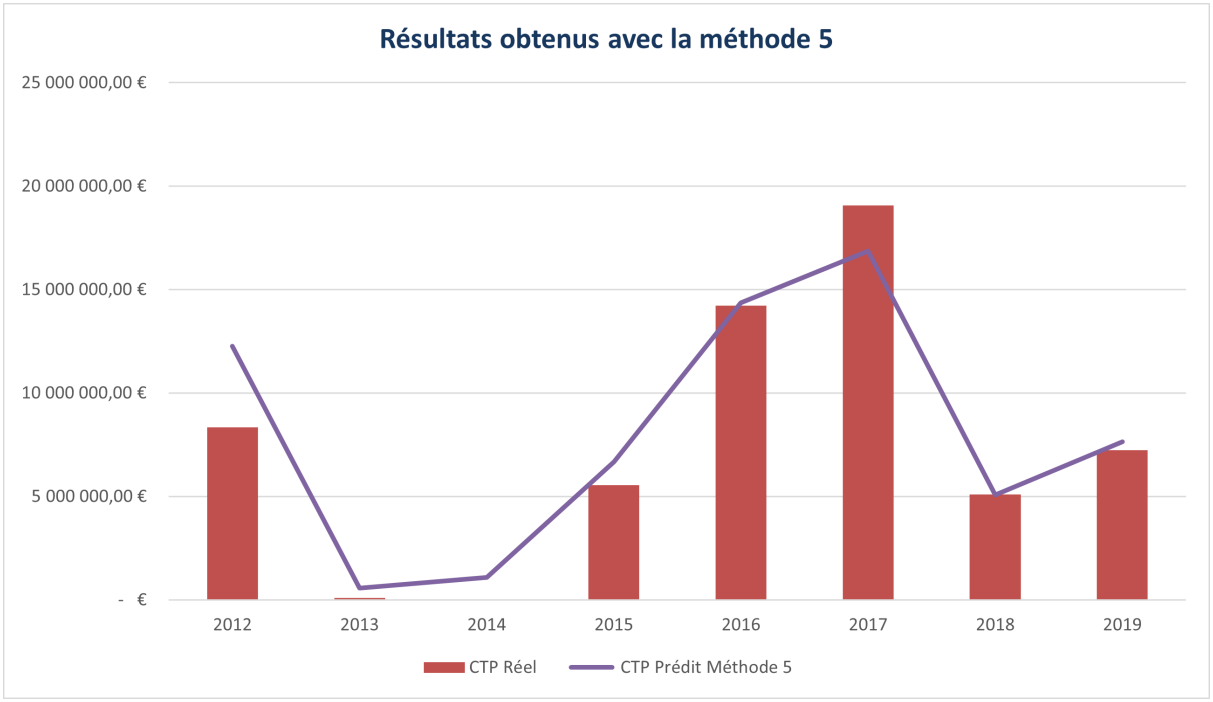

FIGURE 6.5 – Charge sinistres de la Sécheresse avec un coût moyen qui dépend de l'intensité

Constats :

— Cette méthode ne surestime pas les années 2013 et 2014.

— Les années 2013, 2014, 2015, 2016, 2017 et 2018 sont bien définies.

— Néanmoins, l'année 2017 (forte en Sécheresse) est sous-estimée.

6.1.6 Comparaison des résultats des différentes méthodes

Ces différentes méthodes exposées, le graphe ci-dessous permet de comparer les résultats obtenus. Par souci de clarté, la courbe de la deuxième méthode n'est pas affichée sur le graphique parce qu'elle suit presque la même allure que la courbe de la première méthode (CM global).

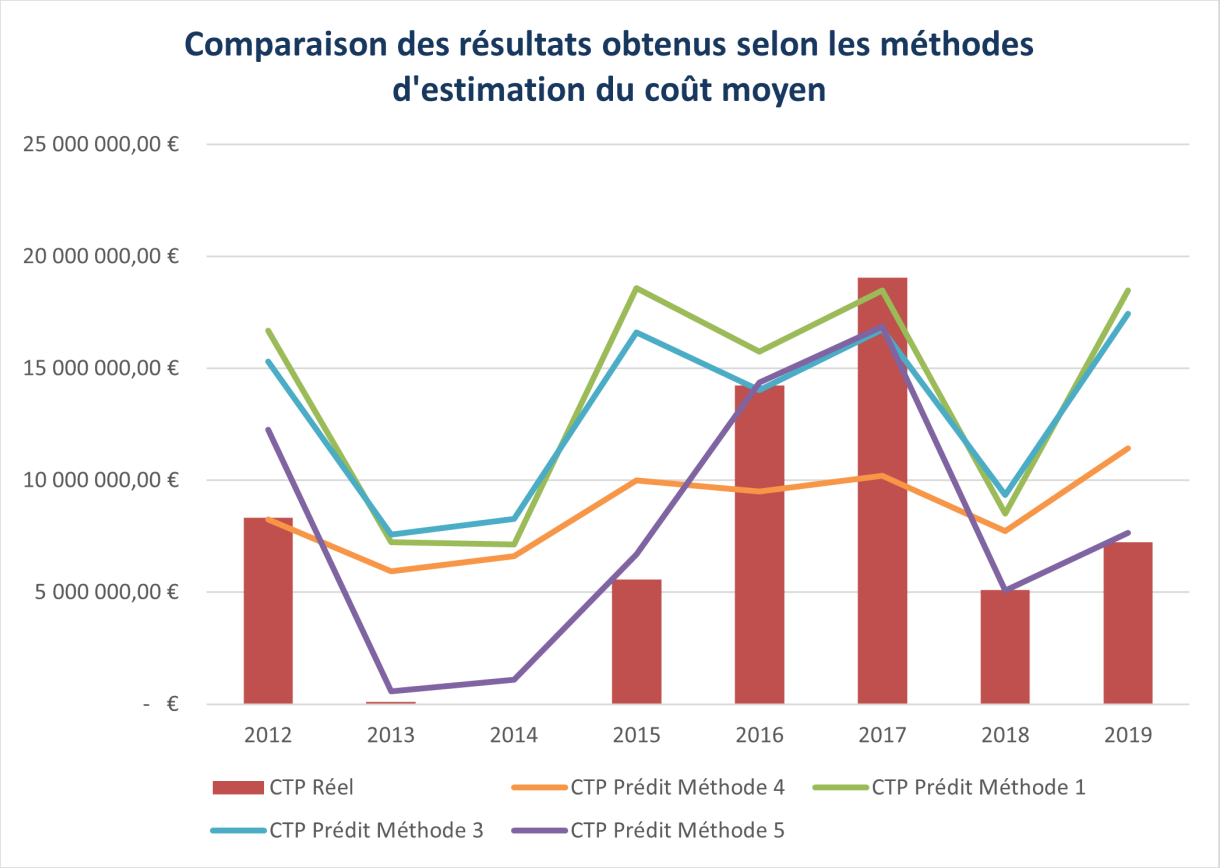

FIGURE 6.6 – Comparaison des résultats obtenus selon les méthodes d'estimation du coût moyen

Constats :

- Les méthodes 1 et 3 estiment bien les survenances fortes en Sécheresse (2016 et 2017 par exemple).
- La méthode 4 ne fait pas beaucoup varier le CTP prédit. Ceci est dû à l'utilisation du S/C (facteur entre 0 et 1).
- La dernière méthode permet de ne pas surestimer les années 2013 et 2014, Cela montre l'importance d'étudier l'intensité des évènements prédits.

Au regard des résultats, la méthode 5 parait être la méthode la plus adéquate pour estimer le CTP Sécheresse en période d'arrêtés des comptes.

6.2 Estimation de la sinistralité pour les années 2020 et 2021

Le modèle prédictif 5 étant retenu, ce dernier est à présent appliqué aux exercices de survenance 2020 et 2021. Ces années ne font pas partie des exercices de survenance de la base d'apprentissage. Donc, le modèle n'a aucune information relative à ces deux années.

Afin d'appliquer le modèle, une base de données identique (mêmes variables explicatives) à la base d'apprentissage est créée pour les années 2020 et 2021.

Le tableau ci-dessous illustre les résultats obtenus :

Exercice de survenance	**CTP Sécheresse estimée**	**CTP Sécheresse vu fin juillet 2022**
2020	2.5 M€	1.7 M€
2021	6 M€	800k€

TABLE 6.3 – Mise en application du modèle pour les années 2020 et 2021

Constats :

— L'estimation pour l'année 2020 semble se rapprocher de la charge réelle de la Sécheresse pour le portefeuille de GOC. Cette estimation vient conforter le modèle.

— L'estimation pour l'année 2021 est une valeur anticipée. En effet, les premiers arrêtés CatNat Sécheresse relatifs à l'année 2021 commencent tout juste à être publiés dans le JO. Cela viendra conforter ou non le modèle dans ses prédictions et dans les estimations obtenues.

Conclusion

Dans ce chapitre, plusieurs méthodes pour estimer la charge sinistres de la Sécheresse ont été proposées.

Les deux premières méthodes consistaient à calculer un coût moyen globale par arrêté. Ces coûts ont montré qu'ils surestimaient certains exercices de survenance. Néanmoins, ils ont su estimer les années fortes en Sécheresse.

Afin de tenir compte d'avantage de la présence de GOC sur les différentes parties du territoire, les biens assurées ont été approchés par les cotisations CatNat. En effet, ces cotisations appelées dépendent de la nature du risque assuré et reflètent, de fait, la valeur des biens. Les résultats obtenus étaient là aussi non suffisamment concluants, mais cette méthode a conduit à penser à une analogie aux modèles de simulation des réassureurs pour les tempêtes, à savoir générer un taux de destruction sur des sommes assurées.
La dernière méthode, quant à elle, a présenté l'importance d'étudier l'intensité des sécheresses. En effet, la charge liée à ce risque dépend de son intensité. De plus, cette dernière varie d'année en année. Les estimations obtenues permettent, à chaque fois, d'avoir une évaluation du risque Sécheresse pour le portefeuille de GOC pour un exercice de survenance en période d'inventaire.

Néanmoins ces estimations ne sont pas parfaites. Plusieurs perspectives et améliorations sont possibles afin de les améliorer. C'est ce qui va être présenté dans la suite "les sensibilités dans l'approche proposée".

Troisième partie

Perspectives et sensibilités de l'étude

Chapitre 7

Sensibilités dans l'approche proposée

Introduction

L'objectif de ce mémoire était d'estimer la charge sinistres de l'exercice courant de la Sécheresse sur bâtiments en période d'inventaire pour le portefeuille de GOC. L'originalité de l'étude résidait au niveau de l'approche. Pour pallier la non-connaissance de l'indicateur SWI [1] des données spécifiques au portefeuille de GOC, via l'assurance MRC sur récoltes, ont été exploitées. Ainsi, l'approche proposée se basait à la fois sur une connaissance du risque Sécheresse et son impact sur le portefeuille de l'entreprise mais aussi sur une phase de modélisation couplant l'utilisation et le développement de modèles statistiques (régression logistique, forêts aléatoires, arbres de décisions). Toutefois un regard critique doit être apportée aux travaux.
Ce chapitre est mené en trois temps :

1. La sensibilité des données utilisées,
2. La sensibilité des modèles développés,
3. La sensibilité des modèles actuels.

7.1 Les données utilisées

L'étude se proposait d'utiliser trois types de données :

1. Les données historiques des arrêtés CatNat Sécheresse,
2. Les données géologiques,
3. Les données sinistres en assurance MRC.

Tout d'abord, les données historiques des arrêtés CatNat Sécheresse sont fiables et absolument nécessaires à l'étude. Grâce à ces données, une connaissance du passé est possible et la variable à prédire a été créée.

Ensuite, l'utilité et l'importance des données géologiques (c'est l'un des critères de reconnaissance d'un arrêté CatNat Sécheresse) ont été illustrés par les résultats des modèles. Plus le sol est exposé au phénomène RGA, plus la probabilité qu'une Sécheresse survienne augmente. Ces données sont fiables et nécessaires à l'étude.

Enfin, pour pallier le second critère de reconnaissance (le SWI uniforme), des données MRC ont été utilisées. Elles apportent un certain nombre d'avantages mais aussi d'inconvénients.
Tout d'abord, ces données sont propres à GOC, donc l'entreprise n'a pas à payer pour les obtenir. Ensuite, ces données permettent de mettre en relation deux métiers du portefeuille et de créer une

1. Soil Wetness Index, critère météorologique pour la reconnaissance des arrêtés

certaine corrélation entre eux. Mais, l'avantage le plus important réside dans l'anticipation des publications.
A noter, depuis l'inventaire du premier semestre 2022, le modèle est en cours d'utilisation.
Néanmoins, ces données peuvent parfois ne pas être fiables. En effet, l'assurance MRC couvre les agriculteurs face à plusieurs aléas climatiques. Par exemple, si une terre agricole a subi de la grêle, cela aura pour conséquence la destruction de la culture et la non-déclaration des futurs aléas climatiques. Pour illustrer les propos, GOC a subi sur cet exercice 2022 un premier semestre très grêlifère, ayant pour conséquence la destruction totale de certaines cultures.

FIGURE 7.1 – Impact de la grêle sur les cultures agricoles

A la suite d'évènements de ce type, un certain nombre de sinistres Sécheresse ne seront pas disponibles. Donc, cela aura un impact négatif sur la qualité et la fiabilité des données utilisées.

D'autres types de données n'ont pas été utilisées dans ce mémoire. Les données météorologiques sont probablement importantes et pourraient être très utiles à l'étude. En effet, la survenance des Sécheresse peut dépendre des températures enregistrées ou de l'humidité constatée.
Ces éléments complémentaires auraient pu rendre le modèle plus robuste.

7.2 Sensibilités des modèles développés

Les modèles développés présentent quelques faiblesses :

— Tout d'abord, un historique de 10 ans peut sembler important, mais selon différents échanges, c'est un historique relativement court. Les modèles gagneraient (et gagneront) en performance avec l'ajout d'années complémentaires,

— la non-intégration dans le modèle du caractère social et économique de la sollicitation de la commission interministérielle,

— enfin, et le point n'est pas neutre, le modèle permet d'évaluer une charge à une date N, mais cette évaluation n'évolue plus par la suite. En effet, les données en entrée de modèles sont figées une fois l'exercice d'inventaire fini. Les déclarations de sinistres arrivant en juillet N+1, GOC utilise, à ce moment-là, différentes méthodes d'évaluations de charge sinistres, dont la méthode Chain-Ladder afin d'évaluer un CTP potentiellement plus adéquate.

A travers les faiblesses illustrées ci-dessus, d'autres éléments peuvent améliorer les évaluations :

— Tout d'abord, si la survenance des sécheresses sur le territoire de GOC est prédite, l'intensité des évènements semble une bonne piste d'amélioration (un modèle autre qu'une fonction polynomiale pourrait être plus pertinent).

— Dans l'actualité du moment, un modèle inflationniste pourrait également être une piste à intégrer.

7.3 Sensibilité du modèle Chain-Ladder [2]

Jusqu'à présent, des évaluations externes et des méthodes basées des triangles de liquidation étaient utilisées pour provisionner le risque Sécheresse. Cependant, des problématiques à la méthode des triangles ressortent :

— la première est qu'il faut nécessairement des sinistres dans le triangle en période d'inventaire. Or, la parution tardive des arrêtés CatNat fait que les sinistres n'apparaissent bien souvent qu'en N+1 après l'année d'inventaire

— la deuxième réside dans les limites de l'utilisation de la méthode Chain-Ladder.

Afin de mieux illustrer ce second point, le principe général de cette méthode est étayé ci-dessous.

Triangulation des données

Le paiement d'un sinistre peut se dérouler sur plusieurs années. L'exercice de survenance de ce dernier correspond à l'année d'origine. Une année de développement quant à elle représente une année entre l'exercice de survenance et la dernière année de paiements.
Cette méthode suppose qu'un sinistre prend au maximum $n+1$ années pour être complètement payé, c'est à dire, n années après l'exercice de survenance.
Le principe de cette approche consiste à estimer une charge sinistres des sinistres survenus lors d'une année à partir d'un triangle de liquidation.
Un triangle de liquidation (appelé aussi triangle de paiements ou de développement) $(d_{i,j})$ avec :

— i l'année de survenance, $i \in \{0, ..., n\}$

— j le délai de développement, $j \in \{0, ..., n-i\}$

— $d_{i,j}$ la charge des sinistres survenus en année i payés en j

est un triangle représenté de cette façon :

2. Cette section est basée sur le mémoire de Noémie ROSE : "Provisionnement en assurance non-vie : Utilisation de modèles paramétriques censurés" (16) et sur les notes de cours "Mathématiques des assurances non-vie" de Xavier GRUCHET. (14).

Exercice de survenance	**Année de développement**							
	0	1	...	j	...	$n-i$	...	n
0	$d_{0,0}$	$d_{0,1}$	...	$d_{0,j}$	...	$d_{0,n-i}$	...	$d_{0,n}$
...	...	...	...	...	...	...	...	
i	$d_{i,0}$	$d_{i,1}$	...	$d_{i,j}$	...	$d_{i,n-i}$		
...	...	...	...	...	...			
$n-j$	$d_{n-j,0}$	$d_{n-j,1}$	...	$d_{n-j,j}$				
...	...	...	...					
$n-1$	$d_{n-1,0}$	$d_{n-1,1}$						
n	$d_{n,0}$							

TABLE 7.1 – Représentation d'un triangle de liquidité

Il s'agit d'un triangle décumulé, c'est-à-dire, la charge $d_{i,j}$ correspond à un paiement effectué lors de l'année $i+j$. Lorsque $i+j > n$, les $d_{i,j}$ ne sont plus observables. A partir de ce triangle, le triangle cumulé peut être calculé.
Soit $D_{i,j} = \sum_{k=1}^{j} d_{i,k}$, le montant cumulé des sinistres d'année d'origine i payés en $i+j$. $(D_{i,j})$ est alors appelé le triangle cumulé.
Pour un exercice de survenance i, $D_{i,n}$ représente la charge sinistres des sinistres.
Le montant $R_i = D_{i,n} - D_{i,n-1}$ représente la provision relative à cet exercice. L'ensemble des provisions est alors calculé en sommant les provisions de chaque année, $R = \sum_{i=1}^{n} R_i$.

Développement du triangle avec Chain-Ladder

Pour pouvoir calculer les provisions, il faut développer le triangle de paiement, c'est-à-dire, estimer les montants $D_{i,j}$ lorsque $i+j > n$ inconnu jusque-là. Pour ce faire, la méthode de "Chain-Ladder" avec une vision déterministe est utilisée.
La méthode de "Chain-Ladder" se base sur l'hypothèse suivante : il existe des facteurs de développement qui ne dépendent que de l'année de développement j qui contrôlent l'évolution des charges.
Autrement dit, $\frac{D_{0,j+1}}{D_{0,j}} = \frac{D_{1,j+1}}{D_{1,j}} = ... = \frac{D_{n-j-1,j+1}}{D_{n-j-1,j}}$
Les facteurs de développement de Chain-Ladder $(f_0, f_1, ..., f_n)$ s'écrivent alors de la façon suivante :

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j} D_{i,j+1}}{\sum_{i=1}^{n-j} D_{i,j}}$$

Avec $j \in \{0, ..., n\}$.
Ensuite, pour chaque exercice de survenance $i \in \{0, ..., n\}$ et année de développement $j \in \{n-i, ..., n\}$ un estimateur de Chain-Ladder est calculé de la façon suivante :

$$\hat{D}_{i,j} = D_{i,n-i} * \prod_{k=n-i+1}^{j} f_j = \hat{D}_{i,j-1} * f_j$$

Une fois le triangle cumulé développé, les réserves $R_i = \hat{D}_{i,n}\hat{} - \hat{D}_{i,n-1}$ sont calculées.

Limites de la méthode

Les conditions pour que la méthode de Chain-Ladder soit performante sont proposées par (16) :

— le passé soit régulier,

— le futur et le présent soient peu différents du passé,

— la branche d'assurance étudiée soit peu volatile.

Or, dans le cadre des sinistres CatNat Sécheresse, le passé est loin d'être régulier. En effet, les coûts dépendent de plusieurs facteurs (météorologique, géologique, publication des arrêtés,..) très variables d'année en année. Par exemple, en 2012 il y a eu 153 Sécheresse tandis qu'en 2013, il n'y a eu que 13 Sécheresse survenues. La figure ci-dessous illustre l'évolution des charges sinistres Sécheresse entre les années de survenance 2009 et 2020.

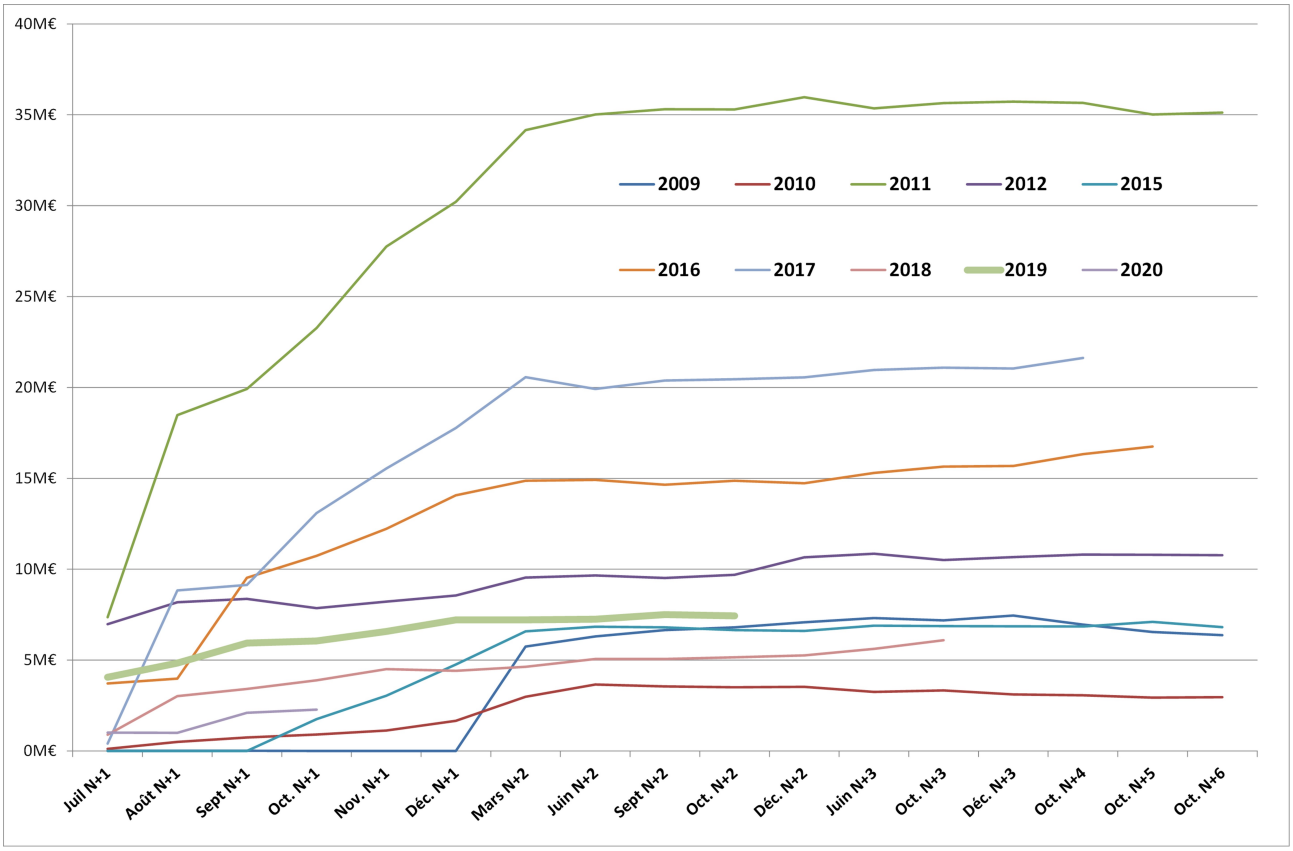

FIGURE 7.2 – Évolution des charges sinistres Sécheresse par année de survenance

Les courbes des différents exercices de survenance ne suivent pas les mêmes tendances (les coefficients Chain-Ladder associés à ces CTP sont très hétérogènes) rendant, de fait, non valide la régularité nécessaire à l'application de cette méthode.
De plus, dans un contexte de réchauffement climatique et d'inflation, la fréquence des sinistres Sécheresse et donc les coûts relatifs à ces évènements ont tendance invalide également l'hypothèse d'une régularité par rapport au passé. Enfin, le risque est relativement volatil.
De part ces 3 aspects, le provisionnement du risque Sécheresse ne semble pas totalement adapté à la méthode Chain-Ladder.

Conclusion

Ce mémoire propose une approche originale pour provisionner le risque Sécheresse en période d'inventaire grâce à l'anticipation des publications des arrêtés.
Cependant, des points d'amélioration se dégagent assez nettement, tant dans la base de données nécessaire à la projection (un historique court de 10 ans, des données météorologiques non intégrées, ...) que dans les modèles eux-mêmes (CTP figé, non intégration du caractère social et économique de la commission ministérielle, ...).

Chapitre 8

Perspectives face au risque Sécheresse

Introduction

Face à la forte progression des évènements Sécheresse et de leurs impacts sur les compagnies d'assurance, de réassurance et les pouvoirs publics, des questions se posent :

— l'aléa Sécheresse peut-il sortir du régime d'indemnisation des CatNat ?

— des démarches de reconnaissance pour une indemnisation plus rapide seront-elles proposées ?

— existe-t-il une politique efficace de prévention contre ce risque ?

Pour répondre à ces questions, la Cour des comptes [1] a été saisie par le CEC [2] des politiques publiques de l'Assemblée Nationale.
A l'issue de son enquête, la Cour a publié un rapport en février 2022 dans lequel elle présente ses remarques et recommandations quant à ce phénomène.
De plus, une réforme sur la loi nommée "ELAN" a été introduite.

En parallèle, le 29 juillet 2022, l'ordonnance n°2022-1075 a été circularisée. Sous l'égide du Ministère de l'Agriculture et de la Souveraineté Alimentaire, les subventions prévues dans le cadre de l'assurance Multi-Risques Climatiques sur Récoltes sont en cours de réforme. L'idée et l'objectif de cette évolution est d'inciter davantage les exploitants agricoles à se protéger face à la croissance et l'intensité des évènements climatiques de toute nature.

Ce chapitre va tout d'abord s'intéresser aux recommandations de la haute juridiction. En deuxième lieu, il présentera les perspectives possibles en lien avec la loi ELAN.

8.1 Les recommandations de la Cour des comptes

Pour la haute juridiction, le phénomène de retrait gonflement des sols argileux est devenu universel. Il touche toutes les parties du territoire métropolitain. De plus, la durée des évènements devient de plus en plus longue. La Cour estime alors que la Sécheresse géotechnique n'est pas assimilable à une catastrophe naturelle. Elle *"invite l'Etat à réexaminer la qualification de catastrophe naturelle à donner au phénomène"*, tout en estimant que l'indemnisation des sinistres Sécheresse *"gagnerait à ne plus relever du régime CatNat"*.
Néanmoins, malgré ces remarques, la Cour n'a pas formulé de recommandations portant sur l'indemnisation des sinistres car elle estime *"qu'une clarification des axes de réformes en cours est nécessaire avant la refonte du régime d'indemnisation des CatNat."*
De plus, la Cour a écarté l'idée de recours à un système purement assurantiel. Selon elle, cela engen-

1. Cour des comptes. Sols argileux et catastrophes naturelles. 2022. (5)
2. Comité d'Evaluation et de Contrôle

drerait des inégalités entre les assurés et limiterait l'assurabilité de ce risque.

Constat :
L'écart du recours à un système purement assurantiel est un coup dur pour les compagnies d'assurances. Ces dernières militaient pour avoir une garantie spécifique (avec des critères propres à l'aléa) dans les produits Multi-Risques Habitation (MRH).
Une réforme du régime serait également la bienvenue. Une reconnaissance plus rapide pourrait améliorer la visibilité des assureurs quant au provisionnement à faire pour ce risque.

D'un autre côté, la Cour a insisté sur la nécessité de renforcer les mesures de prévention contre ce phénomène. Elle suggère notamment d'instaurer un dispositif de prévention mieux adapté avec de nouvelles mesures. Elle propose tout d'abord de renforcer la connaissance du risque Sécheresse auprès des particuliers et de les sensibiliser face aux conséquences de ce phénomène. Selon elle, il est important qu'un acheteur de bien immobilier puisse connaître la nature du sol de l'habitation et des risques futurs qu'il encourt.
Dans ce cadre, la cour propose de *"Mettre en place un dispositif de contrôle et de sanction"* contre les constructions neuves en zones argileuse, ainsi que *"d'intégrer le phénomène RGA dans l'état des risques naturels et technologiques (ERNT) accessible à tout acquéreur ou locataire d'un bien immobilier"*.

Constat :
L'amélioration des mesures de prévention contre la Sécheresse serait bénéfique pour les compagnies d'assurances. Cela engendrerait moins de charge sinistres. De plus, si le propriétaire ou le locataire d'une habitation veut s'assurer, l'assureur peut lui demander les études de sol et de construction réalisées. Cela aiderait l'assureur à avoir une meilleure connaissance du risque.

Enfin, la Cour des comptes propose le développement et l'accélération des projets de recherches pour l'obtention de mesures de remédiation applicables aux anciennes constructions qui sont exposées au phénomène RGA. L'une des idées consiste à construire des puits. Lors d'un évènement Sécheresse, le puits qui est relié à la nappe phréatique viendrait arroser la terre.

8.2 La réforme de la loi ELAN [3]

La loi ELAN [4] est une loi entrée en vigueur le 23 décembre 2018 qui prévoyait de nombreux textes d'application sur les constructions.
Dans sa nouvelle réforme, cette loi vise à faciliter les démarches de reconnaissance de la Sécheresse et accélérer l'indemnisation des sinistrés. Cette loi fait partie des travaux législatifs en cours. Le gouvernement devra d'ailleurs remettre dans les prochains mois, un rapport, tant sur les pistes à privilégier pour améliorer les mesures de prévention, que sur l'élaboration d'un régime d'indemnisation spécifique pour les sinistres de ce péril.

La création d'un régime spécifique à la Sécheresse semble être une bonne perspective et pourrait permettre la création d'une cotisation spécifique adaptée à l'aléa (ce qui n'est pas le cas aujourd'hui avec la cotisation globale CatNat).

3. A. ABBADIE. L'argus de l'assurance. Vers un régime spécifique pour la sécheresse ?. 2022. (1)
4. Évolution du Logement de l'Aménagement et du Numérique

Conclusion

Le régime d'indemnisation des CatNat a montré quelques limites après 40 ans d'existence. Des évolutions semblent nécessaires afin d'alléger les impacts sur les compagnies d'assurance, de réassurance et les pouvoirs publics.
Les travaux de réformes en cours laissent entrevoir de meilleures perspectives dans la maîtrise et la gestion de ce risque.

En parallèle de ces éléments, l'ordonnance n°2022-1075 récemment circularisée propose des modifications dans le mode de distribution des contrats d'assurance Climatiques sur Récoltes. Ces évolutions ont pour objectif affiché de protéger davantage d'exploitations agricoles.
En lien avec les travaux réalisés, cette réforme laisse entrevoir, pour le modèle mis en place, une plus grande robustesse dans les données MRC utilisées et de fait, de meilleures évaluations à venir.

Conclusion Générale

Les incertitudes liées au changement climatique de la planète et l'impact récurrent des sécheresses sur les résultats de GOC ont conduit à engager une réflexion quant aux méthodes utilisées pour estimer la charge sinistres de ce péril en période d'inventaire.

Jusqu'à présent, des évaluations externes et des méthodes basées sur des triangles de liquidation étaient utilisées pour provisionner le risque Sécheresse.
Le fait de ne pas maîtriser les évaluations externes était une problématique pour GOC. De même, l'estimation par la méthode des triangles génère également des inconvénients :

- Le premier se trouve dans les limites de l'utilisation de la méthode Chain-Ladder. Le risque est relativement volatil et les coefficients sont peu stables (comme présenté dans le chapitre 7.3.).
- Le deuxième réside dans la nécessité d'avoir des sinistres dans le triangle en période d'inventaire. Or, la parution tardive des arrêtés CatNat fait que les sinistres n'apparaissent bien souvent qu'en N+1 après l'année d'inventaire. En effet, et pour rappel, pour qu'un arrêté CatNat Sécheresse figure au JO, 2 critères doivent être validées par la Commission interministériel :
    1. Le premier est un critère prédisposition géologique (disponible à tout moment),
    2. Le second est un critère de déclenchement météorologique (le SWI [5] uniforme, dévoilé uniquement une fois l'arrêté paru)

L'objectif de ce mémoire était de proposer de nouvelles approches plus adaptées au risque.

Le Groupe Groupama commercialise aujourd'hui une assurance Multi-Risques Climatiques sur Récoltes (il en est d'ailleurs l'acteur principal avec 45% des cotisations assurantielles en France). Ce produit permet aux exploitants de bénéficier (sans aucune publication au JO à la différence des bâtiments) d'une indemnisation en cas de perte de rendement sur une culture agricole, dès lors que cette dernière ait été altérée par un évènement climatique, comme la sécheresse par exemple.

Ainsi, l'originalité de l'étude résidait non seulement dans l'utilisation de nouveaux types de modèles, mais, avant tout, dans le remplacement du second critère de déclenchement (le SWI uniforme), par les données sinistres en MRC.

Trois étapes ont été nécessaires à la définition de la nouvelle approche :

- Bien évidemment, la création d'une base d'étude adéquate et cohérente (pas de valeur à nulle par exemple pour les zones sans culture agricole) était primordiale,
- Ensuite, trois modèles statistiques [6] ont été élaborés et optimisés afin de définir une probabilité qu'un arrêté CatNat Sécheresse pour un code postal survienne,
- Enfin, différentes méthodes d'estimation d'un coût moyen ont été élaborées [7].

5. Soil Wetness Index, critère météorologique pour la reconnaissance des arrêtés
6. Régression logistique, arbre de décision, forêt aléatoire
7. C'est probablement un des points qui mériterait une analyse plus approfondie

Aux vus des différents éléments apportés dans ce rapport, et même si des points mériteraient d'être améliorés [8], cette nouvelle approche répond globalement bien à l'évaluation d'un CTP en période d'inventaire pour un risque très erratique.

Le modèle est en cours d'utilisation depuis l'arrêté des comptes du premier semestre 2022. L'estimation de la charge sinistres sur les exercices plus récents 2020 et 2021 semble cohérente avec les évaluations actuelles [9].

Ce mémoire s'est aussi intéressé à de possibles évolutions de ce risque, notamment à travers les recommandations de la Cour des comptes et la réforme en cours de la loi ELAN qui probablement pourrait conduire à externaliser la Sécheresse du régime CatNat.

Au-delà des points d'améliorations mentionnés dans ce rapport, la perspective d'une élaboration tarifaire spécifique constituerait un prolongement intéressant à ces travaux.

8. L'intégration de données météorologiques ou bien le caractère socioéconomique des arrêtés dans le calibrage des modèles permettrait probablement une amélioration des estimations

9. Chain-Ladder sur 2020 et évaluation externe sur 2021

Table des figures

Liste des tableaux

Bibliographie

Articles

[1] A.ABADIE. Vers un régime spécifique pour la Sécheresse ?. L'argus de l'Assurance. 2022

[2] N. BONNEFOY. Rapport d'information fait au nom de la mission d'information sur la gestion des risques climatiques et l'évolution de nos régimes d'indemnisation. 2019

[3] BRGM. Cartographie de l'aléa retrait-gonflement des sols argileux dans le département de la marne. 2008

[4] CAISSE CENTRALE DE RÉASSURANCE. Le régime d'indemnisation des catastrophes naturelles. 2011

[5] COUR DES COMPTES. Sols argileux et catastrophes naturelles. 2022

[6] FRANCE ASSUREURS. L'assurance des évènements naturels en 2020. 2022

[7] GEOCONFLUENCES. Région agricole, petite région agricole (au sens statistiques, en France). 2022

[8] GROUPAMA D'OC. Rapport sur la Solvabilité et la Situation Financière. 2021

[9] MISSION RISQUES NATURELS (MRN). La Sécheresse géotechnique. 2018

[10] ORGANISATION MÉTÉOROLOGIQUE MONDIALE. Atlas de la mortalité et des pertes économiques dues à des phénomènes météorologiques, climatiques et hydrologiques extrêmes (1970-2019). 2021

[11] STATISTA. Les catastrophes naturelles dans le monde. 2021

[12] WORLD ECONOMIC FORUM. Global Risks Report. 17th Edition. 2022

Polycopiés de Cours

[13] P. AILLIOT. Notes de cours sur les méthodes de régression. EURIA. Université de Brest. 2016

[14] X. GRUCHET. Notes de cours sur les mathématiques de l'assurance non-vie. EURIA. Université de Brest. 2021

[15] L. ROUVIERE. Cours Régression Logistique avec R. Université Rennes 2.

Mémoire

[16] N. ROSE. Provisionnement en assurance non-vie : Utilisation de modèles paramétriques censurés. Mémoire ISUP. 2009

Ouvrages

[17] P-A. CORNILLON et E. MATZNER-LOBER. Régression Théorie et applications. Édition Presses Universitaires de Rennes. 2007

[18] P-A. CORNILLON et al. Statistiques avec R : 2ème version augmentée. Édition Presses Universitaires de Rennes. 2010

[19] R. GENUER et J-M. POGGI. Les forêts aléatoires avec R. Édition Presses Universitaires de Rennes. 2019

Sites Internet

[20] B. BOEHMKE et B. GREENWELL. Hands-On Machine Learning with R. https://bradleyboehmke.github.io/HOML

[21] T. GIRAUD et H. PECOUT. Cartographie avec R. https://rcarto.github.io/cartographie_avec_r/index.html#fnref1

[22] J. LARMARANGE. Régression logistique binaire, multinomiale et ordinale. https://larmarange.github.io/analyse-R/regression-logistique.html

[23] MACHINE LEARNING GOOGLE COURSES. Classification : ROC Curve and AUC. 2022. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=fr

[24] MÉTÉO FRANCE. Descriptif du SWI uniforme CATNAT. MÉTÉO FRANCE. Descriptif du SWI uniforme CATNAT. https://donneespubliques.meteofrance.fr/client/document/doc_swi_catnat_277.pdf

[25] B. NATANELIC. ML : Précision, F1-Score, Courbe ROC, que choisir ?. https://beranger.medium.com/ml-accuracy-pr%C3%A9cision-f1-score-courbe-roc-que-choisir-5d4940b854d7